Documenta Mathematica

Journal der Deutschen Mathematiker-Vereinigung

Extra Volume ICM 1998

Proceedings

of the

International Congress

of

Mathematicians

Berlin 1998

August 18 – 27

Vol. III: Invited Lectures

Documenta Mathematica, Journal der Deutschen Mathematiker-Vereinigung, publishes research manuscripts out of all mathematical fields and is refereed in the traditional manner.

Documenta Mathematica is published on the World Wide Web under the address:

http://www.mathematik.uni-bielefeld.de/documenta

Manuscripts should be submitted as TEX files by e-mail to one of the editors. Hints for manuscript preparation can be found under the above WWW-address.

This volume was produced electronically, using various dialects of TEX, at the Fakultät für Mathematik, University of Bielefeld, Germany.

ISSN 1431-0635 Print ISSN 1431-0643 Internet

Print: Publishing On Demand by Geronimo GmbH, D-83026 Rosenheim, Germany
http://www.geronimo.de

ICM 1998

Invited Forty-Five Minute Lectures

at the Section Meetings

Contents of Volumes II and III

In case of several authors, Invited Speakers are marked with a *.
The author index is at the end of each of these two volumes.

SECTION 4. ALGEBRAIC GEOMETRY

SECTION 5. DIFFERENTIAL GEOMETRY AND GLOBAL ANALYSIS

SECTION 6. TOPOLOGY

SECTION 10. PARTIAL DIFFERENTIAL EQUATIONS

SECTION 11. MATHEMATICAL PHYSICS

SECTION 14. MATHEMATICAL ASPECTS OF COMPUTER SCIENCE

SECTION 15. NUMERICAL ANALYSIS AND SCIENTIFIC COMPUTING

SECTION 16. APPLICATIONS

Section 17. Control Theory and Optimization

Section 18. Teaching and Popularization of Mathematics

Section 19. History of Mathematics

Section 10

Partial Differential Equations

In case of several authors, Invited Speakers are marked with a *.

Vortices in Ginzburg-Landau Equations

Fabrice Bethuel

Abstract. GL models were first introduced by V.Ginzburg and L.Landau around 1950 in order to describe superconductivity. Similar models appeared soon after for various phenomena: Bose condensation, superfluidity, non linear optics. A common property of these models is the major role of topological defects, termed in our context vortices.

In a joint book with H.Brezis and F.Helein, we considered a simple model situation, involving a bounded domain Ω in R^2, and maps v from Ω to R^2. The Ginzburg-Landau functional, then writes

$$E_\epsilon(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{4\epsilon^2}\int_\Omega (1-|v|^2)^2$$

Here ϵ is a parameter describing some characteristic lenght. We are interested in the study of stationary maps for that energy, when ϵ is small (and in the limit ϵ goes to zero). For such map the potential forces $|v|$ to be close to 1 and v will be almost S^1-valued. However at some point $|v|$ may have to vanish, introducing defects of topological nature, the vortices. An important issue is then to determine the nature and location of these vortices.

We will also discuss recent advances in more physical models like superconductivity, superfluidity, as well as for the dynamics: as previously the emphasis is on the behavior of the vortices.

1991 Mathematics Subject Classification: 35J20, 35J55, 35Q99, 35Q55, 35B98
Keywords and Phrases: Ginzburg-Landau equations, superconductivity, vorticity, evolution equations

1 Introduction

Ginzburg-Landau functionals were introduced around 1950 by V.Ginzburg and L.Landau in order to model energy states of superconducting materials and their phase transitions. Related functionals appeared soon therafter in various fields as superfluidity, Bose condensation, nonlinear optics, fluid mechanics and particle physics. A common feature of these models is that they involve non convex

potentials, which allow the existence of topological defects for stationary states: here we will mainly focus on two-dimensional situations, where theses defects are often termed vortices. In recent years, very importants efforts have been devoted to their study from a mathematical point of view: we will try here to survey parts of these works.

We begin with a simple model, which was studied extensively, in particular in a joint book with H.Brezis and F.Helein [BBH]. Consider a smooth bounded domain in R^2 (for instance a disk), and complex valued functions v on Ω (i.e maps v from Ω to R^2). The simplest possible Ginzburg-Landau functional then takes the form, for these functions

$$E_\epsilon(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{4\epsilon^2}\int_\Omega (1-|v|^2)^2 \, .$$

Here ϵ is a parameter describing some characteristic lenght and we will mainly be interested in the case ϵ is small and in the limit ϵ tends to zero. The potential $V(v) = \epsilon^{-2}(1-|v|^2)$ forces $|v|$, for critical maps for E_ϵ to be close to 1 and therefore, stationnary (or low energy) maps will be almost S^1-valued. However, at some points $|v|$ may have to vanish, introducing "defects".

To have a well-posed mathematical problem, we prescribe next Dirichlet boundary conditions: let g be a smooth map from $\partial\Omega$ to S^1, and prescribe v to be equal to g on $\partial\Omega$. Therefore we introduce the Sobolev space

$$H^1_g(\Omega; R^2) = \{v \in H^1(\Omega; R^2),\, v = g \text{ on } \partial\Omega\} \, .$$

It is then easy to verify that E_ϵ is a C^∞ functional on H^1_g, and that its critical points verify the Ginzburg-Landau equation

$$\Delta v = \frac{1}{\epsilon^2} v(1-|v|^2) \text{ on } \Omega, \quad v = g \text{ on } \partial\Omega. \tag{1}$$

Standard elliptic estimate show that, any solution to (1) is smooth, that

$$|v| \leq 1 \text{ on } \Omega \quad \text{(maximum principle)}, \tag{2}$$

$$|\nabla v| \leq \frac{C}{\epsilon} \text{ on } \Omega \quad \text{for C, some constant depending on } g, \tag{3}$$

$$\frac{1}{4\epsilon^2}\int_\Omega (1-|v|^2)^2 \leq C \, , \quad \text{provided } \Omega \text{ is starshaped.} \tag{4}$$

Since E_ϵ is strictly positive, one easily verifies that it achieves its infimum k_ϵ on H^1_g and hence (1) possesses minimizing solutions (not necessarily unique). We will denote u_ϵ these solutions.

2 Asymptotic analysis of minimisers

The winding number d of g (as map from $\partial\Omega$ to S^1) is crucial in this analysis, forcing, when $d \neq 0$, vortices to appear.

2.1 THE CASE $d = 0$.

In this case, there exists ψ from $\partial\Omega$ to R such that $g = \exp i\psi$. Next let φ_* be the solution of $\Delta\varphi_* = 0$ on Ω, $\varphi_* = \psi$ on $\partial\Omega$ and consider $u_* = \exp i\varphi_*$. Clearly u_* is S^1-valued, so that

$$E_\epsilon(u_*) = \frac{1}{2}\int_\Omega |\nabla u_*|^2 = \frac{1}{2}\int_\Omega |\nabla\varphi_*|^2$$

is bounded independently on ϵ. Hence k_ϵ remains bounded as $\epsilon \to 0$. It is that easy to show that $u_\epsilon \to u_*$ in H^1. Finally in [BBH2] we carried out more refined asymptotics, in particular

$$\|u_* - u\|_{L^\infty} \leq C\epsilon^2.$$

2.2 THE CASE $d \neq 0$.

We may assume, for instance $d > 0$. In this case there are no maps in H^1_g which are S^1-valued (the fact that there are no continuous S^1-valued maps reduces to standard degree theory). In particular $k_\epsilon \longrightarrow +\infty$, and we are facing a singular limit. Since u_ϵ is smooth, the topology of the boundary data forces u_ϵ to vanish somewhere in Ω. The points where u_ϵ vanishes play an important role: the Dirichlet energy will concentrate in there neighborhood, accounting for the divergence of k_ϵ. In [BBH], we established

THEOREM 1 *i) There exists a constant $C > 0$ depending only on g such that*

$$|k_\epsilon - \pi d|\log\epsilon|| \leq C\,,\ \forall\ 0 < \epsilon < 1\,. \tag{5}$$

ii) The map u_ϵ has exactly d zeroes, provided ϵ is sufficiently small (these result relies on a work by P.Baumann, N.Carlson and D.Philips [BCP]) .
iii) There exists exactly d points $a_1, a_2, \cdots, a_d$ in Ω such that up to a subsequence , $\epsilon_n \to 0$,

$$u_{\epsilon_n} \longrightarrow u_*, \text{ on any compact subset of } \Omega\backslash\bigcup_{i=1}^{d}\{a_i\}\,,$$

where

$$u_* = \prod_{i=1}^{d}\frac{z - a_i}{|z - a_i|}\exp i\varphi \ (\varphi \text{ being a harmonic function}).$$

In particular, the winding number around each singularity is $+1$.
iv) The configuration a_i is not arbitrary, but minimizes on $\Omega^d\backslash\Delta$ (where Δ denotes the diagonal) a renormalized energy which has the form

$$W_g(a_1, \cdots, a_d) = \pi\sum_{i\neq j}\log|a_i - a_j| + \text{ boundary conditions.} \tag{6}$$

v) The energy has the expansion, as $\epsilon \to 0$

$$k_\epsilon = \pi d|\log\epsilon| + W_g(a_1, \cdots, a_d) + d\gamma_0 + o(1)$$

where γ_0 is some absolute constant.

REMARKS 1) Theorem 1 was established in [BBH] under the additional assumption that Ω is starshaped. This assumption was removed by M.Struwe [Str] (see also Del Pino-Felmer [DF]).
2) Similar results have been obtained by André and Shafrir, when the potential depends also on x, [AS], in [BR] for the abelian Higgs models, and in [HJS] for a self-dual model.
3) Hardt and Lin have studied in [HL] a different singular limit problem, with the same renormalized energy.
4) A three dimensional analog was studied by Rivière in [R].

3 ASYMPTOTICS FOR NON MINIMIZING SOLUTION

A similar analysis can be carried out for solution which are not necessarily minimizing. Assume Ω is starshaped. Then, we have, [BBH], for v_ϵ solution to (1):

THEOREM 2 *i) There exists some constant $C > 0$, such that, for $0 < \epsilon < 1$*

$$E(v_\epsilon) \leq C(|\log \epsilon| + 1)\,.$$

ii) there exists a subsequence ϵ_n, l points $a_1, \cdots, a_l$ and l integer $d_1, \cdots, d_l$ such that

$$v_{\epsilon_n} \longrightarrow v_* = \prod_{i=1}^{l} \left(\frac{z - a_i}{|z - a_i|}\right)^{d_i} \exp i\varphi, \text{ where } \varphi \text{ is harmonic.}$$

iii) The configuration $(a_i,\, d_i)$ is critical for the renormalized energy.

Note that an important difference between minimizing and non-minimizing solutions is that, for the later one, the multiplicity of vortices has not to be $+1$, and the vortices of opposite degree might coexist.

4 THE EXISTENCE PROBLEM

In view of Theorem 2, a natural question is to determine whether non-minimizing solutions do really exist, and if one is able to prescribe the multiplicity of the vortices. We begin with an elementary example.

4.1 AN EXAMPLE:

Take $\Omega = D^2$ and $g(\theta) = \exp id\theta$ (here $(r,\, \theta)$ denote polar coordinate). In view of the symmetries, one can find a solution $v(r,\, \theta)$ of the form $v_d(r,\, \theta) = f_d(r) \exp id\theta$, where f_d verifies the ODE

$$r^2 f^{\text{“}} + r f^{'} - d^2 f + \frac{1}{\epsilon^2} r^2 f(1 - f^2) = 0\,, \quad f(0) = 0\,, \quad f(1) = 1\,.$$

Computing the energy of these solutions, one sees that they are of order $\pi d^2 |\log \epsilon|$: hence, if $|d| \geq 2$, and ϵ is sufficiently small they are non minimizing. [In the case $d = 1$, v is minimizing thanks to results by P. Mironescu [Mi] and Pacard and Rivière [PaR]].

Actually, for large d, there are much more solutions. Indeed, the Morse Index of the solution v_d is of order $|d|^2$, for large d (see [AB1], [BeH]). Therefore, using symmetries and the index theory of Faddell and Rabinowitz [FR] (a Lyusternik-Schnirelmann theory in the presence of compact group actions), one obtains the existence of at least $\mu_0|d|^2$ orbits of solutions, for large d, where μ_0 is some positive constant (the orbit of a solution v is the set $\{\exp(-id\alpha)v(\exp i\alpha z),\ \alpha \in [0,\ 2\pi[\ \}$).

4.2 Variational methods

A complete Morse theory for (1) has yet to be constructed. In view of (6), one might expect that the level sets for E_ϵ are related to the level sets of W_g on $\Sigma = \Omega_d \setminus \Delta$, and hence that the topologie of Σ might yield solution for (1). This idea was introduced in [AB1], and then extended by Zhou and Zhou [ZZ]: they proved that (1) has at least $|d|+1$ distinct solutions, for sufficiently small ϵ. They are using crucially the fact that the cuplenght of Σ is (at least), $|d|-1$, a result due to V. Arnold [Ar].

We conjecture actually that the number of solutions is much higher. In order to find solutions with vortices of higher multiplicity, one has also to take into account vortices of opposite charges and also the fact that they might annihilate. For that reason, $\Omega^d \setminus \Delta$ is no longer the good model, and one has to turn to spaces as studied by D.Mc. Duff [McD].

Remark: Another construction of (stable) solutions has been introduced in [Li1].

5 Superconductivity

We turn now to the original model for superconductivity, as introduced by Ginzburg and Landau. Here Ω represents a superconducting sample, h_{ex} denotes the external applied magnetic field. The functional to minimize is now

$$F_\epsilon(u, A) = \frac{1}{2}\int_\Omega |\nabla_A u|^2 + |dA - h_{ex}|^2 + \frac{1}{4\epsilon^2}\int_\Omega (1-|u|^2)^2\,.$$

Here $A = A_1 dx_1 + A_2 dx_2$ is a connection accounting for electromagnetic effects, and u represents a condensated wave function for Cooper pairs of electrons, the carrier of superconductivity. In the above renormalized units, $|u|^2$ represents the density of Cooper pairs, so that if $|u| \simeq 0$ the sample is in the normal state, whereas if $|u| \simeq 1$ the material is in the superconducting state. We will see that for certain applied fields h_{ex}, the two states may coexist in the same sample (phase transition of second order). This model leads therefore to many interesting mathematical questions, often related to physical experiments.

5.1 Non simply connected domains

In this case, permanent currents have been observed, even when $h_{ex} = 0$. Jimbo, Morita and Zhai [JMZ], Rubinstein and Sternberg [RS] and Almeida [Al1] have

related this fact to the existence of configurations minimizing the energy in a topological sector. The threshold energy between different sectors is established in [Al2] and corresponds precisely to the energy of a vortex.

When the external field is non zero interesting phenomena occur (the Little-Parks effect), which have been studied in particular by Berger and Rubinstein ([BgR]).

5.2 Critical fields

Suppose ϵ is small, and let Ω be an arbitrary domain. For $h_{ex} = 0$, the minimizing solution is clearly (up to a gauge transformation) $u = 1$, $A = 0$. It is observed that, until h_{ex} reaches a critical field H_{c_1}, the minimizing solution has no vortex (called a Meissner solution). For $h_{ex} > H_{c_1}$, vortices appear, and their number increases with h_{ex}. Finally, for $h_{ex} > H_{c_2}$, another critical field, superconductivity dissapears, and the minimizing solution is $u = 0$.

Stable solutions near H_{c_1} have been thoroughly investigated by S. Serfaty [S1, S2]. In particular the location of the vortices is determined, and it is proved that many branches of solutions corresponding to various numbers of vortices, coexist at the same time. For larger fields, homogenized equations for the vortex distribution have been proposed and studied (see for instance Chapman, Rubinstein and Schatzman [CRS]).

Finally very precise estimates have been obtained in the one dimensional case by C. Bolley and B. Helffer (see[BoH]), for different critical fields and values of ϵ.

6 Evolution equations

Various evolution equations corresponding to the Ginzburg-Landau system have been studied. For the heat-flow equation related to (1), Lin [Li2] has shown that the vortices evolve according to the gradient flow of the renormalised energy (see also [JS]), in a suitable renormalized unit of time. The Schrödinger equation (termed also Gross-Pitaevskii equation)

$$i\frac{\partial u}{\partial t} = \Delta u + u(1 - |u|^2) \tag{7}$$

is of special importance, since it appears as a model for superfluids, Bose condensation, nonlinear optics. It is also related to fluid mechanics, because if $u = \rho \exp i\varphi$, then $\nabla\varphi$ can be interpretated as the velocity in a compressible Euler equation, ρ^2 being the density (with a suitable choice for the pressure).

The dynamics of vortices (on bounded domains) was derived by Colliander and Jerrard [CJ], as the simplectic gradient for the renormalized energy (see also [LX]).

When the domain is R^2, Ovchinnikov and Sigal [OS1] have shown that when the initial data has two vortices of the same sign (and hence infinite GL energy), radiation takes place and the vortices repulse. The existence and behavior of travelling waves solutions to (7) has been widely considered in the physical litterature (see for instance Jones, Putterman and Roberts [JPR], Pismen and

Nepomnyashchy [PN], Josserand and Pomeau [P]). This solutions have the forme $u(x,t) = U(x_1 - ct\,,\, x_2)$ where U is a function on R^2. For $0 < c^2 < 2$, non constant finite energy solutions exists (rigourous proofs are provided in [BS1], [BS2]). When c is small, these solution possess two vortices with degrees $+1$ and -1, the distance separating the vortices is proportional to the inverse of the speed c. The limiting speed $\sqrt{2}$ represents the speed of sound (see [OS2], also for the role of Cherenkov radiation). Stability of these travelling waves has been studied in the physical literature: mathematical proofs are still to be provided as well as for the three dimensional case (vortex rings).

REFERENCES

[Al1] L. Almeida, *Topological sectors for Ginzburg-Landau energies*, preprint 1997.

[Al2] L. Almeida, *Transition energies for Ginzburg-Landau functionals*, preprint 1998.

[Ar] V. Arnold, *The cohomology ring of the colored bride group*, Math Notes, 5 (1969), 138-140 [translation from Mat.Zametski, 5 (1969)].

[AB1] L. Almeida and F. Bethuel, *Multiplicity results for the Ginzburg-Landau equation in presence of symetries*, Houston J. of Math., 23 (1997), 733-764.

[AB2] L. Almeida and F. Bethuel, *Topological methods for the Ginzburg-Landau equation*, J.Math. pures et Appliquées (1998).

[AS] N. André et I. Shafrir, *Minimization of the Ginzburg-Landau functional with weight*, C.R. Acad. Sci. Paris 321 (1995), 999-1004.

[BBH] F. Bethuel, H. Brezis and F. Helein, *Ginzburg-Landau vortices*, Birkhäuser, (1994).

[BBH2] F. Bethuel, H. Brezis and F. Helein, *Asymptotics for a minimization of a Ginzburg Landau functional*, Calc. Var. and PDE, 1 (1993), 123-148.

[BCP] P. Baumann, N. Carlson and D. Phillips, *On the zeros of solutions to Ginzburg-Landau type systems*, SIAM J. Math. Anal., 24 (1993), 1283-1293.

[BeH] F. Bethuel and B. Helffer, *Stability of radial solutions to Ginzburg-Landau equations*, preprint 1998.

[BoH] C. Bolley and B. Helffer, *Proof of De Gennes formula for the superheating field in the weak kappa-limit* , Annales de l'IHP, Analyse Non Linéaire 14 (1997), 597-614.

[BR] F. Bethuel and T. Rivière, *A minimization problem related to superconductivity*, Annales IHP, Analyse Non Linéaire, 12 (1995), 243-303.

[BgR] J. Berger and J. Rubinstein, *Topology of the order parameter in the Little Parks experiment*, Phys. Rev. Lett. 75 (1995), 320-322.

[BS1] F. Bethuel and J.C. Saut, *Travelling waves for the Gross-Pitaevskii equation I*, to appear in Annales IHP, Phys. Théo.

[BS2] F. Bethuel and J.C. Saut, *Travelling waves for the Gross-Pitaevskii equation II*, preprint 1998.

[CJ] J. E. Colliander and R. L. Jerrard, *Ginzburg-Landau vortices: weak stability and Schrödinger equation dynamics*, preprint 1997.

[DF] M. Del Pino and P. Felmer, *Local minimizers of the Ginzburg-Landau energy*, to appear.

[FR] E. Fadell and P. Rabinowitz, *Generalized cohomological index theories*, Invent. Math., 45 (1978), 139-174.

[JMZ] S. Jimbo, Y. Morita and J. Zhai, *Ginzburg-Landau equations and stable steady state solutions in a non-trivial domain*, preprint.

[JP] C. Josserand and Y. Pomeau, *Generation of vortices in a model superfluid He^4 by the KP instability*, Europhysics Letters, 30 (1995), 43-48.

[JPS] C. Jones, S.J. Putterman and P.H. Roberts, *Motion in Bose condensate V*, J. Phys. A., 19 (1986), 2991-3011.

[JS] R.L. Jerrard and H.M. Soner, *Dynamics of Ginzburg-Landau vortices*, to appear in Arch. Rat. Mech. Anal. (1998).

[HJS] M.C. Hong, J. Jost and M. Struwe, *Asymptotic limits of a Ginzburg-Landau type functional*, to appear.

[HL] R. Hardt and F.H. Lin, *Singularities for p-energy minimizing unit vectorfields on planar domains*, Calc. Var. 3 (1995), 311-341.

[Li1] F.H. Lin, *Solutions of Ginzburg-Landau equations and critical points of the renormalized energy*, Annales IHP, Analyse Non Linéaire, 12 (1995), 599-622.

[Li2] F.H. Lin, *Some dynamical properties of Ginzburg-Landau vortices*, Comm. Pure. Appl. Math., 49 (1996), 323-359.

[LX] F.H. Lin and J.X. Xin, *On the incompressible fluid Limit and the vortex motion Law of the non-linear Scrödinger equation*, preprint 1998.

[McD] D. Mac Duff, *Configuration spaces of positive and negative particles*, Topology, 14 (1974), 91-107.

[Mi] P. Mironescu, *Les minimiseurs locaux pour l'équation de Ginzburg-Landau sont à symétrie radiale*, C.R. Acad. Sci. Paris 323 (1996), 593-598.

[OS1] Y.N. Ovchinnikov and J.M. Sigal, *The Ginzburg-Landau equation III. Vortex dynamics*, preprint 1997.

[OS2] Y.N. Ovchinnikov and J.M. Sigal, *Long time behavior of Ginzburg-Landau vortices*, preprint 1998.

[PN] L. Pismen and Nepomnyashchy, *Stability of vortex rings in a model of superflow*, Physica D, (1993), 163-171.

[PaR] F. Pacard and T. Rivière, *A uniqueness result for the minimizers of Ginzburg-Landau functional*, preprint 1998.

[R] T. Rivière, *Line vortices in the $U(1)$-Higgs models*, Cont. Opt. Calc. Var. 1 (1995/96), 77-167 (electronic).

[RS] J. Rubinstein and P. Sternberg, *Homotopy clasification of minimizers of the Ginzburg-Landau energy and the existence of permanent currents*, Comm. Math. Phys. 179 (1996), 257-264.

[S1] S. Serfaty, *Local minimizers for the Ginzburg-Landau Energy near Critical Magnetic fields*, preprint 1997.

[S2] S. Serfaty, *Stable configuration in superconductivity: uniqueness, multiplicity and Vortex-Nucleation*, preprint 1998.

[Str] M. Struwe, *On the asymptotic behavior of the Ginzburg-Landau model in 2 dimensions*, J. Diff. Integr. Equ., 7 (1994), 1613-1624, Erratum 8 (1995), 224.

Université Paris-Sud
Lab. d'Analyse Numérique
et E.D.P
91405 Orsay Cedex
France

Phenomena of Compensation and Estimates for Partial Differential Equations

Frédéric Hélein

Abstract. Quantities like the Jacobian determinant of a mapping play an important role in several partial differential equations in Physics and Geometry. The algebraic structure of such nonlinearities allow to improve slightly the integrability or the regularity of these quantities, sometimes in a crucial way. Focused on the instance of $\frac{\partial a}{\partial x}\frac{\partial b}{\partial y} - \frac{\partial a}{\partial y}\frac{\partial b}{\partial x}$, where a and $b \in H^1(\mathbb{R}^2)$, we review some results obtained on that quantity for 30 years and applications to partial differential equations arising in Geometry, in particular concerning the conformal parametrisations of constant mean curvature surfaces and the harmonic mappingss between Riemannian manifolds.

1991 Mathematics Subject Classification: 35, 43, 49, 53
Keywords and Phrases: Compensation phenomena, Harmonic maps

For 30 years, many remarkable properties concerning some nonlinear quantities like Jacobian determinants of mappings or the scalar product of a divergencefree vector field by the gradient of a function has been observed and used. One instance is the continuity with respect to the weak convergence in L^2. The basic example is the following : if $a_k \rightharpoonup a$ weakly and $b_k \rightharpoonup b$ weakly in $H^1(\mathbb{R}^m)$, then $\{a_k, b_k\}_{\alpha\beta} := \frac{\partial a_k}{\partial x^\alpha}\frac{\partial b_k}{\partial x^\beta} - \frac{\partial a_k}{\partial x^\beta}\frac{\partial b_k}{\partial x^\alpha}$ converges to $\{a, b\}_{\alpha\beta}$ in the distribution sense. The discovery and the study of such properties is the subject of the theory of compensated compactness of F. Murat and L. Tartar [Mu], which became a powerful tool in the theory of homogeneisation and the study of quasiconvex functionals. These technics has been recently enlarged, after R. Di Perna, by P. Gérard [Gé] and L. Tartar [Ta2] independentely in a microlocal context.

We want to tell here a story parallel to compensated compactness' one.

1 H-surfaces

It began with the study of surfaces of constant mean curvature H in the Euclidean space $\mathbb{R}^3$. Let D^2 be the unit disk in the plane $\mathbb{R}^2$. A local conformal parametrisation $X \in H^1(D^2, \mathbb{R}^3)$ satisfies

$$\Delta X = 2H\frac{\partial X}{\partial x} \times \frac{\partial X}{\partial y} \text{ weakly in } H^1(D^2, \mathbb{R}^3), \tag{1}$$

where $V \times W$ is the standard vectorial product in $\mathbb{R}^3$. H. Wente proved that each weak solution of (1) is smooth ($\mathcal{C}^\infty$) [W1]. The crucial step of his proof this to prove that a solution of (1) is continuous. It relies on the particular structure of the right-hand side of (1). For instance, the first component :

$$\Delta X^1 = 2H \left(\frac{\partial X^2}{\partial x} \frac{\partial X^3}{\partial y} - \frac{\partial X^2}{\partial y} \frac{\partial X^3}{\partial x} \right)$$

is a Jacobian determinant. Later in the the beginning of the eighties, in papers from H. Wente [W2] and H. Brezis, J.-M. Coron [BrC], it became clear that the main point in Wente's proof relies on the following. Let $a, b \in H^1(D^2, \mathbb{R})$ and $\phi \in L^1(D^2, \mathbb{R})$ be a weak solution of

$$\left\{ \begin{array}{rcll} -\Delta \phi & = & \{a,b\} := \dfrac{\partial a}{\partial x}\dfrac{\partial b}{\partial y} - \dfrac{\partial a}{\partial y}\dfrac{\partial b}{\partial x} & \text{on } D^2 \\ \phi & = & 0 & \text{on } \partial D^2. \end{array} \right. \tag{2}$$

Then ϕ is actually in $H^1(D^2) \cap \mathcal{C}^0(D^2)$ and we have the following : there exist some positive constants C_∞ and C_2 such that

$$||\phi||_{L^\infty} \leq C_\infty ||da||_{L^2} ||db||_{L^2}, \tag{3}$$

$$||d\phi||_{L^2} \leq C_2 ||da||_{L^2} ||db||_{L^2}. \tag{4}$$

Both estimations are not true in general if we replace the right hand side of (2) by an arbitrary bilinear function of a and b : we would then only obtain that $\phi \in W^{1,p} \cap L^q$ with $1 \leq p < 2$ and $1 \leq q < \infty$. Here the algebraic structure of $\{a, b\}$ is very important and allows us to do many manipulations such as

$$\{a,b\} = \frac{\partial}{\partial x}\left(a \frac{\partial b}{\partial y} \right) - \frac{\partial}{\partial y}\left(a \frac{\partial b}{\partial x} \right)$$

- the basic trick in the proof.

REMARK *Estimates (3) and (4) lead to other inequalities, similar to the isoperimetric inequality in $\mathbb{R}^3$, see [BrC].*

2 ESTIMATES IN REFINED SPACES

In the beginning of the eighties, L. Tartar observed other nice properties on $\{a, b\}$ in the framework of fluid dynamics [Ta1]. And in 1989, S. Müller showed that if u is any function in $W^{1,m}(\mathbb{R}^m, \mathbb{R}^m)$ such that $\det(du)$ is nonnegative a.e. , then, $\det(du)\log(1 + \det(du)) \in L^1(\mathbb{R}^m)$, which improves slightly the naive observation that $\det(du) \in L^1(\mathbb{R}^m)$ [Mü]. We say that $\det(du)$ is in $L^1\log L^1(\mathbb{R}^m)$. The proof of that fact relies also on the use of the isoperimetric inequality in $\mathbb{R}^m$. Notice that if $m = 2$ and $u = (a, b)$, then $\det(du)$ is just $\{a, b\}$.

A few time later, R. Coifman, P.-L. Lions, Y. Meyer and S. Semmes proved actually that if u is any function in $W^{1,m}(\mathbb{R}^m, \mathbb{R}^m)$, then $\det(du)$ belongs to the

generalized *Hardy space* $\mathcal{H}^1(\mathbb{R}^m)$ [CLMS]. It includes S. Müller's result, for it was known that any nonnegative function in $\mathcal{H}^1(\mathbb{R}^m)$ is in $L^1\log L^1(\mathbb{R}^m)$. These authors obtained similar results: for instance, if $B \in L^2(\mathbb{R}^m, \mathbb{R}^m)$ is a divergence free vector field and $V \in H^1(\mathbb{R}^m, \mathbb{R})$, then

$$\nabla V.B \in \mathcal{H}^1(\mathbb{R}^m), \tag{5}$$

the exact analog of the "div-curl lemma" of F. Murat and L. Tartar [Mu].

To make sense it is worth to say what is the generalized Hardy space (see [St]). Several definition coexists. One is the following. Let $f \in L^1(\mathbb{R}^m)$, define

$$f^\star(x) := \sup_{t>0} \left| \int_{\mathbb{R}^m} f(x-y)\phi(\frac{y}{t})\frac{dy}{t^m} \right|,$$

where $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^m)$ is a function such that $\int_{\mathbb{R}^m} f\phi = 1$. Then

$$\mathcal{H}^1(\mathbb{R}^m) := \{f \in L^1(\mathbb{R}^m)/f^\star \in L^1(\mathbb{R}^m)\}.$$

We endow this space with the norm

$$||f||_{\mathcal{H}^1} = ||f||_{L^1} + ||f^\star||_{L^1}.$$

Notice that, through as theorem of C. Fefferman and E. Stein, $BMO(\mathbb{R}^m)$ is the dual space of $\mathcal{H}^1(\mathbb{R}^m)$ ([F], [FSt]). The main property of $\mathcal{H}^1(\mathbb{R}^m)$ is that there exists many linear operators (like the Riesz transform) which are continuous on L^p spaces for $1 < p < \infty$, but not on L^1. But these operators are continuous on $\mathcal{H}^1(\mathbb{R}^m)$.

3 Applications to partial differential equations in geometry

Many applications of these properties have been obtained in the theory of harmonic maps.

Harmonic maps into a sphere

A first example is my result on the regularity of weakly harmonic maps between a two dimensional domain Ω and the two-sphere $S^2 \subset \mathbb{R}^3$ [H1]. These are maps $u \in H^1(\Omega, S^2) := \{v \in H^1(\Omega, \mathbb{R}^3)/|v| = 1 \text{ a.e. }\}$ which are weak solutions of

$$\Delta u + u|du|^2 = 0, \text{ weakly in } H^1(\Omega, \mathbb{R}^3). \tag{6}$$

Here, no Jacobian determinant appears at first glance and the knowledge that $u|du|^2 \in L^1$ is unuseful. The point is to use another equivalent form of the equation which is the conservation law

$$\frac{\partial}{\partial x}\left(u\times\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(u\times\frac{\partial u}{\partial y}\right)=0,\ \text{weakly in } H^1(\Omega,\mathbb{R}^3). \tag{7}$$

This relation was already observed and used independentely by several authors ([Che], [Sh], [KRS]). Assume without loss of generality that Ω is simply connected. We can "integrate" this equation and we deduce that $\exists B\in H^1(\Omega,\mathbb{R}^3)$ such that

$$\left\{\begin{array}{rcl}\dfrac{\partial B}{\partial x} & = & u\times\dfrac{\partial u}{\partial y}\\ \dfrac{\partial B}{\partial y} & = & -u\times\dfrac{\partial u}{\partial x}.\end{array}\right. \tag{8}$$

Now, using the fact that $|u|^2=1$ a.e., which implies that $\langle u,\frac{\partial u}{\partial x}\rangle=\langle u,\frac{\partial u}{\partial y}\rangle=0$, we can rewrite (6) as

$$\begin{array}{rcl}-\Delta u^i & = & \langle u^i\dfrac{\partial u}{\partial x},\dfrac{\partial u}{\partial x}\rangle+\langle u^i\dfrac{\partial u}{\partial y},\dfrac{\partial u}{\partial y}\rangle\\ & = & \langle u^i\dfrac{\partial u}{\partial x}-u\dfrac{\partial u^i}{\partial x},\dfrac{\partial u}{\partial x}\rangle+\langle u^i\dfrac{\partial u}{\partial y}-u\dfrac{\partial u^i}{\partial y},\dfrac{\partial u}{\partial y}\rangle.\end{array}$$

We recognize in the last expression components of $u\times\frac{\partial u}{\partial x}$ and $u\times\frac{\partial u}{\partial y}$. Thus, using (8),

$$-\Delta u^i=-\{u^j,B^k\}-\{B^j,u^k\}, \tag{9}$$

for any (i,j,k) which is a circular permutation of $(1,2,3)$. Now equation (9) is similar to (1) and allows us to prove continuity of u using Wente's estimate. The smoothness of u follows from the classical elliptic theory.

This result generalizes in a straightforward way if we replace the target manifold S^2 by a sphere of arbitrary dimension or a homogeneous manifold, once one realized that the conservation law (7) is a consequence of the symmetries of S^2, using *Noether's theorem* (see [H2]).

This result has also been extended to to the case where the domain Ω is also of higher dimension by L. C. Evans [E]. He proved that, if Ω is an open subset of $\mathbb{R}^m$ is a weakly stationary map into a sphere, then u is smooth in $\Omega\setminus\mathcal{S}$, where $\mathcal{S}$ is a closed subset whose Hausdorff measure of dimension $m-2$ vanishes - a weakly stationary map is a weakly harmonic map satisfying the extra condition that $\int_\Omega|d(u\circ\phi_t)|^2=\int_\Omega|du|^2+o(t)$, for all smooth family of diffeomorphisms ϕ_t acting on Ω, such that ϕ_0 is the identity mapping.

Evans' proof relies on the same arguments, plus the following: the extra condition leads to a monotonicity formula which provides an estimate in BMO. On the other hand, equations like (9) gives estimates in Hardy spaces, through the results of [CLMS]. These estimates complete exactly because of the duality

between $\mathcal{H}^1$ and BMO.

REMARK *It is possible to avoid to use the difficult duality result about $\mathcal{H}^1$ and BMO by direct estimates obtained by S. Chanillo [Cha]. Even more recently, more direct proofs without using that duality has been constructed by P. Hajlasz, P. Strzelecki [HS] and A. Chang, L. Wang, P. Yang [CWY].*

HARMONIC MAPS INTO ARBITRARY MANIFOLDS

It has been possible to extend the previous results for weakly harmonic maps into arbitrary manifolds $\mathcal{N}$. The difficulty is that in general $\mathcal{N}$ is not symmetric and we cannot apply Noether's theorem to construct conservation laws. In dimension 2, I did prove that weakly harmonic maps on a surface, into an arbitrary smooth compact manifold without boundary is smooth, generalizing the preceeding results for spheres [H3]. After, F. Bethuel generalized Evans' result to weakly stationary maps into arbitrary manifolds [Be].

Let $\mathcal{N}$ be a smooth compact Riemannian manifold without boundary. Thanks to the Nash-Moser theorem, we can assume that $\mathcal{N}$ is isometrically embedded in $\mathbb{R}^N$. We define $H^1(\Omega, \mathcal{N})$ to be the set of functions u in $H^1(\Omega, \mathbb{R}^N)$ such that $u \in \mathcal{N}$ a.e. Then weakly harmonic maps $u \in H^1(\Omega, \mathcal{N})$ are the solutions in the distribution sense of the system

$$\Delta u + A(u)(du, du) = 0, \tag{10}$$

where $A(u)(.,.)$ is the second fundamental form of the embedding of $\mathcal{N}$ in $\mathbb{R}^N$. It is a bilinear form on the tangent space to $\mathcal{N}$ at u, with values in the normal subspace to $\mathcal{N}$ at u. Such maps are critical points of the restriction of the functional

$$E(u) = \int_\Omega |du|^2 dx$$

on $H^1(\Omega, \mathcal{N})$. In proving regularity results, the point is to exploit the Euler-Lagrange equation with suitable test-functions, which in some sense are able to measure, to calibrate the possible wild behaviour of a given weak solution. One instance of wild behaviour we have in mind is like the map $(x, y) \longmapsto (\cos(\log(r)), \sin(\log(r)),0)$, from $\mathbb{R}^2$ to S^2, where $r = \sqrt{x^2 + y^2}$: it is harmonic on $\mathbb{R}^2 \setminus \{0\}$ and its image turns along a great circle faster and faster as (x, y) goes to 0. One would like to prove that such a singularity (or something which looks asymptotically like that) does not exists (it actually has an infinite energy). So how to measure such a wild winding ? If $\mathcal{N}$ is S^2, we just take the test function $u \times \phi$, where $\phi \in H^1 \cap L^\infty(\Omega, \mathbb{R}^3)$ and we recover the trick given by Noether's theorem in writing the equation as the conservation law (7). In other cases, we need to construct test functions doing the same job, namely calibrating the possible winding of u. This obtained by using an orthonormal frame on $\mathcal{N}$, moving along u in the "more parallel way". This last requirement means that, although it is not possible in general to construct a covariantly parallel moving frame, it is possible to minimize

the covariant derivative of that moving frame along u. The good news are that the obstruction for constructing a covariantly parallel moving frame along u is the curvature of $\mathcal{N}$ or more precisely the pull-back of the curvature two-form by u. But this pull-back is just a combination of two-order minors of the kind $\{a, b\}$, in the Hardy space! This is done by the following construction.

We start with a given smooth orthonormal moving frame $\tilde{e}(m) = (\tilde{e}_1, ..., \tilde{e}_n)(m)$ defined globally on $\mathcal{N}$ (m being here a point on $\mathcal{N}$), a smooth section of the bundle $\mathcal{F}$ of orthonormal tangent frames over $\mathcal{N}$. In many cases, such a section does not exists globally, because of topological obstructions. Nevertheless, it is possible to reduce ourself to such a situation, through some geometrical argument. Then, for any map $u \in H^1(\Omega, \mathcal{N})$, we consider the composed moving frame $\tilde{e} \circ u$, a section of the pull-back bundle $u^\star \mathcal{F}$, together with all the gauge transformations of $\tilde{e} \circ u$, i.e. for all $R \in H^1(\Omega, SO(n))$, we consider the new frame $e^R(z) = \tilde{e} \circ u(z).R(z)$ for a.e. $z \in \Omega$, or

$$e^R_a(z) = \sum_{b=1}^{n} \tilde{e}_b[u(z)].R^b_a(z).$$

We choose among all e^R's those who minimize the functional

$$F(e^R) := \int_\Omega \sum_{a,b=1}^{n} [\langle \frac{\partial e^R_a}{\partial x}, e^R_b \rangle^2 + \langle \frac{\partial e^R_a}{\partial y}, e^R_b \rangle^2] dxdy.$$

We call a *Coulomb moving frame* such a frame. It satisfies the Euler-Lagrange equation

$$\frac{\partial}{\partial x} \langle \frac{\partial e^R_a}{\partial x}, e^R_b \rangle + \frac{\partial}{\partial y} \langle \frac{\partial e^R_a}{\partial y}, e^R_b \rangle = 0, \tag{11}$$

another conservation law. This equation can be used as (7): some manipulations shows that $\exists A^a_b \in H^1(\Omega)$ such that

$$\begin{cases} \dfrac{\partial A^b_a}{\partial x} &= \langle \dfrac{\partial e^R_a}{\partial y}, e^R_b \rangle \\ \dfrac{\partial A^b_a}{\partial y} &= -\langle \dfrac{\partial e^R_a}{\partial x}, e^R_b \rangle, \end{cases}$$

and that ΔA^b_a is a sum of Jacobian determinants of the type $\{a, b\}$. Namely ΔA^b_a times the volume form on Ω is the pull-back by u of a closed two-form on $\mathcal{N}$ related to the curvature form. This improves slightly the regularity of e^R. In particular, we deduce that the L^2 connection coefficients $\langle \frac{\partial e^R_a}{\partial x}, e^R_b \rangle$ and $\langle \frac{\partial e^R_a}{\partial y}, e^R_b \rangle$ are in fact in the Lorentz space $L^{(2,1)}$, a slight refinement of the usual L^2 space (actually it is the dual space to $L^{(2,\infty)}$, known as weak L^2) (see [StW], [Hu], [BL]). Notice that the above construction did not use at all the hypothesis that u is weakly harmonic.

Now, if we assume that u is weakly harmonic, we will work with the projection of equation (10) on the Coulomb moving frame. We hence get a first

order, Cauchy-Riemann system $\frac{\partial \alpha^a}{\partial \overline{z}} = \sum_{b=1}^{n} \omega_b^a \alpha^b$, where the α^a's are complex numbers representing the derivatives of u and the ω_b^a's are also complex numbers representing connection coefficients. The preliminary work on the Coulomb moving frame ensures us that the ω_b^a's are in $L^{(2,1)}$, instead of L^2. This is enough to prove that u is locally Lipschitz and then that u is smooth.

The regularity theorem of F. Bethuel combines in a delicate way these arguments and Evans' ones. For more details on all of that, see [Be] and [H4].

CONFORMAL PARAMETRISATIONS OF SURFACES

In her thesis, T. Toro, proved the surprising (and difficult) result that the graph of a map ϕ in $H^2(\Omega, \mathbb{R})$, where Ω is an open subset of $\mathbb{R}^2$, is a Lipschitz submanifold, i.e. that there exists local bilipschitz parametrisations of the graph of ϕ. Actually she proved the more general result that this is true for any surface Σ whose mean curvature is a L^2 function on Σ [Tor]. Then, a simpler approach has been found by S. Müller and V. Švérak [MüŠ]. They proved that if Σ is a surface whose mean curvature function belongs to $L^2(\Sigma)$, then a conformal parametrisation of Σ is a bilipschitz function. Their result follows from the observation that, for a local conformal parametrisation $X : D^2 \longrightarrow \Sigma$, if we denote (e_1, e_2) an orthonormal frame such that $dX = e^f(e_1 dx + e_2 dy)$, then

$$\Delta f = u^\star \Omega, \tag{12}$$

where Ω is the curvature two-form on Σ. Thus Δf looks like a Jacobian determinant $\{a, b\}$ and the Wente estimate, or the Coifman, Lions, Meyer, Semmes results implies boundedness of f in L^∞, meaning that X is Lipschitz.

4 THE BEST CONSTANTS

Going back to Wente's result on the disk D^2, it is natural to generalize this inequality to arbitrary two-dimensional domain Ω in the plane, or on a Riemannian surface $(\mathcal{M}, g)$ and to look for the best constants in (3) and (4). If

$$-\Delta_g \phi = \{a, b\} \text{ on } \mathcal{M}, \tag{13}$$

we call

$$\mathcal{C}_\infty(\mathcal{M}, g) = \inf\{osc(\phi)/\phi \text{ is a solution of (13)},$$
$$\text{where } (a, b) \in H^1(\mathcal{M}, \mathbb{R}^2), ||da||_{L^2}^2 + ||db||_{L^2}^2 = 2\},$$

$$\mathcal{C}_2(\mathcal{M}, g) = \inf\{||d\phi||_{L^2}^2/\phi \text{ is a solution of (13)},$$
$$\text{where } (a, b) \in H^1(\mathcal{M}, \mathbb{R}^2), ||da||_{L^2}^2 + ||db||_{L^2}^2 = 2\}.$$

A priori, $\mathcal{C}_\infty(\mathcal{M}, g)$ and $\mathcal{C}_2(\mathcal{M}, g)$ should depend on $\mathcal{M}$ and on the metric g. A first observation is that (13) is invariant under conformal transformations of $(\mathcal{M}, g)$. Thus $\mathcal{C}_\infty(\mathcal{M}, g)$ and $\mathcal{C}_2(\mathcal{M}, g)$ depend only on the conformal structure of $(\mathcal{M}, g)$. Moreover, F. Bethuel and J.-M. Ghidaglia proved that these constants were bounded by a universal one [BeG]

Recently the precise evaluation of these constants were completed by S. Baraket and P. Topping for $\mathcal{C}_\infty(\mathcal{M}, g)$ [Ba], [Top] and by Y. Ge for $\mathcal{C}_2(\mathcal{M}, g)$ [Ge]. We have that

- $\mathcal{C}_\infty(\mathcal{M}, g) = \frac{1}{2\pi}$, for all $(\mathcal{M}, g)$.
- $\mathcal{C}_2(\mathcal{M}, g) = \sqrt{\frac{3}{16\pi}}$ if $\partial\mathcal{M}$ is non empty and $\mathcal{C}_2(\mathcal{M}, g) = \sqrt{\frac{3}{32\pi}}$ if $\partial\mathcal{M}$ is empty.

Both result relies on the optimal isoperimetric inequality (on the plane for $\mathcal{C}_\infty(\Omega)$ and in $\mathbb{R}^3$ for $\mathcal{C}_2(\Omega)$).

BACK TO THE BEGINNING

The search for the optimal constant $\mathcal{C}_2(\mathcal{M}, g)$ leads to a variational problem very similar to the search for the optimal constant in Sobolev embedding of $H^1(\mathbb{R}^m)$ in $L^{\frac{2m}{m-2}}(\mathbb{R}^m)$. First this problem is invariant under conformal transformations. Moreover critical points of the functional $||d\phi||^2_{L^2}$ under the constraint that $||da||^2_{L^2} + ||db||^2_{L^2} = 2$, satisfies the following Euler-Lagrange equation: there exists a Lagrange multiplier $\lambda \in (0, \infty)$ such that

$$u = \begin{pmatrix} \sqrt{\lambda}a \\ \sqrt{\lambda}b \\ \lambda\phi \end{pmatrix}$$

is a weak solution of

$$\Delta u = 2\frac{\partial u}{\partial x} \times \frac{\partial u}{\partial y},$$

the equation of conformal parametrisations of constant mean curvature surfaces (see [H4], [Ge]). Hence we are led to another variational formulation of that geometrical problem. Y. Ge obtained several existence results on this problem, by constructing minimizing and non-minimizing solutions [Ge].

5 REFERENCES

[Ba] S. Baraket, *Estimations of the best constant involving the L^∞ norm in Wente's inequality*, Ann. Fac, Sciences Toulouse (1997).

[BL] J. Bergh, J. Löfström, *Interpolation spaces, an introduction*, Springer Verlag Berlin, 1976.

[Be] F. Bethuel, *On the singular set of stationary harmonic maps*, Manuscripta Mathematica 78 (1993), 417-443.

[BeG] F. Bethuel, J.-M. Ghidaglia, *Improved regularity for solutions of elliptic equations involving Jacobian and applications*, J. Maths pures et Appliquées 72 (1993), 441-475.

[BrC] H. Brezis, J.-M. Coron, *Multiple solutions of H-systems and Rellich's conjecture*, Comm. Pure Appl. Math. 37 (1984), 149-187.

[CWY] S.-Y. A. Chang, L. Wang, P. C. Yang, *Regularity of harmonic maps*, preprint 1998.

[Cha] S. Chanillo, *Sobolev inequalities involving divergence free maps*, Comm. PDE's 16 (1991), 1969-1994.

[Che] Y. Chen, *The weak solutions to the evolution problems of harmonic maps*, Math. Zeitschrift 201 (1989), 69-74.

[CLMS] R. Coifman, P.-L. Lions, Y. Meyer, S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pure Appl. 72 (1993), 247-286.

[E] L. C. Evans, *Partial regularity for stationary harmonic maps into spheres*, Arch. Rat. Mech. Anal. 116 (1991), 101-163.

[F] C. Fefferman, *Characterisation of bounded mean oscillations*, Bull. Am. Math. Soc. 77 (1971), 585-587.

[FSt] C. Fefferman, E. Stein, *H^p spaces of several variables*, Acta Mathematica 129 (1972), 137-193.

[Ge] Y. Ge, *Estimation of the best constant involving the L^2 norm in Wente's inequality and compact H-surfaces into Euclidean space*, COCV (1998).

[Gé] P. Gérard, *Microlocal defect measures*, preprint Unversité Paris-Sud (1990).

[HS] P. Hajlasz, P. Strzelecki, *Subelliptic p-harmonic maps into spheres and the ghost of Hardy spaces*, preprint 1997.

[H1] F. Hélein, *Régularité des applications faiblement harmoniques entre une surface et une sphère*, C. R. Acad. Sci. Paris 311 (1990), 519-524.

[H2] F. Hélein, *Regularity of weakly harmonic maps from a surface into a manifold with symmetries*, Manuscripta Mathematica 70 (1991), 293-318.

[H3] F. Hélein, *Régularité des applications faiblement harmoniques entre une surface et une variété riemannienne*, C. R. Acad. Sci. Paris 312 (1990), 591-596.

[H4] F. Hélein, *Applications harmoniques, lois de conservation et repères mobiles*, Diderot éditeur, 1996; *Harmonic maps, conservations laws and moving frames*, Diderot éditeur, 1997.

[Hu] R.A. Hunt, *On L(p,q) spaces*, L'enseignement mathématique XII (1966), 249-276.

[KRS] J. Keller, J. Rubinstein, P. Sternberg, *Reaction-diffusion processes and evolution to harmonic maps*, SIAM J. Appl. Math. 49 n. 6 (1989), 1722-1733.

[Mü] S. Müller, *Higher integrability of determinants and weak convergence in L^1*, J. Reine Angewandte Math. 412 (1990), 20-34.

[MüŠ] S. Müller, V. Švérak, *On surfaces of total finite curvature*, J. Diff. Geometry 42 (1995), 229-258.

[Mu] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. 53 (1978), 1092-1099.

[Sh] J. Shatah, *Weak solutions and developments of singularities of the SU(2) σ-model*, Comm. Pure and Appl. Math. 41 (1988), 459-469.

[St] E. Stein, *Harmonic analysis*, Princeton University Press, Princeton, 1993.

[StW] E. Stein, G. Weiss, *Introduction to Fourier analysis on Euclidean spaces*, Princeton University Press, Princeton, 1970.

[Ta1] L. Tartar, *Remarks on oscillations and Stokes' equation*, in Lec. Notes in Physics 230, Springer 1984.

[Ta2] L. Tartar, *H-measures and applications*, in Proc. Int. Cong. Math. Kyoto, Japan 1990, MSJ (1991).

[Top] P. Topping, *The optimal constant in Wente's L^∞ estimates*, Comm. Math. Helv. (1997).

[Tor] T. Toro, *Surfaces with generalized second fundamental form in L^2 are Lipschitz manifolds*, J. Diff. Geometry 39 (1994), 65-101.

[W1] H. Wente, *An existence theorem for surfaces of constant mean curvature*, J. Math. Anal. Appl. 26 (1969), 318-344.

[W2] H. Wente, *Large solutions to the volume constraint Plateau problem*, Arch. Rat. Mech. Anal. 75 (1980), 59-77.

Frédéric Hélein
CMLA, ENS de Cachan
61 avenue du Président Wilson
94235 Cachan Cedex, France
helein@cmla.ens-cachan.fr

Viscosity Solutions of Elliptic Partial Differential Equations

Robert R. Jensen

Abstract. In my talk and its associated paper I shall discuss some recent results connected with the uniqueness of viscosity solutions of nonlinear elliptic and parabolic partial differential equations. By now, most researchers in partial differential equations are familiar with the definition of viscosity solution, introduced by M. G. Crandall and P. L. Lions in their seminal paper, "Condition d'unicité pour les solutions generalisées des équations de Hamilton-Jacobi du premier order," *C. R. Acad. Sci. Paris* 292 (1981), 183–186. Initially, the application of this definition was restricted to nonlinear first order partial differential equations—i.e., Hamilton-Jacobi-Bellman equations—and it was shown that viscosity solutions satisfy a maximum principle, implying uniqueness. In 1988 an extended definition of viscosity solution was applied to second order partial differential equations, establishing a maximum principle for these solutions and a corresponding uniqueness result. In the following years numerous researchers obtained maximum principles for viscosity solutions under weaker and weaker hypotheses. However, in all of these papers it was necessary to assume some minimal modulus of spatial continuity in the nonlinear operator, depending on the regularity of the solution, and to assume either uniform ellipticity or strong monotonicity in the case of elliptic operators. The results I shall discuss are related to attempts to weaken these assumptions on the partial differential operators—e.g., operators with only measurable spatial regularity, and operators with degenerate ellipticity.

1991 Mathematics Subject Classification: 35, 49, 60
Keywords and Phrases: nonlinear, elliptic, partial differential equations, viscosity solution, stochastic process

1 Viscosity solutions: a brief history

Although the history of viscosity solutions begins in 1981/83, depending on your individual bias, an important precursor is found in the work of S. N. Kruzkov. In fact, it's noted in [12] that, "analogies with S. N. Krukov's theory of scalar conservation laws ([29]) provided guidance for the notion [of viscosity solutions]

and its presentation." In this context one should also mention L. C. Evans [16], which developed techniques that serendipitously anticipated the introduction of viscosity solutions.

M. G. Crandall and P. L. Lions announced the discovery of viscosity solutions in 1981 ([10]). Complete proofs and details were presented shortly after this in their landmark paper [11]. However, the definition of viscosity solution used in this paper bears little resemblance to any of those we now employ. It is in M. G. Crandall, L. C. Evans and P. L. Lions [9] where we first see a systematic use of one of the now familiar definitions of viscosity solutions. P. L. Lions was quick to grasp the potential in extending the notion of viscosity solutions to more general PDEs—[10] and [11] only deal with first order Hamilton-Jacobi-Bellman equations. His papers, [30] and [31], are the first attempts to extend the first order results of [11] to second order equations. Using stochastic control theory, he was able to prove a maximum principle for viscosity solutions of convex (or concave) nonlinear second order Hamilton-Jacobi equations.

It was five years later that methods were developed which extended the theory of viscosity solutions to fully nonlinear second order elliptic PDEs. In the first of these papers R. Jensen [24] proved a maximum principle for Lipschitz viscosity solutions to the fully nonlinear second order elliptic PDE on a bounded domain $\Omega \subset \mathbf{R}^n$

$$F(u, Du, D^2u) = 0 \quad \text{in} \quad \Omega \tag{1}$$

Next, in a short note R. Jensen, P. .L. Lions, and P. E. Souganidis [28] removed the hypothesis of Lipschitz continuity from the viscosity solution. At about the same time, using the ideas in [24], N. Trudinger proved $C^{1,\alpha}$ regularity for viscosity solutions of uniformly elliptic problems ([35]), and a maximum principle for such solutions ([36]). Then H. Ishii [20] made an important contribution by removing the assumption of spatial independence in the PDE. I.e., the maximum principle could now be applied to viscosity solutions of

$$F(x, u, Du, D^2u) = 0 \quad \text{in} \quad \Omega \tag{2}$$

Finally, in concurrently developed papers H. Ishii and P. L. Lions [22], and R. Jensen [25] significantly extended [20] giving very general (and in [25], a rather complicated technical) conditions under which a maximum principle holds for viscosity solutions of (2). In particular, suppose the functions $F(x, t, p, M)$ appearing in (2) is given by the formula

$$F(x,t,p,M) = \min_{\beta \in \mathcal{B}} \left\{ \max_{\gamma \in \mathcal{C}} \left\{ a_{il}^{\beta\gamma}(x) a_{jl}^{\beta\gamma}(x) m_{ij} + b_i^{\beta\gamma}(x) p_i - c^{\beta\gamma}(x) t - h^{\beta\gamma}(x) \right\} \right\} \tag{3}$$

where $M = (m_{ij})$, $p = (p_1, \ldots, p_n)$ and summation is implicit over the indices $i, j,$ and l. Then we have from [25]

COROLLARY 5.11. *Let F be the function defined by* (3) *and assume $\left\{\left(a_{rs}^{\beta\gamma}(x)\right)\right\}$ are uniformly Lipschitz continuous in $\overline{\Omega}$, $\left\{\left(b_i^{\beta\gamma}(x)\right)\right\}$ are uniformly Lipschitz continuous in $\overline{\Omega}$, and $\left\{\left(c^{\beta\gamma}(x)\right)\right\}$ and $\left\{\left(h^{\beta\gamma}(x)\right)\right\}$ are equicontinuous in $\overline{\Omega}$. If u is a*

viscosity subsolution of (2), *v is a viscosity supersolution of* (2), *and*

$$F(x,t,p,M) - F(x,s,q,N) \leq \max\{K_1 \text{trace}(M-N), K_2(s-t)\} + K_3|p-q| \quad (4)$$

then

$$\sup_{\Omega}(u-v)^+ \leq \sup_{\partial\Omega}(u-v)^+ \quad (5)$$

We also have two other corollaries from [25] which demonstrate the link between the spatial dependence of F and the regularity of the viscosity solution.

COROLLARY 5.14. *Let F be the function defined by* (3) *and assume $\{(a_{rs}^{\beta\gamma}(x))\}$ are uniformly Hölder continuous with exponent $\gamma(> 1/2)$ in $\overline{\Omega}$, $\left\{\left(b_i^{\beta\gamma}(x)\right)\right\}$ are uniformly Hölder continuous with exponent $2\gamma - 1$ in $\overline{\Omega}$, and $\{(c^{\beta\gamma}(x))\}$ and $\{(h^{\beta\gamma}(x))\}$ are equicontinuous in $\overline{\Omega}$. If u is a viscosity subsolution of* (2), *v is a viscosity supersolution of* (2), *either u or v is Hölder continuous with exponent $\alpha > 2 - 2\gamma$, and* (4) *holds, then*

$$\sup_{\Omega}(u-v)^+ \leq \sup_{\partial\Omega}(u-v)^+ \quad (6)$$

COROLLARY 5.16. *Let F be the function defined by* (3) *and assume $\{(a_{rs}^{\beta\gamma}(x))\}$ are uniformly Hölder continuous with exponent $\gamma(\leq 1/2)$ in $\overline{\Omega}$, $\left\{\left(b_i^{\beta\gamma}(x)\right)\right\}$ are equicontinuous in $\overline{\Omega}$, and $\{(c^{\beta\gamma}(x))\}$ and $\{(h^{\beta\gamma}(x))\}$ are also equicontinuous in $\overline{\Omega}$. If u is a viscosity subsolution of* (2), *v is a viscosity supersolution of* (2), *either u or v is in $C^{1,\alpha}(\Omega)$ for some $\alpha \geq \frac{1-2\gamma}{1-\gamma}$, and* (4) *holds, then*

$$\sup_{\Omega}(u-v)^+ \leq \sup_{\partial\Omega}(u-v)^+ \quad (7)$$

While the preceding results are not sharp, they do indicate how the assumption of greater regularity of the viscosity solution allows us to reduce the regularity in the spatial dependence of F necessary to prove a maximum principle. Specifically, in conjunction with regularity results about the gradient (e.g., [35]), one obtains a fairly general maximum principle (compare [36]).

It was also during this period that L. Caffarelli's famous paper [3] on interior *a priori* estimates for viscosity solutions appeared. It was in this paper that Caffarelli extended the classical $W^{2,p}$, $C^{1,\alpha}$, and$C^{2,\alpha}$ interior estimates, using the Aleksandov-Bakelman-Pucci maximum principle, the Calderon-Zygmund decomposition lemma, and an extremely clever application of the Krylov-Safonov Harnack inequality. By eschewing the traditional approach used for linear PDEs—singular integral operator theory—he obtains results which are powerful enough to apply to fully nonlinear uniformly elliptic operators.

2 VISCOSITY SOLUTIONS: RECENT RESULTS

The two most exciting (or depressing, depending on your point of view) recent results are a pair of counterexamples due to N. Nadirashvili. The first ([32]), is an example of nonuniqueness for linear uniformly elliptic PDEs with bounded, measurable coefficients. I.e., consider the equation

$$\left.\begin{aligned} &\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} = f(x) \quad \text{in} \quad \Omega \subset \mathbf{R}^n \\ &u|_{\partial\Omega}(x) = g(x) \end{aligned}\right\} \tag{8}$$

If $(a_{ij}(x))$ are bounded, measurable and uniformly elliptic, f is bounded and measurable, and g is bounded and continuous we may define a solution of (8) as a limit of solutions of

$$\left.\begin{aligned} &\sum_{i,j=1}^{n} a^k_{ij}(x)\frac{\partial^2 u^k}{\partial x_i \partial x_j} = f(x) \quad \text{in} \quad \Omega \subset \mathbf{R}^n \\ &u^k|_{\partial\Omega}(x) = g(x) \end{aligned}\right\} \tag{9}$$

where $\left\{\left(a^k_{ij}(x)\right)\right\}$ are smooth and converge almost everywhere to $(a_{ij}(x))$. The sequence $\left\{u^k\right\}$ is equicontinuous due to Krylov's Hölder continuity estimates. Hence, the sequence has accumulation points. We may view these accumulation points as "good" solutions of (8). If there is only one accumulation point no matter what approximating sequence we use, then (in some sense) the "good" solution of (8) is unique.

Under certain conditions it is possible to prove that "good" solutions of (8) are unique. For example, M. C. Cerutti, L. Escauriaza, and E. B. Fabes [6] prove this if the set of discontinuities of $(a_{ij}(x))$ is countable with at most one accumulation point. M. Safonov [34] proves uniqueness if the set of discontinuities of $(a_{ij}(x))$ has sufficiently small Hausdorff dimension. In this connection R. Jensen [27] defines a measure theoretic notion of viscosity solution and proves that viscosity solutions and "good" solutions are equivalent. A continuous function $u \in C(\overline{\Omega})$ is a viscosity subsolution of (8) if for any $\phi \in C^2(\Omega)$ such that $(u-\phi)(x) \geq (u-\phi)(y)$ for all $y \in \Omega$ and for all $\eta > 0$

$$\limsup_{\varepsilon\to 0} \frac{1}{\varepsilon^n} \int_{\mathrm{B}(x,\varepsilon)} \left[\sum_{i,j=1}^{n} a_{ij}(y)\left(\frac{\partial^2\phi}{\partial x_i \partial x_j}(x) + \eta\delta_{ij}\right) - f(y)\right]^{+} dy > 0 \tag{10}$$

it's a viscosity supersolution if for any $\phi \in C^2(\Omega)$ such that $(u-\phi)(x) \leq (u-\phi)(y)$ for all $y \in \Omega$ and for all $\eta > 0$

$$\limsup_{\varepsilon\to 0} \frac{1}{\varepsilon^n} \int_{\mathrm{B}(x,\varepsilon)} \left[\sum_{i,j=1}^{n} a_{ij}(y)\left(\frac{\partial^2\phi}{\partial x_i \partial x_j}(x) - \eta\delta_{ij}\right) - f(y)\right]^{-} dy > 0 \tag{11}$$

and it's a viscosity solution if it's both a subsolution and a supersolution. It's relatively easy to see that a "good" solution is always a viscosity solution. Amazingly, it's also possible to show that if u is a viscosity solution of (8), then there is

a sequence of coefficients $\left\{\left(a_{ij}^k(x)\right)\right\}$ converging to $(a_{ij}(x))$ such that the solutions $\left\{u^k\right\}$ of (9) converge to u.

It follows that viscosity solutions is the "right" or "natural" space to work in when studying solutions of (8). The counterexample of [32] shows that multiple viscosity solutions of (8) *do* can exist. I.e., viscosity solutions of (8) are *not* unique. Still, [27] has some interesting consequences. For example, suppose $(a_{ij}(x))$ are continuous. Then we know from the general theory of linear PDEs that there is a solution $w \in W^{2,p}(\Omega) \cap C(\overline{\Omega})$ for any $p > n$. Such solutions are unique and stable. It now follows from [27] that if u is a viscosity solution of (8), then $u = w$. Thus, if $(a_{ij}(x))$ are continuous, then viscosity solutions of (8) are in $W^{2,p}(\Omega)$. In a pair of papers related to [27], [5] and [8], L. Caffarelli, , M. G. Crandall, M. Kocan, P. Soravia, and A. Święch examine the notion of a L^p-viscosity solutions. In the context of (8) a function $u \in W^{2,p}(\Omega)$ for $p > n/2$ is a L^p-viscosity subsolution of (8) if for any $\phi \in W^{2,q}_{loc}(\Omega)$ such that $q > p$ and $(u-\phi)(y)$ has a local max at $y = x$ then

$$\operatorname{ess\,lim\,sup}_{y \to x} \left\{ \sum_{i,j=1}^{n} a_{ij}(y) \frac{\partial^2 \phi}{\partial x_i \partial x_j}(y) - f(y) \right\} \geq 0 \tag{12}$$

it's a L^p-viscosity supersolution if for any $\phi \in W^{2,q}_{loc}(\Omega)$ such that $q > p$ and $(u - \phi)(y)$ has a local min at $y = x$ then

$$\operatorname{ess\,lim\,inf}_{y \to x} \left\{ \sum_{i,j=1}^{n} a_{ij}(y) \frac{\partial^2 \phi}{\partial x_i \partial x_j}(y) - f(y) \right\} \leq 0 \tag{13}$$

and it's a L^p-viscosity solution if it's both a subsolution and a supersolution. The authors prove a variety of interesting results concerning such solutions. In particular they they show that that such solutions are twice differentiable almost everywhere, they examine the relationship between various definitions of viscosity solutions (in the measurable context), and they extend and generalize the results in [27]. One of the tools in their analysis is the interesting paper of L. Escauriaza ([15]), which extends the classical Aleksandrov-Bakelman-Pucci maximum principle.

Nadirashvili's second counterexample, [33], shows that there is a smooth function F such that the solution of (2) is *not* C^2. This is important because this result shows that the $C^{2,\alpha}$ regularity theory—the Schauder estimates—of linear PDEs doesn't hold for fully nonlinear PDEs, underscoring the importance of the theory of viscosity solutions to elliptic PDEs. Applications of viscosity solutions to degenerate elliptic and parabolice PDEs also underscore their importance. One of the more widely known applications has been to the problem of motion by mean curvature. The idea of embedding the hypersurface as a level set of some initial value and evolving the initial data by the appropriate degenerate parabolic PDE goes back to L. C. Evans and J. Spruck [19], and Y. G. Chen, Y. Giga, and S. Goto [7]. Showing that the level set's evolution was independent of the particular initial data used, they were able to prove existence and uniqueness results for the motion

by mean curvature problem. These results have been expanded on and generalized in L. C. Evans [18], and H. Ishii, and P. E. Souganidis [23].

In a different vein R. Jensen [26] studied a highly nonlinear degenerate elliptic PDE in the context of L^∞ minimization and the limit of the p-Laplacian as p goes to infinity. Recently this operator has also been connected to the Monge-Kantorovich problem of optimal transport, and (I have been told) to image processing. The problem studied in [26] is to find the "best" Lipschitz extension into Ω of the boundary data $g(x)$. This is reduced to the problem of existence and uniqueness of the nonlinear PDE

$$\left.\begin{array}{l} \displaystyle\sum_{i,j=1} \frac{\frac{\partial u}{\partial x_i}}{|Du|}(x)\frac{\frac{\partial u}{\partial x_j}}{|Du|}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} = 0 \quad \text{in} \quad \Omega \subset \mathbf{R}^n \\ u|_{\partial\Omega}(x) = g(x) \end{array}\right\} \tag{14}$$

It is easy to see that (14) is both degenerate elliptic and singular at $Du(x) = 0$. Never the less, it was shown that viscosity solutions of (14) exist and also satisfy a maximum principle. Hence, they are unique. Furthermore, for this problem there are also counterexamples to the existence of classical solutions. In fact, the best regularity for this problem appears to be $C^{1,\alpha}$, but a proof of this remains open.

REFERENCES

[1] Alvarez, O. and A. Tourin, Viscosity solutions of nonlinear integro-differential equations, *Ann. Inst. H. Poincaré Anal. Non Linéaire* 13 (1996), 293–317.

[2] Barles, G. and J. Burdeau, The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problems, *Comm. Partial Differential Equations* 20 (1995), 129-178.

[3] Caffarelli, L., Interior *a priori* estimates for solutions of fully nonlinear equations, *Annals Math.* 130 (1989), 180–213.

[4] Caffarelli, L. and X. Cabré, *Fully nonlinear elliptic equations,* American Mathematical Society Colloquium Publications, 43 (1995).

[5] Caffarelli, L., M. G. Crandall, M. Kocan, and A. Święch, On viscosity solutions of fully nonlinear equations with measurable ingredients, *Comm. Pure Appl. Math.* 49 (1996) 365–397.

[6] Cerutti, M. C., L. Escauriaza, and E. B. Fabes, Uniqueness in the Dirichlet problem for some elliptic operators with discontinuous coefficients, *Ann. Mat. Pura Appl.* 163 (1993), 161–180.

[7] Chen, Y. G., Y. Giga, and S. Goto, Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations, *J. Differential Geom.* 33 (1991), 749–786.

[8] Crandall, M. G., M. Kocan, P. Soravia, and A. Święch, On the equivalence of various weak notions of solutions of elliptic PDEs with measurable ingredients, *Progress in elliptic and parabolic partial differential equations* (*Capri, 1994*), 136–162, *Pitman Res. Notes Math. Ser.,* 350, *Longman, Harlow,* 1996.

[9] Crandall, M. G., L. C. Evans, and P. L. Lions, Some properties of viscosity solutions of Hamilton-Jacobi equations, *Trans. Amer. Math. Soc.* 282 (1984), 487–502.

[10] Crandall, M. G. and P. L. Lions, Condition d'unicité pour les solutions generalisées des équations de Hamilton-Jacobi du premier order, *C. R. Acad. Sci.* 292 (1981), 183–186.

[11] Crandall, M. G. and P. L. Lions, Viscosity solutions of Hamilton-Jacobi equations, *Trans. Amer. Math. Soc.* 277 (1983), 1–42.

[12] Crandall, M. G., H. Ishii and P. L. Lions, User's guide to viscosity solutions of second order partial differential equations, *Bull. Amer. Math. Soc.(N.S.)*27 (1992), 1–67.

[13] Cranny, T. R., Nonclassical solutions of fully nonlinear elliptic PDEs, *Instructional Workshop on Analysis and Geometry, Part I* (*Canberra,* 1995), 121–136, *Proc. Centre Math. Appl. Autstralia Nat. Univ.,* 34, *Australia Nat. Univ., Canberra,* 1996.

[14] Cranny, T. R., On the uniqueness of solutions of the homogeneous curvature equations, Ann. Inst. H. Poincaré Anal. Non Linéaire 13 (1996), 619–630.

[15] L. Escauriaza, $W^{2,n}$ a priori estimates for solutions to fully nonlinear equations, *Indiana Univ. Math. J.* 42 (1993), 413–423.

[16] Evans, L. C., A convergence theorem for solutions of nonlinear second order elliptic equations, *Indiana Univ. Math. J.* 27 (1978), 875–887.

[17] Evans, L. C., Regularity for fully nonlinear elliptic equations and motion by mean curvature, *Viscosity solutions and applications (Montecatini Terme, 1995)*, 98–133, *Lecture Notes in Math.,* 1660, *Springer, Berlin,* 1997.

[18] Evans, L. C., A geometric interpretation of the heat equation with multivalued initial data, *SIAM J. Math. Anal.* 27 (1996), 932–958.

[19] Evans, L. C. and J. Spruck, Motion of level sets by mean curvature.I, *J. Differential Geom.* 33 (1991), 635–681.

[20] Ishii, H., On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDE's, *Comm. Pure Appl. Math.* 42 (1989) 14–45.

[21] Ishii, H., Degenerate parabolic PDEs with discontinuities and generalized evolutions of surfaces, *Adv. Differential Equations* 1 (1996), 51–72.

[22] Ishii, H. and P. L. Lions, Viscosity solutions of fully nonlinear second-order elliptic partial differential equations, *J. Differential Equations* 83 (1990), 26–78.

[23] Ishii, H. and P. E. Souganidis, Generalized motion of noncompact hypersurfaces with velocity having arbitrary growth on the curvature tensor, *Tôhoku Math. J.* 47 (1995), 227–250.

[24] Jensen, R., The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations, *Arch. Rat. Mech. Anal.* 101 (1988), 1–27.

[25] Jensen, R., Uniqueness criteria for viscosity solutions of fully nonlinear elliptic partial differential equations, *Indiana Univ. Math. J.* 38 (1989), 629–667.

[26] Jensen, R., Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient, *Arch. Rat. Mech. Anal.* 123 (1993) 51–74.

[27] Jensen, R., Uniformly Elliptic PDEs with Bounded, Measurable Coefficients, *J. Fourier Anal. Appl.* 2 (1996), 237–259.

[28] Jensen, R., P. L. Lions, and P. E. Souganidis, A uniqueness result for viscosity solutions of second order fully nonlinear partial differential equations, *Proc. Amer. Math. Soc.* 102 (1988), 975-987.

[29] Kruzkov, S. N., First order quasilinear equations in several independent variables, *Math. USSR-Sb.* 10 (1970), 217–243.

[30] Lions, P. L., Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part 1: The dynamic programming principle and applications, *Comm. Partial Differential Equations* 8 (1983), 1101-1174.

[31] Lions, P. L., Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part 2: Viscosity solutions and uniqueness, *Comm. Partial Differential Equations* 8 (1983), 1229-1276.

[32] Nadirashvili, N., Nonuniqueness in the martingale problem and the Dirichlet problem for uniformly elliptic operators, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* 24 (1997), 537–550.

[33] Nadirashvili, N., Nonclassical Solutions to Fully Nonlinear Elliptic Equations, preprint.

[34] Safonov, M. V., On a weak uniqueness for some elliptic equations, *Comm. Partial Differential Equations* 19 (1994, 943–957.

[35] Trudinger, N. S., Hölder gradient estimates for fully nonlinear elliptic equations, *Proc. Roy. Soc. Edinburgh Sect. A* 108 (1988), 57–65.

[36] Trudinger, N. S., Comparison principles and pointwise estimates for viscosity solutions of nonlinear elliptic equations, *Rev. Mat. Iberoamericana* 4 (1988), 453–468.

[37] Trudinger, N. S., The Dirichlet problem for the prescribed curvature equations, *Arch. Rat. Mech. Anal.* 111 (1990) 153–179.

[38] Wang, L., A maximum principle for elliptic and parabolic equations with oblique derivative boundary problems, *J. Partial Differential Equations* 5 (1992), 23–27.

Robert R. Jensen
Dept. Math. and Comp. Sciences
Loyola University Chicago
Chicago, IL 60626
U.S.A.
rrj@math.luc.edu

Minimal Regularity Solutions of Nonlinear Wave Equations

Hans Lindblad

Abstract.

Inspired by the need to understand the complex systems of non-linear wave equations which arise in physics, there has recently been much interest in proving existence and uniqueness for solutions of nonlinear wave equations with low regularity initial data.

We give counterexamples to local existence with low regularity data for the typical nonlinear wave equations. In the semi-linear case these are sharp, in the sense that with slightly more regularity one can prove local existence.

We also present join work with Georgiev and Sogge proving global existence for a certain class of semi-linear wave equation. This result was a conjecture of Strauss following an initial result of Fritz John. We develop weighted Strichartz estimates whose proof uses techniques from harmonic analysis taking into account the symmetries of the wave equation.

1991 Mathematics Subject Classification: 35L70
Keywords and Phrases: Non-linear wave equations, hyperbolic equations, local existence, low regularity solutions, Strichartz estimates

Introduction.

Recently there has been much interest in proving existence and uniqueness of solutions of nonlinear wave equations with low regularity initial data. One reason is that many equations from physics can be written as a system of nonlinear wave equations with a conserved energy norm. If one can prove local existence and uniqueness assuming only that the energy norm of initial data is bounded then global existence and uniqueness follow. Therefore it is interesting to find the minimal amount of regularity of the initial data needed to ensure local existence for the typical nonlinear wave equations.

We give counterexamples to local existence with low regularity data for the typical nonlinear wave equations. In the semi-linear case the counterexamples are sharp, in the sense that with slightly more regularity one can prove local existence. It is natural to look for existence in Sobolev spaces, since the Sobolev norms are more or less the only norms that are preserved for a linear wave equation. The counterexamples involve constructing a solution that develops a singularity along

a characteristic for all positive times. In the quasi-linear case it also involves controlling the geometry of the characteristic set. The norm is initially bounded but becomes infinite for all positive times, contradicting the existence of a solution in the Sobolev space. The counterexamples are half a derivative more regular than what is predicted by a scaling argument. The scaling argument use the fact that the equations are invariant under a scaling to obtain a sequence of solutions for which initial data is bounded in an appropriate Sobolev norm. The counterexamples were not widely expected since for several nonlinear wave equations one does obtain local existence down to the regularity predicted by scaling.

On the other hand, the classical local existence theorems for nonlinear wave equations are not sharp in the semi-linear case. These results were proved using just the energy inequality and Sobolev's embedding theorem. Recently they were improved using space-time estimates for Fourier integral operators known as Strichartz' estimates, and generalizations of these. There are many recent results in this field, for example work by Klainerman-Machedon[13-15], Lindblad-Sogge[24], Grillakis[6] Ponce-Sideris[26] and Tataru. In particular, Klainerman-Machedon proved that for equations satisfying the 'null condition', one can go down to the regularity predicted by the scaling argument mentioned above. In joint work with Sogge[24] we prove local existence with minimal regularity for a simple class of model semi-linear wave equations. There are related results for KdV and nonlinear Schrödinger equations, for example in work by Bourgain and Kenig-Ponce-Vega.

Whereas the techniques of harmonic analysis were essential in improving the local existence results, the Strichartz estimates are not the best possible global estimates since they do not catch the right decay as time tends to infinity if the initial data has compact support. The classical method introduced by Klainerman [11,12] to prove global existence for small initial data is to use the energy method with the vector fields coming from the invariances of the equation. However, this method requires much regularity of initial data and also the energy method alone does not give optimal estimates for the solution since it is an estimate for derivatives. We will present joint work with Georgiev and Sogge giving better global estimates using techniques from harmonic analysis taking into account the invariances or symmetries of the wave equation. We obtain estimates with mixed norms in the angular and spherical variables, with Sogge[24], and weighted Strichartz' estimates with Georgiev and Sogge[4]. Using these new estimates we prove that a certain class of semi-linear wave equations have global existence in all space dimensions. This was a conjecture by Strauss, following an initial result by John.

1. Counterexamples to local existence.

We study quasi-linear wave equations and ask how regular the initial data must be to ensure that a local solution exists. We present counterexamples to local existence for typical model equations. Greater detail of the construction can be found in Lindblad [20-23]. In the semi-linear case the counter examples are sharp in the sense that for initial data with slightly more regularity a local solution exists. This was shown recently in Klainerman-Machedon [13-15], Ponce-Sideris[26] and Lindblad-Sogge[24] using space time estimates know as Strichartz' estimates and refinements of these. However for quasi-linear equations it is still unknown what the optimal result is; there is a gap between the counterexamples and a recent

improvement on the existence result by Tataru[42] and Bahouri-Chemin[1].

Consider the Cauchy problem for a quasi-linear wave equation:

$$\begin{aligned} \Box u = G(u, u', u''), \quad (t,x) \in S_T = [0,T) \times \mathbb{R}^n, \\ u(0,x) = f(x), \quad u_t(0,x) = g(x), \end{aligned} \tag{1.1}$$

where G is a smooth function which vanishes to second order at the origin and is linear in the third variable u''. (Here $\Box = \partial_t^2 - \sum_{i=1}^n \partial_{x_i}^2$.) Let $\dot{H}^\gamma$ denote the homogeneous Sobolev space with norm $\|f\|_{\dot{H}^\gamma} = \| \, |D_x|^\gamma f\|_{L^2}$ where $|D_x| = \sqrt{-\Delta_x}$ and set

$$||u(t,\cdot)||_\gamma^2 = \int \left(\big| \, |D_x|^{\gamma-1} u_t(t,x) \big|^2 + \big| \, |D_x|^\gamma u(t,x) \big|^2 \right) dx. \tag{1.2}$$

We want to find the smallest possible γ such that

$$(f,g) \in \dot{H}^\gamma(\mathbb{R}^n) \times \dot{H}^{\gamma-1}(\mathbb{R}^n), \tag{1.3}$$

$$\operatorname{supp} f \cup \operatorname{supp} g \subset \{x; |x| \le 2\} \tag{1.4}$$

implies that we have a local distributional solution of (1.1) for some $T > 0$, satisfying

$$(u, \partial_t u) \in C_b\big([0,T]; \dot{H}^\gamma(\mathbb{R}^n) \times \dot{H}^{\gamma-1}(\mathbb{R}^n) \big). \tag{1.5}$$

To avoid certain peculiarities concerning non-uniqueness we also require that u is a *proper solution*:

Definition 1.1. We say that u is a *proper solution* of (1.1) if it is a distributional solution and if in addition u is the weak limit of a sequence of smooth solutions u_ε to (1.1) with data $(\phi_\varepsilon * f, \phi_\varepsilon * g)$, where $\phi_\varepsilon(x) = \phi(x/\varepsilon)\varepsilon^{-n}$ for some function ϕ satisfying $\phi \in C_0^\infty$, $\int \phi \, dx = 1$.

Even if one has smooth data and hence a smooth solution there might still be another distributional solution which satisfies initial data in the space given by the norm (1.2). In fact, $u(t,x) = 2H(t-|x|)/t$ satisfies $\Box u = u^3$ in the sense of distribution theory. If $\gamma < 1/2$ then $||u(t,\cdot)||_\gamma \to 0$ when $t \to 0$ by homogeneity. Since $u(t,x) = 0$ is another solution with the same data it follows that we have non-uniqueness in the class (1.5) if $\gamma < 1/2$. Definition 1.1 picks out the smooth solution if there is one.

Our main theorem is the following:

THEOREM 1.2. *Consider the problem in* 3 *space dimensions,* $n = 3$*, with*

$$\begin{aligned} \Box u = \big(D^l u \big) D^{k-l} u, \quad D = (\partial_{x_1} - \partial_t), \\ u(0,x) = f(x), \quad u_t(0,x) = g(x), \end{aligned} \tag{1.6}$$

where $0 \le l \le k - l \le 2$*,* $l = 0, 1$*. Let* $\gamma = k$*. Then there are data* (f,g) *satisfying (1.3)-(1.4), with* $\|f\|_{\dot{H}^\gamma} + \|g\|_{\dot{H}^{\gamma-1}}$ *arbitrarily small, such that (1.6) does not have any proper solution satisfying (1.5) in* $S_T = [0,T) \times \mathbb{R}^3$ *for any* $T > 0$*.*

Remark 1.3. It follows from the proof of the theorem above that the problem is ill-posed if $\gamma = k$. In fact there exists a sequence of data $f_\varepsilon, g_\varepsilon \in C_0^\infty(\{x; |x| \le 1\})$

with $\|f_\varepsilon\|_{\dot H^\gamma} + \|g_\varepsilon\|_{\dot H^{\gamma-1}} \to 0$ such that if T_ε is the largest number such that (1.6) has a solution $u_\varepsilon \in C^\infty([0,T_\varepsilon)\times\mathbb{R}^3)$, we have that either $T_\varepsilon \to 0$ or else there are numbers $t_\varepsilon \to 0$ with $0 < t_\varepsilon < T_\varepsilon$ such that $\|u_\varepsilon(t_\varepsilon,\cdot)\|_\gamma \to \infty$. It also follows from the proof of the Theorem that either there is no distributional solution satisfying (1.5) with $\gamma = k$ or else we have non-uniqueness of solutions in (1.5).

Remark 1.4. By a simple scaling argument one gets a counterexample to well-posedness, but it has lower regularity than our counterexamples:

$$\gamma < k + \frac{n-4}{2}. \tag{1.7}$$

Indeed, if u is a solution of (1.6) which blows up when $t = T$ then $u_\varepsilon(t,x) = \varepsilon^{k-2}u(t/\varepsilon, x/\varepsilon)$ is a solution of the same equation with lifespan $T_\varepsilon = \varepsilon T$ and $\|u_\varepsilon(0,\cdot)\|_\gamma = \varepsilon^{k-2+n/2-\gamma}\|u(0,\cdot)\|_\gamma \to 0$ if γ satisfies (1.7). By contrast, our counterexamples are designed to concentrate in one direction, close to a characteristic. It appears that our construction has a natural generalization to any number of space dimensions n, with the initial data lying in $\dot H^\gamma$,

$$\gamma < k + \frac{n-3}{4}. \tag{1.8}$$

Remark 1.5. In Klainerman-Machedon[13,15] it was proved that for semi-linear wave equations satisfying the "null condition" one can in fact get local existence for data having the regularity (1.7) predicted by the scaling argument.

Now, there is a unique way to write (1.6) in the form

$$\sum_{j,k=0}^{3} g^{jk}(u)\partial_{x_j}\partial_{x_k}u = F(u,Du) \tag{1.9}$$

where $x_0 = t$ and $g^{jk}(u)$ are symmetric. In the semi-linear case $g^{jk} = m^{jk}$, where m^{jk} is given by (1.10). We now define the notion of a *domain of dependence*.

Definition 1.6. Assume that $\Omega \subset \mathbb{R}_+ \times \mathbb{R}^3$ is an open set equipped with a Lorentzian metric $g_{jk} \in C(\Omega)$ such that inverse g^{jk} satisfies

$$\sum_{j,k=0}^{3} |g^{jk} - m^{jk}| \le 1/2, \quad \text{where} \quad \begin{cases} m^{00} = 1, \;\; m^{jj} = -1, \; j > 0 \\ m^{jk} = 0, \quad \text{if} \quad j \neq k \end{cases}. \tag{1.10}$$

Then Ω is said to be a *domain of dependence for the metric* g_{ij} if for every compact subset $K \subset \Omega$ there exists a smooth function $\phi(x)$ such that the open set $\mathcal{H} = \{(t,x);\, t < \phi(x)\}$ satisfies

$$\overline{\mathcal{H}} \subset \Omega, \qquad K \subset \mathcal{H} \tag{1.11}$$

and $\partial\mathcal{H}$ is space-like, i.e.

$$\sum_{j,k=0}^{3} g^{jk}(t,x)N_j(x)N_k(x) > 0, \quad \text{if} \quad t = \phi(x), \quad N(x) = \big(1, -\nabla_x\phi(x)\big). \tag{1.12}$$

Since a solution u to (1.6) gives rise to a unique metric g_{jk} we say that Ω is a *domain of dependence for the solution* u if it is a domain of dependence for g_{jk}.

LEMMA 1.7. *There is an open set $\Omega \subset \mathbb{R}_+ \times \mathbb{R}^3$ and a solution $u \in C^\infty(\Omega)$ of (1.6) such that Ω is a domain of dependence and writing*

$$\Omega_t = \{x; (t,x) \in \Omega\}, \tag{1.13}$$

we have that $\partial\Omega_0$ is smooth,

$$\int_{\Omega_t} \left((\partial_{x_1} - \partial_t)^k u(t,x)\right)^2 dx = \infty, \quad t > 0, \qquad \text{and} \tag{1.14}$$

$$\sum_{|\beta| \le k} \int_{\Omega_t} \left(\partial^\beta u(t,x)\right)^2 dx < \infty, \quad \text{when} \quad t = 0, \quad \text{where} \tag{1.15}$$

$\beta = (\beta_0, ..., \beta_3)$ and $\partial^\beta = \partial^{\beta_0}/\partial x_0^{\beta_0} \cdots \partial^{\beta_3}/\partial x_3^{\beta_3}$. Furthermore in the quasi-linear case, $k - l = 2$, the norms $\|D^l u\|_{L^\infty(\Omega)}$ can be chosen to be arbitrarily small.

Proof of Theorem 1.2. By Lemmas 1.7 we get a solution $\overline{u}$ in a domain of dependence Ω with initial data $\overline{u}(0,x) \in H^k(\Omega_0)$ and $\overline{u}_t(0,x) \in H^{k-1}(\Omega_0)$. We can extend these to $f \in H^k(\mathbb{R}^3)$ and $g \in H^{k-1}(\mathbb{R}^3)$, see Stein[36]. If there exist a proper solution u of (1.6) in $S_T = [0,T] \times \mathbb{R}^3$ with these data, it follows from Definition 1.1 and Lemma 1.8 that u is equal to $\overline{u}$ in $S_T \cap \Omega$, contradicting (1.5).

LEMMA 1.8. *Suppose $u \in C^\infty(\Omega)$ is a solution to (1.6) where Ω is a domain of dependence. In the quasi-linear case, $k - l = 2$, assume also that $\|D^l u\|_{L^\infty(\Omega)} \le \delta$. Suppose also that $u_\varepsilon \in C^\infty(S_T)$, where $S_T = [0,T) \times \mathbb{R}^3$, and u_ϵ are solutions of (1.6) with data $(f_\varepsilon, g_\varepsilon)$ where $f_\varepsilon \to f$ and $g_\varepsilon \to g$ in $C^\infty(K_0)$ for all compact subsets of K_0 of $\Omega_0 = \{x; (0,x) \in \Omega\}$. Then $u_\varepsilon \to u$ in $\Omega \cap S_T$.*

It is essential that Ω is a domain of dependence for Lemma 1.8 to be true; one needs exactly the condition (1.12) in order to be able to use the energy method.

Let us now briefly describe how to construct the solution u and the domain of dependence Ω in Lemma 1.7. First we find a solution $u_1(t,x_1)$ for the corresponding equation in one space dimension, (1.16), which develops a certain singularity along a non time like curve $x_1 = \mu(t)$, with $\mu(0) = 0$. The initial data (1.17)-(1.18) has a singularity when $x_1 = 0$ and because of blow-up for the nonlinear equations, the singularity that develops for $t > 0$ is stronger than the singularity of data. Then $u(t,x) = u_1(t,x_1)$ is a solution of (1.6) in the set $\{(t,x);\, x_1 > \mu(t)\}$. The singularity of data is however too strong for the integral in (1.15) over this set to be finite when $t = 0$. Therefore we will construct a smaller domain of dependence, Ω, satisfying (1.20), such that the curve $x_1 = \mu(t)$, $x_2 = x_3 = 0$, still lies on $\partial\Omega$.

One can find rather explicit solution formulas for the one dimensional equations;

$$(\partial_{x_1} + \partial_t)(\partial_{x_1} - \partial_t)u_1(t,x_1) + (\partial_{x_1} - \partial_t)^l u_1(t,x_1)(\partial_{x_1} - \partial_t)^{k-l} u_1(t,x_1) = 0. \tag{1.16}$$

By choosing particular initial data

$$\begin{aligned}
&u_1(0,x_1) = \chi''(x_1), && \partial_t u_1(0,x_1) = 0, && \text{if } k = 0, l = 0,\\
&u_1(0,x_1) = -\chi'(x), && \partial_t u_1(0,x_1) = \chi''(x_1) + \chi'(x_1)^2, && \text{if } k = 1, l = 0,\\
&u_1(0,x_1) = 0, && \partial_t u_1(0,x_1) = -\chi^{(3-k)}(x_1), && \text{if } k \ge 2,
\end{aligned} \tag{1.17}$$

$$\text{where} \quad \chi(x_1) = \int_0^{x_1} -\varepsilon |\log|s/4||^\alpha \, ds, \quad 0 < \alpha < 1/2, \varepsilon > 0 \tag{1.18}$$

we get a solution

$$u_1 \in C^\infty(\Omega^1), \quad \text{where} \quad \Omega^1 = \{(t,x_1); \mu(t) < x_1 < 2-t\} \subset \mathbb{R}_+ \times \mathbb{R}^1 \tag{1.19}$$

for some function $\mu(t)$ with $\mu(0) = 0$, such that Ω^1 is a domain of dependence and such that $u_1(t,x_1)$ has a singularity along $x_1 = \mu(t)$. One sees this from the solution formulas which can be found in Lindblad[22,23]. Essentially what is happening is that the initial data (1.17)-(1.18) has a singularity when $x_1 = 0$. For the linear equation, $u_{tt} - u_{x_1x_1} = 0$, the singularity would just have propagated along a characteristic, however the nonlinearity causes the solution to increase and this strengthens the singularity for $t > 0$. (This is the same phenomena that causes blow-up for smooth initial data.)

Define $\Omega \subset \mathbb{R}_+ \times \mathbb{R}^3$ to be the largest domain of dependence for the metric obtained from the solution $u(t,x) = u_1(t,x_1)$ (see (1.9)), such that

$$\Omega \subset \Omega^1 \times \mathbb{R}^2, \quad \Omega_0 = \{x; (0,x) \in \Omega\} = B_0 = \{x; |x-(1,0,0)| < 1\}. \tag{1.20}$$

(It follows from Definition 1.6 that the union and intersection of a finite number of domains of dependence is a domain of dependence so indeed a maximal domain exists.) It follows that $u(t,x) = u_1(t,x_1)$ is a solution of (1.6) in Ω satisfying (1.17) in Ω_0. The initial data (1.17)-(1.18) was chosen so that (1.15) just is finite if $t = 0$

Let Ω_t be as in (1.13) and

$$S_t(x_1) = \{(x_2,x_3) \in \mathbb{R}^2;\ (x_1,x_2,x_3) \in \Omega_t\}, \quad a_t(x_1) = \int_{S_t(x_1)} dx_2\, dx_3. \tag{1.21}$$

With this notation the integral in (1.14) becomes

$$\int_{\mu(t)}^{2-t} a_t(x_1)\big((\partial_{x_1} - \partial_t)^k u_1(t,x_1)\big)^2\, dx_1. \tag{1.22}$$

The proof that this integral is infinite consists of estimating the two factors in the integrand from below, close to $x_1 = \mu(t)$.

In the semi-linear case the metric g^{jk} is just m^{jk} so Ω^1 is a domain of dependence if and only if $\mu'(t) \geq 1$ and it follows that $\Omega = \Omega^1 \times \mathbb{R}^2 \cap \Lambda$, where $\Lambda = \{(t,x); |x-(1,0,0)| + t < 1\}$. Hence for $x_1 > \mu(t)$; $S_t(x_1) = \{(x_2,x_3);\ (x_1-1)^2 + x_2^2 + x_3^2 < (1-t)^2\}$ so then $a_t(x_1) = \pi(2-t-x_1)(x_1-t)$. Also, the specific solution formulas are relatively simple. In particular if $k-l = l = 1$ then its easy to verify that

$$(\partial_{x_1} - \partial_t)u_1(t,x_1) = \frac{\chi'(x_1-t)}{1+t\chi'(x_1-t)}, \quad u_1(0,x) = 0 \tag{1.23}$$

satisfies (1.16)-(1.17) when $1 + t\chi'(x_1-t) > 0$. Since $\chi'(0+) = -\infty$ and $\chi'' > 0$ it follows that there is a function $\mu(t)$, with $\mu'(t) > 1$ and $\mu(0) = 0$, such that $1 + t\chi'(x_1-t) = 0$, when $x_1 = \mu(t)$. Hence $1 + t\chi'(x_1-t) \leq C(t)(x_1-\mu(t))$ so

$$\int_{\mu(t)}^{1/2} a_t(x_1)\big((\partial_{x_1} - \partial_t)u_1(t,x_1)\big)^2\, dx_1 \geq \int_{\mu(t)}^{1/2} \frac{(x_1-t)\, dx_1}{C(t)^2(x_1-\mu(t))^2} = \infty. \tag{1.24}$$

However, in the quasi-linear case, estimating $a_t(x_1)$ from below requires a detailed analysis of the characteristic set $\partial\Omega$ for the operator (1.25), see Lindblad[23].

$$\partial_t^2 - \sum_{i=1}^3 \partial_{x_i}^2 - V(\partial_{x_1} - \partial_t)^2, \text{ where } V = (\partial_{x_1} - \partial_t)^l u_1. \tag{1.25}$$

2. GLOBAL EXISTENCE

We will present sharp global existence theorems in all dimensions for small-amplitude wave equations with power-type nonlinearities. For a given "power" $p > 1$, we shall consider nonlinear terms F_p satisfying

$$\big|\,(\partial/\partial u)^j F_p(u)\,\big| \le C_j |u|^{p-j},\ j = 0, 1. \tag{2.1}$$

The model case, of course, is $F_p(u) = |u|^p$. If $\mathbb{R}^{1+n}_+ = \mathbb{R}_+ \times \mathbb{R}^n$, and if $f, g \in C_0^\infty(\mathbb{R}^n)$ are fixed, we shall consider Cauchy problems of the form

$$\begin{cases} \Box u = F_p(u),\ \ (t,x) \in \mathbb{R}^{1+n}_+ \\ u(0,x) = \varepsilon f(x),\ \partial_t u(0,x) = \varepsilon g(x), \end{cases} \tag{2.2}$$

where $\Box = \partial^2/\partial t^2 - \Delta_x$. Our goal is to find, for a given n, the range of powers for which one always has a global weak solution of (2.2) if $\varepsilon > 0$ is small enough.

In 1979, John [9] showed that for $n = 3$, (2.2) has global solutions if $p > 1 + \sqrt{2}$ and $\varepsilon > 0$ is small. He also showed that when $p < 1 + \sqrt{2}$ and $F_p(u) = |u|^p$ there is blow-up for most small initial data, see also [17]. It was shown later by Schaeffer [28] that there is blowup also for $p = 1 + \sqrt{2}$. After Johns work, Strauss made the conjecture in [38] that when $n \ge 2$, global solutions of (2.2) should always exist if ε is small and p is greater than a critical power p_c that satisfy

$$(n-1)p_c^2 - (n+1)p_c - 2 = 0,\ \ p_c > 1. \tag{2.3}$$

This conjecture was shortly verified when $n = 2$ by Glassey [5]. John's blowup results were then extended by Sideris [30], showing that for all n there can be blowup for arbitrarily small data if $p < p_c$. In the other direction, Zhou [43] showed that when $n = 4$, in which case $p_c = 2$, there is always global existence for small data if $p > p_c$. This result was extended to dimensions $n \le 8$ in Lindblad and Sogge [25]. Here it was also shown that, under the assumption of spherical symmetry, for arbitrary $n \ge 3$ global solutions of (2.2) exist if $p > p_c$ and ε is small enough. For odd spatial dimensions, the last result was obtained independently by Kubo [16]. The conjecture was finally proved in all dimensions by Georgiev-Lindblad-Sogge[4]. Here we will present that argument.

We shall prove Strauss conjecture using certain "weighted Strichartz estimates" for the solution of the linear inhomogeneous wave equation

$$\begin{cases} \Box w(t,x) = F(t,x), \quad (t,x) \in \mathbb{R}^{1+n}_+ \\ w(0,\,\cdot\,) = \partial_t w(0,\,\cdot\,) = 0. \end{cases} \tag{2.6}$$

This idea was initiated by Georgiev [3]. We remark that we only have to consider powers smaller than the conformal power $p_{\text{conf}} = (n+3)/(n-1)$ since it was already known that there is global existence for larger powers. See, e.g., [24].

Let us, however, first recall the inequality for (2.6), that John [9] used;

$$\|t(t-|x|)^{p-2} w\|_{L^\infty(\mathbb{R}^{1+3}_+)} \le C_p \|t^p (t-|x|)^{p(p-2)} F\|_{L^\infty(\mathbb{R}^{1+3}_+)},$$
$$\text{if}\ \ F(t,x) = 0,\ t - |x| \le 1,\ \text{and}\ \ 1 + \sqrt{2} < p \le 3.$$

Unfortunately, no such pointwise estimate can hold in higher dimensions due to the fact that fundamental solutions for $\Box$ are no longer measures when $n \ge 4$. Despite this, it turns out that certain estimates involving simpler weights which are invariant under Lorentz rotations (when $R = 0$) hold;

THEOREM 2.1. *Suppose that $n \geq 2$ and that w solves the linear inhomogeneous wave equation (2.6) where $F(t,x) = 0$ if $|x| \geq t + R - 1$, $R \geq 0$. Then*

$$\|((t+R)^2 - |x|^2)^{\gamma_1} w\|_{L^q(\mathbb{R}^{1+n}_+)} \leq C_{q,\gamma} \|((t+R)^2 - |x|^2)^{\gamma_2} F\|_{L^{q/(q-1)}(\mathbb{R}^{1+n}_+)}, \tag{2.7}$$

provided that $2 \leq q \leq 2(n+1)/(n-1)$ and

$$\gamma_1 < n(1/2 - 1/q) - 1/2, \text{ and } \gamma_2 > 1/q. \tag{2.8}$$

One should see (2.7) as a weighted version of Strichartz [39,40] estimate;

$$\|w\|_{L^{2(n+1)/(n-1)}(\mathbb{R}^{1+n}_+)} \leq C \|F\|_{L^{2(n+1)/(n+3)}(\mathbb{R}^{1+n}_+)}. \tag{2.9}$$

If one interpolates between this inequality and (2.7), one finds that the latter holds for a larger range of weights (see also our remarks for the radial case below). However, for the sake of simplicity, we have only stated the ones that we will use.

Let us now give the simple argument showing how our inequalities imply the proof of Strauss conjecture. Let $u_{-1} \equiv 0$, and for $m = 0, 1, 2, 3, \ldots$ let u_m be defined recursively by requiring

$$\begin{cases} \Box u_m = F_p(u_{m-1}) \\ u_m(0,x) = \varepsilon f(x), \ \partial_t u_m(0,x) = \varepsilon g(x), \end{cases}$$

where $f, g \in C_0^\infty(\mathbb{R}^n)$ vanishing outside the ball of radius $R-1$ centered at the origin are fixed. Then if $p_c < p \leq (n+3)/(n-1)$, we can find γ satisfying

$$1/p(p+1) < \gamma < ((n-1)p - (n+1))/2(p+1). \tag{2.10}$$

Set

$$A_m = \|((t+R)^2 - |x|^2)^\gamma u_m\|_{L^{p+1}(\mathbb{R}^{1+n}_+)}. \tag{2.11}$$

Because of the support assumptions on the data, domain of dependence considerations imply that u_m, and hence $F_p(u_m)$, must vanish if $|x| > t + R - 1$. It is also standard that the solution u_0 of the free wave equation $\Box u_0 = 0$ with the above data satisfies $u_0 = O(\varepsilon(1+t)^{-(n-1)/2}(1 + |t - |x||)^{-(n-1)/2})$. Using this one finds that $A_0 = C_0 \varepsilon < \infty$. It follows from (2.10) that

$$\gamma < n(1/2 - 1/q) - 1/2, \text{ and } p\gamma > 1/q, \quad \text{if} \quad q = p+1, \tag{2.12}$$

so if we apply (2.7) to the equation $\Box(u_m - u_0) = F_p(u_{m-1})$ we therefore obtain

$$\begin{aligned} &\|((t+R)^2 - |x|^2)^\gamma u_m\|_{L^{p+1}} \\ &\quad \leq \|((t+R)^2 - |x|^2)^\gamma u_0\|_{L^{p+1}} + C_1 \|((t+R)^2 - |x|^2)^{p\gamma} |u_{m-1}|^p\|_{L^{(p+1)/p}} \\ &\quad = \|((t+R)^2 - |x|^2)^\gamma u_0\|_{L^{p+1}} + C_1 \|((t+R)^2 - |x|^2)^\gamma u_{m-1}\|^p_{L^{p+1}}, \end{aligned}$$

i.e. $A_m \le A_0 + C_1 A_{m-1}^p$. From this we can inductively deduce that $A_m \le 2A_0$, for all m, if $A_0 = C_0\varepsilon$ is so small that $C_1(2A_0)^p \le A_0$. Similarly, we can get bounds for differences showing that $\{u_m\}$ is a Cauchy sequence in the space associated with the norm (2.11), so the limit exists and satisfies (2.2).

The proof of Theorem 2.1 uses a decomposition into regions, where the weights $(t^2 - |x|^2)$ are essentially constant, together with the invariance of the norms and the equation under Lorentz transformations. In each case we get the desired estimate by using analytic interpolation, Stein[35], between an $L^1 \to L^\infty$ and an $L^2 \to L^2$ estimate with weights, for the Fourier integral operators associated with the wave equation. See [4] for the complete proof and further references. In [4] we also prove a stronger scale invariant weighted Strichartz estimate under the assumption of radial symmetry. This assumption was later removed by Tataru[41]

THEOREM 2.2. *Let n be odd and assume that F is spherically symmetric and supported in the forward light cone $\{(t,x) \in \mathbb{R}^{1+n} : |x| \le t\}$. Then if w solves* (2.6) *and if* $2 < q \le 2(n+1)/(n-1)$

$$\|(t^2-|x|^2)^{-\alpha} w\|_{L^q(\mathbb{R}^{1+n}_+)} \le C_\gamma \|(t^2-|x|^2)^{\beta} F\|_{L^{q/(q-1)}(\mathbb{R}^{1+n}_+)}, \tag{2.13}$$
$$\text{if } \beta < 1/q, \quad \alpha+\beta+\gamma = 2/q, \quad \text{where} \quad \gamma = (n-1)(1/2-1/q).$$

REFERENCES

1. Bahouri-Chemin, oral communication (1998).
2. D. Christodoulou and S. Klainerman, *The global nonlinear stability of the Minkowski space-time*, Princeton University Press, 1993.
3. V. Georgiev, *Weighted estimate for the wave equation*, Nonlinear Waves, Proceedings of the Fourth MSJ International Research Institute, vol. 1, Hokkaido Univ., 1996, pp. 71–80.
4. V. Georgiev, H. Lindblad and C. Sogge, *Weighted Strichartz estimates and global existence for semi-linear wave equations*, Amer. J. Math. **119** (1997), 1291–1319.
5. R. Glassey, *Existence in the large for $\Box u = F(u)$ in two dimensions*, Math. Z. **178** (1981).
6. M. Grillakis, *A priory estimates and regularity of nonlinear waves Proceedings of the international Congress of Mathematics, (Zürich, 1994)* (1995), Birkäuser, 1187–1194.
7. L. Hörmander, *Fourier integrals I*, Acta Math. **127** (1971), 79–183.
8. ———, *Lectures on nonlinear hyperbolic differntial equations*, Springer Verlag, 1997.
9. F. John, *Blow-up of solutions of nonlinear wave equations in three space dimensions*, Manuscripta Math. **28** (1979), 235–265.
10. F. John and S. Klainerman, *Almost global existence to nonlinear wave equations in three space dimensions*, Comm. Pure Appl. Math. **37** (1984), 443-455.
11. S. Klainerman, *Uniform decay estimates and the Lorentz invariance of the classical wave equation*, Comm. Pure Appl. Math. **38** (1985), 321-332.
12. ———, *Long time behaviour of solutions to nonlinear wave equations Proceedings of the International Congress of Mathematics, (Warsaw, 1983)* (1984), PWN, Warsaw, 1209-15.
13. S. Klainerman and M. Machedon, *Space-time estimates for null-forms and the local existence theorem*, Comm. Pure and Appl. Math. (1993), 1221-1268.
14. ———, *Finite Energy solutions of the Yang-Mills equations in $\mathbf{R}^{3+1}$*, Ann. of Math **142** (1995), 39–119.
15. ———, *Smoothing estimates for null forms and applications*, Duke Math. J. **81** (1996).
16. H. Kubo, *On the critical decay and power for semi-linear wave equations in odd space dimensions*, preprint.
17. H. Lindblad, *Blow up for solutions of $\Box u = |u|^p$ with small initial data*, Comm. Partial Differential Equations **15** (1990), 757–821.

18. ———, *On the lifespan of solutions of nonlinear wave eq.*, Comm. Pure Appl. **43** (1990).
19. ———, *Global solutions for nonlinear wave equations with small initial data*, Comm. Pure Appl. **45** (1992), 1063-1096.
20. ———, *A sharp counterexample to local existence of low regularity solutions to nonlinear wave equations*, Duke Math. J. **72** (1993), 503–539.
21. ———, *Counterexamples to local existence for nonlinear wave equations. Journées "Équations aux Dérivées Partielles" (Saint-Jean-de-Monts, 1994)*, École Polytech., Palaiseau (1994), Exp. No. X.
22. ———, *Counterexamples to local existence for semi-linear wave equations*, Amer. J. Math. **118** (1996), 1–16.
23. ———, *Counterexamples to local existence for quasi-linear wave equations*, preprint.
24. H. Lindblad and C. D. Sogge, *On existence and scattering with minimal regularity for semi-linear wave equations*, J. Funct. Anal. **130** (1995), 357–426.
25. ———, *Long-time existence for small amplitude semi-linear wave equations*, Amer. J. Math. **118** (1996), 1047–1135.
26. G. Ponce and T. Sideris, *Local regularity of nonlinear wave equations in three space dimensions*, Comm. Partial Differential Equations **18** (1993), 169–177.
27. J. Rauch, *Explosion for some Semilinear Wave Equations*, Jour. of Diff. Eq. **74 (1)** (1988).
28. J. Schaeffer, *The equation $\Box u = |u|^p$ for the critical value of p*, Proc. Royal Soc. Edinburgh **101** (1985), 31–44.
29. I. Segal, *Space-time decay for solutions of wave equations*, Adv. Math. **22** (1976), 305–311.
30. T. Sideris, *Nonexistence of global solutions to semi-linear wave equations in high dimensions*, Comm. Partial Diff. Equations **12** (1987), 378–406.
31. J. Shatah and A. Shadi Tahvildar-Zadeh, *On the Cauchy Problem for Equivariant Wave Maps*, CPAM **47** (1994), 719-754.
32. Sogge, *On local existence for nonlinear wave equations satisfying variable coefficient null conditions*, Comm. Partial Diff. Equations **18** (1993), 1795–1823.
33. ———, *Lectures on nonlinear wave equations*, International Press, Cambridge,, 1995.
34. C. D. Sogge and E. M. Stein, *Averages of functions over hypersurfaces: Smoothness of generalized Radon transforms*, J. Analyse Math. **54** (1990), 165–188.
35. E. M. Stein, *Interpolation of linear operators*, Trans. Amer. Math. Soc. **83** (1956), 482–492.
36. E. M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Univ..
37. W. Strauss, *Nonlinear scattering theory*, Scattering theory in mathematical physics, Reidel, Dordrect, 1979, pp. 53–79.
38. ———, *Nonlinear scattering at low energy*, J. Funct. Anal. **41** (1981), 110-133.
39. R. Strichartz, *A priori estimates for the wave equation and some applications*, J. Funct. Analysis **5** (1970), 218–235.
40. ———, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J. **44** (1977), 705–714.
41. Tataru, *Strichartz estimates in the hyperbolic space and global existence for the semi-linear wave equation*, preprint (1997).
42. ———, *oral communication* (1998).
43. Y. Zhou, *Cauchy problem for semi-linear wave equations with small data in four space dimensions*, J. Diff. Equations **8** (1995), 135–14444.

Hans Lindblad
UCSD Department of Mathematics
9500 Gilman Drive
La Jolla, CA 92093-0112
USA
lindblad@math.ucsd.edu

Fourier Analysis of Null Forms and Non-linear Wave Equations

M. Machedon

Abstract. The non-linear terms of many equations, including Wave Maps and Yang-Mills have a special, "null", structure. In joint work with Sergiu Klainerman, I use techniques of Fourier Analysis, such as generalizations and refinements of the restriction theorem applied to null forms to study the optimal Sobolev space in which such non-linear wave equations are well posed.

1991 Mathematics Subject Classification: 35, 42
Keywords and Phrases: Null forms, restriction theorem

The following notation will be used: repeated indices are summed, $x^\alpha, 0 \leq \alpha \leq n$ are the coordinates $t, x^i, 1 \leq i \leq n$, ∇_α are the usual derivatives, and indeces are raised or lowered according to the Minkowski metric $-1, 1, \cdots, 1$ (i. e. raising or lowering the 0 index changes sign).

Wave maps are functions $\phi : \mathbb{R}^{n+1} \to M$ from Minkowski space $\mathbb{R}^{n+1}$ to a Riemannian manifold M with metric g which arise as critical points of the Lagrangian

$$\int_{R^{n+1}} \big(\nabla_\alpha \phi, \nabla^\alpha \phi\big)_g \tag{1}$$

The Euler-Lagrange equations of the above, written in coordiantes on M are

$$\Box \phi^i + \Gamma^i_{jk}(\phi)\big(\nabla_\alpha \phi^j, \nabla^\alpha \phi^k\big) = 0 \tag{2}$$

where $\Gamma^i_{j,k}$ are the Christoffel symbols. We see the first null form

$$Q_0(\phi, \psi) = \nabla_\alpha \phi \nabla^\alpha \psi$$

arising as part of the non-linear term. There is more going on than one can read off from (2). In the special case of $M = S^{k-1} \subset \mathbb{R}^k$ the equation (2) can also be written as

$$\Box \phi + \phi\big(\nabla_\alpha \phi \cdot \nabla^\alpha \phi\big) = 0 \tag{3}$$

with constraint $|\phi| = 1$. Take dot product with ϕ_t. Because of the constraint, the non-linear term drops out and we get conservation of energy, just as for the linear wave equation. In the special case of $n = 2$, it is still an open question whether the Cauchy problem (2) is well posed globally in time. Related to that is the question whether (2) is well posed locally in time for small data in H^1. Because of conservation of energy, such a result would prove global in time regularity for solutions of (2) with smooth data with small energy. There is a lot of evidence that using both the null condition and the geometric condition used in (3) the wave map equation should be well posed locally in time for Cauchy data in $H^{n/2}$. There has been a lot of work in recent years on this question. We currently have the result for $H^{n/2+\epsilon}$, see [K-M 4], [K-S], [G] . These results use only the null condition, and such a result fails by half a derivative for equations not satisfying the null condition, see [L]. Also, the sharp $H^{n/2}$ result cannot be true for general equations of the type (2), without using geometric information about the target manifold, as the example of geodesic solutions shows: Let $\gamma(t)$ be a geodesic on M which blows up in finite time, and let ψ be a solution of the homogeneous wave equations $\Box\psi = 0$ with $H^{n/2}$ data. Now, $\phi = \gamma(\psi)$ is a solution of (2). Since the supremum of ψ can become large, ϕ can blow up instantly.

The definitive result on equations of the type (2) which does not take the geometric condition into account is well posedness in the Besov space $B^{2,1}_{n/2}$, due to Daniel Tataru [T2]. For related applications of the geometric condition see [F-M-S], [Sh].

The Yang-Mills equations are non-linear analogues of the Maxwell equations. Let G be one of the classical compact Lie groups, and g its Lie algebra. The unknown is a connection potential $A_\alpha : \mathbb{R}^{n+1} \to g$, such that the corresponding covariant derivative $D_\alpha = \partial_\alpha + [A_\alpha,]$ satisfies

$$D^\alpha F_{\alpha,\beta} = 0$$

where the curvature $F_{\alpha,\beta} = [D_\alpha, D_\beta]$

Here we have gauge freedom: if A_α is a solution, and O is a G-valued function, then $OA_\alpha O^{-1} - \partial_\alpha O O^{-1}$ is also a solution. Thus we may impose an additional gauge condition on A_α. We choose the Coulomb gauge : $\partial^i A_i = 0$. Then we have

$$\Box A_i = -2[A_j, \partial_j A_i] + [A_j, \partial_i A_j] + \cdots$$

together with an elliptic equation for A_0. The dots turn out to be less important terms. We will now identify the null forms in the right hand side. They will involve $Q_{ij}(\phi, \psi) = \nabla_i\phi\nabla_j\psi - \nabla_j\phi\nabla_i\psi$. In fact using the divergence condition on A to express it as curl$B$, the first term is of the type $Q_{ij}(B, A)$. Similarly, the curl of the second term is of the type $Q_{ij}(A, A)$, which is all the information we need since the divergence of the whole right hand side is 0. Thus a simplified model for Yang-Mills is

$$\Box A = Q_{ij}((-\Delta)^{-1/2}A, A) + (-\Delta)^{-1/2}Q_{ij}(A, A) \tag{4}$$

The indeces of A are not important, and have been supressed.

The Yang-Mills equations in 3+1 dimensions are sub-critical. There is a conserved energy, and our local existence result implies that the time of existence of a smooth solution depends only on the energy of the initial data (and the solution stays as smooth as it started in this interval). The argument is complicated by gauge dependance, and the fact that energy differs form the H^1 norm by a lower order term, see [K-M3]. The global existence result was already known, due to Eardley and Moncrief [E-M]. However, our new techniques also give global existence in the energy space. It was shown by M. Keel [Ke], along the same lines, that there is global regularity for Yang-Mills coupled with a critical power Higgs field. This is a new global existence result, accessible only through our new local estimates.

In 4+1 dimensions, Yang-Mills are critical, and it was shown by Klainerman and Tataru that they are well posed in $H^{1+\epsilon}$ [K-T]. See also [K-M8] for a related result.

Following is a summary of the main estimates used in the above proofs.

Recall the classical Strichartz inequality gives the (optimal) estimate for a solution of $\Box\phi = 0$

$$\|(\nabla\phi)^2\|_{L^3(\mathbf{R}^3)} \leq C\big(\|\phi(0,\cdot)\|^2_{H^{3/2}(\mathbf{R}^2)} + \|\phi_t(0,\cdot)\|^2_{H^{1/2}(\mathbf{R}^2)}\big)$$

However, for a null form we have

$$\|Q(\phi,\phi)\|_{L^2(\mathbf{R}^3)} \leq C\big(\|\phi(0,\cdot)\|_{H^{5/4}(\mathbf{R}^2)} + \|\phi_t(0,\cdot)\|_{H^{1/4}(\mathbf{R}^2)}\big)$$

The proof is based on writing the L^2 norm of the quadratic form as the L^2 norm of a convolution of measures supported on the light cone, on the Fourier transform side. The symbol of the null form kills the worst singularity in the convolution. This has been generalized to the variable coefficient case by C. Sogge [So]. Some ideas in the proof were also used in [Sc-So].

Using this type of estimate one can prove that (2) is well posed in $H^{(n+1)/2}$, which is already non-trivial, is only true for equations satisfying some kind of null condition (for n=2, 3), but is not optimal. Also, the same techniques give local existence for finite energy data for Yang-Mills in 3+1 dimensions.

To get to the optimal result, that the Wave Map equation (2) is well posed in $H^{n/2+\epsilon}$ we have to make extensive use of the spaces $H_{s,\delta}$ used by Bourgain for KdV [B]; see also [Be]:

$$\|\phi\|_{s,\delta} \ = \|w_+^s w_-^\delta \tilde{\phi}\|_{L^2(d\tau d\xi)}$$

where $w_+(\tau,\xi) = 1+|\tau|+|\xi|$, $w_-(\tau,\xi) = 1+||\tau|-|\xi||$, and $\tilde{\phi}$ denotes the space-time Fourier transform. Also, let $D_\pm$ be the operator with symbol $w_\pm$. There are two advantages in working with these spaces. Functions in $H_{s,\delta}$ with $\delta > 1/2$ satisfy the same Strichartz-Pecher estimates that solutions of $\Box\phi = 0$ with H^s Cauchy data would.

In 3+1 dimensions, for instance,

$$\|\phi\|_{L^\infty(dt)L^2(dx)} \leq C\|\phi\|_{0,\delta} \tag{5a}$$

is the energy estimate, and all estimates obtained by interpolating it with the (false) end-point result

$$\|\phi\|_{L^2(dt)L^\infty(dx)} \leq C\|\phi\|_{1,\delta} \tag{5b}$$

are true.

Also, the argument is simplified if one also notices that, for $\delta < 1/2$ and p defined by $\frac{1}{p} = \frac{1}{2} - \delta$,

$$\|\phi\|_{L^p(dt)L^2(dx)} \leq C\|\phi\|_{0,\delta} \tag{5c}$$

See [T1] for a general treatment of these spaces.

The second advantage of the spaces $H^{s,\delta}$ is that the solution to $\Box\phi = F$ with Cauchy data f_0, f_1 satisfies

$$\|\chi(t)\phi\|_{s,\delta} \leq C\bigg(\|F\|_{s-1,\delta-1} + \|f_0\|_{H^s} + \|f_1\|_{H^{s-1}}\bigg)$$

where χ is a smooth cut-off function in time. In order to solve $\Box\phi = Q(\phi,\phi)$ for small time it suffices to solve the integral equation

$$\phi = \chi(t)\bigg(W * Q + W(f_0) + \partial_t W(f_1)\bigg) \tag{6}$$

W is the fundamental solution of $\Box$. This idea also goes back to Bourgain. See also [K-P-V].

In order to show that the equation (2) is well posed in H^s, for $s > 3/2$, in 3+1 dimensions, it suffices to prove an inequality of the form

$$\|Q_0(\phi,\psi)\|_{s-1,\delta-1} \leq C\|\phi\|_{s,\delta}\|\psi\|_{s,\delta} \tag{7}$$

where $\delta > 1/2$

The symbol of the null form Q_0 is

$$\tau\lambda - \xi\cdot\eta = \frac{1}{2}\bigg((\tau+\lambda)^2 - |\xi+\eta|^2 - \tau^2 + |\xi|^2 - \lambda^2 + |\eta|^2\bigg)$$

Using this, the left hand side of (4) is dominated by the sum of terms, a typical one being

$$\|D_-^{\delta-1}((D_+^s D_-^{1/2}\phi)(D_+^{1/2}\psi))\|$$

Estimate this norm by duality, integrating against $F \in L^2$:

$$\int D_-^{\delta-1}((D_+^s D_-^{1/2}\phi)(D_+^{1/2}\psi))F$$
$$= \int (D_+^s D_-^{1/2}\phi)(D_+^{1/2}\psi))D_-^{\delta-1}F$$

The first term is in L^2, the second one in $H_{s-1/2,\delta}$ and the third one in $H_{0,1-\delta}$. Thus, it suffices to show $H_{0,1-\delta} \cdot H_{s-1/2,\delta} \subset L^2$. This is true, and follows from (5 a, b, c). In fact, for there exist p close to ∞, $q > 2$, close to 2, $\frac{1}{p} + \frac{1}{q} = \frac{1}{2}$ such that the first term is in $L^p(dt)L^2(dx)$ and the second one in $L^q(dt)L^\infty(dx)$. The original argument of [K-M4] used convolutions of measures.

An problem related to Yang-Mills, worked out in [K-M6], to show that the model

$$\Box\phi = Q_{ij}(\phi, \phi) \tag{8}$$

is well posed in $H^{3/2+\epsilon}$ in 3+1 dimensions.

The analogue of (7) is not true. There is an estimate for the symbol $|\xi \times \eta| \leq |\xi|^{1/2}|\eta|^{1/2}|\xi+\eta|^{1/2}(w_-(\tau,\xi)+w_-(\lambda,\eta)+w_-(\tau+\lambda,\xi+\eta)^{1/2}$, but after distributing derivatives as above one has to bound a troublesome term

$$\|D_-^{-1/2}((D_-^{1/2}D_+^{1/2}\phi)(D_+^s\psi))\|_{L^2}$$

By duality, this would correspond to an estimate

$$D^{-(s-1/2)}\Big(H_{0,1/2} \cdot H_{0,1/2}\Big) \subset L^2$$

($s > 3/2$). This is false, the counterexample is an adaptation of an old construction due to A. Knapp. There are other useful estimates along these lines which are true, and which are needed for (4), (8), see [K-M5], [K-T]. In 3+1 dimensions the (barely false) end-point estimates are are

$$D^{-1/2}\Big(H_{1/4,\delta} \cdot H_{1/4,\delta}\Big) \subset L^2 \tag{9a}$$

and

$$D^{-1}\Big(H_{1/2,\delta} \cdot H_{1/2,\delta}\Big) \subset L^1(dt)L^\infty(dx) \tag{9b}$$

Back to (5), we are foced to make stronger assumptions on our norms. A simplification of the original argument in [K-M6], used in [K-T], is to require, (modulo an ϵ) that, in addition to $\phi \in H_{s,1/2}$, ϕ should also satisfy

$$\|\phi\|_* = \inf\{\|F\|_{L^1(dt)L^\infty(dx)},\ |\widetilde{(D_-^{1/2}D_+^{1/2}\phi)}| \leq |\tilde{F}|\} < \infty$$

These norms are constructed so that we recover (7)

$$\|Q_{ij}(\phi,\psi)\|_{s-1,\delta-1} \leq C\big(\|\phi\|_{s,\delta} + \|\phi\|_*\big)\big(\|\psi\|_{s,\delta} + \|\psi\|_*\big) \tag{7'}$$

and it turns out also

$$\|D_+^{-1}D_-^{-1}Q_{ij}(\phi,\psi)\|_* \leq C\big(\|\phi\|_{s,\delta} + \|\phi\|_*\big)\big(\|\psi\|_{s,\delta} + \|\psi\|_*\big)$$

These types of modified norms also work for the model Yang-Mills problem (4), to prove well posedness in $H^{1+\epsilon}$ in $4+1$ dimensions. To prove the necessary estimates one must use the analogues of 9a, 9b in $4+1$ dimensions.

References

[Be] M. Beals Self-spreading and strength of singularities for solutions to semilinear wave equations, Annals of Math 118, 1983 no1 187-214

[B] J. Bourgain Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, I, II, Geom. Funct. Analysis 3, (1993), 107-156, 202-262.

[C-Z] D. Christodoulou, A. Tahvildar-Zadeh, On the regularity properties of spherically symmetric wave maps, Comm. Pure Appl. Math. 46, (1993) 1041-1091.

[F-M-S] A. Freire, S. Muller, M Struwe, Weak convergence of harmonic maps, Invent. Math. 130, 1997, no3, 589-617

[G] M. Grillakis, A priori estimates and regularity of non-linear waves, Proceedings of ICM, Zurich, 1994.

[H] F. Helein, Regularity of weakly harmonic maps from a surface into a manifold with symmetries, Manuscripta Math. 70, (1991), 203-218.

[Ke] M. Keel, Global existence for critical power Yang-Mills-Higgs equations in R^{3_1} Comm. in PDE 22 (1997) no 7-8 1161 – 1225

[K-P-V] K. Kenig, G. Ponce, L. Vega The Cauchy problem for the Korteveg-De Vries equation in Sobolev spaces of negative indices, Duke Math Journal 71, no 1, pp 1-21 (1994)

[K-M1] S. Klainerman and M. Machedon Space-time estimates for null forms and the local existence theorem, Comm. Pure Appl. Math, vol XLVI, 1221-1268 (1993)

[K-M2], S. Klainerman and M. Machedon On the Maxwell-Klein-Gordon equation with finite energy, Duke Math Journal, vol. 74, no. 1 (1994)

[K-M3] S. Klainerman and M. Machedon Finite energy solutions of the Yang-Mills equations in $\mathbf{R}^{3+1}$, Annals of Math. 142, 39-119 (1995)

[K-M4] S. Klainerman and M. Machedon Smoothing estimates for null forms and applications, Duke Math Journal, 81, no 1, in celebration of John Nash, 99-133 (1996) Also 1994 IMRN announcement.

[K-M5] S. Klainerman and M. Machedon with appendices by J. Bourgain and D. Tataru, Remark on the Strichartz inequality, International Math Research Notices no 5, 201-220 (1996).

[K-M6] S. Klainerman and M. Machedon Estimates for null forms and the spaces $H_{s,\delta}$ International Math Research Notices no 17, 853-865 (1996).

[K-M7] S. Klainerman and M. Machedon On the regularity properties of a model problem related to wave maps, Duke Math Journal 87, (1997) no 3, 553–589

[K-M8] S. Klainerman and M. Machedon On the optimal local regularity for gauge field theories, Differential and Integral Equations 10, (1997) no. 6, 1019–1030

[K-S] S. Klainerman, S. Selberg Remark on the optimal regularity for equations of Wave Maps type, Comm PDE. 22 (1997) no 5-6, 901-918

[K-T] S. Klainerman and D. Tataru, On the local regularity for Yang-Mills equations in $\mathbf{R}^4 + 1$, preprint

[L] H. Lindblad Counterexamples to local existence for semi-linear wave equations, Amer. J. Math, 118 (1996) no 1 1-16

[Sh] J. Shatah Weak Solutions and development of singularities for the $SU(2)\sigma$-Model, Comm. Pure Appl. Math, vol XLI 459-469 (1988)

[Sch-So] W. Schlag, C Sogge Local smoothing estimates related to the circular maximal theorem. Math Res. Letters 4 (1997) no 1-15

[So] C. Sogge On local existence for non-linear wave equation satisfying variable coefficient null condition, Comm PDE 18 (1993) n0o 11 1795-1821

[T1] D. Tataru, The X^s_θ spaces and unique continuation for solutions to the semi-linear wave equation, Comm. PDE, 21 (5-6), 841-887 (1996)

[T2] D. Tataru, Local and global results for wave maps, preprint

M. Machedon
University of Maryland
College Park, Md 20742
USA

Blow-up Phenomena for Critical Nonlinear Schrödinger and Zakharov Equations

Frank Merle

Abstract. In this paper, we review qualitative properties of solutions of critical nonlinear Schrödinger and Zakharov equations which develop a singularity in finite time.

1991 Mathematics Subject Classification: 35Bxx, 35Qxx
Keywords and Phrases: Blow-up, Critical, Schrodinger

I. *The Problem*

We are interested in the formation of singularities in time, in Hamiltonian systems of infinite dimension, and with infinite speed of propagation. A prototype is the nonlinear Schrödinger equation

$$\left\{ \begin{array}{rcl} iu_t & = & -\Delta u - |u|^{p-1}u, \\ u(0) & = & u_0, \end{array} \right. \tag{1}$$

for $(x,t) \in R^N \times [0,T)$ and u=0 at infinity. This equation appears in various situations in physics (plasma physics, nonlinear optics,...see [20] for example). Because of its importance in physics, we are interested in the case where $p-1 = 4/N$ and $N = 2$. We will consider

$$iu_t = -\Delta u - |u|^{\frac{4}{N}}u. \tag{2}$$

Equation (1) has Galilean, scaling, and translation invariances. In the case $p = \frac{4}{N} + 1$, the nonlinear equation has the same structure as the linear equation: it has one more invariance (the conformal invariance): if $u(t)$ is a solution of (2) then $v(t) = \frac{1}{t^{N/2}} e^{+i\frac{|x|^2}{4t}} \overline{u}\left(\frac{1}{t}, \frac{x}{t}\right)$ is also a solution of (2). Thus, there are three invariants of the motion in this case: the mass $|u|_{L^2}$, the energy $E(u) = \frac{1}{2}\int_{R^N} |\nabla u|^2 dx - \frac{1}{\frac{4}{N}+2}\int_{R^N} |u|^{\frac{4}{N}+2} dx$, and the energy of v, $E(v)$.

A more refined physical model is also considered: the Zakharov equation (nonlinear Schrödinger equation coupled with the wave equation). Because of the coupling, all invariances disappear. The system is

$$\left\{ \begin{array}{rcl} iu_t & = & -\Delta u + nu, \\ n_t & = & -\nabla \cdot v, \\ \frac{1}{c_0^2} v_t & = & -\nabla(n + |u|^2), \end{array} \right. \tag{3}$$

where $(x,t) \in R^2 \times [0,T)$.

We note formally that if $c_0 = +\infty$, system (3) reduces to equation (2) in dimension two. There are two invariants: $|u|_{L^2}$, and $H(u,n,v) = \int |\nabla u|^2 dx + \int n|u|^2 dx + \frac{1}{2}\int n^2 dx + \frac{1}{2c_0{}^2}\int |v|^2 dx$.

The first natural question concerns the local wellposedness of the equations in time. The natural spaces for this equation are spaces where the conserved quantities are defined. For the Schrödinger equation, H^1 local wellposedness has been proved in [10], [11], [14]. The use of Strichartz estimates (of space-time nature, where the role of space and time are similar) leads to the result in L^2 in [8] (L^2 is optimal in some sense, see [2]). This space will play a crucial role in the analysis below. See [4],[5] in the periodic case.

For the system (3), the coupling between the two equations creates several difficulties. In energy space, that is $(u,n,n_t) \in H_1 = H^1 \times L^2 \times L^2$, the local wellposedness was proved in [3],[9]. The problem to be solved in the analogue of L^2 for the Schrödinger equation is still open (an intermediate space was found in [9]).

The problem we are interested in concerns the description of solutions of equations (2),(3) which develop a singularity in finite time (or blow up in finite time). That is, solutions such that in the time dynamics, the nonlinear terms play an important role. This question is important from the physical point of view. Indeed, equations (2) or (3) appear as simplifications of more complex models. In particular, one hopes that the simplification is relevant for regular solutions, and that close to the singularity, the neglected terms will play a role . Blow up in finite time means that the regular regime where the approximation is carried out is unstable in time, and close to singularity, a transitory regime appears. From the description of this transitory regime, one can hope to find the new dynamics relevant from the physical point of view. In particular, a crucial question, after the *existence* of singularity in finite time, is to describe *how* this singulary forms.

For equation (2), there are two elementary results about existence of blow-up solutions.

On one hand, in 1972 Zakharov derived in [33] (see also [13],[28]) a Pohozaev type identity for the nonlinear Schrödinger equation: let $u_0 \in \Sigma$ where $\Sigma = H^1 \cap \{xu_0 \in L^2\}$; then for all t, $u(t) \in \Sigma$ and

$$\frac{d^2}{dt^2}\int |x|^2|u|^2 dx = 16E(u_0). \tag{4}$$

It follows that if $E(u_0) < 0$ then $u(t)$ blows up in finite time. Note that the power appearing in (2) is the smallest power such that blow-up occurs in H^1.

On the other hand, the elliptic theory established in the 80's ([1],[31],[17], [30]) yields the existence of one explicit solution of (2), periodic in time and of the form $P(t,x) = e^{it}Q(x)$, where Q is the unique positive solution (up to translation) of the equation

$$u = \Delta u + |u|^{\frac{4}{N}}u, \tag{5}$$

whose L^2 norm is characterized by the Gagliardo-Niremberg inequality

$$\forall v \in H^1, \frac{1}{\frac{4}{N}+2}\int |v|^{\frac{4}{N}+2}dx \leq \frac{1}{2}\int |\nabla v|^2 dx \left(\frac{\int |v|^2 dx}{\int Q^2 dx}\right)^{\frac{2}{N}}. \tag{6}$$

From the conformal invariance, we have that

$$S(t,x) = \frac{1}{t^{N/2}} e^{-\frac{i}{t}+i\frac{|x|^2}{4t}} Q\left(\frac{x}{t}\right) \tag{7}$$

is a blow-up solution of equation (2). This is in some sense the only explicit blow-up solution for the critical Schrödinger equation.

Until the 90's, no rigorous results on blow-up were known for the Zakharov equation.

II. *Results for Nonlinear Critical Schrödinger Equations.*

II.1 *Characterization of the minimal blow-up solution.*

The first task is to define a notion of smallness such that u_0 small implies no blow-up. In the case $u_0 \in H^1$, energy conservation and (6) yield that if $|u_0|_{L^2} < |Q|_{L^2}$, the solution is globally defined. Moreover, we note that the blow-up solution $S(t)$ is such that $|S(t)|_{L^2} = |Q|_{L^2}$. The natural question is to characterize all minimal blow-up solutions in L^2 of equation (2).

a) *The result.*

We have the following theorem

THEOREM 1 ([25]),([26])

Let $u_0 \in H^1$. Assume that $|u(t)|_{L^2} = |Q|_{L^2}$ and that $u(t)$ blows up in finite time. Then, up to invariance of equation (2),

$$u(t) = S(t). \tag{8}$$

That is, there are $x_0 \in R^N$, $x_1 \in R^N$,$T \in R$, $\theta \in R$, and $\omega \in R^+$ such that

$$u(t) = e^{i(-\frac{\omega^2}{t-T}+\frac{|x-x_0|^2}{4(t-T)})}\left(\frac{\omega}{t-T}\right)^{\frac{N}{2}} Q\left(\frac{(x-x_0)\omega}{t-T} - x_1\right). \tag{9}$$

Let us give some idea of the proof. Various arguments in the proof will apply in other contexts, giving qualitative information about blow-up solutions. Consider a blow-up solution of minimal mass $u(t)$, and denote by T its blow-up time.

- *Localization results on the singularity.* Using rough variational estimates, we show that there exist $\tilde{\rho}$, $\tilde{\theta}$, $\tilde{x}$ such that as t $\to$ T, $u(t) \sim e^{i\tilde{\theta}}\tilde{\rho}^{\frac{N}{2}} Q((x-\tilde{x})\tilde{\rho})$ in H^1. Then from refined geometrical estimates around Q, there are $\rho(t)$, $\theta(t)$, $x(t)$ such that $u(t) - e^{i\theta(t)}\rho(t)^{\frac{N}{2}} Q((x-x(t))\rho(t))$ is bounded in H^1. In particular, $|u(t,x+x(t))|^2 \rightharpoonup |Q|^2_{L^2}\delta_{x=0}$ as t $\to$ T. In the radial case, a different approach can be used to show that for all radial blow-up solutions the behavior outside the origin is mild.

- *Local virial identities.* Using time variation of $\int \psi(x)|u(t,x)|^2 dx$, where ψ is a localized function, we then show that $u(t)$ and u_0 decay at infinity. That is,

$u(t) \in \Sigma$ and $|x||u(t,x)|$ can be controlled in L^2 at infinity, uniformly in time. Moreover, it is shown that the singularity point $x(t)$ has a limit as t$\to$ T (for example the origin).

- *Conclusion using the minimality condition.* Let us consider the polynomial in time of degree two $p(t) = \int |x|^2 |u(t,x)|^2 dx$. From the previous steps, $p(T) = 0$. Using the minimality condition, we show that $p'(T) = 0$. By explicit calculation, we check that the energy of a transformation of the initial data u_0 is zero with an L^2 norm equal to $|Q|_{L^2}$, which is the variational characterization of Q up to invariance of the elliptic equation. This concludes the proof.

b) *Application to asymptotic behavior for globally defined solutions* [26].

The conformal invariance and the nonblow-up result of Theorem 1 yield a decay result in time for solutions defined for all time. Indeed, the nonlinear term can be seen as a perturbation localized in time for the linear Cauchy problem, for initial data such that $u_0 \in \Sigma$ and $|u_0|_{L^2} \leq |Q|_{L^2}$, except for the two solutions $P(t)$ and $S(t)$ (and the ones related via the invariances). More precisely, as t $\to +\infty$, the nonlinear solution behaves as a solution of the linear Schrödinger equation (scattering theory can be carried out: $u(t,x) \sim U(t)u_\infty$ as $t \to +\infty$, where U(t) is the free semigroup).

Note that the set of initial data such that this behavior occurs is open, which implies the following: for all u_0 different from $P(t)$ and $S(t)$ such that $|u_0|_{L^2} = |Q|_{L^2}$, there is a ball in L^2 such that if the initial data is inside the ball, the solution does not blow up. It is optimal since the virial identity yields that for all $\epsilon > 0$, if $u_0 = (1+\epsilon)S(-1)$ or $(1+\epsilon)P(-1)$ then the solution blows up in finite time.

II.2 *Qualitative properties of blow-up solutions*

a) *Concentration results in L^2.*

In this subsection, we show that the blow-up phenomena may be observed in L^2 and do not depend on the space where the Cauchy theory is applied. Let us assume first that $u_0 \in H^1$, then from [22], [12], we have

- *Concentration in L^2*: there are $x(t)$ and $\rho(t) \to +\infty$ such that

$$\liminf |u(t)|_{L^2(|x-x(t)| \leq \rho(t)^{-1})} \geq |Q|_{L^2}. \tag{10}$$

- *asymptotic compactness in L^2*: for any sequence $t_n \to T$ there is a subsequence t_n and an $H \in H^1$ with $|H|_{L^2} \geq |Q|_{L^2}$ such that in $H^1 - weak$

$$\rho_n^{\frac{N}{2}} u(t_n, (x - x_n)\rho_n) \rightharpoonup H. \tag{11}$$

We do not know if H or $|H|_{L^2}$ depends on the sequence (except for some partial results in the radial case).

Let us now assume that $u_0 \in L^2$, and $N = 1, 2$; energy arguments no longer apply in this case. Nevertheless, in [6], [27], refinement of Strichartz' Inequality (implying that the Cauchy problem can be solved in $X \supset L^2$), harmonic analysis techniques, and the use of the conformal invariance allow us to obtain concentration in L^2 and asymptotic compactness properties in L^2 up to the invariance of the equation. That is, there is an $\alpha_0 > 0$ such that for a subsequence t_n and an

$H \in L^2$ with $|H|_{L^2} \geq \alpha_0$, there are parameters a_n, b_n, x_n, ρ_n, where $\rho_n \to +\infty$, such that in $L^2 - weak$

$$e^{ia_n x + ib_n |x|^2} \rho_n^{\frac{N}{2}} u(t_n, (x - x_n)\rho_n) \rightharpoonup H. \tag{12}$$

Note, from the invariance of the equation, the solution with initial data $e^{iax+ib|x|^2} c^{\frac{N}{2}} H((x-d)c)$ can be written in terms of the solution with initial data H .
It is an open problem to prove $\alpha_0 = |Q|_{L^2}$.

b) *Construction of blow-up solutions from $S(t)$.*

Here we describe constructions of solutions which behave like $S(t)$ at the blow-up point. Another problem will be to construct if possible other types of blow-up solutions (with for example a different blow-up rate, see [18],[19]). Let $x_1, \ldots, x_n$ be given points of R^N. In [21], a blow-up solution is constructed such that the blow-up set is exactly the points $x_1, \ldots, x_n$ and as t→T, $u(t) \sim \Sigma \omega_i^{\frac{N}{2}} S(t, (x - x_i)\omega_i)$ in L^2, where the ω_i are sufficiently large.

In the case $N = 2$, for u^* very regular such that $\partial^\alpha u^*(0) = 0$ for $|\alpha| \leq \alpha_0$, in [7] the existence of a solution u(t) is proved such that $u(t) \sim S(t, x) + u^*(x)$ in L^2 at the blow-up. An open problem is to reduce α_0 to 1 or 2.

c) *Giving a sense to the equation after blow-up.* [26]

We are interested in giving a sense to the equation after the blow-up time. We consider the case of a minimal blow-up solution, that is, after renormalization $u(t) = S(t, x)$ for $t < 0$.
Let $\epsilon > 0$, and set

$$u_\epsilon(t, x) = (1 - \epsilon)S(-1, x) + O(\epsilon^2) \text{ in } \Sigma.$$

We have that $|u_\epsilon|_{L^2} < |Q|_{L^2}$, thus $u_\epsilon(t)$ is defined for all time. The question is what happens in the limit as $\epsilon \to 0$ after the blow-up time (for $t > 0$). Using the characterization of the minimal blow-up solution and a family of auxiliary variational problems in Σ, we have the following result:

THEOREM 2 ([26]) *There is a $\theta(\epsilon) \in R$ continuous in ϵ such that*

$$\begin{array}{rcll} u_\epsilon(t) & \to & S(t) \text{ in } H^1 & \text{for} \quad t < 0 \\ |u_\epsilon(0)|^2 & \rightharpoonup & |Q|_{L^2}^2 \delta_{x=0} & \\ e^{-i\theta(\epsilon)} u_\epsilon(t) & \to & S(t) \text{ in } H^1 & \text{for} \quad t > 0. \end{array}$$

We then prove that as $\epsilon \to 0$, the omega-limit set of $e^{i\theta(\epsilon)}$ is S^1. From this result, the omega limit set of u_ϵ is $\{u_\theta \mid \theta \in S^1\}$, where

$$u_\theta(t) = S(t) \text{ for } t < 0 \text{ and } u_\theta(t) = e^{i\theta} S(t) \text{ for } t > 0.$$

In particular, we first show that the singularity is unstable in time.
In addition, from the blow-up, the physical phenomenon loses its deterministic character (but just up to one parameter in S^1). In addition, this result seems in some sense independent of the approximation. Therefore, the physics (which is not understood close to the singularity) has in some sense no influence on the

behavior of the solution after the blow-up, at least in the case of the minimal blow-up solution.

III. *Results for the Zakharov System.*

III.1 *No blow-up under smallness conditions*
As in the case of the critical Schrödinger equation, for initial data (u_0, n_0, v_0) in H_1, if $|u_0|_{L^2} < |Q|_{L^2}$, then there is no blow-up. Moerover for any blow-up solution, as t goes to the blow-up time, $u(t)$ concentrates in L^2 to a magnitude of at least $|Q|_{L^2}$ (see (11)).

At the critical mass level, $|u_0|_{L^2} = |Q|_{L^2}$, there is still a periodic solution (and the family it generates) given by

$$\tilde{P}(t) = (u(t), n(t), v(t)) = (P(t), -Q^2, 0). \tag{13}$$

Using the coupling between the equations, one can prove ([12]) that there are no blow-up solutions of (3) such that

$$|u_0|_{L^2} = |Q|_{L^2}. \tag{14}$$

III.2 *Existence of a family of explicit blow-up solutions* ([12]).
In fact the family of blow-up solutions of type S, for $\omega > 0$

$$S_\omega(t, x) = \omega^{\frac{N}{2}} S(t\omega^2, x\omega) \tag{15}$$

does not disappear. From bifurcation type arguments at $\omega = +\infty$ and index theory, we construct an explicit family of blow-up solutions of equation (3) of structure similar to that of $S(t)$ (where n and $|u|^2$ are of the same order): for all $\omega > 0$,

$$(u_\omega(t, x), n_\omega(t, x)) = \left((\frac{\omega}{t}) e^{-\frac{i\omega^2}{t} + i\frac{|x|^2}{4t}} P_\omega(\frac{x\omega}{t}), \frac{\omega^2}{t^2} N_\omega(\frac{x\omega}{t}) \right), \tag{16}$$

where (P_ω, N_ω) are radial solutions of the following equation, where $r = |x|$

$$\begin{cases} P + NP = \Delta P, \\ \frac{1}{{c_0}^2\omega^2}(r^2 \frac{\partial^2 N}{\partial r^2} + 6r\frac{\partial N}{\partial r} + 6N) - \Delta N = \Delta P^2. \end{cases} \tag{17}$$

Note that when $\omega = +\infty$ (17) reduces to (5). It is then proved that $\{|P_\omega|_{L^2}\} = (|Q|_{L^2}, +\infty)$, which has several consequences:

- *There are no minimal blow-up solutions in L^2 for the Zakharov equation.* Indeed, for all $\epsilon > 0$, there is a blow-up solution such that $|u_0|_{L^2} = |Q|_{L^2} + \epsilon$ and there are no blow-up solutions such that $|u_0|_{L^2} \leq |Q|_{L^2}$. The situation is different from the Schrödinger equation.

- *Any $c > |Q|_{L^2}$ can be a concentration mass*: there is a blow-up solution such that at the blow-up, $|u(t, x)|^2 \rightharpoonup c\delta_{x=0}$.

- Using these explicit solutions as ω becomes large, we are able to prove that the periodic solution $\tilde{P}(t)$ *is unstable* in the following sense: in all neighborhoods of it, there is a blow-up solution.

III.3 *Existence of a large class of blow-up solutions* [24].

As for the critical Schrödinger equation, a natural question to ask is, For Hamiltonians $H_0 < 0$, does the solution blow-up? (which will produce a large class of blow-up solutions). No Pohozaev identity was known until recently. In [24], the following identity was derived

$$\frac{d^2}{dt^2}M(t) = 2H(u_0, n_0, v_0) + \frac{1}{{c_0}^2}\int |v|^2 dx, \tag{18}$$

where

$$M(t) = \frac{1}{4}\int |x|^2|u|^2 dx + \frac{1}{{c_0}^2}\int_{R^2\times[0,t]} n(x.v)dxdt. \tag{19}$$

Note that if $c_0 = +\infty$ then relation (19) reduces to (4). The nature of the obstruction to global existence is slightly different from that in equation (2). Indeed, in [23], it is shown that for any blow-up solution, $M(t) \to -\infty$ as t goes to the blow-up time. Nevertheless, by localization techniques, it is proved in the radial case that if $H_0 < 0$, then the solution blows up in finite time or infinite time (and is concentrated in L^2 at the blow-up).

As a corollary, all periodic solutions of type $(u, n) = (e^{it}W(x), -W^2)$ where W is a solution of (5) are unstable since $H(e^{it}W(x), -W^2, 0) = 0$.

III.4 *Toward the structural stability of* S [23].

Let us measure the blow-up rate by the H^1 norm $|\nabla u(t)|_{L^2}$. An important problem is to understand the type of rates at the blow-up time and their stability. For equation (2), the blow-up of S (that is of the minimal blow-up solution) is $\frac{1}{|t|}$. We expect that minimality is related to stability. It seems not to be the case; in [18], [19] a blow-up rate of the type $\frac{|Log|log|t|||^{\frac{1}{2}}}{|t|^{\frac{1}{2}}}$ is observed numerically.

Nevertheless, we show the following result for the Zakharov equation (relating minimality to structural stability). Consider any blow-up solution of (3) (with any finite c_0), then

$$|\nabla u(t)|_{L^2} \geq \frac{c}{|t|}. \tag{20}$$

We note that this lower bound is optimal since the solution (u_ω, n_ω) blows up with this rate. Therefore, if we consider the refined equation from the physical point of view, the solution with blow-up rate $\frac{|Log|log|t|||^{\frac{1}{2}}}{|t|^{\frac{1}{2}}}$ disappears (even if c_0 is very large).

In [29], the same blow-up rate that was observed for S is seen, and seems numerically stable. It is an open problem to prove that all blow-up solutions of the Zakharov equation blow up with the same rate as S (the upper bound remains to be proved).

REFERENCES

[1] Berestycki, H.; Lions, P.-L. Nonlinear scalar field equations.I, II. Arch. Rational Mech. Anal. 82 (1983), no. 4, 313–345 and 347–375.

[2] Birnir, B.; Kenig, C.; Ponce, G.; Svanstedt, N.; Vega, L.; On the ill-posedness of the IVP for the generalized Korteweg-de Vries and nonlinearSchrödinger equations. J. London Math. Soc. (2) 53 (1996), no. 3, 551–559.

[3] Bourgain, J.; Colliander, J. On wellposedness of the Zakharov system. Internat. Math. Res. Notices 1996, no. 11, 515–546.

[4] Bourgain, J.; Harmonic analysis and nonlinear partial differential equations. Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zurich, 1994), 31–44, Birkhäuser, Basel, 1995.

[5] Bourgain, J.; Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. I.Schrodinger equations. Geom. Funct. Anal. 3 (1993), no. 2, 107–156.

[6] Bourgain, J.; Refinement of Strichartz' Inequality and applications to 2D-NLS with critical Nonlinearity, I.R.R.N., n.5, (1998) 253-283.

[7] Bourgain, J.; Wang, Preprint.

[8] Cazenave, T.; Weissler, F.; Some remarks on the nonlinear Schrödinger equation in the critical case. Nonlinear semigroups, partial differential equations and attractors (Washington, DC, 1987), 18–29, Lecture Notes in Math., Springer, Berlin-New York, 1989.

[9] Ginibre, J.; Tsutsumi, Y.; Velo, G.; On the Cauchy problem for the Zakharov system. J. Funct. Anal. 151 (1997), no. 2, 384–436.

[10] Ginibre, J.; Velo, G.; The global Cauchy problem for the nonlinear Schrödinger equation revisited. Ann. Inst. H. Poincare Anal. Non Lineaire 2 (1985), no. 4, 309–327.

[11] Ginibre, J.; Velo, G.; On a class of nonlinear Schrödinger equations. I. The Cauchy problem, general case. J. Funct. Anal. 32 (1979), no. 1, 1–32.

[12] Glangetas, L.; Merle, F. Existence of self-similar blow-up solutions for Zakharov equation in dimension two. I., Concentration properties of blow-up solutions and instability results for Zakharov equation in dimension two. II. Comm. Math. Phys. 160 (1994), no.1 and 2, 173–215 and 349–389.

[13] Glassey, R. T. On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations. J. Math. Phys. 18 (1977), no. 9, 1794–1797.

[14] Kato, T.;On nonlinear Schrödinger equations. Ann. Inst. H. Poincare Phys. Theor. 46 (1987), no. 1, 113–129.

[15] Kenig, C.; Ponce, G.; Vega, L.; On the Zakharov and Zakharov-Schulman systems. J. Funct. Anal. 127 (1995), no. 1, 204–234.

[16] Kenig, C.; Ponce, G.; Vega, L.; Well-posedness and scattering results for the generalized Korteweg-de Vries equation via the contraction principle. Comm. Pure Appl. Math. 46 (1993), no. 4, 527–620.

[17] Kwong, M. K.; Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbf{R}^n$. Arch. Rational Mech. Anal. 105 (1989), no. 3, 243–266.

[18] Landman, M. J.; Papanicolaou, G. C.; Sulem, C.; Sulem, P.-L.; Rate of blowup for solutions of the nonlinear Schrödinger equation at critical dimension. Phys. Rev. A (3) 38 (1988), no. 8, 3837–3843.

[19] LeMesurier, B. J.; Papanicolaou, G. C.; Sulem, C.; Sulem, P.-L.; Local structure of the self-focusing singularity of the nonlinear Schrödinger equation. Phys. D 32 (1988), no. 2, 210–226.

[20] Mac laughlin D., Lecture in Park-city, I.A.S. summer school 1995.

[21] Merle, F.; Construction of solutions with exactly k blow-up points for the Schrödinger equation with critical nonlinearity. Comm. Math. Phys. 129 (1990), no. 2, 223–240.

[22] Merle, F.; Tsutsumi, Yoshio L^2 concentration of blow-up solutions for the nonlinear Schrödinger equation with critical power nonlinearity. J. Differential Equations 84 (1990), no. 2, 205–214.

[23] Merle, F.; Lower bounds for the blowup rate of solutions of the Zakharov equation in dimension two. Comm. Pure Appl. Math. 49 (1996), no. 8, 765–794.

[24] Merle, F.; Blow-up results of virial type for Zakharov equations. Comm. Math. Phys. 175 (1996), no. 2, 433–455.

[25] Merle, F. Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equations with critical power. Duke Math. J. 69 (1993), no. 2, 427–454.

[26] Merle, F.; On uniqueness and continuation properties after blow-up time of self-similar solutions of nonlinear Schrödinger equation with critical exponent and critical mass. Comm. Pure Appl. Math. 45 (1992), no. 2, 203–254.

[27] Merle, F.; Vega,L.; Compactness at blow-up time for L^2 solutions of the critical Nonlinear Schrödinger equation in 2D, I.M.R.N. (1998), no. 8, 399–425.

[28] Ogawa, T.; Tsutsumi, Y.; Blow-up of H^1 solution for the nonlinear Schrödinger equation. J. Differential Equations 92 (1991), no. 2, 317–330.

[29] Papanicolaou, G. C.; Sulem, C.; Sulem, P.-L.; Wang, X. P.; Singular solutions of the Zakharov equations for Langmuir turbulence. Phys. Fluids B 3 (1991), no. 4, 969–980.

[30] Strauss, W.; Existence of solitary waves in higher dimensions. Comm. Math. Phys. 55 (1977), no. 2, 149–162.

[31] Weinstein, M. I.; Nonlinear Schrödinger equations and sharp interpolation estimates. Comm. Math. Phys. 87 (1982/83), no. 4, 567–576.

[32] Weinstein, M. I.; On the structure and formation of singularities in solutions to nonlinear dispersive evolution equations. Comm. Partial Differential Equations 11 (1986), no. 5, 545–565.

[33] Zakharov, V.E.; Collapse of langmuir waves, J.E.T.P 35 (1972), 908-914.

Frank Merle
Mathematiques
Université de Cergy-Pontoise
2, Avenue A. Chauvin, Pontoise
95302 Cergy-Pontoise
France

On Nonlinear Dispersive Equations

Gustavo Ponce

Introduction: I shall describe some of the recent developments in the application of harmonic analysis to non-linear dispersive equations. In recent years this subject has generated an intense activity and many new results have been proved. My contribution to this field has been made in collaboration with Carlos E. Kenig and Luis Vega. Their scientific inspiration, which has been so rewarding for me, is surpassed only by the warmth of their friendship.

We shall be concerned with the initial value problem (IVP) for nonlinear dispersive equations of the form

$$\begin{cases} \partial_t u = iP(\nabla_x)u + F(u), \quad t \in \mathbb{R}, \quad x \in \mathbb{R}^n, \\ u(x,0) = u_0(x), \end{cases} \tag{1}$$

where $P(D)$ is the constant coefficient operator defined by its real symbol $P(i\xi)$ and $F(\cdot)$ represents the nonlinearity.

We shall concentrate our attention in the following two problems:

Problem A: The problem of the minimal regularity of the data u_0 which guarantees that the IVP (1) is well-posed.

Problem B: The existence and uniqueness for the IVP (1) for some dispersive models for which classical approaches do not apply.

Let us first consider Problem A. Our notion of well-posedness includes existence, uniqueness, persistence, i.e. if $u_0 \in \mathcal{X}$ function space then the corresponding solution describes a continuous curve in $\mathcal{X}$, and lastly continuous dependence of the solution upon the data. Thus, solutions of (1) induce a dynamical system on $\mathcal{X}$ by generating a continuous flow, see [Kt].

We use classical the Sobolev spaces $\mathcal{X} = H^s(\mathbb{R}^n) = (1-\Delta)^{-s/2}L^2(\mathbb{R}^n),\ s \in \mathbb{R}$ to measure the regularity of the data.

To illustrate our arguments we consider the IVP for the generalized Korteweg-de Vries (gKdV) equation

$$\begin{cases} \partial_t u + \partial_x^3 u + u^k \partial_x u = 0, \qquad t, x \in \mathbb{R}, \ \ k \in \mathbb{Z}^+, \\ u(x,0) = u_0(x). \end{cases} \tag{1.1}$$

For $k = 1$ (KdV) the equation in (1.1) was derived by Korteweg-de Vries as a model for long waves propagating in a channel. Later, the cases $k = 1, 2$ were found to be relevant in several physical situations. Also they have been studied because of their relation to inverse scattering theory and to algebraic geometry (see [Mi] and references therein).

Local well-posedness results imply global ones via the conservation laws

$$I_2(u) = \int_{-\infty}^{\infty} u^2(x,t)dx, \qquad I_3(u) = \int_{-\infty}^{\infty} ((\partial_x u)^2 - c_k u^{k+2})(x,t)dx, \tag{1.2}$$

satisfied by solutions of (1.1), (for $k = 1, 2$ there are infinitely many I_j's, see [Mi]).

Concerning the local well-posedness of the IVP (1.1) our first result is the following.

THEOREM 1.1 ([KePoVe3]).
The IVP (1.1) is locally well-posed in $H^s(\mathbb{R})$ if

$$\begin{cases} k = 1 \quad \textit{and} \quad s > 3/4, \\ k = 2 \quad \textit{and} \quad s \geq 1/4, \\ k = 3 \quad \textit{and} \quad s > 1/12, \\ k \geq 4 \quad \textit{and} \quad s \geq (k-2)/4k.- \end{cases} \tag{1.3}$$

Observe that if $u(\cdot)$ solves the equation in (1.1) then $u_\lambda(x,t) = \lambda^{2/k} u(\lambda x, \lambda^3 t)$ is also a solution with data $u_\lambda(x,0) = \lambda^{2/k} u_0(\lambda x)$ and

$$\|D_x^s u_\lambda\|_2 = c\lambda^{s-(k-4)/2k}. \tag{1.4}$$

Thus, for $s = (k-4)/2k$ the above norm is independent of λ. The result in Theorem 1.1 for $k \geq 4$ correspond to the scaling value in (1.4) and has been shown to be optimal, see [KePoVe3] and [BKPSV]. Theorem 1.1 and the conservation laws in (1.2) imply the global well-posedness of (1.1) with $u_0 \in H^s(\mathbb{R})$, $s \geq 1$ and $k = 1, 2, 3$. For $k \geq 4$ the existence of global solution for data $u_0 \in H^1(\mathbb{R})$ of arbitrary size is unknown.

To explain our result with more details we choose the case $k = 2$, i.e. the modified Korteweg-de Vries (mKdV) equation.

THEOREM 1.2 ([KePoVe3]).
Let $k = 2$. Then for any $u_0 \in H^{1/4}(\mathbb{R})$ there exist

$$T = c\,\|D_x^{1/4} u_0\|_2^{-4}, \tag{1.5}$$

and a unique strong solution $u(t)$ of the IVP (1.1) satisfying

$$u \in C([-T,T] : H^{1/4}(\mathbb{R})), \tag{1.6}$$

and

$$\|D_x^{1/4}\,\partial_x u\|_{L_x^\infty L_T^2} \;+\; \|u\|_{L_x^4 L_T^\infty} < \infty. \tag{1.7}$$

Moreover, the map data $\to$ solution, from $H^{1/4}(\mathbb{R})$ into the class defined by (1.6)–(1.7) is locally Lipschitz.

In addition, if $u_0 \in H^{s'}(\mathbb{R})$ with $s' > s$, then the above results hold with s' instead of s in the same time interval $[-T, T]$.-

The properties (1.6)–(1.7) guarantee the uniqueness of the solution and that the nonlinear term is well defined, i.e. it is at least a distribution.

In [Ka], T. Kato established the existence of a global weak solution for the IVP (1.1) with $k = 1, 2, 3$ and data $u_0 \in L^2(\mathbb{R}^2)$. In [GiTs], Ginibre-Tsutsumi showed that if $(1 + |x|)^{3/8} u_0 \in L^2(\mathbb{R})$ then IVP (1.1) with $k = 2$ has a unique solution. Since the operator $\Gamma = x - 3t\partial_x^2$ commutes with the linear part of the equation in (1.1) one sees that Theorem 1.2 and the result in [GiTs] complement each other. Also the estimate of the life span of the local solution in (1.5) agrees with that given by the scaling argument in (1.4).

The proof of Theorem 1.2 is based on the following two sharp linear estimates, in which we introduced the notation

$$U(t)v_0(x) = \int_{-\infty}^{\infty} e^{i(t\xi^3 + x\xi)} \widehat{v}_0(\xi)\, d\xi. \tag{1.8}$$

In [KeRu], Kenig-Ruiz proved that

$$\left(\int_{-\infty}^{\infty} \sup_{[-1,1]} |U(t)v_0|^4 dx \right)^{1/4} \leq c \|D_x^{1/4} v_0\|_2, \tag{1.9}$$

and that both indexes in (1.8), i.e. 4, 1/4 are optimal. In [KePoVe1], we showed that there exists $c > 0$ such that for any $x \in \mathbb{R}$

$$\left(\int_{-\infty}^{\infty} |\partial_x U(t)v_0|^2 dt \right)^{1/2} = c \|v_0\|_2. \tag{1.10}$$

This is a sharp version of the local smoothing effects first established by T. Kato [Ka] for solutions of the KdV equation, see also [KuFr].

In [Bo1], J. Bourgain showed that the IVP for the KdV ($k = 1$ in (1.1)) is locally (consequently globally) well-posed in L^2. His proof relies on the use of the spaces $X_{s,b}$, i.e. the completion of $S(\mathbb{R}^2)$ respect to the norm

$$\|F\|_{X_{s,b}} = \|(1 + |\tau - \xi^3|)^b (1 + |\xi|)^s \widehat{F}(\xi, \tau)\|_{L^2(\mathbb{R}^2)}. \tag{1.11}$$

These spaces were introduced by M. Beals [Be] in his study of propagation of singularities for solutions to semi-linear wave equations, and have been successfully used in several related works. In [KlMa] and subsequent works, Klainerman-Machedon used them to study the minimal regularity problem on the data for systems of nonlinear wave equations with nonlinearities satisfying a special structure.

In [KePoVe4], we proved that the IVP for the KdV equation ($k = 1$ in (1.1)) is locally well-posed in $H^s(\mathbb{R})$, $s > -3/4$.

THEOREM 1.3 ([KePoVe4]).

Let $s \in (-3/4, 0]$. Then there exists $b \in (1/2, 1)$ such that for any $u_0 \in H^s(\mathbb{R})$ there exist $T = T(\|u_0\|_{H^s}) > 0$ (with $T(\rho) \to \infty$ when $\rho \to 0$) and a unique solution $u(t)$ of the IVP (1.1) in the time interval $[-T, T]$ satisfying

$$u \in C([-T, T] : H^s(\mathbb{R})), \tag{1.12}$$

$$u \in X_{s,b} \subseteq L^\infty_{x,\ell\mathrm{oc}}(\mathbb{R} : L^2_t(\mathbb{R})), \tag{1.13}$$

and

$$\partial_x(u^2) \in X_{s,b-1}, \qquad \partial_t u \in X_{s-3,b-1}. \tag{1.14}$$

Moreover, the map data $\to$ solution from $H^s(\mathbb{R})$ into the class defined by (1.12)-(1.14) is locally Lipschitz.

In addition, if $u_0 \in H^{s'}(\mathbb{R})$ with $s' > s$, then the above results hold with s' instead of s in the same time interval $[-T, T]$.-

The method of proof of Theorem 1.3 is based on bilinear estimates involving the spaces $X_{s,b}$ and elementary techniques. These techniques were motivated by the work of C. Fefferman [Fe] for the $L^4(\mathbb{R}^2)$ estimate for the Bochner-Riesz operator.

In [KePoVe4], we also established that for the case of the mKdV ($k = 2$ in (1.1)) the argument based on multilinear estimates and the use of $X_{s,b}$-spaces does not improve our result in Theorem 1.2.

The gap between the KdV result ($s > -3/4$) and that for the mKdV ($s \geq 1/4$) is somehow consistent with the Miura transformation, i.e. if v solves the mKdV equation then $u = c_1 v^2 + c_2 \partial_x v$ solves the KdV equation.

The method of proof in [KePoVe3], [Bo1],[KePoVe4], is based on the contraction principle which combined with the Implicit Function Theorem shows that the map data $\to$ solution is smooth.

In [Bo2], J. Bourgain proved that if one requires the map data $\to$ solution be smooth (C^3 suffices) then our results for the KdV ($s > -3/4$) in [KePove4] and for the mKdV ($s \geq 1/4$) in [KePoVe3] are optimal. In particular, it follows that these results cannot be improved by using only an iteration argument.

Regarding the global well-posedness of the IVP for the KdV and mKdV equations we have the following recent results. In [FoLiPo], Fonseca-Linares-Ponce showed that the IVP for the mKdV equation is globally well-posed (although not

necessarily globally bounded) in $H^s(\mathbb{R})$, $s \in (3/5, 1)$. In [CoSt], Colliander-Staffilani proved the IVP for the KdV equation is globally well-posed (although not necessarily globally bounded) in $H^s(\mathbb{R})$, $s \in (-3/20, 0)$. The proofs combine ideas in [Bo3] and Theorems 1.2-1.3 described above.

Problem B.

We begin by considering the IVP for nonlinear Schrödinger equations of the form

$$\begin{cases} \partial_t u = i\mathcal{L}u + P(u, \nabla_x u, \overline{u}, \nabla_x \overline{u}), & t \in \mathbb{R},\ x \in \mathbb{R}^n, \\ u(x,0) = u_0(x), \end{cases} \tag{2.1}$$

where $\mathcal{L}$ is a non-degenerate constant coefficient, second order operator

$$\mathcal{L} = \sum_{j \le k} \partial_{x_j}^2 - \sum_{j > k} \partial_{x_j}^2, \qquad \text{for some} \quad k \in \{1, .., n\}, \tag{2.2}$$

and $P : \mathbb{C}^{2n+2} \to \mathbb{C}$, is a polynomial of the form

$$P(z) = P(z_1, .., z_{2n+2}) = \sum_{l_0 \le |\alpha| \le d} a_\alpha z^\alpha, \qquad l_0 \ge 2. \tag{2.3}$$

When a special form of the nonlinear term P is assumed, for example,

$$D_{\partial_{x_j} u} P, \qquad \text{are real for } \ j = 1, .., n, \tag{2.4}$$

standard energy estimates provide the desired result. In this case, the dispersive part of the equation, the operator $\mathcal{L}$, does not play any role. Another technique used to overcome the "loss of derivatives" introduced by the nonlinear term is to present the problem in a suitable analytic function spaces, see [SiTa], [Hy].

In [KePoVe2], we proved that (2.1) is locally well-posed for "small" data, in $H^s(\mathbb{R}^n)$, for s large enough, when $l_0 \ge 3$ in (2.3), and in a weighted version of it, if $l_0 = 2$ in (2.3). This result applies to the general form of $\mathcal{L}$ in (2.2). The main idea is to use in the integral equation version of the IVP (2.1)

$$u(t) = e^{it\mathcal{L}} u_0 + \int_0^t e^{i(t-t')\mathcal{L}} P(u, \nabla_x u, \overline{u}, \nabla_x \overline{u})(t') dt', \tag{2.5}$$

and the following estimates,

$$\begin{cases} (i) \quad |||D^{1/2} e^{it\mathcal{L}} u_0|||_T \equiv \sup\limits_{\mu \in \mathbb{Z}^n} \Big(\displaystyle\int_0^T \int_{Q_\mu} |D^{1/2} e^{it\mathcal{L}} u_0|^2 dx dt \Big)^{1/2} \le c \|u_0\|_2, \\ (ii)\ |||\nabla_x \displaystyle\int_0^t e^{i(t-t')\mathcal{L}} F(t') dt'|||_T \le c |||F|||'_T, \end{cases} \tag{2.6}$$

where $\{Q_\mu\}_{\mu \in \mathbb{Z}^n}$ is a family cubes of side one with disjoint interiors covering $\mathbb{R}^n$, and $D = (-\Delta)^{1/2}$.

The local smoothing effect in (i) was proven by Constantin-Saut [CnSa], Sjölin [Sj], and Vega [Ve]. We proved the inhomogeneous version (ii) in [KePoVe2].

It is essential the gain of one derivative in (2.6) (ii). This allows to use the contraction principle in (2.5) and avoid the "loss of derivatives". However, the $|||\cdot|||$ norm forces the use of its dual

$$|||G|||'_T \equiv \|G\|_{l^1_\mu(L^2(Q_\mu\times[0,T]))} = \sum_{\mu\in\mathbb{Z}^n} (\int_0^T \int_{Q_\mu} |G(x,t)|^2 dxdt)^{1/2}. \tag{2.7}$$

This factor cannot be made small by taking T small, except if $G(t)$ is small at $t=0$. It is here where the restriction on the size of the data appears.

In [HyOz], for the one dimensional case $n=1$, Hayashi-Ozawa removed the smallness assumption on the size of the data in [KePoVe2]. They used a change of variable to obtain an equivalent system with a nonlinear term independent of $\partial_x u$, which can be treated by the standard energy method.

In [Ch], for the elliptic case $\mathcal{L}=\Delta$, H. Chihara removed the size restriction on the data in any dimension. The change of variable in this case involves pseudo-differential operators ψ.d.o's. A main step in his proof is a diagonalization method in which the assumption on the ellipticity of $\mathcal{L}$ is essential.

In [KePoVe5], we removed the size restriction for the general form of the operator $\mathcal{L}$ in (2.1).

THEOREM 2.1 ([KePoVe5]). *There exist $s=s(n;P)>0$, and $m=m(n;P)>0$, such that for any $u_0\in H^s(\mathbb{R}^n)\cap L^2(\mathbb{R}^n:|x|^{2m}dx)$ the IVP (2.1) has a unique solution $u(\cdot)$ defined in the time interval $[0,T]$ satisfying that*

$$u\in C([0,T]:H^s(\mathbb{R}^n)\cap L^2(\mathbb{R}^n:|x|^{2m}dx)),\ \ \textit{and}\ \ |||J^{s+1/2}u|||_T<\infty. \tag{2.8}$$

If $s'>s$, then the above results hold, with s' instead of s, in the same time interval $[0,T]$.

Moreover, the map data $\to$ solution from $H^s(\mathbb{R}^n)\cap L^2(\mathbb{R}^n:|x|^{2m}dx)$ into the class in (2.8) is locally continuous.-

Our argument of proof uses the Calderón-Vaillancourt class [CaVa]. This was suggested by the work of J. Takeuchi [Tk]. In order to extend the argument in [KePoVe2] to prove Theorem 2.1, we need to show that, under appropriate assumptions on the smoothness and decay of the coefficients $b_{k,j}=(b_{k,1},.,b_{k,n})$, $k=1,2$, $j=1,.,n$, the IVP for the linear Schrödinger equation with variable coefficient lower order terms

$$\begin{cases} \partial_t v = i\mathcal{L}v + b_1(x)\cdot\nabla_x v + b_2(x)\cdot\nabla_x\overline{v} + F(x,t), \quad t\in\mathbb{R},\ x\in\mathbb{R}^n, \\ v(x,0)=v_0\in H^s(\mathbb{R}^n), \end{cases} \tag{2.9}$$

has a unique solution $v\in C([0,T]:H^s(\mathbb{R}^n))$ such that

$$\sup_{[0,T]}\|v(t)\|_{H^s} + |||J^{s+1/2}v|||_T \le c(b_1;b_2;T)(\|v_0\|_{H^s} + |||J^{s-1/2}F|||'_T). \tag{2.10}$$

Equations of the form described in (2.1) with $\mathcal{L}$ non-elliptic arise in several situations. For example, in the study of water wave problems, Davey-Stewartson [DS], and Zakharov-Shulman [ZaSc] systems, in ferromagnetism, Ishimori system [Ic], as higher dimension completely integrable model, see [AbHa].

Consider the Davey-Stewartson (DS) system

$$\begin{cases} i\partial_t u + c_0\partial_x^2 u + \partial_y^2 u = c_1|u|^2u + c_2 u\partial_x\varphi, \\ \partial_x^2\varphi + c_3\partial_y^2\varphi = \partial_x|u|^2, \\ u(x,y,0) = u_0(x,y), \end{cases} \tag{2.11}$$

where $u = u(x,y,t)$ is a complex-valued function, $\varphi = \varphi(x,y,t)$ is a real-valued function, (when $(c_0,c_1,c_2,c_3) = (-1,1,-2,1)$ or $(1,-1,2,-1)$ the system in (1.1) is known in inverse scattering as the DSI and DSII respectively).

In the case $c_3 < 0$, $c_0 < 0$, (i.e. the equation in (1.6) is essentially not semi-linear, and the dispersive operator is non elliptic) the only available existence results are for analytic data, Hayashi-Saut [HySa], or "small" data, Linares-Ponce [LiPo]. For other results for the DS system we refer to [GhSa], [HySa], [LiPo] and references therein.

The IVP for the Ishimori system can be written as

$$\begin{cases} i\partial_t u + \partial_x^2 u \mp \partial_y^2 u = \frac{2\overline{u}((\partial_x u)^2 - (\partial_y u)^2)}{1+|u|^2} + ib(\partial_x\varphi\partial_y u - \partial_y\varphi\partial_x u), \\ \partial_x^2\varphi \pm \partial_y^2\varphi = 4i\ \frac{\partial_x u\ \partial_y\overline{u} - \partial_x\overline{u}\ \partial_y u}{(1+|u|^2)^2}, \\ u(x,y,0) = u_0(x,y). \end{cases} \tag{2.12}$$

The $(-,+)$ case was studied by A. Souyer [So]. The case $(+,-)$ in (2.11) was first studied by Hayashi-Saut [HySa] in a class of analytic functions which allowed them to obtain local and global existence for small analytic data. In [Hy2], N. Hayashi removed the analyticity assumptions in [HySa] by establishing the local existence and uniqueness of solutions of the IVP (2.12), for the case $(+,-)$, with small data u_0 in the weighted Sobolev space $H^4(\mathbb{R}^2) \cap L^2((x^2+y^2)^4dxdy)$.

In a forthcoming article [KePoVe6] we remove the smallness assumption in [Hy2]. In particular, we prove the local existence and uniqueness of solutions of the IVP (2.12) with $(+,-)$ sign for data of arbitrary size in a weighted Sobolev space. Several problems have to be overcome to extend our approach in [KePoVe5] to this case. First, we have to deal with operators which are ψ.d.o. only in one variable. In particular, to establish the local smoothing effects described in (2.6) we shall need the operator valued version of the sharp Gårding inequality. Another difficulty of our approach is that for the linearized system associated to (2.12) the coefficients of the first order terms do not decay in both variables. One has terms of the form $a(x,y)\partial_x u$ where the coefficient $a(\cdot)$ is a smooth function with decay only in the x-variable. However, a careful analysis, consistent with Mizohata's condition in [Mz], shows that this one variable decay suffices.

REFERENCES

[AbHa] Ablowitz, M. J., and Haberman, R., *Nonlinear evolution equations in two and three dimensions*, Phys. Rev. Lett. **35** (1975), 1185–1188.

[Be] Beals, M., *Self-spreading and strength of singularities for solutions to semilinear wave equations*, Ann. Math. **118** (1983), 187–214.

[BKPSV] Birnir, B., Kenig, C. E., Ponce, G., Svanstedt, N., and Vega, L., *On the ill-posedness of the IVP for the generalized Korteweg-de Vries and nonlinear Schrödinger equations*, J. London Math. Soc. **53** (1996), 551–559.

[Bo1] Bourgain, J., *Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations*, Geometric and Functional Anal. **3** (1993), 107-156, 209-262.

[Bo2] Bourgain, J., *Periodic Korteweg-de Vries equation with measure as initial data*, Selecta Math. (N.S.) **3** (1997), 115–159.

[Bo3] Bourgain, J., *Refinements of Strichartz's inequality and applications to 2D-NLS with critical nonlinearity*, Int. Math. Research Not. **5** (1998), 253-283.

[CaVa] Calderón, A. P., and Vaillancourt, R., *A class of bounded pseudodifferential operators*, Proc. Nat. Acad. Sci. USA. **69** (1972), 1185-1187.

[Ch] Chihara, H., *The initial value problem semilinear Schrödinger equations*, Publ. RIMS. Kyoto Univ. **32** (1996), 445-471.

[CoSt] Colliander, J. and Steffilani, G., *private communication.*

[CnSa] Constantin, P. and Saut, J. C., *Local smoothing properties of dispersive equations*, J. Amer. Math. Soc. **1** (1989), 413–446.

[DS] Davey, A., and Stewartson, K., *On three dimensional packets of surface waves*, Proc. R. Soc. A **338** (1974), 101-110.

[Fe] Fefferman, C., *A note on spherical summation multipliers*, Israel J. Math. **15** (1973), 44–52.

[FoLiPo] Fonseca, G., Linares, F., and Ponce, G., *Global well-posedness for the modified Korteweg-de Vries equation*, pre-print.

[GiTs] Ginibre, J. and Tsutsumi, Y., *Uniqueness for the generalized Korteweg-de Vries equations*, SIAM J. Math. Anal. **20** (1989), 1388–1425.

[GhSa] Ghidaglia, J. M., and Saut, J. C., *On the initial value problem for the Davey-Stewartson System*, Nonlinearity **3** (1990), 475–506.

[Hy1] Hayashi, N., *Global existence of small analytic solutions to nonlinear*

Schrödinger equations, Duke Math. J **62** (1991), 575–592.

[Hy2] Hayashi, N., *Local existence in time of small solutions to the Ishimori system*, preprint.

[HyOz] Hayashi, N., and Ozawa, T., *Remarks on nonlinear Schrödinger equations in one space dimension*, Diff. Integral Eqs **2** (1994), 453–461.

[HySa] Hayashi, N., and Saut, J-C., *Global existence of small solutions to the Davey-Stewartson and the Ishimori systems*, Diff. Integral Eqs **8** (1995), 1657-1675.

[Is] Ishimori, Y., *Multi vortex solutions of a two dimensional nonlinear wave equation*, Progr. Theor. Phys **72** (1984), 33–37.

[Ka] Kato, T., *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, Advances in Math. Supp. Studies, Studies in Applied Math. **8** (1983), 93–128.

[KePoVe1] Kenig, C. E., Ponce, G. and Vega, L., *Oscillatory integrals and regularity of dispersive equations*, Indiana University Math. J. **40** (1991), 33–69.

[KePoVe2] Kenig, C. E., Ponce, G., and Vega, L., *Small solutions to nonlinear Schrödinger equations*, Annales de l'I.H.P. **10** (1993), 255–288.

[KePoVe3] Kenig, C. E., Ponce, G., and Vega, L., *Well-posedness and scattering results for the generalized Korteweg-de Vries equation via the contraction principle*, Comm. Pure Appl. Math. **46** (1993), 527–620.

[KePoVe4] Kenig, C. E., Ponce, G. and Vega, L., *A bilinear estimate with applications to the KdV equation*, J. Amer. Math. Soc. **9** (1996), 573–603.

[KePoVe5] Kenig, C. E., Ponce, G. and Vega, L., *Smoothing effects and local existence theory for the generalized nonlinear Schrödinger equations*, to appear in Inventiones Math.

[KePoVe6] Kenig, C. E., Ponce, G. and Vega, L., *On the initial value problem for the Ishimori system*, to appear.

[KeRu] Kenig, C. E. and Ruiz, A., *A strong type (2,2) estimate for the maximal function associated to the Schrödinger equation*, Trans. Amer. Math. Soc. **280** (1983), 239–246.

[KlMa] Klainerman, S., and Machedon, M., *Space-time estimates for null forms and the local existence theorem*, omm. Pure Appl. Math. **46** (1993), 1221–1268.

[KuFr] Kruzhkov, S. N., and Faminskii, A. V., *Generalized solutions of the Cauchy problem for the Korteweg-de Vries equation*, Math. U.S.S.R. Sbornik **48** (1984), 93–138.

[LiPo] Linares, F., and Ponce, G., *On the Davey-Stewartson systems*, Annales de l'I.H.P. Analyse non linéaire **10** (1993), 523–548.

[Mi] Miura, R. M., *The Korteweg-de Vries equation: A survey of results*, SIAM review **18** (1976), 412–459.

[Mz] Mizohata, S., *On the Cauchy problem*, Notes and Reports in Math. in Science and Engineering, Science Press& Academic Press **3** (1985).

[SiTa] Simon, J., and Taflin, E., *Wave operators and analytic solutions for systems of nonlinear Klein-Gordon equations and of non-linear Schrödinger equations*, Comm. Math. Phys. **99** (1985), 541–562.

[Sj] Sjölin, P., *Regularity of solutions to the Schrödinger equations*, Duke Math. J. **55** (1987), 699–715.

[So] Souyer, A., *The Cauchy problem for the Ishimori equations*, J. Funct. Anal. **105** (1992), 233–255.

[Tk] Takeuchi, J., *Le problème de Cauchy pour certaines équations aux derivées partielles du type de Schrödinger, VIII; symétrisations indépendendantes du temps*, C. R. Acad. Sci. Paris **t315, Série 1** (1992), 1055-1058.

[Ve] Vega, L., *The Schrödinger equation: pointwise convergence to the initial date*, Proc. Amer. Math. Soc. **102** (1988), 874–878.

[ZaSc] Zakharov, V. E., and Schulman, E. I., *Degenerated dispersion laws, motion invariant and kinetic equations*, Physica **1D** (1980), 185-250.

Gustavo Ponce
Department of Mathematics
University of California
Santa Barbara, CA 93106, USA

Inverse Boundary Value Problems for Partial Differential Equations

Dedicated to the memory of my father and to the memory of Warren Ambrose

Gunther Uhlmann

Abstract. We survey the role of complex geometrical optics solutions to partial differential equations in the solution of several inverse boundary value problems.

1991 Mathematics Subject Classification: 35R30, 35P05, 35J05, 78A70
Keywords and Phrases: Inverse boundary problems, complex geometrical optics, Dirichlet to Neumann map, inverse conductivity problem

0. Introduction

Inverse boundary problems are a class of problems in which one seeks to determine the internal properties of a medium by performing measurements along the boundary of the medium. These inverse problems arise in many important physical situations, ranging from geophysics to medical imaging to the non-destructive evaluation of materials.

The appropriate mathematical model of the physical situation is usually given by a partial differential equation (or a system of such equations) inside the medium. The boundary measurements are then encoded in a certain boundary map. The inverse boundary problem is to determine the coefficients of the partial differential equation inside the medium from knowledge of the boundary map.

In this paper we will survey part of the significant progress which has been made in the last twenty years in this area. Many of the advances have been a consequence of the construction of *complex geometrical optics* solutions for the class of partial differential equations under consideration. The prototypical example of an inverse boundary problem is the inverse conductivity problem, also called electrical impedance tomography, first proposed by A. P. Calderón [7]. In this case the boundary map is the voltage to current map; that is, the map assigns to a voltage potential on the boundary of a medium the corresponding induced current flux at the boundary of the medium. The inverse problem is to recover the electrical conductivity of the medium from the boundary map.

We will also discuss in this paper other examples of inverse boundary problems, including examples associated to the Schrödinger equation in the presence of a magnetic field, Maxwell's equations and the Lamé system of elasticity. The unifying theme of the paper is the role of complex geometrical optics solutions in inverse boundary value problems and our selection of problems reflects this choice. We list a series of basic open problems in the field. For an account of the close connection between inverse boundary value problems and inverse scattering problems at a fixed energy see [40]. Another important omission is the discussion of inverse boundary value problems for hyperbolic equations, in particular the Boundary Control Method. See the review paper [4] for more details.

1. The inverse conductivity problem for an isotropic conductivity

Let $\Omega \subseteq \mathbb{R}^n$ be a bounded domain with smooth boundary (many of the results are valid for Lipschitz boundaries). We denote by γ the conductivity of Ω, which we assume is in $L^\infty(\Omega)$ and strictly positive. The potential u in Ω with voltage f on $\partial\Omega$ satisfies

$$L_\gamma u = \operatorname{div}(\gamma \nabla u) = 0 \text{ in } \Omega; \quad u|_{\partial\Omega} = f. \tag{1.1}$$

The voltage to current map, or Dirichlet to Neumann map (DN), is defined by

$$\Lambda_\gamma(f) = \left(\gamma \frac{\partial u}{\partial \nu}\right)\Big|_{\partial\Omega}, \tag{1.2}$$

where u is the solution of (1.1), and ν denotes the unit outer normal to $\partial\Omega$.

The inverse problem is to determine γ knowing Λ_γ. More precisely we want to study properties of the map

$$\gamma \longrightarrow^{\Lambda} \Lambda_\gamma. \tag{1.3}$$

Note that $\Lambda_\gamma : H^{\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$ is bounded. We can divide this problem into several parts.

a) Injectivity of Λ (identifiability).
b) Continuity of Λ and its inverse if it exists (stability).
c) What is the range of Λ? (characterization problem).
d) Formula to recover γ from Λ_γ (reconstruction).
e) Give a numerical algorithm to find an approximation. of the conductivity given a finite number of voltage and current measurements at the boundary (numerical reconstruction).

In this section we outline the proof of the following identifiability result proven in [36].

1.1 Theorem. *Let $n \geq 3$. Let $\gamma_1, \gamma_2 \in C^2(\overline{\Omega})$ be strictly positive functions in $\bar{\Omega}$ such that $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$. Then $\gamma_1 = \gamma_2$ in $\bar{\Omega}$.*

Sketch of the proof. Using Green's thorem it is easy to prove that

$$Q_\gamma(f) := \int_\Omega \gamma |\nabla u|^2 dx = \int_{\partial\Omega} \Lambda_\gamma(f) f dS, \tag{1.4}$$

where u is the solution of (1.1). In other words $Q_\gamma(f)$ is the quadratic form associated to the selfadjoint linear map $\Lambda_\gamma(f)$, i.e., to know $\Lambda_\gamma(f)$ or $Q_\gamma(f)$ for all $f \in H^{\frac{1}{2}}(\partial\Omega)$ is equivalent. $Q_\gamma(f)$ measures the energy needed to maintain the potential f at the boundary.

Formula (1.4) suggests that instead of prescribing voltage measurements at the boundary to determine the conductivity in the interior, we find solutions of the equation (1.1). This is the point of view of Calderón [7] in his analysis of the linearized problem at a constant conductivity.

To find these solutions we first reduce the problem to studying the Schrödinger equation at zero energy. Let $\gamma \in C^2(\overline{\Omega})$ be a positive function. We have

$$\gamma^{-\frac{1}{2}} L_\gamma \gamma^{-\frac{1}{2}} u = (\Delta - q)u, \qquad q = \frac{\Delta\sqrt{\gamma}}{\sqrt{\gamma}}. \tag{1.5}$$

For any $q \in L^\infty(\Omega)$ we can define the set of Cauchy data

$$C_q = \{(f, g); f = u|_\Omega, g = \frac{\partial u}{\partial \nu}|_\Omega, u \in H^1(\Omega) \text{ solution of (1.1)}\}$$

If 0 is not a Dirichlet eigenvalue of $\Delta - q$ then C_q is the graph of a map which is, by definition, the DN map. Theorem 1.1 follows from Theorem 1.2 and the fact that Λ_γ determines both γ at the boundary and the normal derivative of γ at the boundary (see [15], [37]).

1.2 THEOREM. *Assume $q_i \in L^\infty(\Omega), i = 1, 2$ and $C_{q_1} = C_{q_2}$. Then $q_1 = q_2$.*

Sketch of the proof of Theorem 1.2. The key result is the construction of complex geometrical optics solutions to the Schrödinger equation. This was motivated by Calderón's analysis of the linearized problem at a constant conductivity [7].

1.1 LEMMA. *Let $q \in L^\infty(\mathbb{R}^n)$ with compact support. Let $\rho \in \mathbb{C}^n$ with $\rho \cdot \rho = 0$. Let $-1 < \delta < 0$. Then if $|\rho| \geq C(\delta) sup_{x\in\mathbb{R}^n}|(1 + |x|^2)q(x)|$ for some $C(\delta) > 0$, there exists a unique solution of $(\Delta - q)u = 0$ in $\mathbb{R}^n$ of the form*

$$u = e^{x\cdot\rho}(1 + \psi_q(x, \rho)) \tag{1.6}$$

with $\psi_q(\cdot, \rho) \in L^2_\delta(\mathbb{R}^n)$. Moreover $\|\psi_q(\cdot, \rho)\|_{L^2_\delta(\mathbb{R}^n)}$ goes to 0 as $|\rho|$ goes to infinity. A more precise estimate is proven in [36]. *(Here $L^2_\delta(\mathbb{R}^n)$ denotes the weighted L^2 space with norm $\|f\|^2_{L^2_\delta(\mathbb{R}^n)} = \int(1 + |x|^2)^\delta |f(x)|^2 dx$.)*

Let $q_i \in L^\infty(\Omega)$ as in the statement of Theorem (1.2). We define $q_i = 0$ in $\mathbb{R}^n - \Omega$. Let $\rho_i, i = 1, 2$ as in Lemma (1.1) with $\rho_1 = \eta + i(k+l)$, $\rho_2 = -\eta + i(k-l)$ with $\eta, k, l \in \mathbb{R}^n$ satisfying $\langle \eta, k\rangle = \langle k, l\rangle = \langle \eta, l\rangle = 0$, $|\eta|^2 = |k|^2 + |l|^2$ and $|l| \geq R_i$, with R_i sufficiently large so that Lemma 1.1 is valid for $q_i, i = 1, 2$ (here we use $n \geq 3$). We take

$$u_i = e^{x\cdot\rho_i}(1 + \psi_{q_i}(x, \rho_i)), \quad i = 1, 2. \tag{1.7}$$

The next important ingredient is the following identity which follows easily from Green's theorem.

1.2 LEMMA. *Let* $q_i \in L^\infty(\Omega)$, $i = 1, 2$ *and* $C_{q_1} = C_{q_2}$. *Then*

$$\int_\Omega (q_1 - q_2) u_1 u_2 = 0 \tag{1.8}$$

for every solution $u_i \in H^1(\Omega)$ *of* $(\Delta - q_i)u_i = 0$ *in* $\mathbb{R}^n$.

Now we plug (1.7) into (1.8). Taking the limit as $|l| \to \infty$, we easily conclude that the Fourier transform of q_1 and q_2 coincide.

In order to construct ψ_q as in (1.6) we solve the equation

$$\Delta_\rho \psi_q = q(1 + \psi_q) \text{ with } \Delta_\rho f = e^{-x\cdot\rho}\Delta(e^{x\cdot\rho} f). \tag{1.9}$$

We note that the characteristic variety of Δ_ρ is a codimension two real submanifold. We can construct an inverse Δ_ρ that satisfies the following estimate proven for $n \geq 3$ in [36] and for $n = 2$ in [35].

$$\|\Delta_\rho^{-1}\|_{\delta+1,\delta} \leq \frac{C}{|\rho|} \tag{1.10}$$

with $-1 < \delta < 0$, C is a positive constant, and $\| \ \|_{\delta+1,\delta}$ denotes the operator norm.

Using the complex geometrical optics solutions of Lemma 1.1 Alessandrini proved stability estimates for the map (1.3). A reconstruction method using these solutions was proposed in [19], [25]. We remark that the construction of the solutions (1.6) is in the whole of $\mathbb{R}^n$. Complex geometrical solutions in compact sets have been constructed in [8], [10].

Theorem 1.1 extends to non-linear conductivities [29]. Theorem 1.2 extends to the non-linear Schrödinger equation under some additional assumptions on the potential [14]. These results use a linearization procedure due to Isakov [11].

MAXWELL'S EQUATIONS.

One obtains the conductivity equation (1.1) if one neglects the time variation of the electromagnetic field in Maxwell's equations. We now describe the boundary map in this case.

Let $\Omega \subseteq \mathbb{R}^3$ be a bounded domain with smooth boundary. The electromagnetic field (E, H) satisfies the frequency domain Maxwell's equation which are given by

$$\operatorname{rot} E = i\omega\mu H, \quad \operatorname{rot} H = (-i\omega\varepsilon + \sigma)E \text{ in } \Omega \tag{1.11}$$

where $\omega > 0$ is the time-harmonic frequency of the field, $\varepsilon > 0$ denotes the electrical permittivity, $\mu > 0$ the magnetic permeability, and $\sigma \geq 0$ the conductivity. We assume that all the functions are smooth. The boundary map is given by

$$\Lambda_{\varepsilon,\mu,\sigma}(\omega) : \nu \wedge E|_{\partial\Omega} \to \nu \wedge H|_{\partial\Omega}$$

where E, H satisfies (1.11). A global identifiability result was proven in this case in [26]. The proof was simplified in [27], where the problem is reduced to constructing

geometrical optics solutions for a Schrödinger equation with q an 8×8 matrix. Lemma 1.1 applies also in this case.

Open problem 1. How much smoothness should one assume on the conductivity for Theorem 1.1 to be valid? R. Brown extended Theorem 1.1 to conductivities in $C^{\frac{3}{2}+\epsilon}(\overline{\Omega})$, with ϵ any positive number. The natural conjecture is that the theorem holds for Lipschitz conductivities since unique continuation is valid in this case. There are no known counterexamples for rough conductivites. Kohn and Vogelius proved identifiability for piecewise real-analytic conductivities [16]. In [12] the case of a conductivity having a jump discontinuity across the boundary of a subdomain is considered.

Open problem 2. Is Theorem 1.1 valid if we measure the DN map only on part of the boundary?

Open problem 3. Is it possible to characterize the boundary values of the complex geometrical optics solutions (1.6)? This might have implications in the characterization and reconstruction problem.

Open problem 4. Is it possible to develop the reconstruction method based on the complex geometrical optics solutions into a convergent numerical algorithm?

Open problem 5. (The anisotropic case.) Conductivities may depend also on direction. Muscle tissue in the human body is an example. In this case the conductivity is represented by a positive definite matrix. It seems like a difficult problem to find complex geometric optics solutions in the anisotropic case. Moreover, it is not true that the DN map in this case determines uniquely the conductivity. See [38] for a discussion of the obstruction to identifiability in this case. The case of real analytic conductivities was considered in [17]. The case of quasilinear real-analytic anisotropic conductivities is discussed in [31]. For further results see [38].

2. The two dimensional case

Nachman proved in [20] that, in the two dimensional case, one can uniquely determine conductivities in $W^{2,p}(\Omega)$ for some $p > 1$ from Λ_γ. An essential part of Nachman's argument is the construction of the complex geometrical optics solutions (1.6) for all complex frequencies $\rho \in \mathbb{C}^2$, $\rho \cdot \rho = 0$, for potentials of the form (1.5). Then he applies the $\overline{\partial}$-method in inverse scattering, pioneered in one dimension by Beals and Coifman [2] and extended to higher dimensions by several authors (see [25] for further discussions and applications of the $\overline{\partial}$ method). The analog of Theorem 1.2 is open, in two dimensions, for a general potential $q \in L^\infty(\Omega)$. We outline a different approach to [20] that allows less regular conductivities.

The inverse conductivity problem.

We describe here an extension of Nachman's result to $W^{1,p}(\Omega), p > 2$, conductivities by Brown and the author [6]. We follow an earlier approach of Beals and Coifman [3], who studied scattering for a first order system whose principal part is $\begin{pmatrix} \overline{\partial} & 0 \\ 0 & \partial \end{pmatrix}$.

2.1 THEOREM. *Let $n = 2$. Let $\gamma \in W^{1,p}(\Omega), p > 2$, γ strictly positive. Assume $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$. Then $\gamma_1 = \gamma_2$ in $\overline{\Omega}$.*

We first reduce the conductivity equation to a first order system. We define the scalar potential q and matrix potential Q by

$$q = -\frac{1}{2}\partial \log \gamma, \quad Q = \begin{pmatrix} 0 & q \\ \overline{q} & 0 \end{pmatrix}. \tag{2.1}$$

We let D be the operator

$$D = \begin{pmatrix} \overline{\partial} & 0 \\ 0 & \partial \end{pmatrix}. \tag{2.2}$$

An easy calculation shows that if u satisfies the conductivity equation $\operatorname{div}(\gamma \nabla u) = 0$, then

$$D \begin{pmatrix} v \\ w \end{pmatrix} - Q \begin{pmatrix} v \\ w \end{pmatrix} = 0 \quad \text{with} \quad \begin{pmatrix} v \\ w \end{pmatrix} = \gamma^{\frac{1}{2}} \begin{pmatrix} \partial u \\ \overline{\partial} u \end{pmatrix}. \tag{2.3}$$

In [6] matrix solutions of (2.3) are constructed which have the form

$$u_k = m(z,k) \begin{pmatrix} e^{izk} & 0 \\ 0 & e^{-i\overline{z}k} \end{pmatrix}, \tag{2.4}$$

where $z = x_1 + ix_2$, $k \in \mathbb{C}$, with $m \to 1$ as $|z| \to \infty$ in a sense to be described below. To construct m we solve the integral equation

$$m - D_k^{-1} Q m = 1 \tag{2.5}$$

where, for a matrix-valued function A,

$$D_k A = E_k^{-1} D E_k A; \quad E_k A = A^d + \Lambda_k^{-1} A^{\text{off}}; \quad \Lambda_k(z) = \begin{pmatrix} e^{i(z\overline{k}+\overline{z}k)} & 0 \\ 0 & e^{-i(zk+\overline{z}\overline{k})} \end{pmatrix}.$$

Here A^d denotes the diagonal part of A and A^{off} the antidiagonal part.

The next result gives the solvability of (2.5) in an appropriate space.

2.1 LEMMA. *Let $Q \in L^p(\mathbb{R}^2), p > 2$, and compactly supported. Assume that Q is a hermitian matrix. Choose r so that $\frac{1}{p} + \frac{1}{r} > \frac{1}{2}$ and then β so that $\beta r > 2$. Then the operator $(I - D_k^{-1}Q)$ is invertible in $L^r_{-\beta}$. Moreover the inverse is differentiable in k in the strong operator topology. Here L^r_β denotes a weighted L^r space.*

Lemma 2.1 implies the existence of solutions of the form (2.4) with $m - 1 \in L^r_{-\beta}(\mathbb{R}^2)$ with β, r as in Lemma 2.1. The next step, following the $\overline{\partial}$ method, consists in relating $\frac{\partial}{\partial \overline{k}} m(z,k)$ and scattering data that in turn is determined from the DN map. For more details see [6].

Problem 5 has been solved in the anisotropic case in two dimensions for sufficiently smooth conductivities. By using isothermal coordinates, one can reduce the anisotropic case to the isotropic case, and therefore construct complex geometrical optics solutions in this case (see [34].) The case of quasilinear anisotropic conductivities is considered in [31].

OPEN PROBLEM 6: THE POTENTIAL CASE. Problems 1-4 are also open for the inverse conductivity problem in two dimensions. As we mentioned at the beginning of this section, the analog of Theorem 1.2 is unknown at present for a general potential $q \in L^\infty(\Omega)$. By Nachman's result it is true for potentials of the form $q = \frac{\Delta u}{u}$ with $u \in W^{2,p}(\Omega), u > 0$ for some p, $p > 1$. Sun and Uhlmann proved generic uniqueness for pairs of potentials in [32]. The semilinear case, under additional assumptions on the potential, was considered in [13]. In [33] it is shown that one can determine the singularities of an L^∞ potential from its Cauchy data.

3. First order perturbations of the Laplacian

There are several inverse boundary value problems associated to first order perturbations of the Laplacian. We consider briefly here an inverse boundary value problem associated to the Lamé system in elasticity theory.

We first discuss how to construct complex geometrical optics solutions for any scalar first order perturbation of the Laplacian.

Let $L_N = \Delta + N(x, D)$ with $N(x, D)$ a first order differential operator in $\mathbb{R}^n$ with smooth coefficients with compact support. We attempt to construct solutions u_ρ of $L_N u_\rho = 0$ of the form $u_\rho = e^{x\cdot\rho} m_\rho$. The equation for $m(x, \rho)$ is $M_\rho m_\rho := (\Delta_\rho + N_\rho) m_\rho = 0$ where $N_\rho f = e^{-x\cdot\rho} N(e^{x\cdot\rho} f)$ and Δ_ρ as in (1.9).

The difficulty in finding m_ρ is that the operator $\Delta_\rho^{-1} N_\rho$ contain terms that don't decay in $|\rho|$ in any reasonable norm. We get around this difficulty by conjugating the operator $\Delta_\rho + N_\rho$ to an operator that behaves like a zeroeth order perturbation of Δ_ρ. This idea is motivated by formula (1.5). To do this we consider pseudodifferential operators depending on a complex parameter [28]. For these operators the variable ρ behaves like the variable ξ. More precisely, we define $Z = \{\rho \in \mathbb{C}^n - 0; \rho \cdot \rho = 0\}$ and $A \in L^m(\mathbb{R}^n, Z)$ if we can write

$$Af(x) = \int e^{i\langle x,\xi\rangle} a_\rho(x, \xi) \widehat{f}(\xi) d\xi, \quad f \in C_0^\infty(\mathbb{R}^n), \text{ where } a_\rho \in S^m(\mathbb{R}^n, Z), i.e.$$

.

$$\sup_{x\in K} |\partial_x^\alpha \partial_\xi^\beta a_\rho(x, \xi)| \leq C_{\alpha,\beta,K}(1 + |\xi| + |\rho|)^{m-|\beta|}, \forall K \subset\subset \mathbb{R}^n.$$

.

We have that $\Delta_\rho \in L^2(\mathbb{R}^n, Z), N_\rho \in L^1(\mathbb{R}^n, Z)$. The key result proved in [21] is that one can conjugate $\Delta_\rho + N_\rho$ to $\Delta_\rho + C_\rho$, with $C_\rho \in L^0(\mathbb{R}^n, Z)$.

3.1 LEMMA. *Let $K \subset\subset \mathbb{R}^n$ be a compact subset. Let $M_\rho(x, D)$ be as defined above. Then there exist A_ρ, $B_\rho \in L^0(\mathbb{R}^n, Z)$ such that*

$$M_\rho A_\rho = B_\rho (\Delta_\rho + C_\rho), \tag{3.1}$$

where $C_\rho \in L^0(\mathbb{R}^n, Z)$. Moreover $\phi A_\rho \phi$ and $\phi B_\rho \phi$ are invertible on $L^2(K)$ for large $|\rho|$ for all $\phi \in C_0^\infty(\mathbb{R}^n)$ with $\phi = 1$ on K.

Now it is easy to construct many solutions l_ρ of $(\Delta_\rho + C_\rho)l_\rho = 0$ in any compact set since the operator $\phi C_\rho \phi$ is bounded on $L^2(\mathbb{R}^n)$, with operator norm independent of $|\rho|$ being a pseudodifferential operator of order zero depending on the parameter ρ (see [28] for more details on thiese operators.) Therefore, by the intertwining property (3.1), $m_\rho = A_\rho l_\rho$ is a solution of $M_\rho m_\rho = 0$. The construction of A_ρ, B_ρ is quite explicit.

In the paper [23], building on early work of Sun [30], these complex geometrical optics solutions were used to prove a global identifiability result for an inverse boundary value problem associated to the Schrödinger equation in the presence of smooth magnetic potential and electric potential. C. Tolmasky reduced the regularity needed in [39] to just one derivative for the magnetic potential, and a bounded electric potential, by using non-smooth symbols depending on the complex parameter ρ. The paper [18] also uses these solutions to prove a global identifiability result for Maxwell's equations in chiral media by reducing this case to a first order system perturbation of the Laplacian.

AN INVERSE BOUNDARY VALUE PROBLEM FOR THE ELASTICITY SYSTEM.

An inverse boundary value problem arising in the mechanics of materials is to determine the elastic parameters of a medium by making displacement and traction measurements at the boundary of the medium. We describe briefly below the boundary map in this case.

Let $\Omega \subseteq \mathbb{R}^n$ be a bounded open set with smooth boundary. We consider Ω as an elastic, isotropic, inhomogeneous medium with Lamé parameters λ, μ. The generalized Hooke's law states that under the assumption of no body forces acting on Ω, the displacement u satisfies

$$(Lu)_i = (L_{\lambda,\mu}u)_i = \sum_{j,k,l=1}^{n} \frac{\partial}{\partial x_j} C_{ijkl} \frac{\partial}{\partial x_l} u_k = 0 \text{ in } \Omega, \quad i = 1, \dots, n, \tag{3.2}$$

$$u|_{\partial\Omega} = f$$

where

$$C_{ijkl} = \lambda \delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{ij} + \delta_{il}\delta_{jk}) \quad (1 \le i, j, k, l \le n), \tag{3.3}$$

with δ_{ij} the Kronecker delta and $(Lu)_i$ denotes the i-th component of Lu.

$C = (C_{ijkl})$ is the elastic tensor. The boundary value problem (3.2) has a unique solution under the strong convexity condition $\mu > 0, n\lambda + 2\mu > 0$ in $\bar{\Omega}$.

The Dirichlet to Neumann map is defined in this case by

$$(\Lambda_{\lambda,u}(f))_i = \sum_{l,k,l=1}^{n} \nu_j C_{ijkl} \left. \frac{\partial u_k}{\partial x_l} \right|_{\partial\Omega}, \quad i = 1, ..., n \tag{3.4}$$

where $\nu = (\nu_1, \dots, \nu_n)$ is the unit outer normal to $\partial\Omega$ and u is the solution of (3.2). Physically the DN map sends the displacement at the boundary to the traction at the boundary. The following global identifiability result was proven in [21].

3.1 THEOREM. *Let $n \geq 3$. Let $(\lambda_i, \mu_i) \in C^\infty(\bar{\Omega}) \times C^\infty(\bar{\Omega}), i = 1, 2$ satisfy the strong convexity condition (3.3). Assume $\Lambda_{(\lambda_1,\mu_1)} = \Lambda_{(\lambda_2,\mu_2)}$. Then $(\lambda_1, \mu_1) = (\lambda_2, \mu_2)$ in $\bar{\Omega}$.*

The proof of Theorem 3.1 follows the general outline of the proof of Theorem 1.2. Namely, one proves an identity similar to (1.8) by using Green's theorem. Second, one reduces the elasticity system to a first order system (a more direct way to do this was given in [9]). Now one constructs geometrical optics solutions for the elasticity system using Lemma 3.1, which also applies to the first order system under consideration. The details of this outline can be found in [21].

OPEN PROBLEM 7. The analog of problems 1-5 are also open for the elasticity system. The analog of Theorem 2.1 is not known for the elasticity system in two dimensions. It is known that one can uniquely identify from the DN map Lamé parameters close to constant (see [22].) The methods of section 2 might be useful to prove a global identifiability result in this case.

ACKNOWLEDGEMENTS

The author would like to acknowledge support from the National Science Foundation and the Office of Naval Research, for the research presented in this paper.

I would also like to thank Eduardo Chappa, Rafe Mazzeo and Hart Smith for their helpful comments on an earlier version of the manuscript.

REFERENCES

1. Alessandrini, G.,, *Stable determination of conductivity by boundary measurements*, Appl. Anal. (1988), 153–172.
2. Beals R. and Coifman R., *Transformation spectrales et equation d' evolution non-lineares*, Seminaire Goulaouic-Meyer-Schwarz, exp. 21 (1981-1982).
3. ———, *The spectral problem for the Davey-Stewarson and Ishimori hierarchies*, Nonlinear evolution equations: Integrability and spectral methods, Manchester University Press (1988), 15–23.
4. Belishev M., *Boundary control in reconstruction of manifolds and metrics (the BC method)*, Inverse Problems **13** (1997), R1-R45.
5. Brown R., *Global uniqueness in the impedance imaging problem for less regular conductivities*, SIAM J. Math. Anal. **27** (1996), 1049–1056.
6. Brown R. and Uhlmann G., *Uniqueness in the inverse conductivity problem with less regular conductivities in two dimensions*, Comm. P.D.E. **22** (1997), 1009-1027.
7. A. P. Calderón, *On an inverse boundary value problem*, Soc. Brasileira de Matemática (1980), 65–73.
8. Hähner P, *A periodic Faddeev-type solution*, J. Diff. Eq. **128** (1996), 300–308.
9. Ikehata M., *A remark on an inverse boundary value problem arising in elasticity*, preprint.
10. Isakov V., *Completeness of products of solutions and some inverse problems for PDE*, J. Diff. Eq. **119** (1992), 59–70.
11. ———, *On uniqueness in inverse problems for semilinear parabolic equations*, Arch. Rat. mech. Anal. **92** (1993), 1–12.
12. ———, *On uniqueness of recovery of discontinuity conductivity*, Comm. Pure Appl. Math. **41** (1988), 865–877.
13. Isakov V. and Nachman A., *Global uniqueness for a two dimensional semilinear elliptic inverse problem*, Trans. AMS **347** (1995), 3375–3390.
14. Isakov V. and Sylvester J., *Global uniqueness for a semilinear elliptic inverse problem*, Comm. Pure Appl. Math. **47** (1994), 1403–1410.

15. Kohn R. and Vogelius M., *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math. **37** (1984), 289–298.
16. ———, *Determining conductivity by boundary measurements II. Interior results*, Comm. Pure Appl. Math **38** (1985), 643-667.
17. Lee J. and Uhlmann G., *Determining anisotropic real-analytic conductivities by boundary measurements*, Comm. Pure Appl. Math. **42** (1989), 1097–1112.
18. McDowall S., *An Electrodynamic inverse problem in chiral media*, Trans. AMS, to appear.
19. Nachman A., *Reconstruction from boundary measurements*, Annals of Math. **128** (1988), 71–96.
20. ———, *Global uniqueness for a two-dimensional inverse boundary value problem*, Annals of Math. **143**.
21. Nakamura G. and Uhlmann G., *Global uniqueness for an inverse boundary value problem arising in elasticity*, Invent. Math. **118** (1994), 457–474.
22. ———, *Identification of Lamé parameters by boundary measurements*, Amer. J. of Math. **115** (1993), 1161–1187.
23. Nakamura G., Sun Z. and Uhlmann G., *Global identifiability for an inverse problem for the Schrödinger equation in a magnetic field*, Math. Annalen **303** (1995), 377–388.
24. Novikov R., *Multidimensional inverse spectral problems for the equation* $-\Delta\psi + (v(x) - Eu(x))\psi = 0$, Functional Analysis and its Applications **22** (1988), 263–272.
25. Novikov R. and Henkin G., $\overline{\partial}$*- equation in the multidimensional inverse scattering problem*, Russian Math. Surveys **42** (1987), 109–180.
26. Ola P., Päivärinta L. and Somersalo E., *An inverse boundary value problem in electrodynamics*, Duke Math. J. **70** (1993), 617–653.
27. Ola P. and Somersalo E., *Electromagnetic inverse problems and generalized Sommerfeld potentials*, SIAM J. Appl. Math. **56** (1996), 1129–1145.
28. Shubin, M. A., *Pseudodifferential operators and spectral theory*, Springer Series in Soviet Mathematics, Springer-Verlag, 1987.
29. Sun Z., *On a quasilinear boundary value problem*, Math. Z. **221** (1996), 293–305.
30. ———, *An inverse boundary value problem for Schrödinger operator with vector potentials*, Trans. AMS **338** (1993), 953–969.
31. Sun Z. and Uhlmann G., *Inverse problems in quasilinear anisotropic media*, Amer. J. Math. **119** (1997), 771–797.
32. ———, *Generic uniqueness for an inverse boundary value problem*, Duke Math. J. **62** (1991), 131–155.
33. ———, *Recovery of singularities for formally determined inverse problems*, Comm. Math. Physics **153** (1993), 431–445.
34. Sylvester, J., *An anisotropic inverse boundary value problem*, Comm Pure Appl. Math. **43** (1990), 202–232.
35. Sylvester J. and Uhlmann G., *A uniqueness theorem for an inverse boundary value problem in electrical prospection Comm. Pure Appl. Math.* **39** (1986), 91–112.
36. ———, *A global uniqueness theorem for an inverse boundary value problem*, Annals of Math. **125** (1987), 153–169.
37. ———, *Inverse boundary value problems at the boundary–continuos dependence*, Comm. Pure Appl. Math. **41** (1988), 197–221.
38. ———, *Inverse problems in anisotropic media*, Contemp. Math. **122** (1991), 105–117.
39. Tolmasky C., *Exponentially growing solutions for non smooth first-order perturbations of the Laplacian*, SIAM J. Math. Anal. **29** (1998), 116–133.
40. Uhlmann G., *Inverse boundary value problems and applications*, Astérisque **207** (1992), 153–211.

Gunther Uhlmann
Department of Mathematics
Box 354350
University of Washington
Seattle, WA 98195-4350, USA

Scattering Theory: Some Old and New Problems

D. Yafaev

Abstract. Scattering theory is, roughly speaking, perturbation theory of self-adjoint operators on the (absolutely) continuous spectrum. It has its origin in mathematical problems of quantum mechanics and is intimately related to the theory of partial differential equations. Some recently solved problems, such as asymptotic completeness for the Schrödinger operator with long-range and multiparticle potentials, as well as open problems, are discussed. We construct also potentials for which asymptotic completeness is violated. This corresponds to a new class of asymptotic solutions of the time-dependent Schrödinger equation. Special attention is paid to the properties of the scattering matrix, which is the main observable of the theory.

1991 Mathematics Subject Classification: Primary 35J10, 47A75; Secondary 81U20
Keywords and Phrases: wave operators, asymptotic completeness, the N-particle Schrödinger operator, new channels of scattering, the scattering matrix

1. Basic notions. Let H_0 and H be self-adjoint operators on Hilbert spaces $\mathcal{H}_0$ and $\mathcal{H}$, respectively. Let $P_0^{(ac)}$ be the orthogonal projection on the absolutely continuous subspace $\mathcal{H}^{(ac)}(H_0)$ of H_0 and $J : \mathcal{H}_0 \to \mathcal{H}$ be a bounded operator. The main problem of mathematical scattering theory (see e.g. [23] or [31]) is to show the existence of the strong limits

$$W^{\pm} = W^{\pm}(H, H_0; J) = s - \lim_{t \to \pm\infty} \exp(iHt) J \exp(-iH_0 t) P_0^{(ac)}, \qquad (1)$$

known as the wave operators. If the limits (1) exist, then the wave operators enjoy the intertwining property $HW^{\pm} = W^{\pm}H_0$, so their ranges are contained in $\mathcal{H}^{(ac)}(H)$. In the most important case $\mathcal{H}_0 = \mathcal{H}$, $J = Id$, the limit (1) is isometric and is denoted $W^{\pm}(H, H_0)$. The operator $W^{\pm}(H, H_0)$ is said to be complete if its range coincides with $\mathcal{H}^{(ac)}(H)$. This is equivalent to the existence of $W^{\pm}(H_0, H)$. In terms of the operators (1) the scattering operator is defined by $\mathbf{S} = (W^+)^* W^-$. It commutes with H_0 and hence reduces to multiplication by the operator-function $S(\lambda)$, known as the scattering matrix, in a representation of $\mathcal{H}_0$ which is diagonal for H_0.

In scattering theory there are two essentially different approaches. One of them, the trace-class method, makes no assumptions about the "unperturbed"

operator H_0. Its basic result is the Kato-Rosenblum theorem (and its extension due to M. Birman and D. Pearson), which guarantees the existence of $W^{\pm}(H, H_0; J)$ if the perturbation $V = HJ - JH_0$ belongs to the trace class. According to the Weyl-von Neumann-Kuroda theorem this condition cannot be relaxed in the framework of operator ideals even in the case $J = Id$. The second, smooth, method relies on a certain regularity of the perturbation in the spectral representation of the operator H_0. There are different ways to understand regularity. For example, in the Friedrichs model [8] V is an integral operator with smooth kernel. Another possibility is to assume that $V = K^*K_0$ where K is H-smooth (in the sense of T. Kato which, roughly speaking, means that the function $||K \exp(-Ht)f||^2$ is integrable on $\mathbb{R}$ uniformly for $||f|| \leq 1$) and K_0 is H_0-smooth.

The assumptions of trace-class and smooth scattering theory are quite different. Thus it would be desirable to develop a theory unifying the trace-class and smooth approaches. Of course this problem admits different interpretations, but it becomes unambiguously posed in the context of applications, especially to differential operators. Suppose that $\mathcal{H} = L_2(\mathbb{R}^d)$, $H_0 = -\Delta + V_0(x)$, $H = H_0 + V(x)$ where V_0 and V are real bounded functions and $V(x) = O(|x|^{-\rho})$ as $|x| \to \infty$. Trace-class theory shows that the wave operators $W^{\pm}(H, H_0)$ exist (and are complete) if V_0 is an arbitrary bounded function and $\rho > d$. Smooth theory requires an explicit spectral analysis of the operator H_0, which is possible for special V_0 only (the simplest case $V_0 = 0$) but imposes the less stringent assumption $\rho > 1$ on the perturbation V. This raises

PROBLEM 1 *Let $d > 1$. Do the wave operators $W^{\pm}(H, H_0)$ exist for arbitrary $V_0 \in L_\infty(\mathbb{R}^d)$ and V satisfying the bound $V(x) = O(|x|^{-\rho})$, assuming only that $\rho > 1$?*

In the event of a positive solution of Problem 1, wave operators would be automatically complete under its assumptions. We conjecture, on the contrary, that Problem 1 has a negative solution. Moreover, we expect that the absolutely continuous part of the spectrum is no longer stable in the situation under consideration.

2. THE MULTIPARTICLE SCHRÖDINGER OPERATOR. One of the important problems of scattering theory is the description of the asymptotic behaviour of N interacting quantum particles for large times. The complete classification of all possible asymptotics (channels of scattering) is called asymptotic completeness. Let us recall the definition of generalized N-particle Hamiltonians introduced by S. Agmon. Consider the self-adjoint Schrödinger operator $H = -\Delta + V(x)$ on the Hilbert space $\mathcal{H} = L_2(\mathbb{R}^d)$. Suppose that some finite number α_0 of subspaces X^α of $X := \mathbb{R}^d$ are given and let x^α, x_α be the orthogonal projections of $x \in X$ on X^α and $X_\alpha = X \ominus X^\alpha$, respectively. We assume that $V(x) = \sum_{\alpha=1}^{\alpha_0} V^\alpha(x^\alpha)$, where V^α is a real function of the variable x^α satisfying the short-range condition

$$|V^\alpha(x^\alpha)| \leq C(1 + |x^\alpha|)^{-\rho}, \quad \rho > 1. \tag{2}$$

Many intermediary results are valid also for long-range pair potentials satisfying

$$|V^\alpha(x^\alpha)| + (1 + |x^\alpha|)\,|\nabla V^\alpha(x^\alpha)| \leq C(1 + |x^\alpha|)^{-\rho}, \quad \rho > 0. \tag{3}$$

The two-particle Hamiltonian H is recovered if $\alpha_0 = 1$ and $X^1 = X$. The three-particle problem is distinguished from the general situation by the condition that $X_\alpha \cap X_\beta = \{0\}$ for $\alpha \neq \beta$. Clearly, $V^\alpha(x^\alpha)$ tends to zero as $|x| \to \infty$ outside of any conical neighbourhood of X_α but $V^\alpha(x^\alpha)$ is constant on planes parallel to X_α. Due to this property the structure of the spectrum of H is much more complicated than in the two-particle case.

Let us consider linear sums $X^a = X^{\alpha_1} + X^{\alpha_2} + \ldots + X^{\alpha_k}$ of the subspaces X^{α_j}. Without loss of generality, one can suppose that X coincides with one of the X^a. We denote by $\mathcal{X}$ the set of all subspaces X^a with $X^0 := \{0\} \in \mathcal{X}$ included in it but X excluded. Let x^a and x_a be the orthogonal projections of $x \in X$ on the subspaces X^a and $X_a = X \ominus X^a$, respectively. The index a (or b) labels all subspaces $X^a \in \mathcal{X}$ and, in the multiparticle terminology, a parametrizes decompositions of an N-particle system into noninteracting clusters; x^a is the set of "internal" coordinates of all clusters, while x_a describes the relative motion of clusters.

For each a define an auxiliary operator $H_a = -\Delta + V^a$ with a potential $V^a = \sum_{X^\alpha \subset X^a} V^\alpha$, which does not depend on x_a. In the representation $L_2(X) = L_2(X_a) \otimes L_2(X^a)$, $H_a = -\Delta_{x_a} \otimes I + I \otimes H^a$, where $H^a = -\Delta_{x^a} + V^a$. The operator H^a corresponds to the Hamiltonian of clusters with their centers-of-mass fixed at the origin, $-\Delta_{x_a}$ is the kinetic energy of the center-of-mass motion of these clusters, and H_a describes an N-particle system with interactions between different clusters neglected. Eigenvalues of the operators H^a are called thresholds for the Hamiltonian H. We denote by Υ the set of all thresholds and eigenvalues of the Hamiltonian H. Let P^a be the orthogonal projection in $L_2(X^a)$ on the subspace spanned by all eigenvectors of H^a. Then $P_a = I \otimes P^a$ commutes with the operator H_a. Set also $H_0 = -\Delta, P_0 = I$. The basic result of scattering theory for N-particle Schrödinger operators is the following

Theorem 2 *Let assumption* (2) *hold. Then the wave operators* $W_a^\pm = W^\pm(H, H_a; P_a)$ *exist and are isometric on the ranges* $R(P_a)$ *of projections* P_a. *The subspaces* $R(W_a^\pm)$ *are mutually orthogonal, and scattering is asymptotically complete:*

$$\bigoplus_a R(W_a^\pm) = \mathcal{H}^{(ac)}, \quad \mathcal{H}^{(ac)} = \mathcal{H}^{(ac)}(H).$$

The spectral theory of multiparticle Hamiltonians starts with the following basic result (see [19], [22]). It is formulated in terms of the auxiliary operator $A = \sum(x_j D_j + D_j x_j)$, $D_j = -i\partial_j$, $j = 1, \ldots, d$. In what follows $E(\Lambda)$ is the spectral projection of the operator H corresponding to a Borel set $\Lambda \subset \mathbb{R}$ and Q is the operator of multiplication by $(x^2 + 1)^{1/2}$.

Theorem 3 *Let each pair potential* V^α *be a sum of two functions satisfying assumptions* (2) *and* (3), *respectively. Then eigenvalues of* H *may accumulate only at the thresholds of* H, *so the "exceptional" set* Υ *is closed and countable. Furthermore, for every* $\lambda \in \mathbb{R} \setminus \Upsilon$ *there exists a small interval* $\Lambda_\lambda \ni \lambda$ *such that the Mourre estimate for the commutator holds, i.e.,*

$$i([H, A]u, u) \geq c\|u\|^2, \quad c = c_\lambda > 0, \quad u \in E(\Lambda_\lambda)\mathcal{H}. \tag{4}$$

Finally, for any compact interval Λ such that $\Lambda \cap \Upsilon = \emptyset$ and any $r > 1/2$, the operator $Q^{-r}E(\Lambda)$ is H-smooth (the limiting absorption principle). In particular, the operator H does not have singularly continuous spectrum.

In the case $N = 2$ the limiting absorption principle suffices for construction of scattering theory but, for $N > 2$, one needs additional analytical information corresponding in some sense to the critical case $r = 1/2$. However, the operator $Q^{-1/2}$ is definitely not H-smooth even in the free case $H = -\Delta$. Hence we construct differential operators which improve the fall-off of functions $(\exp(-iHt)f)(x)$ for large t and x. Denote by $\langle \cdot, \cdot \rangle$ the scalar product in the space $\mathbb{C}^d$. Let $\nabla_a = \nabla_{x_a}$ be the gradient in the variable x_a and let $\nabla_a^\perp$,

$$(\nabla_a^\perp u)(x) = (\nabla_a u)(x) - |x_a|^{-2}\langle (\nabla_a u)(x), x_a\rangle x_a,$$

be its orthogonal projection in X_a on the plane orthogonal to the vector x_a. Let $\mathbf{\Gamma}_a$ be a closed cone in $\mathbb{R}^d$ such that $\mathbf{\Gamma}_a \cap X_b = \{0\}$ if $X_a \not\subset X_b$. Let $\chi(\mathbf{\Gamma}_a)$ denote its characteristic function. Our main analytical result is the following:

THEOREM 4 *Suppose that the assumptions of Theorem 3 hold. Then for any a, the operator $G_a = \chi(\mathbf{\Gamma}_a)Q^{-1/2}\nabla_a^\perp E(\Lambda)$ is H-smooth.*

In particular, for the free region $\mathbf{\Gamma}_0$, where all potentials V^α are vanishing, the operator $\chi(\mathbf{\Gamma}_0)Q^{-1/2}\nabla^\perp E(\Lambda)$ is H-smooth. By analogy with the radiation conditions in the two-particle case (see e.g. [24]), we refer to the estimates of Theorem 4 as radiation estimates.

Our proof of Theorem 4 hinges on the commutator method. To that end, we construct a first-order differential operator $M = \sum(m_j D_j + D_j m_j)$, $m_j = \partial m/\partial x_j$, such that, for any a, the commutator $[H, M]$ satisfies the estimate

$$i[H, M] \geq c_1 G_a^* G_a - c_2 Q^{-2r}, \quad r > 1/2, \quad c_1, c_2 > 0, \tag{5}$$

locally (that is, sandwiched by $E(\Lambda)$). Here the "generating" function m is real, smooth and homogeneous of degree 1 for $|x| \geq 1$. It is completely determined by the geometry of the problem, that is by the collection of subspaces X^α. Roughly speaking, we set $m(x) = \mu_a|x_a|$ in a neighbourhood of every subspace X_a with neighbourhoods of all subspaces $X_b \not\supset X_a$ removed from it. In particular, $m(x) = |x|$ in a free region where all potentials are vanishing. It is important that $m(x)$ is a convex function, which implies that $i[H_0, M] \geq 0$ (up to an error $O(|x|^{-3})$). The arguments of [18] show that H-smoothness of the operator G_a is a direct consequence of estimate (5) and of the limiting absorption principle.

For the proof of asymptotic completeness we first consider auxiliary wave operators

$$W^\pm(H, H_a; M^a E_a(\Lambda)), \quad W^\pm(H_a, H; M^a E(\Lambda)), \tag{6}$$

where the "identifications" M^a are again first-order differential operators with suitably chosen "generating" functions m^a. It is important that $m^a(x)$ equals zero in some conical neighbourhood of every X_α such that $X_a \not\subset X_\alpha$. Hence coefficients of the operator $(V - V^a)M^a$ vanish as $O(|x|^{-\rho})$, $\rho > 1$, at infinity. The analysis

of the commutator $[H_a, M^a]$ relies on Theorem 4. This shows that the "effective perturbation" $HM^a - M^a H_a$ can be factorized into a product $K^* K_a$ where K is H-smooth and K_a is H_a-smooth (locally), which implies the existence of the wave operators (6). The final step of the proof is to choose functions m^a in such a way that $\sum_a m^a = m$ and to verify that the range of the operator $W^{\pm}(H, H; ME(\Lambda))$ coincides with the subspace $E(\Lambda)\mathcal{H}$. This is again a consequence of the Mourre estimate (4).

In the three-particle case Theorem 2 was first obtained, under some additional assumptions, by L. Faddeev [7] (see also [10], [27]), who used a set of equations he derived for the resolvent of H. The optimal formulation in the three particle case is due to V. Enss [5]. The approach to asymptotic completeness relying on the Mourre estimate goes back to I. Sigal and A. Soffer [25]. Our proof given in [32] is closer to that of G. M. Graf [11]. In contrast to [11] we fit N-particle scattering theory into the standard framework of the smooth perturbation theory. Its advantage is that it admits two equivalent formulations: time-dependent as discussed above, and stationary, where unitary groups are replaced by resolvents. This allows us [33, 34] to obtain stationary formulas for the basic objects of the theory: wave operators, scattering matrix, etc. The approaches of [7] and of [25, 11, 32] are quite different and at the moment there is no bridge between them.

There are several Hamiltonians similar to the N-particle Schrödinger operator H for which the methods of [11] or [32] can be tried.

Problem 5 *Develop scattering theory for the discrete version of H (the Heisenberg model) acting in the space $L_2(\mathbb{Z}^d)$. The same question is meaningful for the generalization of $H = H_0 + V$ where $H_0 = -\sum \Delta_k$, $k = 1, \dots, N$, is replaced by a more general differential operator, say, by $\sum(-\Delta_k)^2$.*

Radiation estimates similar to those of Theorem 4 are crucial also for different proofs due to S. Agmon, T. Ikebe, H. Kitada, Y. Saito (see e.g. [12] and also [14, 37]) of asymptotic completeness for the two-particle Schrödinger operator with a long-range potential and for scattering on unbounded obstacles [3, 13]. Actually, only the case of the Dirichlet boundary condition was considered in [3, 13]. Therefore the following question naturally arises:

Problem 6 *Develop scattering theory for the operator $H = -\Delta$ in the complement of an unbounded domain Ω (for example, of a paraboloid) with Neumann or more general boundary conditions on Ω.*

3. Long-range pair potentials. If H is the two-particle Schrödinger operator with a short-range potential $V(x) = O(|x|^{-\rho})$, $\rho > 1$, then, by the definition of the wave operator, for any $f^0 \in L_2(\mathbb{R}^d)$ and $f^{\pm} = W^{\pm}(H, H_0) f^0$,

$$(\exp(-iHt) f^{\pm})(x) = \exp(i\Phi(x,t))(2it)^{-d/2} \hat{f}^0(x/(2t)) + o(1), \quad t \to \pm\infty, \quad (7)$$

where $\Phi_0(x,t) = x^2(4t)^{-1}$, $\hat{f}^0$ is the Fourier transform of f^0 and $o(1)$ denotes a function whose norm tends to zero as $t \to \pm\infty$. For long-range potentials satisfying the condition

$$|D^{\kappa} V(x)| \le C(1 + |x|)^{-\rho - |\kappa|}, \quad \rho > 0, \quad \forall \kappa, \quad (8)$$

the relation (7) can be used for the definition of the modified wave operator $\tilde{W}^{\pm}$: $f^0 \mapsto f^{\pm}$. (Actually, many results in the long-range case remain valid if (8) is satisfied for $|\kappa| \leq 2$ only but we shall not dwell upon this.) In this case however the phase function $\Phi(x,t) = \Phi_V(x,t)$ depends on a potential V and is constructed as an approximate solution of the corresponding eikonal equation. It follows from asymptotic completeness that relation (7) is fulfilled for every $f \in \mathcal{H}^{(ac)}$. The asymptotics (7) shows that, for $f \in \mathcal{H}^{(ac)}$, the solution $(\exp(-iHt)f)(x)$ "lives" in the region where $|x| \sim |t|$.

Similarly, in the N-particle short-range case, for any $f_a^0 \in L_2(X_a)$ and $f_{a,k}^{\pm} = W_a^{\pm}(\psi^{a,k} \otimes f_a^0)$,

$$(\exp(-iHt)f_{a,k}^{\pm})(x) = \psi^{a,k}(x^a)\exp(i\Phi_{a,k}(x_a,t))(2it)^{-d_a/2}\hat{f}_a^0(x_a/(2t)) + o(1), \quad (9)$$

where $d_a = \dim X_a$, $\Phi_{a,k}(x_a,t) = x_a^2(4t)^{-1} - \lambda^{a,k}t$. Theorem 2 implies that every $f \in \mathcal{H}^{(ac)}$ is an orthogonal sum of vectors $f_{a,k}^{\pm}$ satisfying (9). For N-particle systems with long-range pair potentials V^{α} the result is almost the same if condition (8) with some $\rho > \sqrt{3} - 1$ is fulfilled for all functions $V^{\alpha}(x^{\alpha})$. In this case again every $f \in \mathcal{H}^{(ac)}$ is an orthogonal sum of vectors $f_{a,k}^{\pm}$ satisfying (9) with suitable functions $\Phi_{a,k}(x_a,t)$. This result (asymptotic completeness) was obtained by V. Enss [6] for $N = 3$ and extended by J. Dereziński [4] (his method is different from [6] and uses some ideas of I. Sigal and A. Soffer) to an arbitrary number of particles (see also [26]).

4. NEW CHANNELS OF SCATTERING. It turns out that for some three- (and N-) particle systems with pair potentials satisfying (8) for $\rho < 1/2$, there exist channels of scattering different from (9). We rely on the following general construction [35, 36]. Suppose that $\mathbb{R}^d = X_1 \oplus X^1$, $\dim X_1 = d_1$, $\dim X^1 = d^1$, $d_1 + d^1 = d$, but we do not make any special assumptions about a potential $V(x) = V(x_1, x^1)$. Let us introduce the operator $H^1(x_1) = -\Delta_{x^1} + V(x_1, x^1)$ acting on the space $L_2(X^1)$. Suppose that $H^1(x_1)$ has a negative eigenvalue $\lambda(x_1)$, and denote by $\psi(x_1, x^1)$ a corresponding normalized eigenfunction. In interesting situations the function $\lambda(x_1)$ tends to zero slower than $|x_1|^{-1}$. Let us consider it as an "effective" potential energy and associate to the long-range potential $\lambda(x_1)$ the phase function $\Phi = \Phi_{\lambda}$. We prove, under some assumptions, that for every $g \in L_2(X_1)$ there exists an element $f^{\pm} \in \mathcal{H}^{(ac)}$ such that

$$(\exp(-iHt)f^{\pm})(x) = \psi(x_1, x^1)\exp(i\Phi(x_1,t))(2it)^{-d_1/2}g(x_1/(2t)) + o(1) \quad (10)$$

as $t \to \pm\infty$. The mapping $\check{g} \mapsto f^{\pm}$ ($\check{g}$ is the inverse Fourier transform of g) defines the new wave operator $\mathcal{W}^{\pm}$. It is isometric on $L_2(X_1)$, and $H\mathcal{W}^{\pm} = \mathcal{W}^{\pm}(-\Delta_{x_1})$. The ranges of $\mathcal{W}^{\pm}$ and of $\tilde{W}^{\pm}$ are orthogonal if both of these wave operators exist. The existence of solutions of the time-dependent Schrödinger equation with asymptotics (10) requires rather special assumptions which are naturally formulated in terms of eigenfunctions $\psi(x_1, x^1)$. Typically the asymptotic behaviour of $\psi(x_1, x^1)$ as $\lambda(x_1) \to 0$ has a certain self-similarity:

$$\psi(x_1, x^1) = |x_1|^{-\sigma d^1/2}\Psi(|x_1|^{-\sigma}x^1) + o(1) \quad (11)$$

for some $\Psi \in L_2(X^1)$ and $\sigma > 0$. We prove the asymptotics (10) if (11) is fulfilled for $\sigma < 1/2$. On the other hand, simple examples show that (11) for $\sigma \geq 1/2$ does not ensure existence of solutions with the asymptotics (10). It is important that $\psi(x_1, x^1)$ can be chosen as an *approximate* solution of the equation $H^1(x_1)\psi(x_1) - \lambda(x_1)\psi(x_1) = 0$.

Let us first give an example of a two-body long-range potential (see [36], for more general classes) for which the completeness of the modified wave operator is violated. Let

$$V(x_1, x^1) = -v(< x_1 >^q + < x^1 >^q)^{-\rho/q}, \quad \rho \in (0,1),\ q \in (0,2),\ v > 0, \quad (12)$$

where we use the notation $< y >= (1 + |y|^2)^{1/2}$. The function (12) is infinitely differentiable and $V(x) = O(|x|^{-\rho})$ as $|x| \to \infty$. The bound (8) is fulfilled for arbitrary κ off any conical neighbourhood of the planes X_1 and X^1. This suffices for the existence of the modified wave operator $\tilde{W}^{\pm}$. If $q = 2$, then $V(x_1, x^1)$ is a radial function, so $\tilde{W}^{\pm}$ is complete.

THEOREM 7 *Let a potential V be defined by* (12) *where $1 - \rho < q < 2(1 - \rho)$. Let Λ be any eigenvalue and Ψ be a corresponding eigenfunction of the operator $K = -\Delta_{x^1} + v\rho q^{-1}|x^1|^q$ in the space $L_2(X^1)$. Define the function $\psi(x_1, x^1)$ and the "potential" $\lambda(x_1)$ by the equations*

$$\psi(x_1, x^1) = |x_1|^{-\sigma/2}\Psi(|x_1|^{-\sigma}x^1), \quad \lambda(x_1) = -v|x_1|^{-\rho} + \Lambda|x_1|^{-2\sigma}, \quad (13)$$

where $\sigma = (\rho + q)(2 + q)^{-1}$ and set $\Phi = \Phi_\lambda$. Then the wave operator $\mathcal{W}^{\pm}$ exists and the subspaces $R(\mathcal{W}^{\pm})$, $R(\tilde{W}^{\pm})$ are orthogonal.

Let us now consider the Schrödinger operator, which describes three one-dimensional particles with one of three pair interactions equal to zero. The following result was obtained in [35].

THEOREM 8 *Let $V(x) = V^1(x^1) + V^2(x^1 - x_1)$ where $d^1 = d_1 = 1$. Suppose that $V^1 \geq 0$ is a bounded function, $V^1(x^1) = 0$ for $x^1 \geq 0$ and $V^1(x^1) = v_1|x^1|^{-r}$, $v_1 > 0$, $r \in (0,2)$, for large negative x^1. Suppose that a bounded function V^2 satisfies for some $\rho \in (0, 1/2)$ and $v_2 > 0$ one of the two following conditions*: 1^0 *$V^2(x^2) = -v_2|x^2|^{-\rho}$ for large positive x^2*; 2^0 *$V^2(x^2) = v_2|x^2|^{-\rho}$ for large negative x^2. Let Λ be any eigenvalue and Ψ a corresponding eigenfunction of the equation $-\Psi'' + |v_2|\rho x^1\Psi = \Lambda\Psi$ for $x^1 \geq 0$, $\Psi(0) = 0$, extended by 0 to $x^1 \leq 0$. Define the function $\psi(x_1, x^1)$ and the "potential" $\lambda(x_1)$ by the equations* (13), *where $\sigma = (\rho + 1)/3$ and $v = -v_2$, $x_1 < 0$ in the case 1^0, $v = v_2$, $x_1 > 0$ in the case 2^0. Put $\Phi = \Phi_\lambda$. Then the wave operator $\mathcal{W}^{\pm}$ defined by equality* (10) *exists for any $g \in L_2(\mathbb{R}_{\mp})$ in the first case and for any $g \in L_2(\mathbb{R}_{\pm})$ in the second case. Moreover, the subspaces $R(\mathcal{W}^{\pm})$, $R(\tilde{W}_0^{\pm})$ and $R(\tilde{W}_\alpha^{\pm})$, $\alpha = 1, 2$, are orthogonal.*

We emphasize that for $f \in R(\mathcal{W}^{\pm})$ the solution $u(t) = \exp(-iHt)f$ of the Schrödinger equation "lives" for large $|t|$ in the region where $x_1 \sim -|t|$ in the case 1^0 or $x_1 \sim |t|$ in the case 2^0 and $x^1 \sim |t|^\sigma$ for $\sigma \in (1/3, 1/2)$. Such solutions describe a physical process where a pair of particles (say, the first and the second)

interacting by the potential V^1 are relatively close to one another and the third particle is far away. This pair is bound by a potential depending on the position of the third particle, but this bound state is evanescent as $|t| \to \infty$. Thus solutions $u(t)$ for $f \in R(\mathcal{W}^\pm)$ are intermediary between those for $f \in R(\tilde{W}_0^\pm)$ and $f \in R(\tilde{W}_\alpha^\pm)$.

There is however a gap between the cases when asymptotic completeness holds and when it is violated. Hence the following questions arise.

PROBLEM 9 *Is the scattering asymptotically complete when* $\rho \in [1/2, \sqrt{3}-1)$? *The same question for all* $\rho < \sqrt{3}-1$ *if particles are, say, three-dimensional.*

Note that, under some additional assumptions, asymptotic completeness for all $\rho > 1/2$ was checked in [28, 9]. In the cases when new channels are constructed one can expect that all possible asymptotics of the time-dependent Schrödinger equation have either the form (9) or (10). In a somewhat similar situation a result of such type was established in [30]. Thus, we formulate

PROBLEM 10 *To prove* (*for example, under the assumptions of Theorems* 7 *and* 8) *generalized asymptotic completeness, that is, that the ranges of all wave operators constructed exhaust* $\mathcal{H}^{(ac)}(H)$.

5. THE SCATTERING MATRIX. For the two-particle Schrödinger operator $H = -\Delta + V(x)$ in the space $L_2(\mathbb{R}^d)$, the scattering matrix $S(\lambda)$, $\lambda > 0$, is a unitary operator on the space $L_2(\mathbb{S}^{d-1})$. If V is a short-range potential, then the operator $S(\lambda) - Id$ is compact so the spectrum of $S = S(\lambda)$ consists of eigenvalues $\mu_n^\pm = \exp(\pm i\theta_n^\pm)$, $\pm\theta_n^\pm > 0$, lying on the unit circle $\mathbb{T}$ and accumulating only at the point 1. Moreover, the asymptotics of the scattering phases $\theta_n^\pm$ is determined by the asymptotics of the potential $V(x)$ at infinity and is given by the Weyl type formula. The following assertion was established in [2].

THEOREM 11 *Let* $V(x) = v(x|x|^{-1})|x|^{-\rho} + o(|x|^{-\rho})$, $\rho > 1$, $v \in C^\infty(\mathbb{S}^{d-1})$, *as* $|x| \to \infty$. *Then* $n^\gamma \theta_n^\pm(\lambda) \to \Omega_\pm$ *as* $n \to \infty$, *where* $\gamma = (\rho-1)(d-1)^{-1}$ *and* $\Omega_\pm$ *is some explicit functional of* v *and* ρ, λ.

The situation is drastically different for long-range potentials. Note that in this case modified wave operators can be defined [14, 15] by equality (1) where $J_\pm$ is a suitable pseudo-differential operator. It depends on the sign of t. The existence and completeness of the operators $W_\pm(H, H_0; J_\pm)$ follow immediately from Theorem 4, which fits the long-range scattering into the theory of smooth perturbations. In the long-range case, $S(\lambda) - Id$ is no longer compact. Moreover, its spectrum covers the whole unit circle. For simplicity we give the precise formulation only for the case $\rho > 1/2$ (see [37], for details).

THEOREM 12 *Let condition* (8) *with* $\rho > 1/2$ *hold. Suppose that the function*

$$\mathbf{V}(\omega, b) = \int_{-\infty}^{\infty} \Big(V(t\omega) - V(b + t\omega) \Big) dt, \quad |\omega| = 1, \quad \langle \omega, b \rangle = 0, \tag{14}$$

satisfies the condition $|\mathbf{V}(\omega_0, t_n b_0)| \to \infty$ *for some point* ω_0, b_0 *and some sequence* $t_n \to \infty$. *Then for all* $\lambda > 0$ *the spectrum of the scattering matrix* $S(\lambda)$ *covers the unit circle.*

Our study of the scattering matrix relies on its stationary representation (in terms of the resolvent). First, using the so called microlocal or propagation estimates [20, 17, 16], we show that, up to an integral operator with C^∞-kernel, S can be considered as a pseudo-differential operator with explicit principal symbol

$$\mathbf{s}(\omega, b; \lambda) = \exp\Big(i2^{-1}\lambda^{-1/2}\mathbf{V}(\omega, \lambda^{-1/2}b)\Big), \quad |\omega| = 1, \quad \langle \omega, b\rangle = 0. \qquad (15)$$

If $\rho \le 1$, this is an oscillating function as $|b| \to \infty$, which implies Theorem 12. Note that in the short-range case the principal symbol of S equals 1 which corresponds to the Dirac-function in its kernel. In the long-range case this singularity disappears.

The kernel $s(\omega, \omega')$ of S (the scattering amplitude) is [1, 15] a C^∞-function off the diagonal. Its diagonal singularity is given by the Fourier transform of the symbol (15). It turns out [37] that for an asymptotically homogeneous function $V(x)$ of order $-\rho$, $\rho < 1$, the kernel s is a sum of a finite number of terms $s_j = w_j \exp(i\psi_j)$, where the moduli $w_j(\omega, \omega')$ and the phases $\psi_j(\omega, \omega')$ are asymptotically homogeneous functions, as $\omega - \omega' \to 0$, of orders $-(d-1)(1+\rho^{-1})/2 < -d+1$ and $1-\rho^{-1}$, respectively. Thus S is more singular than the singular integral operator. In the case $\rho = 1$, the modulus $w = |s|$ is asymptotically homogeneous of order $-d+1$, and the phase ψ of s has a logarithmic singularity on the diagonal.

In the N-particle case results on the structure of the scattering matrix S are scarce. Let us mention the one by R. Newton [21] (see also [29], for an elementary proof), which asserts that in the 3-particle case, $S = S_1 S_2 S_3 \tilde{S}$, where S_α is the scattering matrix for the Hamiltonian with only one pair interaction V^α and the operator $\tilde{S} - Id$ is compact. We conclude with

PROBLEM 13 *Extend the above result to an arbitrary* N.

REFERENCES

[1] S. Agmon, Some new results in spectral and scattering theory of differential operators in $\mathbb{R}^n$, Seminaire Goulaouic Schwartz, Ecole Polytechnique, 1978.

[2] M. Sh. Birman and D. R. Yafaev, Asymptotics of the spectrum of the scattering matrix, J. Soviet Math. 25 no. 1 (1984), 793-814.

[3] P. Constantin, Scattering for Schrödinger operators in a class of domains with noncompact boundaries, J. Funct. Anal. 44 (1981), 87-119.

[4] J. Dereziński, Asymptotic completeness of long-range quantum systems, Ann. Math. 138 (1993), 427-473.

[5] V. Enss, Completeness of three-body quantum scattering, in: *Dynamics and processes*, P. Blanchard and L. Streit, eds., Springer Lecture Notes in Math. 1031 (1983), 62-88.

[6] V. Enss, Long-range scattering of two- and three-body quantum systems, in: *Journées EDP,* Saint Jean de Monts (1989), 1-31.

[7] L. D. Faddeev, *Mathematical Aspects of the Three Body Problem in Quantum Scattering Theory*, Israel Program of Sci. Transl., 1965.

[8] L. D. Faddeev, On the Friedrichs model in the theory of perturbations of the continuous spectrum, Amer. Math. Soc. Transl. Ser.2 62 (1967).

[9] C. Gérard, Asymptotic completeness of 3-particle long-range systems, Invent. Math. 114 (1993), 333-397.

[10] J. Ginibre and M. Moulin, Hilbert space approach to the quantum mechanical three body problem, Ann. Inst. H.Poincaré, A 21(1974), 97-145.
[11] G. M. Graf, Asymptotic completeness for N-body short-range quantum systems: A new proof, Comm. Math. Phys. 132 (1990), 73-101.
[12] L. Hörmander, *The analysis of linear partial differential operators* IV, Springer-Verlag, 1985.
[13] E. M. Il'in, Scattering by unbounded obstacles for elliptic operators of second order, Proc. of the Steklov Inst. of Math. 179 (1989), 85-107.
[14] H. Isozaki, H. Kitada, Modified wave operators with time-independent modifies, J. Fac. Sci, Univ. Tokyo, 32 (1985), 77-104.
[15] H. Isozaki, H. Kitada, Scattering matrices for two-body Schrödinger operators, Sci. Papers College Arts and Sci., Univ. Tokyo, 35 (1985), 81-107.
[16] A. Jensen, Propagation estimates for Schrödinger-type operators, Trans. Amer. Math. Soc. 291 (1985), 129-144.
[17] A. Jensen, E. Mourre, P. Perry, Multiple commutator estimates and resolvent smoothness in quantum scattering theory, Ann. Inst. Henri Poincaré, ph. th., 41 (1984), 207-225.
[18] T. Kato, Smooth operators and commutators, Studia Math. 31(1968), 535-546.
[19] E. Mourre, Absence of singular spectrum for certain self-adjoint operators, Comm. Math. Phys. 78 (1981), 391-400.
[20] E. Mourre, Opérateurs conjugués et propriétés de propagation, Comm. Math. Phys. 91 (1983), 279-300.
[21] R. Newton, The three particle S-matrix, J. Math. Phys. 15 (1974), 338-343.
[22] P. Perry, I. M. Sigal and B. Simon, Spectral analysis of N-body Schrödinger operators, Ann. Math. 144 (1981), 519-567.
[23] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* III, Academic Press, 1979.
[24] Y. Saito, *Spectral Representation for Schrödinger Operators with Long-Range Potentials*, Springer Lecture Notes in Math. 727, 1979.
[25] I. M. Sigal and A. Soffer, The N-particle scattering problem: Asymptotic completeness for short-range systems, Ann. Math. 126(1987), 35-108.
[26] I. M. Sigal and A. Soffer, Asymptotic completeness of N-particle long range systems, Journal AMS 7 (1994), 307-333.
[27] L. E. Thomas, Asymptotic completeness in two- and three-particle quantum mechanical scattering, Ann. Phys. 90 (1975), 127-165.
[28] X. P. Wang, On the three body long range scattering problems, Reports Math. Phys. 25(1992), 267-276.
[29] D. R. Yafaev, On the multichannel scattering theory in two spaces, Theor. Math. Phys. 37 (1978), 867-874.
[30] D. R. Yafaev, Scattering theory for time-dependent zero-range potentials, Ann. Inst. H.Poincaré, ph.th., 40 (1984), 343-359.
[31] D. R. Yafaev, *Mathematical Scattering Theory*, Amer. Math. Soc., 1992.
[32] D. R. Yafaev, Radiation conditions and scattering theory for N-particle Hamiltonians, Comm. Math. Phys. 154 (1993), 523-554.
[33] D. R. Yafaev, Eigenfunctions of the continuous spectrum for the N-particle Schrödinger operator, in: *Spectral and scattering theory*, M. Ikawa, ed., Marcel Dekker, Inc. (1994), 259-286.
[34] D. R. Yafaev, Resolvent estimates and scattering matrix for N-particle Hamiltonians, Int. Eq. Op. Theory 21 (1995), 93-126.
[35] D. R. Yafaev, New channels of scattering for three-body quantum systems with long-range potentials, Duke Math. J. 82 (1996), 553-584.
[36] D. R. Yafaev, New channels in the two-body long-range scattering, St. Petersburg Math. J. 8 (1997), 165-182.
[37] D. R. Yafaev, The scattering amplitude for the Schrödinger equation with a long-range potential, Comm. Math. Phys. 191 (1998), 183-218.

D. Yafaev, Department of Mathematics, University Rennes-1
Campus Beaulieu, 35042, Rennes, France, yafaev@univ-rennes1.fr

Section 11

Mathematical Physics

In case of several authors, Invited Speakers are marked with a *.

Spectral Statistics

Eugene Bogomolny

Abstract. The relation between statistical properties of energy eigenvalues of deterministic systems and the distribution of periodic orbits is discussed.

1991 Mathematics Subject Classification: 81Q50, 15A52, 11M26
Keywords and Phrases: Quantum chaos, random matrices, spectral statistics

1 Introduction

The purpose of this paper is to attract attention to the subject of statistical distribution of deterministic sequences. In quantum chaos problems they can be the eigenvalues of a quantum mechanical problem, in number theory the natural choice is the imaginary parts of non-trivial zeros of zeta functions, etc.

The important point that makes this field interesting is the observation that statistical distributions of completely different sequences are, to a large extent, universal depending only on very robust properties of the system considered. The origin of such universal laws remains unclear.

In the fifties Wigner and later Dyson (see articles in [1] and the review [2]), based on a physical idea that 'complicated' means 'random', have proposed to consider the Hamiltonian of heavy nuclei as a random matrix taken from a certain ensemble characterized only by symmetry properties. The duality: 'Hamiltonian $\longleftrightarrow$ random matrix' has been proved very useful [3], [4] and stimulated the development of random matrix theory [5]. Later it was understood that the same idea can also be applied to low-dimensional quantum systems and the accepted conjectures are: (i) local statistical behaviour of energy levels of classically integrable systems is close to the Poisson distribution [7], (ii) energy levels of classically chaotic systems are distributed as eigenvalues of random matrices from the standard random matrix ensembles [6]. One of these ensembles (Gaussian Unitary Ensemble (GUE)) seems to describe the local spectral distribution of non-trivial zeros of zeta functions of number theory [8]-[11].

The volume of numerical evidences in the favor of these conjectures is impressive (see e.g. [3], [9], [4]) but the full mathematical proof even in the simplest cases is still lacking.

In this paper we shall discuss a straightforward method to attack this problem based on trace formulae.

2 Trace formulae

The Gutzwiller trace formula [12] states that the density of eigenvalues for a quantum system can be written as a sum of a smooth ($\bar{d}$) term and an oscillating part

$$d^{(osc)}(E) = \sum_{ppo} \sum_{n=1}^{\infty} A_{p,n} \exp(\frac{i}{\hbar} n S_p(E)) + c.c. \ , \tag{1}$$

given by a sum over primitive periodic orbits and their repetitions. Here S_p is the classical action calculated along one of such orbits,

$$A_{p,n} = \frac{T_p}{2\pi\hbar |Det(M_p^n - 1)|^{1/2}} \exp(-i\frac{\pi}{2} n \mu_p),$$

M_p is the monodromy matrix around the orbit, T_p is its period, and μ_p is the Maslov index. For the motion on constant negative curvature surfaces generated by discrete groups this formula coincides with the Selberg trace formula but for generic systems it represents only the first term of a formal expansion on the Planck constant.

Similar expresion exists also for the Riemann zeta function. For the density of nontrivial Riemann zeros (assuming $s_n = \frac{1}{2} + iE_n$)

$$d^{(osc)}(E) = -\frac{1}{\pi} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} \Lambda(n) \cos(E \log n), \tag{2}$$

where $\Lambda(n) = \log p$, if n is a power of a prime p, and $\Lambda(n) = 0$ otherwise.

3 Correlation functions

The n-point correlation function of energy levels is defined as the probability of having n levels at prescribed positions

$$R_n(\epsilon_1, \epsilon_2, \ldots, \epsilon_n) = < d(E+\epsilon_1) d(E+\epsilon_2) \ldots d(E+\epsilon_n) >, \tag{3}$$

where the brackets $< \ldots >$ denote the smoothing over an energy window

$$< f(E) > = \int f(E') \sigma(E - E') dE', \tag{4}$$

with an appropriate weighting function $\sigma(E)$ centered near zero.

In particular, the 2-point correlation function has the form

$$\begin{aligned} R_2(\epsilon_1, \epsilon_2) \quad &= \bar{d}^2 + \sum_{p_i, n_i} A_{p_1,n_1} A^*_{p_2,n_2} < \exp(\frac{i}{\hbar}(n_1 S_{p_1}(E) - n_2 S_{p_2}(E))) > \\ &\times \exp(\frac{i}{\hbar}(n_1 T_{p_1}(E)\epsilon_1 - n_2 T_{p_2}(E)\epsilon_2)) + c.c. \end{aligned} \tag{5}$$

The terms with the sum of actions are assumed to be washed out by the smoothing procedure.

4 Diagonal approximation

Berry in [13] proposed to estimate the above sum by taking into account only terms with exactly the same actions which leads to the following expression for the two-point correlation form factor (the Fourier transform of R_2)

$$K^{(diag)}(t) = 2\pi \sum_{p,n} |A_{p,n}|^2 \delta(t - nT_p(E)) + c.c., \tag{6}$$

where the sum is taken over all periodic orbits with exactly the same action.

Using the Ruelle-Bowen-Sinai measure on periodic orbits (called in physical literature the Hannay-Ozorio de Almeida sum rule [14]) one finds that for ergodic systems

$$K^{(diag)}(t) = g\frac{t}{2\pi}, \tag{7}$$

where g is the mean multiplicity of periodic orbits. For generic systems without time reversal invariance $g = 1$ and for systems with time reversal invariance $g = 2$ and this result coincides with the small-t behaviour of form factor of classical ensembles.

Unfortunately, $K^{(diag)}(t)$ grows with t but the exact form factor for systems without spectral degeneracies should tend to $\bar{d}$ when $t \to \infty$. This contradiction clearly indicates that the diagonal approximation cannot be correct for all values of t and more complicated tools are needed to obtain the full form factor.

5 Beyond the diagonal approximation

We begin to discuss the calculation of off-diagonal terms on the example of the Riemann zeta function where more information is available and then we shall generalize the method to dynamical systems.

The connected two-point correlation function of the Riemann zeros is

$$R_2(\epsilon_1, \epsilon_2) = \frac{1}{4\pi^2} \sum_{n_1,n_2} \frac{\Lambda(n_1)\Lambda(n_2)}{\sqrt{n_1 n_2}} < e^{iE\log(n_1/n_2)+i(\epsilon_1 \log n_1 - \epsilon_2 \log n_2)} > + c.c. \tag{8}$$

The diagonal terms correspond to $n_1 = n_2$ and

$$R_2^{(diag)}(\epsilon) = -\frac{1}{4\pi^2}\frac{\partial^2}{\partial\epsilon^2} \log(|\zeta(1+i\epsilon)|^2 \Phi^{(diag)}(\epsilon)), \tag{9}$$

where $\epsilon = \epsilon_1 - \epsilon_2$ and the function $\Phi^{(diag)}(\epsilon)$ is given by a convergent sum over prime numbers

$$\Phi^{(diag)}(\epsilon) = \exp(-\sum_p \sum_{m=1}^{\infty} \frac{m-1}{m^2 p^m} e^{im\log p\epsilon} + c.c.). \tag{10}$$

When $\epsilon \to 0$, $R_2(\epsilon) \to -(2\pi^2\epsilon^2)^{-1}$ which agrees with the smooth GUE result.

The term $\exp(iE\log(n_1/n_2))$ oscillates quickly if n_1 is not close to n_2. Denoting $n_1 = n_2 + h$ and expanding smooth functions on h one gets

$$R_2^{(off)}(\epsilon) = \frac{1}{4\pi^2}\sum_{n,d}\frac{\Lambda(n)\Lambda(n+h)}{n} < e^{iE(h/n)+i\epsilon\log n} > + c.c. \tag{11}$$

The main problem is clearly seen here. The function $F(n,h) = \Lambda(n)\Lambda(n+h)$ changes irregularly as it is nonzero only when both n and $n+h$ are powers of prime numbers. Fortunately, the dominant contribution to the two-point correlation function comes from the mean value of this function

$$\alpha(h) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\Lambda(n)\Lambda(n+h), \tag{12}$$

and its explicit expression follows from the famous Hardy–Littlewood conjecture [15]

$$\alpha(h) = \sum_{(p,q)=1} e^{-2\pi i\frac{p}{q}h}\left(\frac{\mu(q)}{\psi(q)}\right)^2, \tag{13}$$

where the sum is taken over all coprime integers q and $p < q$, $\mu(n)$ and $\psi(n)$ are the Mobius and the Euler functions respectively.

Using this expression for $\alpha(h)$ and performing the sum over all h one obtains

$$R_2^{(off)}(\epsilon) = \frac{1}{4\pi^2}|\zeta(1+i\epsilon)|^2 e^{2\pi i\bar{d}\epsilon}\Phi^{(off)}(\epsilon) + c.c., \tag{14}$$

where function $\Phi^{(off)}(\epsilon)$ is given by a convergent product over primes

$$\Phi^{(off)}(\epsilon) = \prod_p(1 - \frac{(1-p^{i\epsilon})^2}{(p-1)^2}). \tag{15}$$

In the limit of small ϵ, $R_2^{(off)}(\epsilon) \to (e^{2\pi i\bar{d}\epsilon} + e^{-2\pi i\bar{d}\epsilon})/(2\pi\epsilon)^2$ which corresponds exactly to the GUE result.

The above calculations demonstrate how one can compute the two-point correlation function through the knowledge of pair-correlation function of periodic orbits. For the Riemann case one can prove under the same conjectures[1] that all n-point correlation functions of Riemann zeros tend to the corresponding GUE results [16].

The interesting consequence of the above formula is the expression for the two-point form factor

$$K^{(off)}(t) = \frac{1}{4\pi^2}\sum_{(p,q)=1}\left(\frac{\mu(q)}{\psi(q)}\right)^2(\frac{q}{p})\delta(t - 2\pi\bar{d} - \log\frac{q}{p}), \tag{16}$$

which means that the off-diagonal two-point form factor is a sum over δ-functions in special points equal the Heisenberg time ($T_H = 2\pi\bar{d}$) plus a difference of periods

[1] Really only a smoothed version of the Hardy-Littlewood conjecture is needed.

of two pseudo-orbits (linear combinations of periodic orbits). This set is dense but the largest peaks correspond to the shortest pseudo-orbits. Similarly the two-point diagonal form factor is the sum of δ functions at the positions of periodic orbits

$$K^{(diag)}(t) = \frac{1}{4\pi^2} \sum_{p,m} \frac{\log^2 p}{p^m} \delta(t - m \log p). \tag{17}$$

The smooth values corresponding to the random matrix predictions appear only after a smoothing of these functions over a suitable interval of t.

6 Arithmetical systems

Similar behavior has been observed in a completely different model, namely for the distribution of eigenvalues of the Laplace–Beltrami operator for the modular domain [17]. It was shown that in this model the two-point correlation form factor can be written in the following form

$$K(t) = \frac{1}{\pi^3 k} \sum_{(p,q)=1} \left| \frac{q}{p} \beta(p,q) \right|^2 \delta(t - t_{p,q}), \tag{18}$$

where

$$t_{p,q} = \frac{2}{k} \ln \frac{kq}{\pi p}, \text{ and } \beta(p,q) = \frac{S(p,p;q)}{q^2 \prod_{\omega|q}(1 - \omega^{-2})}.$$

The product is taken over all prime divisors of q and $S(p,p;q)$ is the Kloosterman sum

$$S(n,m;c) = \sum_{d=1}^{c-1} \exp(2\pi i (nd + md^{-1})/c).$$

This model belongs to the so-called arithmetical models corresponding to the motion on constant negative curvature surfaces generated by arithmetic groups. For all these models due to the exponential multiplicities of periodic orbits one expects [18] that the spectral statistics will tend to the Poisson distribution though from a classical point of view all these models are the best known examples of classically chaotic motion. Using the above expression one can prove this statement for the modular group.

7 Construction of the density of states from finite number of periodic orbits

The main difficulty in using trace formulae is their divergent character. The diagonal approximation consists, in some sense, on computing the density of states from a sum over a finite number of periodic orbits but this sum cannot produce δ-function singularities. There exists an artificial method [19] which permits to avoid this difficulty. Its main ingredient is the Riemann-Siegel form of the zeta function

$$\zeta(1/2 - iE) = z_T(E) + e^{2\pi i \bar{N}(E)} z_T^*(E), \tag{19}$$

where instead of the correct Riemann-Siegel expansion one uses a truncated product over periodic orbits

$$z_T(E) = \prod_{\log p<T} (1 - p^{-1/2+iE})^{-1}.$$

The density of zeros for function (19) takes the form

$$D_T(E) = d_T(E) \sum_{k=-\infty}^{\infty} (-1)^k e^{2\pi i k \bar{N}(E)} \left(\frac{z_T^*(E)}{z_T(E)} \right)^k, \tag{20}$$

where $d_T(E)$ is the density of state truncated at $\log p < T$.

Assuming that T is of the order of the Heisenberg time, $T_H = 2\pi\bar{d}$, and $\bar{d} \to \infty$ after some algebra we get

$$R_2^{(off)}(\epsilon_1, \epsilon_2) = \bar{d}^2 e^{2\pi i \bar{d}\epsilon} < \frac{z_T^*(E+\epsilon_1) z_T(E+\epsilon_2)}{z_T(E+\epsilon_1) z_T^*(E+\epsilon_2)} > + c.c. \tag{21}$$

The last step consists in performing the energy average of this expression. As logarithms of primes are not commensurable, the energy average of any smooth function of $\exp(iE \log p_j)$ equals its phase average

$$< f >= \int_0^{2\pi} \cdots \int_0^{2\pi} f(e^{i\phi_1}, \ldots, e^{i\phi_M}) \prod_{j=1}^{M} \frac{d\phi_j}{2\pi}. \tag{22}$$

This is essentially equivalent to the random phase approximation, or to the ergodic theorem for quasi-periodic functions with non-commensurable periods, or to the strict diagonal approximation.

For the Riemann zeta function the total contribution equals

$$R_2(\epsilon) = C^2 \exp(2\pi i \bar{d}\epsilon) |\zeta(1 + i\epsilon)|^2 \Phi^{(off)}(\epsilon) + c.c. \ , \tag{23}$$

where

$$\Phi^{(off)}(\epsilon) = \prod_p (1 - \frac{(1 - p^{i\epsilon})^2}{(p-1)^2}), \tag{24}$$

$\epsilon = \epsilon_2 - \epsilon_1$, and $C = \bar{d} \prod_p (1 - 1/p)$. All products in these expressions include prime numbers up to $\ln p = T$. The first two products converge when $T \to \infty$ and only the last one requires a regularization. But our parameter T has not yet been fixed. Let us choose it in such a way that

$$2\pi\bar{d} \prod_{\ln p<T} (1 - \frac{1}{p}) = 1. \tag{25}$$

The same factor appears in the statistical approach to prime numbers (see discussion in [15]) and can be considered as a renormalisation of formally divergent sums. After this renormalization we get exactly the same formula (14) as has been derived in the previous section using the Hardy-Littlewood conjecture about the pairwise distribution of prime numbers.

8 Off-diagonal terms for dynamical systems

For 2-dimensional dynamical systems the only difference with the Riemann case is that the truncated zeta function $z_T(E)$ contains now an infinite product over m

$$z_T(E) = \prod_{T_p<T} \prod_{m=0}^{\infty} (1 - \frac{e^{iS_p(E)/\hbar - i\pi\mu_p/2}}{|\Lambda_p|^{1/2}\Lambda_p^m}), \tag{26}$$

where Λ_p is the largest eigenvalue of the monodromy matrix.

The simplest and most natural assumption is that in generic systems without time-reversal invariance periodic orbits up to period T are linearly non-commensurable (as primes). Under this conjecture after some algebra we obtain that when $T \to \infty$

$$R_2(\epsilon) = \frac{e^{2\pi i \bar{d}\epsilon}}{4\pi^2} |\gamma^{-1} Z_{cl}(i\epsilon)|^2 \prod_p < R_p > \left| \frac{Z_p(i\epsilon)}{Z_p(0)} \right|^2 + c.c. \tag{27}$$

and

$$< R_p >= \sum_{n=0}^{\infty} \frac{(a;q)_n^2}{(q;q)_n^2} y^n, \tag{28}$$

where $(a;q) = (1-a)(1-aq)\dots(1-aq^{n-1})$, $a = e^{-i\tau_p}$, $q = \Lambda_p^{-1}$, $y = |\Lambda_p|^{-1} e^{i\tau_p}$, and $\tau_p = l_p\epsilon/k$. $Z_{cl}(s)$ is a classical zeta function, $Z_{cl}(s) = \prod_p Z_p(s)^{-1}$ with $Z_p(s) = 1 - e^{\tau_p s}/|\Lambda_p|$.

The maximum period T is determined from the condition

$$2\pi\bar{d} \prod_{T_p<T} Z_p(0) = \frac{1}{|\gamma|}, \tag{29}$$

where γ is the residue of $Z_{cl}(s)$ at $s = 0$ ($Z_{cl}(s) \to \gamma/s$ when $s \to 0$). As above this renormalization fixes T to be of the order of T_H and ensures that, when $\epsilon \to 0$, $R_2(\epsilon)$ tends to the GUE result.

9 Random matrix universality

There exists another method of semiclassical calculation of off-diagonal part of correlation functions which demonstrates that if such formulae exist they coincide with the above obtained expressions.

According to the naive trace formula the density of states is

$$d(E) = \tilde{d}(E) + \eta(E), \tag{30}$$

where $\tilde{d}(E)$ is the truncated density of states computed from a set of short-period orbits with period $T_p < T$ (now we shall assume that $T \ll T_H$) and $\eta(E)$ is (unknown) part of the density constructed from high-period orbits.

Let us try now to construct a random matrix ensemble which has the mean density of eigenvalues exactly equals $\tilde{d}(E)$. In principle, the necessary potential can be computed from the Dyson equation

$$\int \frac{\tilde{d}(t)}{x-t}dt = \frac{1}{2}V'(x). \tag{31}$$

But the explicit form of this potential is irrelevant as under quite general conditions the resulting distribution does not depend on the explicit form of this function (provided it corresponds to the so-called definite momentum problem [21]) and all correlation functions depend only on the kernel $K_N(x,y)$ which in the bulk of the spectrum in the limit $N \to \infty$ tends to

$$K(x,y) = \frac{\sin \pi(N(x) - N(y))}{\pi(x-y)}, \tag{32}$$

where $N(x) = \int^x \tilde{d}(x')dx'$ is the mean staircase function.

Hence, the two-point correlation function will take the form

$$R_2(\epsilon_1,\epsilon_2) = < \tilde{d}(E+\epsilon_1)\tilde{d}(E+\epsilon_2) - \frac{\sin^2 \pi(\tilde{N}(E+\epsilon_1) - \tilde{N}(E+\epsilon_2))}{\pi^2(\epsilon_1-\epsilon_2)^2} > . \tag{33}$$

As $\tilde{d}(E)$ is known it is possible to perform the smoothing over the appropriate energy window. Using the same transformations as above one can show that under the assumption $T \ll T_H$ the dependence of T will disappear and one gets the same formulae as above.

10 Conclusion

The heuristic arguments presented in this paper demonstrate how, in principle, the existence of the trace formula and certain natural conjectures about the distribution of periodic orbits (or primes) combine together to produce universal local statistics. In particular, for systems without the time-reversal invariance the assumption that low-period orbits are non-commensurable leads to the GUE statistics (at least for 2-point correlation function). The close relation between diagonal (9) and off-diagonal (14) terms (first observed for disordered systems in [22]) suggests the existence of a certain unified principle. The best candidate for it is the 'unitarity' property of the trace formula, namely, that the distribution of periodic orbits should be such that the corresponding eigenvalues will be real. In some sense certain long-period orbits are connected to the short ones and the investigation of this connection may clarify the origin of universal spectral statistics. The interesting question is what conjectures about periodic orbits are necessary to obtain correlation functions for systems with time-reversal invariance where almost all periodic orbits appear in pairs with exactly the same action.

References

[1] C.E. Porter, *Statistical Theories of Spectra: Fluctuations*, (Academic Press, New York, 1965).

[2] E.P. Wigner, Random Matrices in Physics, SIAM Review, 9 (1967) 1-23.

[3] O. Bohigas, Random matrix theories and chaotic dynamics, in *Chaos and Quantum Physics*, M.-J. Giannoni, A. Voros, and J. Zinn-Justin, eds. (Elsevier, Amsterdam, 1991), 87-199.

[4] T.Guhr, A. Müller-Groeling, and H. Weidenmüller, Random Matrix Theories in Quantum Physics: Common Concepts, Phys. Rep. 299 (1998) 189-425.

[5] M.L. Mehta, *Random Matrices*, (Academic Press, New York, 1991).

[6] O. Bohigas, M.-J. Giannoni, and C. Schmit, Characterization of Chaotic Quantum Spectra and Universality of Level Fluctuation Laws, Phys. Rev. Lett. 52 (1984) 1-4.

[7] M.V. Berry and M. Tabor, Level clustering in the regular spectrum, Proc. R. Soc. London A 356 (1977) 375-394.

[8] H.L. Montgomery, The pair correlation of zeros of the zeta function, in *Analytical Number Theory*, Proc. Symp. Pure Math. 24, Amer. Math. Soc. (1973) 181-193.

[9] A.M. Odlyzko, The 10^{20}th Zero of the Riemann Zeta function and 70 Million of its Neighbors, ATT Bell Laboratories Preprint (1989).

[10] Z. Rudnik and P. Sarnak, Zeroes of principal L-functions and random matrix theory, Duke Math. J. 81 (1996) 269-322.

[11] N.M. Katz and P. Sarnak, *Random Matrices, Frobenius Eigenvalues, and Monodromy*, (1997) to be published.

[12] M.C. Gutzwiller, Periodic Orbits and Clasical Quantization Conditions, J. Math. Phys. 12 (1971) 343-358.

[13] M.V. Berry, Semi-classical Theory of Spectral Rigidity, Proc. R. Soc. London, A 400 (1985) 229-251.

[14] J.H. Hannay and A.M. Ozorio de Almeida, Periodic orbits and a correlation function for the semiclassical density of states, J. Math. Phys. A17 (1984) 3429-3440.

[15] G.H. Hardy and J.E. Littlewood, Some problems of 'Partitio Numerorum'; III: On the expression of a number as a sum of primes, Acta Mathematica 44 (1922) 1-70.

[16] E. Bogomolny and J. Keating, Random matrix theory and the Riemann zeros I: three- and four-point correlations, Nonlinearity 8 (1995) 1115-1131; *ibid*, II: n-point correlations, Nonlinearity 9 (1996) 911-935.

[17] E. Bogomolny, F. Leyvraz, and C. Schmit, Distribution of Eigenvalues for the Modular Group, Commun. Math. Phys. 176 (1996) 577-617.

[18] E. Bogomolny, B. Georgeot, M.-J. Giannoni, and C. Schmit, Chaotic Billiards Generated by Arithmetic Groups, Phys. Rev. Lett. 69 (1992) 1477-1480; *ibid*, Arithmetical Chaos, Phys. Rep. 291 (1997) 219-324.

[19] E. Bogomolny and J. Keating, Gutzwiller's Trace Formula and Spectral Statistics: Beyond the Diagonal Approximation, Phys.Rev. Lett. 77 (1996) 1472-1475.

[20] E. Bogomolny and C. Schmit, Semiclassical Computations of Energy Levels, Nonlinearity 6 (1993) 523-547.

[21] E. Bogomolny, O. Bohigas, and M. Pato, On the distribution of eigenvalues of certain matrix ensembles, Phys. Rev. E 85 (1996) 639-779.

[22] A.V. Andreev and P.L. Altshuler, Spectral Statistics beyond Random Matrix Theory, Phys. Rev. Lett. 75 (1995) 902-905.

Eugene Bogomolny
DPT, CNRS-URA 62,
Institut de Physique Nucléaire,
91406 Orsay Cedex, France,
bogomol@ipno.in2p3.fr

Scaling Algebras in Local Relativistic Quantum Physics

Detlev Buchholz

Abstract. The novel method of scaling algebras allows one to compute and classify the short distance (scaling) limit of any local relativistic C^*–dynamical system and to determine its symmetry structure. The approach is based on an adaptation of ideas from renormalization group theory to the C^*–algebraic setting.

Local relativistic quantum physics [1] in a pseudo-Riemannian spacetime manifold $(\mathcal{M}, g)$ can conveniently be described by C^*–dynamical systems $(\mathcal{A}, \alpha)$, where $\mathcal{A}$ is a C^*–algebra, describing the physical observables in $\mathcal{M}$, and α a representation of the isometry group of $(\mathcal{M}, g)$ by automorphisms of $\mathcal{A}$. The principle of Einstein causality is implemented in this setting by specifying a net (pre-cosheaf) of subalgebras of $\mathcal{A}$ which are labelled by the open, relatively compact regions $\mathcal{O} \subset \mathcal{M}$,

$$\mathcal{O} \longmapsto \mathcal{A}(\mathcal{O}),$$

such that algebras corresponding to causally disjoint regions commute with each other. A theory is fixed by specifying a dynamical system which is subject to these constraints.

In high energy physics the structure of the observables in very small spacetime regions $\mathcal{O}$ (at "small spacetime scales") is of great interest. It can be explored with the help of scaling algebras which have been introduced in [2] by adopting ideas from the theory of the renormalization group. We outline this method for the case where $(\mathcal{M}, g)$ is d–dimensional Minkowski space $\mathbb{R}^{\,\mathrm{d}}$, equipped with its standard Lorentzian metric. The corresponding isometry group is the Poincaré group $\mathcal{P}_+^\uparrow$ whose elements (Λ, x) are composed of Lorentz transformations and spacetime translations.

For the analysis of the short distance properties of a theory one first proceeds from the given net and automorphisms $(\mathcal{A}, \alpha)$ at spacetime scale $\lambda = 1$ (in appropriate units) to the corresponding nets $(\mathcal{A}^{(\lambda)}, \alpha^{(\lambda)})$ describing the theory at arbitrary scale $\lambda \in \mathbb{R}_+$. This is accomplished by setting for given λ

$$\mathcal{O} \longmapsto \mathcal{A}^{(\lambda)}(\mathcal{O}) \doteq \mathcal{A}(\lambda\mathcal{O}), \quad \alpha^{(\lambda)}_{\Lambda,x} \doteq \alpha_{\Lambda,\lambda x}.$$

The latter nets are in general not isomorphic to each other for different values of λ, so they are to be regarded as different theories.

In addition one needs a way of comparing observables at different scales. To this end one considers certain specific functions $\underline{A}$ of the scaling parameter whose values $\underline{A}_\lambda$ are, for given λ, regarded as observables in the theory $(\mathcal{A}^{(\lambda)}, \alpha^{(\lambda)})$. With this idea in mind one is led to the following construction.

Consider the C^*–algebra $\mathrm{L}^\infty(\mathcal{A})$ of functions $\underline{A} : \mathbb{R}_+ \longmapsto \mathcal{A}$ for which the algebraic operations are pointwise defined and which have finite norm $||\underline{A}|| = \sup_\lambda ||\underline{A}_\lambda||$. The Poincaré group $\mathcal{P}_+^\uparrow$ acts on $\mathrm{L}^\infty(\mathcal{A})$ by automorphisms $\underline{\alpha}_{\Lambda,x}$ which are given by

$$\big(\underline{\alpha}_{\Lambda,x}(\underline{A})\big)_\lambda \doteq \alpha^{(\lambda)}_{\Lambda,x}(\underline{A}_\lambda).$$

We restrict attention to the subalgebra of $\mathrm{L}^\infty(\mathcal{A})$ on which these automorphisms act strongly continuously. Moreover, we introduce a local net structure on this subalgebra by setting

$$\mathcal{O} \longmapsto \underline{\mathcal{A}}(\mathcal{O}) \doteq \{\underline{A} : \underline{A}_\lambda \in \mathcal{A}^{(\lambda)}(\mathcal{O}),\ \lambda \in \mathbb{R}_+\}.$$

The *scaling algebra* $\underline{\mathcal{A}}$ is then defined as the inductive limit of the local algebras $\underline{\mathcal{A}}(\mathcal{O})$. It is easily checked that $(\underline{\mathcal{A}}, \underline{\alpha})$ is again a local C^*–dynamical system which is completely fixed by the given net.

The physical states in the underlying theory are described by a folium of positive linear and nomalized functionals $\omega \in \mathcal{A}^*$ which are locally normal with respect to each other [1]. Their structure at small spacetime scales can be analyzed with the help of the scaling algebra as follows. Given ω, one defines its lift to the scaling algebra at scale $\lambda \in \mathbb{R}_+$ by setting

$$\underline{\omega}_\lambda(\underline{A}) \doteq \omega(\underline{A}_\lambda), \quad \underline{A} \in \underline{\mathcal{A}}.$$

If $\underline{\pi}_\lambda$ denotes the GNS–representation of $\underline{\mathcal{A}}$ induced by $\underline{\omega}_\lambda$ one considers the net

$$\mathcal{O} \longmapsto \underline{\mathcal{A}}(\mathcal{O})/\ker \underline{\pi}_\lambda, \quad \underline{\alpha}^{(\lambda)},$$

where ker means "kernel" and $\underline{\alpha}^{(\lambda)}$ is the induced action of the Poincaré transformations $\underline{\alpha}$ on this quotient. It is important to notice that this net is isomorphic to the underlying theory $(\mathcal{A}^{(\lambda)}, \alpha^{(\lambda)})$ at scale λ. This insight leads to the following canonical definition of the scaling limit of the theory: One first considers the limit(s) of the net of states $\{\underline{\omega}_\lambda\}_{\lambda \searrow 0}$. By standard compactness arguments, this net has always a non–empty set $\{\underline{\omega}_0\}$ of limit points. The following facts about these limit states have been established in [2]:

PROPOSITION 1. The set $\{\underline{\omega}_0\}$ does not depend on the chosen physical state ω.

PROPOSITION 2. Each $\underline{\omega}_0$ is a vacuum state on $(\underline{\mathcal{A}}, \underline{\alpha})$, i.e. a ground state with respect to the time evolution which is invariant under Poincaré transformations. Moreover, in $\mathrm{d} > 2$ dimensional Minkowski space theories these vacua are pure states.

Denoting the GNS–representation corresponding to given $\underline{\omega}_0$ by $(\underline{\pi}_0, \underline{\mathcal{H}}_0)$ one then defines in complete analogy to the case $\lambda > 0$ the net

$$\mathcal{O} \longmapsto \mathcal{A}^{(0)} \doteq \underline{\mathcal{A}}(\mathcal{O})/\ker \underline{\pi}_0, \quad \alpha^{(0)} \doteq \underline{\alpha}^{(0)}$$

which is to be interpreted as *scaling limit of the underlying theory.* The various steps in this construction can be summarized in the diagram

$$(\mathcal{A}, \alpha) \longrightarrow (\mathcal{A}^{(\lambda)}, \alpha^{(\lambda)}) \longrightarrow (\underline{\mathcal{A}}, \underline{\alpha}) \longrightarrow \{(\mathcal{A}^{(0)}, \alpha^{(0)})\}.$$

It is now possible to classify the scaling limits as follows [2].

Classification: Let $(\mathcal{A}, \alpha)$ be a net with properties specified above. There are the following mutually exclusive possibilities for the structure of the scaling limit theory induced by the corresponding scaling limit states $\{\underline{\omega}_0\}$.

1. The nets $(\mathcal{A}^{(0)}, \alpha^{(0)})$ are all isomorphic to the trivial net $(\mathbb{C}\cdot 1, \mathrm{id})$ (classical scaling limit)
2. The nets $(\mathcal{A}^{(0)}, \alpha^{(0)})$ are all isomorphic and the algebras $\mathcal{A}^{(0)}$ are non–abelian (quantum scaling limit)
3. Not all of the nets $(\mathcal{A}^{(0)}, \alpha^{(0)})$ are isomorphic (degenerate scaling limit)

Theories with a quantum scaling limit are of primary physical interest. For this class there holds the following statement on the enhancement of symmetries at small scales [2].

Proposition 3. The scaling limit nets $(\mathcal{A}^{(0)}, \alpha^{(0)})$ of theories with a quantum scaling limit admit an automorphic action of the scaling transformations $\mathbb{R}_+$.

Simple examples in this class are nets generated by non–interacting quantum fields in $\mathrm{d} = 3$ and 4 dimensions [3]. For a discussion of the other cases see [4].

The fact that the scaling limit theories $(\mathcal{A}^{(0)}, \alpha^{(0)})$ exhibit all features of local nets of observable algebras allows one to apply standard methods for their analysis and physical interpretation. For the determination of the symmetries appearing in the scaling limit one can rely in the case of $\mathrm{d} > 2$ dimensional Minkowski space on the Doplicher–Roberts reconstruction theorem [5]. The necessary prerequisite for its application is the following result established in [2].

Proposition 4. If a local net complies with the special condition of duality (modular covariance) of Bisognano and Wichmann, the same holds true for its scaling limit.

By the results of Doplicher and Roberts [5] one can then recover from the outer local endomorphisms of the scaling limit net $(\mathcal{A}^{(0)}, \alpha^{(0)})$ a compact group $G^{(0)}$ whose irreducible representations are in one–to–one correspondence to the set of physical states which appear in the scaling limit and carry a localizable charge. Moreover, there exists an extension of the scaling limit net to a field net $(\mathcal{F}^{(0)}, \alpha^{(0)})$ on which $G^{(0)}$ acts by automorphisms and which implements the action of the local endomorphisms. The vacuum representation of this field net describes the charged

physical states and can be used to analyse their detailed properties. In particular one can determine from it the particle content of the scaling limit theory, which corresponds to the set of non–trivial irreducible representations of the Poincaré group $\mathcal{P}_+^\uparrow$ appearing in the vacuum sector of $(\mathcal{F}^{(0)}, \alpha^{(0)})$.

Of special interest is the comparison of the particle and symmetry content of the underlying theory and of its scaling limit [6]. Depending on the theory, there may be particles at finite scales which disappear in the scaling limit, particles which survive in this limit and particles which only come into existence at very small scales. These possibilities correspond exactly to the features of the various particle like structures which are observed in high energy collision experiments. Intuitive physical notions, such as quark, gluon, colour symmetry and confinement thus acquire an unambiguous mathematical meaning in the present setting [6].

The extension of the short distance analysis to local nets on spacetimes $(\mathcal{M}, g)$ with a large isometry group, such as de Sitter space, is straightforward. For theories on spacetimes with small isometry groups it is however less clear how to define corresponding scaling algebras which consist of sufficiently regular elements. An interesting proposal to solve this problem has been made by Verch [7]. In this approach the resulting scaling limits turn out to be local C^*-dynamical systems in the (Minkowskian) fibers of the tangent bundle of $(\mathcal{M}, g)$. For a further classification of these theories it would be of interest to analyze the transport between the corresponding dynamical systems in the various fibers which is induced by the underlying dynamics. This "quantum connection" should also contain relevant information on the presence of local gauge symmetries in the underlying theory.

REFERENCES

[1] R. Haag: *Local Quantum Physics*, Springer 1992

[2] D. Buchholz, R. Verch: Rev. Math. Phys. 8 (1995) 1195–1240

[3] D. Buchholz, R. Verch: hep-th 9708095 (to appear in Rev. Math. Phys.)

[4] D. Buchholz: Ann. I. H. Poincaré 64 (1996) 433–460

[5] S. Doplicher, J.E. Roberts: Commun. Math. Phys. 131 (1990) 51–107

[6] D. Buchholz: Nucl. Phys. B469 (1996) 333–356

[7] R. Verch: pp. 564–577 in *Operator Algebras and Quantum Field Theory*, International Press 1997

Detlev Buchholz
Institut für Theoretische Physik
Universität Göttingen
Bunsenstraße 9
D-37073 Göttingen, Germany

Finite-Size Scaling in Percolation

J. T. Chayes

Abstract. This work is a detailed study of the phase transition in percolation, in particular of the question of finite-size scaling: Namely, how does the critical transition behavior emerge from the behavior of large, finite systems? Our results rigorously locate the proper window in which to do critical computation and establish features of the phase transition. This work is a finite-dimensional analogue of classic work on the critical regime of the random graph model of Erdös and Rényi.

1991 Mathematics Subject Classification: 82B43, 82B26, 60K35, 05C80
Keywords and Phrases: percolation, phase transitions, finite-size scaling

1. Introduction

This paper gives an overview and discussion of some recent results of Borgs, Chayes, Kesten and Spencer [BCKS2] on finite-size scaling and incipient infinite clusters in percolation.

We consider bond percolation in a finite subset Λ of the hypercubic lattice $\mathbb{Z}^d$. Nearest-neighbor bonds in Λ are occupied with probability p and vacant with probability $1-p$, independently of each other. Let p_c denote the bond percolation threshold in $\mathbb{Z}^d$, namely the value of p above which there exists an infinite connected cluster of occupied bonds. As a function of the size of the box Λ, we determine the scaling window about p_c in which the system behaves critically. For our purposes, criticality is characterized by the behavior of the distribution of sizes of the largest clusters in the box. We show how these clusters can be identified with the so-called incipient infinite cluster—the cluster of infinite expected size which appears at p_c. It turns out that these results can be established axiomatically from hypotheses which are mathematical expressions of the purported scaling behavior in critical percolation. Moreover, these hypotheses can be explicitly verified in two dimensions. In this brief overview, I will omit all details of the proofs of the [BCKS2] results, focusing instead on the motivation, the hypotheses and a few of the implications of these results. The reader is referred to [BCKS1] and [BCKS2] for more details and for related results which are not included here. Some of the discussion here closely parallels that of [CPS].

2. The Motivation

The motivation for the [BCKS2] work was threefold.

The Random Graph Model

The original motivation for this work was to obtain an analogue of known results on the so-called random graph model of Erdös and Rényi ([ER1], [ER2]; see also [B2]). The random graph model is simply the percolation model on the complete graph, i.e., it is a model on a graph of N sites in which each site is connected to each other site independently with uniform probability $p(N)$. Physicists would call this a mean-field percolation model. It turns out that the model has particularly interesting behavior if $p(N)$ scales like $p(N) \approx c/N$ with $c = \Theta(1)$. Here, as usual, $f = \Theta(N^\alpha)$ means that there are nonzero, finite constants c_1 and c_2, of equal sign, such that $c_1 N^\alpha \leq f \leq c_2 N^\alpha$.

Let $W^{(i)}$ denote the random variable representing the size of the i^{th} largest cluster in the system. Erdös and Renyi showed that the model has a *phase transition* at $c = 1$ characterized by the behavior of $W^{(1)}$. It turns out that, with probability one,

$$W^{(1)} = \begin{cases} \Theta(\log N) & \text{if} \quad c < 1 \\ \Theta(N^{2/3}) & \text{if} \quad c = 1 \\ \Theta(N) & \text{if} \quad c > 1. \end{cases}$$

Moreover, for $c > 1$, $W^{(1)}/N \to \theta(c) > 0$, while for $c = 1$, $W^{(1)}$ has a nontrivial distribution (i.e., $W^{(1)}/N^{2/3} \nrightarrow$ constant), again with probability one. The smaller clusters have the same behavior as the largest for $c \leq 1$, but different behavior for $c > 1$: For $i > 1$, $W^{(i)} = \Theta(\log N)$ for all $c \neq 1$, while at $c = 1$, $W^{(i)} = \Theta(N^{2/3})$. The $\Theta(N)$ cluster for $c > 1$ is clearly the analogue of the infinite cluster in percolation on finite-dimensional graphs; here it is called the *giant component*. As we will see, the $\Theta(\log N)$ clusters are analogues of finite clusters in ordinary percolation. The $\Theta(N^{2/3})$ clusters will turn out to be the analogues of the so-called *incipient infinite cluster* in percolation. The work on the regime $c \neq 1$ appeared already in the original papers of Erdös and Rényi ([ER1], [ER2]); the correct behavior for $c = 1$ was derived many years later by Bollobás [B1].

In the past decade, there has been a great deal of work and remarkable progress on the random graph model. Much of this work culminated in the combinatoric tour de force of Janson, Knuth, Luczak and Pittel [JKLP]. Using remarkably detailed calculations, it was shown that shown that the correct parameterization of the critical regime is

$$p(N) = \frac{1}{N} + \frac{\lambda_N}{N^{4/3}},$$

in the sense that if $\lim_{N\to\infty} |\lambda_N| < \infty$, then $W^{(i)} = \Theta(N^{2/3})$ for all i, and furthermore each $W^{(i)}$ has a nontrivial distribution (which was actually calculated in [JKLP]). On the other hand, if $\lim_{N\to\infty} \lambda_N = -\infty$, then $W^{(2)}/W^{(1)} \to 1$ with probability one, whereas if $\lim_{N\to\infty} \lambda_N = +\infty$, then $W^{(2)}/W^{(1)} \to 0$ and $W^{(1)}/N^{2/3} \to +\infty$ with probability one. The largest component in the regime with $\lambda_N \to +\infty$ is called the *dominant component*. As we will see, it has an analogue in ordinary percolation.

The initial motivation for the [BCKS2] work was to find a finite-dimensional analogue of the above results. To this end, we considered d-dimensional percolation in a box of linear size n, and hence volume $N = n^d$. We asked how the size of the largest cluster in the box behaves as a function of n for $p < p_c$, $p = p_c$ and $p > p_c$. Also, we asked whether there is a window $p(n)$ about p_c such that the system has a nontrivial cluster size distribution within the window.

Finite-Size Scaling

The considerations of the previous paragraph lead us immediately to the question of *finite-size scaling* (FSS). Phase transitions cannot occur in finite volumes, since all relevant functions are polynomials and thus analytic; nonanalyticities only emerge in the infinite-volume limit. What quantities should we study to see the phase transition emerge as we go to larger and larger volumes?

Before the [BCKS2] work, this question had been addressed rigorously only in systems with first-order transitions—transitions at which the correlation length and order parameter are discontinuous ([BK], [BI]). Finite-size scaling at second-order transitions is more subtle due to the fact that the order parameter vanishes at the critical point. For example, in percolation it is believed that the infinite cluster density vanishes at p_c. However, physicists routinely talk about an incipient infinite cluster at p_c. This brings us to our third motivation.

The Incipient Infinite Cluster

At p_c, there is no infinite cluster with probability one, but the expected size of the cluster of the origin is infinite. Physicists call this finite object of infinite expected size, the *incipient infinite cluster* (IIC).

In the mid-1980's there were two attempts to construct rigorously an object that could be identified as an incipient infinite cluster. Kesten [K] proposed to look at the conditional measure in which the origin is connected to the boundary of a box centered at the origin, by a path of occupied bonds: $P_p^n(\cdot) = P_p(\cdot \mid 0 \leftrightarrow \partial[-n,n]^d)$. Here, as usual, $P_p(\cdot)$ is product measure at bond density p. Observe that, at $p = p_c$, as $n \to \infty$, $P_p^n(\cdot)$ becomes mutually singular with respect to the unconditioned measure $P_p(\cdot)$. Nevertheless, Kesten found that

$$\lim_{n\to\infty} P_{p_c}^n(\cdot) = \lim_{p \searrow p_c} P_p(\cdot \mid 0 \leftrightarrow \infty).$$

Moreover, Kesten studied properties of the infinite object so constructed and found that it has a nontrivial fractal dimension which agrees with the fractal dimension of the physicists' incipient infinite cluster.

Another proposal was made by Chayes, Chayes and Durrett [CCD]. They modified the standard measure in a different manner than Kesten, replacing the uniform p by an inhomogeneous $p(b)$ which varies with the distance of the bond b from the origin:

$$p(b) = p_c + \frac{c}{1 + \mathrm{dist}(0,b)^\zeta}.$$

The idea was to enhance the density just enough to obtain a nontrivial infinite object. [CCD] found that when $\zeta = 1/\nu$, where ν is the so-called correlation length

exponent, the measure $P_{p(x)}$ has some properties reminiscent of the physicists' incipient infinite cluster.

In the work to be discussed here, [BCKS2] propose yet a third rigorous incipient cluster—namely the largest cluster in a box. This is, in fact, exactly the definition that numerical physicists use in simulations. Moreover, it will turn out to be closely related to the IICs constructed by Kesten and Chayes, Chayes and Durrett. Like the IIC of [K], the largest cluster in a box will have a fractal dimension which agrees with that of the physicists' IIC. Also, the [BCKS2] proofs rely heavily on technical estimates from the IIC construction of [K]. More interestingly, the form of the scaling window $p(n)$ for the [BCKS2] problem will turn out to be precisely the form of the enhanced density used to construct the IIC of [CCD].

3. Definitions and Preliminaries

We briefly review some standard definitions and notation for percolation on $\mathbb{Z}^d$ (see e.g., [CPS]). Let $C(x)$ denote the occupied cluster of the site $x \in \mathbb{Z}^d$, and let $|C(x)|$ denote its size. The order parameter is the infinite cluster density

$$P_\infty(p) = P_p(|C(0)| = \infty),$$

and the standard susceptibility is the expected finite cluster size

$$\chi^{\text{fin}}(p) = E_p(|C(0)|, |C(0)| \neq \infty).$$

Here, as usual, E_p denotes expectation with respect to P_p. The finite cluster point-to-point connectivity function is

$$\tau^{\text{fin}}(x, y; p) = P_p(C(x) = C(y), |C(x)| < \infty),$$

The exponential rate of decay of this connectivity defines the correlation length $\xi(p)$:

$$1/\xi(p) = -\lim_{|x|\to\infty} \frac{1}{|x|} \log \tau^{\text{fin}}(0, x; p)$$

where the limit is taken with x along a coordinate axis. Another point-to-point connectivity, which for $p > p_c$ behaves much like τ^{fin}, is

$$\tau^{\text{cov}}(x, y; p) = P_p(|C(x)| = \infty, |C(y)| = \infty) - P_\infty^2(p).$$

Notice that

$$\chi^{\text{fin}}(p) = \sum_x \tau^{\text{fin}}(0, x; p).$$

Similarly, we can define another susceptibility,

$$\chi^{\text{cov}}(p) = \sum_x \tau^{\text{cov}}(0, x; p).$$

Another connectivity function is the point-to-box connectivity function

$$\pi_n(p) = P_p(\exists x \in \partial[-n, n]^d \text{ s.t. } C(0) = C(x)).$$

We also introduce the quantity

$$s(n) = (2n)^d \, \pi_n(p_c) \, .$$

It will turn out that $s(n)$ represents the size of the largest critical clusters on scale n. Finally, the cluster size distribution is described by

$$P_{\geq s}(p) = P_p(|C(0)| \geq s).$$

We next recall the definitions of some of the standard power laws expected to characterize the scaling behavior of relevant quantities in percolation, noting that the existence of these power laws has not yet been rigorously established in low dimensions. We define $F(p) \approx |p - p_c|^\alpha$ to mean $\lim_{p \to p_c} \log F(p)/\log|p - p_c| = \alpha$, and implicitly assume that the approach is identical from above and below threshold, unless noted otherwise. Similarly, we use the notation $F(n) \approx n^\alpha$ to mean $\lim_{n\to\infty} \log G(n)/\log n = \alpha$. The power laws of relevance to us are

$$P_\infty(p) \approx |p - p_c|^\beta, \quad p > p_c \, ,$$

$$\chi^{\text{fin}}(p) \approx |p - p_c|^{-\gamma} \, ,$$

$$\xi(p) \approx |p - p_c|^{-\nu} \, ,$$

$$P_{\geq s}(p_c) \approx s^{-1/\delta}$$

and

$$\pi_n(p_c) \approx n^{-1/\rho} \, .$$

Note that the last relation implies

$$s(n) \approx n^{d_f} \qquad \text{with} \qquad d_f = d - 1/\rho \, .$$

Here we use the notation d_f to indicate that the power law of $s(n)$ characterizes the fractal dimension of the incipient infinite cluster.

For rigorous work, it is often convenient to replace the correlation length by the finite-size scaling correlation length, $L_0(p)$, introduced in [CCF]. Define the rectangle crossing probability: $R_{L,M}(p) = P_p\{ \exists$ occupied bond crossing of $[0, L] \times [0, M] \cdots \times [0, M]$ in the 1-direction$\}$. Observing that, for $p < p_c$, $R_{L,3L}(p) \to 0$ as $L \to \infty$, we define

$$L_0(p) = L_0(p, \epsilon) = \min\{L \geq 1 \mid R_{L,3L}(p) \leq \epsilon\} \quad \text{if } p < p_c \, .$$

It can be shown [CCF] that the scaling behavior of $L_0(p, \epsilon)$ is essentially the same as that of the standard correlation length $\xi(p)$: for $0 < \epsilon < a(d)$, there exist constants $c_1 = c_1(d)$, $c_2 = c_2(d, \epsilon) < \infty$ such that

$$\frac{1}{L_0(p,\epsilon)} \;\leq\; \frac{1}{\xi(p)} \;\leq\; \frac{c_1 \log L_0(p,\epsilon) + c_2}{L_0(p,\epsilon) - 1}, \quad p < p_c \, .$$

Hereafter we will assume that $\epsilon < a(d)$; we usually suppress the ϵ-dependence in our notation. For $p > p_c$, [BCKS2] define $L_0(p, \epsilon)$ in terms of finite-cluster crossings in an annulus; the reader is referred to [BCKS2] for precise definitions and properties of the resulting length. Another important quantity in the high-density phase of percolation is the surface tension $\sigma(p)$; see [ACCFR] for the precise definition. By analogy with the definition of a finite-size scaling correlation length below threshold, [BCKS2] define a finite-size scaling inverse surface tension as

$$A_0(p) = A_0(p, \epsilon) = \min\{L^{d-1} \geq 1 \mid R_{L,3L}(p) \geq 1 - \epsilon\} \quad \text{if } p > p_c\,.$$

Again, see [BCKS2] for properties of $A_0(p)$.

4. The Scaling Axioms and the Results

The [BCKS2] results are established under a set of axioms which we can explicitly verify in two dimensions and which we expect to be true whenever the dimension does not exceed the upper critical dimension d_c (presumably $d_c = 6$). We call these axioms the *Scaling Axioms* since they are characterizations of the scaling behaviors implicitly assumed in the physics literature. In this section, we will review the axioms and a few of the results from [BCKS2]. Much of this treatment is taken almost verbatim from a preliminary version of [BCKS2] and [CPS].

The Scaling Axioms

Several of the axioms involve the length scales $L_0(p)$ and $A_0(p)$, and therefore implicitly involve the constant ϵ. [BCKS2] assume that the axioms are true for all $\epsilon < \epsilon_0$, where $\epsilon_0 = \epsilon_0(d)$ depends on a so-called rescaling lemma.

The axioms are written in terms of the equivalence symbol $\asymp$. Here $F(p) \asymp G(p)$ means that $C_1 F(p) \leq G(p) \leq C_2 F(p)$ where $C_1 > 0$ and $C_2 < \infty$ are constants which do not depend on n or p, as long as p is uniformly bounded away from zero or one, but which may depend on the constants ϵ, $\tilde{\epsilon}$ or x appearing explicitly or implicitly in the axioms. The [BCKS2] scaling axioms are

(I) $L_0(p) \to \infty$ as $p \downarrow p_c$;

(II) For $0 < \widetilde{\epsilon} < \epsilon_0$, $x \geq 1$ and $p > p_c$,
$A_0(p) \asymp L_0^{d-1}(p) \asymp L_0^{d-1}(p, \widetilde{\epsilon}; x)$;

(III) There are constants $D_1 > 0, D_2 < \infty$ such that
$D_1 \leq \pi_n(p)/\pi_n(p_c) \leq D_2$ if $n \leq L_0(p)$;

(IV) There are constants $D_3 > 0, \rho_1 > 2/d$, such that
$\pi_{kn}(p_c)/\pi_n(p_c) \geq D_3 k^{-1/\rho_1}, \quad n, k \geq 1$;

(V) There exists a constant D_4 such that for $p > p_c$,
$\chi^{\text{cov}}(p) \leq D_4 L_0^d(p) \pi^2_{L_0(p)}(p_c)$ and $\chi^{\text{fin}}(p) \leq D_4 L_0^d(p) \pi^2_{L_0(p)}(p_c)$;

(VI) For $p > p_c$,
$\pi_{L_0(p)}(p_c) \asymp P_\infty(p)$;

(VII) There exist constants $D_5, D_6 < \infty$ such that for $p < p_c$ and $k \geq 1$,
$P_{\geq ks(L_0(p))}(p) \geq D_5 e^{-D_6 k} P_{\geq s(L_0(p))}(p)$.

Let us briefly discuss the interpretation of the axioms. The first tells us that the approach to p_c is critical—i.e., continuous or second-order—from above p_c. Axiom (II) is the assumption of equivalence of length scales above p_c: The second part of it asserts the equivalence of the finite-size scaling lengths at various values of $x \geq 1$ and $\epsilon \in (0, \epsilon_0)$. The first part of it, i.e. $A_0(p) \asymp L_0^{d-1}(p)$, is called Widom scaling. It is equivalent to a hyperscaling relation the surface tension and correlation length exponents.

The third axiom formalizes a central element of the conventional scaling wisdom. Scaling theory asserts that whenever the system is viewed on length scales smaller than the correlation length, it behaves as it does at threshold. Axiom (III) asserts that this is the case for the connectivity function $\pi(p)$. Axiom (IV) implies that the connectivity function $\pi_n(p)$ has a bound of power law behavior at threshold. Of course, scaling theory assumes a pure power law with exponent $-1/\rho$. Axioms (V) and (VI) imply hyperscaling and scaling relations among the critical exponents. In terms of exponents, (V) is equivalent to the hyperscaling relation $d\nu = 2\beta + \gamma$, while (VI) is equivalent to the scaling relation $\nu/\rho = \beta$. Finally, Axiom (VII) gives a bound on the exponential decay rate of the cluster size distribution below p_c.

THEOREM 0 ([BCKS2]). *The Scaling Axioms (I)–(VII) hold in dimension $d = 2$.*

The proof of this theorem is technically quite complicated. It involves essentially the most complicated constructions which have been done for two-dimensional percolation.

A Few Results

In order to state the [BCKS2] results, we need to define a scaling window in which the system behaves critically, i.e. an analogue of the function $p(N)$ in the random graph problem. For us, this is described by the function

$$g(p,n) := \begin{cases} -\frac{n}{L_0(p)} & \text{if } p < p_c \\ 0 & \text{if } p = p_c \\ \frac{n}{L_0(p)} & \text{if } p > p_c. \end{cases}$$

It will turn out that a sequence of systems with density p_n behaves critically, subcritically, or supercritically— as far as size of large clusters is concerned— in finite boxes if, as $n \to \infty$, $g(p_n, n)$ remains bounded, tends to $-\infty$, or tends to ∞, respectively. If this is the case, we say that the sequence of systems is inside, below or above the scaling window, respectively.

We again use the symbol $\asymp$, this time for two sequences a_n and b_n of real numbers . We write $a_n \asymp b_n$ if $0 < \liminf_{n\to\infty} a_n/b_n \leq \limsup_{n\to\infty} a_n/b_n < \infty$.

Our first theorem characterizes the scaling window in terms of the *expectation* of the largest cluster sizes.

THEOREM 1 ([BCKS2]). *Suppose that Axioms (I)–(VII) hold.*

i) If $\{p_n\}$ is inside the scaling window, i.e., if $\limsup_{n\to\infty} |g(p_n, n)| < \infty$, and $i \in \mathbb{N}$, then

$$E_{p_n}\{W_{\Lambda_n}^{(i)}\} \asymp s(n)\,.$$

ii) If $\{p_n\}$ is below the scaling window, i.e., $g(p_n, n) \to -\infty$, then

$$E_{p_n}\{W^{(1)}_{\Lambda_n}\} \asymp s(L_0(p_n)) \log \frac{n}{L_0(p_n)} \,.$$

iii) If $\{p_n\}$ is above the scaling window, i.e., $g(p_n, n) \to \infty$, then

$$\frac{E_{p_n}\{W^{(1)}_{\Lambda_n}\}}{|\Lambda_n| P_\infty(p_n)} \to 1 \qquad as \qquad n \to \infty \,,$$

and

$$\frac{E_{p_n}\{W^{(2)}_{\Lambda_n}\}}{|\Lambda_n| P_\infty(p_n)} \to 0 \qquad as \qquad n \to \infty \,.$$

Assuming the existence of critical exponents and monotonicity of various quantities, Theorem 1 says that the scaling window is of the form

$$p_n = p_c \pm \frac{c}{n^{1/\nu}},$$

that inside the window

$$W^{(1)} \approx n^{d_f}, \quad W^{(2)} \approx n^{d_f}, \quad \cdots$$

while above the window

$$\begin{aligned} W^{(1)} &\approx n^d P_\infty \,, \\ W^{(1)}/n^{d_f} &\to \infty \,, \\ W^{(2)}/W^{(1)} &\to 0, \end{aligned}$$

and below the window

$$W^{(1)}/n^{d_f} \to 0$$

where, in fact,

$$W^{(1)} \approx \xi^{d_f} \log n/\xi \,.$$

The above results hold in expectation.

[BCKS2] also prove analogues of statements (i)–(iii) of the theorem for convergence in probability, rather than in expectation. Furthermore, within the scaling window, we get results on the distribution of cluster sizes which show that the distribution does not go to a delta function. This is to be contrasted with the behavior above the window, where the cluster size distribution approaches its expectation, with probability one. All of these additional results require some delicate second moment estimates. The reader is referred to [BCKS2] for precise statements of these results and for their proofs.

One final result is worth mentioning, since it is used in the proofs of the other results and is of interest in its own right. It concerns the number of clusters on scales $m < n$. Before stating the result, it should be noted that, due to statement

(i) of Theorem 1, the "incipient infinite cluster" inside the scaling window is not unique, in the sense that $W^{(2)}_{\Lambda_n}$ is of the same scale as $W^{(1)}_{\Lambda_n}$. This should be contrasted with the behavior of $W^{(2)}_{\Lambda_n}/W^{(1)}_{\Lambda_n}$ above the scaling window (see statement (iii)), a remnant of the uniqueness of the infinite cluster above p_c. The next theorem relates the non-uniqueness of the "incipient infinite cluster" inside the scaling window to the property of scale invariance at p_c. Basically, it says that the number of clusters of scale m in a system of scale n is a function only of the ratio n/m. How can this hold on *all* scales m? The only way it can be true is if the system has a fractal-like structure with smaller clusters inside holes in larger clusters. The theorem concerns the number $N_\Lambda(s_1, s_2)$ of clusters with size between s_1 and s_2.

THEOREM 2 ([BCKS2]). *Assume that Axioms (I)–(IV) are valid. Let $\{p_n\}$ lie inside the scaling window. Then there exist strictly positive, finite constants σ_1, σ_2, C_1 and C_2 (all depending on the sequence $\{p_n\}$) such that*

$$C_1\left(\frac{n}{m}\right)^d \le E_{p_n}\left\{N_{\Lambda_n}(s(m), s(km))\right\} \le C_2\left(\frac{n}{m}\right)^d,$$

provided m and k are strictly positive integers with $k \ge \sigma_1$ and $\sigma_2 m \le n$.

5. INTERPRETATION OF THE RESULTS

How can we understand the form of the window? As explained earlier, the system is expected to behave critically whenever the length scale is less than the correlation length. Indeed, this is the content of Axiom (III). But the boundary of this region is given by

$$n \approx \tilde{\lambda}\xi \approx \tilde{\lambda}|p - p_c|^{-\nu}, \qquad \text{i.e.} \quad p \approx p_c \pm \frac{\lambda}{n^{1/\nu}},$$

where $\tilde{\lambda}$, λ are constants. This is of course precisely the content of Theorem 1.

What would these results say if we attempted to apply them in the case of random graph model (to which they of course do not rigorously apply)? Let us use the hyperscaling relation $d\nu = \gamma + 2\beta$ and the observation that the volume N of our system is just n^d, to rewrite the window in the form

$$p_n = p_c \pm \frac{\lambda}{n^{1/\nu}} = p_c\left(1 \pm \frac{c}{n^{1/\nu}}\right) = p_c\left(1 \pm \frac{c}{N^{1/d\nu}}\right) = p_c\left(1 \pm \frac{c}{N^{1/(\gamma+2\beta)}}\right).$$

Similarly, let us use the hyperscaling relation $d_f/d = \delta/(1+\delta)$ to rewrite the size of the largest cluster as

$$W^{(1)} \approx n^{d_f} \approx N^{d_f/d} \approx N^{\delta/(1+\delta)}.$$

Noting that the random graph model is a mean-field model, we expect (and in fact it can be verified) that $\gamma = 1$, $\beta = 1$ and $\delta = 2$. Using also $p_c = 1/N$, we have a window of the form

$$p(N) = \frac{1}{N} \pm \frac{c}{N^{4/3}},$$

and within that window

$$W^{(1)} \approx N^{2/3},$$

just the values obtained in the combinatoric calculations on the random graph model.

The results also have implications for finite-size scaling. Indeed, the form of the window tells us precisely how to locate the critical point, i.e. it tells us the correct region about p_c in which to do critical calculations. Similarly, $W^{(1)} \approx N^{2/3}$ tells us how to extrapolate the scaling of clusters in the critical regime.

Finally, the results tell us that we may use the largest cluster in the box as a candidate for the incipient infinite cluster. Within the window, it is not unique, in the sense that there are many clusters of this scale. However, outside the window (even including a region where p is not uniformly greater than p_c as $n \to \infty$), there is a unique cluster of largest scale. This is the analogue of what is called the *dominant component* in the random graph problem.

References

[ACCFR] M. Aizenman, J. T. Chayes, L. Chayes, J. Fröhlich and L. Russo, *On a sharp transition from area law to perimeter law in a system of random surfaces*, Commun. Math. Phys. **92** (1983), 19–69.

[B1] B. Bollobás, *The evolution of random graphs*, Trans. Am. Math. Soc. **286** (1984), 257–274.

[B2] B. Bollobás, *Random Graphs*, Academic Press, London, 1985.

[BCKS1] C. Borgs, J. T. Chayes, H. Kesten and J. Spencer, *Uniform boundedness of critical crossing probabilities implies hyperscaling*, preprint.

[BCKS2] C. Borgs, J. T. Chayes, H. Kesten and J. Spencer, *The birth of the infinite cluster: finite-size scaling in percolation*, preprint.

[BI] C. Borgs and J. Imbrie, *Crossover finite-size scaling at first order transitions*, J. Stat. Phys. **69** (1992), 487–537.

[BK] C. Borgs and R. Kotecký, *A rigorous theory of finite-size scaling at first-order phase transitions*, J. Stat. Phys. **61** (1990), 79–110.

[CCD] J. T. Chayes, L. Chayes and R. Durrett, *Inhomogeneous percolation problems and incipient infinite clusters*, J. Phys. A: Math. Gen. **20** (1987), 1521–1530.

[CCF] J. T. Chayes, L. Chayes and J. Fröhlich, *The low-temerature behavior of disordered magnets*, Commun. Math. Phys. **100** (1985), 399–437.

[CPS] J. T. Chayes, A. Puha and T. Sweet, *Independent and dependent percolation*, Proceedings of the PCMI Summer School on Probability Theory, AMS, Providence, 1998.

[ER1] P. Erdös and A. Rényi, *On random graphs I*, Publ. Math. (Debrecen) **6** (1959), 290-297.

[ER2] P. Erdös and A. Rényi, *On the evolution of random graphs*, Magy. Tud. Akad. Mat. Kut. Intéz. Közl. **5** (1960), 17–61.

[JKLP] S. Janson, D. E. Knuth, T. Luczak and B. Pittel, *The birth of the giant component*, Random Struc. Alg. **4** (1993), 233–358.

[K] H. Kesten, *The incipient infinite cluster in two-dimensional percolation*, Probab. Th. Rel. Fields **73** (1986), 369–394.

Jennifer Tour Chayes
Microsoft Research
One Microsoft Way
Redmond, WA 98052
jchayes@microsoft.com

Extended Dynamical Systems

P. Collet

Abstract. We discuss the dynamics of dissipative systems defined in unbounded space domains, and in particular results on the existence of the semi-flow of evolution, on the development structures due to instabilities, and on large time behaviors.

1991 Mathematics Subject Classification: 35K, 58F, 76E, 76R.
Keywords and Phrases: Amplitude equations, renormalization group, attractors, ϵ-entropy.

I. Introduction.

Extended dynamical systems deal with the time evolutions of systems where the spatial extension is important. One of the remarkable achievement of the theory of dynamical systems was the proof that if one considers a system in a bounded domain (for example a two dimensional incompressible fluid), then the global attracting set is a compact set with finite dimensional Hausdorff dimension although the phase space of the system is infinite dimensional (see [R.], [T.] etc.). It turns out however that the dimension of the attractor grows with the spatial extension of the system, and attractors of large dimension are not very easy to analyze at the present time. Also the theory of dynamical systems does not provide easily information about the spatial structure of the solutions.

This is not so important for small spatial extension where this spatial structure is rather simple. However systems with large spatial extension develop interesting spatial structures. One of the most common of these structures are the waves on a sea excited by a gentle wind (see [M.] for a discussion of the present status of knowledge of this major phenomenon). As a simple criteria we will say that a system is extended if the spatial size of the system is much larger than the typical size of the structures. As will be explained below, in many interesting situations the scale of the structures is well defined. For example, even during the wildest storms, the wavelength of the waves is much smaller than the size of the sea (even of some large lakes).

Note that for spatially extended systems defined in large but bounded spatial domains, all the large time information about spatial structures should be available from the study of the attractor(s), invariant measures etc. The problem is that it is not easy at all to extract spatial information out of the non linear structure of the dynamics in phase space. This is why Physicists have been looking for a long

time for a direct approach which emphasizes the search for structures and their evolution.

There are other difficulties with large systems which are of more technical nature. When one tries to use bifurcation theory for example, one is lead to study the spectrum of the linearized time evolution around a stationary solution. When the spatial extension of the system becomes large, the spectrum although discrete (in a bounded domain) becomes very dense. In general this implies that bifurcation theory gives results only on a very small range of parameters. This has also unpleasant experimental consequences since a small variation of the parameter near criticality can result in a large number of eigenvalues becoming unstable.

Physicists have dealt with these difficulties since a long time. By analogy with Statistical Mechanics, one may hope that if a system has a large spatial extension, its behavior may be well approximated by the behavior of a system with infinite extension (the so called thermodynamic limit). Of course one may expect corrections from far away boundaries.

This assumption has an important technical consequence. When studying the spectrum of the linearized evolution in bounded domains one should use some adequate basis of functions (Fourier series etc.). In unbounded domains, at least for operators with constant coefficients, the spectra is easily obtained using Fourier transform which is a much more convenient tool. A lot of important results have been obtained this way by Physicists (see for example [Ch.]). In fact one could remark that whenever Fourier transform is used in a Physical problem to deal with spatial dependence, an assumption of infinite spatial extension has been made.

Note that contrary to the case of spatially finite systems, extended systems lead in general to continuous spectrum for the linearized evolution. Another difficulty, almost never mentioned in the Physics literature is the nature of phase space of extended systems. Since spectral theory will be an important tool, one would imagine working in a phase space which is for example a Sobolev space. Several interesting works have been done in that direction. However this is not the phase space one would like to use. For example, such a space does not contain waves. Therefore more natural and interesting phase spaces should be like L^∞.

The rest of this paper is organized as follows. In section 2 we will present some results on the global existence of the time evolution in extended domains. In section 3 we will discuss the instabilities of homogeneous solutions and see how they can lead to the appearance of structures at well defined scales. Finally in section 4 we will present some results dealing with questions of large time asymptotic.

II. Global Existence of the Semi-Flow.

The first mathematical question with extended systems is the problem of global existence of the semi flow of time evolution. As mentioned in the introduction, one of the difficulty is that we want to deal with a rather large phase space containing in particular wave-like solutions which do not tend to zero at infinity. Several results have been obtained in the case of Sobolev phase spaces, however the methods do not seem to apply to the phase spaces which are required for the general Physical applications. Very few results are available for only bounded initial conditions, and we will briefly describe some of these results. As in the case of dynamical systems,

it is convenient to work with some simplified models which exhibit the essential phenomena without the complexity of the real equations. One such model is the so called Swift-Hohenberg equation. This equation gives the time evolution of a real field $u(t,x)$ and is given in one space dimension by

$$\partial_t u = \eta u - (1+\partial_x^2)^2 u - u^3 \,, \tag{SH}$$

where η is a real parameter. This equation was derived by Swift and Hohenberg as a model for the onset of convection [S.H.].

Another popular model whose importance will become clearer below is the so called complex Ginzburg-Landau equation. This equation describes the time evolution of a complex field $A(t,x)$ and is given by

$$\partial_t A = (1+i\alpha)\Delta A + A - (1+i\beta)A|A|^2 \,, \tag{CGL}$$

where α and β are two real parameters.

The basic problem is to prove that these equations have (nice) solutions for all time if we start with an initial condition which is only bounded (and somewhat regular). The case of the Ginzburg-Landau equation was treated first in [C.E.1] in dimension one and generalized in [C.1] and [G.V.]. Regularity of the solution was obtained in [C.3] and [T.B.]. We summarize the results in the following theorem.

Theorem II.1 ([C.1],[C.3]). *In dimension one and two, for any complex valued function A_0 of the space variable x, bounded and uniformly continuous, there is a unique solution A of the (CGL) equation with initial condition A_0. This function A is for all times bounded and uniformly continuous. Moreover, there is a positive constant $T = T(A_0,\alpha,\beta)$, and two positive constants $C = C(\alpha,\beta)$ and $h = h(\alpha,\beta)$ such that for any $t > T$, the function $A(t,\cdot)$ extends to a function analytic in the strip $|\Im z| \le h$ and satisfying*

$$\sup_{|\Im z|\le h} |A(t,z)| \le C \,.$$

In other words, the dynamics contracts the large fields to a universal invariant ball, and moreover regularity develops. In the case of dimension three and higher, estimates are presently only available for a restricted range of parameters, we refer to the original publications for more details. A similar result holds for the Swift-Hohenberg equation. The case of reaction-diffusion equations may prove more difficult to deal with, see [C.X.].

As mentioned above, in order to prove such a result one has to use techniques which are rather different from the case of bounded domains. Theorem II.1 has been proven using a local energy estimate. We only indicate the basic starting point. Note that the local (in time) existence and boundedness of the solution follows easily from the usual techniques using the contraction mapping principle. The main goal is therefore to obtain some global a-priori estimate.

Let φ be a regular function tending to zero sufficiently fast at infinity. For example

$$\varphi(x) = \frac{1}{(1+|x|^2)^d}$$

where d is the dimension. The idea is to probe the size of a bounded function by looking at the L^2 norm of this function multiplied by some translate of φ. This is reminiscent of the so called amalgam spaces.

The basic quantity to estimate is the number

$$I(t) = \sup_{x_0} I(t, x_0)$$

where

$$I(t, x_0) = \int |A(t, x)|^2 \, \varphi(x - x_0) \, dx \; .$$

After some simple algebra and integration by parts, one gets easily

$$\frac{d}{dt} I(t, x_0) \leq K - \int |A(t, x)|^2 \, \varphi(x - x_0) \, dx = K - I(t, x_0) \; .$$

where K is some constant which depends only on φ and the coefficients α and β. This differential inequality tells us that after some time the quantity $I(t, x_0)$ will settle forever below $2K$. Note also that the time it takes to reach this situation can be bounded above by a quantity which depends only on $\|A_0\|_{L^\infty}$ (and of course on the coefficients α and β of the equation). The rest of the proof is based on similar but more involved estimates. We refer the reader to the original papers for more details.

III. Instabilities and Structures.

As mentioned above, instability in extended systems leads very often to the development of structures with a well defined wave length. We will illustrate this phenomenon on the one dimensional Swift-Hohenberg equation (S.H.). We first observe that for any value of the parameter η, the homogeneous function $u(t, x) = 0$ is a stationary solution. If we linearize the evolution around this solution, we get the equation

$$\partial_t v = \eta v - (1 + \partial_x^2)^2 v \; ,$$

which describes the linear time evolution of small perturbations of the homogeneous solution. This equation can be explicitly solved by taking the Fourier transform in x. One gets

$$\partial_t \hat{v}(t, k) = \omega(\eta, k) \hat{v}(t, k) \tag{III.1}$$

where

$$\omega(\eta, k) = \eta - (1 - k^2)^2 \tag{III.2} \; .$$

It is then easy to see that if $\eta < 0$, the solution tends to zero ($\omega < 0$), whereas if $\eta > 0$ some Fourier modes are exponentially amplified ($\omega > 0$ for k near ± 1). As mentioned above, one should be careful with the interpretation of this result in direct space since we want to work in a phase space of functions which do not decay at infinity and whose Fourier transforms are in general distributions. Nevertheless this trivial analysis will give the right intuition. It is indeed possible to prove that

for the complete non-linear (S.H.) equation and $\eta < 0$, bounded initial conditions relax to zero.

The case $\eta > 0$ is of course more interesting and reminiscent of bifurcation theory. Recall that the main idea of bifurcation theory is that in phase space, the dominant part of the bifurcated branch is along the subspace of the linear problem which becomes unstable. The amplitude in that direction(s) varying slowly. For $\eta > 0$ small, we have in the (SH) equation two bands of modes of width $\mathcal{O}(\eta^{1/2})$ around ± 1 which are unstable. All other modes are linearly damped. This follows easily from (III.1) and (III.2). By analogy with standard bifurcation theory, one may expect that the coefficient in the unstable directions will vary slowly in space and time. This is indeed what can be proven.

Theorem III.1. *There are positive numbers R, η_0, $C_1, \cdots, C_4$ such that if $\eta \in]0, \eta_0[$, if $u_0(x)$ is a real bounded uniformly continuous function such that $\|u_0\|_{L^\infty} < R$, there is a positive number $T_1 = T_1(u_0, \eta)$ such that for any $t > T_1$, the solution u of (S.H.) with initial condition u_0 satisfies*

$$\|u(t, \cdot)\|_{L^\infty} < C_1 \eta^{1/2} .$$

Moreover, for any $t > T_1$, there is a solution $B(s, y)$ of the real Ginzburg-Landau equation

$$\partial_s B = \partial_y^2 B + B - B|B|^2 \qquad \text{(G.L.)}$$

such that for any $t \le \tau \le t + C_2 \eta^{-1} \log \eta^{-1}$ we have

$$\|u(\tau, \cdot) - \mathcal{B}(\tau, \cdot)\|_{L^\infty} < C_3 \eta^{1/2 + C_4} ,$$

where

$$\mathcal{B}(\tau, x) = 3^{-1/2} \eta^{1/2} e^{ix} B(\eta(\tau - t), \eta^{1/2} x/2) + c.c.$$

This result is similar to what can be obtained in bifurcation theory using normal forms. It says that the function reconstructed from the normal form (here the (G.L.) equation) reproduces well the true evolution during a large time. However, as for normal forms, we cannot expect this to be true forever since small errors due to truncation of the normal form are likely to be amplified by the unstable dynamics.

The idea of amplitude equation is rather old, and we refer to [C.H.] for references. Several versions of the above result (or similar ones) have been published in [C.E.1], [v.H.], [K.M.S.], [S.]. The original idea of shadowing of trajectories is due to Eckhaus [E.]. A new proof using a dynamical renormalization group was given in [C.2] for the case of discrete evolution equations. The renormalization group method has several advantages. First of all it provides a systematic and rigorous approach to multi-scale analysis. It also provides a proof of the above theorem in one step. In the above formulation, it was convenient to separate the initial contraction regime from the subsequent shadowing result. It turns out in the proof that they are different manifestations of the same renormalization effect.

Also one gets some information on the initial contraction phase. The fact that the size of the initial condition R can be chosen independent of η is important and in a sense is optimal since one cannot expect the result to hold for initial conditions of size much larger than unity without further hypothesis since the dynamics may well have another fixed point of order one (although for the particular case of the (S.H.) equation one can prove global attraction). Last but not least, renormalization group produces universal results. This is a nice substitute for the generic arguments of finite dimensional bifurcation theory. This explains why the Ginzburg-Landau equation appears so often in the study of instabilities of extended systems. It turns out that the associated fixed point of the renormalization group is the equation

$$\partial_s B = \partial_y^2 B - B|B|^2 \ ,$$

which is invariant by the rescaling $s \to L^2 s$, $y \to Ly$, $B \to LB$. The relevant unstable manifold parameterized by a real number σ is the (unnormalized) G.L. equation

$$\partial_s B = \partial_x^2 B + \sigma B - B|B|^2 \ .$$

We refer to [B.K.1] and [C.2] for more details and references on the renormalization group ideas and to [A.] page 212 for a general program.

IV. Large Time Behavior.

For dissipative systems in bounded domains, various notions of attractors have been introduced. One tries to describe in the phase space an invariant set which captures all the asymptotic dynamics. Various results have been proven about the compactness and finite Hausdorff dimension of such objects. For extended systems we cannot hope for such results and the definition of the global attracting set has to be modified due to the lack of compactness. Mielke and Schneider [M.S.] following an idea of Feireisl have proposed to define a global attracting set using two different topologies. One is a global topology (of the type L^∞), the other one is a local topology where one recovers compactness. We give below a variant of their result for the (CGL) equation (for other equations see [M.S.]).

THEOREM IV.1. *For the (CGL) equation in dimension 1 and 2, there is a set $\mathcal{A}$ of functions analytic in a strip of width h around the real space and satisfying*

$$\sup_{|\Im z|<h} |A| < C,$$

where h and C are as in Theorem II.1 and such that

1) $\mathcal{A}$ is closed in L^∞,
2) $\mathcal{A}$ is invariant by space translations,
3) $\mathcal{A}$ is invariant by the semi flow (S_t) of evolution of the (CGL) equation (namely $S_t(\mathcal{A}) = \mathcal{A}$ for any $t > 0$),
4) $\mathcal{A}$ is compact in $L^\infty(Q)$ for any cube Q,
5) $\mathcal{A}$ attracts any bounded set of L^∞, namely if B is a bounded set in L^∞, the L^∞ distance between $S_t(B)$ and $\mathcal{A}$ tends to zero when t tends to infinity.

We refer to [M.S.] for the proof. We mention however that although 4) is trivial from the analyticity of the functions in $\mathcal{A}$, this compactness property is

crucial in the proof of 3) and 5) together with the fact that the L^∞ norm of a function on the whole line is obtained by taking the sup of the L^∞ norms of the function on the translates of a fixed cube.

Once the global attractor of a dynamical system has been identified, one can try to give some geometrical description of this object. As mentioned before, for systems in bounded domains one tries to prove that the attractor has finite Hausdorff dimension. A natural question for extended domains is whether there is a good notion of dimension per unit volume of space. In this direction, Ghidaglia and Heron [G.H.] have given for the (CGL) equation in finite domain an upper bound on the Hausdorff dimension of the attractor which is proportional to the length of the domain in space dimension one and proportional to the surface in space dimension two. However contrary to the case of statistical mechanics, it is not clear at this moment whether a sub-additive result holds for the dimension. The main difficulty is to connect the dimension of attractors for the union of two domains.

We have recently considered this question with J.-P. Eckmann from another point of view related to signal analysis. For simplicity I will only discuss one dimensional systems although the results are true in any dimension. We start directly with the system in an unbounded domain, but we observe it in a finite window, for example the interval $[-L, L]$. This is quite natural in view of the above definition of attractor. Note however that since the functions on the attractor $\mathcal{A}$ are analytic they will be seen in any interval. This implies that the dimension of $\mathcal{A}$ in $L^\infty([-L, L])$ is infinite. Kolmogorov and Tikhomirov have studied a similar situation for some spaces of analytic functions [K.T.]. They have defined following Shannon the ϵ-entropy per unit length H_ϵ as follows. Let $\mathcal{B}$ be a subset of $L^\infty(\mathrm{R})$. For a fixed $\epsilon > 0$ one defines $N_L(\epsilon)$ as the smallest number of balls of radius at most ϵ (in $L^\infty([-L, L])$) needed to cover $\mathcal{B}$. The ϵ entropy per unit length of $\mathcal{B}$ is defined by

$$H_\epsilon(\mathcal{B}) = \lim_{L\to\infty} \frac{\log_2 N_L(\epsilon)}{L} ,$$

provided the limit exists. One is then interested at the behavior of this quantity when ϵ tends to zero. Note the exchange of limits with respect to the usual definition of dimension. For the attractor $\mathcal{A}$, if one fixes L and let ϵ tends to zero, one gets an infinite dimension. In other words, for a fixed precision ϵ, if the size of the window is too small, one gets the impression of an object of infinite dimension. As the result below indicates, there is however a cross-over length which depends on the precision ϵ beyond which one sees a finite dimension per unit length (at this fixed precision). Kolmogorov and Tikhomirov in [K.T.] proved the following estimates.

For the set $\mathcal{E}_\sigma(C)$ of entire functions satisfying

$$|f(z)| \le Ce^{\sigma|\Im z|}$$

one has

$$H_\epsilon(\mathcal{E}_\sigma(C)) \approx \frac{2\sigma}{\pi} \log_2(1/\epsilon) ,$$

where $\approx$ means that the ratio of the two quantities tend to 1 when ϵ tends to zero.

For the set $\mathcal{S}_\sigma(C)$ of functions analytic in a strip of width h around the real axis and satisfying

$$\sup_{|\Im z|\leq h} |f(z)| \leq C$$

one has

$$H_\epsilon(\mathcal{S}_h(C)) \approx \frac{1}{\pi h}(\log_2(1/\epsilon))^2 \, .$$

We refer to [K.T.] for the proof of these two statements.

From the previous result on the analyticity of the functions in $\mathcal{A}$ one would expect a growth of the ϵ-entropy proportional to $(\log \epsilon)^2$. It turns out that there is in a sense far less functions in $\mathcal{A}$, and in the sense of ϵ-entropy we have indeed a finite dimension per unit length.

THEOREM [C.E.4]. *There is a number $c = c(\alpha, \beta) > 1$ such that for the (CGL) in dimension 1 and 2 we have*

$$c^{-1} \log_2(1/\epsilon) \leq H_\epsilon(\mathcal{A}) \leq c \log_2(1/\epsilon) \, .$$

Note that some functions belonging to $\mathcal{A}$ are known which are not entire. We have also obtained recently with J.-P. Eckmann a proof of existence of the topological entropy per unit volume. Moreover this quantity can also be obtained from a discrete sampling of the solutions (see [C.E.5]).

V. CONCLUSIONS.

Extended systems occur naturally in many natural questions. They appear in Physics, Chemistry, Biology, Ecology and other sciences as soon as the spatial extension of the system becomes important. We refer to [C.H.] [B.N.] and [Mu.] for some examples.

From the mathematical point of view there are many open problems. The understanding of the evolution of structures and the occurrence of spatio temporal chaos are the most challenging. There are very few results in these area where even numerical simulations are difficult to perform. As in the case of finite dimensional dynamical systems, there are two main trends of research up to now.

In the first one, one tries to understand the spatial structure of the solutions. This is quite natural near the onset of instability of the homogeneous state, where the structures play a dominant role. We refer to [B.N.] for a review of this approach. Even near onset there are important questions which are not understood. For example in dimension 2 or larger, the amplitude should be a distribution on the unit circle and up to now a global derivation of an amplitude equation has not been performed. The analysis has only been achieved under various symmetry assumptions which strongly restrict the solutions, although these solutions with symmetries are the ones which appear in experiments. In a different perspective, particular solutions with interesting physical meaning have been constructed (see for example [C.E.3] for more details).

The second trend of research is more of statistical nature and is concerned with asymptotic time evolution. The existence of interesting invariant measures is still an open problem. Beyond numerical simulations some analogies have been drawn in the spatio temporal intermittency transition with directed percolation (see [B.P.V.] for a review). Even in the case where there is no spatio temporal chaos, the asymptotic state may be non trivial as in the phase ordering kinetics problem (see [B.] where consequences of scaling hypothesis are developed). We mention however that in the problem of coupled lattice maps, interesting invariant measures have been constructed ([B.K.2] and references therein).

Finally I refer to the conclusion of Bowman and Newell in their RPM Colloquia [B.N.] for a statement on the future of this field.

References.

[A.] P.W.Anderson. *Basic Notions of Condensed Matter Physics*. Benjamin-Cummings 1984.

[B.] A.J.Bray. Theory of phase-ordering kinetics. Advances in Physics 43, 357-459 (1994).

[B.K.1] J.Bricmont, A.Kupiainen. Renormalizing Partial Differential Equations, in *XIth International congress on mathematical physics*, D.Iagolnitzer ed. International Press Incorporated, Boston (1995).

[B.K.2] J.Bricmont, A.Kupiainen. High temperature expansions and dynamical systems. Commun. Math. Phys. 178, 703-732 (1996).

[B.P.V.] P.Bergé, Y.Pomeau, C.Vidal. *L'Espace Chaotique*. Paris, Hermann 1998.

[B.N.] C.Bowman, A.C. Newell. Natural patterns and wavelets. Rev. Mod. Phys. 70, 289-301 (1998).

[Ch.] S.Chandrasekhar. *Hydrodynamic and Hydromagnetic Stability*. New-York, Dover 1981.

[C.1] P.Collet.Thermodynamic limit of the Ginzburg-Landau equation. Nonlinearity 7, 1175-1190 (1994).

[C.2] P.Collet. Amplitude equation for lattice maps. A renormalization group approach. J.Stat. Phys 90, 1075-1105 (1998).

[C.3] P.Collet. Non linear parabolic evolutions in unbounded domains. In *Dynamics, Bifurcations and Symmetries*, pp 97-104, P.Chossat editor. Nato ASI 437,Plenum, New York, London 1994.

[C.E.1] P.Collet, J.-P.Eckmann. The time dependent amplitude equation for the Swift-Hohenberg problem. Commun. Math. Phys. 132, 135-153 (1990).

[C.E.2] P.Collet, J.-P.Eckmann. Space-Time behavior in problems of hydrodynamic type: a case study. Nonlinearity 5, 1265-1302 (1992).

[C.E.3] P.Collet, J.-P.Eckmann. *Instabilities and Fronts in Extended Sytems*. Princeton University Press, Princeton 1990.

[C.E.4] P.Collet, J.-P.Eckmann. Extensive properties of the Ginzburg-Landau equation. Preprint (1998).

[C.E.5] P.Collet, J.-P.Eckmann. The definition and measurement of the topological entropy per unit volume in parabolic pde's. Preprint (1998).

[C.X.] P.Collet J.Xin. Global Existence and Large Time Asymptotic Bounds of L^∞ Solutions of Thermal Diffusive Combustion Systems on R^n. Ann.

Scuela Norm. Sup. 23, 625-642 (1996).

[C.H.] M.C.Cross, P.C.Hohenberg. Pattern formation outside of equilibrium. Rev. Mod. Phys. 65, 851-1112 (1993).

[E.] W.Eckhaus. The Ginzburg-Landau manifold is an attractor. J. Nonlinear Sci. 3, 329-348 (1993).

[G.H.] J.M.Ghidaglia, B.Heron. Dimension of the attractors associated to the Ginzburg-Landau Partial Differential Equation. Physica 28 D, 282-304 (1987).

[G.V.] J.Ginibre, G.Velo. The Cauchy problem in local spaces for the complex Ginzburg-Landau equation. II: contraction methods. Commun. Math. Phys. 187, 45-79 (1997).

[K.M.S.] P.Kirrmann,A.Mielke, G.Schneider. The validity of modulation equations for extended systems with cubic nonlinearities. Proc. R. Soc. Edinb. 122, 85-91 (1992).

[K.T.] A.N.Kolmogorov, V.M.Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. In *Selected works of A.N.Kolmogorov, Vol III.* A.N. Shiryayev ed. Dordrecht, kluwer 1993.

[L.N.P.] J.Legas, A.Newell, T.Passot. Order parameter equations for patterns. Ann. Rev. of Fluid Mech. 25, 399-453 (1993).

[M.S.] A.Mielke, G.Schneider. Attractors for modulation equation on unbounded domains-existence and comparison. Nonlinearity 8, 743-768 (1995).

[M.] J.Miles. Generation of surface waves by wind. Appl. Mech. Rev. 50, R5-R9 (1997).

[Mu.] J.D.Murray. *Mathematical biology*. Berlin, Springer, 1993.

[R.] D.Ruelle Characteristic exponents for a viscous fluid subjected to time dependent forces. Commun. Math. Phys. 93, 285-300 (1984).

[S.] G.Schneider. Global existence via Ginzburg-Landau formalism and pseudo-orbits of Ginzburg-Landau approximations. Preprint Hannover (1993).

[S.H.] J.Swift, P.Hohenberg. Hydrodynamic fluctuations at the convective instability. Phys. Rev. A15, 319 (1977).

[T.] R.Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. New-York, Springer (1988).

[T.B.] P.Takac, P.Bollerman, A.Doelman, A.van Harten, E.S.Titi. Analyticity of essentially bounded solutions to semilinear parabolic systems and validity of the Ginzburg-Landau equation. SIAM J. Math. Anal. 27, 424-448 (1996).

[vH.] A.van Harten. On the validity of the Ginzburg-Landau's equation. Journal of Nonlinear Sciences 1, 397-422 (1992).

P. Collet
Centre de Physique Théorique
CNRS UMR 7644, Ecole Polytechnique
F-91128 Palaiseau Cedex (France)

The Mathematics of Fivebranes

Robbert Dijkgraaf

Abstract.

Fivebranes are non-perturbative objects in string theory that generalize two-dimensional conformal field theory and relate such diverse subjects as moduli spaces of vector bundles on surfaces, automorphic forms, elliptic genera, the geometry of Calabi-Yau threefolds, and generalized Kac-Moody algebras.

1991 Mathematics Subject Classification: 81T30
Keywords and Phrases: quantum field theory, elliptic genera, automorphic forms

1 Introduction

This joint session of the sections Mathematical Physics and Algebraic Geometry celebrates a historic period of more than two decades of remarkably fruitful interactions between physics and mathematics. The 'unreasonable effectiveness', depth and universality of quantum field theory ideas in mathematics continue to amaze, with applications not only to algebraic geometry, but also to topology, global analysis, representation theory, and many more fields. The impact of string theory has been particularly striking, leading to such wonderful developments as mirror symmetry, quantum cohomology, Gromov-Witten theory, invariants of three-manifolds and knots, all of which were discussed at length at previous Congresses.

Many of these developments find their origin in two-dimensional conformal field theory (CFT) or, in physical terms, in the first-quantized, perturbative formulation of string theory. This is essentially the study of sigma models or maps of Riemann surfaces Σ into a space-time manifold X. Through the path-integral over all such maps a CFT determines a partition function Z_g on the moduli space $\mathcal{M}_g$ of genus g Riemann surfaces. String amplitudes are functions $Z(\lambda)$, with λ the string coupling constant, that have asymptotic series of the form

$$Z(\lambda) \sim \sum_{g\geq 0} \lambda^{2g-2} \int_{\mathcal{M}_g} Z_g. \tag{1}$$

But string theory is more than a theory of Riemann surfaces. Recently it has become possible to go beyond perturbation theory through conceptual breakthroughs such as string duality [23] and D-branes [19]. Duality transformations

can interchange the string coupling λ with the much better understood geometric moduli of the target space X. D-Branes are higher-dimensional extended objects that give rise to special cycles $Y \subset X$ on which the Riemann surface can end, effectively leading to a *relative* form of string theory.

One of the most important properties of branes is that they can have multiplicities. If k branes coincide a non-abelian $U(k)$ gauge symmetry appears. Their 'world-volumes' carry Yang-Mills-like quantum field theories that are the analogues of the two-dimensional CFT on the string world-sheet. The geometric realization as special cycles (related to the theory of calibrations) has proven to be a powerful tool to analyze the physics of these field theories. The mathematical implications are just starting to be explored and hint at an intricate generalization of the CFT program to higher dimensions.

This lecture is a review of work done on one of these non-perturbative objects, the fivebrane, over the past years in collaboration with Erik Verlinde, Herman Verlinde and Gregory Moore [4, 5, 8, 7]. I thank them for very enjoyable and inspiring discussions.

2 FIVEBRANES

One of the richest and enigmatic objects in non-perturbative string theory is the so-called fivebrane, that can be considered as a six-dimensional cycle Y in space-time. Dimension six is special since, just as in two dimensions, the Hodge star satisfies $*^2 = -1$ and one can define *chiral* or 'holomorphic' theories. The analogue of a free chiral field theory is a 2-form 'connection' B with a self-dual curvature H that is locally given as $H = dB$ but that can have a 'first Chern class' $[H/2\pi] \in H^3(Y, \mathbb{Z})$. (Technically it is a Deligne cohomology class, and instead of a line bundle with connection it describes a 2-gerbe on Y.) A system of k coinciding fivebranes is described by a 6-dimensional conformal field theory, that is morally a $U(k)$ non-abelian 2-form theory. Such a theory is not known to exist at the classical level of field equations, so probably only makes sense as a quantum field theory.

One theme that we will not further explore here is that (at least for $k = 1$) the fivebrane partition function Z_Y can be obtained by quantizing the intermediate Jacobian of Y, very much in analogy with the construction of conformal blocks by geometric quantization of the Jacobian or moduli space of vector bundles of a Riemann surface [24]. This leads to interesting relations with the geometry of moduli spaces of Calabi-Yau three-folds and topological string theory. In fact there is even a seven-dimensional analogue of Chern-Simons theory at play.

The fivebrane theory is best understood on manifolds of the product form $Y = X \times T^2$, with X a 4-manifold. In the limit where the volume of the two-torus goes to zero, it gives a $U(k)$ Yang-Mills theory as studied in [21]. In that case the partition function computes the Euler number of the moduli space of $U(k)$ instantons on X. In the $k = 1$ case this relation follows from the decomposition of the 3-form

$$H = F_+ \wedge dz + F_- \wedge d\overline{z} \tag{2}$$

with $F_\pm$ (anti)-self-dual 2-forms on X. In this way holomorphic fields on T^2 are coupled to self-dual instantons on X. The obvious action of $SL(2, \mathbb{Z})$ on T^2

translates in a deep quantum symmetry (S-duality) of the 4-dimensional Yang-Mills theory.

Actually, the full fivebrane theory is much richer than a 6-dimensional CFT. It is believed to be a six-dimensional *string* theory that does not contain gravity and that reduces to the CFT in the infinite-volume limit. We understand very little about this new class of string theories, other than that they can be described in certain limits as sigma models on instanton moduli space [20, 7, 1, 25]. As we will see, this partial description is good enough to compute certain topological indices, where only so-called BPS states contribute.

3 Conformal field theory and modular forms

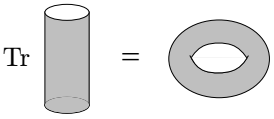
One of the striking properties of conformal field theory is the natural explanation it offers for the modular properties of the characters of certain infinite-dimensional Lie algebras such as affine Kac-Moody algebras. At the hart of this explanation—and in fact of much of the applications of quantum field theory to mathematics—lies the equivalence between the Hamiltonian and Lagrangian formulation of quantum mechanics [22]. For the moment we consider a *holomorphic* or chiral CFT.

In the Hamiltonian formulation the partition function on an elliptic curve T^2 with modulus τ is given by a trace over the Hilbert space $\mathcal{H}$ obtained by quantization on $S^1 \times \mathbb{R}$. For a sigma model with target space X, this Hilbert space will typically consist of L^2-functions on the loop space $\mathcal{L}X$. It forms a representation of the algebra of quantum observables and is $\mathbb{Z}$-graded by the momentum operator P that generates the rotations of S^1. For a chiral theory P equals the holomorphic Hamiltonian $L_0 = z\partial_z$. The character of the representation is then defined as

$$Z(\tau) = \mathrm{Tr}_{\mathcal{H}} q^{P - \frac{c}{24}} \tag{3}$$

with $q = e^{2\pi i \tau}$ and c the central charge of the Virasoro algebra. The claim is that this character is always a suitable modular form for $SL(2,\mathbb{Z})$, *i.e.*, it transforms covariantly under linear fractional transformations of the modulus τ.

In the Lagrangian formulation $Z(\tau)$ is computed from the path-integral over maps from T^2 into X. The torus T^2 is obtained by gluing the two ends of the cylinder $S^1 \times \mathbb{R}$, which is the geometric equivalent of taking the trace.



Modularity is therefore built in from the start, since $SL(2,\mathbb{Z})$ is the 'classical' automorphism group of the torus T^2.

The simplest example of a CFT consists of c free chiral scalar fields $x : \Sigma \to V \cong \mathbb{R}^c$. Ignoring the zero-modes, the chiral operator algebra is then given by an infinite-dimensional Heisenberg algebra that is represented on the graded Fock space

$$\mathcal{H}_q = \bigotimes_{n>0} S_{q^n} V. \tag{4}$$

Here we use a standard notation for formal sums of (graded) symmetric products

$$S_q V = \bigoplus_{N\geq 0} q^N S^N V, \qquad S^N V = V^{\otimes N}/S_N. \tag{5}$$

The partition function is then evaluated as

$$Z(\tau) = q^{-\frac{c}{24}} \prod_{n>0} (1-q^n)^{-c} = \eta(q)^{-c} \tag{6}$$

and is indeed a modular form of $SL(2,\mathbb{Z})$ of weight $-c/2$ (with multipliers if $c \not\equiv 0 \mod 24$.) The 'automorphic correction' $q^{-c/24}$ is interpreted as a regularized sum of zero-point energies that naturally appear in canonical quantization.

4 String theories and automorphic forms

The partition function of a *string* theory on a manifold Y will have automorphic properties under a larger symmetry group that reflects the 'stringy' geometry of Y. For example, if we choose $Y = X \times S^1 \times \mathbb{R}$, with X compact and simply-connected, quantization will lead to a Hilbert space $\mathcal{H}$ with a natural $\mathbb{Z}\oplus\mathbb{Z}$ gradation. Apart from the momentum P we now also have a winding number W that labels the components of the loop space $\mathcal{L}Y$. Thus we can define a two-parameter character

$$Z(\sigma,\tau) = \mathrm{Tr}_{\mathcal{H}}\big(p^W q^P\big), \tag{7}$$

with $p = e^{2\pi i\sigma}, q = e^{2\pi i\tau}$, with both σ,τ in the upper half-plane $\mathbb{H}$. We claim that $Z(\sigma,\tau)$ is typically the character of a generalized Kac-Moody algebra [2] and an automorphic form for the arithmetic group $SO(2,2;\mathbb{Z})$.

The automorphic properties of such characters become evident by changing again to a Lagrangian point of view and computing the partition function on the compact manifold $X \times T^2$. The T-duality or 'stringy' symmetry group of T^2 is

$$SO(2,2;\mathbb{Z}) \cong PSL(2,\mathbb{Z}) \times PSL(2,\mathbb{Z}) \ltimes \mathbb{Z}_2, \tag{8}$$

where the two $PSL(2,\mathbb{Z})$ factors act on (σ,τ) by separate fractional linear transformations and the mirror map $\mathbb{Z}_2$ interchanges the complex structure τ with the complexified Kähler class $\sigma \in H^2(T^2,\mathbb{C})$. This group appears because a string moving on T^2 has both a winding number $w \in \Lambda = H_1(T^2;\mathbb{Z})$ and a momentum vector $p \in \Lambda^*$. The 4-vector $k = (w,p)$ takes value in the even, self-dual Narain lattice $\Gamma^{2,2} = \Lambda \oplus \Lambda^*$ of signature $(2,2)$ with quadratic form $k^2 = 2w\cdot p$ and automorphism group $SO(2,2;\mathbb{Z})$.

In the particular example we will discuss in detail in the next sections, where the manifold X is a Calabi-Yau space, there will be an extra $\mathbb{Z}$-valued quantum number and the Narain lattice will be enlarged to a signature $(3,2)$ lattice. Correspondingly, the automorphic group will be given by $SO(3,2,\mathbb{Z}) \cong Sp(4,\mathbb{Z})$.

5 Quantum mechanics on the Hilbert scheme

As we sketched in the introduction, in an appropriate gauge the quantization of fivebranes is equivalent to the sigma model (or quantum cohomology) of the moduli space of instantons. More precisely, quantization on the six-manifold $X \times S^1 \times \mathbb{R}$, gives a graded Hilbert space

$$\mathcal{H}_p = \bigoplus_{N \geq 0} p^N \mathcal{H}_N, \tag{9}$$

where $\mathcal{H}_N$ is the Hilbert space of the two-dimensional supersymmetric sigma model on the moduli space of $U(k)$ instantons of instanton number N on X. If X is an algebraic complex surface, one can instead consider the moduli space of stable vector bundles of rank k and $ch_2 = N$. This moduli space can be compactified by considering all torsion-free coherent sheaves up to equivalence. In the rank one case it coincides with the Hilbert scheme of points on X. This is a smooth resolution of the symmetric product $S^N X$. (We note that for the important Calabi-Yau cases of a $K3$ or abelian surface the moduli spaces are all expected to be hyper-Kähler deformations of $S^{Nk}X$.)

The simplest type of partition function will correspond to the Witten index. For this computation it turns out we can replace the Hilbert scheme by the more tractable orbifold $S^N X$. For a smooth manifold M the Witten index computes the superdimension of the graded space of ground states or harmonic forms, which is isomorphic to $H^*(M)$, and therefore equals the Euler number $\chi(M)$.

For an orbifold M/G the appropriate generalization is the *orbifold* Euler number. If we denote the fixed point locus of $g \in G$ as M^g and centralizer subgroups as C_g, this is defined as a sum over the conjugacy classes $[g]$

$$\chi_{orb}(M/G) = \sum_{[g]} \chi_{top}(M^g/C_g). \tag{10}$$

For the case of the symmetric product $S^N X$ this expression can be straightforwardly computed, as we will see in the next section, and one finds

Theorem 1 [13] — *The orbifold Euler numbers of the symmetric products $S^N X$ are given by the generating function*

$$\chi_{orb}(S_p X) = \prod_{n>0} (1 - p^n)^{-\chi(X)}.$$

Quite remarkable, if we write $p = e^{2\pi i \tau}$, the formal sum of Euler numbers is (almost) a modular form for $SL(2, \mathbb{Z})$ of weight $\chi(X)/2$. This is in accordance with the interpretation as a partition function on $X \times T^2$ and the S-duality of the corresponding Yang-Mills theory on X [21].

A much deeper result of Göttsche tells us that the same result holds for the Hilbert scheme [9]. In fact, in both cases one can also compute the full cohomology and express it as the Fock space, generated by an infinite series of copies of $H^*(X)$

shifted in degree [10]

$$H^*(S_pX) \cong \bigotimes_{n>0} S_{p^n} H^{*-2n+2}(X). \tag{11}$$

Comparing with (4) we conclude that the Hilbert space of ground states of the fivebrane is the Fock space of a chiral CFT. This does not come as a surprise given the remarks in the introduction. One can also derive the action of the corresponding Heisenberg algebra using correspondences on the Hilbert scheme [16].

6 The elliptic genus

We now turn from particles to strings. To compute the fivebrane string partition function on $X \times T^2$, we will have to study the two-dimensional supersymmetric sigma model on the moduli space of instantons on X. Instead of the full partition function we will compute again a topological index — the elliptic genus. Let us briefly recall its definition.

For the moment let X be a general complex manifold of dimension d. Physically, the elliptic genus is defined as the partition function of the corresponding $N = 2$ supersymmetric sigma model on a torus with modulus τ [15]

$$\chi(X; q, y) = \mathrm{Tr}_{\mathcal{H}}\Big((-1)^{F_L+F_R} y^{F_L} q^{L_0 - \frac{d}{8}}\Big), \tag{12}$$

with $q = e^{2\pi i\tau}$, $y = e^{2\pi iz}$, z a point on T^2. Here $\mathcal{H}$ is the Hilbert space obtained by quantizing the loop space $\mathcal{L}X$ (formally the space of half-infinite dimensional differential forms). The Fermi numbers $F_{L,R}$ represent (up to an infinite shift that is naturally regularized) the bidegrees of the Dolbeault differential forms representing the states. The elliptic genus counts the number of string states with $\overline{L}_0 = 0$. In terms of topological sigma models, these states are the cohomology classes of the right-moving BRST operator Q_R. In fact, if we work modulo Q_R, the CFT gives a *cohomological* vertex operator algebra.

Mathematically, the elliptic genus can be understood as the S^1-equivariant Hirzebruch χ_y-genus of the loop space of X. If X is Calabi-Yau the elliptic genus has nice modular properties under $SL(2, \mathbb{Z})$. It is a weak Jacobi form of weight zero and index $d/2$ (possibly with multipliers). The coefficients in its Fourier expansion

$$\chi(X; q, y) = \sum_{m \geq 0,\, \ell} c(m, \ell) q^m y^\ell \tag{13}$$

are integers and can be interpreted as indices of twisted Dirac operators on X. For a $K3$ surface one finds the unique (up to scalars) Jacobi form of weight zero and index one, that can expressed in elementary theta-functions as

$$\chi(K3; q, y) = 2^3 \cdot \sum_{\text{even } \alpha} \vartheta_\alpha^2(z; \tau)/\vartheta_a^2(0; \tau). \tag{14}$$

7 Elliptic genera of symmetric products

We now want to compute the elliptic genus of the moduli spaces of vector bundles, in particular of the Hilbert scheme. Again, we first turn to the much simpler symmetric product orbifold $S^N X$.

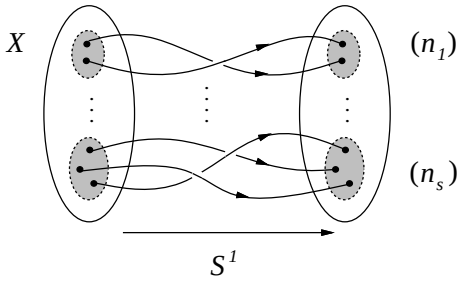
The Hilbert space of a two-dimensional sigma model on any orbifold M/G decomposes in sectors labeled by the conjugacy classes $[g]$ of G, since $\mathcal{L}(M/G)$ has disconnected components of twisted loops satisfying

$$x(\sigma + 2\pi) = g \cdot x(\sigma), \qquad g \in G. \tag{15}$$

In the case of the symmetric product orbifold X^N/S_N these twisted sectors have an elegant interpretation [8]. The conjugacy classes of the symmetric group S_N are labeled by partitions of N,

$$[g] = (n_1)\cdots(n_s), \qquad \sum_i n_i = N, \tag{16}$$

where (n_i) denotes an elementary cycle of length n_i. A loop on $S^N X$ satisfying this twisted boundary condition can therefore be visualized as



As is clear from this picture, one loop on $S^N X$ is *not* necessarily describing N loops on X, but instead can describe $s \leq N$ loops of length $n_1, \ldots, n_s$. By length n we understand that the loop only closes after n periods. Equivalently, the action of the canonical circle action is rescaled by a factor $1/n$.

In this way we obtain a 'gas' of strings labeled by the additional quantum number n. The Hilbert space of the formal sum $S_p X$ can therefore be written as

$$\mathcal{H}(S_p X) = \bigotimes_{n>0} S_{p^n} \mathcal{H}_n(X). \tag{17}$$

Here $\mathcal{H}_n(X)$ is the Hilbert space obtained by quantizing a single string of length n. It is isomorphic to the subspace $P \equiv 0 \pmod{n}$ of the single string Hilbert space $\mathcal{H}(X)$. From this result one derives

Theorem 2 [8] — *Let X be a Calabi-Yau manifold, then the orbifold elliptic genera of the symmetric products $S^N X$ are given by the generating function*

$$\chi_{orb}(S_p X; q, y) = \prod_{n>0,\, m\geq 0,\, \ell} (1 - p^n q^m y^\ell)^{-c(nm,\ell)}.$$

In the limit $q \to 0$ the elliptic genus reduces to the Euler number and we obtain the results from §5. Only the constant loops survive and, since twisted loops then localize to fixed point sets, we recover the orbifold Euler character prescription and Theorem 1.

8 Automorphic forms and generalized Kac-Moody algebras

The fivebrane string partition function is obtained from the above elliptic genus by including certain 'automorphic corrections' and is closely related to an expression of the type studied extensively by Borcherds [3] with the infinite product representation[1]

$$\Phi(\sigma,\tau,z) = p^a q^b y^c \prod_{(n,m,\ell)>0} (1-p^n q^m y^\ell)^{c(nm,\ell)} \tag{18}$$

For general Calabi-Yau space X it can be shown, using the path-integral representation, that the product Φ is an automorphic form of weight $c(0,0)/2$ for the group $SO(3,2,\mathbb{Z})$ for a suitable quadratic form of signature $(3,2)$ [12, 14, 18].

In the important case of a $K3$ surface Φ is the square of a famous cusp form of $Sp(4,\mathbb{Z}) \cong SO(3,2,\mathbb{Z})$ of weight 10,

$$\Phi(\sigma,\tau,z) = 2^{-12} \prod_{even\ \alpha} \vartheta[\alpha](\Omega)^2 \tag{19}$$

the product of all even theta-functions on a genus-two surface Σ with period matrix

$$\Omega = \begin{pmatrix} \sigma & z \\ z & \tau \end{pmatrix}, \qquad \det \operatorname{Im} \Omega > 0. \tag{20}$$

Note that Φ is the 12-th power of the holomorphic determinant of the scalar Laplacian on Σ, just as η^{24} is on an elliptic curve. The quantum mechanics limit $\sigma \to i\infty$ can be seen as the degeneration of Σ into a elliptic curve.

In the work of Gritsenko and Nikulin [11] it is shown that Φ has an interpretation as the denominator of a generalized Kac-Moody algebra. This GKM algebra is constructed out of the cohomological vertex algebra of X similar as in the work of Borcherds. This algebra of BPS states is induced by the string interaction, and should also have an algebraic reformulation in terms of correspondences as in [12].

9 String interactions

Usually in quantum field theory one first quantizes a single particle on a space X and obtains a Hilbert space $\mathcal{H} = L^2(X)$. Second quantization then corresponds to taking the free symmetric algebra $\bigoplus_N S^N\mathcal{H}$. Here we effective reversed the order of the two operations: we considered quantum mechanics on the 'second-quantized' manifold $S^N X$. (Note that the two operations do not commute.) In this

[1] Here the positivity condition means: $n, m \geq 0$, and $\ell > 0$ if $n = m = 0$. The 'Weyl vector' (a,b,c) is defined by $a = b = \chi(X)/24$, and $c = -\frac{1}{4}\sum_\ell |\ell| c(0,\ell)$.

framework it is possible to introduce interactions by deforming the manifold $S^N X$, for example by considering the Hilbert scheme or the instanton moduli space. It is interesting to note that there is another deformation possible.

To be concrete, let X be again a $K3$ surface. Then $S^N X$ or $\mathrm{Hilb}^N(X)$ is an Calabi-Yau of complex dimension $2N$. Its moduli space is unobstructed and 21 dimensional — the usual 20 moduli of the $K3$ surface plus one extra modulus. This follows essentially from

$$h^{1,1}(S^N X) = h^{1,1}(X) + h^{0,0}(X). \tag{21}$$

The extra cohomology class is dual to the small diagonal, where two points coincide, and the corresponding modulus controls the blow-up of this $\mathbb{Z}_2$ singularity. Physically it is represented by a $\mathbb{Z}_2$ twist field that has a beautiful interpretation, that mirrors a construction for the 10-dimensional superstring [6] — it describes the joining and splitting of strings. Therefore the extra modulus can be interpreted as the string coupling constant λ [7, 25].

The geometric picture is the following. Consider the sigma model with target space $S^N X$ on the world-sheet $\mathbb{P}^1$. A map $\mathbb{P}^1 \to S^N X$ can be interpreted as a map of the N-fold *unramified* cover of $\mathbb{P}^1$ into X. If we include the deformation λ the partition function has an expansion

$$Z(\lambda) \sim \sum_{n \geq 0} \lambda^n Z_n, \tag{22}$$

where Z_n is obtained by integrating over maps with n simple branch points. In this way higher genus surfaces appear as non-trivial N-fold branched covers of $\mathbb{P}^1$. The string coupling has been given a geometric interpretation as a modulus of the Calabi-Yau $S^N X$.

It is interesting to note that this deformation has an alternative interpretation in terms of the moduli space of instantons, at least on $\mathbb{R}^4$. The deformed manifold with $\lambda \neq 0$ can be considered as the moduli space of instantons on a *non-commutative* version of $\mathbb{R}^4$ [17].

References

[1] O. Aharony, M. Berkooz, S. Kachru, N. Seiberg, and E. Silverstein, *Matrix Description of Interacting Theories in Six Dimensions,* Phys. Lett. B420 (1998) 55–63, `hep-th/9707079.`

[2] R. Borcherds, *Monstrous moonshine and monstrous Lie superalgebras,* Invent. Math, 109 (1992) 405–444.

[3] R. Borcherds, *Automorphic forms on $O_{s+2,2}(R)$ and infinite products,* Invent. Math. 120 (1995) 161.

[4] R. Dijkgraaf, E. Verlinde and H. Verlinde, *BPS spectrum of the five-brane and black hole entropy,* Nucl. Phys. B486 (1997) 77–88, `hep-th/9603126`; *BPS quantization of the five-brane,* Nucl. Phys. B486 (1997) 89–113, `hep-th/9604055.`

[5] R. Dijkgraaf, E. Verlinde and H. Verlinde, *Counting dyons in $N = 4$ string theory,* Nucl. Phys. B484 (1997) 543, `hep-th/9607026`.

[6] R. Dijkgraaf, E. Verlinde, and H. Verlinde, *Matrix string theory,* Nucl. Phys. B500 (1997) 43–61, hep-th/9703030.

[7] R. Dijkgraaf, E. Verlinde, and H. Verlinde, *5D Black holes and matrix strings,* Nucl. Phys. B506 (1997) 121–142. hep-th/9704018.

[8] R. Dijkgraaf, G, Moore, E. Verlinde, and H. Verlinde, *Elliptic genera of symmetric products and second quantized strings,* Commun. Math. Phys. 185 (1997) 197–209, hepth/9608096.

[9] L. Göttsche, *The Betti numbers of the Hilbert Scheme of Points on a Smooth Projective Surface,* Math.Ann. 286 (1990) 193–207.

[10] L. Göttsche and W. Soergel, *Perverse sheaves and the cohomology of Hilbert schemes of smooth algebraic surfaces*, Math. Ann. 296 (1993), 235–245.

[11] V.A. Gritsenko and V.V. Nikulin, *Siegel automorphic form corrections of some Lorentzian Kac-Moody algebras,* Amer. J. Math. 119 (1997), 181–224, alg-geom/9504006.

[12] J. Harvey and G. Moore, *Algebras, BPS states, and strings,* Nucl. Phys. B463 (1996) 315–368, hep-th/9510182; *On the algebra of BPS states,* hep-th/9609017.

[13] F. Hirzebruch and T. Höfer, *On the Euler Number of an Orbifold,* Math.Ann. 286 (1990) 255.

[14] T. Kawai, $N = 2$ *Heterotic string threshold correction, K3 surface and generalized Kac-Moody superalgebra,* Phys. Lett. B372 (1996) 59–64, hep-th/9512046.

[15] P.S. Landweber Ed., *Elliptic Curves and Modular Forms in Algebraic Topology* (Springer-Verlag, 1988), and references therein.

[16] H. Nakajima, *Heisenberg algebra and Hilbert schemes of points on projective surfaces,* alg-geom/9507012.

[17] N. Nekrasov and A. Schwarz, *Instantons on noncommutative* $\mathbb{R}^4$ *and (2,0) superconformal six-dimensional theory,* hep-th/9802068.

[18] C.D.D. Neumann, *The elliptic genus of Calabi-Yau 3- and 4-folds, product formulae and generalized Kac-Moody Algebras,* J. Geom. Phys. to be published, hep-th/9607029.

[19] J. Polchinski, *Dirichlet-branes and Ramond-Ramond charges,* Phys. Rev. Lett. 75 (1995) 4724–4727, hep-th/9510017.

[20] A. Strominger and C. Vafa, *Microscopic origin of the Bekenstein-Hawking entropy,* Phys. Lett. B379 (1996) 99–104, hep-th/9601029.

[21] C. Vafa and E. Witten, *A strong coupling test of S-duality,* Nucl. Phys. B431 (1994) 3–77, hep-th/9408074.

[22] E. Witten, *Geometry and physics,* Plenary Lecture, ICM, Berkeley (1988).

[23] E. Witten, *String theory in various dimensions,* Nucl. Phys. B443 (1995) 85, hep-th/9503124.

[24] E. Witten, *Five-brane effective action in M-theory,* J. Geom. Phys. 22 (1997) 103–133, hep-th/9610234.

[25] E. Witten, *On the conformal field theory of the Higgs branch,* J. High Energy Phys. 07 (1997) 3, hep-th/9707093.

Robbert Dijkgraaf
Department of Mathematics
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands
rhd@wins.uva.nl

On the Problem of Stability for Near to Integrable Hamiltonian Systems

Antonio Giorgilli

Abstract. Some recent applications and extensions of Nekhoroshev's theory on exponential stability are presented. Applications to physical systems concern on the one hand realistic evaluations of the regions where exponential stability is effective, and, on the other hand, the relaxation time for resonant states in large, possibly infinite systems. Extensions of the theory concern the phenomenon of superexponential stability of orbits in the neigbourhood of invariant KAM tori.

1991 Mathematics Subject Classification: Primary 58F10; Secondary 70F15, 70H05, 70K20
Keywords and Phrases: Perturbation theory, Nekhoroshev theory, exponential stability.

1. Overview

According to Poincaré ([26], tome I, chapt. I, § 13) the general problem of dynamics is the investigation of a canonical system of differential equations with Hamiltonian

$$H(p,q,\varepsilon) = h(p) + \varepsilon f(p,q,\varepsilon) \ , \tag{1}$$

where $(p,q) \in \mathcal{G} \times \mathbf{T}^n$ are action–angle variables, $\mathcal{G} \subset \mathbf{R}^n$ is open, ε is a small parameter and n is the number of degrees of freedom. The functions h and f are assumed to be analytic in all arguments; in particular the perturbation $f(p,q,\varepsilon)$ can be expanded in power series of ε in a neighbourhood of $\varepsilon = 0$. Many physical systems may be described by a Hamiltonian of the form above; the most celebrated one is the planetary system with its natural and (which is of interest now) artificial bodies.

My aim here is to illustrate some results concerning the stability of such systems. The word "stability" is used here in a wide sense, which includes a considerable weakening of the traditional concept investigated, e.g., by Lyapounov. I will pay particular attention to quantities that remain almost constant for a time that increases faster than any inverse power of ε as $\varepsilon \to 0$. Following Littlewood, I will refer to stability estimates of this kind as *exponential stability*.

It is well known that for $\varepsilon = 0$ the unperturbed system $h(p)$ is trivially integrable, since the orbits lie on invariant tori parameterized by the actions p,

and the flow is typically quasiperiodic with frequencies $\omega(p) = \frac{\partial h}{\partial p}$. It has been proven by Poincaré that for $\varepsilon \neq 0$ the system is generically non–integrable (see [26], chapt. V). This is due to the existence of resonances among the frequencies, i.e., relations of the form $\langle k, \omega(p) \rangle = 0$ with $0 \neq k \in \mathbf{Z}^n$.

It was only after the year 1954 that a significant advance of our knowledge was made with the celebrated theorem of Kolmogorov[18], Arnold[1] and Moser[23]. They proved the existence of a set of invariant tori of large relative measure, thus assuring stability in probabilistic sense. Almost at the same time, Moser[22] and Littlewood[19][20] introduced the methods leading to exponential stability. Several years later a general formulation was given by Nekhoroshev, who proved that the action variables p remain almost invariant for a time that increases exponentially with the inverse of the perturbation ε; more precisely, one has

$$\big|p(t) - p(0)\big| < B\varepsilon^b \quad \text{for} \quad |t| < T_* \exp\big((\varepsilon_*/\varepsilon)^a\big) \tag{2}$$

for some constants B, T_*, ε_*, $a \leq 1$ and $b < 1$ (see [24], [25], [3], [4], [21], [14]).

My purpose here is to report on some progress made during the last decade. I will address in particular the following points: (a) the actual relevance of exponential stability for physical systems; (b) the extension of the concept of exponential stability to systems with a very large number of degrees of freedom, and possibly to infinite systems; (c) some relations between KAM and Nekhoroshev's theory, and in particular a stronger stability result that I will call *superexponential stability*.

Both KAM theorem and Nekhoroshev's theorem apply provided the size ε of the perturbation is smaller than a critical value, ε_* say. On the other hand, the problem of finding realistic estimates for the critical value ε_* is generally a very hard one: the analytical estimates available are useless for a practical application to a physical model, and only in a few, very particular models realistic results have been obtained. One such case concerns the stability of the Lagrangian point L_4 of the restricted problem of three bodies in the Sun–Jupiter case. I discuss in sect. 2 how realistic estimates may be obtained by complementing the analytical scheme with explicit calculation of perturbation series.

For systems with a large number of degrees of freedom one is confronted with the problem that all estimates seem to indicate that Nekhoroshev's theorem looses significance for $n \to \infty$ because the constants T_*, ε_* and a tend to zero. As a typical example let us consider a system of identical diatomic molecules moving on a segment and interacting via a short range analytic potential; this may be considered as a one–dimensional model of a gas, the main simplification being that the rotational degrees of freedom of the molecules are not taken into account. The model admits a natural splitting into two subsystems, i.e., the translational motions and the internal vibrations of the molecules, with a coupling due to collisions. According to the equipartition principle, every degree of freedom would get the same average energy. However, it was already suggested by Boltzmann that this should be true only if one considers time averages over a sufficiently long time (relaxation time). Boltzmann's suggestion was that such a time would increase with the frequency of the internal vibrations, becoming of the order of days or centuries (see [9]); a few years later Jeans suggested that the relaxation time could

increase exponentially with the frequency, possibly becoming of the order of billions of years (see [17]). I discuss in sect. 3 how far the suggestion of Boltzmann and Jeans may be dynamically justified if one relinquishes the request that all actions be constant, and pays attention only to the transfer of energy between the two subsystems. For a discussion of the relevance of the exponential stability in statistical mechanics see [7] and [10] and the references therein.

Finally, it is interesting from the theoretical viewpoint to investigate the behaviour of the orbits in the neighbourhood of an invariant KAM torus. I discuss this point in sect. 4 by illustrating how KAM theorem may be obtained by using Nekhoroshev's theorem as a basic iteration step. As a straightforward consequence one gets the result that in most of the phase space the orbits are stable for a time that is much longer than the exponential time predicted by Nekhoroshev. Indeed, the exponential time in (2) is replaced by $\exp\big(\exp(1/\varrho)\big)$, where ϱ is the distance from an invariant KAM torus. This is what I call *superexponential stability*.

2. The triangular Lagrangian equilibria

It is known that in a neighbourhood of an elliptic equilibrium the Hamiltonian may be given the form

$$H(x,y) = \frac{1}{2}\sum_{l=1}^{n} \omega_l \left(x_l^2 + y_l^2\right) + \sum_{s>2} H_s(x,y)\ , \tag{3}$$

where $\omega \in \mathbf{R}^n$ is the vector of the harmonic frequencies and H_s is a homogeneous polynomial of degree s in the canonical variables $(x,y) \in \mathbf{R}^{2n}$. The stability of the equilibrium $x = y = 0$ for the system (3) is a trivial matter if all frequencies ω have the same sign, e.g., they are all positive. For, in this case the classical Lyapounov's theory applies since the Hamiltonian has a minimum at the origin. This simple argument does not apply if the frequencies do not vanish but have different signs.

The stability over long times has been investigated by Birkhoff using the method of normal form going back to Poincaré (see [26], tome II, chapt. IX, § 125). Assuming that there are no resonance relations among the frequencies ω, via a near the identity canonical transformation $(x,y) \to (x',y')$ the Hamiltonian is given the normal form up to a finite order $r > 2$

$$H^{(r)}(x',y') = \frac{1}{2}\sum_{l=1}^{n} \omega_l p'_l + Z^{(r)}(p') + \mathcal{R}^{(r)}(x',y')\ , \tag{4}$$

where $p'_l = ({x'}_l^2 + {y'}_l^2)/2$ are the new actions, $Z^{(r)}$ is at least quadratic in p' and the unnormalized remainder $\mathcal{R}^{(r)}$ is a power series starting with terms of degree $r+1$ in x', y'. If we forget the remainder then the system is integrable and the motion is quasiperiodic on invariant tori, since $Z^{(r)}$ depends only on the new actions. Birkhoff's remark was that the normalized Hamiltonian $H^{(r)}$ is convergent in a

neighbourhood of the origin, e.g., in some polydisk of radius ϱ (that may depend on r) and center at the origin, i.e.,

$$\Delta_\varrho = \left\{ (x,y) \in \mathbf{R}^{2n} \,:\, \sqrt{x_j^2 + y_j^2} < \varrho \right\} \,. \tag{5}$$

Hence, the size of the remainder may be estimated by $C_r \varrho^{r+1}$, with some constant C_r that Birkhoff did not try to evaluate. He concluded that the dynamics given by the integrable part of the Hamiltonian is a good approximation of the true dynamics up to a time of order $O(\varrho^{-r})$; on this remark he based his theory of *complete stability* (see [8], chapt. IV, § 2 and § 4).

It was pointed out by Poincaré that the series produced by perturbation expansions have an asymptotic character (see [26], tome II, chapt. VIII). Now this fact lies at the basis of the exponential stability. Indeed the constant C_r is expected to grow at least as $O(r!)$, so that the size of the remainder is $O\big(r!\varrho^{r+1}\big)$. Having fixed ϱ (i.e., the domain of the initial data) one chooses $r \sim 1/\varrho$, and by a straightforward use of Stirling's formula one gets $|\mathcal{R}| = O\big(\exp(-1/\varrho)\big)$. By working out the analytical estimates one gets for the unperturbed actions $p_l = (x_l^2 + y_l^2)/2$ the following bound (see [13] or [12]):

THEOREM: *Let the frequencies ω satisfy the diophantine condition*

$$\big|\langle k, \omega \rangle\big| \geq \gamma |k|^{-\tau} \quad \text{for } 0 \neq k \in \mathbf{Z}^n \,. \tag{6}$$

Then there exists a ϱ_* such that for all orbits satisfying $\big(x(0), y(0)\big) \in \Delta_\varrho$ one has

$$\big|p(t) - p(0)\big| = O(\varrho^3) \quad \text{for } |t| < T = O\big(\exp(1/\varrho^{1/(\tau+1)})\big) \,.$$

For a practical application the problem is that the estimated value of ϱ_* may be ridiculously small. A better evaluation may be obtained by explicitly calculating all series involved in the normalization process up to some (not too low) order. This just requires some elementary algebra on computer.

The Hamiltonian is expanded in power series as in (3) up to some order r, and then is given a normal form at the same order. The explicit transformation of coordinates and the new action variables p' as functions of the old coordinates can be constructed, too. Moreover, in a polydisk Δ_ϱ we may evaluate the quantity

$$D(\varrho, r) = \sup_{(x',y') \in \Delta_\varrho} \big|\dot{p}'\big| = \sup_{(x',y') \in \Delta_\varrho} \big|\big\{p', \mathcal{R}^{(r)}\big\}\big| \,;$$

to this end, the expression of the lowest order term of the remainder $\mathcal{R}^{(r)}$ may be used. Having fixed a polydisk Δ_{ϱ_0} containing the initial data we conclude that the orbit can not escape from a polydisk Δ_ϱ, with an arbitrary $\varrho > \varrho_0$, for $|t| < \tau(\varrho_0, \varrho, r)$, where

$$\tau(\varrho_0, \varrho, r) = \frac{\varrho^2 - \varrho_0^2}{2D(\varrho, r)} \,. \tag{7}$$

This produces an estimate depending on the arbitrary quantities ϱ and r. Let ϱ_0 and r be fixed; then, in view of $D(\varrho, r) \sim C_r \varrho^{r+1}$, the function $\tau(\varrho_0, \varrho, r)$, considered as function of ϱ only, has a maximum for some value ϱ_r. This looks quite odd, because one would expect τ to be an increasing function of ϱ. However, recall that (7) is just an estimate; looking for the maximum means only that we are trying to do the best use of our poor estimate. Let us now keep ϱ_0 constant, and calculate $\tau(\varrho_0, \varrho_r, r)$ for increasing values of $r = 1, 2, \dots$, with ϱ_r as above. Since C_r is expected to grow quite fast with r we expect to find a maximum of $\tau(\varrho_0, \varrho_r, r)$ for some optimal value $r_{\rm opt}$. Thus, we are authorized to conclude that *for every ϱ_0 we can explicitly evaluate the positive constants $\varrho(\varrho_0) = \varrho_{r_{\rm opt}}$ and $T(\varrho_0) = \tau(\varrho_0, \varrho(\varrho_0), r_{\rm opt})$ such that an orbit with initial point in the polydisk Δ_{ϱ_0} will not escape from Δ_ϱ for $|t| < T(\varrho_0)$.*

In order to show that the method above may be effective let me consider the triangular Lagrangian point L_4 of the restricted problem of three bodies, with particular reference to the Sun–Jupiter case. In the planar case the frequencies are $\omega_1 \sim 0.99676$ and $\omega_2 \sim -0.80464 \times 10^{-1}$; hence the standard Lyapounov theory does not apply.

The procedure above has been worked out by expanding all functions in power series up to order 35. One may look in particular for a value of ϱ_0 such that $T(\varrho_0)$ is the estimated age of the universe. The result is that ϱ_0 is roughly 0.127 times the distance L_4–Jupiter; this is certainly a realistic result. A comparison with the known Trojan asteroids shows that four of them are inside the region which assures stability for the age of the universe (see [16] for a complete report).

3. On the conjecture of Boltzmann and Jeans

Let us consider a canonical system with analytic Hamiltonian

$$H(p, x, \pi, \xi) = \hat{h}(p, x) + h_\omega(\pi, \xi) + f(p, x, \pi, \xi) \ , \tag{8}$$

where

$$h_\omega(\pi, \xi) = \frac{1}{2} \sum_{l=1}^{\nu} \left(\pi_l^2 + \omega_l^2 \xi_l^2 \right) \ , \quad (\pi, \xi) \in \mathbf{R}^{2\nu}$$

is the Hamiltonian of a system of harmonic oscillators, $\hat{h}(p, x)$ is the Hamiltonian of a generic n–dimensional system, and $f(p, x, \pi, \xi)$ a coupling term which is assumed to be of order ξ, and so to vanish for $\xi = 0$.

This model was suggested by the numerical study of the system of diatomic molecules mentioned in sect. 1 (see [5] and [6]). In that case $\hat{h}(p, x)$ represents the translational degrees of freedom, and $h_\omega(\pi, \xi)$ describes the internal vibrations of the molecules. Since the molecules are identical, all frequencies coincide.

The identification of a perturbation parameter in the system (8) goes as follows. Write $\omega = \lambda\Omega$ with large λ and Ω of the same order of the inverse of a typical time scale of the constrained system (for example the characteristic time for the collision of two molecules, which is non zero if the interaction potential is regular); then transform the variables according to $\pi = \pi' \sqrt{\lambda\Omega}$ and $\xi = \xi' / \sqrt{\lambda\Omega}$,

and assume the total energy of the subsystem h_ω to be finite, so that the variables (π', ξ') turn out to be confined in a disk of size $1/\sqrt{\lambda}$. Then the Hamiltonian may be given the form, omitting primes,

$$H(p, x, \pi, \xi, \lambda) = \hat{h}(p, x) + \lambda h_\Omega(\pi, \xi) + \frac{1}{\lambda} f_\lambda(p, x, \pi, \xi)$$
$$h_\Omega(\pi, \xi) = \frac{1}{2} \sum_{l=1}^{\nu} \Omega_l \left(\pi_l^2 + \xi_l^2\right)$$

(here, a straightforward computation would give $\lambda^{-1/2}$ in front of f, but f itself turns out to be of order $\lambda^{-1/2}$, since it vanishes for $\xi = 0$). Here too the main technical tool is the reduction of the Hamiltonian to a normal form. Precisely, via a near to identity canonical transformation $(p, x, \pi, \xi) \to (p', x', \pi', \xi')$ the Hamiltonian is given the form

$$H'(p', x', \pi', \xi', \lambda) = \lambda h_\Omega(\pi', \xi') + \hat{h}(p', x') + Z(p', x', \pi', \xi', \lambda) + \mathcal{R}(p', x', \pi', \xi', \lambda) \ ,$$

where Z is in normal form in the sense that $\{h_\Omega, Z\} = 0$. Thus h_Ω is an approximate first integral. The normalization process is performed until the remainder is exponentially small in the parameter $1/\lambda$. This requires an optimal choice of the number of normalization steps, as in the case of the elliptic equilibrium.

THEOREM: *Assume that all frequencies ω are equal. Then there are positive constants T_* and λ_* such that for every $\lambda > \lambda_*$ one has*

$$\begin{aligned} \left|h_\Omega(\pi, \xi) - h_\Omega(\pi', \xi')\right| &= O(\lambda^{-2}) \ ; \\ \left|h_\Omega(t) - h_\Omega(0)\right| &= O(\lambda^{-1}) \quad \text{for } |t| < T_* \exp\left(\frac{\lambda}{\lambda_*}\right) \ . \end{aligned} \tag{9}$$

The remarkable fact is that the exponent a that appears in the general form (2) of the exponential estimate is 1, no matter of the number n of degrees of freedom. This removes the worst dependence on n, and is in complete agreement with the numerical calculations in [5].

In the case of the diatomic gas there is still a dependence on n in the constants T_* and λ_*, which turn out to be $O(1/n^2)$ (the number of two–body interaction terms in the perturbation). Such a dependence could hardly be removed on a purely dynamical basis, because the possibility that all molecules collide together at all times may not be excluded. This is clearly unrealistic. A complete proof of the conjecture of Boltzmann and Jeans could perhaps be obtained by complementing the dynamical theory with statistical considerations.

The result above has been extended to further situations, including the case of infinite systems. As an example, consider a modification of the celebrated nonlinear chain of Fermi, Pasta and Ulam[11] in which the equal masses are replaced by alternating heavy and light masses. It is known that the spectrum splits into two well separated branches, called the acoustical and the optical one. Moreover the optical frequencies are very close to each other. The whole system may thus

be considered as composed of two separate subsystems, and the subsystem h_ω of the optical frequencies may still be considered as a system of oscillators with the same frequency: the small difference can consistently be considered as part of the perturbation. In this case it has been proven that the exponential estimate applies also to the case of an infinite chain, provided the *total* energy is sufficiently small (see [2]). Strictly speaking, this is not enough for the application to the problem of equipartition of energy in statistical mechanics, since in that case one is interested in initial data with fixed *specific* energy. However, the discrepancy is still due to the fact that we are working on a purely dynamical basis. For, the possibility that the whole energy of the optical subsystem remains concentrated on a single oscillator for a long time is not excluded. Here too one should include statistical considerations.

4. Superexponential stability

Let us go back to considering the Hamiltonian (1). I will need to consider the action variables in a domain $\mathcal{G}_\varrho = \bigcup_{p\in\mathcal{G}} B_\varrho(p)$, where ϱ is a positive parameter, $\mathcal{G} \subset \mathbf{R}^n$ is open, and $B_\varrho(p)$ denotes the open ball of radius ϱ and center p. The phase space is $\mathcal{D} = \mathcal{G}_\varrho \times \mathbf{T}^n$.

If the unperturbed Hamiltonian $h(p)$ is non degenerate, then the construction of the normal form for the Hamiltonian can not be performed globally on the action domain $\mathcal{G}_\varrho$. For, the small denominators $\langle k, \omega(p)\rangle$ (with $k \in \mathbf{Z}^n$ and $\omega(p) = \frac{\partial h}{\partial p}$) may generically vanish in a set of points that is dense in $\mathcal{G}_\varrho$. This fact lies at the basis of Poincaré's proof of nonexistence of uniform first integrals (see [26], chapt. V).

The way out of this problem is based on: (a) a Fourier cutoff of the perturbation, i.e., only a finite number of Fourier modes is considered during the process of normalizing the Hamiltonian, and (b) the construction of the normal form in local nonresonance domains where the small denominators are far enough from zero. The burden of constructing the nonresonance domains is taken by the so called *geometric part* of the proof of Nekhoroshev's theorem: basically, the original domain $\mathcal{G}_\varrho$ is covered by subdomains corresponding to *known* resonances of different multiplicity $0, 1, \ldots, n$, where multiplicity zero corresponds to the region free from resonances. The domains so constructed are open because only a *finite* number of resonances is taken into account; this is a consequence of the Fourier cutoff. The normal form is *local* to each domain, and depends on the resonances that appear on it. Nekhoroshev's theorem on exponential stability follows by proving that every orbit is confined inside a local nonresonance domain for an exponentially long time.

The result that I'm going to illustrate is based on iteration of Nekhoroshev's theorem. Let me first state the result. Let φ^t be the canonical flow generated by the Hamiltonian (1). A n–dimensional torus $\mathcal{T}$ will be said to be (η, T)–*stable* in case one has $\mathrm{dist}(\varphi^t P, \mathcal{T}) < \eta$ for all $|t| < T$ and for every $P \in \mathcal{T}$. The formal statement is the following

THEOREM: *Consider the Hamiltonian (1), and assume that the unperturbed Hamiltonian $h(p)$ is convex. Then there exists $\varepsilon^* > 0$ such that for all $\varepsilon < \varepsilon^*$ the following statement holds true: there is a sequence $\{\mathcal{D}^{(r)}\}_{r\geq 0}$ of subsets of $\mathcal{D}$, with $\mathcal{D}^{(0)} = \mathcal{D}$, and two sequences $\{\varepsilon_r\}_{r\geq 0}$ and $\{\varrho_r\}_{r\geq 1}$ of positive numbers satisfying*

$$\varepsilon_0 = \varepsilon \ , \quad \varepsilon_r = O\big(\exp(-1/\varepsilon_{r-1})\big) \ ,$$
$$\varrho_0 = \varrho \ , \quad \varrho_r = O(\varepsilon_{r-1}^{1/4}) \ ,$$

such that for every $r \geq 0$ one has:

(i) $\mathcal{D}^{(r+1)} \subset \mathcal{D}^{(r)}$;

(ii) $\mathcal{D}^{(r)}$ *is a set of n–dimensional tori diffeomorphic to* $\mathcal{G}^{(r)}_{\varrho_r} \times \mathbf{T}^n$;

(iii) $\mathrm{Vol}(\mathcal{D}^{(r+1)}) > (1 - O(\varepsilon_r^a))\,\mathrm{Vol}(\mathcal{D}^{(r)})$ *for some positive $a < 1$;*

(iv) $\mathcal{D}^{(\infty)} = \bigcap_r \mathcal{D}^{(r)}$ *is a set of invariant tori for the flow φ^t, and moreover one has* $\mathrm{Vol}(\mathcal{D}^{(\infty)}) > (1 - O(\varepsilon_0^a))\,\mathrm{Vol}(\mathcal{D}^{(0)})$;

(v) for every $p^{(r)} \in \mathcal{G}^{(r)}$ *the torus* $p^{(r)} \times \mathbf{T}^n \subset \mathcal{D}^{(r)}$ *is* $(\varrho_{r+1}, 1/\varepsilon_{r+1})$*–stable;*

(vi) for every point $p^{(r)} \in \mathcal{G}^{(r)}$ *there exists an invariant torus* $\mathcal{T} \subset B_{\varrho_r}(p^{(r)}) \times \mathbf{T}^n$.

Let me illustrate the main points of the proof (for a complete proof see [15]). A careful reading of the geometric part of Nekhoroshev's theorem allows one to extract the following information: there exists a subset $\mathcal{D}^{(1)}$ of phase space characterized by absence of resonances of order smaller than $O(1/\varepsilon)$; such a domain is the union of open balls of positive radius ϱ_1, and its complement has measure $O(\varepsilon^{1/4})$. Moreover, in this subset one may introduce new action–angle variables, (p', q') say, which give the Hamiltonian the original form (1), but with a perturbation of size $\varepsilon_1 = O\big(\exp(-1/\varepsilon)\big)$.

Nekhoroshev's theorem can be applied again to the new Hamiltonian in the open domain $\mathcal{D}^{(1)}$, thus allowing one to construct a second nonresonant domain $\mathcal{D}^{(2)}$ characterized by absence of resonances of order smaller than $O(1/\varepsilon_1) = O\big(\exp(-1/\varepsilon)\big)$. Such a procedure can be iterated infinitely many times, and this gives the sequence $\mathcal{D}^{(r)}$ of subdomains of phase space, the existence of which is stated in the theorem. Nekhoroshev's stability estimates hold in every such domain, with stability times exponentially increasing at every step.

The sequence $\mathcal{D}^{(r)}$ of domains converges to a set $\mathcal{D}^{(\infty)}$ of invariant tori. This part of the proof is just an adaptation of Arnold's proof of KAM theorem and the set of invariant tori so obtained is similar to Arnold's one.

Let me finally explain how superexponential stability arises. Properties (v) and (vi) imply that every $(\varrho^{r+1}, 1/\varepsilon_{r+1})$–stable torus is ϱ_r–close to an invariant torus. In view of the form of the sequences ϱ_r and ε_r given in the statement of our theorem one has

$$\varepsilon_{r+1} = O(1/\exp(1/\varepsilon_r)) = O(1/\exp(\exp(1/\varepsilon_{r-1}))) = O(1/\exp(\exp(1/\varrho_r))) \ .$$

In view of this remark we may say that *in the neighbourhood of an invariant torus the natural perturbation parameter is the distance ϱ from the torus, and the diffusion speed is bounded by a superexponential of the inverse of the distance from an invariant torus.*

Acknowledgements. My interest in problems related to classical perturbation theory for Hamiltonian systems and to its applications to the problem of statistical mechanics arose from a seminar of L. Galgani, whom I consider as a master and a friend. Most of the results discussed here are the fruit of a long collaboration with him and G. Benettin, and, more recently, with D. Bambusi and A. Morbidelli. I want to express here my deep gratitude to all of them.

References

[1] Arnold, V. I.: *Small denominators and problems of stability of motion in classical and celestial mechanics*, Usp. Math. Nauk **18** N.6, 91 (1963); Russ. Math. Surv. **18** N.6, 85 (1963).

[2] Bambusi, D. and Giorgilli, A.: *Exponential stability of states close to resonance in infinite dimensional Hamiltonian Systems,* J. Stat. Phys, **71**, 569 (1993).

[3] Benettin, G., Galgani, L. and Giorgilli, A.: *A proof of Nekhoroshev's theorem for the stability times in nearly integrable Hamiltonian systems.* Cel. Mech., **37**, 1–25 (1985).

[4] Benettin, G. and Gallavotti, G.: *Stability of motions near resonances in quasi–integrable Hamiltonian systems.* Journ. Stat. Phys., **44**, 293 (1986).

[5] Benettin, G., Galgani, L. and Giorgilli, A.: *Exponential law for the equipartition times among translational and vibrational degrees of freedom*, Phys. Lett. A **120**, 23–27 (1987).

[6] Benettin, G., Galgani, L. and Giorgilli, A.: *Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory, part II*, Comm. Math. Phys, **121**, 557–601 (1989).

[7] Benettin, G., Carati, A. and Gallavotti, G.: *A rigorous implementation of the Jeans Landau Teller approximation for adiabatic invariants*, Nonlinearity **10**, 479–505 (1997).

[8] Birkhoff, G. D.: *Dynamical systems*, New York (1927).

[9] Boltzmann, L.: Nature **51**, 413–415 (1895).

[10] Carati, A. and Galgani, L.: *Planck's formula in classical mechanics*, preprint (1998).

[11] Fermi, E., Pasta, J. and Ulam, S.: *Studies of nonlinear problems*, Los Alamos document LA–1940 (1955).

[12] Giorgilli, A.: *Rigorous results on the power expansions for the integrals of a Hamiltonian system near an elliptic equilibrium point*, Ann. Ist. H. Poincaré, **48**, 423–439 (1988).

[13] Giorgilli, A., Delshams, A., Fontich, E., Galgani, L. and Simó, C.: *Effective stability for a Hamiltonian system near an elliptic equilibrium point, with an application to the restricted three body problem.* J. Diff. Eqs., **20**, (1989).

[14] Giorgilli, A. and Zehnder, E.: *Exponential stability for time dependent potentials*, ZAMP (1992).

[15] Giorgilli, A. and Morbidelli, A.: *Invariant KAM tori and global stability for Hamiltonian systems*, ZAMP **48**, 102–134 (1997).
[16] Giorgilli, A. and Skokos, Ch.: *On the stability of the Trojan asteroids*, Astron. Astroph. **317**, 254–261 (1997).
[17] Jeans, J.H.: *On the vibrations set up in molecules by collisions*, Phil. Magazine **6**, 279 (1903).
[18] Kolmogorov, A. N.: *Preservation of conditionally periodic movements with small change in the Hamilton function*, Dokl. Akad. Nauk SSSR, **98**, 527 (1954).
[19] Littlewood, J.E.: *On the equilateral configuration in the restricted problem of three bodies*, Proc. London Math. Soc.(3) **9**, 343–372 (1959).
[20] Littlewood, J.E.: *The Lagrange configuration in celestial mechanics*, Proc. London Math. Soc.(3) **9**, 525–543 (1959).
[21] Lochak, P.: *Canonical perturbation theory via simultaneous approximations*, Usp. Math. Nauk. **47**, 59–140 (1992). English transl. in Russ. Math. Surv.
[22] Moser, J.: *Stabilitätsverhalten kanonischer differentialgleichungssysteme,* Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl IIa, nr.6, 87–120 (1955).
[23] Moser, J.: *On invariant curves of area–preserving mappings of an annulus*, Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl II, 1–20 (1962).
[24] Nekhoroshev, N. N.: *Exponential estimates of the stability time of near–integrable Hamiltonian systems.* Russ. Math. Surveys, **32**, 1 (1977).
[25] Nekhoroshev, N. N.: *Exponential estimates of the stability time of near–integrable Hamiltonian systems, 2.* Trudy Sem. Petrovs., **5**, 5 (1979).
[26] Poincaré, H.: *Les méthodes nouvelles de la mécanique céleste*, Gauthier–Villars, Paris (1892).

Antonio Giorgilli
Dipartimento di Matematica
Via Saldini 50
20133 MILANO (Italy)
antonio.giorgilli@mi.infn.it

Stability of Matter in Classical and Quantized Fields

Gian Michele Graf

Abstract. In recent years considerable activity was directed to the issue of stability in the case of matter interacting with an *electromagnetic field*. We shall review the results which have been established by various groups, in different settings: relativistic or non-relativistic matter, classical or quantized electromagnetic fields. Common to all of them is the fact that electrons interact with the field both through their charges *and* the magnetic moments associated to their spin. Stability of non-relativistic matter in presence of magnetic fields requires that $Z\alpha^2$ (where Z is the largest nuclear charge in the system) as well as the fine structure constant α itself, do not exceed some critical value. If one imposes an ultraviolet cutoff to the field, as it occurs in unrenormalized quantum electrodynamics, then stability no longer implies a bound on α, $Z\alpha^2$. An important tool is given by Lieb–Thirring type inequalities for the sum of the eigenvalues of a one–particle Pauli operator with an arbitrary inhomogeneous magnetic field.

1991 Mathematics Subject Classification: 81-02
Keywords and Phrases: Stability of matter

Introduction

Ordinary matter consists of molecules and atoms which are largely empty inside. Yet matter does not shrink. A related — and more fundamental — aspect of stability is the fact that the energy per particle is bounded below, independently of the number of particles. This is what is usually referred to as stability of matter. It should be stressed that it goes well beyond the stability of individual atoms. Basic thermodynamic properties such as extensivity (e.g., two moles of water occupy with good approximation twice the volume occupied by a single mole) also depend on this property. These topics are reviewed in [19, 20].

Stability of matter could not hold without quantum mechanics and, in particular, without the uncertainty principle, but the Pauli principle and screening properties of the interaction (Coulomb) potential are equally important (see [34] for the consequences of tampering with these tenets). The first instance where stability was established, by Dyson and Lenard [9], is non-relativistic matter consisting of N electrons which move in the field of M nuclei having fixed but arbitrary positions. We denote by $q_i = -1$, resp. $q_i = Z$, the charge of an electron

$(i = 1, \dots, N)$, resp. of a nucleus $(i = N+1, \dots, N+M)$. According to the Pauli principle a (pure) state of the N electrons is given by a normalized wave function

$$\Psi \in \bigwedge_{i=1}^{N} \mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^2) \tag{1}$$

in the N-fold antisymmetric tensor product of the single particle Hilbert space $\mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^2)$. Here, $\mathbb{C}^2$ accounts for the spin of the electron, whose role is however unessential so far. The Hamiltonian is, in appropriate units,

$$H = \sum_{i=1}^{N} t_i + V_{\mathrm{c}} \ , \tag{2}$$

where the kinetic energy of a single electron is $t = p^2$, $p = -\mathrm{i}\nabla$ and the index i refers to the variables of the i-th electron. The Coulomb potential V_{c} is

$$V_{\mathrm{c}} = \sum_{i<j} \frac{q_i q_j}{|x_i - x_j|} \ .$$

Theorem 1. *There is a constant $C(Z)$ independent of the position of the nuclei, such that*

$$H \geq -C(Z)(N+M) \ . \tag{3}$$

Subsequently, Lieb and Thirring [27] obtained a much better constant $C(Z)$ which is of order unity for $Z \approx 1$. They also provided a simpler proof, thereby linking (3) to stability of Thomas-Fermi theory. (See however [17] for a short proof closer in spirit to [9]).

In recent years considerable activity was directed to the issue of stability in the case of matter interacting with an electromagnetic field, which brings the model closer to physical reality. Results have been established by various groups, in different settings: relativistic or non-relativistic matter, classical or quantized electromagnetic fields.

Stability and instability in classical magnetic fields

To begin with, consider the addition of a classical, external magnetic field $B = \nabla \wedge A$. There, stability — uniformly in the magnetic vector potential A — persists [1, 7] if the field is included through minimal substitution, i.e., for $t = D^2$, $D = p + A$. This follows by means of the diamagnetic inequality. To actually describe matter in magnetic fields one must however also add the interaction of the electrons with the field through their spins or, more precisely, through the associated magnetic moments. The corresponding kinetic energy is

$$t = D^2 + \frac{g}{2} B \cdot \sigma \ ,$$

where $\sigma = (\sigma_1, \sigma_2, \sigma_3)$ are the Pauli matrices and g is known as the gyromagnetic factor. Its physical value is $g = 2$, as long as radiative corrections from quantum electrodynamics are neglected. Stability (3) extends straightforwardly to any $g < 2$, while for $g > 2$ the Hamiltonian is not even bounded below. In the critical case $g = 2$, to which we shall henceforth restrict, the kinetic energy may be written as

$$t = D^2 + B \cdot \sigma = \not{D}^2 , \qquad \not{D} = D \cdot \sigma .$$

Dynamical spins confer new aspects to the issue of stability. A first indication of this is the following: Whereas the equation $D\psi = 0$ admits (by the uncertainty principle) only $\psi = 0$ as a solution in $\mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^2)$, there exist [30] field configurations A such that $\not{D}\psi = 0$ has non-trivial solutions called zero-modes. This effectively invalidates the uncertainty principle and, as a result, stability as defined above. To see this, just consider the case $N = M = 1$ with Hamiltonian

$$H_A = \not{D}_A^2 - Z|x|^{-1} .$$

By scaling both the field and its zero-mode,

$$A_\lambda(x) = \lambda^{-1} A(x/\lambda) , \qquad \psi_\lambda(x) = \lambda^{-3/2} \psi(x/\lambda) , \tag{4}$$

we obtain $\not{D}_{A_\lambda} \psi_\lambda = 0$ and

$$(\psi_\lambda, H_{A_\lambda} \psi_\lambda) = -Z\lambda^{-1} (\psi, |x|^{-1} \psi) , \tag{5}$$

which can be made arbitrarily large and negative by letting $\lambda \to 0$.

However, a proper formulation of stability should incorporate the field energy

$$H_{\mathrm{cf}} = \frac{1}{8\pi\alpha^2} \int B(x)^2 d^3x \tag{6}$$

into the Hamiltonian:

$$H = \sum_{i=1}^{N} t_i + V_{\mathrm{c}} + H_{\mathrm{cf}} . \tag{7}$$

Here $\alpha > 0$ is the fine structure constant. The physical value of this dimensionless parameter is $\alpha = e^2/\hbar c \approx 1/137$. Note that under (4) the magnetic field scales as $B_\lambda(x) = \lambda^{-2} B(x/\lambda)$, so that H_{cf} scales as λ^{-1}, just as the Coulomb energy (5). Thus already from the case $N = M = 1$ one sees that stability for (7) may hold only if $Z\alpha^2$ is sufficiently small. Another necessary condition is that α itself be small enough. To see the latter, consider $N = 1$ and M large. As above, let the electron be in a zero-mode of a fixed field A. Distribute the many nuclei according to some limiting density, e.g., uniformly over a ball. The repulsion energy between the nuclei is $\leq C_1 (ZM)^2$, and the attraction of the electron $\leq -C_2(ZM)$, with C_1, $C_2 > 0$ independent of Z, M. By minimizing the sum of the two bounds we obtain $(\psi, V_{\mathrm{c}} \psi) \leq -C_2^2/4C_1$ for $ZM = C_2/2C_1$. Thus,

$$(\psi, H\psi) \leq -\frac{C_2^2}{4C_1} + \frac{1}{8\pi\alpha^2} \int B(x)^2 d^3x < 0$$

for α large enough. Since both the Coulomb and the field energy scale the same way, the expectation value of the Hamiltonian can in fact be made arbitrarily large and negative. The above two conditions are in fact sufficient for stability:

Theorem 2. *The Hamiltonian (7) is stable, i.e.,*

$$H \geq -C(N+M) \,,$$

provided α and $Z\alpha^2$ are small enough.

The theorem was first established by Fefferman [12], for $Z = 1$. Soon thereafter, Lieb, Loss, and Solovej [23] found a simpler proof which furthermore ensures stability at physical values of the parameters Z, α and produces a realistic lower bound $-C$ on the energy per particle. An additional improvement of Lieb, Siedentop and Solovej [24, 25] and Loss [29] yields the following sufficient condition for stability:

$$\frac{\pi}{2} Z + 2.7919 Z^{2/3} + 1.2987 \leq 0.2153 \alpha^{-2} \,. \tag{8}$$

In particular, for $\alpha = 1/137$ stability holds if $Z \leq 2264$. Precursors of Theorem 2 are found in [16, 21], where the cases $N = 1$ and $M = 1$, resp. $N = 1$ or $M = 1$, were proved.

Let us present the proof of Theorem 2 given in [25], but for brevity we shall not keep track of best constants. The stability of (7) is brought into relation with stability of an apparently unrelated Hamiltonian $H_{\rm rel}$, namely that of relativistic matter without dynamical spins. It is defined by (2), but with $t = \alpha^{-1}|D|$. The corresponding stability result was proven in [8, 15, 28, 22].

Theorem 3.

$$H_{\rm rel} \geq 0 \,, \tag{9}$$

if α and $Z\alpha$ are sufficiently small.

Note that $H_{\rm rel}$ can be uniformly bounded below only if it is non-negative, since both its terms scale as λ^{-1}. Explicitly, stability is assured [22] if the l.h.s. of (8) does not exceed α^{-1}. On the other hand, $H_{\rm rel}$ is unbounded below [18] if $Z\alpha > 2/\pi$.

The other ingredients of the proof of Theorem 2 are:

- The Birman-Koplienko-Solomyak inequality [3]: For any operators A, $B \geq 0$,

$$\mathrm{tr}(A-B)_+ \leq \mathrm{tr}(A^2 - B^2)_+^{1/2} \,, \tag{10}$$

  where $s_+ = \max(s, 0)$, provided the operator on the r.h.s. is trace class.
- The Lieb-Thirring estimate [27]:

$$\mathrm{tr}(-h)_+^\gamma \leq L_\gamma \int v(x)^{\gamma + \frac{3}{2}} d^3x \tag{11}$$

  for $\gamma \geq 0$ and any Schrödinger operator $h = D^2 - v$ on $\mathrm{L}^2(\mathbb{R}^3)$ with $v = v(x) \geq 0$. The l.h.s. can be written as $\sum_k |e_k|^\gamma$, where $e_k < 0$ are the negative eigenvalues of h.

Let us denote by $\tilde{\alpha}$ the fine structure constant in $H_{\rm rel}$, to avoid confusion. Using (9), the first two terms in (7), $H_{\rm m} = \sum_{i=1}^{N} \not{p}_i^2 + V_{\rm c}$, can be estimated as

$$H_{\rm m} \geq \sum_{i=1} (\not{p}^2 - \tilde{\alpha}^{-1}|D|)_i \geq -\operatorname{tr}(\tilde{\alpha}^{-1}|D| - 2\beta|\not{p}|)_+ - \beta^2 N \ ,$$

for any $\beta > 0$. Here we used $\not{p}^2 \geq 2\beta|\not{p}| - \beta^2$ and the Pauli principle. Now (10) can be used to bound the trace (setting $4\beta^2 = 2\tilde{\alpha}^{-2}$) as

$$\tilde{\alpha}^{-1} \operatorname{tr}(D^2 - 2\not{p}^2)_+^{1/2} = \tilde{\alpha}^{-1} \operatorname{tr}(-D^2 - 2B\cdot\sigma)_+^{1/2} \leq 2\tilde{\alpha}^{-1} L_{1/2} \int 4B(x)^2 d^3x \ ,$$

where, in the last step, we used $-B\cdot\sigma \leq |B|$ and (11). Summing up, one obtains

$$H = H_{\rm cf} + H_{\rm m} \geq \Big(\frac{1}{8\pi\alpha^2} - \frac{8L_{1/2}}{\tilde{\alpha}}\Big) \int B(x)^2 d^3x - \frac{1}{2}\tilde{\alpha}^{-2} N \ ,$$

showing that stability holds for $\alpha^2 \leq \tilde{\alpha}/(64\pi L_{1/2})$.

Finally, Lieb, Siedentop and Solovej [24, 25] considered relativistic matter with dynamical spins. The appropriate kinetic energy is given by the Dirac operator

$$t = D\cdot\alpha + \beta m$$

acting on $\mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^4)$, where $m \geq 0$ is the mass and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, β are the Dirac matrices. Except for this modification, the many-body Hamiltonian $H_{\rm Dirac}$ is still given by (7). Clearly $H_{\rm Dirac}$, just as t, is unbounded below, but the proper interpretation, going in essence back to Dirac, is 'to fill the Fermi sea' for t. In other words, one should only consider expectation values for $H_{\rm Dirac}$ in states

$$\Psi \in \bigwedge_{i=1}^{N} \mathfrak{h}_+ \ ,$$

where $\mathfrak{h}_+ \subset \mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^4)$ is the positive spectral subspace for t.

Theorem 4.

$$(\Psi, H_{\rm Dirac}\Psi) \geq 0$$

(uniformly also in $m \geq 0$), provided α and $Z\alpha$ are small enough.

For $\alpha = 1/137$ stability holds up to $Z \leq 56$. The proof is related to the one sketched above.

Stability and instability in quantized electromagnetic fields

We shall consider only the case of non-relativistic matter. The model is formally still defined by the Hamiltonian (7), but with the following changes. First, the Hilbert space now is $\mathcal{H} = \mathcal{H}_{\rm m} \otimes \mathcal{F}$, where $\mathcal{H}_{\rm m}$ is the Hilbert space (1) for matter and $\mathcal{F}$, the Hilbert space for the field, is the bosonic Fock space over $\mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^2)$.

Here, $\mathbb{C}^2$ accounts for the helicity of the photon. Second, the ultraviolet-cutoff electromagnetic vector potential in the Coulomb gauge is given by

$$A_\Lambda(x) = A_-(x) + A_-(x)^* \,, \quad A_-(x) = \frac{\alpha^{1/2}}{2\pi} \int_{|k|\le\Lambda} |k|^{-1/2} \sum_{\lambda=\pm} a_\lambda(k) e_\lambda(k) \mathrm{e}^{\mathrm{i}kx} d^3k \,,$$

where $\Lambda < \infty$ is the cutoff. For each k, the direction of propagation $\hat{k} = k/|k|$ and the polarizations $e_\pm(k) \in \mathbb{C}^3$ are orthonormal. The operators $a_\lambda(k)^*$ and $a_\lambda(k)$ are creation and annihilation operators on $\mathcal{F}$ and satisfy canonical commutation relations

$$[a_\lambda(k)^\#, a_{\lambda'}(k')^\#] = 0 \,, \qquad [a_\lambda(k), a_{\lambda'}(k')^*] = \delta_{\lambda\lambda'}\delta(k-k') \,.$$

The vacuum state $\Omega \in \mathcal{F}$, $(\Omega,\Omega) = 1$, is distinguished by $a_\lambda(k)\Omega = 0$, for all $k \in \mathbb{R}^3$. The kinetic energy in (7), $t = \not{D}^2$, is now defined with $D = p + A_\Lambda(x)$. Finally, the quantum field energy is

$$H_{\mathrm{qf}} = \alpha^{-1} \int |k| \sum_{\lambda=\pm} a_\lambda(k)^* a_\lambda(k) d^3k \,. \tag{12}$$

This completes the definition of the Hamiltonian, which we denote by H_Λ. To see how (12) relates to the previous definition (6), we introduce the (tranverse) electric field $E(x) = -\mathrm{i}[H_{\mathrm{qf}}, A_\Lambda(x)]$ and the magnetic field $B(x) = (\nabla \wedge A_\Lambda)(x)$. Then,

$$H_{\mathrm{qf}} = \frac{1}{8\pi\alpha^2} \int :E(x)^2 + B(x)^2: \, d^3x \,, \tag{13}$$

where $: \ldots :$ denotes Wick ordering; explicitly, $:B(x)^2: = B(x)^2 - (\Omega, B(x)^2\Omega)$, and analogously for $E(x)^2$. In contrast to (6), the integrand of (13) may also take negative (expectation) values.

Let us remark that the model represents, apart from the cutoff needed to make it well-defined, a physically correct description of the coupled system consisting of matter and field, since the Hamiltonian yields the correct equations of motion. The spectral theory of a similar model is discussed in [2].

The stability of Theorem 2 carries over to this situation [6, 5], but not with the same explicit bounds.

Theorem 5. *For any $\Lambda > 0$,*

$$H_\Lambda \ge -C(\alpha, Z, \Lambda)(N + M) \,, \tag{14}$$

for small enough α, $Z\alpha^2$, with $C(\alpha, Z, \Lambda) = \mathrm{const} \cdot \tilde{Z} \max(\tilde{Z}, \alpha^{1/4}\Lambda)$ and $\tilde{Z} = Z+1$.

Actually, the ultraviolet cutoff prevents the instability explained before Theorem 2. As a result, the restriction to small values of α, $Z\alpha^2$ may be dropped, as shown by Fefferman [13] and Fefferman, Fröhlich and Graf [14]:

Theorem 5'. *For any α, Z, Λ, the estimate (14) holds with $C(\alpha, Z, \Lambda) = \mathrm{const} \cdot \tilde{Z}(1 + \beta^5 \log\beta)(\beta^{-2}\tilde{Z} + \Lambda)$ with $\beta = \tilde{Z}\alpha^2 + 1$.*

This fact is not of direct physical significance, however. Rather, one should consider a renormalized Hamiltonian

$$H_{\Lambda,\mathrm{ren}} = \sum_{i=1}^{N} m_\Lambda^{-1} \not{p}_i^2 + V_\mathrm{c} + H_\mathrm{qf} - \mu_\Lambda N \ , \tag{15}$$

where the mass m_Λ and the chemical potential μ_Λ are to be chosen so that the energy of a one electron state with small total momentum p is p^2. It appears conceivable that stability for (15) holds uniformly in Λ, for small enough α, $Z\alpha^2$.

The proof of Theorem 5 can be reduced to stability statements for matter in classical, external fields [12, 4], but with a different expression for the field energy H_cf than before. For reasons related to the vacuum energy subtraction mentioned above, the classical field energy (6) should be replaced by

$$H_\mathrm{cf} = \frac{1}{8\pi\alpha^2} \int_U B(x)^2 d^3x \ , \tag{16}$$

where the integration is now restricted to a small neighborhood U of the nuclei. A similar expression [13, 5], involving also the field gradient, occurs in the proof of Theorem 5'.

Magnetic Lieb-Thirring type inequalities

An issue of related, but also independent interest is found in Lieb-Thirring inequalities corresponding to (11) for Pauli, rather than Schrödinger, Hamiltonians, i.e., for $h = \not{p}^2 - v$ on $\mathrm{L}^2(\mathbb{R}^3, \mathbb{C}^2)$. (We shall focus on $\gamma = 1$, corresponding to the sum of the negative eigenvalues of h). The first such estimate, by Lieb, Solovej and Yngvason [26] applies to constant magnetic fields $B(x) = B$.

Theorem 6. *For constant fields,*

$$\sum_k |e_k| \le a_\delta \int v(x)^{5/2} d^3x + b_\delta |B| \int v(x)^{3/2} d^3x \ , \tag{17}$$

for any $0 < \delta < 1$, *with* $a_\delta = 0.3119\,\delta^{-2}$ *and* $b_\delta = 0.2123(1-\delta)^{-1}$.

The second term represents the contribution of the lowest Landau level, i.e., of the lowest (degenerate) eigenvalue of $\not{p}^2$, whereas the higher levels are accounted for by the familiar first term. Note that a generalization to arbitrary non-constant fields cannot be obtained by just pulling $|B(x)|$ in (17) under the integral sign. Such a bound would be too small (for small v), since, due to the possible existence of zero-modes $\not{p}\psi = 0$, the bound has to be at least $(\psi, v\psi)$.

Estimates for non-constant fields are due to Erdős [10], followed by [23, 32, 33, 4, 5, 31]. Some of them are useful in proofs of stability of matter. In this context we mention the bound of Lieb, Loss and Solovej [23]:

Theorem 7.

$$\sum_k |e_k| \le a_\delta \int v(x)^{5/2} d^3x + b_\delta \Big(\int B(x)^2 d^3x \Big)^{3/4} \Big(\int v(x)^4 d^3x \Big)^{1/4} \ , \tag{18}$$

for any $0 < \delta < 1$, *with* $a_\delta = 0.0654\,\delta^{-1}$ *and* $b_\delta = 0.1005\,\delta^{-5/8}(1-\delta)^{-3/8}$.

One may be tempted to believe that the second term could be replaced by $\int |B(x)|^{3/2} v(x)\, d^3x$, which would imply (18) by Hölder's inequality. It turns out — essentially by arguments of Erdős [10] — that this is not true: The interplay between the field $B(x)$ and the potential $v(x)$ is not strictly local. It is however possible to define an effective scalar field $b(x) \geq 0$ which allows for a semi–local version of (18). This is of interest in connection with the definition (16) and is the content of the following result of Bugliaro et al. [4]:

Theorem 8.

$$\sum_k |e_k| \leq C' \int v(x)^{5/2} d^3x + C'' \int b(x)^{3/2} v(x) d^3x \,, \tag{19}$$

$$\int b(x)^2 d^3x \leq C \int B(x)^2 d^3x \,. \tag{20}$$

In particular, the two estimates together imply (18), except for the constants. The construction of $b(x)$ can be explained as follows. The interplay between the field B and V takes place on a length scale $r(x)$ which depends on B itself (see below), and $b(x)^2$ is the average of $B(y)^2$ over that length scale:

$$b(x)^2 = \int r(x)^{-3} \varphi\Big(\frac{y-x}{r(x)}\Big) B(y)^2 d^3y \,,$$

with appropriate decay of $\varphi(z) \geq 0$ as $|z| \to \infty$. To determine $r(x)$, note that in the constant field case it is proportional to $|B|^{-1/2}$, the radius of a Landau orbit in the lowest Landau level. In the general case, it is determined self-consistently as $r(x) = b(x)^{-1/2}$. A different definition of $b(x)$ due to Sobolev [32, 33], which motivated the one just presented, also implies (19), but not (20).

Yet another generalization of (17) aims at estimating the contributions of the field gradient $\nabla \otimes B = (\partial_i B_j)_{i,j=1,2,3}$. This was done by Erdős and Solovej [11] and, under somewhat different conditions, by Bugliaro, Fefferman and Graf [5]. To this end a length scale $l(x)$ is introduced which is related to $\nabla \otimes B$ in a similar way as $r(x)$ is related to B.

Theorem 9.

$$\sum e_i \leq C' \int V(x)^{3/2} (V(x) + \widehat{B}(x)) d^3x + C'' \int V(x) P(x)^{1/2} (P(x) + \widehat{B}(x)) d^3x \,,$$

where $\widehat{B}(x)$ *is the average of* $|B(y)|$ *over a ball of radius* $l(x)$ *centered at* x, *and* $P(x) = l(x)^{-1}(r(x)^{-1} + l(x)^{-1})$.

By the variational principle, this estimate implies a bound on the density $n(x) = \sum_j |\psi_j(x)|^2$ of orthonormal zero-modes ψ_j of $\not{D}$. The bound is

$$n(x) \leq C'' P(x)^{1/2} (P(x) + \widehat{B}(x)) \,,$$

and, as it should, it vanishes in the case of a homogeneous magnetic field.

References

[1] J. Avron, I.Herbst, B. Simon, Schrödinger operators with magnetic fields: I. General interactions. Duke Math. J. 45, 847-883 (1978).

[2] V. Bach, J. Fröhlich, I. M. Sigal, Mathematical theory of nonrelativistic matter and radiation. Lett. Math. Phys. 34, 183–201 (1995).

[3] M. S. Birman, L. S. Koplienko, M. Z. Solomyak, Estimates for the spectrum of the difference between fractional powers of two-selfadjoint operators. Soviet Mathematics 19, 1-6 (1975).

[4] L. Bugliaro, C. Fefferman, J. Fröhlich, G. M. Graf, J. Stubbe, A Lieb-Thirring bound for a magnetic Pauli Hamiltonian, Commun. Math. Phys. 187, 567-582 (1997).

[5] L. Bugliaro, C. Fefferman, G. M. Graf, A Lieb-Thirring bound for a magnetic Pauli Hamiltonian, II, to appear in Rev. Math. Iberoamericana.

[6] L. Bugliaro, J. Fröhlich, G. M. Graf, Stability of quantum electrodynamics with non-relativistic matter, Phys. Rev. Lett. 77, 3494-3497 (1996).

[7] J. M. Combes, R. Schrader, R. Seiler, Classical bounds and limits for energy distributions of Hamiltonian operators in electromagnetic fields. Ann. Phys. 111, 1-18 (1978).

[8] J. G. Conlon, The ground state energy of a classical gas, Commun. Math. Phys. 94, 439-458 (1984).

[9] F. J. Dyson, A. Lenard, Stability of matter, I, II, J. Math. Phys. 8, 423-434 (1967); 9, 698-711 (1967).

[10] L. Erdős, Magnetic Lieb-Thirring inequalities, Commun. Math. Phys. 170, 629-668 (1995)

[11] L. Erdős, J. P. Solovej, Semiclassical eigenvalue estimates for the Pauli operator with strong non-homogeneous magnetic fields. I. Non-asymptotic Lieb-Thirring type estimates. To appear in Duke Math. Jour.

[12] C. Fefferman, Stability of Coulomb systems in a magnetic field. Proc. Natl. Acad. Sci. USA 92, 5006-5007 (1995).

[13] C. Fefferman, On electrons and nuclei in a magnetic field, Adv. Math. 124, 100-153 (1996).

[14] C. Fefferman, J. Fröhlich, G. M. Graf, Stability of ultraviolet-cutoff quantum electrodynamics with non-relativistic matter. Commun. Math. Phys. 190, 309-330 (1997).

[15] C. Fefferman, R. de la Llave, Relativistic stability of matter. I. Rev. Math. Iberoamericana 2, 119-215 (1986).

[16] J. Fröhlich, E. H. Lieb, M. Loss, Stability of Coulomb systems with magnetic fields I: The one-electron atom, Commun. Math. Phys. 104, 251-270 (1986).

[17] G. M. Graf, Stability of matter through an electrostatic inequality. Helv. Phys. Acta 70, 72-79 (1997).

[18] I. Herbst, Spectral theory of the operator $(p^2 + m^2)^{1/2} - Ze^2/r$. Commun. Math. Phys. 53, 285-294 (1977).

[19] E. H. Lieb, The stability of matter, Rev. Mod. Phys. 48, 553-569 (1976).

[20] E. H. Lieb, The stability of matter: From atoms to stars, Bull. Amer. Math. Soc. 22, 1-49 (1990).

[21] E. H. Lieb, M. Loss, Stability of Coulomb systems with magnetic fields II: The many-electron atom and the one-electron molecule, Commun. Math. Phys. 104, 271-282 (1986).
[22] E. H. Lieb, M. Loss, H. Siedentop, Stability of relativistic matter via Thomas-Fermi theory, Helv. Phys. Acta 69, 974-984 (1996).
[23] E. H. Lieb, M. Loss, J. P. Solovej, Stability of matter in magnetic fields, Phys. Rev. Lett. 75, 985-989 (1995).
[24] E. H. Lieb, H. Siedentop, J. P. Solovej, Stability of relativistic matter with magnetic fields. Phys. Rev. Lett. 79, 1785-1788 (1997).
[25] E. H. Lieb, H. Siedentop, J. P. Solovej, Stability and instability of relativistic electrons in classical electromagnetic fields. Jour. Stat. Phys. 89, 37-59 (1997).
[26] E. H. Lieb, J. P. Solovej, J. Yngvason, Asymptotics of heavy atoms in high magnetic fields: II. Semiclassical regions, Commun. Math. Phys. 161, 77-124 (1994).
[27] E. H. Lieb, W. Thirring, Bound for the kinetic energy of fermions which proves the stability of matter, Phys. Rev. Lett. 35, 687-689 (1975). Errata 35, 1116 (1975).
[28] E. H. Lieb, H.-T. Yau, The stability and instability of relativistic matter, Commun. Math. Phys. 118, 177-213 (1988).
[29] M. Loss, Stability of matter with magnetic fields, Proceedings of the XIIth International Congress on Mathematical Physics, Brisbane, Australia (1997).
[30] M. Loss, H.-T. Yau, Stability of Coulomb systems with magnetic fields: III. Zero energy bound states of the Pauli operator, Commun. Math. Phys. 104, 283-290 (1986).
[31] Z. Shen, On the moments of negative eigenvalues for the Pauli operator, preprint (1997).
[32] A. V. Sobolev, On the Lieb-Thirring estimates for the Pauli operator, Duke Math. J. 82, 607-635 (1996).
[33] A. V. Sobolev, Lieb-Thirring inequalities for the Pauli operator in three dimensions. IMA Vol. Math. Appl. 95, Springer (1997).
[34] W. Thirring, Introduction to 'The stability of matter: From atoms to stars, Selecta of Elliott H. Lieb', 2nd ed., Springer (1997).

Gian Michele Graf
Theoretische Physik
ETH-Hönggerberg
CH-8093 Zürich
gmgraf@itp.phys.ethz.ch

Rogers-Ramanujan Identities: A Century of Progress from Mathematics to Physics

Alexander Berkovich and Barry M. McCoy

Abstract. In this talk we present the discoveries made in the theory of Rogers-Ramanujan identities in the last five years which have been made because of the interchange of ideas between mathematics and physics. We find that not only does every minimal representation $M(p,p')$ of the Virasoro algebra lead to a Rogers-Ramanujan identity but that different coset constructions lead to different identities. These coset constructions are related to the different integrable perturbations of the conformal field theory. We focus here in particular on the Rogers-Ramanujan identities of the $M(p,p')$ models for the perturbations $\phi_{1,3},\ \phi_{2,1},\ \phi_{1,2}$ and $\phi_{1,5}$.

1991 Mathematics Subject Classification: 11P57, 82A68
Keywords and Phrases: Rogers-Ramanujan identities, lattice models of statistical mechanics, conformal field theory, affine Lie algebras

1 Introduction

In 1894 L.J. Rogers [1] proved the following identities for $a = 0, 1$ between infinite series and products valid for $|q| < 1$

$$\sum_{n=0}^{\infty} \frac{q^{n(n+a)}}{(q)_n} = \prod_{n=1}^{\infty} \frac{1}{(1-q^{5n-1-a})(1-q^{5n-4+a})}$$
$$= \frac{1}{(q)_\infty} \sum_{n=-\infty}^{\infty} (q^{n(10n+1+2a)} - q^{(5n+2-a)(2n+1)}) \text{ with } (q)_n = \prod_{j=1}^{n}(1-q^j). \quad (1)$$

For about the first 85 years after their discovery interest in these identities and their generalizations was confined to mathematicians and many ingenious proofs and relations with combinatorics, basic hypergeometric functions and Lie algebras were discovered by MacMahon, Rogers, Schur, Ramanujan, Watson, Bailey, Slater, Gordon, Göllnitz, Andrews, Bressoud, Lepowsky and Wilson and by 1980 there were over 130 isolated identities and several infinite families of identities known.

The entry of these identities into physics occurred in the early '80's when Baxter [2], Andrews, Baxter and Forrester [3, 4], and the Kyoto group [5] encountered (1) and various generalizations in the computation of order parameters of

certain lattice models of statistical mechanics. A further glimpse of the relation to physics is seen in the development of conformal field theory by Belavin, Polyakov and Zamolodchikov [6] and the form of computation of characters of representations of Virasoro algebra by Kac [7], Feigin and Fuchs [8] and Rocha-Caridi [9]. The occurrence of (1) in this context led Kac [10] to suggest that "every modular invariant representation of Vir should produce a Rogers-Ramanujan type identity."

The full relation, however, between physics and Rogers-Ramanujan identities is far more extensive than might be supposed from these first indications. Starting in 1993 the authors [11]-[17] have fused the physical insight of solvable lattice models in statistical mechanics with the classical work of the first 85 years and the recent developments in conformal field theory to greatly enlarge the theory of Rogers-Ramanujan identities. In this talk we will summarize the results of this work and present some of the current results. Our point of view will be dictated by our background in statistical mechanics but we will try to indicate where alternative viewpoints exist. Hopefully in this way some of the inevitable language barriers between physicists and mathematicians can be overcome.

2 What is a Rogers-Ramanujan Identity?

The work of the last 5 years originating in physics problems has provided a new framework and point of view in the study of Rogers-Ramanujan identities. The emphasis is not the same as in the earlier mathematical investigations and thus it is worthwhile to discuss generalities before the presentation of detailed results.

2.1 Sums instead of products

The equation (1) is the equality of three objects; an infinite sum involving $(q)_n$, an infinite product, and a second sum with $(q)_\infty$ in the denominator. For the first 85 years since (1) was proved it was the equality of the first infinite series with the infinite product which was called the Rogers-Ramanujan identity. The second sum while present in the intermediate steps of the proofs was always eliminated in favor of the product by use of the triple or pentuple product formula. The first important insight that was recognized when Rogers-Ramanujan identities arose in physics is that, contrary to this long history, it is not the product but rather the second sum on the right which arises in the statistical mechanical and conformal field theory applications. Indeed by now it is true that in most cases where we have generalizations of the identities between the two sums a product form is not known. Consequently by Rogers-Ramanujan identity we will mean the equality of the sums without further reference to possible product forms.

2.2 Polynomials instead of infinite series

The second insight which is also present in the very first papers on the connection of Rogers-Ramanujan identities with physics [2, 3, 4] is the fact that the physics will often lead to polynomial identities (with an order depending on an integer L) which yield infinite series identities as $L \to \infty$. The polynomial generalization of

(1) is the identity first proven in 1970 [18]

$$F_a(L,q) = B_a(L,q) \tag{2}$$

where

$$F_a(L,q) = \sum_{n=0}^{\infty} q^{n(n+a)} \begin{bmatrix} L-n-a \\ n \end{bmatrix} \tag{3}$$

and

$$B_a(L,q) = \sum_{n=-\infty}^{\infty} (-1)^n q^{n(5n+1+2a)/2} \begin{bmatrix} L \\ \lfloor \frac{1}{2}(L-5n-a) \rfloor \end{bmatrix} \tag{4}$$

where $\lfloor x \rfloor$ denotes the integer part of x and the Gaussian polynomials (q-binomial coefficients) are defined for integer m, n by

$$\begin{bmatrix} n \\ m \end{bmatrix} = \begin{cases} \frac{(q)_n}{(q)_m (q)_{n-m}} & 0 \le m \le n \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The identity (1) is obtained by using $\lim_{n\to\infty} \begin{bmatrix} n \\ m \end{bmatrix} = 1/(q)_m$ It is generalizations of the polynomial identity (2) which we will call a Rogers-Ramanujan identity.

2.3 THE GENERALIZATIONS OF $F_a(L,q)$

All known generalizations of $F_a(L,q)$ can be written in terms of the following function [12]

$$f = \sum_{\text{restrictions}} q^{\frac{1}{2}\mathbf{m}\mathbf{B}\mathbf{m} - \frac{1}{2}\mathbf{A}\mathbf{m}} \prod_{\alpha=1}^{n} \begin{bmatrix} ((1-\mathbf{B})\mathbf{m} + \frac{\mathbf{u}}{2})_\alpha \\ m_\alpha \end{bmatrix} \tag{6}$$

where $\mathbf{m}, \mathbf{u}$ and $\mathbf{A}$ are n dimensional vectors and $\mathbf{B}$ is an $n \times n$ dimensional matrix and the sum is over all values of the variables m_α possibly subject to some restrictions (such as being even or odd). In many cases the q-binomials are defined by (5) but there do occur cases in which an extended definition

$$\begin{bmatrix} m+n \\ m \end{bmatrix} = \begin{cases} \frac{(q^{n+1})_m}{(q)_m} & \text{for } m \ge 0,\ n \text{ integers} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

which allows n to be negative needs to be used.

The function (6) has the interpretation as the partition function for a collection of n differerent species of free massless (right moving) fermions with a linear energy momentum relation $e(P_j^\alpha)_\alpha = vP_j^\alpha$ where the momenta are quantized in units of $2\pi/M$ and are chosen from the sets

$$P_j^\alpha \in \{P_{\min}^\alpha(\mathbf{m}), P_{\min}^\alpha(\mathbf{m}) + \frac{2\pi}{M}, P_{\min}^\alpha(\mathbf{m}) + \frac{4\pi}{M}, \cdots, P_{\max}^\alpha(\mathbf{m})\} \tag{8}$$

with the Fermi exclusion rule $P_j^\alpha \neq P_k^\alpha$ for $j \neq k$ and all $\alpha = 1, 2, \cdots, n$,

$$P_{\min}^\alpha(\mathbf{m}) = \frac{\pi}{M}[((\mathbf{B}-1)\mathbf{m})_\alpha - \mathbf{A}_\alpha + 1] \text{ and } P_{\max}^\alpha = -P_{\min}^\alpha + \frac{2\pi}{M}(\frac{\mathbf{u}}{2} - \mathbf{A})_\alpha \tag{9}$$

where if some $u_\alpha = \infty$ the corresponding $P^\alpha_{\max} = \infty$. The $F_a(L,q)$ of (3) is regained in the very special case of $n = 1,\ \mathbf{B} = 2, \mathbf{u} = 2L - 2a$ and $\frac{1}{2}\mathbf{A} = -a$. Because of the Fermi exclusion rule we call these sums which generalize $F_a(L,q)$ Fermi forms. The generalization which the selection rule (8) makes over the usual exclusion rule of fermions is of great physical importance in the physics of the fractional quantum Hall effect [19].

2.4 The generalizations of $B_a(L,q)$

The first polynomial found which generalizes $B_a(L,q)$ is $B^{(p,p')}_{r,s}(L,a,b;q)$ given by[3, 4]

$$\sum_{j=-\infty}^{\infty}\left(q^{j(jpp'+rp'-sp)}\begin{bmatrix} L \\ \frac{L+a-b}{2} - jp' \end{bmatrix} - q^{(jp+r)(jp'+s)}\begin{bmatrix} L \\ \frac{L-a-b}{2} - jp' \end{bmatrix}\right). \tag{10}$$

with $L + a - b$ even. When $L \to \infty$ this polynomial reduces to

$$\lim_{L\to\infty} B^{(p,p')}_{r,s}(L,a,b;q) = \frac{1}{(q)_\infty}\sum_{j=-\infty}^{\infty}\left(q^{j(jpp'+rp'-sp)} - q^{(jp+r)(jp'+s)}\right) \tag{11}$$

which is (multiplied by $q^{\Delta^{(p,p')}_{r,s}-c/24}$) the well known character [8, 9] of the minimal model $M(p,p')$ of the Virasoro algebra with central charge $c = 1-6(p-p')^2/pp'$ and conformal dimension $\Delta^{(p,p')}_{r,s} = [(rp'-sp)^2 - (p-p')^2]/4pp'$ $(1 \le r \le p-1,\ 1 \le s \le p'-1)$. In the method of Feigin and Fuchs [8] this formula is obtained by modding out null vectors from the Fock space of one free boson. For this reason we call generalizations of $B_a(L,q)$ bosonic forms.

When $p = 2, p' = 5, r = 1$ and $s = 2 - a$ the character (11) is identical with the righthand side of (1). This is the original inspiration for the belief that there is a connection between conformal field theory and Rogers-Ramanujan identities.

Moreover we note that the relation between the exclusion rules (8) with the character formula (11) provided by Rogers-Ramanujan identities explains why conformal field theory and related Kac-Moody algebra [20] methods have been successfully applied to the fractional quantum Hall effect. In particular the Rogers-Ramanujan identities of [21] guarantee that starting from the $U(1)$ Kac-Moody algebra description of edge states in the fractional quantum hall effect [20] there must be corresponding description in terms of fermionic quasiparticles.

But unlike the generalizations of $F_a(L,q)$ there are other quite distinct generalizations of $B_a(L,q)$ which have been found to occur. One of the more widely studied uses, instead of q-binomials (5), the q-trinomials of Andrews and Baxter [22]

$$\begin{pmatrix} L \\ A \end{pmatrix}_2^p = \sum_{j=0}^{\infty} q^{j(j+A-p)}\frac{(q)_L}{(q)_j(q)_{j+A}(q)_{L-2j-A}} \tag{12}$$

and replaces (10) by either $B^{(1)(p,p')}_{r,s}(L,a,b;q)$ given by

$$\sum_{j=-\infty}^{\infty}\left[q^{j(pp'j+rp'-sp)}\begin{pmatrix} L \\ 2pj+a-b \end{pmatrix}_2^0 - q^{(jp+r)(jp'+s)}\begin{pmatrix} L \\ 2pj+a+b \end{pmatrix}_2^0\right], \tag{13}$$

which appear in the computation of the order parameters of the dilute A models [23], or $B_{r,s}^{(2)(p,p')}(L,a,b;q)$ given by

$$\sum_{j=-\infty}^{\infty}\left[q^{j(pp'j+rp'-sp)}\binom{L}{p'j+a-b}_2^0-q^{(jp+r)(jp'+s)}\binom{L}{p'j+a+b}_2^0\right]. \tag{14}$$

These q-trinomials have the property that $\lim_{L\to\infty}\binom{L}{A}_2^0=\frac{1}{(q)_\infty}$ and thus we see that although the polynomials $B_{r,s}^{(1)(p,p')}(L,a,b;q)$ and $B_{r,s}^{(2)(p,p')}(L,a,b;q)$ are not the same as $B_{r,s}^{(p,p')}(L,a,b;q)$ all three polynomials have the the same $L\to\infty$ limit (11). Further generalizations to q-multinomials have also been investigated [24, 25, 26, 27].

2.5 Proof by L-difference equations

The polynomial Roger-Ramanujan identities which generalize (2) are proven by demonstrating that the generalizations of $F_a(L,q)$ and $B_a(L,q)$ each satisfy the same difference equation in the variable L and are explicitly identical for suitably small values of L. Thus (2) is proven by demonstrating [18] that both $F_a(L,q)$ and $B_a(L,q)$ satisfy

$$h(L,q)=h(L-1,q)+q^{L-1}h(L-2,q)\quad \text{for} L\geq a+2 \tag{15}$$

and that they are identical for $L=a,a+1$. We refer to such equations as L-difference equations.

For the Fermi forms (6) the L-difference equations are derived by the general technique of telescopic expansions [13] which uses the two recursion relations for q-binomial coefficients (5)

$$\begin{bmatrix}n\\m\end{bmatrix}=\begin{bmatrix}n-1\\m-1\end{bmatrix}+q^m\begin{bmatrix}n-1\\m\end{bmatrix}=q^{n-m}\begin{bmatrix}n-1\\m-1\end{bmatrix}+\begin{bmatrix}n-1\\m\end{bmatrix} \tag{16}$$

which hold for all positive integers m,n or the identical recursion relations for generalized q-binomial coefficients (7) which hold for all integer m,n without restriction.

For the Bose form (10) which involves q-binomials the recursion relation (16) is sufficient to derive an L-difference equation but for the Bose forms (13) and (14) which involve q-trinomials we need not only the trinomial recursion relations such as

$$\binom{L}{A}_2^1=q^{L-1}\binom{L-1}{A}_2^1+q^A\binom{L-1}{A+1}_2^0+\binom{L-1}{A-1}_2^0 \tag{17}$$

but also so-called "tautological" equations such as

$$\binom{L}{A-1}_2^1-q^{A-1}\binom{L}{A+1}_2^1=\binom{L}{A-1}_2^0-q^{2A}\binom{L}{A+1}_2^0 \tag{18}$$

which reduce to trivialities when $q=1$. These "tautological" identities are what make the results involving q-trinomials more intricate to prove.

3 RESULTS FOR MINIMAL MODELS $M(p,p')$

The irreducible representations $M(p,p')$ with central charge less than one are parameterized by two relatively prime integers p and p' and the characters are given by (11). Thus the suggestion of Kac [10] can by taken to mean that each bosonic form of the character has a fermionic form. We have recently proven [14, 15] that such identities do indeed exist, even generalized to polynomial identities, for all p and p'.

But there is much more to the theory than this. The minimal models $M(p,p')$ can be realized in terms of the coset construction of fractional level [28, 29]

$$\frac{(A_1^{(1)})_1 \times (A_1^{(1)})_m}{(A_1^{(1)})_{m+1}} \quad \text{with } m = \frac{p}{p'-p} - 2 \text{ or } -\frac{p'}{p'-p} - 2. \tag{19}$$

However, these constructions are not unique and as an example we note that the model $M(3,4)$ in addition to the coset (19) with $m=1$ has the representation $(E_8^{(1)})_1 \times (E_8^{(1)})_1/(E_8^{(1)})_2$. It may thus be asked whether or not the Rogers Ramanujan identity is a unique property of the model $M(p,p')$ or is it a property of the several different coset constructions. For the $M(3,4)$ it is known that just as there are two coset constructions so there are two very different fermionic representations of the characters. For example

$$\chi_{1,1}^{(3,4)} = \sum_{\substack{m=0 \\ m \text{ even}}}^{\infty} \frac{q^{\frac{m^2}{2}}}{(q)_m} = \sum_{n_1,\cdots,n_8=0}^{\infty} q^{\mathbf{n}\mathbf{C}_{E_8}^{-1}\mathbf{n}} \prod_{j=1}^{8} \frac{1}{(q)_j}. \tag{20}$$

Thus it is natural to extend the suggestion of Kac to the conjecture that *to every coset construction of conformal field theory there exists a Rogers-Ramanujan polynomial identity.*

Physically there are even more reasons to make such a conjecture. Conformal field theories represent integrable massless systems. But it is not needed for a system to be massless for it to be integrable and it is known [30] that the operators $\phi_{1,3}$, $\phi_{2,1}, \phi_{1,5}$ and $\phi_{1,2}$ provide integrable massive perturbations of $M(p,p')$ whenever they are relevant. Each of these massive models has a fermionic quasi-particle spectrum which is a basis of states in the Hilbert space. As a basis this is independent of mass and thus still is a basis in the massless limit. We identity these quasi-particles with the fermionic representations (6). But the different massive perturbations will in general have a differerent number of quasi-particles and thus each integrable perturbation is expected to give a different fermionic form and hence a different Rogers-Ramanujan identity. However, even though at the level of the field theory these characters are the same at the level of finite statistical mechanical models the polynomials will be different. Thus we expect that each coset will lead to a different polynomial identity.

In the remainder of this section we will summarize how much of this conjecture has been proven.

3.1 THE PERTURBATION $\phi_{1,3}$

The integrable perturbation $\phi_{1,3}$ corresponds to the coset (19) and the bosonic polynomial is the original $B^{(p,p')}_{r,s}(L,a,b;q)$ (10) of [3, 4].

For the unitary case $M(p,p+1)$ the Rogers-Ramanujan identities were first proven in [13]. Here the matrix $\mathbf{B}$ is $\frac{1}{2}$ the Cartan matrix of A_{p-2}

$$B_{j,k} = \frac{1}{2}C_{A_{p-2}}|_{j,k} = \delta_{j,k} - \frac{1}{2}\delta_{j,k+1} - \frac{1}{2}\delta_{j,k-1} \quad 1 \le j,k \le p-2 \tag{21}$$

and $u_j = L\delta_{j,1}$ for $r = s = 1$ The general case of arbitrary p and p' is treated in [14, 15] and here $\mathbf{B}$ is a "fractional" generalization of a Cartan matrix which is obtained from the analysis of Bethe's Ansatz equations of the XXZ spin chain of Takahashi and Suzuki [31]. There are families of r, s for which the vector $\mathbf{A}$ is known but results for all cases have not been explicitly written down although an algorithm exists which allows the identity for any r, s to be found. For $p' = p+1$ only the conventional binomial coefficients (5) are needed and the Fermi form consists of a single term of the form (6). However, for general values of p' the modified binomials (7) arise and in addition there are many values of r, s where the Fermi form consists of a linear combination of terms of the form (6). It is essentially the existence of these linear combinations which makes the complete set of results difficult to explicitly write down.

3.2 THE PERTURBATIONS $\phi_{2,1}$ AND $\phi_{1,5}$

Rogers-Ramanujan identities for the character with the minimal conformal dimension for the integrable perturbations $\phi_{2,1}$ and $\phi_{1,5}$ have recently been obtained [16] for models $M(p,p')$ by means of the recently discovered [17] trinomial analogue of Bailey's lemma and some computer tested conjectures. For the unitary case $M(p,p+1)$ we have just completed the proof of the identities for all values of r and s. When $2p > p'$ the perturbation $\phi_{2,1}$ is relevant and the bosonic form $B^{(1)}$ of (13) appears in the identities. We also have identities for $\frac{p'}{3} < p < \frac{p'}{2}$ where the perturbation $\phi_{1,5}$ is relevant and the bosonic form $B^{(2)}$ of (14) is used.

For the unitary case $M(p,p+1)$ the matrix $\mathbf{B}$ is of dimension $p-1$ where

$$\begin{aligned} B_{j,k} &= \frac{1}{2}C_{A_{p-2}}|_{j,k} \quad 2 \le j,k \le p-2 \\ B_{0,0} &= B_{1,1} = 1, \quad B_{0,2} = -B_{2,0} = 1/2 \quad B_{1,2} = B_{2,1} = -1/2 \end{aligned} \tag{22}$$

and zero otherwise and $u_j = 2L\delta_{j,0}$ for $r = s = 1$. This matrix differs significantly from the $p-2$ dimensional matrix (21) in that it is not symmetric.

The matrices $\mathbf{B}$ are also known [16] for the nonunitary cases $p' \neq p+1$. However, in many of these nonunitary cases a new phenomena arises not seen in the $\phi_{1,3}$ perturbations, namely that there can be several different fermionic representations (with different dimensions of the $\mathbf{B}$ matrix) of the same bosonic polynomial.

3.3 THE PERTURBATION $\phi_{1,2}$

The final case of integrable perturbations is $\phi_{1,2}$ but this case is not nearly so well understood. For the three very special unitary cases of cases $M(3,4)$, $M(4,5)$ and $M(6,7)$ Rogers-Ramanujan identities are known [11, 32] where the $\mathbf{B}$ matrices are twice the inverse of the Cartan matrix of E_8, E_7 and E_6 respectively and the bosonic form is obtained from (13) with the replacement $p \to p+1$ in the q-trinomials. Beyond these nothing further seems to be known.

4 HOW MANY IDENTITIES?

We demonstrated in [14, 15] that every $M(p,p')$ yields a set of Rogers-Ramanujan identities. But we also found that there are more than one identity for each $M(p,p')$. The question then arises of how many fermionic representations there are for the characters of each model $M(p,p')$ The answer to this is not known and the scope of the problem is perhaps most vividly shown by considering the three state Potts model $M(5,6)$ where in addition to the identities for the $\phi_{2,1}$ perturbation discussed above there is another set of identities which are a special case of the "parafermionic" identities first found by Lepowsky and Primc [33] in 1985 where the matrix $\mathbf{B}$ is twice the inverse Cartan matrix of A_2 and in the limit $L \to \infty$, $\mathbf{u} \to \infty$. This perturbation is also for the $\phi_{2,1}$ perturbation but has two quasi-particles instead of the four quasi-particles of (22). One may speculate that this has something to do with the difference between A and D modular invariants, but the actual explanation and interpretation of this fact is not known nor is it known if such extra representations exist for other models. If this is part of the explanation then we must enlarge the conjecture of sec. 3 to account for the possible modular invariants. But even this suggestion will not give an explanation for all of the various identities found for the nonunitary $\phi_{2,1}$ perturbations in [16]. The full range of Rogers-Ramanujan identities is by no means yet understood and it is anticipated that both in the mathematics and in the physics there is much still left to be discovered.

ACKNOWLEDGEMENT This work is supported in part by the National Science Foundation of the USA under DMR 9703543

REFERENCES

[1] L.J. Rogers, Second memoir on the expansion of certain infinite products,Proc. Lond. Math. Soc. 25 (1894) 318.

[2] R.J. Baxter, Rogers-Ramanujan identities in the hard hexagon model, J. Stat. Phys. 26 (1981) 427.

[3] G.E. Andrews, R.J. Baxter and P.J. Forrester, Eightvertex SOS model and generalized Rogers-Ramanujan-type identities, J. Stat. Phys. 35 (1984) 193.

[4] P.J. Forrester and R.J. Baxter, Further exact solutions of the eightvertex SOS model and generalizations of the Rogers-Ramanujan identities, J. Stat. Phys. 38 (1985) 435.

[5] E. Data, M. Jimbo, A. Kuniba, T. Miwa and M. Okado, Exactly solvable SOS models: local height probabilities and theta function identities, Nucl. Phys. B290 (1987) 231.

[6] A.A. Belavin, A.M. Polyakov and A.B. Zamolodchikov, Infinite conformal symmetry in two-dimensional quantum field theory, Nucl. Phys. B241 (1984) 333.

[7] V.G. Kac, *Contravariant form for infinite-dimensional Lie algebras and superalgebras,* Lect. Notes in Phys. 94 (1979) 441.

[8] B.L. Feigin and D.B. Fuchs, Verma modules over the Virasoro algebra, Funct. Anal. Appl. 17 (1983) 241.

[9] A. Rocha-Caridi, in *Vertex Operators in Mathematics and Physics,* ed. J. Lepowsky, S. Mandelstam and I.M. Singer (Springer, Berlin 1985) 451.

[10] V.G. Kac, Modular invariance in mathematics and physics, in *Proceedings of the AMS Centennial Symposium*, (1992 American Mathematical Society) 337.

[11] R. Kedem, T.R. Klassen, B.M. McCoy and E. Melzer, Fermionic quasi-particle representations for characters of $(G^{(1)})_1 \times (G^{(1)})_1/(G^{(1)})_2$, Phys. Lett. B 304 (1993) 263.

[12] R. Kedem, T.R. Klassen, B.M. McCoy and E. Melzer, Fermionic sum representations for conformal field theory characters, Phys. Lett. B307 (1993) 68.

[13] A. Berkovich, Fermionic counting of RSOS states and Virasoro character formulae for the unitary minimal series $M(\nu, \nu+1)$: Exact results, Nucl. Phys. B431 (1994) 315.

[14] A. Berkovich and B.M. McCoy, Continued fractions and fermionic representations for characters of $M(p, p')$ minimal models, Lett. Math. Phys. 37 (1996) 49.

[15] A. Berkovich, B.M. McCoy and A. Schilling, Rogers-Schur-Ramanujan type identities for $M(p, p')$ minimal models of conformal field theory, Comm. Math. Phys. 191 (1998) 325.

[16] A. Berkovich, B.M. McCoy and P.A. Pearce, The perturbation $\phi_{2,1}$ and $\phi_{1,5}$ of the minimal models $M(p, p')$ and the trinomial analogue of Bailey's lemma, Nucl. Phys. B519[FS] (1998) 597.

[17] G.E. Andrews and A. Berkovich, A trinomial analogue of Bailey's lemma, Comm. Math. Phys. 192 (1998) 2451.

[18] G.E. Andrews, A polynomial identity which implies the Rogers-Ramanujan identities, Scripta Mathematica 28 (1970) 297.

[19] R.B. Laughlin, Anomalous quantum Hall effect; An incompressible quantum fluid with fractionally charged excitations, Phys. Rev. Letts. 50 (1983) 1395.

[20] X-G Wen, Theory of the edge states in fractional quantum Hall effect, Int. J. Mod. Phys. B6 (1992) 1711.

[21] R.A.J. van Elburg and K. Schoutens, Quasi-particles in fractional quantum hall effect edge theories, condmat/9801272.

[22] G.E. Andrews and R.J. Baxter, Lattice gas generalizations of the hard-hexagon model III. q-trinomial coefficients, J. Stat. Phys. 47 (1987) 297.

[23] S.O. Warnaar, P.A. Pearce, K.A. Seaton, and B. Nienhuis, Order parameters of the dilute A models, J. Stat. Phys. 74 (1994) 469.

[24] G.E. Andrews, Schur's theorem, Capparelli's conjecture and q-trinomial coefficients, in *The Rademacher legacy to mathematics*, G.E. Andrews et al eds., Cont. Math. 166 (1994) 141.

[25] A.N. Kirillov, Dilogarithm Identities, Prog. Theor. Phys. Suppl. 118 (1995) 61.

[26] A. Schilling, Multinomials and polynomial bosonic forms for the branching functions of $\widehat{su}(2)_m \times \widehat{su}(2)_N/\widehat{su}(2)_{m+N}$, Nucl. Phys. B 467 (1996) 247.

[27] S.O. Warnaar, The Andrews-Gordon identities and q multinomial coefficients, Comm. Math. Phys. 184 (1997) 203.

[28] V. Kac and M. Wakimoto, Modular invariant representations of infinite-dimensional Lie algebras and superalgebras, Proc. Nat. Acad. Sci. USA 85 (1988) 4956.

[29] P. Mathieu and M.A. Walton, Fractional-level Kac-Moody algebras and non-unitary coset conformal field theories, Prog. Theor. Phys. Suppl. 102 (1990) 229.

[30] A.B. Zamolodchikov, Higher-order integrals of motion in two-dimensional models of field theory with a broken conformal symmetry, JETP Lett. 46 (1987) 160.

[31] M. Takahashi and M. Suzuki, One-dimensional anisotropic Heisenberg model at finite temperature, Prog. Theor. Phys. 48 (1972) 2187.

[32] S.O. Warnaar and P.A. Pearce, Exceptional structure of the dilute A_3 model: E_8 and E_7 Rogers-Ramanujan identities, J. Phys. A27 (1994) L891.

[33] J. Lepowsky and M. Primc *Structure of the standard modules for the affine Lie algebra* $A_1^{(1)}$, Cont. Math. vol. 46, (Providence, RI 1985).

Alexander Berkovich
Institute for Theoretical Physics
State University of New York
Stony Brook,NY,11794-3840, USA
alexb@max.physics.sunysb.edu

Barry M. McCoy
Institute for Theoretical Physics
State University of New York
Stony Brook, NY, 11794-3840
mccoy@max.physics.sunysb.edu

Metastability and the Ising Model

Roberto H. Schonmann

Abstract. We present recent results on a classical non-equilibrium statistical mechanics problem, in the context of a well-studied idealized interacting particle system, called kinetic Ising model. The problem concerns the speed and the patterns of relaxation of statistical mechanical systems in the proximity of the phase-transition region, and is related to the problem of understanding the metastable behavior of systems in such regions.

1991 Mathematics Subject Classification: 60K35 82B27
Keywords and Phrases: kinetic Ising model, stochastic Ising model, Glauber dynamics, metastability, relaxation, nucleation, droplet growth, Wulff shape, large deviations, asymptotic expansion

It is well known that a ferromagnetic material which is in equilibrium under a negative external magnetic field relaxes to equilibrium very slowly after the magnetic field is switched to a small positive value. A detailed mathematical analysis of such a phenomenon can only be performed on simplified models. It is widely accepted that an appropriate model for this and many other purposes is a kinetic Ising model: a Markov process which endows the traditional Ising model with a particular stochastic dynamics. On each vertex of an infinite lattice $\mathbb{Z}^d$, we have variables (called spins) which take the values -1 or $+1$. The system evolves in continuous time as a Markov process which is time-reversible and has as invariant measures the classical Gibbs measures of statistical mechanics. When the temperature parameter, T, appearing in the definition of the model is small enough, there is a phase transition which takes place when the external field parameter, h, changes sign (this corresponds to the change from a negative to a positive orientation of most spins). The question then arises of how the system relaxes to equilibrium when h is small and positive, and the system is initially in an equilibrium distribution corresponding to a small negative value of h.

Simulations have shown that the relaxation mechanism is driven by the behavior of the clusters (droplets) of $+1$-spins which form initially in the sea of -1-spins. While small droplets tend to shrink and disappear, large ones tend to grow and are responsible for the relaxation. This phenomenon has long been understood on non-rigorous heuristic grounds, and can be used to predict for instance the order of magnitude, as $h \searrow 0$, of the relaxation time for the process. The prediction is that the relaxation time grows as $\exp(\lambda h^{d-1})$, where λ is a constant which

can be computed. The value of λ is, in particular, related to the equilibrium surface tension of the Ising model through the Wulff construction, which solves a variational problem.

In this note we will overview rigorous results of the type described above and also some important extensions. A thorough review of metastability, even in the context of the kinetic Ising models, is far beyond the scope of this note. Here we will limit ourselves to the main results in the papers [Sch1] and [SS], which concern metastability in the vicinity of the phase transition region. A great deal of recent progress on metastability of the kinetic Ising models stems from the fact that these models also display metastable behavior away from this region, at low enough temperature. For a detailed discussion of relations between the various manifestations of metastability of kinetic Ising models we refer the reader to [Sch2], where further reference to the literature can also be found. More recent progress in this direction is contained in [Nev], [BC], [DS], [CO], [CL] and references therein. For a paper which reports on extensive numerical studies directly related to the mathematical work reviewed here, we refer the reader to [RTMS].

The precise definition of the kinetic Ising models is lengthy and somewhat technical. It can be found, e.g., in [Sch1] and [SS]. For the purpose of this note it is best to just give a somewhat intuitive description. At each site of the lattice $\mathbb{Z}^d$ there is a variable (spin) which can take the value -1 or $+1$. The configuration on the complete lattice is then an element of the space $\Omega = \{-1, +1\}^{\mathbb{Z}^d}$. The system evolves in time, with spins flipping back and forth, at rates which depend on the state of nearby spins. The system as a whole is a Markov process with state space Ω. The interaction among spins is driven by an energy function (Hamiltonian) formally defined on Ω by

$$H_h(\sigma) = -\frac{1}{2} \sum_{x,y \text{ n.n.}} \sigma(x)\sigma(y) - \frac{h}{2} \sum_x \sigma(x),$$

where "x, y n.n." means that x and y are nearest neighbors in $\mathbb{Z}^d$, i.e., they are separated by Euclidean distance 1, $h \in \mathbb{R}$ is the external field and $\sigma \in \Omega$ is a generic configuration.

Formally, Gibbs distributions are defined as probability distributions μ over Ω, with

$$\mu(\sigma) = \frac{\exp(-H_h(\sigma)/T)}{\text{Normalization}},$$

where $T = 1/\beta > 0$ is the temperature. When $h \neq 0$ or $T > T_c = T_c(d)$ it is known that there is a unique Gibbs distribution, which then describes the system in equilibrium and will be denoted by $\mu_{T,h}$. In $d = 1$, $T_c = 0$, but for $d \geq 2$, $T_c > 0$. The segment $\{\{0\} \times (0, T_c)\}$ of the phase diagram $h \times T$ corresponds then to the phase-transition region. For these values of the parameters there are multiple Gibbs distributions; one of them corresponds to a limit of Gibbs distributions under $h < 0$ (resp. $h > 0$) as $h \nearrow 0$ (resp. $h \searrow 0$), and is called the $(-)$-phase (resp. the $(+)$-phase), represented by $\mu_{T,-}$ (resp. $\mu_{T,+}$). Expectations with respect to Gibbs measures will be denoted in the standard fashion

$$\langle f \rangle_{T,h} = \int f d\mu_{T,h}.$$

Of particular interest is the magnetization $m(T,h) = \langle\sigma(0)\rangle_{T,h}$. Away from the phase-transition region, $m(T,\cdot)$ is analytic. It is nevertheless believed that for $T < T_c$ this function has no analytic continuation from $h < 0$ to $h > 0$. This result has been proved indeed for low enough T in [Isa].

The time evolution which defines the kinetic Ising model as a Markov process on Ω is given by a generator L of the form given next. Intuitively, when the configuration is σ, the spin at each site $x \in \mathbb{Z}^d$ is flipping at a rate $c(x,\sigma)$.

$$(Lf)(\sigma) = \sum_{x\in\mathbb{Z}^d} c(x,\sigma)(f(\sigma^x) - f(\sigma)).$$

Here $f : \Omega \to \mathbb{R}$ is supposed to be a local observable, i.e., to depend only on the spin at finitely many sites of the lattice, σ^x is the configuration obtained from σ by flipping the spin at the site x, and $c(x,\sigma)$ is called the rate of flip of the spin at the site x when the system is in the state σ. The rates $c(x,\sigma)$ are supposed to satisfy certain conditions, the main one of them being called detailed balance or reversibility, and formally given by

$$\mu(\sigma)c(x,\sigma) = \mu(\sigma^x)c(x,\sigma^x).$$

This assures that the Gibbs distributions are invariant for the process. Other conditions are that the rates are invariant under translations of the lattice, are of finite range of dependency, are monotone in the configuration and external field, and are uniformly bounded above and below when T is fixed and $|h|$ is small. Several choices can be made for the rates, satisfying all this conditions. To give a few examples, we introduce

$$\Delta_x H_h(\sigma) = H_h(\sigma^x) - H_h(\sigma).$$

Common choices of rates are:
Example 1) *Metropolis Dynamics*

$$c_{T,h}(x,\sigma) = \exp(-\beta(\Delta_x H_h(\sigma))^+),$$

where $(a)^+ = \max\{a, 0\}$ is the positive part of a.

Example 2) *Heat Bath Dynamics*

$$c_{T,h}(x,\sigma) = \frac{1}{1+\exp(\beta\Delta_x H_h(\sigma))}.$$

Example 3)

$$c_{T,h}(x,\sigma) = \exp\left(-\frac{\beta}{2}\Delta_x H_h(\sigma)\right).$$

If in the kinetic Ising model the initial configuration is selected at random according to a probability measure ν, then the resulting process is denoted by $(\sigma^\nu_{T,h;t})_{t\geq 0}$. When ν is concentrated on the configuration with all spins -1, we will denote this process by $(\sigma^-_{T,h;t})_{t\geq 0}$. The probability measure on the space of trajectories of the process will be denoted by $\mathbb{P}$, and the corresponding expectation by $\mathbb{E}$.

The following is the main result of [Sch1].

THEOREM 1. *For each dimension $d \geq 2$ there is $T_0 > 0$ such that for every temperature $T \in (0, T_0)$ the following happens. There are constants $0 < \lambda_1(T) \leq \lambda_2(T) < \infty$ such that if we let $h \searrow 0$ and $t \to \infty$ together, then for every local observable f*

i) $\mathbb{E}(f(\sigma^-_{T,h;t})) \to \langle f \rangle_{T,-}$ *if* $\limsup h^{d-1} \log t < \lambda_1(T)$.

ii) $\mathbb{E}(f(\sigma^-_{T,h;t})) \to \langle f \rangle_{T,+}$ *if* $\liminf h^{d-1} \log t > \lambda_2(T)$.

Explicit estimates on the values of $\lambda_1(T)$ and $\lambda_2(T)$ were also given in [Sch1]. The theorem above was conjectured by Aizenman and Lebowitz in [AL], where they proved a similar result for certain deterministic cellular automata evolving from initial random configurations selected according to translation invariant product measures. Actually they conjectured the stronger result, which states that also $\lambda_1(T) = \lambda_2(T) =: \lambda_c(T)$.

Theorem 1 was greatly improved in [SS] in the case in which $d = 2$. In particular in this paper the conjecture by Aizenman and Lebowitz was fully vindicated in this case. A somewhat simplified and partial statement of the main result in [SS] is as follows.

THEOREM 2. *Suppose $d = 2$ and $T < T_c$. There is a constant $\lambda_c = \lambda_c(T)$ such that for every probability distribution $\nu = \mu_{T,h'}$, $h' < 0$, the following happens.*

i) If $0 < \lambda < \lambda_c$, then for each $n \in \{1, 2, ...\}$ and for each local observable f,

$$\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;\exp(\lambda/h)}\right)\right) = \sum_{j=0}^{n-1} \frac{1}{j!} \frac{d^j \langle f \rangle_{T,\hat{h}}}{d\hat{h}^j}\bigg|_{\hat{h}=0_-} h^j + O(h^n)$$

for $h > 0$, where $O(h^n)$ is a function of f and h which satisfies $\limsup_{h \searrow 0} |O(h^n)|/h^n < \infty$.

ii) If $\lambda > \lambda_c$, then for any finite positive C there is a finite positive C_1 such that for every local observable f,

$$\left|\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;\exp(\lambda/h)}\right)\right) - \langle f \rangle_{T,h}\right| \;\leq\; C_1 \; ||f||_\infty \; \exp\left(-\frac{C}{h}\right),$$

for all $h > 0$.

The value of $\lambda_c(T)$ can be written in terms of other quantities which are related to the equilibrium distributions of the Ising model. This expression and its meaning, which are of great relevance, will be presented later in the paper. Next we compare Theorems 1 and 2 and explain some of their content.

Three of the ways in which Theorem 2 improves on Theorem 1 when $d = 2$ are: 1) There is a single constant λ_c separating the regimes (i) and (ii). 2) The temperature is now only required to be below T_c. 3) The initial distribution is much more general than in Theorem 1, where it was supposed to be concentrated on the configuration with all spins down. It is natural indeed to start from an equilibrium state at a small negative h, change it to a small positive h and observe the evolution of the system afterwards.

To illustrate and clarify the main way in which Theorem 2 improves further the statement in Theorem 1, let us take the local observable given by $f(\sigma) = \sigma(0)$ and $n = 2$. We have then, when $0 < \lambda < \lambda_c$

$$\mathbb{E}\left(\sigma^{\nu}_{T,h;\exp(\lambda/h)}(0)\right) = -m^* + \chi h + O(h^2),$$

when $h > 0$. Here

$$m^* = m^*(T) = \langle\sigma(0)\rangle_{T,+} = -\langle\sigma(0)\rangle_{T,-},$$

is the spontaneous magnetization, and

$$\chi = \chi(T) = \left.\frac{d\langle\sigma(0)\rangle_{T,h}}{dh}\right|_{h=0_-} = \left(\frac{\beta}{2}\right)\sum_{x\in\mathbb{Z}^2}\left\{\langle\sigma(0)\sigma(x)\rangle_{T,-} - \langle\sigma(0)\rangle_{T,-}\langle\sigma(x)\rangle_{T,-}\right\},$$

is the susceptibility at $h = 0_-$. This means that when $h > 0$ is small the function $-m^* + \chi h$ is a better approximation to $\mathbb{E}\left(\sigma^{\nu}_{T,h;\exp(\lambda/h)}(0)\right)$ than the constant function identical to $-m^* = \langle f\rangle_{T,-}$. This function $-m^* + \chi h$ is the smooth linear continuation into the region $h \geq 0$ of the function which to $h < 0$ associates the equilibrium expectation $\langle f\rangle_{T,h}$. Similar interpretations can be given for larger values of n and arbitrary f. In this sense Theorem 2 shows that the dynamics gives meaning to arbitrarily smooth metastable continuations of the distributions $\mu_{T,h}$, $h < 0$, into the region $h > 0$, inspite of the absence of an analytic continuation.

In the Physics literature (see, e.g., [BM]), one sometimes relates the metastable relaxation of a system to the presence of a "plateau" in the graph corresponding to the time evolution of a quantity of the type of $\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;t}\right)\right)$. Of course, strictly speaking there is no "plateau", and generically the slope of such a function is never 0. Still, from the experimental point of view a rough "plateau" can be seen and described as follows. In a relatively short time $\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;t}\right)\right)$ seems to converge to a value close to $\langle f\rangle_{T,-}$; after this, one sees an apparent flatness in the relaxation curve over a period of time which may be quite long compared with the time needed to first approach this value. But eventually the relaxation curve starts to deviate from this almost constant value and move towards the true asymptotic limit, close to $\langle f\rangle_{T,+}$. The experimentally almost flat portion of the relaxation curve is referred to as a "plateau". Theorem 2 can be seen to some extent as giving some precise meaning to such a "plateau", and we discuss now two ways in which this can be done. First note that if $0 < \lambda' < \lambda'' < \lambda_c$, then from Part (i) of the Theorem we have

$$\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;\exp(\lambda'/h)}\right)\right) - \mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;\exp(\lambda''/h)}\right)\right) \to 0,$$

faster than any power of h. Observe that we are considering times which are of different order of magnitudes, when h is small, and still we are observing a

nearly constant $\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;t}\right)\right)$. For a second way in which Theorem 2 can be seen as expressing the presence of a "plateau", we can think of plotting $\mathbb{E}\left(f\left(\sigma^{\nu}_{T,h;t}\right)\right)$ versus $\log(t)$, rather than versus t. This is somewhat the natural graph to consider, if one is interested in the order of magnitude of the relaxation time. If the $\log(t)$-axis is drawn in the proper scale, amounting to replacing it with $h\log(t)$, then, when h is small, Theorem 2 tells us that the graph should be close to that of a step function which jumps at the point λ_c, from the value $\langle f\rangle_{T,-}$ to the value $\langle f\rangle_{T,+}$.

The relation between the constant $\lambda_c(T)$ and some quantities related to the equilibrium Ising model can best be explained by presenting an heuristic reasoning which lies behind Theorems 1 and 2. The heuristics is presented next in the case $d = 2$. For more on this heuristics including a different way of approaching it and some of its history see [RTMS].

The first ingredient of the heuristics is the idea of looking at an individual droplet of the stable phase (roughly the $(+)$-phase, since h is small) in a background given by the metastable phase (roughly the $(-)$-phase). Let S be the shape of that droplet, which a priori can be arbitrary. Say that l^2 is the volume (i.e., the number of sites) of the droplet, and let us find an expression for the free-energy of such a droplet. This free-energy may be seen as coming from two main contributions. There should be a bulk term, proportional to l^2. This term should be obtained by multiplying l^2 by the difference in free-energy per site between the $(+)$-phase and the $(-)$-phase in the presence of a small magnetic field $h > 0$. This difference in the free-energy per site of the two phases should come only from the term in the Hamiltonian which couples the spins to the external field and should therefore be given by $2m^*h/2 = m^*h$. The other relevant contribution to the free-energy of the droplet should come from its surface, where there is an interface between the $(+)$-phase and the $(-)$-phase. This contribution is proportional to the length of the interface, which is of the order of l. It should be multiplied by a constant w_{S} which depends on the shape of the droplet. This constant w_{S} represents the excess free-energy per unit of length integrated over the surface of the droplet when its scale is changed so that its volume becomes 1. Adding the pieces, we obtain for the free-energy of the droplet the expression

$$\Phi_{\mathrm{S}}(l) = -m^*hl^2 + w_{\mathrm{S}}l.$$

The two terms in this expression become of the same order of magnitude, in case l is of the order of $1/h$. Therefore, it is natural to write $l = b/h$, with a new variable $b \geq 0$. This yields

$$\Phi_{\mathrm{S}}(b/h) = \frac{\phi_{\mathrm{S}}(b)}{h},$$

where

$$\phi_{\mathrm{S}}(b) = -m^*b^2 + w_{\mathrm{S}}b.$$

This very simple function takes the value 0 at $b = 0$, grows with b on the interval $[0, B^{\mathrm{S}}_c,]$, where $B^{\mathrm{S}}_c = B^{\mathrm{S}}_c(T) = \frac{w_{\mathrm{S}}}{2m^*}$, reaching its absolute maximum $\phi_{\mathrm{S}}(B^{\mathrm{S}}_c) = \frac{(w_{\mathrm{S}})^2}{4m^*} = A^{\mathrm{S}}(T) = A^{\mathrm{S}}$ at the end of this interval. Then it decreases with b on the semi-infinite interval $[B^{\mathrm{S}}_c, \infty)$, converging to $-\infty$ as $b \to \infty$.

Metastability is then "understood" from the fact that systems in contact with a heat bath move towards lowering their free-energy, so that the presence of a free-energy barrier which needs to be overcome in order to create a large droplet of the stable phase with any shape keeps the system close to the metastable phase. Subcritical droplets are constantly being created by thermal fluctuations, in the metastable phase, but they tend to shrink, as dictated by the free-energy landscape. On the other hand, once a supercritical droplets is created due to a larger fluctuation, it will grow and drive the system to the stable phase, possibly colliding and coalescing in its growth with other supercritical droplets created elsewhere. As a function of h, the linear size of a critical droplet, B_c^{S}/h, blows up as $h \searrow 0$. One can then, in a somewhat circular, but heuristically-meaningful way, say that the macroscopic free-energy of droplets is indeed a relevant object of consideration. One can also hope then that sharp theorems could be conjectured and possibly proven regarding the asymptotic behavior of quantities of interest in the limit $h \searrow 0$.

Regarding the shape of the droplet, the height of this barrier is minimized by minimizing the value of the constant w_{S}. It is a fact (see [DKS]) that indeed one can introduce a well defined surface tension function between the $(+)$-phase and the $(-)$-phase, and that it produces a single convex shape S which minimizes w_S. This shape is called the Wulff shape. We will simplify the notation by omitting the subscript S when it is the Wulff shape. In particular,

$$B_c = \frac{w}{2m^*}, \quad A = \frac{w^2}{4m^*}.$$

Based on the expression above for the free-energy barrier, one predicts the rate of creation of supercritical droplets with center at a given place to be $\exp\left(\frac{-\beta A}{h}\right)$.

In what follows now we write d instead of 2, to make the role of the dimension clear in the geometric argument which comes next. We are concerned with an infinite system, and we are observing it through a local function f, which depends, say, on the spins in a finite set $\mathrm{Supp}(f)$. For us the system will have relaxed to equilibrium when $\mathrm{Supp}(f)$ is covered by a big droplet of the plus-phase, which appeared spontaneously somewhere and then grew, as discussed above. We want to estimate how long we have to wait for the probability of such an event to be large. If we suppose that the radius of supercritical droplets grows with a speed v, then we can see that the region in space-time where a droplet which covers $\mathrm{Supp}(f)$ at time t could have appeared is, roughly speaking, a cone with vertex in $\mathrm{Supp}(f)$ and which has as base the set of points which have time-coordinate 0 and are at most at distance tv from $\mathrm{Supp}(f)$. The volume of such a cone is of the order of $(vt)^d t$. The order of magnitude of the relaxation time, t_{rel}, before which the region $\mathrm{Supp}(f)$ is unlikely to have been covered by a large droplet and after which the region $\mathrm{Supp}(f)$ is likely to have been covered by such an object can now be obtained by solving the equation

$$(vt_{\mathrm{rel}})^d \, t_{\mathrm{rel}} \, \exp\left(-\frac{\beta A}{h}\right) = 1.$$

This gives us

$$t_{\rm rel} = v^{-d/(d+1)} \ \exp\left(\frac{\beta A}{(d+1)\ h}\right).$$

In order to use this relation to predict the way in which the relaxation time scales with h, one needs to figure out the way in which v scales with h. If we suppose, for instance, that v does not scale with h, or at least that if it goes to 0, as $h \searrow 0$, it does it so slowly that

$$\lim_{h\searrow 0} \ h^{d-1} \log v = 0, \tag{1}$$

then we can predict that

$$t_{\rm rel} \simeq \exp\left(\frac{\beta A}{(d+1)\ h}\right) = \exp\left(\frac{\lambda_c}{h}\right),$$

where

$$\lambda_c = \frac{\beta A}{d+1} = \frac{\beta w^2}{(d+1)\ 4m^*} = \frac{\beta w^2}{12\ m^*}. \tag{2}$$

The heuristics above may seem extremely crude. Potentially the interaction between droplets could spoil the whole picture and lead to a much faster decay. In the opposite direction, even if the droplet picture makes sense, their speed of growth could be so slow that (1) could fail an therefore the relaxation time would be much larger than predicted above.

One of the major contributions of [SS] is to prove that indeed λ_c in Theorem 2 is given by (2). This means that close to the phase transition region the time evolution can be well described in first approximation by the highly simplified droplet dynamics.

Acknowledgements: It is a pleasure to thank Senya Shlosman for the collaboration in [SS] and other related projects. This work was supported by the N.S.F. grant DMS-9703814.

REFERENCES

[AL] M. Aizenman and J. L. Lebowitz, *Metastability effects in bootstrap percolation*, J. Phys. A **21** (1988), 3801–3813.

[BC] G. Ben Arous and R. Cerf, *Metastability of the three dimensional Ising model on a torus at very low temperatures*, Electronic Journal of Probability **1** (1996), 1–55.

[BM] K. Binder and H. Müller-Krumbhaar, *Investigation of metastable states and nucleation in the kinetic Ising model*, Physical Review B **9** (1974), 2328–2353.

[CL] E. N. M. Cirillo and J. Lebowitz, *Metastability in the two-dimensional Ising model with free boundary conditions*, J. of Stat. Phys. **90** (1998), 211–226.

[CO] E. N. M. Cirillo and E. Olivieri, *Metastability and nucleation for the Blume-Capel model – different mechanisms of transition*, J. of Stat. Phys. **83** (1996), 473–554.

[DS] P. Dehghanpour and R. H. Schonmann, *Metropolis dynamics relaxation via nucleation and growth*, Commun. Math. Phys. **188** (1997), 89–119.

[DKS] R. L. Dobrushin, R. Kotecký and S. B. Shlosman, *Wulff construction: a global shape from local interaction*, AMS translations series, Providence (Rhode Island), 1992.

[Isa] S. N. Isakov, *Nonanalytic features of the first order phase transition in the Ising model*, Commun. Math. Phys. **95** (1984), 427–443.

[Nev] E. J. Neves, *A discrete variational problem related to Ising droplets at low temperatures.*, J. of Stat. Phys. **80** (1995), 103–123.

[RTMS] P. A. Rikvold, H. Tomita, S. Miyashita, and S. W. Sides, *Metastable lifetimes in a kinetic Ising model: dependence on field and system size*, Phys. Rev. E **49** (1994), 5080–5090.

[Sch1] R. H. Schonmann, *Slow droplet-driven relaxation of stochastic Ising models in the vicinity of the phase coexistence region*, Commun. Math. Phys. **161** (1994), 1–49.

[Sch2] R. H. Schonmann, *Theorems and conjectures on the droplet driven relaxation of stochastic Ising models*, in *Probability theory of spatial disorder and phase transition*, G. Grimmett, ed., Kluwer Publ. Co (1994), 265–301.

[SS] R. H. Schonmann and S. B. Shlosman, *Wulff droplets and the metastable relaxation of kinetic Ising models*, Commun. Math. Phys. (to appear).

Roberto H. Schonmann
Mathematics Department
UCLA
Los Angeles CA 90095
U.S.A.

Space of Local Fields in Integrable Field Theory and Deformed Abelian Differentials

Feodor A. Smirnov[1]

Abstract. In this talk I consider the space of local operators in integrable field theory. This space allows two different descriptions. The first of them is due to conformal field theory which provides a universal picture of local properties in quantum field theory. The second arises from counting solutions to form factors equations. Considering the example of the restricted Sine-Gordon model I show that these two very different descriptions give the same result. I explain that the formulae for the form factors are given in terms of deformed hyper-elliptic integrals. The properties of these integrals, in particular the deformed Riemann bilinear relation, are important for describing the space of local operators.

1 Quantum field theory in two dimensions.

Consider a massive relativistic quantum field theory (QFT) in two dimensional Minkowski space M^2. For $x = (x_0, x_1) \in M^2$ we put $x^2 = x_0^2 - x_1^2$. Let us take for simplicity the case when there is only one stable particle of mass m in the spectrum. To this particle we associate the creation-annihilation operators $a^*(\beta), a(\beta)$ where the rapidity β parameterizes the energy-momentum of particle: $p_0(\beta) = m\cosh\beta$, $p_1(\beta) = m\sinh\beta$. The only non-vanishing commutator is

$$[a(\beta_1), a^*(\beta_2)] = \delta(\beta_1 - \beta_2)$$

The space of states of the theory is the Fock space created by the action of an arbitrary finite number of operators $a^*(\beta)$ on the vacuum $|0\rangle$ which is annihilated by $a(\beta)$. We denote this space by $\mathcal{H}_p$. The action of the operators of energy and momentum P_μ in $\mathcal{H}_p$ is defined by $P_\mu|0\rangle = 0$, $[P_\mu, a^*(\beta)] = p_\mu(\beta)a^*(\beta)$.

In local QFT there exist local operators $\mathcal{O}_i(x) = e^{iP_\mu x_\mu}\mathcal{O}_i(0)e^{-iP_\mu x_\mu}$ acting in the space $\mathcal{H}_p$ and satisfying

$$[\mathcal{O}_i(x), \mathcal{O}_j(x')] = 0 \qquad \text{for} \qquad (x - x')^2 < 0$$

Obviously, these local operators create a linear space which will be denoted by $\mathcal{H}_o$. The Lehmann-Symanzik-Zimmerman axiomatic requires the existence of two special local operators. One of them is the symmetric energy-momentum tensor $T_{\mu\nu}$ such that $\partial_\mu T_{\mu\nu} = 0$ and $P_\mu = \int T_{\mu 0}(x)dx_1$. The other one is the interpolating field $\phi(x)$ weakly

[1] Membre du CNRS

approaching, when $x_0 \to \pm\infty$, the free "out" and "in" fields constructed via the creation-annihilation operators $a^*_{out}(\beta), a_{out}(\beta)$, $a^*_{in}(\beta), a_{in}(\beta)$. It is required that they are unitary equivalent: $a_{out}(\beta) = \mathcal{S}a_{in}(\beta)\mathcal{S}^{-1}$ with the unitary operator $\mathcal{S}$ (S-matrix) which leaves invariant the vacuum and one particle state, and commutes with P_μ. We identify the original operator $a(\beta)$ with $a_{in}(\beta)$.

This axiomatic has an obvious generalization to the case when the particle has internal (isotopic) degrees of freedom, and to the case of fermionic statistic or even generalized statistic, the latter case is also possible in two dimensions.

Let us consider in some details the space of local operators $\mathcal{H}_o$. The general philosophy teaches us that in order to understand the structure of this space one has to consider the ultra-violet (short-distance) limit of the original QFT. At least intuitively this idea is quite natural. The ultra-violet limit of massive QFT is described by a certain conformal field theory (CFT). The spaces of local operators of two theories must coincide, and, since the CFT in two dimensions allows in many cases a complete solution [1], we get a description of "universality classes" of two-dimensional QFT.

In the conformal case the theory essentially splits into two chiral sectors, which means that any operator $\mathcal{O}(x)$ can be rewritten as $\mathcal{O}^-(x_-)\mathcal{O}^+(x_+)$ where $x_\pm = x_0 \pm x_1$ are light-cone coordinates. The space of local operators in CFT is described in terms of two Virasoro algebras with generators $\mathcal{L}^\pm_n$ satisfying

$$[\mathcal{L}^\pm_m, \mathcal{L}^\pm_n] = (m-n)\mathcal{L}^\pm_{m+n} + c\,\delta_{m,-n}\tfrac{n^3-n}{12}$$

where the central charge c is an important characteristic of the theory. These Virasoro algebras act on the space $\mathcal{H}_o$ which happens to be organized as follows. There are primary fields ϕ_m satisfying

$$\mathcal{L}^\pm_n\phi_m = 0,\ n<0, \qquad \mathcal{L}^\pm_0\phi_m = \Delta_m\phi_m$$

where Δ_m is the scaling dimension of primary field. Different local operators are obtained by acting with $\mathcal{L}^\pm_{-n}$ on the primary fields. So, the one has

$$\mathcal{H}_o = \bigoplus_m W^-_m \otimes W^+_m$$

where W_m is a Verma module of the Virasoro algebra.

In this talk I shall consider a particular example of CFT with $c<1$. The coupling constant ξ which we use is related to c as follows

$$c = 1 - \tfrac{6}{\xi(\xi+1)}$$

Considering the coupling constant in generic position we have infinitely many primary fields ϕ_m , $m \geq 0$ with scaling dimensions $\Delta_m = -\frac{m}{2} + \frac{m}{2}(\frac{m}{2}+1)\xi$. We shall concentrate on one chirality considering only one Virasoro algebra with generators $\mathcal{L}_n \equiv \mathcal{L}^-_n$. The Verma module W_m has a singular vector on level $m+1$. The irreducible representation of the Virasoro algebra is obtained by factorizing over the Verma submodule created over this singular vector. The vectors from this submodule are called "null-vectors". It must be emphasized that the process of factorizing over the null-vectors has the dynamical meaning of imposing the equations of motion. The latter statement can be clearly

understood in the classical limit $\xi \to 0$ when the chiral CFT gives the classical Korteweg-de-Vris (KdV) hierarchy with second Poisson structure. The space of local operators turns into the space of functions on the phase space of KdV. It is shown in [2] that the null-vectors in that case provide all the equations of motion of the KdV hierarchy.

Let us return to massive QFT. Consider a local operator $\mathcal{O}(x)$ and define

$$f_{\mathcal{O}}(\beta_1, \cdots, \beta_n) = \langle 0 |\, \mathcal{O}(0)\, a^*(\beta_1) \cdots a^*(\beta_n) |0\rangle$$

where $\beta_1 < \cdots < \beta_n$. To other ranges of β's the function $f_{\mathcal{O}}$ is continued analytically. The function $f_{\mathcal{O}}$ is called form factor. The matrix elements of a local operator between two arbitrary states of $\mathcal{H}_p$ can be obtained by certain analytical continuation of the form factors due to crossing symmetry. The dependence on x can be taken into account trivially because the matrix element is taken between the eigen-states of the energy-momentum. Thus the form factors define the local operator completely. On the other hand the set of form factors define a pairing between the spaces $\mathcal{H}_o$ and $\mathcal{H}_p$.

2 INTEGRABLE FIELD THEORY.

The problem of finding the form factors of local operators for any massive QFT looks rather hopeless. However, in the special case of integrable field theory (IFT) this problem can be solved. In IFT the scattering is factorizable which means that every scattering process is reduced to two-particle scattering [3]. The two-particle S-matrix $S(\beta_1 - \beta_2)$ depends analytically on the difference of rapidities. As it has been already said the particle can carry internal degrees of freedom lying in finite-dimensional isotopic space. In that case $S(\beta_1 - \beta_2)$ is an operator acting in the tensor product of the isotopic spaces attached to the particles scattered. The S-matrix must satisfy certain requirements, the most important of which being the Yang-Baxter equation [4].

Consider now the form factors. The first examples of exact form factors in IFT are given in [5]. I gave a complete solution of the problem in a series of papers (partly in collaboration with A.N. Kirillov) summarized in the monograph [6]. If the particles have internal degrees of freedom the form factor takes values in the tensor product of isotopic spaces. It is convenient to consider the form factors as row-vectors. Then we act from the right by the operators like $S(\beta_i - \beta_j)$ (which act non-trivially only in the tensor product of i-th and j-th spaces). It has been shown that for the operator $\mathcal{O}$ to be local it is necessary and sufficient that the following requirements are satisfied [6]:

1. ANALYTICITY. The form factor $f_{\mathcal{O}}(\beta_1, \cdots, \beta_n)$ is a meromorphic function of all its arguments in the finite part of the complex plane.

2. SYMMETRY.

$$f_{\mathcal{O}}(\beta_1, \cdots \beta_i, \beta_{i+1} \cdots, \beta_n) S(\beta_i - \beta_{i+1}) = f_{\mathcal{O}}(\beta_1, \cdots \beta_{i+1}, \beta_i \cdots, \beta_n) \qquad (1)$$

3. TOTAL EUCLIDEAN ROTATION.

$$f_{\mathcal{O}}(\beta_1, \cdots, \beta_{n-1}, \beta_n + 2\pi i) = f_{\mathcal{O}}(\beta_n, \beta_1, \cdots, \beta_{n-1}) \qquad (2)$$

3. ANNIHILATION POLE. In the absence of bound states there are no other singularities in variable β_n in the strip $0 < \mathrm{Im}\beta_n < 2\pi$ but simple poles at the points $\beta_n = \beta_j + \pi i$.

The residue at the one of them ($\beta_n = \beta_{n-1} + \pi i$) is given below, other residues can be obtained from the symmetry property.

$$2\pi i \,\mathrm{res} f_{\mathcal{O}}(\beta_1, \cdots \beta_{n-1}, \beta_n) = \\ = f_{\mathcal{O}}(\beta_1, \cdots \beta_{n-2}) \otimes c_{n-1,n} \, (I - S(\beta_{n-1} - \beta_1) \cdots S(\beta_{n-1} - \beta_{n-2})) \qquad (3)$$

where $c_{n-1,n}$ is a certain vector from the tensor product of n-th and $(n-1)$-th isotopic spaces which is canonically related to the S-matrix.

Two comments are in order here. First, clearly the IFT is completely defined by the S-matrix in agreement with the general idea of Heisenberg. Second, the space $\mathcal{H}_o$ is in one-to-one correspondence with the space of solutions of the system of linear equations (1, 2, 3). So, we have to establish the relation of this description to the one given by CFT.

Let me consider my favorite example of IFT which is the restricted Sine-Gordon model (RSG)[7]. I will not give the traditional Lagrangian definition of the model, instead I shall present the S-matrix which, as it has been said, defines the IFT completely. The particles in RSG are two-component (soliton-antisoliton), so, the S-matrix is an operator acting in $\mathbb{C}^2 \otimes \mathbb{C}^2$ as follows

$$S(\beta) = S_0(\beta) \left(e^{-\frac{\beta}{\xi}} \, R(q) - e^{\frac{\beta}{\xi}} \, \widehat{R}(q)^{-1} \right) \qquad (4)$$

where $\beta = \beta_1 - \beta_2$, ξ is a coupling constant, $S_0(\beta)$ is certain c-number multiplier which is not very relevant for our goals. The matrix $R(q)$ is the R-matrix of the quantum group $U_q(sl_2)$ [8] with $q = \exp(\frac{2\pi i}{\xi})$ acting in the tensor product of two-dimensional representations:

$$R(q) = q^{\frac{1}{4}(\sigma^3 \otimes \sigma^3 + 1)} \left(I + (q^{\frac{1}{2}} - q^{-\frac{1}{2}}) \sigma^+ \otimes \sigma^- \right),$$

where $\sigma^3, \sigma^\pm$ are Pauli matrices. Finally, $\widehat{R}(q) = PR(q)$, where P is the operator of permutation. The S-matrix (4) gives the famous Sine-Gordon (SG) S-matrix found by Zamolodchikov [9].

The RSG model is a sector of SG model. Let us consider the isotopic spaces as spaces of two-dimensional representations of the quantum group. The S-matrix (4) is written in a manifestly $U_q(sl_2)$-invariant form. If one introduces the action of $U_q(sl_2)$ in the space of particles of the SG model, restriction to RSG corresponds to considering $U_q(sl_2)$-invariant subspace. This restriction looks at the first glance as a kinematical one, but it has important dynamical consequences. The space $\mathcal{H}_o$ of RSG corresponds to $U_q(sl_2)$-invariant solutions of the equations (1, 2, 3). From certain physical consideration we know that this space must coincide with the space of operators of CFT with $c = 1 - \frac{6}{\xi(\xi+1)}$ defined above.

One remark should be made. I have said that particles in two dimensional QFT can have generalized statistics which means that their interpolating fields are quasi-local (some phases appear in the commutation relations on the space-like interval). In that case the equations (1, 2, 3) are satisfied for the operators which are not only local, but also mutually local with the interpolating fields, otherwise some minor modification is needed. This is the situation which takes place in RSG model: solitons are particles with generalized statistics. Only the primary fields ϕ_{2m} and their Virasoro descendents are mutually local with the interpolating fields of solitons. For simplicity we shall take for $\mathcal{H}_o$ of RSG the space span by these "truely local" operators.

3 DEFORMED HYPER-ELLIPTIC DIFFERENTIALS.

The formulae for the form factors of the RSG model are given in terms of deformed hyper-elliptic integrals. Let me explain what these integrals are. Consider a hyper-elliptic Riemann surface of genus $n-1$ defined by the equation $c^2 = p(a)$ with $p(a) = \prod_{j=1}^{2n}(a - b_j)$. Take the abelian differentials regular everywhere except at the two points lying over the point $a = \infty$, and having no simple poles. Up to exact forms there are $2n-2$ such differentials, everyone of them is written in the form

$$\omega = \frac{l(a)}{\sqrt{p(a)}} da \tag{5}$$

with some polynomial $l(a)$. Introduce the intersection form

$$\omega_1 \circ \omega_2 = \sum_{a=\infty} \operatorname{res}\ (\omega_1 \Omega_2)$$

where $d\Omega = \omega$. The basis of dual differentials can be constructed as follows. Consider the anti-symmetric polynomial of two variables:

$$c(a_1, a_2) = \sqrt{p(a_1)} \frac{\partial}{\partial a_1} \left(\frac{\sqrt{p(a_1)}}{a_1 - a_2} \right) - \sqrt{p(a_2)} \frac{\partial}{\partial a_2} \left(\frac{\sqrt{p(a_2)}}{a_2 - a_1} \right)$$

For any decomposition of this polynomial of the form

$$c(a_1, a_2) = \sum_{i=1}^{n-1} (r_i(a_1) s_i(a_2) - r_i(a_2) s_i(a_1))$$

the differentials η_i and ζ_i defined by using r_i and s_i respectively in equation (5) are dual:

$$\eta_i \circ \eta_j = 0,\ \zeta_i \circ \zeta_j = 0,\ \eta_i \circ \zeta_j = \delta_{ij}$$

Consider the canonical homology basis with a-cycles α_i and b-cycles β_i. The Riemann bilinear relation (as A. Nakayashiki pointed out to me, the hyper-elliptic case was found by Weierstrass) says that the matrix of periods

$$\mathcal{P} = \begin{pmatrix} \int_\alpha \eta\,, & \int_\beta \eta \\ \int_\alpha \zeta\,, & \int_\beta \zeta \end{pmatrix}$$

belongs to the symplectic group $Sp(2n-2)$.

Now I am going to describe a deformation of these abelian differentials which is needed for the description of RSG form factors. Obviously only tensor product of even number of two-dimensional representations can have a $U_q(sl_2)$-invariant subspace. So, we have only form factors with even number of particles with rapidities $\beta_1, \cdots, \beta_{2n}$. Let us introduce the notations $b_j = \exp(\frac{2\beta_j}{\xi})$, $B_j = \exp(\beta_j)$. Consider two polynomials $l(a)$ and $L(A)$ which can depend respectively on b_j and B_j as parameters. We define the following pairing for these polynomials [10]:

$$\langle l, L \rangle = \int_{-\infty}^{\infty} \Phi(\alpha) l(a) L(A) d\alpha \tag{6}$$

where $a = \exp(\frac{2\alpha}{\xi})$, $A = \exp(\alpha)$. The function $\Phi(\alpha)$ satisfies the equations

$$\begin{aligned} p(aq)\Phi(\alpha + 2\pi i) &= q^2 p(a)\Phi(\alpha), & p(a) &= \prod(a - b_j) \\ P(-AQ)\Phi(\alpha + i\xi) &= QP(A)\Phi(\alpha), & P(A) &= \prod(A + B_j) \end{aligned} \tag{7}$$

where $Q = e^{i\xi}$. We require that $\Phi(\alpha)$ is regular for $0 < \mathrm{Im}\alpha < \pi$, that it behaves as aA when $\alpha \to -\infty$, and that it has the following asymptotics when $\alpha \to +\infty$:

$$\begin{aligned} &\Phi(\alpha) \sim f(a)F(A), \\ &f(a) = a^{-(n-1)}\big(1 + \sum_{k>0} c_k a^{-k}\big), \qquad F(A) = A^{-(2n-1)}\big(1 + \sum_{k>0} C_k A^{-k}\big) \end{aligned}$$

where the coefficients c_k, C_k can be found using (7). These requirements fix the function $\Phi(\alpha)$ completely, the explicit formula is available, but we shall not need it. The following functionals are defined for arbitrary polynomials $l(a)$ and $L(A)$:

$$\mathbf{r}l \equiv \mathrm{res}_{a=\infty}(a^{-1}l(a)f(a)), \qquad \mathbf{R}L \equiv \mathrm{res}_{A=\infty}(A^{-1}L(A)F(A)) \tag{8}$$

What is the relation between the pairing $\langle l, L\rangle$ and the hyper-elliptic integrals? Take the limit when $\xi \to \infty$ keeping b_j finite. In RSG model this is the strong coupling limit. In this limit the integral (6) goes asymptotically to the period of the differential defined by $l(a)$ (5) over a cycle which is fixed by the polynomial $L(A)$. Thus we have a deformation of hyper-elliptic integrals in which the differentials and the cycles enter in much more symmetric way than they do classically. We shall call l and L respectively q-form and q-cycle. The striking feature of this deformation is that it preserves all the important properties of classical hyper-elliptic integrals. Let me explain this point.

After appropriate regularization [6], the pairing $\langle l, L\rangle$ can be defined for every pair of polynomials l and L satisfying $\mathbf{r}l = 0$, $\mathbf{R}L = 0$. However, only a finite number of them give really different results because it can be shown that the value of the integral does not change if we add to l or L polynomials of the form

$$\begin{aligned} &\mathbf{d}[h](a) \equiv a^{-1}\left(p(a)h(a) - p(aq^{-1})h(aq^{-2})\right) \\ &\mathbf{D}[H](A) \equiv A^{-1}\left(P(A)H(A) - P(AQ)H(-A)\right) \end{aligned} \tag{9}$$

where the polynomials h and H are arbitrary. The first polynomial from (9) can be considered as an exact q-form and the second one as a q-boundary. It is easy to see that modulo (9) we have $2n-2$ q-forms and $2n-2$ q-cycles, so, the dimensions of cohomologies and homologies do not change after the deformation.

Consider now two anti-symmetric polynomials:

$$\begin{aligned} c(a_1, a_2) =& \frac{p(a_1)}{a_1(a_1q - a_2)} - \frac{p(a_1q^{-1})}{a_1(a_1q^{-1} - a_2)} - \\ &- \frac{p(a_2)}{a_2(a_2q - a_1)} + \frac{p(a_2q^{-1})}{a_2(a_2q^{-1} - a_1)} \\ C(A_1, A_2) =& \frac{1}{A_1A_2}\Big(\frac{A_1 - A_2}{A_1 + A_2}(P(A_1)P(A_2) - P(-A_1)P(-A_2)) + \\ &+ (P(-A_1)P(A_2) - P(A_1)P(-A_2))\Big) \end{aligned} \tag{10}$$

Suppose that modulo exact q-forms and q-boundaries we have the decompositions:

$$c(a_1,a_2)=\sum_{i=1}^{n-1}\left(r_i(a_1)s_i(a_2)-r_i(a_2)s_i(a_1)\right)$$
$$C(A_1,A_2)=\sum_{i=1}^{n-1}\left(R_i(A_1)S_i(A_2)-R_i(A_2)S_i(A_1)\right)$$

then the following deformed Riemann bilinear relation holds [10].
PROPOSITION. *The matrix*

$$\mathcal{P}=\begin{pmatrix}\langle r,R\rangle, & \langle r,S\rangle\\ \langle s,R\rangle, & \langle s,S\rangle\end{pmatrix}$$

belongs to the symplectic group $Sp(2n-2)$.

4 EXACT FORM FACTORS AND SPACE OF OPERATORS.

The quantum group invariance means that the $2n$-particle form factors in the RSG model belong to $U_q(sl_2)$-invariant subspace of the tensor product $(\mathbb{C}^2)^{\otimes 2n}$. The dimension of this invariant subspace equals $\binom{2n}{n}-\binom{2n}{n-1}$. There is a nice coincidence of dimensions

$$\tbinom{2n}{n}-\tbinom{2n}{n-1}=\tbinom{2n-2}{n-1}-\tbinom{2n-2}{n-3}$$

where the RHS gives the dimension of $(n-1)$-th fundamental irreducible representation of $Sp(2n-2)$ (explicitly described later). This representation is naturally related to the construction of the previous section.

Consider the space $\mathfrak{h}_k$ of anti-symmetric polynomials of k variables $a_1,\cdots,a_k$. We can define the following operators acting between different $\mathfrak{h}_k$:
1. The operator $\mathbf{r}$ acting from $\mathfrak{h}_k$ to $\mathfrak{h}_{k-1}$ by applying the "residue" (8) to one argument, obviously $\mathbf{r}^2=0$.
2. For every $h\in\mathfrak{h}_1$ (a polynomial of one variable) define the operator $\mathbf{d}[h]$ acting from $\mathfrak{h}_{k-1}$ to $\mathfrak{h}_k$ by

$$(\mathbf{d}[h]l_{k-1})(a_1,\cdots,a_k)=\sum_{i=1}^{k}(-1)^i\mathbf{d}[h](a_i)\;l_{k-1}(a_1,\cdots,\widehat{a_i},\cdots,a_k)$$

3. The operator $\mathbf{c}$ acts from $\mathfrak{h}_{k-2}$ to $\mathfrak{h}_k$ by

$$(\mathbf{c}l_{k-2})(a_1,\cdots,a_k)=\sum_{i<j}^{k}(-1)^{i+j}c(a_i,a_j)\;l_{k-2}(a_1,\cdots,\widehat{a_i},\cdots,\widehat{a_j}\cdots,a_k)$$

Denote by $\widehat{\mathfrak{h}}_k$ the following subspace of $\mathfrak{h}_k$:

$$\widehat{\mathfrak{h}}_k=\mathrm{Ker}\big(\mathbf{r}\mid_{\mathfrak{h}_k\to\mathfrak{h}_{k+1}}\big)/\bigoplus_{h\in\mathfrak{h}_1}\mathrm{Im}\big(\mathbf{d}[h]\mid_{\mathfrak{h}_{k-1}\to\mathfrak{h}_k}\big)$$

The space $\widehat{\mathfrak{h}}_k$ is finite-dimensional of dimension $\binom{2n-2}{k}$. The action of the operator $\mathbf{c}$ can be restricted to the spaces $\mathfrak{h}_k$. We denote by $\mathfrak{h}^0_k$ the subspace:

$$\mathfrak{h}^0_k = \widehat{\mathfrak{h}}_k / \mathrm{Im}\big(\mathbf{c}\,|_{\widehat{\mathfrak{h}}_{k-2}\to\widehat{\mathfrak{h}}_k}\big)$$

which is isomorphic to $Sp(2n-2)$-irreducible submodule of maximal dimension in the space of anti-symmetric tensors of rank k. We are interested in the biggest possible: $\mathfrak{h}^0_{n-1}$.

The construction of form factors starts by describing a certain linear isomorphism:

$$(\mathbb{C}^2)^{\otimes 2n}_{\mathrm{inv}} \simeq \mathfrak{h}^0_{n-1} \tag{11}$$

of which we shall not need the explicit form. Using this isomorphism we identify every $e \in (\mathbb{C}^2)^{\otimes 2n}_{\mathrm{inv}}$ with a polynomial $l[e]_{n-1} \in \mathfrak{h}^0_{n-1}$.

Consider now the spaces $\mathfrak{H}_k$ of anti-symmetric polynomials $L_k(A_1, \cdots, A_k)$. The action of operators $\mathbf{R}$, $\mathbf{D}[H]$, $\mathbf{C}$ is defined in exactly the same way as the action of $\mathbf{r}$, $\mathbf{d}[h]$, $\mathbf{c}$ using the formulae (8) and (9). For $L_n \in \mathfrak{H}_n$ and for $l_{n-1} \in \mathfrak{h}^0_{n-1}$ define the pairing:

$$\langle l_{n-1}, L_n\rangle = \int_{-\infty}^{\infty} d\alpha_1 \cdots \int_{-\infty}^{\infty} d\alpha_{n-1} \prod_{i=1}^{n-1} \Phi(\alpha_i) \times l_{n-1}(a_1, \cdots, a_{n-1})(\mathbf{R}L_n)(A_1, \cdots, A_n) \tag{12}$$

The requirement $l_{n-1} \in \mathfrak{h}^0_{n-1}$ together with the existence of q-boundaries and of deformed Riemann bilinear relation leads to the following remarkable consequence. For arbitrary $H \in \mathfrak{H}_1$, $L_{n-1} \in \mathfrak{H}_{n-1}$ and $L_{n-2} \in \mathfrak{H}_{n-2}$

$$\mathbf{D}[H]L_{n-1} \simeq 0, \qquad \mathbf{C}L_{n-2} \simeq 0 \tag{13}$$

where $L_n \simeq 0$ means that for such L_n the integrals (12) vanish for any $l_{n-1} \in \mathfrak{h}^0_{n-1}$.

The form factors must satisfy three equations (1), (2), (3). Consider first the equations (1) and (2) only. Obviously, one can multiply any solution of these two equations by a quasi-constant, i.e. $2\pi i$-periodic symmetric function of β_j which is the same as symmetric Laurent polynomial of B_j. We have

PROPOSITION. *To every $L_n \in \mathfrak{H}_n$ corresponds a solution to (1), (2) belonging to $(\mathbb{C}^2)^{\otimes 2n}_{\mathrm{inv}}$:*

$$f^{L_n}(\beta_1, \cdots, \beta_{2n}) = \sum_e \langle l_{n-1}[e], L_n\rangle e$$

where the sum is taken over a basis of $(\mathbb{C}^2)^{\otimes 2n}_{\mathrm{inv}}$. These solutions span a vector space over the ring of quasiconstants, and the only possible linear dependence of solutions arises from relations (13).

Let me appeal to the strong coupling limit for explaining the meaning of this construction. In this limit the equations (1), (2) turn into certain linear differential equations. These linear differential equations are solved in terms of hyper-elliptic integrals (b_j are the branch points) and, naturally, different solutions are counted by different cycles. So, it is not a wonder that after the deformation the solutions are counted by L_n which have the meaning of deformed cycles as explained above.

With every local operator $\mathcal{O}$ we identify an infinite tower of polynomials $L[\mathcal{O}]_n$ such that

$$f_{\mathcal{O}}(\beta_1,\cdots,\beta_n) = f^{L_n[\mathcal{O}]}(\beta_1,\cdots,\beta_{2n})$$

The polynomials $L_n[\mathcal{O}]$ must be related for different n in order to satisfy the remaining equation (3). We have

PROPOSITION. *The form factors $f_{\mathcal{O}}(\beta_1,\cdots,\beta_n)$ satisfy (3) if and only if the anti-symmetric polynomials $L_n[\mathcal{O}](A_1,\cdots,A_n)$, which are at the same time symmetric Laurent polynomials of the parameters $B_1,\cdots,B_{2n}$, satisfy the recurrence relations:*

$$L_n[\mathcal{O}](A_1,\cdots,A_n|B_1,\cdots,B_{2n})|_{B_{2n}=-B_{2n-1}} \equiv \sum_{i=1}^{n-1}(-1)^i \prod_{j\neq i}(A_j^2 - B_{2n}^2)$$

$$\times\, L_{n-1}[\mathcal{O}](A_1,\cdots,\widehat{A_i},\cdots,A_n|B_1,\cdots,B_{2n-2}) \qquad \left(\mathrm{mod}\prod_{j=1}^{n}(A_j^2 - B_{2n}^2)\right) \quad (14)$$

i.e. the difference between LHS and RHS is divisible by $\prod(A_j^2 - B_{2n}^2)$ as polynomial of A_j.

Recall that the equation (3) concerns the residue at the pole $\beta_{2n} = \beta_{2n-1} + \pi i$ which corresponds to $B_{2n} = -B_{2n-1}$ and $b_{2n} = qb_{2n-1}$. In the strong coupling limit the branch points b_{2n} and b_{2n-1} approach each other, so, we arrive at a singularity of the moduli space. Thus the geometrical analogy of our construction is as follows. With every $2n$-particle space we associate the moduli space of hyper-elliptic curves, the lower moduli space is embedded into the upper one as its singularity. Equation (14) gives a rule for embedding of deformed homologies.

The solution to the relation (14) which describes the primary field ϕ_{2m} is

$$L_n[\phi_{2m}](A_1,\cdots,A_n|B_1,\cdots,B_{2n}) = \prod_{i<j}(A_i^2 - A_j^2)\prod A_i^{2m}\prod B_j^{-m}$$

One can multiply $L_n(\phi_{2m})$ by the polynomials

$$I_{2k-1}(B) = \sum B_j^{2k-1}, \qquad J_{2k}(A|B) = \sum A_i^{2k} - \frac{1}{2}\sum B_j^{2k}$$

which does not spoil the relation (14). It corresponds to the action of operators $\mathcal{I}_{2k-1}$ and $\mathcal{J}_{2k}$ in the space $\mathcal{H}_0$, for example, $L_n[\mathcal{J}_{2k}\mathcal{O}](A|B) = J_{2k}(A|B)L_n[\mathcal{O}](A|B)$. I put forward the following

CONJECTURE. *The space of operators span by $\mathcal{I}_{2k_1-1}\cdots\mathcal{I}_{2k_p-1}\mathcal{J}_{2k_1}\cdots\mathcal{J}_{2k_q}\phi_{2m}$ coincides with the Verma module of Virasoro algebra generated over the primary field.*

Let me say a few words about the meaning of this construction. In RSG model there is an infinite number of local integrals of motion I_{2k-1} which can be written in the form $I_{2k-1} = \int h_{2k}(x)dx_1$ with some local densities $h_{2k}(x)$. For any local operator $\mathcal{O}$ we define an operator $\mathcal{I}_{2k_1-1}\mathcal{O} = [I_{2k-1},\mathcal{O}]$ which is also local. The operator $\mathcal{I}_{2k-1}$ acting on $\mathcal{H}_p$ is the same as before because the eigen-value of I_{2k-1} on $2n$-particle state equals $I_{2k-1}(B)$. The operators $\mathcal{J}_{2k}$ describe certain transverse to the integrals of motion coordinates.

The Verma module of Virasoro algebra is span by the vectors $\mathcal{L}_{k_1} \cdots \mathcal{L}_{k_l} \phi_{2m}$. The operator $\mathcal{L}_0$ defines the grading in this space such that the degree of $\mathcal{L}_k$ equals k. It can be shown that the degrees of $\mathcal{I}_{2k-1}$ and $\mathcal{J}_{2k}$ with respect to the same grading equal respectively $2k-1$ and $2k$. So, the characters of the two graded spaces coincide which makes the above conjecture very plausible. There are other arguments in favour of this conjecture which I cannot explain here.

There is a crucial check for the above conjecture. It has been said that the Verma module of the Virasoro algebra is reducible: there is a submodule of null-vectors which corresponds to the equations of motion of the model. The question is whether it is possible to find these null-vectors describing the space $\mathcal{H}_o$ in terms of $\mathcal{I}_{2k-1}$ and $\mathcal{J}_{2k}$? This can be done because certain local operators vanishes due to the relations (13), moreover, the number of these operators is exactly the same as the number of null-vectors in the Verma module [2]. I think that this statement which links together two very different descriptions of the space $\mathcal{H}_o$ is a good point to finish this talk.

ACKNOWLEDGEMENT. I would like to thank O. Babelon and D. Bernard for fruitful collaboration and for help in preparing this talk.

REFERENCES

[1] A.A. Belavin, A.M. Polyakov, A.B. Zamolodchikov, Nucl. Phys. B241 (1984) p.333.

[2] O. Babelon, D. Bernard, F.A. Smirnov, Comm. Math. Phys. 186(1997) 601

[3] L.D.Faddeev, V.E.Korepin, Physics Reports 420 (1978) pp.1-78

[4] A.B.Zamolodchikov, Al.B.Zamolodchikov, Annals. Phys. 120 (1979) p.253;
B. Berg, M. Karowski, P. Weisz, V. Kurak, Nucl. Phys. 134B (1979) 125

[5] M. Karowski, P. Weisz, Nucl. Phys. 139B (1978) 455

[6] F.A. Smirnov, *Form Factors in Completely Integrable Models of Quantum Field Theory.* Adv. Series in Math. Phys. 14, World Scientific, Singapore (1992)

[7] F.A. Smirnov, Int. J. Mod. Phys. A4 (1989) 4213, *and* Nucl. Phys. 337B (1990) 156;
N.Yu Reshetikhin and F. Smirnov, Comm. Math. Phys. 131 (1990) 157
A. LeClair, Phys. Lett. 230B (1989) 103

[8] V.G. Drinfeld, *Quantum Groups*, Proc. of ICM, Berkeley, CA, New York: Academic Press (1986) 798

[9] A.B.Zamolodchikov, Pisma ZhETF 25 (1977) 499

[10] F.A. Smirnov, Lett.Math.Phys. 36 (1996) 267

Feodor A. Smirnov
Laboratoire de Physique Théorique et Hautes Energies,
Université Pierre et Marie Curie, Tour 16, 1-er étage, 4 place Jussieu,
75252 Paris, France, smirnov@lpthe.jussieu.fr
On leave from Steklov Mathematical Institute, St. Petersburg, Russia

Scaling Limit of Particle Systems, Incompressible Navier-Stokes Equation and Boltzmann Equation

Horng-Tzer Yau*

Abstract. We review recent work on derivations of the Euler, incompressible Navier-Stokes and Boltzmann equations from scaling limits of microscopic dynamics.

1991 Mathematics Subject Classification: Primary 82C22; Secondary 60K35,82C05, 82C10.
Keywords and Phrases: Euler equations, incompressible Navier-Stokes equations, Boltzmann equation, Hamiltonian dynamics, stochastic dynamics, Schrödinger equation.

I. Introduction

Macroscopic equations such as the Euler equations, Navier-Stokes equations or Boltzmann equation are usually derived through a continuum formulation of conservation of mass and momentum or in the last case, by idealizing the collision process. But, they also have a more fundamental origin in the microscopic equations of Newton or Schrodinger. The main question is whether this assertion can be put on a firm mathematical foundation and whether macroscopic concepts such as the viscosity, the nonlinearity, and the dissipation of the entropy can be understood microscopically. There are other important questions about many-body dynamics such as fluctuations, time-dependent correlations and behavior of tagged particles which are naturally formulated only on the microscopic level, but due to the restriction of the length of this review, we shall address only the first question here.

In statistical physics, continuum quantities such as density, velocity, and energy have microscopic versions which assume their macroscopic, deterministic values through the law of large numbers. Therefore, in order the equations describing the evolution of the macroscopic quantities to be exact, certain limits have to be taken, with suitably chosen scalings of space, time, and other macroscopic parameters of the systems. So the first step in the derivation of such equations is a choice of scaling. Denote coordinates by lower case letters (x, t) in the microscopic scale; by capital letters (X, T) in the macroscopic scale. We put the system in a cube of

* Partially supported by U. S. National Science Foundation grants 9703752

size L in d-dimensional space with periodic boundary condition and we will usually assume $d = 3$. Denote the particles by $(x_1, \cdots, x_N, v_1, \cdots, v_N)$ with the density (in the microscopic unit, i.e., number of particles per microscopic unit volume) $\rho = N/L^d$. Let ε be the ratio between the microscopic unit and the macroscopic unit (say, $\varepsilon \sim 10^{-8}$). There are typically three choices of scalings:

$$\left\{ \begin{array}{ll} \text{Grad} & \rho = \varepsilon, (X,T) := (x\varepsilon, t\varepsilon) \\ \text{Euler} & \rho = 1, (X,T) := (x\varepsilon, t\varepsilon) \\ \text{Diffusive} & \rho = 1, (X,T) := (x\varepsilon, t\varepsilon^2) \end{array} \right\} \Longrightarrow \left\{ \begin{array}{lll} & & \text{finite} \\ \text{collisions} & : & \varepsilon^{-1} \\ \text{per particle} & & \varepsilon^{-2} \end{array} \right\} \tag{1.1}$$

The Euler and diffusive limits will be referred to as hydrodynamic limits. The typical number of collisions is finite for the Grad limit; infinite in the hydrodynamic limits. Hence the Grad limit is the closest to free motion without collisions. Essentially due to this feature, O. Lanford [12] proved the convergence of the hard core billiards to the Boltzmann equation in the Grad limit in short time based on the BBGKY hierarchy. Lanford's work, though restrictive in many ways, remains the only rigorous result on the scaling limits of many-body Hamiltonian systems with no unproven assumptions.

II. Euler Equations

At present there is no rigorous derivation of Euler equations from Hamiltonian mechanics. Unlike the Grad limit, the Euler limit involves an infinite number of collisions and the typical behavior is governed by the stationary (equilibrium, invariant) states, which are assumed to be Gibbs in the famous Boltzmann Hypothesis. More precisely:

Boltzmann Hypothesis : The invariant (stationary) measures of many body classical dynamics are *Gibbs* $\sim e^{-\beta H}$. In particular, the typical velocity distributions of different particles are uncorrelated (Weak Boltzmann Hypothesis).

The Boltzmann Hypothesis is strictly speaking *incorrect* because there are singular invariant measures. We believe that these singular invariant measures can be removed by regularity assumption such as *finite specific entropy*, i.e., entropy per microscopic unit volume is finite. The following theorem is a joint work with S. Olla and S. Varadhan [15].

Theorem. *Assume the weak Boltzmann Hypothesis holds for invariant measures with finite specific entropy. Suppose the Euler equation has a smooth solution in* $[0,T]$. *Then the empirical density, velocity. and energy converge to the solution of the Euler equations in* $[0,T]$ *with probability one.*

Recall that classical dynamics are characterized by a Hamiltonian

$$H(x,v) = \frac{1}{2}\sum_{\alpha=1}^{N} \|v_\alpha\|^2 + \sum_{\alpha<\beta\le N} V(x_\alpha - x_\beta) \tag{2.1}$$

with V a two-body potential and the Liouville equation

$$\partial_t f_{N,t}(x_1, \cdots, x_N, v_1, \cdots, v_N) = \mathcal{L}^* f_{N,t} \tag{2.2}$$

where $f_{N,t}$ is the density (w.r.t. the standard Lebesgue measure) of the system at time t and the Liouville operator is given by

$$-\mathcal{L}^* = \mathcal{L} = \sum_{\alpha=1}^{N} \left[\frac{\partial H}{\partial v_\alpha} \frac{\partial}{\partial x_\alpha} - \frac{\partial H}{\partial x_\alpha} \frac{\partial}{\partial v_\alpha} \right]$$

with the adjoint taken w.r.t. the standard Lebesgue measure.

For a given configuration $\omega = (x_1, \cdots, x_N, v_1, \cdots, v_N)$ the empirical density and velocity (which rigorously speaking are measures) are defined by

$$\hat{\rho}_{\epsilon,\omega}(X) = N^{-1} \sum_{\alpha=1}^{N} \delta(X - \varepsilon x_\alpha)\,,$$

$$\hat{v}_{\varepsilon,\omega}(X) = N^{-1} \sum_{\alpha=1}^{N} \delta(X - \varepsilon x_\alpha) v_\alpha\,,$$

where δ is the standard delta function on Euclidean space. We shall say $\hat{\rho}_{\varepsilon,\omega(T/\varepsilon)}(X)$ has a density $\rho(X,T)$ if for any test function J on the unit torus,

$$\int J(\varepsilon x) \hat{\rho}_{\epsilon,\omega(T/\varepsilon)}(X) dX = N^{-1} \sum_{\alpha=1}^{N} J(\,\varepsilon\, x_\alpha(T/\varepsilon)\,) \to \int J(X) \rho(X,T) dX$$

as $\varepsilon \to 0$. Similarly for the velocity,

$$N^{-1} \sum_{\alpha=1}^{N} J(\,\varepsilon\, x_\alpha(T/\varepsilon)\,)\, v_\alpha(T/\varepsilon) \to \int J(X)\, (\rho v)(X,T) dX.$$

To obtain the Euler equation, we differentiate the velocity

$$\frac{d}{dT} \int J(X) \rho(X,T) v(X,T) dX \sim \varepsilon^{-1} \frac{d}{dt} N^{-1} \sum_{\alpha=1}^{N} J(\varepsilon x_\alpha) v_\alpha$$

$$= -(2N)^{-1} \sum_{\alpha=1}^{N} \varepsilon^{-1} J(\varepsilon x_\alpha) \frac{\partial H}{\partial x_\alpha} + \cdots$$

$$= -(2N)^{-1} \sum_{\alpha=1}^{N} \nabla J(\varepsilon x_\alpha) \underbrace{\sum_{\beta \neq \alpha} \frac{x_\alpha - x_\beta}{\varepsilon} \cdot (\nabla V) \left(\frac{x_\alpha - x_\beta}{\varepsilon} \right)}_{\text{MICRO CURRENT}} + \cdots \tag{2.3}$$

(the micro current appearing here is only a main term for illustration of the idea). Recall the Euler equations:

$$\frac{d\rho}{dt} + \nabla\,(\rho v) = 0$$
$$\frac{d\,(\rho v)}{dt} + \nabla\,[\rho v \otimes v + P] = 0$$
$$\frac{d\,(\rho e)}{dt} + \nabla\,[\rho e\, v - v\, P] = 0$$

Here the pressure P is a function of density, velocity and energy and is determined by the equation of state from the equilibrium Gibbs measure. So in order to obtain the Euler equations we need to show that

$$\text{MICRO CURRENT} \to \text{MACRO CURRENT}\,(= P(\hat{\rho}_{\varepsilon,\omega}, \hat{v}_{\varepsilon,\omega}, \hat{e}_{\varepsilon,\omega})) \tag{2.4}$$

in the limit $\varepsilon \to 0$. This equality is understood in the sense of law of large numbers w.r.t. the density of the systems $f_{N,t}$ at time t, i.e.,

$$\begin{aligned} N^{-1} \int f_{N,t}(\omega) \Big| \sum_{\alpha=1}^{N} \nabla J(\varepsilon x_\alpha) \\ \times \Big[\sum_{\beta \neq \alpha} \frac{x_\alpha - x_\beta}{\varepsilon} \cdot (\nabla V) \left(\frac{x_\alpha - x_\beta}{\varepsilon} \right) - P(\hat{\rho}_{\varepsilon,\omega}, \hat{v}_{\varepsilon,\omega}, \hat{e}_{\varepsilon,\omega}) \Big] \Big| d\omega \to 0 \end{aligned} \tag{2.5}$$

where $d\omega = dx_1 dv_1 \cdots dx_N dv_N$.

The density $f_{N,t}$ satisfies the Liouville equation (2.2). At the present time we have essentially no estimate on this equation and the required identity has not been proved. To appreciate the difficulties, we list a few comments on the Liouville equation:

• It conserves L^p norm and positivity (thus f_N can be considered as a probability density) but L^p norm is not useful since $\| f_N \|_p \sim e^{CN}$ which is a huge number.

• There is no elliptic theory for classical dynamics.

• The BBGKY method works only for perturbation of free dynamics and thus is only useful for the Grad limit for which $\rho \sim \varepsilon$.

Instead of approaching it via elliptic estimates or L^p theory, a useful way to establish (2.5) is to consider the ergodic property of the Hamiltonian systems. The key observation, due to Morrey [14], is that (2.5) holds if we replace $f_{N,t}$ by any Gibbs measure (with Hamiltonian H (2.1)), or more generally, if "locally" $f_{N,t}$ is a Gibbs measure of the Hamiltonian H. If we can prove that "locally" $f_{N,t}$ is a equilibrium measure with finite specific entropy, we have proved (2.5) provided that we assume the Boltzmann Hypothesis. So the proof of Theorem 2.1 consists of two main ingredients: 1. Prove that the weak Boltzmann hypothesis implies the Boltzmann hypothesis. 2. Clarify the precise meaning of the word "locally" and eliminate the possibility of meso-scale fluctuation. The method we used for 2 is the relative entropy method.

Recall that for any two probability densities the relative entropy is defined by

$$S(f|g) = \int f \log(f/g) d\omega$$

Suppose f_t is a solution of the Liouville equation and ψ_t is any density. Then

$$\partial_t S(f_t|\psi_t) = \int f_t \left\{ \psi_t^{-1} \left[\mathcal{L}^* - \partial_t \right] \psi_t \right\} d\omega \tag{2.6}$$

This identity can be checked easily from the Liouville equation. It also has a version for Markov processes:

$$\partial_t S(f_t|\psi_t) = -D(f_t|\psi_t) + \int f_t \left\{ \psi_t^{-1} \left[\mathcal{L}^* - \partial_t \right] \psi_t \right\} d\omega \tag{2.7}$$

where $D(f|\psi)$ is the Dirichlet form of f w.r.t. ψ and is nonnegative [21, 15]. Now recall the entropy inequality (or the Jensen inequality) which states that for any function W,

$$\int fW d\omega \leq S(f|\psi) + \log \int \psi \exp(W)\ d\omega$$

Thus from (2.6),

$$\partial_t S(f_t|\psi_t) \leq S(f_t|\psi_t) + \log \int \psi_t \exp\left\{\ \psi_t^{-1}\left[\mathcal{L}^* - \partial_t\right]\psi_t\ \right\} d\omega$$

If we have

$$N^{-1} \log \int \psi_t \exp\left\{\ \psi_t^{-1}\left[\mathcal{L}^* - \partial_t\right]\psi_t\ \right\} d\omega \to 0 \tag{2.8}$$

then the relative entropy can be controlled on the relevant time scale and this will imply the estimate (2.5) and thus the Euler equations. Note that the left hand side of (2.8) is independent of f_t so the remaining argument in [15] can be summarized as showing that (2.8) holds iff ψ_t is a local Gibbs state with density, velocity and energy chosen according to the Euler equations (Note: As it is, (2.8) is incorrect; some arguments using ergodicity of the Hamiltonian dynamics are also needed). This is essentially a dynamical variational approach because we solve the problem by guessing a good trial function which in this case is the local Gibbs state.

III The incompressible Navier-Stokes equations

The Navier-Stokes equations are the next order corrections to the Euler equations. In order to derive them one needs to show that

$$\text{MICRO CURRENT} \to \text{MACRO CURRENT} + \varepsilon\nu\nabla\hat{v}_{\varepsilon,\omega} + o(\varepsilon) \tag{3.1}$$

where the currents are given by (2.3) and (2.4) and ν is the viscosity. Since there is an ε appearing in the viscosity term, (3.1) is in a sense *the next order correction to the Boltzmann hypothesis*! From the expression for the micro current in (2.3), it is hard to even imagine how the viscosity correction arises. This difficulty was recognized decades ago by Dobrushin, Lebowitz, and Spohn, among others. Recent work [20, 7, 8, 13] has given us a good understanding of the nature of (3.1), though a rigorous proof from the Hamiltonian dynamics is still very far off.

The equation for the leading order terms of (3.1) is (2.4) and it holds w.r.t. Gibbs measures in the sense of law of large numbers. The difficulty to justify (2.4) rigorously for Hamiltonian dynamics (i.e. (2.5)) is to prove that the solutions to the Liouville equation are locally stationary and all stationary measures are Gibbs. On the other hand, one can check easily that (3.1) is *incorrect* w.r.t any Gibbs measures with Hamiltonian H. Indeed, (3.1) is a "dynamical identity". It can be interpreted physically via the linear response theory or the Green-Kubo formula (see [17] for an account). A more mathematical interpretation is through the fluctuation-dissipation equation which we now explain.

Roughly speaking, the fluctuation-dissipation equation states that

$$\text{MICRO CURRENT} \to \text{MACRO CURRENT} + \varepsilon\nu\nabla\hat{v}_{\varepsilon,\omega} + \varepsilon\mathcal{L}g + o(\varepsilon) \tag{3.2}$$

for some function $g(\omega)$, where $\mathcal{L}$ is the Liouville operator. In other words, (3.1) is correct only up to a quotient of the image of the Liouville operator. The image of the Liouville operator is understood as fluctuation, negligible in the relevant scale after *time average*: for any bounded function g

$$\varepsilon \int_0^t ds \int f_{s,N}(\omega)\,(\varepsilon\mathcal{L}g)(\omega)\,d\omega = \varepsilon^2 \int [f_{t,N} - f_{0,N}](\omega)\,g(\omega)\,d\omega \sim \varepsilon^2$$

and is thus negligible to the first order in ε, the relevant scale.

It is difficult to work on "next order correction" and thus we turn to the incompressible Navier-Stokes (INS) equations

$$\frac{\partial u}{\partial t} + u\cdot\nabla u = -\nabla p + \nabla\nu\nabla u, \qquad \nabla\cdot u = 0. \tag{3.3}$$

The INS equations are invariant under the *incompressible scaling*,

$$x \to \varepsilon x, \quad t \to \varepsilon^2 t, \quad u \to \varepsilon^{-1} u, \quad p \to \varepsilon^2 p, \tag{3.4}$$

under which (3.2) becomes

$$\text{MICRO CURRENT} \to \text{MACRO CURRENT} + \nu\nabla\hat{v}_{\varepsilon,\omega} + \mathcal{L}g \tag{3.5}$$

Notice that both the viscosity and the function g are unknown and (3.5) determines both. We interpret (3.5) as a decomposition of the space of microscopic currents into a direct sum of the space of macroscopic currents, the gradient of the velocity representing the dissipation and the image of the Liouville operator representing the fluctuation.

Equation (3.5) is extremely difficult to solve as it requires inversion of the Liouville operator. A class of more manageable *stochastic lattice gas* models were introduced in a joint work with R. Esposito and R. Marra [8]. Even for these, (3.5) requires the inversion of a nonsymmetric operator in infinite dimensions with a complex interaction. If the generator $\mathcal{L}$ is symmetric, i.e., the dynamics is reversible, (3.5) can be solved by formulating the problem in an appropriate space so that it reduces to inverting a self-adjoint operator. This formulation, due to S. Varadhan [20], is already quite sophisticated since the terms appearing in (3.5) do not live in a natural space. On the other hand, in order to obtain the INS equations, the dynamics has to retain essential features of the Hamiltonian dynamics; this forces us into nonzero drifts and therefore nonreversibility. The invertibility in the nonsymmetric case is very subtle [13]. Dimension comes into play, and we believe that (3.5) has no solution at all for dimension $d \le 2$.

In the models of [8] particles have velocities in a chosen finite set and at each site of the lattice at most one particle of each velocity is allowed. The dynamics consists of two parts: Random walks and binary collisions between particles. The random walk part of the dynamics requires only that particles with velocity v should have the mean drift v. The binary collisions conserve momentum. Note that conservation of energy is not important here because the INS equations are

equations of velocity alone. The combined dynamics should have good ergodic properties and also restore rotational symmetry in the limit. The restoration of the rotational symmetry is not trivial because the lattice structure breaks the symmetry. Sets of velocities and dynamics satisfying all the requirements can be found in [8].

The main result in [8] states that (3.5) has a solution (in a suitable sense) for $d \geq 3$ and if the INS equations have a strong solution up to a fixed time T then the rescaled empirical velocity densities (measures)

$$v_{\varepsilon,\omega}(X) := \varepsilon^{d-1} \sum_x \delta(X - \varepsilon x) \sum_v v\eta(x,v) \tag{3.6}$$

converges to that solution. Here $\eta(x,v) \in \{0,1\}$ is the number of particles of velocity v at site x. Notice the blowup of the velocity by ε^{-1} in accordance with the scaling (3.4).

The assumption that the INS equations have a strong solutions has a long history in their derivation from more basic models. Derivations of the INS equations from the Boltzmann equation go back already a century to Chapman, Enskog and Hilbert, and were made rigorous in the seventies [4,5]. However the removal of the smoothness assumption has not been so easy. A program [3] of deriving the weak (Leray) solutions from the DiPerna-Lions solutions of the Boltzmann equation remains far from complete, due to a lack of good estimates. Though it was believed that the analysis of particle systems would be even more difficult because they are essentially infinite dimensional, in a joint work with J. Quastel [16] we have been able to remove this obstacle.

Theorem 3.1. *Let P_ε be the distributions of the empirical momentum densities (3.6). Then P_ε are precompact (as a set of probability measures with respect to a suitable topology) and any weak limit is supported entirely on weak solutions of the INS equations satisfying the energy inequality.*

Theorem 3.1 is proven only for $d = 3$. The restriction $d \leq 3$ is for technical reasons; *the restriction $d \geq 3$, however, is intrinsic.* Since the macroscopic velocity is defined through the law of large numbers in statistical physics, it inherits a small fluctuation from the central limit theorem, which is of order $\varepsilon^{d/2}$. When we blow up the velocity by ε^{-1} in the incompressible limit (3.6), this term becomes of order one or larger for dimensions $d \leq 2$ and thus the macroscopic velocity is not well defined in this limit. Note that this argument applies to any dynamics including the Hamiltonian dynamics.

Though (3.5) determines the viscosity, it is important to have an independent characterization, traditionally expressed as a time integral of current-current correlation functions, which up to constants is given by:

$$\nu = \int_0^\infty \big\langle \text{ MICRO CURRENT } (t=0);\ \text{MICRO CURRENT } (t=s) \big\rangle\, ds \tag{3.7}$$

where $\langle f; g \rangle = \langle fg \rangle - \langle f \rangle \langle g \rangle$ is the correlation function and the expectation is w.r.t. lattice gas dynamics starting from *equilibrium*. This is called the Green-Kubo formula and is proved rigorously in [13, 8] for $d \geq 3$. For dimension $d \leq 2$,

the Green-Kubo formula (3.7) diverges, (3.5) has no solution, and the time scaling is faster than diffusive. We are thus forced to conclude that the two dimensional INS equations cannot be obtained as the incompressible limit of *any* microscopic dynamics.

A large deviation principle was also given in [16]. One main step in [16] is a proof of the energy estimate for the INS equations directly from the lattice gas dynamics by implimenting a renormalizatin group analysis. The techincally most demanding points, the large field problems in the standard field theory and the large fluctuation here, are controlled by the entropy method [10] and and the logarithmic Sobolev inequality [22]. The entropy method is an infinite dimensional version of the energy method in PDE; the logarithmic Sobolev inequality plays the role of the usual Sobolev inequalities.

IV Quantum Dynamics

Most problems concerning classical or stochastic dynamics have corresponding quantum versions. They are however often too difficult to study. The classical or stochastic dynamics are governed by the evolution of a probability density; the quantum dynamics by a complex wave function. Although both dynamics are linear, the physics in the quantum case is given by the square of the wave function, breaking the superposition law. Furthermore, the evolution of a wave function is determined by its phase which is very hard to control. We mention here a result on the quantum Lorentz gases [6] to give some flavor of quantum dynamics.

Classical Lorentz gases model a classical particle in an environment of fixed scatterers distributed randomly (or periodically). The question is the time evolution of this particle for a typical configuration of the scatterers. Denote by $\omega = (x_\alpha), \alpha = 1, \cdots, N$, the configuration of scatterers in a cube of width L. We are interested in the Grad limit (1.1) with $\varrho = N/L^d$ denoting the density of the scatterers. The typical number of collisions is now of the order $t\rho \sim 1$. It was proved in [9,19, 1] that its time evolution converges to a linear Boltzmann equation

$$\partial_T F_T(X,V) + V \cdot \nabla_X F_T(X,V) = \int \sigma(U,V)\,[F_T(X,U) - F_T(X,V)]\,dU, \quad (4.1)$$

where F is the phase space density and $\sigma(U,V)$ is the scattering cross section.

The quantum Lorentz gases can be obtained by simply replacing the classical dynamics by the quantum dynamics. More precisely, let $V_0(x)$ be a fixed "nice" function. The Schrödinger equation governing the quantum particle is given by

$$i\partial_t \psi_t = H_{N,L}\psi_t, \quad \psi_{t=0} = \psi_0, \quad (4.2)$$

where the Hamiltonian is given by

$$H_{N,L} = H := -\Delta/2 + V_\omega, \qquad V_\omega(x) = \sum_{\alpha=1}^{N} V_0(x - x_\alpha). \quad (4.3)$$

The classical phase space density of a wave function ψ is defined through the Wigner transform:

$$W_\psi(x,v) := \int \overline{\psi(x + z/2)}\psi(x - z/2)e^{ivz}dz.$$

The scaling is the same as in the classical case,

$$W_\psi^\varepsilon(X,V) := W_\psi(X/\varepsilon, V). \tag{4.4}$$

Notice that the velocity is not rescaled. The Wigner transform typically has no definite sign, and the associated Husimi function or coherent states are needed to define a positive density, but we will not go into these details here.

Let $\psi_{\omega,t}^\varepsilon$ be the solution to the Schrödinger equation (4.2), (4.3) with initial data ψ_0^ε. Suppose that the initial data is of the following form

$$\psi_0^\varepsilon(x) = \varepsilon^{3/2} h(\varepsilon x) e^{iu_0 x},$$

for some smooth functions h so that as $\varepsilon \to 0$ the rescaled Wigner transform $W_{\psi_0^\varepsilon}^\varepsilon(X,V)dXdV$ converges weakly to $|h(X)|^2\delta(V-u_0)dXdV =: F_0(X,V)dXdV$ as probability measures on R^{2d}. Then in dimension $d=3$ and for V_0 small enough (so that there is no bound state) our main result with L. Erdős [6] is that *for any* $T>0$,

$$EW_{\psi_{\omega,T/\varepsilon}^\varepsilon}^\varepsilon W(X,V)dXdV \to F_T(X,V)dXdV$$

weakly as $\varepsilon \to 0$ *and* $F_T(X,V)$ *satisfies the linear Boltzmann equation (3.6) with initial data* $F_0(X,V)$ *and effective collision kernel* σ *given by the quantum scattering operator of the potential* V_0.

A simple example illustrates the difference between the classical and the quantum dynamics. Suppose that the particle in a Lorentz gas has one collision. Classically we simply choose a scatterer and the particle collides with it. In quantum mechanics, we have from the Duhamel formula

$$\psi_t = e^{-itH}\psi_0 = e^{-itH_0}\psi_0 - i\int_0^t e^{-i(t-s)H_0} V e^{-isH_0}\psi_0 \, ds + \cdots$$

where V_ω is the potential given in (4.3) and $H_0 = -\Delta/2$. The one collision term is the second term on the right hand side which, for simplicity, we write as $\sum_{\alpha=1}^N \psi_{t,\alpha}$. Notice that instead of collision with a scatterer in classical dynamics, it is now a sum of collisions with all scatterers! Since we have to square the wave function to get physical quantities, we need to show that the overlaps (or interference) of off-diagonal terms

$$\langle\, \psi_{t,\alpha},\ \psi_{t,\beta}\,\rangle$$

are very small. Stationary phase methods show they are small, but the number of the off-diagonal terms is much larger than that of diagonal terms. The analysis of this competition is very simple in this first term but very complicated in higher order terms. It nevertheless can be carried out rigorously to all orders [18, 11]. However such results are restricted to the weak coupling limit (a semiclassical limit) and short time. Instead we renormalize the perturbation theory so that we can consider the Grad limit to obtain the quantum scattering kernel. Furthermore, we truncate the Duhammel formula and estimate the error terms to remove the short-time restriction and thus we obtain results global in time [6].

ACKNOWLEDGE: It is a great pleasure to thank J. Quastel for critical readings and commments of the manuscript.

BIBLIOGRAPHY

1 C, Boldrighini, L. Bunimovich and Y. Sinai: J. Stat. Phys. 32, 477-501, (1983).
2 C, Boldrighini, R. L. Dobrushin and Y. M. Suhov: J. Stat. Phys. 31 (1983) 577–616.
3 C. Bardos,F. Golse and D. Levermore: Comptes Rendus de l'Acad. Sci., Ser. 1, 11, 727–732, (1989).
4 R. E. Caflisch and G. C. Papanicolaou: Comm. Pure Appl. Math., 32, 589–616, (1979).
5 A. De Masi, R. Esposito and J. L. Lebowitz: Commun. Pure and Appl. Math., 42,1189–1214, (1989).
6 L. Erdős and H.-T. Yau: Linear Boltzmann Equation as Scaling Limit of Quantum Lorentz Gas, to appear in GT-UAB Conf. Proceedings.
7 R. Esposito and R. Marra: Jour. Stat. Phys., 74, 981, (1993).
8 R. Esposito, R. Marra and H.-T. Yau: Rev. Math. Phys., 6, 1233-1267 and Commun. Math. Phys, 182, 396-456, 1996.
9 G. Gallavotti: Nota inteerna n. 358, Univ. di Roma (1970).
10 Guo,M. Papanicolau, G.C., and Varadhan, S. R. S.: Comm. Math. Phys. 118, 31–59, (1988).
11 T. G. Ho, L. J. Landau and A. J. Wilkins: Rev. Math. Phys. 5, 209-298 (1992).
12 O. E. Lanford: Astérisque 40, 117-137 (1976).
13 C. Landim and H.-T. Yau: Prob. Related Field, 108, 321-356 (1997)
14 C. B. Morrey: Comm. Pure Appl. Math., 8, 279-290, (1955).
15 S. Olla, R. Varadhan and H.-T. Yau: Commun. Math. Phys., 155, 523–560, (1993).
16 J. Quastel and H.-T.Yau: Lattice gases, large deviations, and the incompressible Navier-Stokes equations, to appear in Ann. Math.
17 H. Spohn: *Large Scale Dynamics of Interactin Particles*, Springer-Verlag New York (1991).
18 H. Spohn: J. Stat. Phys., 17, p385 (1977).
19 H. Spohn: Commun. Math. Phys. 60, 277-290 (1978).
20 S. R. S. Varadhan, Taniguchi Symp.,p 75-130, ed. K. D. Elworthy and N. Ikeda, Longman Scientific & Technical, (1990).
21 H.-T. Yau: Lett. Math. Phys. 22, 63-80 (1991).
22 H.-T. Yau: Commun. Math. Phys, 181, 367-408, 1997.

Horng-Tzer Yau
Courant Institute of
mathematical sciences
New York University
New York, NY, 10012, USA.
yau@math.nyu.edu

Section 12

Probability and Statistics

In case of several authors, Invited Speakers are marked with a *.

Stochastic Coalescence

David J. Aldous[1]

Abstract. Consider N particles, which merge into clusters according to the rule: a cluster of size x and a cluster of size y merge at (stochastic) rate $K(x,y)/N$, where K is a specified rate kernel. This Marcus-Lushnikov model of coalescence, and the underlying deterministic approximation provided by the Smoluchowski coagulation equations, have an extensive scientific literature. A recent reformulation is the *general stochastic coalescent*, whose state space is the infinite-dimensional simplex (the state $\mathbf{x} = (x_i, i \geq 1)$ represents unit mass split into clusters of masses x_i), and which evolves by clusters of masses x_i and x_j coalescing at rate $K(x_i, x_j)$. Existing mathematical literature (Kingman's coalescent, component sizes in random graphs, fragmentation of random trees) implicitly studies certain special cases. Recent work has uncovered deeper constructions of special cases of the stochastic coalescent in terms of Brownian-type processes. Rigorous study of general kernels has only just begun, and many challenging open problems remain.

1991 Mathematics Subject Classification: 60J25, 60K35
Keywords and Phrases: continuum tree, entrance boundary, fragmentation, gelation, random graph, random tree, Smoluchowski coagulation equation.

1 Introduction

Our topic centers around two closely related models. The *Marcus-Lushnikov process* is the following finite-state continuous-time Markov process [17, 16].

Fix an integer $N \geq 1$ and a *rate kernel* $K(x,y) \geq 0$. Imagine N particles, originally separate, which merge into clusters according to the rule

$$\text{each pair of clusters, sizes } \{m_i, m_j\} \text{ say, coalesces into one cluster}$$

$$\text{of size } m_i + m_j \text{ at rate } K(m_i, m_j)/N.$$

The *general stochastic coalescent* [10] is the continuous-time Markov process whose state space is the infinite-dimensional simplex $\Delta = \{\mathbf{x} = (x_i) : x_i \geq 0, \sum_i x_i = 1\}$,

[1]Research supported by N.S.F. Grant DMS96-22859

where we imagine a state $\mathbf{x}$ as a fragmentation of unit mass into clusters of masses x_i, and the process evolves according to the rule

each pair of clusters, masses $\{x_i, x_j\}$ say, coalesces into one cluster

of mass $x_i + x_j$ at rate $K(x_i, x_j)$.

Provided the rate kernel K is *homogeneous* with some exponent γ

$$K(cx, cy) = c^\gamma K(x, y), \ 0 < c, x, y < \infty \tag{1}$$

we see that the Marcus-Lushnikov process is a special case of the general stochastic coalescent, by taking each particle to have mass $1/N$ and rescaling time.

There is a large literature in various science disciplines (e.g. physical chemistry [9]) on deterministic equations (see section 3) for coalescence. A lengthy survey of deterministic and stochastic models appears in [2]. In particular, there are three special cases which are now well understood: $K(x, y) = 1$ (Kingman's coalescent), $K(x, y) = x + y$ (the additive coalescent), $K(x, y) = xy$ (the multiplicative coalescent). The next five sections focus on five open problems, whose discussion will illustrate some of the known results.

2 The Feller property of the general stochastic coalescent

In making precise the verbal description of the general stochastic coalescent, one would like to prove it is a Markov process with some regularity properties, specifically the *Feller property* that the distribution at time t be weakly continuous as a function of the initial state. Intuitively, this should be true under very weak assumptions, e.g. that $K(x, y)$ is continuous and strictly positive. But a proof is elusive. Evans and Pitman [10] give a proof under stronger assumptions of Lipschitz continuity.

3 Deterministic limits

Studying $t \to \infty$ time asymptotics in these models isn't interesting: the mass all coalesces into a single cluster. Our remaining problems concern different sorts of asymptotics. In the Marcus-Lushnikov process write $\mathbf{ML}^{(N)}(x, t)$ for the (random) number of mass-x clusters at time t. One expects a weak law of large numbers, saying that as $N \to \infty$

$$N^{-1} \ \mathbf{ML}^{(N)}(x, t) \ \to^p \ n(x, t), \ x \geq 1, \ t \geq 0 \tag{2}$$

where the deterministic limit $n(x, t)$ is the solution of the *Smoluchowski coagulation equation*

$$\tfrac{d}{dt} n(x, t) = \tfrac{1}{2} \sum_{y=1}^{x-1} K(y, x-y) n(y, t) n(x-y, t) - n(x, t) \sum_{y=1}^{\infty} K(x, y) n(y, t) \tag{3}$$

with $n(x,0) = 1_{(x=1)}$. It is these equations which have been studied most intensively in the scientific community [9]. From the verbal description of the model we expect solutions to have the property that mass density is preserved:

$$m_1(t) \equiv \sum_{x=1}^{\infty} xn(x,t) = 1 \ \forall t. \tag{4}$$

This is true [23] under the assumption

$$K(x+y) = O(1+x+y) \tag{5}$$

but in general there might be a phase transition called *gelation*: $\exists T_{\text{gel}} < \infty$ such that

$$m_1(t) = 1, \ t \le T_{\text{gel}}; \quad m_1(t) < 1, \ t > T_{\text{gel}}.$$

The physical interpretation of gelation is that after the critical time, a strictly positive proportion of mass lies in infinite-mass clusters, the *gel*. Exact conditions on K for gelation or non-gelation are another open problem, but let us return to the question of proving (2), which provides conceptual justification for the deterministic approximation. Proving (2) for $t < T_{\text{gel}}$ is closely related to proving uniqueness of solutions of (3) up to T_{gel}. While this is not difficult under assumption (5), the gelling case seems intrinsically much harder, in that the natural techniques one tries to use would prove regularity of solutions for all time, whereas by definition a gelling kernel has solutions with a certain non-regularity property. Jeon [13] and Norris [18] contain the latest results on such questions.

4 The emerging giant cluster for a gelling kernel

The study of component sizes in the classical random graph process [7] is essentially the same as the study of the Marcus-Lushnikov process with $K(x,y) = xy$. It has long been known that $T_{\text{gel}} = 1$ and that the $N \to \infty$ behavior around the critical point is as follows. For large A, at time $1 - A/N^{1/3}$ the largest cluster has size $\delta N^{2/3}$ for some small δ, and there are many clusters of similar size; at time $1 + A/N^{1/3}$ the largest cluster has size $DN^{2/3}$ for some large D, and the second-largest cluster has size $\delta N^{2/3}$. In other words, a distinguished *giant cluster* emerges over the time interval $1 \pm O(N^{-1/3})$ and it has size $\Theta(N^{2/3})$. See [15] for an exhaustive analysis. Rescaling size and time near the critical point leads to a limit process, the *standard multiplicative coalescent* [1], which is the $K(x,y) = xy$ case of the general multiplicative coalescent, except that one has to enlarge the state space to the l_2 space $\{\mathbf{x} = (x_i) : x_i \ge 0, \sum_i x_i^2 < \infty\}$. Remarkably, the marginal distribution of the standard multiplicative coalescent at a fixed time can be expressed in terms of excursion-lengths of a certain inhomogeneous reflecting Brownian motion.

No other gelling kernel is understood, so it is a matter of speculation to what extent this behavior holds for general gelling kernels. Heuristic arguments of van Dongen [21] suggest that for exponent $1 < \gamma < 2$ there should be an emerging giant cluster of size $N^{2/(1+\gamma)}$, but the only rigorous theory is some weak results in [3].

5 Self-similarity and entrance boundary

Consider a non-gelling kernel K which is homogeneous with some exponent $\gamma \leq 1$. It is natural [11] to seek a solution of the Smoluchowski coagulation equation which is asymptotically *self-similar* (also called *self-preserving* or *scaling*), in the sense that, as $t \to \infty$,

$$n(x,t) = s^{-2}(t)(\psi(x/s(t)) + o(1)) \text{ uniformly in } x \tag{6}$$

where $\psi(x) \geq 0$ satisfies $\int_0^\infty x\psi(x)\, dx = 1$. As at (4), we want the mean density $\int xn(x,t)\, dx$ to be constant in time, which explains the s^{-2} term in (6). Of course, the interpretation of (6) is that cluster mass scales with time as $s(t)$. Moreover one expects

$$s(t) \propto t^{\frac{1}{1-\gamma}},\ -\infty < \gamma < 1 \qquad s(t) \propto e^{wt} \text{ for some } w,\ \gamma = 1.$$

Outside the special cases, little is known, though under extra conditions it is probably not hard to prove a "tightness" condition weaker than existence of a self-similar limit (cf. [8] for this result in a slightly different model). For the Marcus-Lushnikov process, this question relates to the time period when typical clusters have size $o(N)$ but not $O(1)$. Reformulating in terms of the stochastic coalescent, we are interested in the time period when typical clusters have mass $o(1)$. When $\gamma < 1$ one expects there to be a unique version of the stochastic coalescent on $0 < t < \infty$ such that the maximum cluster size $\to 0$ as $t \to 0$, and that in this version the empirical distribution of cluster sizes tends (as $t \to 0$, and after rescaling by $s(t)$) to the self-similar distribution ψ. In Markov chain jargon, this is a question about the *entrance boundary*, and is easy to verify ([2] section 4.2) in the case $K(x,y) = 1$. Establishing it more generally seems difficult.

Paradoxically, the two other special cases (kernels xy and $x+y$) seem atypical, in that they have rich entrance boundaries, i.e. there are many different ways to start the process with all the mass in infinitesimally small clusters. See [4, 6] for detailed studies.

6 Connections with d-dimensional models

Our models are "mean-field", in that they do not track positions and velocities of particles in d-dimensional space. This does not mean the models are completely divorced from physical reality. Rather, the details of the physical process under study are used to calculate the form of the rate kernel $K(x,y)$. Perhaps the most interesting case to a probabilist is the original 1916 setting of Smoluchowski [22], who considered particles diffusing under thermal noise, i.e. performing physical Brownian motion in three dimensions, and coalescing upon contact. In this case the appropriate kernel in the mean-field model turns out to be

$$K(x,y) = (x^{1/3} + y^{1/3})(x^{-1/3} + y^{-1/3}).$$

The second term reflects the faster diffusion of smaller particles, the first term reflects their smaller cross-section and hence smaller chance of touching. It is natural to conjecture that, in the full model of spherical masses diffusing by Brownian

motion in 3 dimensions and coalescing upon contact (and relaxing to spheres), as time goes to infinity the distribution approximates a Poisson spatial distribution with rescaled mass distribution following the law ψ at (6) predicted by the mean field theory. This has apparently never been studied rigorously, though a less realistic model is treated by Lang and Nguyen [14].

7 The standard additive coalescent

Lest talking about open problems makes it seem little is known, let me end by mentioning a positive result. Cayley's formula says there are N^{N-2} trees on N labeled vertices. Pick such a tree T_∞ at random. To the edges e of T_∞ attach independent exponential(1) r.v.'s ξ_e. Write $F(t)$ for the forest obtained from T_∞ by retaining only the edges e with $\xi_e \le t$. Write $\mathbf{Y}^{(N)}(t)$ for the vector of sizes of the trees comprising $F(t)$. It can be shown that $(\mathbf{Y}^{(N)}(t); 0 \le t < \infty)$ is the Marcus-Lushnikov process associated with the additive kernel $K(x,y) = x + y$. This construction was apparently first explicitly given by Pitman [20], although various formulas associated with it had previously been developed in combinatorics [19, 24] and statistical physics [12]. What is remarkable is that one can take $N \to \infty$ limits in this construction. The (rescaled) limit of the discrete tree is the *continuum random tree* (CRT); the analog of cutting edges is placing marks according to a Poisson process of intensity e^{-t} along the skeleton of the CRT. Cutting the mass-1 CRT at these marks splits it into subtrees of finite mass; write $\mathbf{X}(t)$ for the vector of masses of these subtrees at time t. Then (as we expect by analogy with the discrete case above) the process $(\mathbf{X}(t), -\infty < t < \infty)$ evolves as the stochastic coalescent for $K(x,y) = x+y$. This process, the *standard additive coalescent*, is studied in detail in [5]. The CRT itself can be constructed from Brownian excursion, so ultimately the construction of the standard additive coalescent uses only Brownian and Poisson ingredients.

Acknowledgements. This project owes much to joint work and ongoing discussions with Jim Pitman, Steve Evans and Vlada Limic.

References

[1] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854, 1997.

[2] D. J. Aldous. Deterministic and stochastic models for coalescence: a review of the mean-field theory for probabilists. To appear in *Bernoulli*. Available via homepage http://www.stat.berkeley.edu/users/aldous, 1997.

[3] D. J. Aldous. Emergence of the giant component in special Marcus-Lushnikov processes. *Random Structures and Algorithms*, 12:179–196, 1998.

[4] D. J. Aldous and V. Limic. The entrance boundary of the multiplicative coalescent. *Electron. J. Probab.*, 3:1–59, 1998.

[5] D. J. Aldous and J. Pitman. The standard additive coalescent. Technical Report 489, Statistics Dept, U.C. Berkeley, 1997. To appear in Ann. Probab.

[6] D. J. Aldous and J. Pitman. Inhomogeneous continuum random trees and the entrance boundary of the additive coalescent. In Preparation, 1998.

[7] B. Bollobás. *Random Graphs*. Academic Press, London, 1985.

[8] J. M. C. Clark and V. Katsouros. Stable growth of a coarsening turbulent froth. To appear in J. Appl. Probab., 1998.

[9] R. L. Drake. A general mathematical survey of the coagulation equation. In G.M. Hidy and J.R. Brock, editors, *Topics in Current Aerosol Research (Part 2)*, volume 3 of *International Reviews in Aerosol Physics and Chemistry*, pages 201–376. Pergammon, 1972.

[10] S. N. Evans and J. Pitman. Construction of Markovian coalescents. Technical Report 465, Dept. Statistics, U.C. Berkeley, 1996. Revised May 1997. To appear in *Ann. Inst. Henri Poincaré*.

[11] S. K. Friedlander and C. S. Wang. The self-preserving particle size distribution for coagulation by Brownian motion. *J. Colloid Interface Sci.*, 22:126–132, 1966.

[12] E. M. Hendriks, J. L. Spouge, M. Eibl, and M. Shreckenberg. Exact solutions for random coagulation processes. *Z. Phys. B - Condensed Matter*, 58:219–227, 1985.

[13] I. Jeon. Existence of gelling solutions for coagulation fragmentation equations. *Commun. Math. Phys.*, to appear, 1998.

[14] R. Lang and X. X. Nguyen. Smoluchowski's theory of coagulation in colloids holds rigorously in the Boltzmann-Grad limit. *Z. Wahrsch. Verw. Gebeite*, 54:227–280, 1980.

[15] Svante Janson ; Donald E. Knuth ; Tomasz Łuczak and B. Pittel. The birth of the giant component. *Random Structures and Algorithms*, 4:233 – 358, 1993.

[16] A. A. Lushnikov. Certain new aspects of the coagulation theory. *Izv. Atmos. Ocean Phys.*, 14:738–743, 1978.

[17] A. H. Marcus. Stochastic coalescence. *Technometrics*, 10:133–143, 1968.

[18] J. R. Norris. Smoluchowski's coagulation equation: Uniqueness, non-uniqueness and a hydrodynamic limit for the stochastic coalescent. Statistical Lab., Cambridge. To appear in Adv. Appl. Probab., 1998.

[19] Yu. L. Pavlov. Limit theorems for the number of trees of a given size in a random forest. *Math. USSR Subornik*, 32:335–345, 1977.

[20] J. Pitman. Coalescent random forests. Technical Report 457, Dept. Statistics, U.C. Berkeley, 1996. Available via http://www.stat.berkeley.edu/users/pitman.

[21] P. G. J. van Dongen. Fluctuations in coagulating systems II. *J. Statist. Phys.*, 49:927–975, 1987.

[22] M. von Smoluchowski. Drei Vorträge über Diffusion, Brownsche Bewegung und Koagulation von Kolloidteilchen. *Physik. Z.*, 17:557–585, 1916.

[23] W. H. White. A global existence theorem for Smoluchowski's coagulation equation. *Proc. Amer. Math. Soc.*, 80:273–276, 1980.

[24] A. Yao. On the average behavior of set merging algorithms. In *Proc. 8th ACM Symp. Theory of Computing*, pages 192–195, 1976.

David J. Aldous
University of California
Department of Statistics
367 Evans Hall
Berkeley CA 94720-3860
U.S.A.
aldous@stat.berkeley.edu

State Space Collapse for Queueing Networks

Maury Bramson*

Abstract. The diffusive limits of queueing networks, known as heavy traffic limits, are a topic of continuing interest. An important ingredient in such work is the demonstration of state space collapse, which says that, in the limit, the process must live on an appropriate subspace. In [Wi98b], conditions are given under which state space collapse suffices for heavy traffic limits. Here, we discuss how state space collapse can be reduced to the problem of showing stability for the fluid model which is the deterministic analog of the queueing networks under consideration. We discuss specific cases, such as first-in first-out (FIFO) networks of Kelly type and certain static priority networks.

1991 Mathematics Subject Classification: Primary 60K25.

1 Introduction

Queueing networks constitute a general family of stochastic processes. In such models, one envisions "customers", such as people, products or some task to be performed, as being lined up at the different queues, or *stations*, of a network. When service of a customer at a station is completed, the customer moves to another station or leaves the network. Customers are also assumed to enter the network at various stations. This behavior will, in general, be random, with random variables corresponding to the choice of the next station when service at a station is completed, to the service times at stations, and to the interarrival times for customers entering the network. The evolution of such a network can be formulated as a continuous time Markov process. Two basic topics for queueing networks concern (1) obtaining criteria for when this Markov process is positive recurrent and (2) deciding when a sequence of networks, under diffusive scaling, converges to a reflecting Brownian motion. The criteria, in the two cases, are related. In this survey, we discuss both topics, with emphasis on the latter.

In many situations, it is important to permit more than one type of behavior for customers at a given station. (For example, patients at the receptionist's desk of a doctor's office will follow different rules, depending on whether they are checking in or out.) To allow for this, one distinguishes between different *classes* or *buffers* at a station; customers in the same class are subject to the same random rules for service and routing to the next class. A queueing network is *single class* if only one class is assigned to each station; otherwise, it is *multiclass*. One can

*The author was supported in part by the National Science Foundation.

also classify a queueing network based on whether or not it allows *feedback*, that is, output from a station can eventually become part of its input. This will occur, for example, when customers repeatedly visit a station along some preassigned route. Not surprisingly, answers for (1) and (2) above will be easiest to obtain for single class networks without feedback, and most difficult for multiclass networks with feedback.

The limits in (2), which are referred to as heavy traffic limits (HTL), have been investigated over the past three decades. Presently, HTL theory remains incomplete for multiclass networks. An important concept in this context is state space collapse (SSC). When SSC holds, customers in the different classes at a station occur (asymptotically) in fixed proportions. Such behavior enables one to generalize HTL results from single class networks to multiclass networks. This is done in [Wi98a]. It is also shown there that SSC follows from a somewhat weaker concept, multiplicative state space collapse (MSSC). This work is summarized in the article [Wi98b] in this volume.

Here, we discuss certain settings where one can demonstrate MSSC. These include well-known families of networks, such as first-in first-out networks of Kelly type. More generally, sufficient conditions for MSSC are given by the convergence of the solutions of fluid model equations which are associated with the networks in question. Such criteria hold, for example, for static priority networks.

The remainder of this article is organized as follows. In Section 2, we summarize the basic notation and definitions for queueing networks. Section 3 discusses the stability of queueing networks. Fluid models, the main tool for demonstrating stability, are introduced here. Section 4 discusses heavy traffic limits. Emphasis there is placed on recent work, in [Br98, Wi98a], which employs state space collapse.

2 Notation and Definitions

We make use here of concepts and notation employed in the article [Wi98b] in this volume, which the reader should consult for more detail. The variable j, $j = 1, \ldots, J$, will denote the stations of the network under consideration, and k, $k = 1, \ldots, K$, will denote the classes of the network. We use $\mathcal{C}(j)$ for the set of classes belonging to a station j, and $s(k)$ for the station to which class k belongs. At each station there is a single server. This server will always by *non-idling*, that is, the server will remain busy as long as there are customers present at its station.

The triple $(E(\cdot), V(\cdot), \Phi(\cdot))$ contains the random input of the network. The random vector $E(t) = \{E_k(t),\ k = 1, \ldots, K\}$ denotes the number of external arrivals by time t, $t \geq 0$, and $V(\mathbf{n}) = \{V_k(n_k),\ k = 1, \ldots, K\}$, $\mathbf{n} = (n_1, \ldots, n_K)$, denotes the cumulative service times for the first n_k customers in each class. The random matrix $\Phi(\mathbf{n})$, with rows $\Phi^k(n_k)$, $k = 1, \ldots, K$, denotes the cumulative routing process after n_k departures from each class k. As in [Wi98b], summands of these quantities are assumed, in each case, to be independent and identically distributed, with the different sequences also being independent of one another. The triple (α, M, P) is the deterministic analog of $(E(\cdot), V(\cdot), \Phi(\cdot))$. The mean vector α gives the external arrival rates at the different classes; the $K \times K$ diagonal

matrix M has the mean service times m_k as its diagonal entries. The matrix $P = \{P_{k\ell},\ k,\ell = 1,\dots,K\}$ gives the probability of a customer being routed from one class to another. In many interesting cases, the routing of the queueing network will be deterministic, with all customers entering the system at the same class, and moving along a given route, until they exit from the system. Such networks are referred to as *re-entrant lines*.

We will consider here only open networks, that is, networks for which the matrix

$$Q \stackrel{\text{def.}}{=} (I - P')^{-1} = I + P' + (P')^2 + \dots \tag{2.1}$$

is finite. (" $'$ " denotes the transpose.) This means that customers at any class are capable of ultimately leaving the network. To investigate these networks, one employs the solutions $\lambda_\ell,\ \ell = 1,\dots,K$, of the *traffic equations*

$$\lambda_\ell = \alpha_\ell + \sum_{k=1}^{K} \lambda_k P_{k\ell}, \tag{2.2}$$

or equivalently, in vector form, of $\lambda = \alpha + P'\lambda$. (All vectors in this article are to be interpreted as column vectors.) Solving (2.2), one obtains $\lambda = Q\alpha$. The term λ_k is the *nominal arrival rate* for class k; to avoid degeneracies, we assume that $\lambda_k > 0$ for all k. Employing m and λ, one defines the *traffic intensity* ρ_j for the jth server as

$$\rho_j = \sum_{k \in \mathcal{C}(j)} m_k \lambda_k, \tag{2.3}$$

with ρ being the corresponding vector. The condition $\rho_j < 1,\ j = 1,\dots,J$, is required for each station, when nonempty, to serve customers, in the long run, more rapidly than they enter the station. When this holds, the network is *strictly subcritical*. When $\rho_j = 1$ for each j, the network is referred to as being *critical* or *balanced*.

Associated with each queueing network is a *discipline*, which specifies the order in which customers receive service. We consider here only *head-of-the-line* (HL) disciplines, where only the first customer in each class may receive service at a given time. For multiclass networks, the proportion of service to be devoted to each class needs to be specified. Examples of disciplines which we will discuss are first-in first-out (FIFO), where the first customer at a station receives all of the service irrespective of its class; head-of-the-line proportional processor sharing (HLPPS), where the amount of service allocated to the first customer in each class is proportional to the number of customers in that class, and static priority disciplines, where classes are assigned a strict ranking, and customers of higher ranked classes are always served first. In the setting of re-entrant lines, examples of static priority disciplines are first-buffer-first-served (FBFS) and last-buffer-first-served (LBFS), where customers at the earlier, respectively latter, classes have priority. When the queueing network is single class, and the service and interarrival times are exponentially distributed, it is referred to as a Jackson network. When the restriction on the service and interarrival times is removed, it is called a generalized Jackson network.

Once a discipline has been given, the triple $(E(\cdot), V(\cdot), \Phi(\cdot))$ and the initial data uniquely specify the evolution of a queueing network along each realization. This defines an underlying Markov process. When this process is positive recurrent, the queueing network is said to be *stable*. Depending on the discipline, the description of the state space can be a bit of a notational burden. We avoid such details here.

3 Stability and Fluid Models

A necessary condition for a queueing network to be stable is that it be strictly subcritical. For a long while, it was generally believed that the condition is also sufficient. This is now known to be false ([Br94], [LuKu91], [RySt92] and [Se94]). It is possible for the flow of customers through a network to synchronize so that, at a given time, customers are clustered at specific parts of the network. This permits individual stations to be periodically "starved" for work, which reduces their long-term efficiency. At the end of each additional cycle, the number of customers in the network will then be, on the average, a multiple of the number for the previous cycle, which produces geometric growth (as measured in cycles).

For many disciplines, however, a queueing network is stable whenever it is strictly subcritical. Fluid models are the main tool for showing this. They allow one, in essence, to replace a queueing network with its continuous deterministic analog of mass flowing through the system. It is typically a considerably easier problem to show stability in this deterministic setting. Under mild conditions on the service and interarrival distributions, the stability of the original queueing network will then follow.

The basic idea is to describe the evolution of a queueing network by a set of equations. One then analyzes the solutions of the corresponding set of deterministic equations, where random quantities have been replaced by their means. One needs to show that the "queue length" vector for such solutions is 0 after a fixed time. It then follows that the queueing network is stable.

In order to describe the evolution of a queueing network, one employs random vectors such as $A(t)$, $D(t)$, $W(t)$, $Y(t)$ and $Z(t)$. The vector $A(t)$ denotes the number of arrivals by time t, $D(t)$ denotes the number of departures, and $Z(t)$ is the number of customers at time t. These three quantities are all class vectors, with components corresponding to the individual classes. The vectors $W(t)$ and $Y(t)$ are both station vectors, with $W(t)$ being the immediate workload (the future time required to serve customers currently at each station), and $Y(t)$ is the cumulative idletime. Typically, the choice of exactly which quantities one employs depends on the particular setting. We will denote the corresponding n-tuple by $\mathfrak{X}(t)$; in the above setting,

$$\mathfrak{X}(t) = (A(t), D(t), W(t), Y(t), Z(t)). \tag{3.1}$$

One connects these quantities together by *queueing network equations*, which include

$$A(t) = E(t) + \sum_k \Phi^k(D_k(t)), \tag{3.2}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{3.3}$$

$$W(t) = CV(A(t) + Z(0)) - et + Y(t) \tag{3.4}$$

$$\int_0^\infty 1_{(0,\infty)}(W_j(s))dY_j(s) = 0, \quad j = 1, \ldots, J, \tag{3.5}$$

for $t \geq 0$. Here, e is the J-vector of all 1's, and C is the $J \times K$ matrix with $C_{jk} = 1$ for $k \in \mathcal{C}(j)$, and $C_{jk} = 0$ otherwise. An additional equation or two is required for the discipline of the network. For instance, for the FIFO discipline, one employs

$$D_k(t + W_j(t)) = Z_k(0) + A_k(t), \quad k = 1, \ldots, K, \tag{3.6}$$

for $t \geq 0$.

For our purposes, the exact nature of the equations (3.2)–(3.6) is not too important. One should think of there as being enough equations to determine the evolution of the queueing network. These equations are used in conjunction with their deterministic analogs, known as *fluid model equations*, which are obtained by replacing $(E(\cdot), V(\cdot), \Phi(\cdot))$ by (α, M, P). The analogs of (3.2)–(3.6) are then given by

$$\bar{A}(t) = \alpha t + P'\bar{D}(t), \tag{3.7}$$

$$\bar{Z}(t) = \bar{Z}(0) + \bar{A}(t) - \bar{D}(t), \tag{3.8}$$

$$\bar{W}(t) = CM(\bar{A}(t) + \bar{Z}(0)) - et + \bar{Y}(t), \tag{3.9}$$

$$\int_0^\infty 1_{(0,\infty)}(\bar{W}_j(s))d\bar{Y}_j(s) = 0, \quad j = 1, \ldots, J, \tag{3.10}$$

$$\bar{D}_k(t + \bar{W}_j(t)) = \bar{Z}_k(0) + \bar{A}_k(t), \quad k = 1, \ldots, K, \tag{3.11}$$

for $t \geq 0$. (To distinguish the solutions of the fluid model equations, we employ overbar notation for the variables in this context.) We also write $\bar{\mathfrak{X}}(t)$ for the analog of (3.1). Such solutions are referred to as fluid model solutions. We restrict our attention to solutions with continuous and nonnegative components, where $\bar{A}(t)$, $\bar{D}(t)$ and $\bar{Y}(t)$ are nondecreasing.

The solutions of the equations (3.2)–(3.6) and (3.7)–(3.11) are connected via the *fluid limits* of $\mathfrak{X}(t)$. These are the limits obtained by applying hydrodynamic scaling to $\mathfrak{X}(t)$, i.e., by scaling the weight of individual customers and time proportionately. (We avoid the technical details here.) Fluid limits are solutions of the fluid model equations; solution of the latter will give information about the original queueing network. The fluid model is said to be *stable* if, for a given $\delta > 0$ and all solutions of the fluid model equations, $\bar{Z}(t) = 0$ for $t \geq \delta|\bar{Z}(0)|$. ($|\cdot|$ denotes the sum of the coordinates.) Since the solutions of a fluid model correspond to a queueing network with the randomness removed, stability of the fluid model says that, in essence, the total number of customers in the queueing network has a net negative drift.

Using elementary properties of Markov processes on general state spaces, it is shown in [Da95] that, under mild assumptions on the service and interarrival times, a queueing network is stable whenever the corresponding fluid model is stable. (Versions of these ideas were first employed in [RySt92].) This enables one

to indirectly study a queueing network by means of the corresponding fluid model equations. In particular, the distributions of the service and interarrival times do not occur in this setting. This enables one, for example, to simply demonstrate the stability of strictly subcritical generalized Jackson networks, whereas a direct argument is quite tedious. The stability of strictly subcritical FIFO networks of Kelly type is another application. (The latter condition means that $m_k = m_\ell$ whenever $s(k) = s(\ell)$.) In general, strictly subcritical FIFO networks which are not of Kelly type need not be stable.

4 Heavy Traffic Limits

Some background

In the introduction, we briefly discussed heavy traffic limits. Here, we go into more detail. The basic setup for HTLs consists of a sequence of queueing networks, with the accompanying n-tuples $\mathfrak{X}^r(t)$ and queueing network equations. One scales the quantities $W^r(t)$ and $Z^r(t)$, setting $\hat{W}^r(t) = W^r(r^2t)/r$ and $\hat{Z}^r(t) = Z^r(r^2t)/r$. The goal is to show that

$$\hat{W}^r(\cdot) \Rightarrow W^*(\cdot) \quad \text{as } r \to \infty, \tag{4.1}$$

where $W^*(\cdot)$ is a semimartingale reflecting Brownian motion (SRBM). The functions $\hat{W}^r(\cdot)$ take values in the space of J-dimensional right continuous functions with left limits, which is equipped with the usual Skorokhod topology, and "$\Rightarrow$" denotes weak convergence.

SRBMs and related concepts are defined in [Wi98b]. Intuitively, the SRBM $W^*(\cdot)$ behaves like a Brownian motion in the interior of the orthant $\mathbb{R}^J_+$; its drift and its covariance matrix are given by appropriate limits of the first two moments of the summands of the triples $(E^r(\cdot), V^r(\cdot), \Phi^r(\cdot))$, and by the discipline of the networks. It is confined to $\mathbb{R}^J_+$ by pushing on the boundary in the directions given by a reflection matrix R (also determined by the above quantities), according to the local time spent there. In order for such a process $W^*(\cdot)$ to exist, R needs to be completely-$\mathcal{S}$.

HTLs have been investigated over the past three decades; a summary of the subject is given in [Wi96, Wi98b]. Implicit in the formulation of (4.1) is the assumption that the states of the corresponding networks are, for large r, essentially given by $\hat{W}^r(t)$ at time t. More detailed information about the system, such as $\hat{Z}^r(t)$, should not be necessary to study the evolution of the limit $W^*(t)$. This type of behavior is known as state space collapse. (The term was used in [Re84a]; related ideas go back to [Wh71].) For our purposes, the relevant variant is *multiplicative state space collapse*, that is

$$\frac{\|\hat{Z}^r(\cdot) - \Delta\hat{W}^r(\cdot)\|_T}{\max(\|\hat{W}^r(\cdot)\|_T, 1)} \to 0 \quad \text{in probability} \tag{4.2}$$

as $r \to \infty$. Here, Δ is an appropriate linear map from $\mathbb{R}^J$ to $\mathbb{R}^K$, which depends on the service discipline; $\|\cdot\|_T$ is the uniform norm over $[0, T]$.

HTLs as in (4.1) need not exist, even for standard disciplines such as FIFO. It was shown in [DaNg94, DaWa93, Wh93] that this is the case for certain sequences of FIFO networks; the problem is related to the limiting reflection matrix R not being completely-$\mathcal{S}$. Another potential problem is the lack of MSSC. These problems need to be faced when dealing with multiclass networks with feedback. (When a network is single class, these problems do not arise, and HTLs exist ([Re84b])). This is also the case when the network is *feedforward*, that is, an ordering among the stations is possible so that customers at lower ranked stations always go to higher numbered stations.) The general theory for multiclass networks is presently incomplete. Below, we summarize some recent work on the subject which uses MSSC and the fluid model equations introduced in Section 3.

Reduction to fluid model equations

In [Wi98a, Br98], HTLs are demonstrated for certain families of multiclass networks. The reasoning employed there can be broken into three "modules", which are essentially independent. The first module, which is worked out in [Wi98a], uses MSSC and the completely-$\mathcal{S}$ condition to derive HTLs. Solutions of the balanced fluid model equations corresponding to the limiting triple (α, M, P), obtained from (α^r, M^r, P^r), are employed in the second module. It is shown in [Br98], that MSSC holds whenever such solutions have "nice" asymptotic behavior. The third module consists of deriving the desired asymptotics for these solutions, and verifying that R is completely-$\mathcal{S}$. Both of these conditions, in the last step, are not trivial in general. They are, though, substantial reductions from MSSC. In this subsection, we discuss the appropriate framework for the second module. We also mention some specific disciplines where the conditions in the third module can be verified.

In order to state our results for MSSC, we need to overcome some technical difficulties. The specific discipline must be known in order to be able to write down all of the relevant queueing network or fluid model equations, such as (3.6). If one wishes to state results on MSSC at the general level of HL processes, it is more convenient to instead work with *cluster points*. These are, in the setting of MSSC, the analog of the fluid limits, which were mentioned briefly in Section 3. Rather than complicate matters, we restrict ourselves here to several more concrete families where we can work directly with the corresponding fluid model equations. Also, as in [Wi98b], we assume that $Z^r(0) = 0$ for the sequences of queueing networks under consideration, in order to simplify formulation of the results.

Associated with a sequence of queueing networks are the triples $(E^r(\cdot), V^r(\cdot), \Phi^r(\cdot))$. We assume here that the corresponding means (α^r, M^r, P^r) satisfy

$$\alpha^r \to \alpha, \quad M^r \to M, \quad P^r \to P \quad \text{as } r \to \infty, \tag{4.3}$$

and that the limit (α, M, P) is balanced. One also needs a uniformity condition on the second moments of the service and interarrival distributions for the sequence. The latter conditions can be ensured, for example, by not allowing $E^r(\cdot)$ or $\Phi^r(\cdot)$ to vary with r, and only allowing the components of $V^r(\cdot)$ to vary by scalar multiples, as is done in [Wi98b]. In order to obtain HTLs from MSSC, as in [Wi98a, Wi98b], one will need to strengthen (4.3) so that $r(\rho^r - e) \to \gamma$ as $r \to \infty$, for some γ,

also holds, although this is not needed for MSSC itself. ($R\gamma$ will be the drift of the HTL.)

We first consider a sequence of queueing networks, with a fixed static priority discipline. As mentioned above, we assume that $Z^r(0) = 0$ for all r. We also assume that (4.3) holds, that (α, M, P) is balanced, and that the second moment conditions referred to above hold. Let $\bar{Z}(t)$ denote the queue length for solutions of the corresponding fluid model equations for the specific discipline. We further assume that for all solutions with $|\bar{Z}(0)| \le 1$,

$$|\bar{Z}(t) - \bar{Z}(\infty)| \le H(t) \tag{4.4}$$

holds for a fixed function $H(t)$, with $H(t) \to 0$ as $t \to \infty$, and for appropriate $\bar{Z}(\infty)$ (depending on $\bar{Z}(0)$) of the form

$$\bar{Z}(\infty) = \Delta\bar{W} \quad \text{for some } \bar{W} \in \mathbb{R}^J. \tag{4.5}$$

It is shown in [Br98], that MSSC follows under these conditions. In [BrDa98], (4.4)–(4.5) are verified for several disciplines, such as FBFS and LBFS. Since one can also show that the R matrix is completely-$\mathcal{S}$ in both cases, the corresponding HTLs follow. (HTLs for FBFS networks are also shown in [ChZh96].)

One can also obtain HTLs for sequences of FIFO networks of Kelly type and HLPPS networks by investigating the corresponding fluid models. The basic procedure is the same as above. In each case, one can, in fact, demonstrate (4.4) with $H(t) = B_1 e^{-B_2 t}$, for appropriate B_1 and $B_2 > 0$. MSSC therefore follows. Since the R matrix will always be completely-$\mathcal{S}$ in both cases, (4.1) holds for appropriate $W^*(t)$. The arguments for showing (4.4) for the two models are related. One obtains an entropy function $\mathcal{H}(t)$ which converges exponentially fast to 0; the states with entropy 0 will satisfy (4.5). The function for FIFO fluid models of Kelly type is

$$\mathcal{H}(t) = \sum_k \int_t^{t+\bar{W}_j(t)} h_k(\bar{D}_k'(r))dr. \tag{4.6}$$

Its asymptotic behavior is analyzed in [Br96] by employing the equations (3.7)–(3.11).

So far, we have not identified the linear map Δ, which "lifts" $\mathbb{R}^J$ to $\mathbb{R}^K$. For the above disciplines, this is easy to do, since ΔW, for $W \in \mathbb{R}^J_+$, will be among the states that remain invariant under the evolution of the corresponding fluid model. Clearly, for static priority disciplines, $(\Delta W)_k = 0$ at all coordinates except where k is the lowest ranked class at its station $j = s(k)$, in which case $(\Delta W)_k = W_j/m_k$. For FIFO networks, $(\Delta W)_k = \lambda_k W_j$, where λ is as in (2.2), and for HLPPS networks,

$$(\Delta W)_k = \frac{\lambda_k m_k W_j}{\sum_{\ell \in \mathcal{C}(j)} \lambda_\ell m_\ell^2}. \tag{4.7}$$

One can see why, in principle, MSSC should follow from the limiting behavior of the fluid model solutions, as in (4.4)–(4.5), by comparing the evolution of the

queue length vector $Z(t)$ under hydrodynamic scaling with its behavior under diffusive scaling. (Some poetic license is taken in phrasing the following steps.) Fluid limits, which are solutions of the fluid model equations, arise from hydrodynamic scaling. So, for large t, the components $\tilde{Z}^r_k(t)$ of $\tilde{Z}^r(t) \stackrel{\text{def.}}{=} Z^r(rt)/r$, as $r \to \infty$, will be in the proportions prescribed by Δ. Recalling that $\hat{Z}^r(t) = Z^r(r^2t)/r$, this implies that $\hat{Z}^r(T_r) = \tilde{Z}^r(rT_r)$, as $r \to \infty$, collapses to the subspace given by Δ, if T_r is chosen so that $rT_r \to \infty$ sufficiently slowly as $r \to \infty$. (One needs the growth of rT_r to be slow enough to avoid the contribution of noise from random fluctuations of $Z^r(r^2T_r)$.) One is, moreover, entitled to restart the processes $\tilde{Z}^r(t)$ at times $i = 1, 2, \ldots$, with

$$\tilde{Z}^{r,i}(t) \stackrel{\text{def.}}{=} Z^r(r(t+i))/r. \tag{4.8}$$

Chopping up the interval $[0, r^2T]$, $T > 0$, from the original time scale into rT pieces, it suffices to analyze the fluid limits corresponding to each of these processes in order to demonstrate MSSC. Under the second moment conditions on the service and interarrival distributions that have already been made, the exceptional events where any of these processes is ill behaved, and the desired collapse does not occur, will have small probability for large r. Also, the assumption $\hat{Z}^r(0) = 0$ ensures that $\hat{Z}^r(t)$ remains close to 0 at small times. Therefore, for a typical realization, $\hat{Z}^r(t)$ collapses to the desired subspace for all $t \in [0, T]$. This reasoning (when carefully carried out) will demonstrate MSSC.

References

[Br94] Bramson, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.*, 4, 414–431.

[Br96] Bramson, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems*, 22, 5–45.

[Br98] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, to appear.

[BrDa98] Bramson, M. and Dai, J. (1998). Heavy traffic limits for some queueing networks. In preparation.

[ChZh96] Chen, H. and Zhang, H. (1996). Diffusion approximations for re-entrant lines with a first-buffer-first-served priority discipline. *Queueing Systems*, 23, 177–195.

[Da95] Dai, J. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid models. *Ann. Appl. Probab.*, 5, 49–77.

[DaNg94] Dai, J. and Nguyen, V. (1994). On the convergence of multiclass queueing networks in heavy traffic. *Ann. Appl. Probab.*, 4, 26–42.

[DaWa93] Dai, J. and Wang, Y. (1993). Nonexistence of Brownian models for certain multiclass queueing networks. *Queueing Systems*, 13, 41–46.

[LuKu91] Lu, S.H. and Kumar, P.R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Autom. Control*, 36, 1406–1416.

[Re84a] Reiman, M.I. (1984). Some diffusion approximations with state space collapse. *Proceedings International Seminar on Modeling and Performance Evaluation Methodology*, Lecture Notes in Control and Informational Sciences, F. Baccelli and G. Fayolle (eds.), Springer, New York, 209–240.

[Re84b] Reiman, M.I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9, 441–458.

[RySt92] Rybko, S. and Stolyar, A. (1992). Ergodicity of stochastic processes that describe the functioning of open queueing networks. *Problems Inform. Trans.*, 28, 3–26 (in Russian).

[Se94] Seidman, T.I. (1994). "First come, first served" can be unstable! *IEEE Trans. Automat. Control*, 39, 2166–2171.

[Wh71] Whitt, W. (1971). Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Probab.*, 8, 74–94.

[Wh93] Whitt, W. (1993). Large fluctuations in a deterministic multiclass network of queues. *Management Science*, 39, 1020–1028.

[Wi96] Williams, R.J. (1996). On the approximation of queueing networks in heavy traffic. *Stochastic Networks, Theory and Applications*, Royal Statistical Society Lecture Note Series, F.P. Kelly, S. Zachary, I. Ziedlins (eds.), Clarendon Press, Oxford, 35–56.

[Wi98a] Williams, R.J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems*, to appear.

[Wi98b] Williams, R.J. (1998). Reflecting diffusions and queueing networks. *Proceedings of the International Congress of Mathematicians*, this issue.

Maury Bramson
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

Random and Deterministic Perturbations of Nonlinear Oscillators

Mark I. Freidlin

Abstract. Perturbations of Hamiltonian systems are considered. The long-time behavior of such a perturbed system, even in the case of deterministic perturbations, is governed, in general, by a stochastic process on a graph related to the Hamiltonian. We calculate the characteristics of the process for systems with one degree of freedom and consider some applications and generalizations.

1991 Mathematics Subject Classification: 60H10, 34F05, 35B20, 60J60
Keywords and Phrases: Random perturbations, Hamiltonian systems, PDE's with a small parameter

Consider an oscillator with one degree of freedom:

$$\ddot{q}_t + f(q_t) = 0, \quad q_0 = q, \quad \dot{q}_0 = p. \tag{1}$$

Let $F(q) = \int_0^q f(u)\,du$ be the potential and $H(p,q) = \frac{p^2}{2} + F(q)$ be the Hamilton function of the oscillator. One can rewrite (1) as the system:

$$\dot{p}_t = -f(q_t) = -\frac{\partial H}{\partial q}, \quad \dot{q}_t = p_t = \frac{\partial H}{\partial p}. \tag{2}$$

We assume that the potential $F(q)$ is a smooth generic function: $f(q) = F'(q)$ is assumed to be continuously differentiable, $f(q)$ has a finite number of zeros, $|f(q)| + f'(q)| \neq 0$, and the values of $F'(q)$ at different critical points are different. Let also $\lim_{|q|\to\infty} F(q) = \infty$. A typical example of $H(p,q)$ and of the phase picture is shown in Fig. 1.

Let $C(z) = \{x = (p,q) \in \mathbf{R}^2 : H(x) = z\}$ be the z-level set of $H(x)$. Since $H(x)$ is generic, $C(z)$ consists of a finite number $n = n(z)$ of connected components. Let Γ be the graph homeomorphic to the set of all connected components of the level sets of $H(x)$ provided with the natural topology (see Fig. 1b). The vertices $O_1, \ldots, O_m$ of Γ correspond to the critical points of $H(x)$. Let $I_1, \ldots, I_n$ be the edges of the graph. A vertex $O_k \in \Gamma$ is called exterior if O_k belongs just to one edge. The other vertices are called interior (vertices O_2 and O_4 in Fig. 1b). Each interior vertex belongs to 3 edges. We write $I_i \sim O_k$ if O_k is one of the ends of I_i. The value of the Hamiltonian H and the number of an edge k define a point of Γ, so that the pairs (H,k) form a global coordinate system on Γ. Define

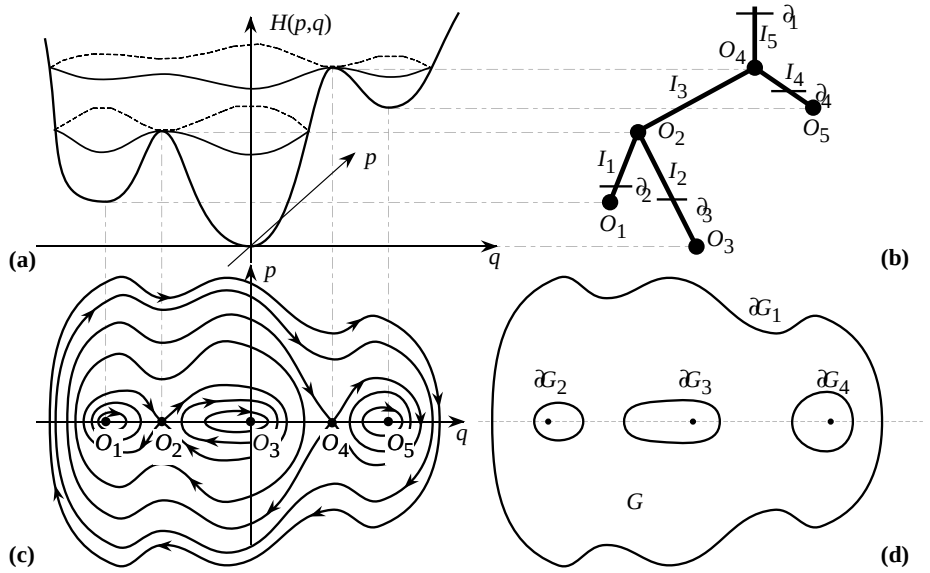


Figure 1.

a metric $\rho(\,\cdot\,,\,\cdot\,)$ on Γ: If $y_1 = (H_1, k)$ and $y_2 = (H_2, k)$ are points of the same edge $I_k \subset \Gamma$, we put $\rho(y_1, y_2) = |H_2 - H_1|$. The distance between any $y_1,\, y_2 \in \Gamma$ is defined as the length of the path connecting y_1 and y_2. Such a path is unique since Γ is a tree.

Consider the map $Y : \mathbf{R}^2 \to \Gamma$, $Y(x) = \big(H(x), k(x)\big) \in \Gamma$, where $k(x)$ is the number of the edge $I_{k(x)} \subset \Gamma$ containing the point of Γ corresponding to the component $C\big(H(x)\big)$ containing $x \in \mathbf{R}^2$. Let $C_k(z) = Y^{-1}(z, k)$, $(z, k) \in \Gamma$. Note that $H(x)$, as well as $k(x)$, are first integrals of system (2): $H(p_t, q_t) \overset{t}{=} H(p_0, q_0)$, $k(p_t, q_t) \overset{t}{=} k(p_0, q_0)$. If $H(p, q)$ has more than one minimum, then these first integrals are independent.

The Lebesgue measure in $\mathbf{R}^2$ is invariant with respect to the flow $X_t \equiv (p_t, q_t)$. If z is not a critical value of $H(x)$, then $C_k(z)$ consists of one periodic trajectory. The normalized invariant density of the flow X_t on $C_k(z)$ with respect to the length element $d\ell$ on $C_k(z)$ is

$$\Big(T_k(z)\big|\nabla H(x)\big|\Big)^{-1}, \quad x \in C_k(z),$$

where

$$T_k(z) = \oint\limits_{C_k(z)} \frac{d\ell}{|\nabla H(x)|}$$

is the period of the revolution along $C_k(z)$.

Consider now the perturbed system:

$$\ddot{q}_t^\varepsilon + f(q_t^\varepsilon) = \varepsilon\beta(\dot{q}_t^\varepsilon, q_t^\varepsilon) + \sqrt{\varepsilon}\sigma(\dot{q}_t^\varepsilon, q_t^\varepsilon) \circ \dot{W}_t. \tag{3}$$

Here W_t is the Wiener process in $\mathbf{R}^1$, functions $\beta(p,q)$ and $\sigma(p,q)$ are supposed to be bounded and continuously differentiable, $0 < \sigma(p,q)$, $0 < \varepsilon \ll 1$. The stochasitc term $\sigma(\dot{q_t^\varepsilon}, q_t^\varepsilon) \circ \dot{W}_t$ in (3) is understood in the Stratanovich sense. The deterministic part of the perturbation $\varepsilon\beta(\dot{q}, q)$ is a kind of friction. A typical and interesting example is $\beta = -\dot{q}$.

Equations (3) can be written as the system

$$\begin{aligned} \dot{p_t^\varepsilon} &= -f(q_t^\varepsilon) + \varepsilon\beta(p_t^\varepsilon, q_t^\varepsilon) + \sqrt{\varepsilon}\sigma(p_t^\varepsilon, q_t^\varepsilon) \circ \dot{W}_t; \\ \dot{q_t^\varepsilon} &= p_t^\varepsilon. \end{aligned} \tag{4}$$

The pair $(p_t^\varepsilon, q_t^\varepsilon) = X_t^\varepsilon$ forms a Markov diffusion process in $\mathbf{R}^2$. The generator A of X_t^ε for a smooth function $g(p,q)$, $(p,q) \in \mathbf{R}^2$, coincides with the differential operator

$$L^\varepsilon g(p,q) = p\frac{\partial g}{\partial q} - f(q)\frac{\partial g}{\partial p} + \varepsilon\beta(p,q)\frac{\partial g}{\partial p} + \frac{\varepsilon}{2}\frac{\partial}{\partial p}\left(\sigma^2(p,q)\frac{\partial g}{\partial p}\right).$$

We are interested in the behavior of the process X_t^ε for $0 < \varepsilon \ll 1$. On any finite time interval $[0,T]$, one can write down an expansion of X_t^ε in the powers of $\sqrt{\varepsilon}$, if $f(q)$, $\beta(p,q)$ and $\sigma(p,q)$ are smooth enough. But, actually, the long time behavior of X_t^ε is, as a rule, of interest. The finite time interval expansion does not help on time intervals of order ε^{-1}, $\varepsilon \downarrow 0$, when the perturbations become essential.

A typical example of a problem of interest is the exit problem. Let G be a bounded domain in $\mathbf{R}^2$. The most interesting case is when G is bounded by trajectories of the non-perturbed system. In Fig. 1, the boundary of the domain G consists of four components ∂G_1, ∂G_2, ∂G_3, ∂G_4. Each of them is a periodic trajectory of system (2). Let $\gamma = Y(G) \subset \Gamma$ and $\partial_i = Y(\partial G_i)$, $i = 1,2,3,4$. Let $\tau^\varepsilon = \min\{t : X_t^\varepsilon \notin G\}$ be the exit time from G. It is not difficult to check that $\tau^\varepsilon \sim \varepsilon^{-1}$ as $\varepsilon \downarrow 0$. Let $\psi(x)$, $x \in \partial G$, be continuous. Calculation of $E_x\tau^\varepsilon = u^\varepsilon(x)$, $P_x\{\tau^\varepsilon < t\} = u^\varepsilon(t,x)$, $E_x\psi(X_\tau^\varepsilon) = v^\varepsilon(x)$, where E_x and P_x mean the expectation and the probability for solutions of (4) starting at $x = (p,q) \in \mathbf{R}^2$, are of interest. Of course, since $X_t^\varepsilon = (p_t^\varepsilon, q_t^\varepsilon)$ is a diffusion process governed by the operator L^ε, one can write down a boundary problem for each of those functions $u^\varepsilon(x)$, $u^\varepsilon(t,x)$, $v^\varepsilon(x)$. Say, $u^\varepsilon(x)$ is the solution of the problem:

$$\begin{aligned} L^\varepsilon u^\varepsilon(p,q) &= p\frac{\partial u^\varepsilon}{\partial q} - f(q)\frac{\partial u^\varepsilon}{\partial p} + \varepsilon\beta(p,q)\frac{\partial u^\varepsilon}{\partial p} + \frac{\varepsilon}{2}\frac{\partial}{\partial p}\left(\sigma^2(p,q)\frac{\partial u^\varepsilon}{\partial p}\right) \\ &= -1, \quad (p,q) \in G, u^\varepsilon(p,q)\big|_{\partial G} = 0. \end{aligned} \tag{5}$$

But even numerical solution of problem (5), because of degeneration of the equation and smallness of $\varepsilon > 0$, is not simple, and the asymptotic approach is the most appropriate.

Since $\tau^\varepsilon \sim \varepsilon^{-1}$, to deal with finite time intervals as $\varepsilon \downarrow 0$, we rescale the time. Put $\tilde{X}_t^\varepsilon = X_{t/\varepsilon}^\varepsilon$, $\tilde{\tau}^\varepsilon = \varepsilon\tau^\varepsilon$. Then $\tilde{X}_t^\varepsilon = (\tilde{p}_t^\varepsilon, \tilde{q}_t^\varepsilon)$ is the solution of the system

$$\begin{aligned} \dot{\tilde{p}_t^\varepsilon} &= -\frac{1}{\varepsilon}f(\tilde{q}_t^\varepsilon) + \beta(\tilde{p}_t^\varepsilon, \tilde{q}_t^\varepsilon) + \sigma(\tilde{p}_t^\varepsilon, \tilde{q}_t^\varepsilon) \circ \dot{\tilde{W}}_t; \\ \dot{\tilde{q}_t^\varepsilon} &= \frac{1}{\varepsilon}\tilde{p}_t^\varepsilon. \end{aligned} \tag{6}$$

Here $\tilde{W}_t^\varepsilon$ is a new Wiener process. We will omit the tilde in the Wiener process.

One can single out the fast and the slow components in the process $\tilde{X}_t^\varepsilon$. The fast component is, basically, the motion along the non-perturbed trajectory. In a vicinity of a periodic trajectory $C_k(z)$, the fast motion, asymptotically as $\varepsilon \downarrow 0$, can be characterized by the invariant density $\Big(T_k(z)\big|\nabla H(x)\big|\Big)^{-1}$, $x \in C_k(z)$.

Taking into account that $H(x)$ and $k(x)$ are first integrals of the non-perturbed system, the slow motion can be described by the projection $Y(\tilde{X}_t^\varepsilon) = \big(H(\tilde{X}_t^\varepsilon), k(\tilde{X}_t^\varepsilon)\big)$ of $\tilde{X}_t^\varepsilon$ on Γ. If we are interested in the asymptotics of $u^\varepsilon(x) = \varepsilon^{-1} E_x \tilde{\tau}^\varepsilon$ as $\varepsilon \downarrow 0$, then it is sufficient to study just the slow component $Y_t^\varepsilon = Y(X_t^\varepsilon)$ as $\varepsilon \downarrow 0$ since $\tilde{\tau}^\varepsilon = \min\{t : Y_t^\varepsilon \notin \gamma\}$, $\gamma = Y(G)$. Therefore, the slow component is, in a sense, the most important for long-time behavior of the process X_t^ε, $0 < \varepsilon \ll 1$. Note, however, that if we are interested in $v^\varepsilon(x) = E_x \psi(X_{\tau^\varepsilon}^\varepsilon)$ and $\psi(x)$ is not a constant on one of the components of ∂G, then the fast component is involved in the behavior of $v^\varepsilon(x)$ as $\varepsilon \downarrow 0$ (compare with [F-W 2] Theorem 2.3 and the remark afterward).

Thus, the problem of long-time behavior of X_t^ε as $\varepsilon \downarrow 0$, to some extent, can be reduced to the asymptotic behavior of the process $Y_t^\varepsilon = Y(\tilde{X}_t^\varepsilon)$ on the graph Γ as $\varepsilon \downarrow 0$.

We prove (see [F-Web 1]) that the process Y_t^ε, $0 \le t \le T$, for any $T < \infty$ converge weakly as $\varepsilon \downarrow 0$ in the space of continuous functions $[0, T] \to \Gamma$ to a continuous Markov process Y_t on Γ. A complete description of all continuous Markov processes on a graph is given in [F-W 1,2]. A continuous Markov process Y_t on $\Gamma = \{I_1, \dots, I_n; O_1, \dots, O_m\}$ is determined by a family of second order elliptic (maybe, generalized) operators $L_1, \dots, L_n$, governing the process inside the edges, and by gluing conditions at the vertices.

To calculate the operator L_k governing the limiting process Y_t inside $I_k \subset \Gamma$, apply the Ito formula to $H(\tilde{X}_t^\varepsilon) \equiv H(\tilde{p}_t^\varepsilon, \tilde{q}_t^\varepsilon)$:

$$H(\tilde{X}_t^\varepsilon) - H(x) = \int_0^t \frac{\partial H}{\partial p}(\tilde{X}_s^\varepsilon)\sigma(\tilde{X}_s^\varepsilon)\, d\, W_s + \frac{1}{2}\int_0^t \sigma^2(\tilde{X}_s^\varepsilon)\frac{\partial^2 H}{\partial p^2}(\tilde{X}_s^\varepsilon)\, ds + \frac{1}{2}\int_0^t \frac{\partial H}{\partial p}\frac{\partial \sigma^2}{\partial p}(\tilde{X}_s^\varepsilon)\, ds + \int_0^t \frac{\partial H}{\partial p}\beta(\tilde{X}_s^\varepsilon)\, ds. \tag{7}$$

The stochastic integral in (7) is taken in Ito sense. Before $H(\tilde{X}_s^\varepsilon)$ changes a little, the trajectory $\tilde{X}_s^\varepsilon$ makes (for $0 < \varepsilon \ll 1$) many rotations along the periodic trajectory of the non-perturbed system. Therefore, the second, the third, and the fourth terms in the right-hand side of (7) are equivalent respectively to

$$\frac{t}{2T(H(x))}\oint_{C_k(H(x))} \frac{\sigma^2(x)H''_{pp}(x)\, d\ell}{|\nabla H(x)|}, \quad \frac{t}{2T(H(x))}\oint_{C_k(H(x))} \frac{\sigma^2(x)'_p H'_p(x)\, d\ell}{|\nabla H(x)|},$$
$$\frac{t}{2T(H(x))}\oint_{C_k(H(x))} \frac{\beta(x)H'_p(x)\, d\ell}{|\nabla H(x)|}, \quad 0 < \varepsilon \ll t \ll 1.$$

To average the stochastic integral in (7), note that because of the selfsimilarity

properties of the Wiener process, this integral is equal to

$$\overline{W}\left(\int_0^t \sigma^2(\tilde{X}_s^\varepsilon)\big(H_p'(\tilde{X}_s^\varepsilon)\big)^2\,ds\right),$$

where $\overline{W}_t$ is an appropriate Wiener process. Using this representation, one can check that the stochastic integral is equivalent to

$$\overline{W}\left(\frac{t}{T_k(H(x))}\oint\limits_{C_k(H(x))}\frac{\sigma^2(x)(H_p'(x))^2\,d\ell}{|\nabla H(x)|}\right),\quad 0<\varepsilon\ll t\ll 1.$$

Using the divergence theorem, we have:

$$\oint\limits_{C_k(z)}\frac{\sigma^2(x)(H_p'(x))^2\,d\ell}{|\nabla H(x)|}=\int\limits_{G_k(z)}\big(\sigma^2(x)H_p'(x)\big)_p'\,dx:=A_k(z),$$

where $G_k(z)$ is the domain in $\mathbf{R}^2$ bounded by $C_k(z)$, $z\in\mathbf{R}^1$. It is easy to check that

$$\frac{dA_k(z)}{dz}=\oint\limits_{C_k(z)}\left[\frac{(\sigma^2(x))_p'H_p'(x)+\sigma^2(x)H_{pp}''(x)}{|\nabla H(x)|}\right]d\ell.$$

Combining all these facts, we conclude from (7) that, starting at a point of $I_k\subset\Gamma$, until the first exit from I_k, the limiting process Y_t is governed by the operator

$$L_k=\frac{1}{2T_k(z)}\frac{d}{dz}\left(A_k(z)\frac{d}{dz}\right)+\frac{1}{T_k(z)}B_k(z)\frac{d}{dz},$$

where

$$B_k(z)=\oint\limits_{C_k(z)}\frac{\beta(x)H_p'(x)}{|\nabla H(x)|}\,d\ell=\int\limits_{G_k(z)}\beta_p'(x)\,dx.$$

In particular, if the perturbation is just the white noise ($\sigma(x)\equiv 1$, $\beta(x)\equiv 0$), then the limiting process in I_k is governed by the operator

$$L_k=\frac{1}{2S_k'(z)}\frac{d}{dz}\left(S_k(z)\frac{d}{dz}\right),$$

where $S_k(z)$ is the area of the domain $G_k(z)\subset\mathbf{R}^2$ bounded by $C_k(z)$; $S_k'(z)=T_k(z)$ is the period of rotation along $C_k(z)$.

To calculate the gluing conditions at the vertices, assume for a moment that $\beta(x)\equiv 0$. Then the Lebesgue measure Λ in the plane is invariant for $\tilde{X}_t^\varepsilon$ for any $\varepsilon>0$. Therefore, the projection $\mu(s)=\Lambda(Y^{-1}(s))$, $s\subset\Gamma$, of the Lebesgue measure on Γ, defined by the mapping $Y:\mathbf{R}^2\to\Gamma$, is invariant for the processes $Y_t^\varepsilon=Y(\tilde{X}_t^\varepsilon)$ on Γ for any $\varepsilon>0$. Thus, the measure $\mu(s)$, $s\subset\Gamma$, is invariant for the limiting process Y_t on Γ. It turns out that among the diffusion processes on Γ governed by operators L_k inside the edges $I_k\subset\Gamma$, there exists just one process for

which the invariant measure coincides with $\mu(s)$. This allows one to calculate the gluing conditions in the case $\beta(x) \equiv 0$. One can check that the exterior vertices are inaccessible for the limit process Y_t, and therefore, no additional gluing conditions should be imposed there. The interior vertices are accessible in a finite time inspite of the degeneration of the diffusion coefficients at the vertices.

To describe the gluing conditions at an interior vertex O_k, note that $Y^{-1}(O_k)$ is a ∞-shaped curve γ shown in Fig. 2. The curve γ consists of the trajectories γ_1, γ_2, and of the equilibrium point O_k of the non-perturbed system. Let G_1 and G_2 be the domains bounded by γ_1 and γ_2, respectively. Let $I_{k_0} \subset \Gamma$ be the edge corresponding to the trajectories surrounding γ (like the trajectory ϕ_0 in Fig. 2);

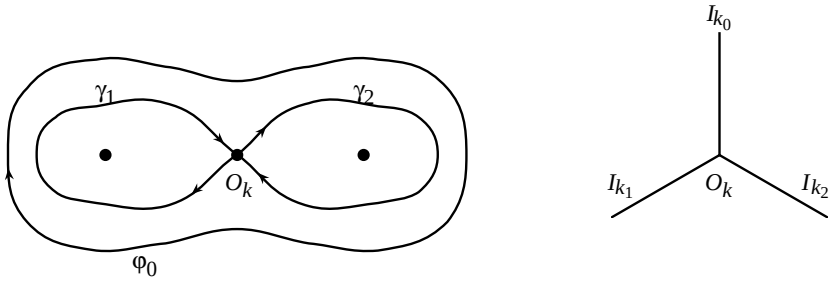


Figure 2.

$I_{k_1} \subset \Gamma$ corresponds to periodic trajectories inside γ_1 which are close to γ_1, and $I_{k_2} \subset \Gamma$ corresponds to trajectories inside γ_2 close to γ_2; I_{k_0}, I_{k_1}, $I_{k_2} \sim O_k$. Put

$$\beta_{ki} = \int\limits_{G_i} \frac{\partial}{\partial p}\left(\sigma^2(p,q)\frac{\partial H(p,q)}{\partial p}\right) dp\, dq, \quad i = 1,2, \quad \beta_{k0} = \beta_{k1} + \beta_{k2}.$$

Then a bounded and continuous on Γ function $u(y)$, $y \in \Gamma$, which is smooth inside the edges, belongs to the domain of definition of the generator A of the limiting process Y_t on Γ if the function $L_k u(z,k)$, $(z,k) \in \Gamma$, is continuous on Γ, and at any interior vertex $O_k \in \Gamma$

$$\beta_{k1} D_1 u(O_k) + \beta_{k2} D_2 u(O_k) = \beta_{k0} D_0 u(O_k),$$

where D_i is the operator of differentiation in z along I_{k_i}, $i = 0,1,2$. The operators L_k together with the gluing conditions at the vertices define the limiting process Y_t on Γ in a unique way.

Now, if $\beta(p,q) \not\equiv 0$ in the perturbation term, one can check, using the Cameron-Martin-Girsanov formula, that the gluing conditions are the same as for $\beta(p,q) \equiv 0$.

To complete the proof, one should also check that the family of processes $Y_t^\varepsilon = Y(\tilde{X}_t^\varepsilon)$, $0 \le t \le T$, is tight in the weak topology and that the limiting process is a Markov one. The tightness follows, roughly speaking, from the at

most linear growth of the coefficients in (7). The Markov property can be proved using some *a priori* bounds for the operator L^ε (see [F-Web1]).

This result allows one to calculate in an explicit form the main terms as $\varepsilon \downarrow 0$ of many interesting characteristics of the process X_t^ε ([F-Web1]). A slight generalization of these results allows one to consider also perturbations of the nonlinear pendulum defined by the equation $\ddot{q}_t + \sin q_t = 0$, ([F-Web2]).

Suppose now that we have just deterministic perturbations: $\sigma(x) \equiv 0$ in equation (6). Let, for brevity, the Hamiltonian have just one saddle point, so that the phase picture for the non-perturbed system is as in Fig. 3a, and let $b'_p(p,q) < 0$, $(p,q) \in \mathbf{R}^2$. The perturbations lead to the picture in Fig. 3b: the perturbed system

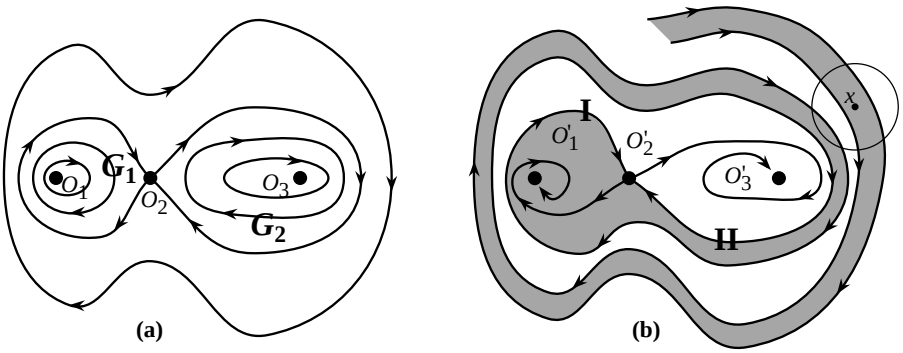


Figure 3.

has a saddle point in a point O'_2 which is close to O_2; the equilibrium points O_1, O_3 will be replaced by asymptotically stable points O'_1, O'_3, which are close to O_1, and O_3, respectively, when $0 < \varepsilon \ll 1$. Two separatrices I and II enter O'_2. They divide the exterior $\mathcal{E}$ of the ∞-shaped curve connected with O_2 in two ribbons. One of these ribbons consists of points attracted to O'_1; another ribbon is attracted to O'_3 (see Fig. 3b). The width of each of these ribbons is of order ε as $\varepsilon \downarrow 0$. When ε becomes smaller, they are moving closer and closer to the ∞-shaped curve. Therefore, any point $x \in \mathcal{E}$ alternatively belongs to a ribbon attracted to either O'_1 or to O'_3 as $\varepsilon \downarrow 0$. This means that the perturbed trajectory X_t^ε starting at $x \in \mathcal{E}$, is attracted alternatively to O'_1 or O'_3 when $\varepsilon \downarrow 0$.

The slow motion of the perturbed system in this case is again the projection on the graph Γ related to $H(x)$: $Y_t^\varepsilon = Y(X^\varepsilon_{t/\varepsilon})$. The averaging procedure shows that the limiting slow motion $\overline{Y}_t$ is a deterministic motion inside each of the edges of the graph Γ:

$$\dot{z}_t = \frac{1}{T_k(z_t)} B_k(z_t), \quad \overline{Y}_t = (z_t, k) \in I_k, \quad k = 1, 2, 3. \tag{8}$$

If we start from a point x with a large enough $H(x)$, and $B_k(z) < 0$ if (z,k) is not a vertex, then the deterministic trajectory hits the vertex O_2 corresponding to the saddle point of $H(x)$ in a finite time. After that, the trajectory of the limiting slow motion goes to one of the two edges attached to O_2 along which H is decreasing.

To which of these two edges the trajectory goes depends on the initial point in a very sensitive way. One can show that the measure of the set of initial points from a neighborhood U of a point x, $U \subset \mathcal{E}$, attracted to O_1 (to O_3) is proportional to $\int\limits_{G_1} \beta_p'(x)\,dx$ ($\int\limits_{G_2} \beta_p'(x)\,dx$) as $\varepsilon \downarrow 0$, where G_1 and G_2 are the left and the right part of the set in $\mathbf{R}^2$ bounded by the ∞-shaped curve. This was briefly mentioned in [A]. The proof is available in [B-F]. If the graph corresponding to $H(x)$ has a more complicated structure and the "friction" $\beta(p,q)$ is allowed to change the sign, the situation can be more complicated: the limiting slow motion can "remember more of its past" (see [B-F]).

There is another way to regularize the problem: Instead of random perturbation of the initial point, one can add a random perturbation to the equation. Let $\sigma(p,q)$ in (6) be replaced by $\sqrt{\kappa}\tilde{\sigma}(p,q)$, where $\kappa > 0$ is a small parameter. Let $\tilde{X}_t^{\varepsilon,\kappa}$ be the solution of (6) with such a replacement. Consider the double limit of the slow component $Y_t^{\varepsilon,\kappa} = Y(\tilde{X}_t^{\varepsilon,\kappa})$, $0 \le t \le T$: first as $\varepsilon \downarrow 0$ for a fixed $\kappa > 0$, and then as $\kappa \downarrow 0$. The first limit gives us the diffusion process Y_t^{κ} on Γ, which was described above. Now we consider the limiting behavior of Y_t^{k}, $0 \le t \le T$, as $\kappa \downarrow 0$. As it is proved in [B-F], this limit (in the sense of weak convergence) exists, independent of the perturbations (of the choice of functions $\sigma(p,q)$) and coincides with the process $\overline{Y}_t$ desribed above: Inside the edges it is a deterministic motion governed by (8), and it branches at each interior vertex O_k to one of the edges attached to O_k, along which H is decreasing, with certain probabilities which are expressed through $H(x)$ and $\beta(x)$ in a way similar to that descried above. The behavior of the limiting slow component after touching an interior vertex O_k is independent now of the past (see [B-F]). The independence of the process $\overline{Y}_t$ of the characteristics of the random perturbations, as well as the fact that the limiting process is the same as occurs if the initial conditions are perturbed, shows that the "randomness" of the limiting slow component is an intrinsic property of the Hamiltonian system and its deterministic perturbations. The random perturbation here is just a way of regularization.

The perturbations in equations (6) are included just in one component. Therefore, the corresponding differential operator $\varepsilon^{-1}L^{\varepsilon}$ is degenerate. This leads to certain additional difficulties in the proof of Markov property for the limiting process. One can consider non-degenerate perturbations and replace the oscillator by an arbitrary Hamiltonian system with one degree of freedom:

$$\dot{\tilde{X}}_t^{\varepsilon} = \frac{1}{\varepsilon}\overline{\nabla} H(\tilde{X}_t^{\varepsilon}) + \beta(\tilde{X}_t^{\varepsilon}) + \sigma(\tilde{X}_t^{\varepsilon}) \circ \dot{W}_t, \quad \tilde{X}_t^{\varepsilon} = x \in \mathbf{R}^2. \tag{9}$$

Here W_t is the Wiener process in $\mathbf{R}^2$, $\beta(x)$ is a smooth bounded vector field in $\mathbf{R}^2$, and $\sigma(x)$ is a 2×2 matrix with smooth bounded entries, $\det \sigma(x) \neq 0$. The Hamiltonian function $H(x)$ is assumed to be smooth, generic, and $\lim_{|x|\to\infty} H(x) = \infty$.

Let $\Gamma = \{I_1, \dots, I_n; O_1, \dots, O_m\}$ be the graph corresponding to $H(x)$ and $Y(x) = \big(H(x), k(x)\big)$ be the corresponding mapping $\mathbf{R}^2 \to \Gamma$. Then one can prove [F-W2,3] that the slow component of the process $\tilde{X}_t^{\varepsilon}$, which is $Y(\tilde{X}_t^{\varepsilon})$, $0 \le t \le T$, converges weakly as $\varepsilon \downarrow 0$ to a diffusion process Y_t on Γ. The process Y_t is governed

inside I_k, $k \in \{1, \dots, n\}$, by the operator

$$L_k = \frac{1}{2T_k(z)} \frac{d}{dz} \left(A_k(z) \frac{d}{dz} \right) + \frac{1}{T_k(z)} B_k(z) \frac{d}{dz}, \quad T_k(z) = \oint\limits_{C_k(z)} \frac{d\ell}{|\nabla H(x)|},$$

$$A_k(z) = \int\limits_{G_k(z)} \operatorname{div} \big(a(x) \nabla H(x) \big) \, dx, \quad a(x) = \sigma(x) \sigma^*(x), \quad B_k(z) = \int\limits_{G_k(z)} \operatorname{div} \beta(z) \, dx. \tag{10}$$

Here $C_k(z) = Y^{-1}(z,k)$, $G_k(z)$ is the domain in $\mathbf{R}^2$ bounded by $C_k(z)$, $(z,k) \in \Gamma \setminus \{O_1, \dots, O_m\}$.

To define the process Y_t for all $t \geq 0$, we should add the gluing conditions at the vertices. The gluing conditions are defined by he domain of definition $D_{\mathfrak{a}}$ of the generator $\mathfrak{A}$ of the process Y_t: a continuous and smooth inside the edges function $f(g)$, $y \in \Gamma$, belongs to $D_{\mathfrak{a}}$, if $L_k f(z,k)$, $y = (z,k) \in \Gamma$, is continuous on Γ and at any interior vertex $O_k \in \Gamma$

$$\sum_{i=1}^{3} \pm \beta_{ki} D_i f(O_k) = 0, \tag{11}$$

where $\beta_{ki} = \lim_{(z,k_i) \to O_k} A_{k_i}(z)$; I_{k_1}, I_{k_2}, $I_{k_3} \sim O_k$; the "+" ("−") sign in front of β_{ki} is taken if H grows (decreases) as the point approaches O_k along I_{k_i}, $i \in \{1,2,3\}$, (see [F-W2,3]).

This result allows one to calculate in a rather explicit form the main term as $\varepsilon \downarrow 0$ of the solution of the following Dirichlet problem:

$$\frac{\varepsilon}{2} \operatorname{div} \big(a(x) \nabla u^\varepsilon(x) \big) + \varepsilon \beta(x) \cdot \nabla u^\varepsilon(x) + \overline{\nabla} H(x) \cdot \nabla u^\varepsilon(x) = 0, x \in G, u^\varepsilon(x) \big|_{\nabla G} = \psi(x).$$

Here $G \subset \mathbf{R}^2$ is as in Fig. 1, $\psi(x)$ is a continuous function on ∂G.

It follows from [F-W2,3] that $\lim u^\varepsilon(x) = v\big(H(x), k(x)\big)$, where $v(z,k)$ is the solution of the Dirichlet problem in $\gamma = Y(G) \subset \Gamma$

$$L_k v(z,k) = 0, (z,k) \in \gamma \setminus \{O_1, \dots, O_m\}, v(\partial_k) = \overline{\psi}_k, \quad k \in \{1,2,3,4\},$$

satisfying the gluing conditions described above. Here $\partial_k = Y(\partial G_k)$, $k = 1,2,3,4$, $\partial\gamma = (\partial_1, \partial_2, \partial_3, \partial_4)$,

$$\psi_k = \left(\oint\limits_{\partial G_k} \frac{a(x) \nabla H(x) \cdot \nabla H(x)}{|\nabla H(x)|} \, d\ell \right)^{-1} \oint\limits_{\partial G_k} \frac{\psi(x) \big(a(x) \nabla H(x) \cdot \nabla H(x) \big)}{|\nabla H(x)|} \, d\ell,$$

$k \in \{1,2,3,4\}$.

The Dirichlet problem in γ can be solved explicitly.

Consider now the case of pure deterministic perturbations: $\sigma(x) \equiv 0$ in (9). Let for brevity $B_k(z)$, defined in (10), be negative if (z,k) is not an exterior vertex. This, in particular, implies that the perturbed system is not Hamiltonian. We can again "regularize" the problem adding small random perturbations to the initial conditions or to the equation and then consider the double limit [B-F].

To consider perturbations of the equation, replace the matrix $\sigma(x)$ in (8) by $\sqrt{\kappa}\sigma(x)$, $\kappa > 0$. Let $\tilde{X}_t^{\varepsilon,\kappa}$ be the solution of equation (8). Consider the projection $Y_t^{\varepsilon,\kappa} = Y(\tilde{X}_t^{\varepsilon,\kappa})$ of $\tilde{X}_t^{\varepsilon,\kappa}$ on Γ. Then, for each $\kappa > 0$, the processes $Y_t^{\varepsilon,\kappa}$, $0 \le t \le T$, converge weakly as $\varepsilon \downarrow 0$ to the process Y_t^{κ} on Γ, which was described above. Let now $\kappa \downarrow 0$. One can check that processes Y_t^{κ}, $0 \le t \le T$, converge weakly to a process $Y_t = (z_t, k_t)$ on Γ as $x \downarrow 0$. Inside any edge $I_k \subset \Gamma$, the process Y_t is deterministic motion governed by equation (8) with $B_k(z)$ defined in (10). If Y_t touches an interior vertex $O_k \in \Gamma$, it leaves O_k without any delay along one of the edges I_{k_1}, $I_{k_2} \sim O_k$, along which H is decreasing, with probabilities P_{k1}, P_{k2};

$$P_{ki} = \frac{|B_{k_i}(O_k)|}{|B_{k_1}(O_k)| + |B_{k_2}(O_k)|}, \quad \left|B_{k_i}(O_k)\right| = \lim_{(z,k_i)\to O_k} \left|B_{k_i}(z)\right|, \quad i = 1, 2,$$

independently of the past [B-F].

A special case of this problem when $a(x)$ is the unit matrix was studied in [W].

If we consider the perturbations of the form

$$\dot{X}_t^{\varepsilon,x} = \overline{\nabla} H(X_t^{\varepsilon,x}) + \varepsilon\beta(X_t^{\varepsilon,x}) + \sqrt{\varepsilon\kappa}\zeta_t,$$

where ζ_t is a stationary process with strong enough mixing properties, and the process ζ_t is not degenerate in a certain sense, then, because of a central-limit-theorem type result, we can expect the same process Y_t as the limit of $Y(X_{t/\varepsilon}^{\varepsilon,\kappa})$ as first $\varepsilon \downarrow 0$ and then $\kappa \downarrow 0$.

These results can be applied to some non-linear problems for second order elliptic and parabolic equations. Consider, for example, reaction-diffusion in a stationary incompressible fluid in $\mathbf{R}^2$:

$$\frac{\partial u^{\varepsilon}(t,x)}{\partial t} = \frac{\varepsilon}{2}\Delta u^{\varepsilon} + \overline{\nabla} H(x)\cdot\nabla u + f(u^{\varepsilon}), \quad t > 0, \quad x \in \mathbf{R}^2, \quad u^{\varepsilon}(0,x) = g(x) \ge 0. \tag{12}$$

Here $H(x)$ is the stream function of a stationary flow. We assume that $H(x)$ is generic and $\lim_{|x|\to\infty} H(x) = \infty$. The initial function is assumed to be continuous. Let for brevity $g(x)$ has a compact support. Let Γ be the graph related to $H(x)$ and $Y(x) : \mathbf{R}^2 \to \Gamma$ be the corresponding mapping. If $f(u) \equiv 0$, it follows from the results formulated in this paper [FW2], that $u^{\varepsilon}(t/\varepsilon, x) \to v(t, Y(x))$, where $v(t,y)$ is the solution of a Cauchy problem on $[0,\infty)\times\Gamma$ with appropriate gluing conditions at the vertices.

But if the reaction term $f(u)$ is included in the equation, one should use a different time scale. Let, for instance, $f(u) = c(u)u$ is of Kolmogorov-Petrovskii-Piskunov type: $c(u) > 0$ for $u < 1$, $c(u) < 0$ for $u > 1$, and $c(0) = \max_{u\ge 0} c(u)$. Then $\lim_{\varepsilon\downarrow 0} u^{\varepsilon}(t/\sqrt{\varepsilon}, x) = w(t, Y(x))$, where $w(t,y)$, $t > 0$, $y = (z,k) \in \Gamma$, is a step function with the values 0 and 1. To describe the set, where $w(t,y)$ is equal to 1, introduce a Riemannian metric ρ on Γ corresponding to the form

$$ds^2 = \frac{T_k(z)}{A_k(z)}dz^2, \quad T_k(z) = \oint\limits_{C_k(z)} \frac{d\ell}{|\nabla H(x)|}, \quad A_k(z) = \int\limits_{G_k(z)} \Delta H\, dx.$$

Note that this form has singularities at the vertices, but those singularities are integrable. Let $\gamma = Y(\operatorname{supp} g) \subset \Gamma$. Then, $w(t,y) = 1$ on the set

$$\left\{ y \in \Gamma : \rho(y,\gamma) < t\sqrt{2c(0)} \right\}, \quad t > 0,$$

and $w(t,y) = 0$ outside of the closure of this set. This is a result of an interplay between the averaging and the large deviations for process $X^\varepsilon_{t/\varepsilon}$, where X^ε_t is the process in $\mathbf{R}^2$ governed by the linear part of the operator in the right-hand side of (12).

Applications of the ideas discussed in this paper to small viscosity asymptotics for the stationary Navier-Stokes equations one can find in [F2].

Applications to an optimal stabilization problem are available in [D-F].

I will briefly consider now some generalizations. First, consider a Hamiltonian system on a two-dimensional torus. A generic Hamiltonian system on a 2-torus has the following structure: it has a finite number of loops such that inside those loops, trajectories behave like in a part of $\mathbf{R}^2$. The exterior $\mathcal{E}$ of the union of the loops is one ergodic class so that the trajectories of the system are dense in $\mathcal{E}$ (see references in [F1]). Therefore, the graph Γ related to this system has a special vertex O_0 which corresponds to the whole set $\mathcal{E}$. Consider now small white noise perturbations of the system. The Lebesgue measure on the torus is invariant for the perturbed process and the projection of this measure on Γ is invariant for the slow component. This implies that the limiting slow component spends at $O_0 \in \Gamma$ a positive time proportional to the relative area of $\mathcal{E}$. Therefore the gluing conditions at O_0 are a little different from the conditions at other vertices or form conditions considered above (see [F-W1], [F1]).

Perturbations of certain Hamiltonian systems on 2-torus may lead also to processes on graphs with loops, but not just trees as in the case of systems in $\mathbf{R}^2$.

Finally, we consider briefly perturbations of Hamiltonian systems with many degrees of freedom:

$$\begin{aligned} &\dot X^\varepsilon_t = \overline{\nabla} H(X^\varepsilon_t) + \sqrt{\varepsilon}\dot W_t + \varepsilon\beta(X^\varepsilon_t), \\ &X^\varepsilon_0 = x \in \mathbf{R}^{2n}, x = (p_1, \dots, p_n; q_1, \dots, q_n). \end{aligned} \tag{13}$$

Here W_t is the $2n$-dimensional Wiener process. $\beta(x)$ is a smooth vector field in $\mathbf{R}^{2n}$, $0 < \varepsilon \ll 1$. If $n > 1$, the non-perturbed system may have additional smooth first integrals: $H_1(x) = H(x)$, $H_2(x)$, $\dots$, $H_\ell(x)$. Let $C(z) = \{x \in \mathbf{R}^{2n} : H_1(x) = z_1, \dots, H_\ell(x) = z_\ell\}$, $z = (z_1, \dots, z_\ell) \in \mathbf{R}^\ell$. If the non-perturbed system X^0_t has a unique "smooth" invariant measure on each $C(z)$, $z \in \mathbf{R}^\ell$, then the slow component can be described by the evolution of the first integrals. In an appropriate time scale, the slow component converges to a diffusion process Y_t, $0 \le t \le T$. The diffusion and drift coefficients of Y_t can be calculated using the standard averaging procedure. We have such an example when considering a system of independent oscillators with one degree of freedom

$$\begin{aligned} &\dot X^\varepsilon_k(t) = \overline{\nabla} H_k(X^\varepsilon_k(t)) + \varepsilon\beta_k(X^\varepsilon_1(t), \dots, X^\varepsilon_n(t)) + \sqrt{\varepsilon}\sigma_k \dot W_k(t), \\ &\quad x_k = (p_k, q_k) \in \mathbf{R}^2, \quad k = 1, \dots, n, \end{aligned} \tag{14}$$

with $H_k(p,q) = a_k p^2 + b_k q^2$ and $a_k,\ b_k > 0,\ k \in \{1,\dots,n\}$, such that the frequencies of the oscillators are incommensurable. Here $W_k(t)$ are independent two-dimensional Wiener processes, σ_k are non-degenerate 2×2 matrices, $\beta_k \in \mathbf{R}^2$. But, in general, if H_k are not quadratic forms, the frequences are changing with the energy and resonances appear. This problem, in the case of deterministic perturbations, was studied by many authors (see [AKN]). The approaches used in the deterministic case allow one to obtain some results on stochastic perturbations as well.

Let $H_k(x)$, $x \in \mathbf{R}^2$, $k = 1,\dots,n$, be generic and $\lim_{|x|\to\infty} H_k(x) = \infty$. Let Γ_k be the graph related to $H_k(x)$ and $Y_k : \mathbf{R}^2 \to \Gamma_k$ be the corresponding mapping. The slow component Y_t^ε of the process $\big(X_1^\varepsilon(t),\dots,X_n^\varepsilon(t)\big) = X_t^\varepsilon$ is defined as the process $Y_t^\varepsilon = \big(Y_1(X_1^\varepsilon(t/\varepsilon)),\dots,Y_n(X_n^\varepsilon(t/\varepsilon))\big)$, on $\Xi = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_n$. Under some mild additional conditions, the processes Y_t^ε, $0 \le t \le T$, converge as $\varepsilon \downarrow 0$ weakly to a process Y_t on Ξ. Inside the n-dimensional pieces of Ξ, where $\sum_{k=1}^n \big|\nabla H_k(x_k)\big| \ne 0$, the process Y_t is described by the averaging procedure. To define the gluing conditions, assume, first, that $\beta_k(x) \equiv 0$, $k = 1,\dots,n$. Then the process X_t^ε is just a collection of n independent processes $X_k^\varepsilon(t)$, each with one degree of freedom. The slow component $Y_k(X_k^\varepsilon(t/\varepsilon))$ of X_k^ε converges, as we already know, to a process $Y_k(t)$ on Γ_k with the gluing conditions described above. Thus, we know what is the limiting slow component for X_t^ε in the case $\beta_k(x) \equiv 0$, $k \in \{1,\dots,n\}$. Using the Cameron-Martin-Girsanov formula, one can check that, if $\beta_k(x) \not\equiv 0$ are bounded and matrices σ_k are non-degenerate, than the gluing conditions will be, in a sense, the same. This allows to give a complete description of the limiting slow component for X_t^ε as a diffusion process on Ξ [F-W4].

Similar to the case of one degree of freedom, this result enable us to show that, under some additional conditions, the long-time behavior of deterministic systems close to Hamiltonian has a stochastic nature. Consider weakly coupled oscillators with one degree of freedom:

$$\begin{aligned} &\dot X_k^\varepsilon(t) = \overline{\nabla} H_k(X_k^\varepsilon(t)) + \varepsilon\beta_k\big(X_1^\varepsilon(t),\dots,X_n^\varepsilon(t)\big),\\ &X_k(0) = x_k \in \mathbf{R}^2, \quad k \in \{1,\dots,n\}, \quad 0 < \varepsilon \ll 1. \end{aligned} \tag{15}$$

The slow motion for this system is the projection of $X^\varepsilon(t) = \big(X_1^\varepsilon(t),\dots,X_n^\varepsilon(t)\big)$ on $\Xi : Y_t^\varepsilon = Y(X_{t/\varepsilon}^\varepsilon)$. As in the one-degree-of-freedom case, the processes Y_t^ε does not converge as $\varepsilon \downarrow 0$. But one can regularize the problem, adding small noise to the equation: Replace σ_k in (14) by $\sqrt{\kappa}\sigma_k$, and let $X_t^{\varepsilon,\kappa}$ be the solution of (14) after this change. The processes $Y_t^{\varepsilon,\kappa} = Y(X_{t/\varepsilon}^{\varepsilon,\kappa})$, $0 \le t \le T$, converge as $\varepsilon \downarrow 0$, for a fixed $\kappa > 0$, to a diffusion process Y_t^κ on Ξ, under some additional conditions. Then one can check that the processes Y_t^κ, $0 \le t \le T$, converge as $\kappa \downarrow 0$ to a process Y_t on Ξ. The process Y_t is deterministic inside the n-dimensional pieces of Ξ and has some stochastic behavior on the edges. The process Y_t is independent of the choice of matrices σ_k, so that it is determined by the intrinsic properties of system (15), but not by the random perturbations.

References:

[A] Arnold, V.I., Small denominators and problems of stability of motion in classical and celestial mechanics, *Russian Math. Surveys*, 18, 6, pp. 86–191, 1963.

[AKN] Arnold, V.I., Kozlov, V.V., and Neishtadt, A.I., Mathematical Aspects of Classical and Celestial Mechanics, in *Dynamical Systems III*, V.I. Arnold editor, Springer 1988.

[BF] Brin, M.I., and Freidlin, M.I., On stochastic behavior of perturbed Hamiltonian systems, *Dynamical Systems and Ergodic Theory*, 1998.

[DF] Dunyak, J. and Freidlin, M., Optimal residence time control of Hamiltonian systems perturbed by white noise, *SIAM J. Control & Opt.*, 36, 1, pp. 233–252, 1998.

[F1] Freidlin, M.I., *Markov Processes and Differential Equations: Asymptotic Problems*, Birkhauser, 1996.

[F2] Freidlin, M.I., Probabilistic approach to the small viscosity asymptotics for Navier-Stokes equations, *Nonlinear Analysis*, 30, 7, pp. 4069–4076, 1997.

[FWeb1] Freidlin, M.I. and Weber, M., Random perturbations of nonlinear oscillators, *The Ann. of Prob.*, 26, 3, pp. 1–43, 1998.

[FWeb2] Freidlin, M.I. and Weber, M., Remark on random perturbations of nonlinear pendulum, submitted to *The Ann. of Appl. Prob.*

[FW1] Freidlin, M.I. and Wentzell, A.D., Diffusion processes on graphs and averaging principle, *The Ann. of Prob.*, 21, 4, pp. 2215–2245, 1993.

[FW2] Freidlin, M.I. and Wentzell, A.D., Random perturbations of Hamiltonian systems, *Mem. of Amer. Math. Soc.*, 109, 523, 1994.

[FW3] Freidlin, M.I. and Wentzell, A.D., On random perturbations of Hamiltonian systems, *Proc. of the Conf. Stoch. Struct. Dyn.*, H. Davoodi and A. Saffar, editors, Puerto Rico, 1995.

[FW4] Freidlin, M.I. and Wentzell, A.D., *Averaging in multi-frequency systems with random perturbations*, in preparation.

[W] Wolansky, G., Limit theorem for a dynamical system in the presence of resonances and homoclinic orbits, *J. Diff. Eq.*, 83, 2, pp. 300–335, 1990.

Mark I. Freidlin
Dept. of Mathematics
Univ. of Maryland
College Park, MD 20742, USA

Bayesian Density Estimation

Jayanta K. Ghosh

Abstract. This is a brief exposition of posterior consistency issues in Bayesian nonparametrics especially in the context of Bayesian Density estimation,

1991 Mathematics Subject Classification: 62A15,62G99.
Keywords and Phrases: Dirichlet mixtures,density estimation

1 Introduction

We describe popular methods of Bayesian density estimation and explore sufficient conditions for the posterior given data to converge to a true underlying distribution P_0 as the data size increases. One of the advantages of Bayesian density estimates is that,unlike classical frequentist methods,choice of the right amount of smoothing is not such a serious problem.

Section 2 provides a general background to infinite dimensional problems of inference such as Bayesian nonparametrics, semiparametrics and density estimation. Bayesian nonparametrics has been around for about twenty five years but the other two areas,specially the last, is of more recent vintage. Section 3 indicates in broad terms why different tools are needed for these three different problems and then Section 4 focuses on our main problem of interest ,namely,positive posterior consistency results for Bayesian density estimation.

2 Background

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with unknown common probability measure P on $(\mathbf{R}, \mathcal{B})$, where $\mathbf{R}$ is the real line and $\mathcal{B}$ the Borel $\sigma-$ field. Typically P lies in some given set of probability measures $\mathcal{P}$. In Bayesian analysis, a statistician puts a probability measure Π on $\mathcal{P}$ equipped with a suitable $\sigma-$ field $\mathcal{B}_{\mathcal{P}}$ and assumes that the unknown P is distributed over $\mathcal{P}$ according to Π and, given P, $X_1, X_2, \ldots, X_n$ are i.i.d. with common distribution $\mathcal{P}$. This completely specifies the joint distribution of the random P and the random Xs. Hence, in principle one can calculate the conditional probability $\Pi(B|X_1, X_2, \ldots, X_n)$ of P lying in some subset B. This is the posterior in distinction with $\Pi(B)$ which is the prior probability of B. Consistency of posterior to be defined below is a sort of partial validation of this method of analysis. We now define posterior consistency at P_0. Suppose unknown to the Bayesian statistician,$X_1, X_2, \ldots, X_n$ are i.i.d. $\sim P_0$,

where P_0 is a given element of $\mathcal{P}$ and not random. Suppose that $\mathcal{P}$ is also equipped with a topology and the topology and $\mathcal{B}_{\mathcal{P}}$ are compatible in the sense that the neighborhoods B of P_0 are $\mathcal{B}_{\mathcal{P}}$ measurable.

DEFINITION: $\Pi(.\,|X_1, X_2, \ldots, X_n)$ is consistent at P_0 if for all neighborhoods B of P_0, as $n \to \infty$,

$$\Pi(B|X_1, X_2, \ldots, X_n) \to 1 \text{ a.s } P_0$$

This property depends on both Π and P_0. It would be desirable to have this property at various P_0's that seem plausible to the Bayesian who is using this posterior.

An old result of Doob shows that such a property holds for all but a $\pi-$null set of P_0's. Unfortunately, this result is too weak to settle whether consistency holds for a particular P_0. It is well known that this property holds for a wide class of priors and all P_0's if $\mathcal{P}$ is finite dimensional,e.g., when $\mathcal{P}$ is the set of all normal distributions $N(\mu, \sigma^2)$ with mean μ and variance σ^2, $-\infty < \mu < \infty, \sigma^2 > 0$. In contrast the answer is usually no when $\mathcal{P}$ is infinite dimensional as in density estimation.

There are three broad classes of infinite dimensional problems —(fully) nonparametric inference like making inference about an unknown distribution function, a semiparametric problem like estimating the point of symmetry of an unknown symmetrical distribution function, and density estimation. The set $\mathcal{P}$ is different for these three cases. In the first case,which is classical, $\mathcal{P}$ is the class of all probability measures on $(\mathbf{R}, \mathcal{B})$. In the third case and, in fact also in the second, we work instead with the set of probability measures P on $(\mathbf{R}, \mathcal{B})$ which have a density f with respect to the Lebesgue measure. In the first two problems the set $\mathcal{P}$ is equipped with the weak topology and the natural tools are the use of tail free priors or a theorem of Schwartz(1965). In the third case the natural topology is that induced by the L_1 or the Hellinger metric. The natural tool is a new theorem that makes use of the notion of metric entropy or packing numbers for the space of densities in addition to one of Schwartz's conditions.

3 NOTATIONS AND OTHER TECHNICALITIES

3.1 NONPARAMETRICS

We start with the nonparametric problem. Let $\mathcal{P}$ be the class of all probability measures on $(\mathbf{R}, \mathcal{B})$; $\mathcal{P}$ be equipped with the weak topology and $\mathcal{B}_{\mathcal{P}}$ the corresponding Borel $\sigma-$ field. Equivalently, $\mathcal{B}_{\mathcal{P}}$ is the smallest $\sigma-$ field which makes the evaluation maps $P \mapsto P(A)$ measurable for each A in $\mathcal{B}$.

The most popular prior on $(\mathcal{P}, \mathcal{B}_{\mathcal{P}})$ is the Dirichlet process due to Ferguson(1973,1974). It is specified by its finite dimensional distributions as follows. Let α be a finite non zero measure on $(\mathbf{R}, \mathcal{B})$. Let $A_1, A_2, \ldots, A_k$ form a measurable partition. Then $P(A_1), P(A_2), \ldots, P(A_k)$ have a finite dimensional Dirichlet distribution with parameters $\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k)$. If $\alpha(A_i) > 0, i = 1, 2, \ldots, k$ then this distribution has a density with respect $(k-1)$ dimensional Lebesgue mea-

sure that has the form

$$\frac{\Gamma(\mathbf{R})}{\prod_1^k \Gamma(\alpha(A_i))} \prod_1^k p_i^{\alpha(A_i)-1}, \qquad 0 < p_i, \sum_1^k p_i = 1$$

If $k = 2$, one gets the beta distribution. Integrating out p_1 one gets

$$E(P(A_i)) = \alpha(A_i)/\alpha(\mathbf{R}) = \bar{\alpha}(\mathbf{A_i}). \tag{1}$$

It can be shown that the posterior given $X_1, X_2, \ldots, X_n$ is again a Dirichlet with $\alpha + \sum_1^n \delta_{X_i}$, in place of α, where δ_{X_i} is the point mass at X_i. Using this fact and (1), one gets immediately,

$$E\left(P(A)|X_1, X_2, \ldots, X_n\right) = \frac{\alpha(\mathbf{R})}{\alpha(\mathbf{R}) + \mathbf{n}} \bar{\alpha}(A) + \frac{n}{\alpha(\mathbf{R}) + \mathbf{n}} \left(\frac{1}{n} \sum \delta_{X_i}(A) \right) \tag{2}$$

which is a convex combination of the prior guess $\bar{\alpha}(A)$ and the frequentist nonparametric maximum likelihood estimate $P_n(A) = \frac{1}{n} \sum \delta_{X_i}(A)$. The weights reflect the Bayesian's confidence in prior guess. One can elicit or choose $\bar{\alpha}(.)$ and $\alpha(\mathbf{R})$ — and hence $\alpha(.)$— from these considerations.

We denote the Dirichlet process by D_α.

PROPOSITION. If Π is D_α and B is a weak neighborhood of true P_0, then $\Pi\left(B|X_1, X_2, \ldots, X_n\right) \to 1$ a.s. (P_0), i.e., posterior consistency holds for all P_0.

At the heart of this fact is the property of being tailfree,vide Ferguson(1974), which allows one to reduce an infinite dimensional problem to a finite dimensional problem and invoke posterior consistency for the latter. This idea as well as the introduction of Dirichlet for another infinite dimensional problem goes back to Freedman(1963).

3.2 SEMIPARAMETRICS

We start with a famous example of Diaconis and Freedman(1986). Suppose we wish to make inference about θ and $P_\theta(.) = P(. - \theta)$ where θ is real and $P(.)$ is symmetric around zero. To put a prior distribution for P_θ one first chooses a P' using a D_α, symmetrizes P' to get P and independently chooses θ. Diaconis and Freedman(1986) show that the posterior for θ need not be consistent in the weak topology.

Various people have observed that semiparametrics should involve probability measures with densities but the Dirichlet assigns probability one to the set of discrete measures. However choosing priors on densities is not enough.

Ghosal, Ghosh and Ramamoorthi(1998) have pointed out that one may argue that the Diaconis–Freedman counter example occurs because of the breakdown of the tailfree property. They show that posterior consistency can be proved provided a condition used by Schwartz(1965) holds. Priors for which posterior consistency holds are exhibited in Ghosal, Ghosh and Ramamoorthi(1998).

The version of Schwartz's(1965) theorem one has to use for this purpose is given below. We now work with $\mathcal{P}$ = the set of probability measures P having a

density f with respect to Lebesgue measure. For two such probability measures P_1, P_2, with densities f_1, f_2 the Kullback–Leibler number $K(P_1, P_2)$ is defined as $\int_{\mathbf{R}} f_1 \log \frac{f_1}{f_2} dx$.

$K(P_1, P_2)$ is always ≥ 0 and may be ∞. It is not a metric but measures the divergence between P_1 and P_2 with the extreme tail of the density playing an important role.

THEOREM 1 *Suppose P_0 belongs to the Kullback–Leibler support of Π, i.e., for all $\delta > 0$,*

$$\Pi\{K(P_0, P) < \delta\} > 0 \tag{3}$$

Then $\Pi(B|X_1, X_2, \ldots, X_n) \to 1$ a.s. (P_0), for all weak neighborhoods B of P_0.

As Ghosal, Ghosh and Ramamoorthi(1998) show property(3)—unlike the tail-free property — continues to hold even with the addition of a finite dimensional parameter.

For later reference as well as completeness we record Schwartz's(1965) theorem in its original form and an extension due to Barron(1988,1998).

THEOREM 2 *Let Π be a prior on $\mathcal{P}$, and $P_0 \in B$. Assume the following conditions:*

1. *$\Pi(K(P_0, P) < \delta) > 0$ for all $\delta > 0$;*
2. *There exists a uniformly consistent sequence of tests for testing $H_0 : P = P_0$ vs. $H_1 : P \in B^c$, i.e., there exists a sequence of tests $\phi_n(X_1, X_2, \ldots, X_n)$ such that as $n \to \infty$,*

$$E_{P_0}\phi_n(X_1, X_2, \ldots, X_n) \to 0 \text{ and } \inf_{P \in B^c} E_P\phi_n(X_1, X_2, \ldots, X_n) \to 1.$$

Then $\Pi(B|X_1, X_2, \ldots, X_n) \to 1$ a.s. P_0.

THEOREM 3 ((BARRON(1988,1998))) *Let Π be a prior on $\mathcal{P}$, and P_0 be in $\mathcal{P}$ and B be a neighborhood of P_0. Assume that $\Pi(K(P_0, P) < \delta) > 0$ for all $\epsilon > 0$. Then the following are equivalent.*

1. *There exists a β_0 such that*

$$P_0\{\Pi(B^c|X_1, X_2, \ldots, X_n) > e^{-n\beta_0} \text{ infinitely often}\} = 0;$$

2. *There exist subsets V_n, W_n of $\mathcal{P}$, positive numbers $c_1, c_2, \beta_1, \beta_2$ and a sequence of tests $\{\phi_n(X_1, X_2, \ldots, X_n)\}$ such that*
   (a) $B^c = V_n \cup W_n$,
   (b) $\Pi(W_n) \leq C_1 e^{-n\beta_1}$,
   (c) $P_0\{\phi_n(X_1, X_2, \ldots, X_n) > 0 \text{ infinitely often}\} = 0$ and $\inf_{P \in V_n} E_P\phi_n \geq 1 - c_2 e^{-n\beta_2}$.

4 DENSITY ESTIMATION

4.1 DIRICHLET MIXTURE OF NORMALS

We illustrate with what seems to be currently the most popular and successful Bayesian method, first proposed by Lo(1984) and implemented in the early nineties via Markov Chain Monte Carlo(1994) by Escobar, Mueller and West (94).

Choose a random $P' \sim D_\alpha$. Since P' is discrete, as observed before,form a convolution with a normal density $N(0,h)$. Let $P = P' * N(0,h)$.

Since the smoothness of P depends on h and one does not know how much smoothness is right, put a prior(usually, inverse gamma)on h also. This completes the specification of a prior,which is often called a Dirichlet mixture of normal. It turns out that for MCMC to be feasible one needs α also to be normal. Simulations and heuristic calculations show that one can improve the rate of convergence by adding a location and scale parameter to α and by putting a prior on these parameters also. The following discussion can handle these refinements as well as general nonnormal α. However for the normal α,one can supplement the discussion below with non trivial heuristic argument that throws light on how convergence takes place. For lack of space the heuristic argument will not be given.

4.2 POSTERIOR CONSISTENCY FOR GENERAL PRIORS

The basic theorem is the following which improves on an earlier result of Barron,Schervish and Wasserman(1997).

Let $\mathcal{P}_0 \subset \mathcal{P}$. For $\delta > 0$, the L_1- metric entropy of $\mathcal{P}_0$, denoted by $J(\delta, \mathcal{P}_0)$ is $\log a(\delta)$, where $a(\delta)$ is the minimum over all k such that there exist $P_1, P_2, \cdots, P_k$ in $\mathcal{P}$ with $\mathcal{P}_\prime \subset \cup_1^k \{P : \|P - P_i\|_1 < \delta\}$.

THEOREM 4 (GHOSAL,GHOSH AND RAMAMOORTHI) *Let* Π *be a prior on* $\mathcal{P}$. *If* $P_0 \in \mathcal{P}$ *and* $\Pi(K(P_0, P) < \epsilon) > 0$ *for all* $\epsilon > 0$. *If for each* $\epsilon > 0$ *there is a* $\delta < \epsilon, c_1, c_2 > 0, \beta < \frac{\epsilon^2}{2}$ *and also* $\mathcal{P}_n$ *such that*

1. $\Pi(\mathcal{P}_n^c) < C_1 e^{-n\beta_1}$ *for large* n
2. $J(\delta, \mathcal{P}_n) < n\beta$

then $\Pi(B|X_1, X_2, \ldots, X_n) \to 1$ *a.s.*$P_0 n$ *for all* L_1*-neighborhoods* B *of* P_0.

The proof of this theorem is based on the result of Barron recorded in Section 3. The first assumption is the condition assumed in Theorem 1 in Section 3 while the two remaining assumptions take care of conditions(2) and (3) of Barron's Theorem.

4.3 APPLICATION TO DIRICHLET MIXTURE OF NORMALS

One has to have two sets of tools to verify the two conditions in Theorem 4. The set or sieve $\mathcal{P}_n$ for verifying the condition is: fix a δ and β as in the theorem then

$$\mathcal{P}_n = \left\{ P = P' * N(0,h); P'[-\sqrt{n}, \sqrt{n}] > 1 - \delta, h > \frac{c(\delta, \beta)}{\sqrt{n}} \right\}$$

Various sufficient conditions which entail application of Theorem 4 are given in Ghosal, Ghosh and Ramamoorthi(97. For example if P_0 is smooth unimodal with finite Shannon entropy and compact support, like the uniform on $[a, b]$ then P_0 belongs to the Kullback–Leibler support of the prior. For unbounded support the tails of P_0 and $\bar{\alpha}$ have to be compatible in a certain way.

4.4 Concluding Remarks

Theorem 4 can also be used to study posterior consistency for Gaussian process priors and Bayesian histograms (Barron(1988,1998) and Ghosh and Ramamoorthi(1998)).

One may also ask whether the Bayes estimate $E(P|X_1, X_2, \ldots, X_n)$ is consistent. It is easy to show that posterior consistency in the weak topology or the topology induced by L_1 norm implies Bayes consistency.

One may also ask questions about rates of convergence and non-informative or default priors which attain a minimax rate of convergence for the posterior or Bayes estimates. This issue is currently under investigation by Ghosal,Ghosh and van der Vaart and by Wassserman and Shen.

A final important remark. In recent work Barron(1998) shows if we focus on the cumulative Kullback-Leibler predictive loss (also called the entropy loss) an elegant consistency theory can be built up using only Kullback-Leibler support.

References

Barron, A. R. (1986). On uniformly consistent tests and Bayes consistency. Unpublished manuscript.

Barron, A. R., Schervish, M. and Wasserman, L. (1996). The consistency of posterior distributions in non parametric problems. Preprint.

Barron, A. R. (1998). Information-theoretic characterizations of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems.*Bayesian Statistics 6*, Editors *J.M. Bernardo,J.O. Berger, A.P. Dawid and A.F.M. Smith* Oxford University press

Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* 14 1–67.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems.*Ann. Statist.* 1 209-230.

Ferguson, T. (1974). Prior distribution on the spaces of probability measures. *Ann. Statist.* 2 615-629.

Freedman, D. (1963). On the asymptotic distribution of Bayes estimates in the discrete case I. *Ann. Math. Statist.* 34 1386–1403.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1998). Consistent semiparametric estimation about a location parameter. Submitted.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997b). Posterior consistency of Dirichlet mixtures in density estimation. Technical report # WS-490, Vrije Universiteit, Amsterdam.

S.Ghosal, J.K. Ghosh and R.V. Ramamoorthi (1997) Consistency issues in Bayesian Nonparametrics . *Asymptotics, Nonparametrics and Time series - a tribute to M.L. Puri* Edited by S.Ghosh.Marcel Dekker

Ghosh, J. K. and Ramamoorthi R. V. (1998). *Lecture notes on Bayesian asymptotics.* Under preparation.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Ann. Statist.* 12 351–357.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* 4 10–26.

West, M., Muller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of uncertainty: A Tribute to D. V. Lindley.* 363–386.

Jayanta K. Ghosh
Indian Statistical Institute
203 B.T. Road
Calcutta 700 035
India

Lattice Point Problems and the Central Limit Theorem in Euclidean Spaces

F. Götze[1]

Abstract. A number of problems in probability and statistics lead to questions about the actual error in the asymptotic approximation of nonlinear functions of the observations. Recently new methods have emerged which provide optimal bounds for statistics of quadratic type. These tools are adaptions of methods which provide sharp bounds in some high dimensional lattice point remainder problems and solve some problems concerning the distribution of values of quadratic forms.

1991 Mathematics Subject Classification: Primary 62E20, 11P21 Secondary 60F05

Keywords and Phrases: Edgeworth expansion, U-statistics, Central Limit Theorem, lattice point rest, ellipsoids, irrational quadratic forms

1. Introduction.

Let $X_1, \ldots, X_n$ denote independent and identically distributed random vectors in R^d, $d \geq 1$.

Example 1.1 Assume that X_1 takes values in the finite set $\{-1, 1\}^d \subset \mathbb{R}^d$ with equal probability 2^{-d}. Write

$$S_n = n^{-1/2}(X_1 + \ldots + X_n).$$

By the Central Limit Theorem (CLT) the sequence of random vectors S_n converges in distribution to a multivariate Gaussian distribution with mean zero and identity covariance matrix. Let $|m|^2 = \langle m, m \rangle$ denote the d-dimensional Euclidean norm and scalar-product. A number of statistical problems require to determine asymptotic approximations for the distribution of test statistics of type

$$T_n = |S_n|^2.$$

It is well known that the distribution function (d.f.) $\mathbf{P}\{T_n \leq v\}$ converges to the χ^2–distribution function with d degrees of freedom, say $\chi(v)$, for all $v \in \mathbb{R}$. In order to measure the error of this approximation we shall use the Kolmogorov distance and would like to determine the optimal exponents $\alpha > 0$ such that for a constant $c > 0$ independent of n

$$\delta_n = \sup_{v \geq 0} |\mathbf{P}\{T_n \leq v\} - \chi(v)| \leq c n^{-\alpha}. \tag{1.1}$$

[1]Research supported by the SFB 343, 'Diskrete Strukturen in der Mathematik', Bielefeld.

Here $T_n \le v$ means that the sum $\sqrt{n}\,S_n$ is contained in a ball $B_{vn} = \{|x| \le \sqrt{v\,n}\}$.

General estimates in the multivariate CLT (Sazonov [Sa], Bhattacharya and Rao [BR]) established the rate $\alpha = 1/2$ uniformly in the class of convex sets. Hence for balls and ellipsoids the achievable rate α should be at least $1/2$.

In Example 1.1 the sum S_n takes values in a lattice. By the local limit theorem its discrete density may be approximated by a Gaussian density such that

$$\mathbf{P}\{S_n = \frac{m}{\sqrt{n}}\} = \varphi_n(m)\big(1 + \mathcal{O}(n^{-1})\big), \quad \varphi_n(m) := \frac{1}{(2\pi n)^{d/2}} \exp\Big\{ -\frac{|m|^2}{2n} \Big\}.$$

Hence bounds in (1.1) can be derived from estimates of

$$\sup_v | \sum_{m \in B_{vn} \cap \mathbb{Z}^d} \varphi_n(m) - \chi(v)|.$$

Since the weights $\varphi_n(m)$ are 'smoothly' depending on m, the problem might be further reduced to the case of constant weights, which leads to a problem about counting the lattice points in B_{vn}. In this way Esseen [E] and Yarnold [Y] have proved

THEOREM 1.2.

$$\mathbf{P}\{T_n \le v\} - \chi(v) = \exp\{-v/2\}\,\Delta(B_{vn}) + \mathcal{O}(n^{-1}). \tag{1.2}$$

Here $\Delta(A)$ denotes the relative lattice point remainder given by

$$\Delta(A) := \frac{\mathrm{vol}_{\mathbb{Z}}\, A - \mathrm{vol}\, A}{\mathrm{vol}\, A}, \tag{1.3}$$

with $\mathrm{vol}_{\mathbb{Z}}\, A$ and $\mathrm{vol}\, A$ denoting the number of points of the standard lattice $\mathbb{Z}^d$ in A and the volume of A respectively.

The relation (1.2) obviously establishes for Example 1.1 an equivalence between bounds in the lattice point remainder problem for ellipsoids and bounds of type (1.1) in the multivariate CLT. Indeed, Landau [L1] and Esseen [E] proved

$$\Delta(B_s) = \mathcal{O}\big(s^{-d/(d+1)}\big) \quad \text{resp.} \quad \delta_n = \mathcal{O}\big(n^{-d/(d+1)}\big).$$

Note though that Esseen's bound holds for balls and *arbitrary* i.i.d. random vectors X_j with finite fourth moment and identity covariance operator, where an equivalence of type (1.2) is not known.

Example 1.1 provides as well lower bounds for the error. Notice that nT_n assumes integer values in the interval $[-dn, dn]$. Distributing probability 1 among these values there exists an integer j such that

$$\mathbf{P}\{T_n = j\,n^{-1}\} \ge c\,n^{-1}, \quad c = 1/(2d+1).$$

Comparing the piecewise constant function $v \mapsto \mathbf{P}\{T_n \le v\}$ with the smooth limit $v \mapsto \chi(v)$, we find the lower bound $\delta_n \ge cn^{-1}$. Hence the rates α in (1.1) are restricted to $1/2 \le \alpha \le 1$.

This lecture is organized as follows. Section 2 contains results in the CLT for quadratic statistics in Euclidean spaces. Corresponding results in lattice point problems are described in section 3. Section 4 contains applications to distributions of values of positive definite and indefinite forms. Finally, in Section 5 we describe inequalities for trigonometric sums which are essential for these results.

A major part of the results in this lecture represents joint work with V. Bentkus.

2. APPROXIMATIONS IN THE CLT FOR QUADRATIC STATISTICS.

The CLT in Euclidean Spaces. Let $X, X_1, X_2, \dots$ be a sequence of i.i.d. random vectors taking values in the d-dimensional Euclidean space $\mathbb{R}^d$ including the case $d = \infty$ of infinite dimensional real Hilbert spaces. We assume that X has mean zero and $|X|$ has a finite second moment. Then the sums S_n converge weakly to a mean zero Gaussian random vector, say G, with covariance equal to the covariance of X. Assume that G is not concentrated on a proper subspace of $\mathbb{R}^d$. Let Q denote a bounded linear operator on $\mathbb{R}^d$. Consider the quadratic form $\mathbb{Q}[x] = \langle Qx, x\rangle$ and assume that Q is non-degenerated, that is $\ker Q = \{0\}$.

The distribution of the quadratic form $\mathbb{Q}[G]$ is determined by its distribution function, say $\chi(v)$, and may be represented up to a shift as the distribution of a finite (resp. infinite) weighted sum of squares of i.i.d. standard Gaussian variables.

Rates of approximation in (1.1) in the CLT for $T_n = \mathbb{Q}[S_n]$ have been intensively studied especially in the infinite dimensional case in view of applications to non parametric goodness-of-fit statistics based on empirical distributions. Unfortunately the techniques of multivariate Fourier inversion of earlier results like that of Esseen [E] cannot be applied here. Several approaches have been developed for this problem.

A probabilistic approach is based on the Skorohod embedding resp. the KMT–method and provided bounds of order $\alpha = 1/4$, Kiefer [Ki], resp. $\mathcal{O}(n^{-1/2}\log n)$, Csörgö [Cs]. An analytic approach is based on a Weyl type inequality for characteristic functions, see (5.4). Using this technique, rates $\alpha = 1 - \varepsilon$ for any $\varepsilon > 0$ have been proved in (1.1), see [G1] and for refinements Bentkus and Zalesskii [BZ] and Nagaev and Chebotarev [NC]. Moreover, using methods like (5.4) the approximation $\chi(v)$ may be refined by asymptotic expansions in (1.1) up to an error of order $\mathcal{O}(n^{-k/2+\varepsilon})$ for *polynomials* of S_n of degree $k \geq 2$, see[G3].

Results providing *optimal* bounds of order $\alpha = 1$ are based on techniques used in related bounds for the corresponding lattice point problems. For *diagonal* quadratic forms and vectors X with *independent* coordinates the rate $\alpha = 1$ was proved for $d \geq 5$ in [BG1]. Here the additive structure of $\mathbb{Q}[x]$ allows to apply a discretization of type (5.5) and a version of the Hardy-Littlewood method of analytic number theory.

New tools described in (5.5)–(5.6) lead to the following result.

THEOREM 2.1. [BG2]. *Let* $\mathbf{E}\, X = 0$ *and* $\beta_4 = \mathbf{E}\,|X|^4 < \infty$. *Assume that* $d \geq 9$ *or* $d = \infty$. *Then*

$$\sup_v \Big|\, \mathbf{P}\big\{\mathbb{Q}[S_n] \leq v\big\} - \mathbf{P}\big\{\mathbb{Q}[G] \leq v\big\}\,\Big| = \mathcal{O}\big(n^{-1}\big). \tag{2.4}$$

The constant in this bound depends on β_4, *the eigenvalues of* Q *and the covariance operator of* G *only.*

REMARK 2.2.

1) For $d = 8$ the bound $\mathcal{O}(n^{-1}\ln^{\delta} n)$ holds with some $\delta > 0$.

2) Similar results like (2.4) hold for $\mathbb{Q}[x-a]$ involving an arbitrary center $a \in \mathbb{R}^d$. Here the approximation by the limit d.f. $\mathbf{P}\big\{\mathbb{Q}[G-a] \leq v\big\}$ needs to be improved by a further expansion term, say $n^{-1/2}\chi_1(v;a)$, which vanishes for $a = 0$.

3) For dimensions $d > 9$ including the case $d = \infty$ *uniform* bounds in (2.4), for $Q = \mathrm{Id}$ say, depend on moments of X and on lower bounds for a *finite* number, say m, of the largest eigenvalues of the covariance operator of X. For such bounds the minimal number $m \le d$ of eigenvalues needed has recently been determined to be $m = 12$, see [GU].

These results can be extended as follows.

U-Statistics. Let $X, X_1, \dots, X_n$ be i.i.d. random variables taking values in an arbitrary measurable space $(\mathcal{X}, \mathcal{B})$ and let $g : \mathcal{X} \to \mathbb{R}$, $h : \mathcal{X}^2 \to \mathbb{R}$ denote real-valued measurable functions. Assume that $h(x, y) = h(y, x)$, for all $x, y \in \mathcal{X}$ and $\mathbf{E}\, h(x, X) = 0$ for almost all $x \in \mathcal{X}$. Consider the so called degenerated U-statistic

$$T_n = \frac{1}{n} \sum\nolimits_{1 \le i < j \le n} h(X_i, X_j) + \frac{1}{\sqrt{n}} \sum\nolimits_{1 \le i \le n} g(X_i), \tag{2.5}$$

and write $\beta_s = \mathbf{E}\,\big|g(X)\big|^s$ and $\gamma_s = \mathbf{E}\,\big|h(X_1, X_2)\big|^s$. Assuming that γ_2 is positive and $\beta_2 + \gamma_2$ is finite, the U-statistic T_n converges to a weighted χ^2–type distribution, say χ. Using a further expansion term, say χ_1, the problem is to derive explicit estimates for the error

$$\delta_n = \sup_v |\mathbf{P}\{T_n \le v\} - \chi(v) - n^{-1/2}\chi_1(v)|. \tag{2.6}$$

Rates of order $\delta_n = o(n^{-1/2})$ have been proved by Korolyuk and Borovskich [KB]. Moreover, for degenerated U-statistics of *any* degree $k \ge 2$ asymptotic approximations have been established up to errors $\delta_n = \mathcal{O}(n^{-k/2+\epsilon})$ in [G2].

Using similar techniques as in Theorem 2.1 the following explicit bound with optimal rate $\alpha = 1$ holds.

THEOREM 2.3. [BG4]. *Let q_j denote the eigenvalues (ordered by decreasing absolute value) of the Hilbert-Schmidt operator induced on $L^2(\mathcal{X})$ by the kernel h. Write $\gamma_{s,r} = \mathbf{E}\,\big(\mathbf{E}\,\big(\big|h(X_1, X_2)\big|^s \,\big|\, X_2\big)\big)^r$ and $\sigma^2 := \gamma_2$. If $q_{13} \ne 0$,*

$$\delta_n \le \frac{C}{n}\Big(\frac{\beta_4}{\sigma^4} + \frac{\beta_3^2}{\sigma^6} + \frac{\gamma_3}{\sigma^3} + \frac{\gamma_{2,2}}{\sigma^4}\Big), \qquad \textit{where } C \le \exp\Big\{\frac{c\sigma}{|q_{13}|}\Big\}. \tag{2.7}$$

REMARK 2.4. 1) In cases where the expansion term χ_1 vanishes the condition $q_9 \ne 0$ suffices to prove a similar bound.

2) The result can be extended to von Mises statistics, i.e. statistics including diagonal terms $h(X_j, X_j) := d(X_j)$, where $d(X)$ has mean zero. This allows to consider as well statistics like $T_n := |S_n - a|^2$.

It is likely that improvements in lattice point approximation problems (see the Conjecture in Section 3) allow to prove error bounds of order $\mathcal{O}(n^{-1})$ in Theorems 2.1 and 2.3 for dimensions $5 \le d \le 8$ as well.

3. Lattice Point Problems.

For a symmetric positive definite matrix Q consider the quadratic form $\mathbb{Q}[x] = \langle Qx, x\rangle$ on R^d and the corresponding ellipsoid

$$E_s := \big\{x \in \mathbb{R}^d : \mathbb{Q}[x] \le s\big\}, \quad \text{for } s \ge 0.$$

Special Ellipsoids. Using similar arguments as for δ_n in Section 1 a corresponding lower bound can be shown for the lattice point remainder $\Delta(E_s)$ (for $Q = Id$), namely

$$\Delta(E_s) = \Omega(s^{-1}), \quad d \geq 1. \tag{3.1}$$

For balls of dimensions $2 \leq d \leq 4$, the lattice point remainder $\Delta(E_s)$ admits sharper lower bounds, e.g.

$$\Omega(s^{-3/4}\log^{1/4} s),\ d=2, \quad \Omega(s^{-1}\log^{1/2} s),\ d=3, \quad \text{and} \quad \Omega(s^{-1}\log\log s),\ d=4,$$

due to Hardy [Ha], Szegö [Sz] and Walfisz [W2] respectively. The upper bound

$$\Delta(E_s) = \mathcal{O}(s^{-1}), \quad d \geq 5 \tag{3.2}$$

has been shown in a number of special cases. It holds for ellipsoids which are *rational*, that is the matrix Q is a multiple of a matrix with rational coefficients. Otherwise Q is called *irrational*. This result is due to Landau [L2] and Walfisz [W1] and depends on the rational coefficients in a non uniform way. For a detailed discussion see the monograph by Walfisz [W2].

For *diagonal* forms $\mathbb{Q}[x] = \sum_{j=1}^{d} q_j x_j^2$ with arbitrary $q_j > 0$, (3.2) is due to Jarnik [J1]. Moreover, if Q is irrational, Jarnik and Walfisz [JW] have shown that the bound

$$\Delta(E_s) = o(s^{-1}), \quad d \geq 5 \tag{3.3}$$

holds and is best possible for general irrational numbers q_j.

General Ellipsoids. For this class Landau [L1] obtained $\Delta(E_s) = \mathcal{O}(s^{-1+\lambda})$ with $\lambda = 1/(d+1)$ for $d \geq 1$, using Dirichlet series methods. His result has been extended by Hlawka [Hl] to convex bodies with smooth boundary and strictly positive Gaussian curvature, and improved to $\mathcal{O}(s^{-1+\lambda})$, with some $\lambda = \lambda(d) > 0$, $\lambda < 1/(d+1)$, by Krätzel and Nowak [KN1, KN2].

Assume without loss of generality that the smallest eigenvalue of Q is 1 and denote the largest eigenvalue by q. Hence $q \geq 1$. The following results provide optimal *uniform* bounds of type (3.2) resp. (3.3) for general ellipsoids.

THEOREM 3.1. [BG3, BG5]. *There is a constant $c > 0$ depending on d only and a function $\rho(s) \in [0,2]$, depending on Q, see (5.2), such that for all $s \geq 1$*

$$\sup_{a\in\mathbb{R}^d} \Delta(E_s + a) \leq c\, q^d\, s^{-1}\big(s^{-\lambda} + \rho(s)\big), \qquad \text{for } d \geq 9, \tag{3.5}$$

where $\lambda \stackrel{\text{def}}{=} \frac{1}{2}\left[\frac{d-1}{2}\right] - 1$, *and*

$$\lim_{s\to\infty} \rho(s) = 0 \qquad \text{if and only if} \qquad Q \text{ is irrational.}$$

If $d = 8$ the bound $\sup_{a\in\mathbb{R}^d} \Delta(E_s + a) \leq c\, q^8 s^{-1} \ln^2(s+1)$ still holds.

The error for generic forms $\mathbb{Q}[x]$ should be much smaller than for rational forms, which can be seen by the following *heuristic* argument. Let $C(m)$ denote the

cube of side length 1 centered at a lattice point $m \in \mathbb{Z}^d$ and let I_s denote the indicator function of E_s. Define ξ_m as function of a randomly chosen Q as $\xi_m = I_s(m) - \int_{C(m)} I_s(x)\,dx$. Then $|\xi_m| \le 1$ and we may assume that the ξ_m have mean zero. Let D_s denote the set of lattice points m such that $C(m)$ intersects ∂E_s. Note that $\xi_m = 0$ for $m \notin D_s$. Then

$$\Delta(E_s)\ \mathrm{vol}\, E_s = \sum_{m\in\mathbb{Z}^d} \xi_m = \sum_{m\in D_s} \xi_m. \tag{3.4}$$

Since E_s has diameter proportional to $r = \sqrt{s}$, the sum in (3.4) extends over $\mathcal{O}(r^{d-1})$ nonzero summands only. If the random variables ξ_m are approximately independent the CLT implies for $r \to \infty$ with probability tending to 1 that (3.4) is smaller than $r^{(d-1)/2}\log r$. Hence one would expect that $\Delta(E_s) = \mathcal{O}(s^{-(d+1)/4}\log s)$. Indeed, Jarnik [J2] proved for $d \ge 4$ an upper bound of order $\mathcal{O}(s^{-d/4+\varepsilon})$ for Lebesgue almost all diagonal forms. For generic forms Landau [L3] established $\Delta(E_s) = \Omega(s^{-(d+1)/4})$. The results described so far suggest the following hypothesis about worst and generic case errors.

CONJECTURE. For any $\epsilon > 0$ the relative lattice point remainder is of order

$$\begin{aligned}
\Delta(E_s + a) &= \mathcal{O}(s^{-1}), && d \ge 5, && \text{for all } Q \text{ and } a,\\
&= o(s^{-1}), && d \ge 5, && \text{for irrational } Q,\\
&= \mathcal{O}(s^{-(d+1)/4+\epsilon}), && d \ge 2, && \text{for Lebesgue almost all } Q \text{ and } a.
\end{aligned}$$

4. Distribution of Values of Quadratic Forms.

Positive Definite Forms. For fixed $\delta > 0$ consider the shells $E_{s+\delta} \setminus E_s = \{x \in \mathbb{R}^s : s \le \mathbb{Q}[x] \le s+\delta\}$. Theorem 3.1 implies

COROLLARY 4.1. *For $d \ge 9$ and irrational Q we have*

$$\lim_{s\to\infty} \frac{\mathrm{vol}_{\mathbb{Z}}\big(E_{s+\delta} \setminus E_s\big)}{\mathrm{vol}(E_{s+\delta} \setminus E_s)} = 1. \tag{4.1}$$

This result may be applied as well to shrinking intervals of size $\delta = \delta(s) \to 0$ as s tends to infinity. The quantity $\mathrm{vol}(E_{s+\delta} \setminus E_s)$ measures the number of values of a positive quadratic form in an interval $(s, s+\delta]$, counting these values according to their multiplicities.

Let s and $n(s)$ denote successive elements of the ordered set $\mathbb{Q}[\mathbb{Z}^d]$ of values of $\mathbb{Q}[m]$. Davenport and Lewis [DL] conjectured that the distance between successive values, that is $n(s) - s$, converges to zero as s tends to infinity for irrational quadratic forms $\mathbb{Q}[x]$ and dimensions $d \ge 5$. They proved in [DL] that there exists a dimension d_0 such that for all $d \ge d_0$ and any given $\varepsilon > 0$ and any lattice point m with sufficiently large norm $|m|$ there exist another lattice point $\overline{m} \in \mathbb{Z}^d$ such that $\big|\mathbb{Q}[m+\overline{m}] - \mathbb{Q}[\overline{m}]\big| < \varepsilon$. This does not rule out the possibility of arbitrary large gaps between possible clusters of values $\mathbb{Q}[m]$, $m \in \mathbb{Z}^d$. This result has been improved by Cook and Raghavan [CR], providing the bound $d_0 \le 995$. Corollary 4.1 now solves this problem for $d \ge 9$.

Define the maximal gap between the values $\mathbb{Q}[m-a]$, $m \in \mathbb{Z}^d$ in the interval $[\tau, \infty)$ as $d(\tau; \mathbb{Q}, a) = \sup_{s\ge\tau}\big(n(s) - s\big)$. Then (4.1) implies

COROLLARY 4.2. [BG5]. *Assume that $d \geq 9$ and that $\mathbb{Q}[x]$ is positive definite. If Q is irrational then* $\sup_{a\in\mathbb{R}^d} d(\tau;\mathbb{Q},a) \to 0$, *as* $\tau \to \infty$.

Indefinite Forms and the Oppenheim conjecture. Assume that Q is irrational and *indefinite*. Consider the infimum value of $\mathbb{Q}[m]$ for nonzero lattice points $m \in \mathbb{Z}^d$

$$M(Q) = \inf\big\{\,\big|\,\mathbb{Q}[m]\,\big|:\ m\neq 0,\ m\in\mathbb{Z}^d\big\}.$$

Oppenheim [O1] conjectured that $M(Q) = 0$, for $d \geq 5$ and irrational *indefinite* Q, and has shown that this implies that the set $\mathbb{Q}[\mathbb{Z}^d]$ is dense in $\mathbb{R}$ for $d \geq 3$, see [O2]. This conjecture has been proved, e.g. for diagonal forms and $d \geq 5$ by Davenport and Heilbronn [DH] and for general forms and $d \geq 21$ by Davenport [Da]. For a review, see Margulis [Mar2]. It has been finally established for all dimensions $d \geq 3$ by Margulis [Mar1].

Let C_s denote a d–dimensional cube of side length $\sqrt{s}$ and center 0. The results of Theorem 3.1 are a consequence of more general asymptotic expansion of $\mu_s\{\mathbb{Q}[x] \leq \beta\}$ in powers of s^{-1} for certain 'smooth' distributions μ_s on $\mathbb{Z}^d$ with support in C_s, see [BG5, Theorem 2.1]. For indefinite forms this result yields the following refinement of Oppenheim's conjecture for dimensions $d \geq 9$.

For a sufficiently small positive constant, say $c_0 = c_0(d)$, let $d(s)$ denote the maximal gap in the finite set of values $\mathbb{Q}[m]$ such that $-c_0 s \leq \mathbb{Q}[m] \leq c_0 s$ and $m \in C_{s/c_0^2} \cap \mathbb{Z}^d$. Then

THEOREM 4.3. [BG5]. *For $d \geq 9$ the maximal gap satisfies*

$$d(s) \ll_d q^{3d/2}\big(s^{-\lambda} + \rho(s)\big) \qquad \text{for } s \geq c_0^{-1} q^{3d/2},$$

with $\rho(s) \leq 2$ defined in (5.2) and λ given in Theorem 3.1.

The *quantitative* version of Oppenheim's conjecture by Dani and Margulis [DM] describes the uniformity of the distribution of the set of values $\mathbb{Q}[\mathbb{Z}^d \cap C_s]$ for star-shaped sets like the cubes C_s introduced above. For a fixed interval $[\alpha,\beta]$ let $V_{\alpha,\beta}$ denote the set of $x \in \mathbb{R}^d$ such that $\mathbb{Q}[x] \in [\alpha,\beta]$. Eskin, Margulis and Mozes proved the following result using ergodic theory for unipotent groups.

THEOREM 4.4. [EMM]. *For any irrational indefinite form Q of signature (p,q) with $q \geq 3$,*

$$\frac{\mathrm{vol}_{\mathbb{Z}}\big(V_{\alpha,\beta}\cap C_s\big)}{\mathrm{vol}\big(V_{\alpha,\beta}\cap C_s\big)} = 1 + o(1), \qquad \text{as } s\to\infty. \tag{4.3}$$

In particular (4.3) holds for all indefinite irrational forms with $d \geq 5$.

Using expansion results for arbitrary forms, the error term in this convergence result can be explicitly estimated for $d \geq 9$, see [BG5, Theorem 2.6].

5. INEQUALITIES FOR CHARACTERISTIC FUNCTIONS AND TRIGONOMETRIC SUMS.

In order to prove the results of Sections 2–4, characteristic functions of $\mathbb{Q}[S_n]$ and weighted trigonometric sums, say $f(t)$, are used. In the latter case the weights are

given by a uniform distribution on the lattice points in the cube C_{2s} smoothed at the boundary of C_{2s} by convolutions with uniform distributions on some sufficiently small cubes, retaining constant weights in the center part $C_s \subset C_{2s}$. A simplified version of these weighted trigonometric sums, used in the explicit bounds of Theorems 3.1 and 4.3, is defined as follows. Let

$$\varphi_a(t;s) = \Big| \big(\mathrm{vol}_{\mathbb{Z}}\, C_s\big)^{-3} \sum_{x_j \in \mathbb{Z}^d \cap C_s} \exp\big\{ i\,t\,\mathbb{Q}[x_1+x_2+x_3-a] \big\} \Big|. \tag{5.1}$$

Note that $\varphi_a(t;s)$ is normalized so that $|\varphi_a(t;s)| \le \varphi_a(0;s) = 1$. Define

$$\gamma\big(s,T\big) = \sup_a \sup_{s^{-1/2} \le t \le T} \varphi_a(t;s).$$

It can be shown that $\lim_{s\to\infty} \gamma\big(s,T\big) = 0$ iff Q is irrational. Finally, given $d \ge 9$ and ε with $0 < \varepsilon < \kappa := 1 - 8/d$, the characteristic $\rho(s)$ of Theorem 3.1 and 4.3 is given by

$$\rho(s) = \inf_{T \ge 1} \Big(T^{-1} + \gamma\big(s,T\big)^{\kappa-\varepsilon}\, T^{\varepsilon} \Big). \tag{5.2}$$

The connections between the probability resp. counting problems and $f(t)$ are made by means of Fourier inversion inequalities based on Beuerling type functions, see Prawitz [Pr], which bound δ_n resp. $|\Delta(E_s)|$ by

$$\int_{-1}^{1} |f(t) - g(t)|\, t^{-1}\, dt + \int_{-1}^{1} \big(|f(t)| + |g(t)|\big)\, dt. \tag{5.3}$$

Here $g(t)$ is the continuous approximation to $f(t)$ replacing the distribution of S_n by a Gaussian distribution resp. the counting measure by the Lebesgue measure.

In the CLT the following version of Weyl's [We] difference scheme for sums of $\mathbb{R}^d$–valued, independent random vectors, say U, V (with identical distribution) and Z, W is used. Let $\bar{X}$ denote an independent copy of X and let $\widetilde{X} = X - \bar{X}$ be its symmetrization. The inequality

$$\Big|\, \mathbf{E}\, \exp\{i\,t\,\mathbb{Q}[U+V+Z+W]\}\Big|^2 \le\ \mathbf{E}\, \exp\{2\,i\,t\,\langle Q\widetilde{U}, \widetilde{Z}\rangle\}, \tag{5.4}$$

now reduces the estimation of $f(t)$ in (5.3) to bounds of order $\mathcal{O}(n^{-1+\varepsilon})$ for conditional linear forms, but in a restricted domain $|t| \le n^{-\epsilon}$ only. This leads to rates $a = 1 - \varepsilon$ in (1.1), see [G1].

In order bound the integral (5.3) by $\mathcal{O}(n^{-1})$, this Weyl step is followed by a discretization step for positive definite functions $H : \mathbb{R}^d \to \mathbb{R}$. For even $n = 2\,l$ and binomial weights $p_n(k) = \binom{n}{l-k}/2^n$, bounds like

$$\mathbf{E}\, H(S_n) \le \frac{1}{n} \sum_{j=1}^{n} \mathbf{E}\, \Big(\sum_{|k| \le l} p_n(k) H\big(k\, n^{-1/2}\, \widetilde{X}_j\big) \Big), \tag{5.5}$$

reduce the support of X to $\mathbb{Z}^d$ and replace characteristic functions of S_n by weighted trigonometric sums.

Finally, for general Q, the desired bounds for weighted trigonometric sums, say $f(t)$, of type (5.1), are based on the following 'correlation' bound

$$\big|f(t)f(t+\varepsilon)\big| \le c\,q^d\left(\big(\varepsilon\, s\big)^{-d/2} + \varepsilon^{d/2}\right) \text{ for all } \ t\in\mathbb{R} \text{ and } \ \varepsilon \ge 0. \tag{5.6}$$

For $t = 0$ we have $f(t) = 1$ and (5.6) becomes a 'double large sieve' estimate for distributions on the lattice, see e.g. Bombieri and Iwaniec [BI]. The inequality (5.6) implies for $t_0 \le t_1$ with $0 < \delta \le |f(t_0)|, |f(t_1)| \le 2\,\delta$ that either

$$|t_0 - t_1| \ \le\ \lambda_r = c_1\delta^{-4/d}s^{-1} \qquad \text{or} \qquad |t_0 - t_1| \ \ge\ \kappa = c_2\delta^{-4/d}.$$

Thus either the arguments t_0 and t_1, where the trigonometric sums are of the same (large) order δ, nearly coincide or their distance has to be 'large' (dependent on δ and d). Hence the set of arguments t, where $f(t)$ assumes values in an interval $[\delta, 2\delta]$ like $A_\delta = \{t \ge v \,:\, \delta \le |f(t)| \le 2\delta\}$ with $v := s^{-2/d}$, may be roughly described as a set of intervals of size at most δ_r separated by 'gaps' of size at least κ. This allows to estimate part of (5.3) approximately as

$$\int_{A_\delta} |f(t)|\,\frac{dt}{t} \ll \sum_{l=0}^{L} \delta\,\lambda_r \frac{1}{v + l\kappa} \ll s^{-1}\delta^{1-8/d}\log\frac{1}{\delta},$$

with some L such that $L\kappa \le 1$. The sum of these parts for $\delta = 2^{-l}$, $l \in \mathbb{N}$ is now of order $\mathcal{O}(s^{-1})$, provided that $d > 8$, which explains the dimensional restriction of this method.

REFERENCES

[BZ] Bentkus, V., and Zalesskiĭ, B., *Asymptotic expansions with nonuniform remainders in the central limit theorem in Hilbert space*, Lithuanian Math. J **25** (1985), 199–208.

[BG1] Bentkus, V. and Götze, F., *Optimal rates of convergence in the CLT for quadratic forms*, Ann. Prob. **24** (1996), 468-490.

[BG2] ———, *Uniform rates of convergence in the CLT for quadratic forms in multidimensional spaces*, Probab. Theory Relat. Fields **109** (1997), 367–416.

[BG3] ———, *On the lattice point problem for ellipsoids*, Acta Arithm. **80** (1997), 101–125.

[BG4] ———, *Optimal bounds in non-Gaussian limit theorems for U-statistics*, Preprint 97–077 SFB 343, Universität Bielefeld (1997) (to appear in Annals Probab. 1999).

[BG5] ———, *Lattice Point Problems and Distribution of Values of Quadratic Forms*, Preprint 97-125, SFB 343, University of Bielefeld (1997).

[BR] Bhattacharya, R. N. and Ranga Rao, R., *Normal Approximation and Asymptotic Expansions*, Wiley, New York, 1986.

[BI] Bombieri, E. and Iwaniec, H., *On the order of* $\zeta(1/2+it)$, Annali Scuola Normale Superiore–Pisa (4) **13** (1986), 449–472.

[CR] Cook, R.J. and Raghavan, S., *Indefinite quadratic polynomials of small signature*, Monatsh. Math. **97** (1984), no. 3, 169–176.

[Cs] Csörgo, S., *On an asymptotic expansion for the von Mises* ω^2*–statistics*, Acta Sci. Math. **38** (1976), 45–67.

[DM] Dani, S. G. and Margulis, G. A., *On orbits of unipotent flows on homogeneous spaces*, Ergod. Theor. Dynam. Syst. **4** (1984), 25–34.

[Da] Davenport, H., *Indefinite quadratic forms in many variables* (II), Proc. London Math. Soc. **8** (1958), no. 3, 109–126.

[DH] Davenport, H. and Heilbronn, H., *On indefinite quadratic forms in five variables forms*, Proc. London Math. Soc. **21** (1946), 185–193.

[DL] Davenport, H. and Lewis, D. J., *Gaps between values of positive definite quadratic forms*, Acta Arithmetica **22** (1972), 87–105.

[EMM] Eskin, A., Margulis, G. A. and Mozes, S., *Upper bounds and asymptotics in a quantitative version of the Oppenheim conjecture*, Ann. of Math. (2) **147** (1998), 93–141.

[Es] Esseen, C.G., *Fourier analysis of distribution functions*, Acta Math. **77** (1945), 1–125.

[G1] Götze, F., *Asymptotic expansions for bivariate von Mises functionals*, Z. Wahrsch. verw. Geb. **50** (1979), 333–355.

[G2] ———, *Expansions for von Mises functionals*, Z. Wahrsch. verw. Geb. **65** (1984).

[G3] ———, *Edgeworth expansions in functional limit theorems*, Ann. Probab. **17** (1989).

[GU] Götze F. and Ulyanov, V. V., *Uniform Approximations in the CLT for Balls in Euclidean Spaces* (1998) (In preparation).

[Ha] Hardy, G. H., *On Dirichlet's divisor problem*, Proc. London Math. Soc. (1916), 1–25.

[Hl] Hlawka, E., *Über Integrale auf konvexen Körpern* I, II, Mh. für Math. **54** (1950), 1–36, 81–99.

[J1] Jarnik, V., *Sur le points á coordonnees entiéres dans les ellipsoides á plusieurs dimensions*, Bull. internat. de l'acad. des sciences de Bohême (1928).

[J2] ———, *Über Gitterpunkte in Mehrdim. Ellipsoiden*, Math. Ann. **100** (1928), 699–721.

[JW] Jarnik, V. and Walfisz, A., *Über Gitterpunkte in Mehrdim. Ellipsoiden*, Math. Z. **32** (1930), 152–160.

[Ki] Kiefer, J., *Skorohod embeding of multivariate r. v. 's and the sample d. f.*, Z. Wahrsch. verw. Geb. **24** (1972), 1–35.

[KB] Koroljuk, V. S. and Borovskich, Yu.,V., *Theory of* U*-statistics*, Kluwer, Dordrecht, 1994.

[KN1] Krätzel, E. and Nowak, G., *Lattice points in large convex bodies*, Mh. Math. **112** (1991), 61–72.

[KN2] ———, *Lattice points in large convex bodies, II*, Acta Arithmetica **62** (1992), 285–295.

[L1] Landau, E., *Zur analytischen Zahlentheorie der definiten quadratischen Formen*, see also [LW], p. 11-29, Sitzber. Preuss. Akad. Wiss. **31** (1915), 458–476.

[L2] ———, *Über Gitterpunkte in mehrdim. Ellipsoiden*, Math. Z. **21** (1924), 126–132.

[L3] ———, *Über die Gitterpunkte in gewissen Bereichen (Vierte Abhandlung)*, see also [LW] , p. 71–84, Nachricht. Königl. Ges. Wiss. Göttingen (1924), 137–150.

[LW] Landau, E. –Walfisz, A., *Ausgewählte Abhandlungen zur Gitterpunktlehre* (A. Walfisz, ed.), Berlin, 1962.

[Mar1] Margulis, G. A., *Discrete subgroups and ergodic theory*, Number theory, trace formulas and discrete groups (Oslo, 1987), Academic Press, Boston, 1989, pp. 377–398.

[Mar2] Margulis, G. A., *Oppenheim conjecture*, Preprint (1997), 1–49.

[NC] Nagaev, S. V. and Chebotarev, V. I., *A refinement of the error estimate of the normal approximation in a Hilbert space*, Siberian Math. J. **27** (1986), 434–450.

[O1] Oppenheim, A., *The minima of indefinite quaternary quadratic forms*, Proc. Nat. Acad. Sci. USA **15** (1929), 724–727.

[O2] ———, *Values of quadratic forms II,III*, Quart. J. Math. Oxford Ser. (2) **4** (1953), 54–59, 60 – 66.

[Pr] Prawitz, H., *Limits for a distribution, if the characteristic function is given in a finite domain*, Skand. AktuarTidskr (1972), 138–154.

[Sa] Sazonov, V. V., *Normal approximation—some recent advances*, Lecture Notes in Math., 879, Springer, Berlin, 1981.

[Sz] Szegö, G., *Beiträge zur Theorie der Laguerreschen Polynome: Zahlentheoretische Anwendungen*, Math. Z. **25** (1926), 388–404.

[W1] Walfisz, A., *Über Gitterpunkte in mehrdim. Ellipsoiden*, Math. Z. **19-** (1924), 300–307.

[W2] ———, *Gitterpunkte in mehrdim. Kugeln*, Warszawa, 1957.

[We] Weyl , H., *Über die Gleichverteilung der Zahlen mod-Eins*, Mathem. Ann. **77** (1915/16), 313–352.

[Y] Yarnold, J., *Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set*, Ann. Math. Stat. **43** (1972), 1566–1580.

Friedrich Götze
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld
Germany
goetze@mathematik.Uni-Bielefeld.DE

Applications of Intentionally Biased Bootstrap Methods

Peter Hall and Brett Presnell

Abstract. A class of weighted-bootstrap techniques, called biased-bootstrap methods, is proposed. It is motivated by the need to adjust more conventional, uniform-bootstrap methods in a surgical way, so as to alter some of their features while leaving others unchanged. Depending on the nature of the adjustment, the biased bootstrap can be used to reduce bias, or reduce variance, or render some characteristic equal to a predetermined quantity. More specifically, applications of bootstrap methods include hypothesis testing, variance stabilisation, both density estimation and nonparametric regression under constraints, 'robustification' of general statistical procedures, sensitivity analysis, generalised method of moments, shrinkage, and many more.

1991 Mathematics Subject Classification: Primary 62G09, Secondary 62G05
Keywords and Phrases: Bias reduction, empirical likelihood, hypothesis testing, local-linear smoothing, nonparametric curve estimation, variance stabilisation, weighted bootstrap

1. Uniform and weighted bootstrap methods

For centuries the sample mean has been recognised as an estimator of the population mean — or in contemporary notation, $\bar{X} = \int x \, d\widehat{F}(x)$ is an estimator of $\mu = \int x \, dF(x)$, where $\widehat{F}$ denotes the empirical distribution function computed using a sample drawn from a distribution F. The idea that the sample median is an estimator of the population median is implicit in work of Galton about 120 years ago. Thus, the notion that a parameter may be regarded as a functional of a distribution function, and estimated by the same functional of the standard empirical distribution, is a rather old one, even though it was perhaps only recognised as a general principle relatively recently.

Efron's (1979) classic paper on the bootstrap vaulted statistical science forward from these simple ideas. Efron saw that when substituting the true F by an estimator $\widehat{F}$, the notion of a 'parameter' could be interpreted much more widely than ever before. It could include endpoints of confidence intervals or critical points of hypothesis tests, as well as error rates of discrimination rules. It could encompass tuning parameters in a wide variety of estimation procedures (even the

nominal levels of intervals or tests can be regarded as tuning parameters), and much more.

Another key ingredient of the methods discussed by Efron (1979) was recognition that in cases where the functional of $\widehat{F}$ could not be computed directly, it could be approximated to arbitrary accuracy by Monte Carlo methods. This differed in important respects from several earlier approaches to 'resampling', as the idea of sampling from the sample has come to be known. In particular, neither Mahalanobis' notion of 'interpenetrating samples', nor Hartigan's (1969) 'subsampling' approach, directly involve drawing a resample *of the same size as the original sample* by sampling with replacement. The methods of Simon (1969, Chapters 23–25) are closer in this respect to the contemporary bootstrap.

The combination of these two ideas — the substitution or 'plug in $\widehat{F}$' rule, and the notion that Monte Carlo methods can be used to surmount computational obstacles — has been little short of revolutionary. When Monte Carlo simulation is employed to compute a standard bootstrap estimator, one samples independently and uniformly from a data set $\mathcal{X} = \{X_1, \ldots, X_n\}$, producing a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ with the property that

$$P(X_i^* = X_j | \mathcal{X}) = n^{-1}\,, \quad 1 \le i, j \le n\,. \tag{1.1}$$

Standard bootstrap methods may be loosely defined as techniques that approximate the relationship between the sample and the population by that between the resample $\mathcal{X}^*$ and the sample $\mathcal{X}$.

The generality of the standard uniform bootstrap may be increased in a number of ways, for example by allowing the resampled values X_i^* to be exchangeable, rather than simply independent, conditional on $\mathcal{X}$ (see e.g. Mason and Newton, 1992); or by retaining the independence but replacing the sampling weight n^{-1} at (1.1) by p_j, say. In the latter case we shall use a dagger instead of the familiar asterisk notation, so that there will be no ambiguity about the procedure we are discussing:

$$P\big(X_i^\dagger = X_j \big| \mathcal{X}\big) = p_j\,, \quad 1 \le i, j \le n\,, \tag{1.2}$$

where $\sum_j p_j = 1$. This 'weighted bootstrap' procedure has been discussed extensively (see e.g. Barbe and Bertail, 1995), usually as a theoretical generalisation of the uniform bootstrap, pointing to a multitude of different modes of behaviour that may be achieved through relatively minor modification of the basic resampling idea.

2. Biased bootstrap methods

In 'standard' settings, where the appropriate way of applying the bootstrap is relatively clear, the uniform bootstrap offers an unambiguous approach to inference. Therein lies part of its attraction — there are no tuning parameters to be selected, for example. However, the lack of ambiguity can also be a drawback. In particular, the rigidity of the conventional bootstrap algorithm makes it relatively difficult to modify uniform-bootstrap methods so as to include constraints on the parameter space. The weighted bootstrap offers a way around this difficulty, by providing an

opportunity for 'biasing' bootstrap estimators so as to fulfill constraints. Moreover, we may interpret the notion of a 'constraint' in a very broad sense, like that of a 'parameter'. Nevertheless, an unambiguous approach to choosing the weights p_i is required. Biased-bootstrap methods provide a solution to that problem.

The biased bootstrap requires two inputs from the experimenter: the distance measure, and the constraints. The first is generic to a wide range of problems, and will be discussed from that viewpoint in section 3. The second is problem-specific, and will be introduced through nine examples in section 4. A general form of the biased bootstrap is to choose the weights p_i so as to minimise distance from the distribution at (1.2) to that at (1.1), subject to the constraints being satisfied (Hall and Presnell, 1998a).

Details of some of the examples in section 4 may be found in Hall and Presnell (1998a,b,c) and Hall, Presnell and Turlach (1998). Examples not treated in section 4 include hypothesis testing, bagging (bootstrap aggregation), shrinkage, and applications involving time series data. The latter may be handled by either modelling the time series as a process with independent disturbances, and applying the biased bootstrap to those; or by using a biased form of the block bootstrap.

Section 5 will consider potential computational issues. Aids to computation include estimating equations, protected Newton-Raphson algorithms, and approximate, sequential linearisation. It will be clear that, using such techniques, biased-bootstrap methods are definitely computationally feasible.

3. Distance measures

For the sake of brevity we shall confine attention to a class of distance measures, the power divergence distances, introduced by Cressie and Read (1984) and Read and Cressie (1988). A wider range has been treated by Corcoran (1998) in the context of Bartlett adjustment of empirical likelihood. See also Baggerley (1998).

Let $p = (p_1, \ldots, p_n)$. For simplicity we assume throughout that $\sum_i p_i = 1$ and each $p_i \geq 0$, although in some cases (e.g. power divergence with index $\rho = 2$) the case where negative p_i's are allowed has computational advantages. Given $\rho \neq 0$ or 1, we may measure the distance between the uniform-bootstrap distribution, $p_{\rm unif} = (n^{-1}, \ldots, n^{-1})$, and the biased-bootstrap distribution (with weight p_i at data value X_i) by

$$D_\rho(p) = \{\rho\,(1-\rho)\}^{-1} \left\{ n - \sum_{i=1}^n (np_i)^\rho \right\}.$$

This quantity is always nonnegative, and vanishes only when $p = p_{\rm unif}$. For $\rho = \frac{1}{2}$, $D_\rho(p)$ is proportional to Hellinger distance. Letting $\rho \to 0$ we obtain

$$D_0(p) = -\sum_{i=1}^n \log(np_i)\,,$$

which equals half Owen's (1988) empirical log-likelihood ratio. Similarly, D_1 may be defined by a limiting argument; it is proportional to the Kullback–Leibler divergence between p and $p_{\rm unif}$ (whereas $D_0(p)$ is proportional to the Kullback–Leibler divergence between $p_{\rm unif}$ and p).

In constructing a biased-bootstrap estimator we would select a value of ρ, and then compute $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ from the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ so as to minimise $D_\rho(p)$, subject to the desired constraints being satisfied. If the parameter value that we wished to estimate was expressible as $\theta(F)$, then its biased-bootstrap estimator would equal $\theta(\widehat{F}_{\hat{p}})$, where $\widehat{F}_p$ denotes the distribution function of the discrete distribution that has mass p_i at data value X_i for $1 \le i \le n$. Usually the value of $\theta(\widehat{F}_{\hat{p}})$ will not be computable directly, but it may always be calculated by Monte Carlo methods, resampling from $\mathcal{X}$ according to the scheme that places weight $\hat{p}_i$ on X_i.

In some instances, for example outlier reduction (section 4.7), there are advantages to using $\rho \neq 0$, since $D_0(p)$ becomes infinite whenever some $p_i = 0$. By way of comparison, Hellinger distance (for example) allows one or more values of p_i to shrink to zero without imposing more than a finite penalty. However, in most other applications we have found that there is little to be gained — and sometimes, something to be lost (see sections 4.1 and 4.2) — by using a value of ρ other than $\rho = 0$.

4. Examples

4.1. Empirical likelihood. The method of empirical likelihood, or EL, was introduced by Owen (1988, 1990). See also Efron (1981). It may be viewed as a special case of the biased bootstrap in which the constraint is $\theta(\widehat{F}_p) = \theta_1$, where $\widehat{F}_p$ denotes the distribution function of the weighted bootstrap distribution with weights p_i, and θ_1 is a candidate value for θ. It is based on the value $\hat{p} = \hat{p}(\theta_1)$ of p that minimises $D_\rho(p)$ subject to $\theta(\widehat{F}_p) = \theta_1$.

One EL approach to constructing an α-level confidence interval for the true value of θ is to take t_α to be the upper α-level quantile of the chi-squared distribution for which the number of degrees of freedom equals the rank of the limiting covariance matrix of the uniform-bootstrap estimator, $\hat{\theta}(\widehat{F}_{p_{\rm unif}})$; and to let the interval be the set of θ_1's such that $D_\rho\{\theta(\widehat{F}_{\hat{p}(\theta_1)})\} \le t_\alpha$. Under regularity conditions that represent only a minor modification of those of Hall and La Scala (1990), this interval may be shown to have asymptotic coverage equal to $1 - \alpha$, no matter what the value of ρ. Using methods of DiCiccio, Hall and Romano (1991) it may be shown that this generalised form of EL is Bartlett-correctable if and only if $\rho = 0$. (Strictly speaking, the term 'likelihood' is appropriate for describing these generalised EL techniques only if $\rho = 0$.) See Baggerley (1998) and Corcoran (1998).

4.2. Variance stabilisation. Here we wish to choose, by empirical means, a transformation $\hat{g}$ which, when applied to a (scalar) parameter estimator $\hat{\theta}$, will implicitly correct for scale. Our method is a biased-bootstrap version of a conventional-bootstrap technique proposed by Tibshirani (1988). It has an advantage over the latter approach in that it does not require selection of any smoothing parameters, or any extrapolation.

As in example 4.1, choose p to minimise $D_\rho(p)$ subject to $\theta(\widehat{F}_p) = \theta_1$. Let $\mathcal{X}^\dagger = \{X_1^\dagger, \ldots, X_n^\dagger\}$ denote a resample drawn by sampling from $\mathcal{X}$ using the weighted bootstrap with weights $\hat{p}_i$, and let $\hat{\theta}^\dagger$ denote the version of $\hat{\theta}$ computed

from $\mathcal{X}^\dagger$ rather than $\mathcal{X}$. Let $\hat{v}(\theta_1) = \mathrm{var}(\hat{\theta}^\dagger|\mathcal{X})$ be the biased-bootstrap estimator of the variance of $\hat{\theta}$ when the true value of θ is θ_1. Write $\hat{g}(\theta)$ for the indefinite integral of $\hat{v}(\theta)^{-1/2}$, with the constant chosen arbitrarily. Using the uniform bootstrap, compute the conditional distribution of $\hat{g}(\hat{\theta}^*) - \hat{g}(\hat{\theta})$ and use it as an approximation to the unconditional distribution of $\hat{g}(\hat{\theta}) - \hat{g}(\theta^0)$, where θ^0 denotes the true parameter value. This enables us to compute confidence intervals for $\hat{g}(\theta^0)$, from which we may calculate intervals for θ^0 by back-transformation. It may be shown that $\rho = 0$ is sufficient for the latter intervals to be second-order accurate.

4.3. Density estimation under constraints. Here we consider kernel-type, biased-bootstrap estimators of the form $\hat{f}_p(x) = \sum_i p_i K_i(x)$, where $K_i(x) = h^{-1}K\{(x - X_i)/h\}$, K is a positive, symmetric kernel, and h is a bandwidth. (The traditional kernel estimator, in which each p_i is replaced by n^{-1}, may be regarded as a uniform-bootstrap estimator of $\theta = E\{K_i(x)\}$.) Constraining the j'th moment of the distribution with density $\hat{f}_p$ to equal the j'th sample moment is equivalent to asking that $\sum_i p_i A_i = a$, where a denotes the sample moment,

$$A_i = \sum_{k=0}^{\langle j/2\rangle} \binom{j}{2k} X_i^{j-2k}\, h^{2k}\, \kappa_{2k}\,,$$

$\langle j/2\rangle$ represents the integer part of $j/2$, and $\kappa_\ell = \int y^\ell\, K(y)\, dy$. Moreover, stipulating that the q'th quantile of the distribution with density $\hat{f}_p$ equal the q'th sample quantile ($\hat{\xi}_q$, say) produces a constraint of the same form, this time with $A_i = L\{(\hat{\xi}_q - X_i)/h\}$ (where L denotes the distribution function corresponding to the density K) and $a = q$. Constraining the interquartile range for $\hat{f}$ to equal its sample value amounts to the obvious linear form in constraints on the 25% and 75% quantiles. See also Chen (1997).

The constraint that entropy equals t, say, has the form

$$-\sum_{i=1}^n p_i \int K_i(x)\, \log\left\{\sum_{j=1}^n p_j\, K_j(x)\right\} dx = t\,.$$

Reducing entropy increases 'peakedness' and reduces spurious bumps in the tails. Combining this observation with the fact that increasing the bandwidth also tends to reduce the number of modes, while decreasing peakedness, we may develop an implicit algorithm (as distinct from the explicit method suggested in section 4.5) for computing a density estimator subject to the constraint of unimodality.

4.4. Correcting Nadaraya-Watson estimator for bias. Suppose data pairs (X_i, Y_i) are generated by the model $Y_i = g(X_i) + \epsilon_i$, where g is the smooth function that we wish to estimate, the design points X_i are random variables with density f, and the errors ϵ_i have zero mean. Then the Nadaraya–Watson estimator of g may be defined by $\tilde{g} = \widehat{\gamma}/\hat{f}$, where $\widehat{\gamma}(x) = n^{-1}\sum_i K_i(x)\, Y_i$ and $\hat{f}(x) = n^{-1}\sum_i K_i(x)$.

The performance of $\tilde{g}$ is generally inferior to that of local-linear estimators, owing to problems of bias. In particular, $\tilde{g}$ is biased for linear functions. To

overcome this difficulty we may use the biased bootstrap to constrain the estimator to be unbiased when g is linear, by insisting that $\sum_i p_i(x)\,(x - X_i)\,K_i(x) = 0$. Thus, $p = (p_1, \ldots, p_n)$ is now a function of location, x. The resulting estimator is

$$\hat{g}(x) = \left\{ \sum_{i=1}^n p_i(x)\,K_i(x)\,Y_i \right\} \Big/ \left\{ \sum_{i=1}^n p_i(x)\,K_i(x) \right\}.$$

It achieves the same minimax efficiency bounds as local-linear smoothing (see e.g. Fan, 1993), and enjoys positivity properties that the latter approach does not.

4.5. Unimodality and monotonicity. Define a continuous density f to be strongly unimodal if there exist points $-\infty < x_1 < x_2 < \infty$ such that (i) f is convex on $(-\infty, x_1)$ and on (x_2, ∞), and (ii) f is concave on (x_1, x_2). In principle we may constrain $\hat{f}_p$ to be a strongly unimodal density estimator, by arguing as follows: (a) for fixed x_1 and x_2, choose $p = p_{x_1 x_2}$ to minimise $D_\rho(p)$ subject to $\hat{f}_p''(x) = \sum_i p_i\,K_i''(x)$ being positive on $(-\infty, x_1)$ and on (x_2, ∞), and negative on (x_1, x_2); (b) choose x_1, x_2 to minimise $D_\rho(p_{x_1 x_2})$ over all possible choices satisfying (a). However, the probability that this is possible does not necessarily converge to 1 as $n \to \infty$, even if the true f is strongly unimodal and considerable latitude is allowed for choice of bandwidth.

On the other hand, a weaker form of unimodality may be successfully imposed. There, we argue as follows: (α) select a candidate $-\infty < x_0 < \infty$ for the mode of $\hat{f}_p$, and choose $p = p_{x_0}$ to minimise $D_\rho(p)$ subject to $\hat{f}'(x_0) = 0$, $\hat{f}''(x_0) \le 0$, and to any point $x \ne x_0$ for which $\hat{f}'(x) = 0$ being a point of inflexion of $\hat{f}_p$; and (β) choose x_0 to minimise $D_\rho(p_{x_0})$ over all possible choices satisfying (α). There is also a version of this method in the context of nonparametric regression, where 'unimodality' of a regression mean is defined in the obvious way.

Likewise, we may use biased-bootstrap methods to impose monotonicity of a function estimator in either the density or regression cases. Confining attention to local-linear estimators for nonparametric regression, we would proceed as follows. Let (X_i, Y_i), for $1 \le i \le n$, denote a sample of independent and identically distributed data pairs. If (X_i, Y_i) is accorded weight p_i then the local-linear estimator of $g(x) = E(Y|X = x)$ equals $\hat{a}$, where $(\hat{a}, \hat{b})$ denotes the pair (a, b) that minimises

$$\sum_{i=1}^n \{Y_i - a - b\,(X_i - x)\}^2\,p_i\,K_i(x)\,.$$

The biased-bootstrap local-linear estimator is $\hat{g}_p = (S_2 T_0 - S_1 T_1)/(S_2 S_0 - S_1^2)$, where

$$S_j(x) = \sum_{i=1}^n (X_i - x)^j\,p_i\,K_i(x)\,, \quad T_j(x) = \sum_{i=1}^n Y_i\,(X_i - x)^j\,p_i\,K_i(x)\,.$$

Suppose we wish to constrain $\hat{g}_p(x)$ to have derivative not less than a given value t, for all x in some interval $\mathcal{I}$. It may be shown that, if the true regression mean g satisfies $g' \ge t$ on $\mathcal{I}$ then, with probability tending to 1 as $n \to \infty$, and for a

wide range of choices of bandwidth, the biased-bootstrap constrained-minimisation problem has a solution, and that the solution has bias and variance of the same order as those of the unconstrained local-linear smoother.

4.6. Bias reduction without violating sign. Let $\hat\theta = \theta(\widehat F_{p_{\rm unif}})$ (possibly vector-valued) denote the uniform-bootstrap estimator of $\theta(F)$, based on data $\mathcal{X} = \{X_1, \ldots, X_n\}$. Suppose we wish to estimate $\psi_0 = \psi(\theta_0)$, where θ_0 is the true value of θ and ψ is a known smooth function. The uniform-bootstrap estimator is $\check\psi = \psi(\hat\theta)$, but is generally biased. The standard uniform-bootstrap bias-reduced estimator is $\widetilde\psi = 2\,\check\psi - E\{\psi(\hat\theta^*)|\mathcal{X}\}$, where $\hat\theta^*$ denotes the uniform-bootstrap version of $\hat\theta$. However, this approach does not necessarily respect the sign of the function ψ. For example, when $\psi(u) \equiv u^2$, and θ is a population mean and $\theta_0 = 0$, the probability that $\widetilde\psi < 0$ converges to 0.68.

A sign-respecting, biased-bootstrap approach to bias reduction may be defined as follows. Let $\hat\theta^\dagger$ denote the version of $\hat\theta$ computed from a resample drawn by sampling at random from $\mathcal{X}$ according to the weighted empirical distribution $\widehat F_p$. A biased-bootstrap approximation to the bias of $\psi(\hat\theta)$ is $\beta(p) = E_p\{\psi(\hat\theta^\dagger)|\mathcal{X}\} - \psi(\hat\theta)$, where E_p denotes expectation with respect to $\widehat F_p$. Choose $p = \hat p$ to minimise $D_\rho(p)$ subject to $\beta(p) = 0$, $\sum_i p_i = 1$ and each $p_i \ge 0$. Then, our biased-bootstrap, bias-reduced, sign-respecting estimator of ψ_0 is $\widehat\psi = \psi(\hat\theta_{\hat p})$, where $\hat\theta_p = \theta(\widehat F_p)$.

It may be shown that, not only does $\widehat\psi$ overcome the sign problem, in cases where the probability that $\widetilde\psi$ has the wrong sign does not converge to 0, $\widehat\psi$ is closer (on average) than $\widetilde\psi$ to ψ_0.

4.7. 'Trimming' or 'winsorising.' Let $\hat\theta_p = \theta(\widehat F_p)$ denote the biased-bootstrap estimator of $\theta = \theta(F)$, and let $\gamma(p, \mathcal{X})$ be a measure of the concentration of the biased-bootstrap distribution with respect to $\hat\theta_p$. For example, in the case of a scalar sample $\mathcal{X}$, and when our interest is in location estimation, we might define

$$\gamma(p, \mathcal{X}) = \sum_{i=1}^n p_i\,(X_i - \bar X_p)^{2k}\,,$$

where $k \ge 1$ is an integer and $\bar X_p = \bar X_p(k)$ minimises $\sum_i\ p_i(X_i - x)^{2k}$ with respect to x. (Taking $k = 1$ we see that $\gamma(p, \mathcal{X})$ is the variance of the biased-bootstrap distribution.) Put $\widehat\gamma = \gamma(p_{\rm unif}, \mathcal{X})$, being the version of the concentration measure in the case of the uniform bootstrap. Given $0 < t \le \widehat\gamma$ we may calibrate the level of concentration by choosing $p = p(t)$ to minimise $D_\rho(p)$ subject to $\gamma(p, \mathcal{X}) = t$. As t decreases, the biased-bootstrap distribution $\widehat F_{p(t)}$ becomes more concentrated.

To avoid the result of calibration being heavily influenced by tail weight of the sampling distribution, we suggest 'inverting' the calibration so that it is on $D_\rho(p)$ rather than $\gamma(p, \mathcal{X})$. That is, given $\xi > 0$ we propose choosing $t = t_\xi$ such that $D_\rho\{p(t)\} = \xi$, and defining $\hat p(\xi) = p(t_\xi)$. In order for this approach to be practicable we require $D_\rho\{p(t)\}$ to be a monotone increasing function of t, which can be verified in many cases.

With this modification it may be shown that, in the case $0 < \rho \le 1$, the biased bootstrap provides a remarkably effective device for reducing the effects of outlying

data values. For example, in the context of univariate location estimation the estimator has a smooth, redescending influence curve, and a breakdown point that may be be located at any desired value $\epsilon \in (0, \frac{1}{2})$ simply by 'trimming' to a known distance (depending only on ϵ) from the empirical distribution. The estimator has an affine-equivariant multivariate form, and has versions for regression and nonparametric regression.

4.8. Sensitivity analysis. The ideas suggested in section 4.7 may be used to develop new, empirical methods for describing influence and sensitivity. For example, one may vary t by an infinitesimal amount, starting at $t = \widehat{\gamma}$, and rank the data values X_i in decreasing order of the amount by which this variation produces a *decrease* in the respective values of p_i. This may be regarded as ranking data values in terms of their influence on concentration, according to the chosen concentration measure. It produces an outlier diagnostic.

An alternative approach is to apply the biased bootstrap with θ equal to a candidate value, θ_1 say, for the parameter, and consider the values of $(\partial/\partial\theta_1)\, p_i(\theta_1)$ evaluated at the uniform-bootstrap estimator $\hat{\theta} = \theta(\widehat{F}_{\rm unif})$. (Of course, the signs of the derivatives convey important information about the nature of sensitivity.) Still another approach is to examine leave-one-out empirical-likelihood ratios computed at biased-bootstrap estimators. These influence diagnostics have potential advantages over traditional techniques; for example, they may be applied to quite arbitrary estimators and parameters.

4.9. Generalised method of moments. The generalised method of moments, or GMM, can provide substantial improvements over the naive method of moments, by reducing the variance of estimators. Versions of the biased bootstrap have already been successfully applied to GMM; see for example Brown and Newey (1995) and Imbens, Johnson and Spady (1998). However, those applications require equations defining the estimators to be of full rank, and the methods can perform poorly when one or more of those equations is (approximately) redundant. Indeed, one may show by example that in such cases, the rate of convergence of GMM estimators can be as slow as $n^{-1/4}$ (where n is sample size), rather than the $n^{-1/2}$ achieved using a much simpler method without a weight matrix in the least-squares step; and that this rate is not improved by iterating GMM. Biased-bootstrap methods can be used to identify redundancy and accommodate it adaptively. The approach involves choosing the weight matrix to minimise a non-asymptotic estimator of mean squared error, and thereby calibrating the standard GMM method so as to obtain nearly-optimal performance. The biased bootstrap is employed to enforce an empirical version of the method-of-moments constraint when defining the mean squared error estimator.

5. Computational issues

By way of notation, let us say that a constraint on p is linear if it may be written in the form $\sum_i p_i A_i = a$, which we denote by (L), where A_i and a depend only on the data, not on p, and may be vectors. (If they were vectors of length ν then we would, in effect, be imposing ν separate linear constraints.) Examples of linear constraints include those encountered in in the context of constraining moments and quantiles

in section 4.3. Particularly for linear constraints, methods described by Owen (1990) and Qin and Lawless (1995), based on estimating equations, generally lead to numerically stable procedures.

It may be shown after a little algebra that under constraint (L), and when the distance function is D_ρ for some $\rho \neq 1$, the resulting p_i's are given by $p_i = p_i(\lambda) = (\lambda_0 + \lambda_1^T A_i)^{1/(1-\rho)}$, where λ_0 is a scalar, λ_1 is a column vector of length ν, and $\lambda = (\lambda_0, \lambda_1)$. (The λ_0 term comes from incorporating the additional condition $\sum_i p_i = 1$. We have not, at this stage, included the constraints $p_i \geq 0$, which in any event hold automatically when $-1 < \rho < 2$.) When $\rho = 1$ we have instead $p_i = \exp(\lambda_0 + \lambda_1^T A_i)$; and for any given ρ, the value of λ is defined by substituting back into (L). Thus, the dimension of the problem has been reduced from n to $\nu + 1$, which remains fixed as n increases. If in addition $\rho = 0$ then it may be shown that $\lambda_0 = n - \lambda_1^T a$, and so dimension reduces further, to ν.

In highly nonlinear problems, where these dimension reduction arguments do not apply, it may be necessary to compute the p_i's directly as the solution to an $(n-1)$-dimensional optimisation problem. For example, we have found that for moderate n a protected Newton-Raphson algorithm performs well in the problem of enforcing unimodality through constraints on entropy. Other approaches, such as the linearisation methods of Wood, Do and Broom (1996), may also be useful in nonlinear problems.

References

Baggerley, K. A. (1998). Empirical likelihood as a goodness of fit measure. *Biometrika*, to appear.

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap*. Springer, Berlin.

Brown, B. W. and Newey, W. K. (1995). Bootstrapping for GMM. Manuscript.

Chen, S. X. (1997). Empirical likelihood-based kernel density estimation. *Austral. J. Statist.* 39, 47–56.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, to appear.

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* 46, 440–464.

DiCiccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* 19, 1053–1061.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. (With Discussion.) *Canad. J. Statist.* 36, 369–401.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* 21, 196–216.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* 58 109–127.

Hall, P. and Presnell, B. (1998a). Intentionally-biased bootstrap methods. *J. Roy. Statist. Soc. Ser. B*, to appear.

Hall, P. and Presnell, B. (1998b). Density estimation under constraints. Manuscript.

Hall, P. and Presnell, B. (1998c). Biased bootstrap methods for reducing the effects of contamination. Manuscript.

Hall, P., Presnell, B. and Turlach, B. (1998). Reducing bias without prejudicing sign. Manuscript.

Hartigan, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* 64, 1303–1317.

Imbens, G. W., Johnson, P. and Spady, R. H. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–358.

Mason, D. M. and Newton, M. A. (1992). A rank statistic approach to the consistency of a general bootstrap. *Ann. Statist.* 20, 1611–1624.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* 18, 90–120.

Qin, J. and Lawless, J. (1995). Estimating equations, empirical likelihood and constraints on parameters. *Canad. J. Statist.* 23, 145–159.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, New York.

Simon, J. L. (1969). *Basic Research Methods in Social Science.* Random House, New York.

Tibshirani, R. (1988). Variance stabilization and the bootstrap. *Biometrika* 75, 433–444.

Wood, A. T. A., Do, K.-A., and Broom, B. M. (1996). Sequential linearization of empirical likelihood constraints with application to U-statistics. *J. Computat. Graph. Statist.* 5, 365–385.

Peter Hall
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200
Australia
peter.hall@anu.edu.au

Brett Presnell
Centre for Mathematics and its Applications
Australian National University
Canberra, ACT 0200
Australia
brett.presnell@anu.edu.au

Oracle Inequalities and Nonparametric Function Estimation

Iain M. Johnstone

Abstract. In non-parametric function estimation, partial prior information about the unknown function is often expressed by a family of models or estimators, among which a choice must be made. Oracle inequalities bound the mean squared error of a given estimator in terms of the (unknowable) best possible choice of model for the unknown function. This survey concentrates on three examples: the James Stein estimator, soft thresholding, and complexity penalized least squares and as illustrations, we describe some consequences for adaptive estimation.

1991 Mathematics Subject Classification: 62G07, 62G20, 62C20
Keywords and Phrases: adaptive estimation, complexity penalty, James-Stein estimator, minimax estimation, thresholding, unconditional basis, wavelet shrinkage

1 Introduction

Statistical theory aims in part to articulate when and why certain applied methods of data analysis succeed. With emergence of large, often instrumentally acquired datasets, recent decades have seen a focus on "nonparametric" models in which the number of model parameters grows with the size of available data. Here we focus on the estimation (or "recovery" or "denoising") of functions observed in additive noise and describe some relatively simple inequalities that encode information on the effect of sparse representation on the quality of estimation.

A common caricature is to posit observed data $y \in \mathcal{R}^n$ with structure $y = \mu + \epsilon z$. Here μ is an unknown function which one desires to "estimate" or "recover", and $z \in \mathcal{R}^n$ is a vector of standard Gaussian noise of known scale ϵ. When expressed in terms of coefficients in an orthonormal basis $\{\psi_i, I \in \mathcal{I}_n\}$, the model becomes

$$y_I = \mu_I + \epsilon z_I \qquad I \in \mathcal{I}_n. \tag{1}$$

Here z_I are independent Gaussian noises of mean zero and variance one. This sequence form of the "Gaussian white noise model", whether finite as here, or infinite, as in Section 1.1 below, is the conceptually and technically simplest model of nonparametric estimation. Extensions to correlated noise and indirect data $y = K\mu + \epsilon z$ are possible, but not covered here.

EXAMPLES 1. (a) The *equispaced, fixed design* in which $y_I = f(t_I) + \sigma z_I$, with f an unknown function defined on $[0,1]$ and $t_I = I/n, I = 1, \dots n$.

(b) An initial segment of the *continuous Gaussian white noise* model. Suppose that W_t is a standard Brownian motion (or sheet) and that one observes $dY_t = f(t)dt + \epsilon dW_t$ for $t \in D$, a compact set in $\mathbb{R}^d$. If $d = 1$ and $D = [0,1]$, this may be interpreted as $Y_t = \int_0^t f(s)ds + \epsilon W_t$ for $0 \le t \le 1$. Take inner products with elements $\{\psi_I, I \in \mathcal{I}\}$ of a complete orthonormal basis for $L_2[0,1]$ and set $y_I = \langle \psi_I, dY \rangle$, $\theta_I = \langle \psi_I, f \rangle$ and $z_I = \langle \psi_I, dW \rangle$. This gives an infinite sequence version of (1), to be used in Section 1.1 below. To recover precisely (1), consider an initial segment of cardinality n of the index set $\mathcal{I}$. A discrete orthogonal wavelet transform of model (i) yields an approximation to this initial segment (after calibrating $\epsilon = \sigma n^{-1/2}$, cf. [10]).

(c) *Redundant regression.* Suppose that there are given vectors (or signals) $x_1, \dots x_p \in \mathbb{R}^n$, and that it is thought useful, for reaons of parsimony, interpretability or otherwise, to represent μ in terms a few of the x_i. Collecting x_i as columns of a "design matrix" $X = [x_1 \cdots x_n]$, one obtains the standard, homoscedastic Gaussian linear regression model $y = X\beta + \epsilon z$. In traditional parametric regression analysis, it is supposed that $p < n$ and that $\mu \in \text{span}\{x_i\}$. However, we specifically consider two "non-parametric" cases: a) $p = n$ and x_i orthogonal (i.e. equivalent to (1)), and b) $p > n$ and *not* orthogonal - here the x_i might be a class of basic signals from a (possibly highly) redundant *dictionary* $\mathcal{D}$ and we seek a parsimonious representation of μ in terms of as few elements of $\mathcal{D}$ as possible.

ASSESSING ERROR. An estimator $\hat{\mu} = \hat{\mu}(y)$ is a function of observed data y: we wish to quantify and compare the quality of estimation as $\hat{\mu}$ varies. Simplest to work with is mean squared error (MSE):

$$r_\epsilon(\hat{\mu}, \mu) = E_\mu |\hat{\mu} - \mu|^2 = \int |\hat{\mu} - \mu|^2 \phi_\epsilon(y - \mu) dy. \tag{2}$$

Here $\phi_\epsilon(z)$ denotes the probability density function of ϵz. The notation $r_\epsilon(\hat{\mu}, \mu)$, mnemonic for "risk", hints at the possible and frequently desirable use of more general error norms $\|\hat{\mu} - \mu\|$ or loss functions $L(\hat{\mu}, \mu)$.

The error $\hat{\mu} - \mu$ is usually decomposed into a zero-mean *stochastic* component $\hat{\mu} - E_\mu \hat{\mu}$ and a deterministic component, the *bias*, $E_\mu \hat{\mu} - \mu$. For quadratic error measures, these components are uncorrelated, so that the MSE is the sum of variance and squared bias terms. In particular, for a *linear* estimator $\hat{\mu}_L(y) = Ly$,

$$r(\hat{\mu}_L, \mu) = \epsilon^2 \text{tr } LL^t + |L\mu - \mu|^2. \tag{3}$$

The quality of approximation of μ by the operator L is thus balanced against the complexity of L, as measured by the variance term, which for example becomes $\epsilon^2 m$ in the case of orthogonal projection onto a subspace of dimension m. Already visible here is the important role that approximation theory plays in analysing the deterministic component of error. For non-linear estimators that, implicitly or explicitly, involve a choice among linear estimators, the analysis of the stochastic term is facilitated by the concentration of measure phenomenon (Section 4).

MODELS AND ESTIMATORS. A model is a subset M of the full parameter space $\mathbb{R}^n$. A family of models $\{M_\alpha, \alpha \in \mathcal{A}\}$ is one device commonly used to represent imperfect and partial information about the unknown μ. Often there is a natural estimator $\hat{\mu}_\alpha$ associated with each model and in this paper we simplistically conflate choice of model with choice of the associated estimator.

EXAMPLES 2. *(a) Spheres and linear shrinkage.* For positive α, let M_α be the sphere $|\mu| = \alpha$: this might correspond to prior information about the signal-to-noise ratio. Natural corresponding estimators are given by *linear shrinkage*: $\hat{\mu}_\alpha = \gamma y$ where $\gamma = \gamma(\alpha)$ is obtained by minimizing the MSE in (3), namely $n\gamma^2 + (1-\gamma)^2|\mu|^2$, on M_α to obtain the Wiener filter $\gamma(\alpha) = \alpha^2/(n+\alpha^2) \in (0,1)$.

(b) Subspaces and projections. In the regression setting of Example 1(c) above, to each subset $J \subset \{1, \ldots, p\}$ of the full variable list is associated a linear model $M_J = \text{span}\,\{x_j, j \in J\}$. The corresponding estimators are orthogonal projections P_J on M_J: these are the least squares estimators on the assumption that $\mu \in M_J$.

IDEAL RISK Given a family $\mathcal{A}$ of models (or corresponding estimators), and for a given unknown μ, the best attainable MSE is given by the *ideal risk*

$$\mathcal{R}_\epsilon(\mu, \mathcal{A}) = \inf_\alpha R(\hat{\mu}_\alpha, \mu).$$

Thus, in example (a), the ideal linear shrinkage risk is

$$\mathcal{R}_\epsilon(\mu, LS) = n\epsilon^2|\mu|^2/(n\epsilon^2 + |\mu^2|). \tag{4}$$

OUTLINE OF PAPER. Of course, μ is not known, and without access to an oracle who divulges the best α, the ideal risk is not attainable by an estimator depending on the data y alone. Nevertheless, it acts as a useful benchmark, and we seek estimators that in an appropriate sense optimally mimick the ideal risk. Such estimators turn out to be non-linear, and in particular, not members of the family $\hat{\mu}_\alpha$. For three settings and estimators, oracle inequalities are presented in Theorems 3, 5 and 8 – we emphasize that the inequalities are non-asymptotic and uniform in character, holding for all n, ϵ and for all $\mu \in \mathbb{R}^n$.

Corresponding lower bounds (although asymptotic in n) show that without some restriction on, or further information about μ, the inequalities cannot be improved, and thus represent in some sense the necessary "price" for searching over a class of models/estimators of a given size.

Oracle inequalities are neither the beginning nor the end of a theory, but when available, are informative tools. For example, Theorems 3, 5 and 8 may also be used to derive asymptotic (i.e. low noise ϵ) results within a framework of adaptive minimax estimation: this class of applications is considered in a connected sequence of "illustrations" in the continuous Gaussian white noise model, which we now introduce.

1.1 ILLUSTRATION: ASYMPTOTIC MINIMAX ESTIMATION.

The continuous Gaussian white noise model is that of Example 1(b). Because of Parseval's inequality $\int_0^1 (\hat{f} - f)^2 = \sum_I (\hat{\theta}_I - \theta_I)^2 = \|\hat{\theta} - \theta\|^2$, estimation error

can equally well be measured in the sequence domain. To evaluate estimators, we use the minimax principle - although inherently conservative and not universally accepted, we find that it leads to clear structures and informative results. Thus, estimators are assessed by their worst case risk over a given Θ. The *minimax risk* measures the best attainable such maximum risk, within a class $\mathcal{E}$ of estimators: $R_{\mathcal{E}}(\Theta,\epsilon) = \inf_{\hat\theta\in\mathcal{E}} \sup_{\theta\in\Theta} r_\epsilon(\hat\theta,\theta)$. The symbols $\mathcal{E} = N, L, D, \ldots$ refer to specific estimator classes: all non-linear, all linear, all threshold rules etc. Finally, estimator $\hat\theta$ is called *asymptotically* $\mathcal{E}-$ *minimax* if

$$\sup_{\theta\in\Theta} r_\epsilon(\hat\theta,\theta) = R_{\mathcal{E}}(\Theta,\epsilon)(1+o(1)), \qquad \epsilon\to 0.$$

In order to describe a flexible and scientifically meaningful class of parameter spaces Θ, we employ a *dyadic sequence* notation, in which $I = (j,k)$, with $j = 0, 1, \ldots$ and $k = 1, \ldots, 2^j$. The primary motivation comes from orthonormal bases of wavelets $\{\psi_{jk}\}$, which, under suitable regularity and decay conditions on the wavelets, and with suitable modifications to handle intervals, form unconditional bases for many function spaces of interest ([22, 15, 3]). Thus their norms may be characterized in terms of conditions on $|\theta_I|$. For example, let χ_I denote the indicator function of the interval $[(k-1)2^{-j}, k2^{-j}]$: the sequence of (quasi-)norms $\|\cdot\|_{\alpha,p}$, defined for $0<\alpha<\infty, 0<p\le\infty$ by

$$\|\theta\|_{\alpha,p}^p = \int_0^1 [\sum_I (2^{aj}|\theta_I|\chi_I)^2]^{p/2}, \qquad a = \alpha + 1/2,$$

are equivalent, (for $p>1$ and $\alpha\in\mathbb{N}$) to the traditional Sobolev norms $\|f\|_{W_p^\alpha}^p = \int_0^1 |f^{(\alpha)}|^p + |f|^p$. In the Hilbertian case $p=2$, these take the simpler form

$$\|\theta\|_{\alpha,2}^2 = \sum_{j\ge 0} 2^{j\alpha}|\theta_j|^2, \qquad |\theta_j|^2 = \sum_k |\theta_{jk}|^2.$$

As parameter spaces, we thus use norm balls: $\Theta_{\alpha,p}(C) = \{(\theta_I) : \|\theta\|_{\alpha,p} \le C\}$, which are analogs of size restrictions on derivatives, but measured in L_p norms.

In practice, the values of (α,p,C) will not be known, and rather than seeking a minimax estimator for a single such $\Theta_{\alpha,p}(C)$, we look for estimates with an adaptive minimaxity property. Thus, suppose that a *scale* of spaces $\mathcal{S} = \{\Theta_\nu(C) : \nu\in\mathcal{V}, C>0\}$ is given, where ν is an order parameter, such as (α,p) above, and C a scale parameter. Then $\hat\theta$ is *adaptively* $\mathcal{E}-$*minimax* if (i) the definition of $\hat\theta$ is independent of (ν,C), and (ii) $\hat\theta$ is asymptotically $\mathcal{E}-$minimax *for all* (ν,C).

2 Linear Shrinkage and Orthogonal Invariance

A celebrated result in parametric statistics, due to Stein [24], is the inadmissibility of the maximum likelihood estimator $\hat\mu^0(y) = y$ in model (1) as soon as $n\ge 3$. Indeed, [17] showed that adaptive linear shrinkage

$$\hat\mu^{JS+}(y) = (1-\hat\gamma)_+ y, \qquad \hat\gamma = (n-2)\epsilon^2/|y|^2,$$

is everywhere better than $\hat{\mu}^0$, in the sense that for all $\mu \in \mathbb{R}^n$, $r_\epsilon(\hat{\mu}^{JS+}, \mu) < r_\epsilon(\hat{\mu}^0, \mu) \equiv n\epsilon^2$. Here $a_+ = \max(a, 0)$. The result was and remains surprising because it can seem counterintuitive that combining data from statistically completely independent problems, represented by each coordinate in (1), leads to better MSE properties.

A simple proof was later given by Stein [25], using his unbiased estimate of risk to show that $\hat{\mu}^{JS}(y) = (1 - \hat{\gamma})y$, necessarily worse than $\hat{\mu}^{JS+}$, satisfies

$$r(\hat{\mu}^{JS}, \mu) = E_\mu\{n - (n-2)^2|y|^{-2}\} < n. \tag{5}$$

(where, for simplicity, $\epsilon = 1$ here.) Using in (5) the fact that the distribution of $|y|^2$ can be represented as the mixture of central chi-squared distributions χ^2_{n+2P} with P distributed as a Poisson variate with mean $|\mu|^2/2$, and applying Jensen's inequality, one obtains our first oracle inequality.

THEOREM 3 ([7]). *In model* (1)*, suppose* $n \geq 3$. *For all* $\mu \in \mathbb{R}^n$,

$$E|\hat{\mu}^{JS} - \mu|^2 \leq 2\epsilon^2 + \frac{(n-2)\epsilon^2|\mu|^2}{(n-2)\epsilon^2 + |\mu|^2}. \tag{6}$$

In view of (4)*, this implies*

$$r_\epsilon(\hat{\mu}^{JS+}, \mu) \leq 2\epsilon^2 + \mathcal{R}_\epsilon(\mu, LS). \tag{7}$$

Thus, the classical James-Stein estimator comes within an additive penalty of $2\epsilon^2$ of mimicking the ideal linear shrinkage estimator. This performance is impressive when calibrated against the minimax risk $R_N(\mathbb{R}^n, \epsilon)$, in this problem $n\epsilon^2$.

However it should be noted that this inequality is orthogonally invariant, and makes no use of the particular basis in which the unknown signal μ is represented.

2.1 ILLUSTRATION: LEVELWISE SHRINKAGE IN THE DYADIC SEQUENCE MODEL.

In the dyadic sequence model of Section 1.1, group coefficients by level j : $y_j = (y_{jk})_{k=1}^{2^j}$. Form a *levelwise* James Stein estimator $\hat{\theta}^{LJS}$ by applying James-Stein shrinkage to y_j: $\hat{\theta}^{LJS}_j = \hat{\theta}^{JS+}(y_j)$, at least for levels j below a cutoff $J = \log_2 \epsilon^{-2}$, above which $\hat{\theta}^{LJS}_j$ simply estimates zero. [Recall the calibration $n = \epsilon^{-2}$ of Example 1(b).] The MSE of the $\hat{\theta}^{LJS}$ may then also be represented levelwise:

$$E\|\hat{\theta}^{LJS} - \theta\|^2 = \sum_{j<J(\epsilon)} E|\hat{\theta}^{JS+}(y_j) - \theta_j|^2 + \sum_{j\geq J(\epsilon)} |\theta_j|^2.$$

The oracle inequality (7) may be applied to each level j in the first sum, while the geometric weights 2^{aj} used to define $\Theta_{\alpha,2}$ imply that the second sum is negligible for small ϵ. For the scale of *Hilbert* spaces $\mathcal{S}_2 = \{\Theta_{\alpha,2}(C) : \alpha > 0, C > 0\}$:

THEOREM 4 ([7]). $\hat{\theta}^{LJS}$ *is adaptively minimax over* $\mathcal{S}_2$.

This recovers and extends a notable result of Efroimovich & Pinsker [14], originally formulated in the Fourier basis. In fact, one verifies relatively easily that $\hat{\theta}$ is adaptively minimax among linear estimates (from the ideal linear shrinkage risk) and then appeals to the celebrated theorem of Pinsker [23]), which shows that for the *ellipsoids* occurring in $\mathcal{S}_2$, linear minimax rules are actually asymptotically minimax among all non-linear estimates.

This levelwise application of an oracle inequality is shows how the dyadic sequence model allows a "lifting" of results from a symmetric and "parametric" setting (an exchangeable multivariate normal law at each level) to a non-parametric, infinite-dimensional model. Other examples of this type may be found in [7, 9].

3 Orthogonal Regression and Thresholding

To this point, we have considered only orthogonally invariant estimators. However, a basic principle is that sparsity of representation of a signal in a given basis leads to better estimation, and to exploit such sparsity, non-linear estimators are needed.

Thus, assume the orthonormal basis leading to coefficients (1) is chosen so that $\{\mu_i\}$ contains few large coefficients, although of course it is not known in advance *which* among the co-ordinates are important.

In this orthogonal regression setting, the least squares subset selection estimators have a simple co-ordinatewise representation: the $j-$th component of $\hat{\mu}_J(y)$ equals y_j if $j \in J$ and 0 otherwise. Thus, the least squares estimators have the form of diagonal projections (DP below). The mean squared error of $\hat{\mu}_J$ is then the sum of terms which measure either variance or bias:

$$r(\hat{\mu}_J, \mu) = \sum_{j \in J} \epsilon^2 + \sum_{j \notin J} \mu_j^2.$$

The ideal risk for among all such diagonal projection estimators can therefore be found by minimizing termwise:

$$\mathcal{R}_\epsilon(\mu, DP) = \inf_J r(\hat{\mu}_J, \mu) = \sum_j \mu_j^2 \wedge \epsilon^2.$$

To quantify sparsity, order the squared magnitudes of the components of μ via $\mu_{(1)}^2 \geq \mu_{(2)}^2 \geq \ldots \geq \mu_{(n)}^2$ and define *compression numbers* $c_j^2 = \sum_{k>j} \mu_{(k)}^2$. The number of large coefficients is measured by $N(\epsilon) = \#\{j : |\mu_j| > \epsilon\}$, and we have

$$\mathcal{R}_\epsilon(\mu, DP) = \epsilon^2 N(\epsilon) + c_{N(\epsilon)}^2,$$

which shows an intimate connection between ideal risk and the compressibility of the signal in this basis.

Various forms of thresholding estimator can be introduced: here we consider soft thresholding:

$$\hat{\mu}_j^{ST}(y) = \mathrm{sgn}(y_j)(|y_j| - \lambda)_+.$$

The key points are that the estimator acts co-ordinatewise and that there is a threshold zone $[-\lambda, \lambda]$ in which the data is interpreted as noise and "discarded".

THEOREM 5 ([6]). *If $\lambda = \sqrt{2\log n}$, then for all $\mu \in \mathbb{R}^n$,*

$$r_\epsilon(\hat{\mu}^{ST}, \mu) \le (2\log n + 1)[\epsilon^2 + \mathcal{R}_\epsilon(\mu, DP)]. \tag{8}$$

Since the logarithmic penalty is of small order relative to n, the result shows that sparsity, as measured by ideal risk, implies good estimation. The bound is valid for all sample sizes and all μ. There has been much work on the choice of threshold λ - the choice given here is attractive for its conservatism: since for independent and identically distributed $N(0,1)$ variates z_i, $P(\max_{1\le j\le n}|z_j| > \sqrt{2\log n}) \to 0$, it follows that $P(\hat{\mu}^{ST} = 0|\mu = 0) \to 1$. For more on these issues and numerical examples, see [11]. Smaller choices of λ, even depending on the data y, lead to better mean squared error in exchange for less conservatism [7]. Natural extensions of Theorems 5 and 6 to correlated noise exist [19]

OPTIMALITY. Absent extra restrictions on μ, the factor $2\log n$ is optimal:

THEOREM 6 ([6]). *As $n \to \infty$,*

$$\inf_{\hat{\mu}} \sup_{\mu\in\mathbb{R}^n} \frac{r(\hat{\mu}, \mu)}{\epsilon^2 + \mathcal{R}_\epsilon(\mu, DP)} \ge (2\log n)(1 + o(1)).$$

The lower bound arises from the difficulty of distinguishing rare true signal components from the also infrequent extremes of the white Gaussian noise z_i. Indeed, suppose $\epsilon = 1$ and that the values μ_i are drawn independently from a two point prior distribution with masses of probability $1 - \delta_n$ at 0 and δ_n at $\bar{\mu}_n$. Choosing $\delta_n = \log n/n$ and $\bar{\mu}_n \sim (2\log \delta_n^{-1})^{1/2}$, it turns out that the posterior distribution of μ_i, having observed even a value of $y_i > \bar{\mu}_n$, is still concentrated on $0 : P(\mu = 0|y = \bar{\mu}_n + z) \approx 1$, for z large and fixed, as $n \to \infty$. Hence, with probability δ_n, the estimator is forced to make an error of order $\bar{\mu}_n^2 \sim 2\log n$.

3.1 ILLUSTRATION: THRESHOLDING IN THE DYADIC SEQUENCE MODEL.

Return to the dyadic sequence model, and apply soft thresholding at $\lambda = \epsilon\sqrt{2\log \epsilon^{-2}}$ to the first $n = \epsilon^{-2}$ coefficients. In other words, $\hat{\theta}_I^T(y) = \eta_{ST}(y_I, \lambda)$ for all I with $j < J(\epsilon)$. Applying the thresholding oracle inequality (8) to the first n co-ordinates,

$$r_\epsilon(\hat{\theta}^T, \theta) \le c \cdot \log \epsilon^{-2} \cdot [\epsilon^2 + \mathcal{R}_\epsilon(\theta, DP)] + \sum_{j\ge J(\epsilon)} |\theta_j|^2 \tag{9}$$

In contrast with the scale $\mathcal{S}_2$ of Section 2.1, consider now a broader scale of Sobolev-type parameter spaces: $\mathcal{S} = \{\Theta_{\alpha,p}(C) : \alpha > 1/p - 1/2, p > 0, C > 0\}$. For such spaces there is a bound relating ideal to minimax risk. First, the geometric weights in the definition imply ([12]) that for $\Theta = \Theta_{\alpha,p}(C)$ and on setting $r = 2\alpha/(2\alpha+1)$,

$$\mathcal{R}_\epsilon(\Theta, DP) := \sup_\Theta \mathcal{R}_\epsilon(\theta, DP) = \sup_{\theta\in\Theta} \sum \theta_I^2 \wedge \epsilon^2 \le c_\alpha C^{2(1-r)}\epsilon^{2r}.$$

Second, the minimax risk over Θ is minorized by that over any inscribed hypercube of dimension m and side length $\epsilon : R_N(\Theta, \epsilon) \ge c_0 m\epsilon^2$. Optimizing over the

dimension m and combining with the previous display, we obtain the basic *ideal to minimax* risk inequality:

$$\mathcal{R}_\epsilon(\Theta, DP) \le c_\alpha C^{2(1-r)}\epsilon^{2r} \le c'_\alpha R_N(\Theta, \epsilon). \tag{10}$$

In combination with the oracle inequality and negligibility of the tail sum in (9), this yields an adaptive *near*-minimaxity property for thresholding:

THEOREM 7. *For all* $\Theta_{\alpha,p}(C) \in \mathcal{S}$,

$$\sup_{\Theta_{\alpha,p}(C)} r_\epsilon(\hat{\theta}^T, \theta) \le c_{\alpha,p} \cdot \log \epsilon^{-2} \cdot R_N(\Theta_{\alpha,p}(C), \epsilon).$$

The term near-minimaxity refers to the logarithmic term in the upper bound, which is negligible with respect to the algebraic rate ϵ^{2r}. In fact, this logarithmic term can also be removed by a lower, data-dependent choice of threshold [7, 18].

Important here is that in contrast to the linear adaptivity of Theorem 4, this result applies for all $p > 0$, and in particular for $p < 2$. These latter spaces contain spatially inhomogeneous functions with localized discontinuities or other singularities. The ability of an estimator to adapt to such functions is in practice more important than the attractive, but limited adaptation of the levelwise James-Stein estimator, and its cousins, the spatially homogeneous kernel methods, even with bandwidth selected from data. This is discussed further in [11].

4 Redundant Dictionaries & Complexity Penalized Model Selection

In seeking a sparse representation for a signal, one may build build rich dictionaries $\mathcal{D} = \{x_1, \ldots, x_p\}$ in various ways: for example by combining many orthonormal bases (as in libraries of wavelet and cosine packets, [4]), or by considering redundant discretizations of continuously parametrized families, or by allowing products (interactions) of many simple elements, such as B-splines with knots at individual data locations (e.g. [16]). In all these cases, the dictionary size p greatly exceeds that data size n, and estimation methods will have to allow for the effects of searching over such a vast domain (in principle, 2^p models).

Recalling Examples 1(c) and 2(b), the data may be represented in the form $y = X\beta + \epsilon z$, where we now assume that $\text{span}(X) = \mathbb{R}^n$. Thus, the models of interest correspond to subsets $J \subset \{1, \ldots, p\}$, $M_J = \text{span}\,\{x_j : j \in J\}$, and $\hat{\mu}_J = P_J y$, orthogonal projection on M_J. The risk of individual projection estimators is given by (3), so the ideal risk of subset selection from dictionary $\mathcal{D}$ becomes

$$\mathcal{R}_\epsilon(\mu, SS(\mathcal{D})) = \min_J r_\epsilon(\hat{\mu}_J, \mu) = \min_J |\mu - P_J\mu|^2 + \epsilon^2 \text{rank}(P_J).$$

To obtain an estimator that mimicks ideal risk, we use the penalized least squares principle. This balances the fit of the estimate, which in the absence of any penalty could be made arbitrarily close to the data, against some measure of complexity of the estimate:

$$\hat{\mu}_P = \text{argmin}_{\tilde{\mu}} |y - \tilde{\mu}|^2 + \epsilon^2 P(\tilde{\mu}).$$

In the orthonormal basis setting, $\hat{\mu}_P$ can be evaluated explicitly when the penalty P has an additive form: for example, $P(\mu) = c\sum \mu_i^2$ implies linear shrinkage, $P(\mu) = 2\lambda \sum |\mu_i|$ implies soft thresholding, and $P(\mu) = \lambda^2 \sum I\{\mu_i \neq 0\}$ implies $\hat{\mu}_{P,i}(y) = y_i I\{|y_i| > \lambda\}$, or hard thresholding. For the *redundant linear model* $y = X\beta + \epsilon z$, we generalize the third case by setting

$$P(\mu) = \lambda^2 N(\mu), \qquad N(\mu) = \min\{|J| : \mu = \sum_{j\in J} \beta_j x_j\}.$$

The resulting penalized least squares estimator may expressed in terms of the residual sums of squares $RSS_J = |y - \hat{\mu}_J|^2$ of the possible models:

$$\min_{\tilde{\mu}} |y - \tilde{\mu}|^2 + \lambda^2\epsilon^2 N(\tilde{\mu}) = \min_J RSS_J + \lambda^2\epsilon^2 \mathrm{rank}(P_J).$$

Hence we call this the *Complexity Penalized Residual Sum of Squares* (CPRSS) estimate. Certain choices of the factor λ lead to well known estimators: $\lambda^2 = 2$ (AIC), $\log p$ (BIC), $2\log n$ (RIC) (For details and references see [8]).

THEOREM 8 ([8]). *Let $\zeta > 1, \beta > 0$ and $\lambda = \lambda_p = \zeta[1 + \sqrt{2(1+\beta)\log(p+1)}]$. Then for all $n, p \geq n$,and $\mu \in \mathbb{R}^n$,*

$$r_\epsilon(\hat{\mu}_{CPRSS}, \mu) \leq L_p[(2+\gamma_p)\epsilon^2 + \mathcal{R}_\epsilon(\mu, SS(\mathcal{D}))], \tag{11}$$

where $L_p = (1-\zeta^{-1})^{-1}\lambda_p^2$, and $\gamma_p = \gamma(p,\beta) \to 0$ as $p \to \infty$.

The penalty factor λ_p^2 is slightly larger than $2\log p$, where p is the cardinality of the dictionary. We emphasize that the result holds for all μ, n and $p \geq n$, and in particular the inequality depends only on p, not n! Building on the remarkable [1], Birgé & Massart are conducting a thorough study of penalties $P(\mu)$ for which such oracle inequalities and improvements hold. While the constant L_p in (11) is certainly not optimal, there is a lower bound similar to Theorem 6:

THEOREM 9 ([8]). *For each fixed $r \in \mathbb{N}$, there exists a sequence of dictionaries $\mathcal{D}_n$ with $p(n) = |\mathcal{D}_n| \asymp n^r$ such that as $n \to \infty$,*

$$\inf_{\hat{\mu}} \sup_{\mu\in\mathbb{R}^n} \frac{E|\hat{\mu} - \mu|^2}{\epsilon^2 + \mathcal{R}_\epsilon(\mu, SS(\mathcal{D}))} \geq [2\log p(n)](1 + o(1)).$$

ROLE OF CONCENTRATION INEQUALITIES. The stochastic part of the proof of Theorem 8 depends on an early example (due to Cirelson-Ibragimov-Sudakov [2, 21]) of what are now in probability called concentration (or deviation) inequalities. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz with $\|f\|_{Lip} = L$. If $Z \sim N_n(0, I)$, then

$$P\{f(Z) \geq Ef(Z) + t\} \leq \exp\{-t^2/2L^2\}.$$

The key points are the Gaussian tail behaviour of $f(Z)$ and the fact that it does not depend on dimension n - hence the dimension-free aspect of Theorem 8. This inequality can then be applied to all projections onto model subsets of cardinality $|J| = \ell$, and then summed over ℓ. Thus, since $f(z) = \|P_J z\|$ has $\|f\|_{Lip} = 1$, and since $Ef(Z) \leq \sqrt{\ell}$, we have, on setting $t = \sqrt{2\ell(1+\beta)\log p}$,

$$P\{\sup_{|J|=\ell} \|P_J z\| \geq \sqrt{\ell} + t\} \leq \binom{p}{\ell} p^{-\ell(1+\beta)} \leq \frac{1}{p^{\ell\beta}\ell!}.$$

4.1 Illustration: Minimaxity for non-standard function classes

The penalized least squares formalism can be applied in situations where no unconditional basis exists. To give a simple example, consider again the model $dY_t = f dt + \epsilon dW_t$, where now $t \in [0,1]^2$, and the *horizon model* for edges in images, studied earlier by, for example, Korostelev and Tsybakov [20]. It is supposed that f takes only the values 0 and 1, and further that the boundary is such that $f(t_1,t_2) = I\{t_2 \le \theta(t_1)\}$. The boundary, or *horizon*, is supposed to be Hölder continuous: more specifically, we say that $f \in \text{HÖLDER}_s(B)$ if $\|\theta\|_\infty + \|\theta^{(r)}\|_\beta \le B$, where $r \in \mathbb{N}$, $\beta = s - r \in (0,1]$ and $\|g\|_\beta = \sup |g(t) - g(t')|/|t - t'|^\beta$. ...]

Dictionaries and minimax risk. While $\mathcal{D}$ is often conceptually infinite, in practice one must work with a family of finite subdictionaries $\mathcal{D}_\epsilon$ with cardinality $m(\epsilon)$ being at most a polynomial function of ϵ^{-2} : $m(\epsilon) \le \beta_1 \epsilon^{-2\beta_2}$. [8] defines a notion of *universal dictionary* for a scale $\mathcal{S} = \{\mathcal{F}_\nu(C)\} of functionclasses$, which has as consequence the same type of *ideal to minimax risk inequality* as used in the orthobasis case (compare (10)): for all $\mathcal{F}_\nu(C) \subset \mathcal{S}$ and $\epsilon < \epsilon(\nu, C)$, there exists $r = r(\nu)$ such that

$$\mathcal{R}_\epsilon(\mathcal{F}_\nu(C), \mathcal{D}_\epsilon) \le K_\nu C^{2(1-r)} \epsilon^{2r} \le K'_\nu R_N(\mathcal{F}_\nu(C), \epsilon).$$

This may then be combined with the oracle inequality of Theorem 8 to obtain adaptive near-minimaxity.

Thus, in the horizon example, we start with a continuum *trapezoid* dictionary, parametrized by $\gamma = (a,b,c,d)$, representing a function taking value 1 on the trapezoid in $[0,1]^2$ with abscissae $a < b$ and corresponding ordinates c, d. Thus $\mathcal{D}_{Trap} = \{T_\gamma : \gamma \in [0,1]^4, b \ge a\}$. To obtain finite subdictionaries, discretize the unit interval into $I_N = \{i/N : 0 \le i \le N\}$ and set $\mathcal{D}_N = \{T_\gamma : \gamma \in I_N^2 \times I_{N^2}^2\}$. Choose $N(\epsilon) = \epsilon^{-2}$, and set $\mathcal{D}_\epsilon = \mathcal{D}_{N(\epsilon)}$. It can be verified [8] that $\mathcal{D}_{Trap}$ is universal for $\mathcal{S} = \{\text{HÖLDER}_s(B) : 0 < s \le 2, 0 < B\}$, with $\nu = s/2, C = B^{1/2}$.

COROLLARY 10 ([8]). *On* $\text{HÖLDER}_s(B)$, *for* $0 < s \le 2$, *and setting* $r = s/(s+1)$,

$$r_\epsilon(\hat{f}_{CPRSS}, f) \le c_0 \cdot \log_2 \epsilon^{-2} \cdot B^{1-r} \epsilon^{2r}.$$

A key remark is that this adaptively (near minimax) rate of convergence is better than the rate attainable using a two dimensional tensor product wavelet basis when $s > 1$.

Nevertheless, a serious practical defect of Theorem 8 is the combinatorial search implicit in the definition of $\hat{\mu}_{CPRSS}$. The development of fast algorithms suitable for specific cases is an active direction of current research [5, 13].

ACKNOWLEDGEMENTS The results described here were obtained in collaboration with David Donoho, and some also in joint work with Gérard Kerkyacharian, Dominique Picard, and Bernard Silverman. Work on this manuscript and talk was supported in part by NSF, NIH and the Guggenheim Foundation.

REFERENCES

[1] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Its Applications*, 00:in press, 1999.

[2] B.S. Cirel'son, I.A. Ibragimov, and V.N. Sudakov. Norm of gaussian sample function. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*, Lecture Notes in Mathematics, 550, pages 20–41, 1976.

[3] A. Cohen, I. Daubechies, and P. Vial. Wavelets and fast wavelet transform on an interval. *Applied Computational and Harmonic Analysis*, 1:54–81, 1993.

[4] R.R. Coifman, Y. Meyer, and M.V. Wickerhauser. Wavelet analysis and signal processing. In B. Ruskai et. al., editor, *Wavelets and their Applications*, pages 153–178. Jones and Bartlett, 1992.

[5] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best-basis selection. *I.E.E.E. Transactions on Information Theory*, 38:713–718, 1992.

[6] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[7] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

[8] D. L. Donoho and I. M. Johnstone. Empirical atomic decomposition. Technical report, Stanford University, 1995.

[9] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 1998. To appear.

[10] D. L. Donoho and I. M. Johnstone. Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica*, 00:in press, 1999.

[11] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57:301–369, 1995. With Discussion.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Universal near minimaxity of wavelet shrinkage. In Pollard D., Torgersen E., and Yang G.L., editors, *Festschrift for L. Le Cam*, pages 183–218. Springer Verlag, 1997.

[13] D.L. Donoho. Wedgelets: Nearly minimax estimation of edges. Technical report, Dept. of Statistics, Stanford University, 1997.

[14] S.Yu. Efroimovich and M.S. Pinsker. A learning algorithm for nonparametric filtering. *Automat. i Telemeh.*, 11:58–65, 1984. (in Russian).

[15] M. Frazier, B. Jawerth, and G. Weiss. *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, 79. American Mathematical Society, Providence, RI, 1991.

[16] J.H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–67, 1991. (with discussion).

[17] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory*, pages 361–380. University of California Press, 1961.

[18] I. M. Johnstone. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statistica Sinica*, 00:in press, 1999.

[19] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B.*, 59:319–351, 1997.

[20] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction: Lecture Notes in Mathematics*. Springer Verlag: New York, 1993.

[21] M. Ledoux. Isoperimetry and gaussian analysis. In P. Bernard, editor, *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilities de Saint Flour, 1994*. Springer Verlag, 1996.

[22] P.G. Lemarié and Y. Meyer. Ondelettes et bases Hilbertiennes. *Revista Matematica Iberoamericana*, 2:1–18, 1986.

[23] M.S. Pinsker. Optimal filtering of square integrable signals in gaussian white noise. *Problems of Information Transmission*, 16:120–133, 1980. originally in Russian in *Problemy Peredatsii Informatsii* 16 52-68.

[24] C. Stein. Efficient nonparametric estimation and testing. In *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pages 187–195. University of California Press, Berkeley, CA., 1956.

[25] Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9:1135–1151, 1981.

Iain M. Johnstone
Department of Statistics
Stanford University
Stanford CA 94305 U.S.A.

Branching Processes, Random Trees and Superprocesses

Jean-François Le Gall

Abstract. We present some recent developments concerning the genealogy of branching processes, and their applications to superprocesses. We also discuss connections with partial differential equations, statistical mechanics and interacting particle systems.

1991 Mathematics Subject Classification: 60J80,60G57,60K35,35J60
Keywords and Phrases: Branching processs, random tree, superprocess, semilinear partial differential equation, Lévy process, voter model, lattice tree.

1 Discrete and continuous genealogical trees

(1.1) *Galton-Watson processes and trees.* A Galton-Watson branching process describes the evolution in discrete time of a population where each individual gives rise, independently of the others, to a random number of children distributed to a given offspring distribution. To be specific, consider an integer $k \geq 0$ representing the initial population, and a probability distribution ν on the set $\mathbb{N}$ of nonnegative integers. The corresponding Galton-Watson process is the Markov chain $(N_n, n \geq 0)$ in $\mathbb{N}$ such that, conditionally on N_n,

$$N_{n+1} \stackrel{(\mathrm{d})}{=} \sum_{i=1}^{N_n} U_i$$

where $U_1, U_2, \ldots$ are independent and distributed according to ν.

It is obvious that the genealogy of such a branching process can be described by k discrete trees. Take $k = 1$ for simplicity. Then the genealogical tree of the population is defined in the obvious way as a random subset $\mathcal{T}$ of $\bigcup_{n=0}^{\infty}(\mathbb{N}^*)^n$, where $\mathbb{N}^* = \{1, 2, 3, \ldots\}$ and $(\mathbb{N}^*)^0 = \{\emptyset\}$ by convention (cf Fig.1 for an example). Here $\emptyset$ labels the ancestor of the population and, for instance, $(3, 2)$ corresponds to the second child of the third child of the ancestor.

Throughout this article, we will concentrate on the critical or subcritical case where $m = \sum_{j=0}^{\infty} j\nu(j) \leq 1$ and we also exclude the (trivial) case where $\nu(\{1\}) = 1$. Then the population becomes extinct in finite time and so the tree $\mathcal{T}$ is a.s. finite.

(1.2) *Continuous-state branching processes.* Continuous-state branching processes (in short, CSBP's) are the continuous analogues of Galton-Watson processes. Formally, a CSBP is a Markov process Y in $\mathbb{R}_+$ whose transition kernels $(P_t(x, dy); t \geq 0, x \in \mathbb{R}_+)$ satisfy the additivity or branching property

$P_t(x+x',\cdot) = P_t(x,\cdot) * P_t(x',\cdot)$. Lamperti [15] has shown that these processes are exactly the scaling limits of Galton-Watson processes. Start from a sequence N^n of Galton-Watson processes with initial values k_n and offspring distributions ν_n depending on n. Suppose that there exists a sequence of constants $a_n \uparrow \infty$ such that

$$\lim_{n\to\infty} \big(\frac{1}{a_n} N^n_{[nt]}, t \geq 0\big) = (Y_t, t \geq 0) \tag{1}$$

in the sense of weak convergence of the finite-dimensional marginals. Then the limiting process Y must be a CSBP, and conversely any CSBP can be obtained in this way.

The distribution of a CSBP can be described analytically as follows. Here again, we restrict our attention to the critical or subcritical situation where $\int y\, P_t(x,dy) \leq x$. Then, the Laplace functional of the kernels $P_t(x,dy)$ must be of the form $\int P_t(x,dy)\, e^{-\lambda y} = \exp(-x u_t(\lambda))$, and the function $u_t(\lambda)$ solves the ordinary differential equation

$$\frac{\partial u_t(\lambda)}{\partial t} = -\psi(u_t(\lambda)) \ , \qquad u_0(\lambda) = \lambda, \tag{2}$$

with a function ψ of the type

$$\psi(u) = \alpha u + \beta u^2 + \int_{(0,\infty)} \pi(dr)\, (e^{-ru} - 1 + ru), \tag{3}$$

where $\alpha, \beta \geq 0$ and π is a σ-finite measure on $(0,\infty)$ such that $\int (r \wedge r^2)\, \pi(dr) < \infty$. Conversely, for any choice of a function ψ of the type (3), there exists an associated CSBP, which we will call the ψ-CSBP.

The case when $\psi(u) = \beta u^2$ (quadratic branching mechanism) is of special importance. The associated process is called the Feller diffusion. It occurs as the limit in (1) when $\nu_n = \nu$ has mean 1 and finite variance, and $k_n \approx \lambda\, n$, $a_n = n$.

In contrast with the discrete setting, it is no longer straightforward to define the genealogical structure of a CSBP. At an informal level, one would like to answer questions of the following type. Suppose that we divide the population at time t in two parts, say green individuals and red individuals. Then which part of the population at time $t+s$ does consist of descendants of green individuals, resp. red individuals ? This should be answered in a consistent way when s and t vary.

(1.3) *The quadratic branching case.* It has been known for some time that the genealogical structure of the Feller diffusion can be coded by excursions of linear Brownian motion. To explain this coding, we will recall a result of Aldous [1].

Start from an offspring distribution ν on $\mathbb{N}$ with mean 1 and finite variance. Consider the Galton-Watson tree with offspring distribution ν, conditioned to have exactly n edges (some mild assumption on ν is needed here so that this conditioning makes sense). Then, provided we rescale each edge by the factor $1/\sqrt{n}$, this conditioned tree, denoted by $\mathcal{T}_{(n)}$, converges in distribution as $n \to \infty$ to the so-called Continuum Random Tree (CRT).

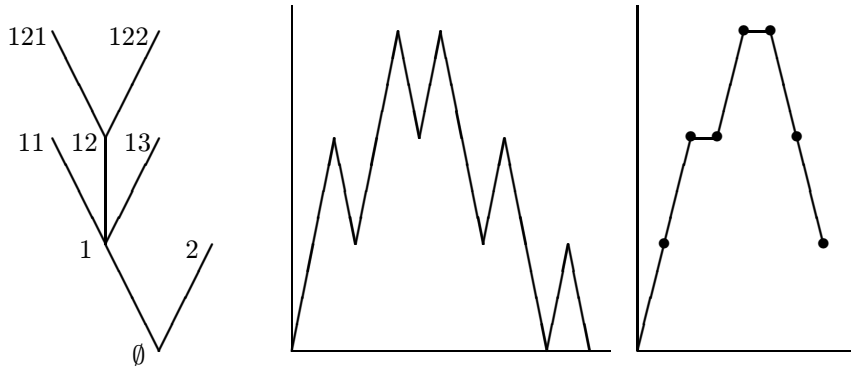
To give a precise meaning to the last statement, we need to say what the CRT is and to explain the meaning of the convergence. The easiest definition of

the CRT is via the coding by a continuous function. Let $e = (e(s), s \geq 0)$ be a continuous function from $\mathbb{R}_+$ into $\mathbb{R}_+$ with compact support and let σ denote the supremum of the support of e. We can then think of this function as coding a "continuous tree" through the following prescriptions:

- Each $s \in [0, \sigma]$ labels a vertex of the tree at generation $e(s)$.
- The vertex s is an ancestor of the vertex s' if $e(s) = \inf_{r \in [s,s']} e(r)$. (In general, the quantity $\inf_{r \in [s,s']} e(r)$ is the generation of the last common ancestor to s and s'.)
- The distance on the tree is $d(s, s') = e(s) + e(s') - 2\inf_{r \in [s,s']} e(r)$, and we identify s and s' if $d(s, s') = 0$.

According to these definitions, the set of ancestors (line of ancestors) of a given vertex s is isometric to the segment $[0, e(s)]$. The lines of ancestors of two vertices s and s' have a common part corresponding to the segment $[0, \inf_{r \in [s,s']} e(r)]$. More generally, for any finite set $s_1, \ldots, s_k$ of vertices, we can make sense of the reduced tree consisting of the lines of ancestors of $s_1, \ldots, s_k$ (see [1] and [17] for more details).

The CRT is the (random) continuous tree that corresponds in the previous coding to the case when the function e is a normalized Brownian excursion (positive Brownian excursion conditioned to have duration 1). Furthermore, the convergence of discrete trees towards the CRT should be understood as follows. Consider for each conditioned tree $\mathcal{T}_{(n)}$, the contour process of the tree (cf Fig.1). Provided that we rescale space by the factor $1/\sqrt{n}$ and space by the factor $1/(2n)$, the contour process of $\mathcal{T}_{(n)}$ converges in distribution towards the normalized Brownian excursion.



Galton-Watson tree contour process height process

Figure 1

To summarize the previous considerations, we can say that the genealogical structure of the Feller diffusion ($\psi(u) = \beta u^2$) is coded by excursions of linear Brownian motion. This fact has appeared in different forms in many articles relating random

walks or linear Brownian motion to branching processes (see in particular Harris [14], Dwass [9], Neveu-Pitman [22], etc.). It is also implicit in the Brownian snake construction of quadratic superprocesses [16], to which we will come back later.

In the next section, we will address the question of extending the previous coding to a general branching mechanism ψ.

2 Coding the genealogy of continuous-state branching processes

(2.1) *The discrete coding.* Consider a sequence $\mathcal{T}_1, \mathcal{T}_2, \dots$ of independent ν-Galton-Watson trees. Write σ_k for the number of vertices (or individuals) in the tree $\mathcal{T}_k$. Then suppose that we enumerate the vertices of the trees $\mathcal{T}_1, \mathcal{T}_2, \dots$ in lexicographical order: We write $\mathcal{T}_k = \{u_{\sigma_1+\cdots+\sigma_{k-1}}, u_{\sigma_1+\cdots+\sigma_{k-1}+1}, \dots, u_{\sigma_1+\cdots+\sigma_k-1}\}$ where $u_{\sigma_1+\cdots+\sigma_{k-1}}, u_{\sigma_1+\cdots+\sigma_{k-1}+1}, \dots, u_{\sigma_1+\cdots+\sigma_k-1}$ are the vertices of the tree $\mathcal{T}_k$ listed in lexicographical order.

Then for every $n \geq 0$, let H_n be the length (or generation) of the vertex u_n. The (random) process $(H_n, n \geq 0)$ is called the discrete height process (cf Fig.1 for an example with one tree). It is a variant of the contour process that was mentioned previously. It is easy to see that the data of the sequence $(H_n, n \geq 0)$ completely determines the sequence of trees and in this sense provides a coding of the trees. The interest of this coding comes from the following elementary lemma.

Lemma 2.1 *There exists a random walk $(S_n, n \geq 0)$ on $\mathbb{Z}$, with initial value $S_0 = 0$ and jump distribution $\mu(k) = \nu(k+1)$ for $k = -1, 0, 1, 2, \dots$, such that, for every $n \geq 0$,*

$$H_n = \mathrm{Card}\{j \in \{0, 1, \dots, n-1\}, S_j = \inf_{j \leq k \leq n} S_k\}. \tag{4}$$

Note that the random walk S is "left-continuous" in the sense that its negative jumps are of size -1 only. This lemma is taken from [19]. Closely related discrete constructions can be found in Borovkov-Vatutin [3] and Bennies-Kersting [2].

(2.2) *The continuous height process.* The previous lemma gives an explicit formula for the height process coding a sequence of Galton-Watson trees in terms of a random walk. Following [19], we will explain how this formula can be generalized to the continuous setting, thus yielding a coding of the genealogy of a CSBP in terms of a Lévy process with no negative jump (the continuous analogue of the left-continuous random walk S).

We start from a Lévy process X with no negative jump. We assume that $X_0 = 0$ and that that X does not drift to $+\infty$. Then the law of X is characterized by its "Laplace transform" $E[\exp(-\lambda X_t)] = \exp(t\psi(\lambda))$ (for $\lambda > 0$), where the possible functions ψ are exactly of the type (3), with the same assumptions on α, β and π. We assume in addition that $\beta > 0$ or $\int r\pi(dr) = \infty$ (or both these properties). This is equivalent to assuming that the paths of X are of infinite variation. (A simpler parallel theory can be developed in the finite variation case.) An important special case is the stable case $\psi(\lambda) = \lambda^{1+b}$, $0 < b < 1$.

Our first aim is to give a continuous analogue of the discrete formula (4). For every fixed $t \geq 0$, we let $X^{(t)} = (X_s^{(t)}, 0 \leq s \leq t)$ be the time-reversed process $X_s^{(t)} = X_t - X_{(t-s)-}$, and $M_s^{(t)} = \sup_{r \leq s} X_r^{(t)}$ be the associated maximum

process. Note that $(X_s^{(t)}, 0 \le s \le t) \overset{(d)}{=} (X_s, 0 \le s \le t)$. The process $M^{(t)} - X^{(t)}$ is a Markov process in $\mathbb{R}_+$ and under our assumptions 0 is a regular point for this Markov process. This enables us to set the following definition.

DEFINITION 2.2 *For every $t \ge 0$, let H_t denote the local time at level 0 and at time t of the process $M^{(t)} - X^{(t)}$. The process $(H_t, t \ge 0)$ is called the ψ-height process.*

A few comments are in order here. First, one needs to specify the normalization of local time. This can be achieved via the following approximation

$$H_t = P - \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^t ds\, 1_{\{M_s^{(t)} - X_s^{(t)} < \varepsilon\}}.$$

Secondly, we have defined H_t for every fixed t, and the measurability properties of the process $(H_t, t \ge 0)$ are not obvious. One can in a canonical way construct a lower-semicontinuous modification of the process $(H_t, t \ge 0)$ (see [19]).

In one special case, namely when $\beta > 0$, one can give a much simpler formula for H_t: If $I_t^s = \inf\{X_r; s \le r \le t\}$, we have $H_t = \beta^{-1} m(\{I_t^s; 0 \le s \le t\})$, where m denotes Lebesgue measure on $\mathbb{R}$ (from this formula one immediately sees that H has continuous paths when $\beta > 0$). In the quadratic case $\psi(\lambda) = \beta\lambda^2$ (X is then a linear Brownian motion), we get that $H_t = \beta^{-1}(X_t - I_t^0)$ is a reflected linear Brownian motion, which agrees with the considerations in (1.3).

We now (informally) claim that H codes the genealogy of a ψ-CSBP "starting with an infinite mass". This should be understood in the sense of the coding of continuous trees via functions as explained previously. (Our present setting is slightly more general because the process H does not always have continuous sample paths.) Analogously to the discrete case, we get the genealogy of a ψ-CSBP starting at $\rho > 0$ by stopping H at $T_\rho = \inf\{t \ge 0, X_t = -\rho\}$.

In what follows, we will give several statements that provide a rigorous justification of the previous informal claim. We first state a "Ray-Knight theorem" that formalizes the naive idea that the number of visits of H at a level a corresponds to the population of the tree at that level.

THEOREM 2.3 [19] *For every $a \ge 0$, the formula*

$$L_t^a = P - \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^t ds\, 1_{\{a < H_s < a+\varepsilon\}}$$

defines a continuous increasing process $(L_t^a, t \ge 0)$. If $T_\rho = \inf\{t \ge 0, X_t = -\rho\}$, the process $(L_{T_\rho}^a, a \ge 0)$ is a ψ-CSBP started at ρ.

When $\psi(u) = \beta u^2$, Theorem 2.3 reduces to a classical Ray-Knight theorem for Brownian local times. In general, Theorem 2.3 can be applied to study the sample path continuity of H.

THEOREM 2.4 [19] *The process H has a continuous modification if and only if $\int^\infty \frac{du}{\psi(u)} < \infty$.*

This condition holds in particular when $\beta > 0$ and in the stable case.

(2.3) *From discrete trees to continuous trees.* Our next result shows that if a sequence of rescaled Galton-Watson processes converges to a ψ-CSBP, the corresponding discrete height processes, suitably rescaled, also converge to the continuous height process H. This is analogous to Aldous' result in the quadratic branching case and proves in some sense that whenever rescaled Galton-Watson processes converge, their genealogical structure also converges to that of the limiting CSBP.

We consider a sequence (ν_n) of offspring distributions and a sequence (a_n) of positive numbers with $\lim a_n = \infty$. For every n let N^n be a Galton-Watson process with offspring distribution ν_n and initial value $N_0^n = [a_n]$.

THEOREM 2.5 [19],[7] *Suppose that the convergence* (1) *holds and that Y is a ψ-CSBP. For every $n \geq 1$, let H^n be the discrete height process associated with a sequence of independent ν_n-Galton-Watson trees. Then,*

$$\lim_{n\to\infty} \big(\frac{1}{n} H^n_{[na_n t]}, t \geq 0\big) = (H_t, t \geq 0) \tag{5}$$

in the sense of weak convergence of finite-dimensional marginals.

The last convergence can be shown to hold in a functional sense, provided that some regularity conditions are satisfied (Duquesne [7]). This reinforcement is important in various applications to invariance principles for functionals of Galton-Watson trees. For instance, one may want to look at the limiting behavior of the reduced tree that consists only of the ancestors of individuals alive at time p. The point is that this reduced tree can be written as an (almost) continuous functional of the discrete height process. Thus the (reinforced) convergence (5) allows one to pass to the limit and to obtain a limiting tree that is a simple functional of the height process H (see [7]).

3 SUPERPROCESSES

(3.1) *The snake construction.* Roughly speaking, superprocesses are obtained by combining a continuous branching mechanism with a Markovian spatial motion. To give a formal definition, consider a function ψ of the type (3) and a Borel right Markov process $(\xi_t, t \geq 0; \Pi_x, x \in E)$ with values in a Polish space E. Let $M_f(E)$ stand for the space of finite measures in E. The (ξ, ψ)-superprocess is the Markov process Z with values in $M_f(E)$ whose transition kernels are determined as follows. For every $0 \leq s < t$ and every bounded continuous function g on E, $E[\exp -\langle Z_t, g\rangle \mid Z_s] = \exp(-\langle Z_s, v_{t-s}\rangle)$, where $(v_t(x), t \geq 0, x \in E)$ is the unique nonnegative solution of the integral equation

$$v_t(x) + \Pi_x \left(\int_0^t ds\, \psi(v_{t-s}(\xi_s)) \right) = \Pi_x(g(\xi_t)). \tag{6}$$

(Compare with (2).) When ξ is a diffusion process with generator L, (6) is the integral form of the partial differential equation $\frac{\partial v}{\partial t} = Lv - \psi(v)$, $v_0 = g$. In the

special case when ξ is Brownian motion in $\mathbb{R}^d$ and $\psi(u) = \beta\, u^2$, Z is called super-Brownian motion (see Perkins [23] for a discussion of super-Brownian motion and related processes).

We will now use our approach to the genealogy of the ψ-CSBP to give a construction of the (ξ, ψ)-superprocess. The idea is to use the height process H to construct in a Markovian way the individual spatial motions of the "particles" of the superprocess. To simplify the presentation, we assume that the condition of Theorem 2.4 holds, so that H has continuous sample paths.

Let us fix a starting point $x \in E$. Conditionally on $(H_s, s \geq 0)$, we define a path-valued (time-inhomogeneous) Markov process $(W_s, s \geq 0)$ whose law is characterized by the following properties:

- For every $s \geq 0$, $W_s = (W_s(t), 0 \leq t \leq H_s)$ is a finite cadlag path in E started at x and defined on the time interval $[0, H_s]$.
- If $s < s'$, $W_{s'}(t) = W_s(t)$ for every $t \leq m(s, s') := \inf_{[s,s']} H_r$, and, conditionally on $W_s(m(s, s'))$, $(W_{s'}(m(s, s') + t), 0 \leq t \leq H_{s'} - m(s, s'))$ is independent of W_s and distributed according to the law of ξ started at $W_s(m(s, s'))$.

Informally, W_s is a path of ξ started at x with length H_s. When H_s decreases, the path erases itself and when H_s increases the path extends itself by following the law of the spatial motion ξ. To summarize the previous properties, we will say that W is the snake driven by H with spatial motion ξ (and initial point x).

The connection with superprocesses is contained in the next theorem, which is essentially the main result of [20]. Recall the definition of L_t^a in Theorem 2.3.

Theorem 3.1 *For every $a \geq 0$, let Z_a be the random measure on E defined by*

$$\langle Z_a, g \rangle = \int_0^{T_\rho} d_s L_s^a \, g(W_s(a)).$$

Then $(Z_a, a \geq 0)$ is a (ξ, ψ)-superprocess started at $\rho\, \delta_x$.

To keep track of the dependence on the initial point x, we will use the notation $\mathbb{P}_x$ for the probability under which W is defined.

(3.2) *The Brownian snake and partial differential equations.* We now concentrate on the quadratic case $\psi(u) = \beta u^2$ and take $\beta = 1/2$ for definiteness. As pointed out previously, the process H is then a (scaled) reflected linear Brownian motion and in particular is Markovian. As a consequence, the process $(W_s, s \geq 0)$, which is now called the Brownian snake, is (time-homogeneous) Markov and indeed verifies the strong Markov property. This plays a crucial role in the applications that are outlined below.

From now on, we suppose that ξ is Brownian motion in $\mathbb{R}^d$. An easy application of the Kolmogorov criterion shows that W has a modification that is continuous with respect to the uniform topology on stopped (continuous) paths.

Our goal is to give some applications of the snake construction to connections between superprocesses and partial differential equations. These connections have

been investigated by Dynkin in a series of important papers (see in particular [10], [11]). The Brownian snake turns out to be a useful tool in the quadratic branching case. The key to the connections with partial differential equations is the next theorem, which reformulates in terms of the Brownian snake a result of Dynkin [10] valid for superprocesses with a more general branching mechanism. We let D be a domain in $\mathbb{R}^d$ and for every path w, we denote by $\tau(w) = \inf\{t \geq 0, w(t) \notin D\}$ the first exit time of D by w (with the convention $\inf \emptyset = \infty$).

THEOREM 3.2 *Let $x \in D$. The limit*

$$\langle Z^D, g \rangle = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^{T_1} ds\, 1_{\{\tau(W_s) < H_s < \tau(W_s)+\varepsilon\}} g(W_s(\tau(W_s)))$$

exists $\mathbb{P}_x$-a.s. for every continuous function on ∂D, and defines a random measure Z^D on ∂D called the exit measure from D. If D is regular (in the classical potential-theoretic sense) and g is continuous and nonnegative on ∂D, the formula

$$u(x) = -\log \mathbb{E}_x(\exp -\langle Z^D, g \rangle) \qquad x \in D \tag{7}$$

defines the unique nonnegative solution of the equation $\Delta u = u^2$ in D with boundary value $u_{|\partial D} = g$.

A nice feature of the probabilistic representation formula (7) is that it can be used to produce many other solutions via suitable passages to the limit. In the setting of our next result, a generalized form of this representation holds for *any* nonnegative solution.

We denote by $\mathcal{R}^D$ the random set $\{W_s(t); 0 \leq s \leq T_1, t \leq \tau(W_s) \wedge H_s\}$.

THEOREM 3.3 [18] *Let D be a domain of class C^2 in $\mathbb{R}^2$. Then, for every $x \in D$, $\mathbb{P}_x$ a.s., the random measure Z^D has a continuous density $z_D(y), y \in \partial D$ with respect to Lebesgue measure on ∂D. Furthermore, the formula*

$$u(x) = -\log \mathbb{E}_x\big(1_{\{\mathcal{R}^D \cap K = \emptyset\}} \exp -\langle \gamma, z_D \rangle\big), \quad x \in D \tag{8}$$

gives a one-to-one correspondence between nonnegative solutions of $\Delta u = u^2$ in D and pairs (K, γ), where K is a closed subset of ∂D and γ is a Radon measure on $\partial D \backslash K$.

In the representation of Theorem 3.3, both K and γ can be determined analytically in terms of the boundary behavior of u: K is the set of points in ∂D where u blows up like the inverse of the squared distance to the boundary, and γ corresponds to the usual trace of u on $\partial D \backslash K$.

The analytic part of Theorem 3.3 has been extended by Marcus and Véron [21] to the equation $\Delta u = u^p$, $p > 1$ in a smooth domain of $\mathbb{R}^d$, provided that $d < \frac{p+1}{p-1}$. (see also Dynkin and Kuznetsov [12],[13]). In the supercritical case $d \geq \frac{p+1}{p-1}$, things become more complicated: One can still define the trace of a general nonnegative solution as a pair (K, γ), but a solution is in general not uniquely determined by its trace, and not all pairs (K, γ) are admissible traces (see [21], [13]). Recently,

Dynkin and Kuznetsov [13] have proposed a finer definition of the trace that might lead to a one-to-one correspondence even in the supercritical case.

A remarkable feature of the connections between superprocesses or snakes and semilinear partial differential equations is the fact that almost all important probabilistic questions correspond to basic analytic problems, and conversely. We will give a last example involving on one hand a Wiener-type test for the Brownian snake and on the other hand solutions with boundary blow-up. We use the notation $c_{2,2}$ for the Sobolev capacity associated with the Sobolev space $W^{2,2}$.

Theorem 3.4 [6] *Let D be a domain in $\mathbb{R}^d$. The following statements are equivalent.*

(i) *There exists a nonnegative solution of $\Delta u = u^2$ in D that blows up everywhere at the boundary.*

(ii) *Let $T = \inf\{s \geq 0, W_s(t) \notin D \text{ for some } t \in (0, H_s]\}$. Then $\mathbb{P}_y(T = 0) = 1$ for every $y \in \partial D$.*

(iii) *$d \leq 3$, or $d \geq 4$ and for every $y \in \partial D$,*

$$\sum_{n=1}^{\infty} 2^{n(d-2)} c_{2,2}(D^c \cap \{z \in \mathbb{R}^d, 2^{-n} \leq |z - y| < 2^{-n+1}\}) = \infty.$$

4 Statistical mechanics and interacting particle systems

(4.1) *Lattice trees.* A lattice tree with n bonds is a connected subgraph of $\mathbb{Z}^d$ with n edges in which there are no loops.

We are interested in a limit theorem that gives information on the typical shape of a lattice tree when n is large. To this end, let $Q_n(d\omega)$ be the uniform probability measure on the set of all lattice trees with n bonds that contain the origin of $\mathbb{Z}^d$. For every tree ω, let $X_n(\omega)$ be the probability measure on $\mathbb{R}^d$ obtained by putting mass $\frac{1}{n+1}$ to each vertex of the rescaled tree $cn^{-1/4}\omega$. Here $c > 0$ is a positive constant.

Provided that the dimension d is large enough, Derbez and Slade [5] proved that the limiting behavior of the law of X_n under Q_n involves a random measure which is closely related to Aldous' CRT. To define this random measure, consider the snake W driven by a normalized Brownian excursion $(e(s), 0 \leq s \leq 1)$, assuming again that the spatial motion is Brownian motion in $\mathbb{R}^d$ (and the initial point is 0). Then the formula

$$\langle \mathcal{I}, f \rangle = \int_0^1 ds\, f(W_s(e(s)))$$

defines a random measure in $\mathbb{R}^d$, sometimes called Integrated Super-Brownian Excursion (ISE).

Theorem 4.1 [5] *For d sufficiently large and for a suitable choice of the constant $c = c(d) > 0$, the law of X_n under Q_n converges weakly to the law of $\mathcal{I}$.*

It is expected that the result holds when $d > 8$ (which is the condition needed to ensure that the topological support of $\mathcal{I}$ is a tree). This is true [5] if one considers "spread-out" trees rather than nearest-neighbor trees. A recent work of Hara and Slade indicates that ISE also appears as a scaling limit of the incipient infinite percolation cluster at the critical temperature, again in high dimensions ($d > 6$).

(4.2) *Interacting particle systems.* A number of recent papers explore the connections between the theory of superprocesses and some of the most classical interacting particle systems. Durrett and Perkins [8] show that the asymptotic behavior of the contact process in $\mathbb{Z}^d$ can be successfully analysed in terms of super-Brownian motion. Here we will concentrate on the classical voter model and follow a work in preparation in collaboration with M. Bramson and T. Cox. Closely related results can be found in a forthcoming article by Cox, Durrett and Perkins.

At each site of $\mathbb{Z}^d$ sits an individual who can have two possible opinions, say 0 or 1. At rate 1 each individual forgets his opinion and gets a new one by choosing uniformly at random one of his nearest neighbors and taking his opinion. Suppose that at the initial time all individuals have type 0, except for the individual at the origin who has type 1. For every $t > 0$, let $\mathcal{U}_t$ denote the set of individuals who have type 1 at time t, and let U_t be the random measure

$$U_t = \sum_{x \in \mathcal{U}_t} \delta_{x/\sqrt{t}}.$$

Then $P[\mathcal{U}_t \neq \emptyset] = P[U_t \neq 0]$ tends to 0 as $t \to \infty$, and the rate of this convergence is known [4]. One may then ask about the limiting behavior of U_t conditionally on $\{U_t \neq 0\}$.

The answer to this question can be formulated in terms of the snake W driven by a Brownian excursion conditioned to hit level 1, with spatial motion given by ($d^{-1/2}$ times) a standard Brownian motion in $\mathbb{R}^d$. We have the following result in dimension $d \geq 3$ (an analogous result holds for $d = 2$).

THEOREM 4.2 *The law of $t^{-1}U_t$ conditionally on $\{U_t \neq 0\}$ converges as $t \to \infty$ to the law of $c_d\mathcal{H}$, where $c_d > 0$ and the random measure $\mathcal{H}$ is defined by*

$$\langle \mathcal{H}, f \rangle = \int_0^\infty dL_s^1 \, f(W_s(1)),$$

where L_s^1 is as previously the local time of the excursion at level 1 and at time s.

To interpret this last theorem, one may say, for the voter model as well as for the (long-range) contact process [8], that the limiting behavior of the process depends on a pseudo-branching structure, which asymptotically comes close to the genealogical structure of the Feller diffusion.

REFERENCES

[1] D.J. Aldous. Ann. Probab. 21 (1993), 248-289.

[2] J. Bennies, G. Kersting. *A random walk approach to Galton-Watson trees.* To appear.

[3] K.A. Borovkov, V.A. Vatutin. J. Appl. Probab. 33 (1996), 614-622.

[4] M. Bramson, D. Griffeath. Z. Wahrsch. verw. Gebiete 53 (1980), 180-196.

[5] E. Derbez, G. Slade. Comm. Math. Physics 198 (1998), 69-104.

[6] J.S. Dhersin, J.F. Le Gall. Probab. Th. Rel. Fields 108 (1997), 103-129.

[7] T. Duquesne. *Théorèmes limites pour le processus d'exploration d'arbres de Galton-Watson*. Preprint (1998)

[8] R. Durrett, E.A. Perkins. *Rescaled contact processes converge to super-Brownian motion for* $d \geq 2$. Preprint (1998)

[9] M. Dwass. Proc. Amer. Math. Soc. 51 (1975), 270-274.

[10] E.B. Dynkin. Probab. Th. Rel. Fields 89 (1991), 89-115.

[11] E.B. Dynkin. Ann. Probab. 21 (1993), 1185-1262.

[12] E.B. Dynkin, S.E. Kuznetsov. *Trace on the boundary for solutions of nonlinear differential equations*. Trans. Amer. Math. Soc., to appear.

[13] E.B. Dynkin, S.E. Kuznetsov. *Fine topology and fine trace on the boundary associated with a class of quasilinear partial differential equations*. Comm. Pure Appl. Math., to appear.

[14] T.E. Harris. Trans. Amer. Math. Soc. 73 (1952), 471-486.

[15] J. Lamperti. Z. Wahrsch. verw. Gebiete 7 (1967), 271-288.

[16] J.F. Le Gall. Probab. Th. Rel. Fields 95 (1993), 25-42.

[17] J.F. Le Gall. Probab. Th. Rel. Fields 96 (1993), 369-383.

[18] J.F. Le Gall. Comm. Pure Appl. Math. 50 (1997), 69-103.

[19] J.F. Le Gall, Y. Le Jan. Ann. Probab. 26 (1998), 213-252.

[20] J.F. Le Gall, Y. Le Jan. *Branching processes in Lévy processes: Laplace functionals of snakes and superprocesses*. Ann. Probab., to appear.

[21] M. Marcus, L. Véron. *The boundary trace of positive solutions of semilinear elliptic equations, I. The subcritical case, II. The supercritical case*. To appear.

[22] J. Neveu, J.W. Pitman. Lecture Notes in Math. 1372, 248-257. Springer, 1989.

[23] E.A. Perkins. *Measure-valued branching diffusions and interactions*. Proceedings of the International Congress of Mathematicians, Zürich 1994. Birkhaäuser, 1995.

Jean-François Le Gall
D.M.I. Ecole Normale Supérieure
45, rue d'Ulm, F-75005 PARIS

Genetic Linkage Analysis: an Irregular Statistical Problem

David Siegmund

Abstract. Linkage analysis, which has the goal of locating genes associated with particular traits in plants or animals (especially inherited diseases in humans), leads to a class of "irregular" statistical problems. These problems are discussed with reference to an idealized model, which serves as a point of departure for more realistic versions of the problem. Some general results, adapted from recent research into "change-point" problems, are presented; and more specific problems arising out of the underlying genetics are discussed.

1991 Mathematics Subject Classification: 62M40, 92D10
Keywords and Phrases: gene mapping, linkage analysis, change point, irregular

1. Introduction. The goal of gene mapping, or linkage analysis, is to locate genes that affect particular traits, especially genes that affect human susceptibility to particular diseases and also genes that affect productivity of agriculturally important species. An artificially simplified, but illuminating genetic model leads to the following class of statistical problems. Observations are available on a doubly indexed set of random variables $Z(c, i\Delta)$, where $c = 1, \cdots, 23$ indexes the set of human chromosomes of genetic lengths ℓ_c and $i\Delta, 0 \le i\Delta \le \ell_c$ are the locations of markers spaced at intermarker distance Δ along each chromosome. For different values of c the random variables are independent. For each fixed c, $Z(c,t)$ is a stationary Gaussian process in t, which satisfies

$$\mathrm{Var}[Z(c,t)] = 1, \quad \mathrm{Cov}[Z(c,s), Z(c,t)] = R(t-s). \tag{1}$$

A case of particular interest is $R(t) = \exp(-\beta|t|)$. For most or perhaps all values of c

$$E[Z(c,t)] = 0 \text{ for all } t, \tag{2}$$

while for some c', $0 < \tau < \ell_{c'}$ and $\xi > 0$

$$E[Z(c',t)] = \xi R(t-\tau). \tag{3}$$

The values of c', τ, and ξ are all unknown. Thus the data consist of a large number of zero mean Gaussian processes observed at equally spaced "time" points. A small

number of these processes are superimposed on a mean value function defining a "peak" of an unknown height ξ at an unknown location τ, and having a known shape R. The statistical problems are to decide which chromosomes, if any, harbor such a location τ and estimate the location by a confidence region. These problems are "irregular" for two reasons: (i) the parameter τ is not identifiable when the nuisance parameter $\xi = 0$; the log likelihood function, which is proportional to $Z(c, \tau)$, is not a smooth function of the parameter τ, even if we are able to make continuous observations in t.

The purpose of this paper is (a) to explain briefly the genetic background of the preceding problems as they relate to mapping human disease genes, (b) propose a framework for their solutions that is useful as a point of departure for discussing more realistic versions of the problems, and (c) describe some alternative models designed to capture the complicating features arising in practice. Special consideration is given to the issue of multiple comparisons that arises through examining the large number of variables $Z(c, i\Delta)$ in searching for the relatively few values of c', t where the expected value is substantially different from 0, and to estimation of τ by confidence regions. Some of these problems can be understood in terms of recent literature on "change-point" problems, to which they are closely related.

2. GENETIC BACKGROUND. Given two related individuals, at a given locus in the genome two alleles are said to be identical by descent if they are inherited from a common ancestor. For example, a pair of half siblings can inherit zero or one allele identical by descent from their common parent, and according to Mendel's laws each of these possibilities has probability 1/2. Genes on different chromosomes segregate independently, while genes on the same chromosome tend to be inherited from the same parental chromosome, and are said to be *linked*. More precisely, if two half siblings share an allele identical by descent at locus t, they will share an allele identical by descent at a locus on a different chromosome with probability 1/2 and at a locus s on the same chromosome with a probability $(1-\phi) \in (1/2, 1)$. This probability is a decreasing function of the distance between s and t.

A pair of siblings can inherit zero or one allele identical by descent from their mother and similarly from their father, hence 0, 1, or 2 overall. For some purposes a single sib pair can be regarded as two independent half sib pairs, but in general siblings require a more complicated analysis. For ease of exposition, we consider only the much simpler case of half siblings.

The basic logic of linkage analysis is that if two relatives, e.g., half siblings or siblings, share an inherited trait, e.g., a disease, that is relatively rare in the population, it is likely that they share an allele predisposing them to the trait that has been inherited identical by descent. Thus the probability of identity by descent for an affected relative pair at a marker locus close to a trait locus is greater than the value given by Mendel's laws (1/2 in the case of half siblings). Our problem is to scan the genome of a sample of affected relatives in search of regions where the identity by descent exceeds the expected proportion by more than can be explained as a chance fluctuation.

A mathematical model for a pair of half siblings is as follows. Let X_t be 1 or 0 according as the half siblings are or are not identical by descent at locus t (on a chromosome c, *which henceforth is suppressed in the notation*). Then for a random pair of half siblings,

$$P\{X_t = 1\} = P\{X_t = 0\} = 1/2 \tag{4}$$

for all t; and for loci s and t on the same chromosome

$$P\{X_s = 1|X_t = 1\} = P\{X_s = 0|X_t = 0\} = 1 - \phi. \tag{5}$$

Assume that τ denotes a genetic locus predisposing to inheritance of the trait (and that there is no other trait locus on the given chromosome). Then for two half siblings sharing a trait in common,

$$P\{X_\tau = 1\} = (1 + \alpha)/2 > 1/2, \tag{6}$$

while the conditional probability (5) continues to hold for loci s, t on the same side of τ. In particular by taking $t = \tau$ in (5) we obtain $P\{X_s = 1\} = [1+\alpha(1-2\phi)]/2$. The value of ϕ in terms of the parameters s and t depends on the model used for the genetic process of recombination. According to the commonly used model suggested by Haldane in 1919,

$$\phi = [1 - \exp(-\beta|t - s|)]/2, \tag{7}$$

and more generally

$$\phi \sim \beta|t - s|/2 \text{ as } |t - s| \to 0. \tag{8}$$

The value of β is determined by the relation of the relative pair. For half siblings it is 0.04 when the units of genetic distance along a chromosome are centimorgans (cM). (One cM is defined as the distance for which the expected number of crossovers per meiosis is 0.01. The average length of a human chromosome is roughly 140 cM. See Suzuki *et al.* for a more thorough discussion.)

Assuming now that one observes identity by descent data for N independent half sibling pairs at marker loci, denoted $i\Delta$, equally spaced at intermarker distance Δ throughout the genome, we form the statistics

$$Z_{i\Delta} = N^{-1/2}\Sigma_{j=1}^{N}[2X_{i\Delta}^{j} - 1], \tag{11}$$

where the summation is over all half sibling pairs. It is possible starting from (11) to address the basic questions of Section 1 (cf. Feingold, 1993, Tu and Siegmund, 1998). A somewhat simpler and more complete analysis is possible if we introduce an additional approximation. It follows from the central limit theorem that as $N \to \infty$ and $\alpha \to 0$ in such a way that $N^{1/2}\alpha \to \xi \geq 0$ the process $Z_{i\Delta}$ defined in (11) converges in distribution to a process, which by (4)-(7) has the properties described in (1) - (3). By an abuse of notation we continue to denote this new process by $Z_{i\Delta}$. Thus we return to the problems already formulated in Section 1.

3. GENOME WIDE FALSE POSITIVE ERROR RATE. If $i\Delta$ in (11) is equal to τ, it follows from (6) that (11) is the score statistic for testing whether $\alpha = 0$; it is also the likelihood ratio statistic in the approximating Gaussian model. Since usually τ is unknown, to test for linkage somewhere on the genome we use

$$\max_c \max_i Z_{i\Delta}. \tag{12}$$

To evaluate approximately the false positive error rate, i.e., the probability under the hypothesis of no linkage throughout the entire genome that (12) exceeds a threshold b, we assume that $b \to \infty$ and $\Delta \to 0$, in such a way that $b\Delta^{1/2}$ converges to a positive constant. Then for a genome wide search

$$P\Big\{\max_c \max_i Z_{i\Delta} > b\Big\} \approx 1 - \exp\{-C\Big[1 - \Phi(b)\Big] - \beta L b \varphi(b) \nu\Big(b\{2\beta\Delta\}^{1/2}\Big)\}. \tag{13}$$

Here Φ and φ are the standard normal distribution function and density function, respectively, C is the number of chromosomes and $L = \Sigma_c \ell_c$ is the total length of the genome in cM. The function ν, which arises in the fluctuation theory of random walks developed by Spitzer in the 1950's, is defined by

$$\nu(x) = 2x^{-2} \exp[-2\Sigma n^{-1} \Phi(-xn^{1/2}/2)]. \tag{14}$$

For small x it is easily evaluated via the relation $\nu(x) = \exp(-\rho x) + o(x^2)$, where $\rho = -\zeta(1/2)/(2\pi)^{1/2} \approx 0.583$, while the series in (14) converges very rapidly for large x. For a numerical example, for markers every $\Delta = 1$ cM and a human genome of 23 chromosomes of average length 140 cM the threshold $b = 3.91$ gives a false positive error rate equal to the conventional 0.05. The approximation (13) was given by Feingold, Brown and Siegmund (1993), as an application of the method of Woodroofe (1976).

4. POWER. To obtain an approximation to the power that we detect a disease locus on a correct chromosome (for simplicity we assume there is at most one on any given chromosome), we first suppose that the disease locus τ is itself a marker locus. We then have the approximation

$$P\Big\{\max_k Z_{k\Delta} \geq b\Big\} \approx 1 - \Phi(b - \xi) + \varphi(b - \xi)\Big[2\nu/\xi - \nu^2/(b + \xi)^2\Big], \tag{15}$$

where $\nu = \nu\Big(b\{2\beta\Delta\}^{1/2}\Big)$, as defined above. The first term in (15) is simply the probability that the process exceeds the threshold b at the disease locus. A disease locus between marker loci needs a similar but more complicated argument involving the (correlated) process $Z_{i\Delta}$ at the two flanking markers. The resulting approximation requires a one dimensional numerical integration for its numerical evaluation.

For the 1 cM intermarker distance and threshold $b = 3.91$ considered in the preceding section, and a disease locus midway between two markers a noncentrality parameter of $\xi = 5.03$ is needed to achieve power of 0.9 to detect the disease locus.

For a given value of the genetic parameter α, this can be converted to a sample size requirement by virtue of the relation $\xi = N^{1/2}\alpha$.

5. Confidence regions. A confidence region can be used to identify a chromosomal region in which to concentrate the search for the exact location of a disease gene. We discuss here two methods that are motivated by the recent literature on "change-point" problems, which have essentially the same structure. These methods are (i) support regions and (ii) Bayesian credible sets. (See Siegmund, 1989, for a review of the change-point literature). Note that as a consequence of the irregularity of this problem, the maximum likelihood estimator of τ is not normally distributed, so it is not correct to use the maximum likelihood estimator plus or minus two estimated standard errors as an approximate 95% confidence interval.

We assume that a disease gene has been correctly identified to lie on a particular chromosome, which contains no other disease gene. For simplicity we assume that the locus τ is exactly a marker locus. Since many investigators type additional markers in the proximity of an apparent disease gene, this latter assumption is often approximately true in practice.

The traditional genetic technique for estimating the location of a disease gene is a support region, which for our purposes can be defined as follows. Given $c > 0$, a support region contains all loci $j\Delta$ such that

$$Z_{j\Delta}^2 \geq \max_i Z_{i\Delta}^2 - c. \tag{16}$$

Within the framework of the approximate Gaussian model, this is equivalent to the standard statistical technique of inverting the likelihood ratio test that $j\Delta$ is the disease locus, to obtain a confidence region. If the problem were regular, which in this case would require that Z_t be twice continuously differentiable in t, the probability of (16) would be given approximately by a χ^2 distribution with one degree of freedom; but that approximation is not correct here. By methods similar to those used to obtain (13) one can approximate the probability of (16) and show that (16) yields an approximate confidence region for the disease locus (Feingold, Brown and Siegmund, 1993, Lander and Kruglyak, 1995, Dupuis and Siegmund, 1998).

Because of the local linear decay near τ displayed in (3), the inequality (16) will be satisfied at all loci within a distance from τ of roughly $c/2\beta\xi^2$. Since ξ is proportional to $N^{1/2}$, the expected size of the support region is proportional to N^{-1}. This stands in contrast to regular problems, where the likelihood function decays quadratically, and the size of a confidence region is proportional to $N^{-1/2}$. It may be shown by more detailed analysis that a value $c \approx 4.5$ corresponds roughly to a 90% confidence interval when $\Delta = 1$ and $\beta = 0.04$. Then for $\xi \approx 5$, the value indicated above that one needs to detect linkage with power about 0.9, the expected size of a support region is about 5 cM. Since this corresponds to about 5×10^6, base pairs, one still needs additional information, invariably of a qualitatively different kind, to locate the gene with precision at the base pair level.

In his study of the closely related change-point problem, Cobb (1978) observed that if ξ were known, the problem of estimation of τ would have essentially

the same structure as estimation of a location parameter. Hence Fisher's (1934) suggestion for estimating a location parameter, to use the conditional distribution of the maximum likelihood estimator given the ancillary statistic, in our case the local rate of decay of the likelihood function, is very attractive. Moreover, this suggestion has minimal computational requirements, since it can be effected by a formal Bayesian credible region based on a uniform prior distribution for τ. To accommodate unknown ξ, one can introduce a prior distribution for ξ or use the profile likelihood function obtained by maximization with respect to ξ for each fixed τ.

Dupuis and Siegmund (1998) have compared these two methods and find that they are roughly comparable, although the former is more robust under a variety of conditions.

6. MULTILOCUS MODELS. There are many additional problems that require a more detailed understanding of the underlying genetics than we have presented so far. In this section we discuss traits involving more than one gene, while in the next we very briefly point out several additional problems.

While some inherited human diseases are governed by a single gene, most of the more common ones having a genetic component, e.g., diabetes, breast cancer, Alzheimer's disease, are known or thought to involve multiple genes. Conceptually the simplest of these are *heterogeneous* traits, where susceptibility increases by virtue of a mutant allele at any one of several loci. It is, of course, possible that the genome scan defined above would identify several disease loci, even though there is no particular effort to do so. Typically a much larger sample size would be required than for a single gene trait having a comparable degree of heritability, since the evidence for linkage is divided among the different disease loci.

Three methods have been suggested to deal with heterogeneous traits: (i) simultaneous search, (ii) conditional search and (iii) homogenization. In simultaneous search, suggested originally by Lander and Botstein (1986), one hypothesizes a specific number, say two, trait loci and searches over combinations of putatitve loci to identify both simultaneously. Because there is a much larger number of multiple comparisons, a suitable threshold under the conditions assumed above would increase from the neighborhood of 4 to about 5 (in searching for two loci). Conditional search, which is appropriate after some trait loci have already been identified, involves stratification of the sample according to the identity by descent status at the (estimated) location of the already discovered loci in order to increase precision in searching for additional trait loci. See Dupuis, Brown and Siegmund (1995) for a theoretical analysis of these two methods. An interesting application of conditional search is contained in Morahan *et al.* (1996), who identified a gene on chromosome two for insulin dependent diabetes by conditioning on the identity by descent status of their sample of sib pairs at the HLA locus on chromosome 6, which had been implicated in several earlier studies.

A third approach to alleviate the problem of heterogeneity is to develop a narrow definition of the disease, in order to make the disease more homogeneous. In some cases this definition can be achieved statistically. A notable success was the identification of the breast cancer gene BRCA1 by defining the trait to be

early onset breast cancer. A recent attempt in the same direction involved a search for a gene contributing to noninsulin dependent diabetes (Mahtani *et al.*, 1996). After failing to find evidence of linkage in the complete study group, the pedigrees in the study were identified with their average level of a quantitative covariate thought to be associated with the trait. The analysis was repeated with only those pedigrees in the most extreme 25% of the distribution of this covariate, then the most extreme 50%, then the most extreme 75%. The genome scan in the most extreme 25% turned up a value that would have been marginally significant at the 0.05 level if the phenotype had been defined *a priori*, but now there is the second dimension of multiple comparisons (i.e., the search over levels of the covariate) to account for.

A suitable model to analyze this two dimensional search within the Gaussian framework introduced above is as follows. Let $Z(t,k)$ for $k = 1, \cdots, m$ be independent identically distributed Gaussian processes in t as defined in Section 1. Here k denotes levels of the covariate and for convenience is assumed to involve equal quantiles of its distribution. Then let

$$S(t,k) = k^{-1/2}\Sigma_{i=1}^{k} Z(t,i).$$

Linkage is detected if

$$\max_{1\le k\le m} \max_{c} \max_{j} S(j\Delta, k) \ge b \tag{17}$$

for a suitable threshold b. Using the method of Siegmund (1988), which generalizes Woodroofe (1976) to multidimensional time, one finds under the hypothesis of no linkage that the probability of (17) is approximately

$$1 - \exp\Big(-\beta L\nu[b(2\beta\Delta)^{1/2}]b^3\phi(b)\int_{bm^{-1/2}}^{\infty} x^{-1}\nu(x)\,dx\Big). \tag{18}$$

For the threshold $b = 3.91$ appropriate for the simple scan of Section 1 when $\Delta = 1$, we find when $m = 4$ that (18) is about 0.15. To obtain a false positive rate of 0.05, one must increase the threshold to $b = 4.2$. Some rough calculations, which should be more carefully analyzed, indicate that if the covariate is effective in "homogenizing" the original sample, one can sometimes achieve substantial gains in power after allowing for the increase in threshold.

In the paper of Mahtani *et al.* (1996) there was the additional problem that the study design required pedigrees to have at least three affecteds and employed a statistic whose distribution under the hypothesis of no linkage is skewed to the right. (See (iii) in Section 7 below.) As a consequence the p-value of their result was about 0.24 after one adjusts for skewness in addition to the two dimensional search.

7. Additional problems. Linkage analysis involves a large number of problems in addition to those discussed above. A few that have been the subject of recent research follow.

(i) The identity by descent data that form the basis of our previous discussion are intrinsically incomplete and require complicated algorithms to process. For

example, for a given relative pair a particular marker may be "informative," so that we can observe the identity by descent status at that marker, or it may be "uninformative." Intermediate possibilities also exist. Since by (5) identity by descent status is correlated at nearby markers, it may be possible to infer that status at an uninformative marker from the status at nearby informative markers. For example, for half siblings it follows from (6) that the likelihood function (for the case of completely informative markers, when the trait locus τ is itself a marker locus) equals

$$\Pi_{j=1}^{N} (1+\alpha)^{X_\tau^j}(1-\alpha)^{1-X_\tau^j}.$$

Let G denote the observed genotypes of all individuals at all markers, and let P_0 denote probability under the hypothesis of no linkage. Then the likelihood function (relative to P_0) when some of the X_τ^j may not be observable is

$$\Pi_{j=1}^{N} E_0[(1+\alpha)^{X_\tau^j}(1-\alpha)^{1-X_\tau^j}|G] = \Pi_{j=1}^{N}[1+\alpha(2Y_\tau^j-1), \tag{19}$$

where $Y_\tau^j = E_0[X_\tau^j|G]$. Kruglyak *et al.* (1996) use hidden Markov chains to calculate the required conditional expectations. Their algorithm works best for a possibly large number of small pedigrees. Additional techniques are required for studies involving large pedigrees, which can make the required calculations extremely onerous (cf. Thompson, 1994). By differentiating (19) one sees that the score statistic for testing $\alpha = 0$ is

$$\hat{Z}_\tau = \Sigma_j[Y_\tau^j - 1/2]/[\Sigma_j \mathrm{Var}(Y_\tau^j)]^{1/2},$$

which reduces to (11) in the case of complete data. Since τ is unknown, we use $\max_c \max_i \hat{Z}_{i\Delta}$ to search the genome for evidence of linkage. By studying the correlation function of $\hat{Z}_{i\Delta}$, Teng and Siegmund (1998) show under certain conditions that a threshold b appropriate for the case of completely informative markers studied above is approximately correct for $\hat{Z}_{i\Delta}$ as well. They also study the effect of incompletely informative markers on the power to detect linkage. These problems are difficult, and pose a number of impediments to a completely satisfactory solution.

(ii) Many traits are defined by quantitative measurement rather than a yes/no dichotomy. Understanding the genetic basis of quantitative traits is also of interest in experimental genetics, e.g., for agriculturally important species or for animal models of human diseases. At the level of abstraction provided by Gaussian approximations one finds that linkage analysis of quantitative traits in humans and in experimental genetics has much in common with the problems discussed above, but many details are quite different–particularly when one considers various breeding designs available in experimental genetics (cf. Lander and Botstein, 1989; Dupuis and Siegmund, 1998).

(iii) The normal approximation suggested in Section 1 is adequate for the simple case of half siblings discussed there, because under the hypothesis of no linkage (11) is symmetrically distributed. In general, particularly when pedigrees contain more than two affecteds or distant affected relatives, the statistic is not symmetrically distributed and the normal approximation can be very poor. For

example, for first cousins the probability of identity by descent at an arbitrary locus is 1/4, so the statistic corresponding to (11) has a distribution skewed to the right; and the approximation (13) is anti-conservative. While it is possible to give approximations based directly on (11) or its analogue in more complex cases, these approximations can be onerous to evaluate numerically. A simple modification of (13) is given by Tu and Siegmund (1998). Let γ be the third moment of Z_t under the hypothesis of no linkage and $\theta = [-1+(1+2b\gamma/N^{1/2})^{1/2}]/\gamma$. Then for a *single* chromosome of genetic length ℓ

$$P\{\max_{0\le i\Delta<\ell} Z_{i\Delta} \ge b\}$$

$$\approx [1-\Phi(b)]\exp(\gamma b^3/6N^{1/2})+\nu\beta\ell b[2\pi(1+\gamma\theta)]^{-1/2}\exp[-N\theta^2(1+2\gamma\theta/3)/2], \quad (20)$$

where $\nu = \nu[b(2\beta\Delta)^{1/2}]$. Note that $\theta \sim b/N^{1/2}$ as either $N \to \infty$ or $\gamma \to 0$, and then (20) reduces to (13). An application of the analogous extension of (18) was described at the end of Section 6.

Acknowledgement. This research was partly supported by NSF Grant DMS-9704324 and by NIH Grant 5 R01 HG00898.

References

Cobb, G.W. (1978). The problem of the Nile: conditional solution to a change-point problem, *Biometrica* 62, 243-251.

Dupuis J., Brown P., Siegmund D. (1995). Statistical methods for linkage analysis of complex traits from high resolution maps of identity by descent. *Genetics* 140, 843-856

Dupuis J. and Siegmund, D. (1998). Statistical methods for mapping quantitative trait loci from a dense set of markers, submitted for publication.

Feingold E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *J. Appl. Probab.* 30, 766-779

Feingold, E., Brown, P.O., Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent, *Am. J. Hum. Genetics*, 53, 234-251

Fisher, R. A. (1934). Two new properties of mathematical likelihood, *Proc. Roy. Soc. A* 144 285-307.

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M. (1996). *An Introduction to Genetic Analysis*, 6th edition, W.H. Freeman and Company, New York.

Kruglyak L., Daly M.J., Reeve-Daly M.P., Lander E.S. (1996). Parametric and non-parametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics* 58, 1347-1363.

Kruglyak, L. and Lander, E.S. (1995). High-resolution genetic mapping of complex traits, *Am. J. Hum. Genet.* 56, 1212-1223.

Lander, E.S. and Botstein, D. (1986). Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphis, *Proc. Nat. Acad. Sci. USA* 83, 7353-7357.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* 121, 185-199.

Lander, E. S. and Schork, N.J. (1994). *Genetic Dissection of complex traits*, Science 265, 2037-2048.

Mahtani, M.M., Widen, E., Lehto, M., Thomas, J., McCarthy, M., Brayer, J., Bryant, B., Chan, G., Daly, M., Forsblom, C., Kanninen, T., Kirby, A., Kruglyak, L., Munnelly, K., Parkkonen, M., Reeve-Daly, M.P., Weaver, A., Brettin, T., Duyk, G., Lander, E.S. and Groop, L.C. (1996). Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families, *Nature Genetics* 14, 90-94.

Morahan, G., Huang, D., Tait, B.D., Colman, P.G., and Harrison, L.C. (1996). Markers on distal chromosome 2q linked to insulin-dependent diabetes mellitus, *Science* 272, 1811-1813.

Risch, N. (1990a,b,c). Linkage strategies for genetically complex traits I, II, III. The power of affected relative pairs, *Am. J. Hum. Genetics* 46, 222-228, 229-241, 242-253.

Siegmund, D. (1988). Approximate tail probabilities for the maxima of some random fields, *Ann. Probab.* 16, 487-501.

Siegmund, D. (1989). Confidence sets in change-point problems, *International Statistical Review* 56, 31-48.

Teng, J. and Siegmund, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers, to appear in *Biometrics*.

Thompson, E. (1994). Monte Carlo likelihood in genetic mapping, *Statist. Sci.* 9, 355-366.

Tu, I.-P. and Siegmund, D. (1998). The maximum of a function of a Markov chain and application to linkage analysis, submitted for publication.

Woodroofe, M. (1976). Frequentist properties of Bayesian sequential tests, *Biometrika* 63, 101–110.

David Siegmund
Department of Statistics
Sequoia Hall
390 Serra Mall
Stanford University
Stanford, CA 94305
USA

Brownian Motion and Random Obstacles

Alain-Sol Sznitman

Abstract. The investigation of Brownian motion and random obstacles exhibits a rich phenomenology and displays paradigms which appear in several other areas of random media. We provide here a brief survey of some recent developments.

1991 Mathematics Subject Classification: 60K40, 82D30

0. Introduction

Much effort has been devoted to the investigation of random media over the last two decades. This field offers a broad selectio n of surprising effects and represents a mathematical challenge. The above applies in particular to the topic of Brownian motio n and random obstacles, which has given rise to new ideas, results and techniques. We shall now explain what the subject is abo ut.

A common example of random obstacles are the soft Poissonian potentials:

$$V(x,\omega) = \sum_i W(x - x_i),\ x \in \mathbb{R}^d\ , \tag{0.1}$$

where $\omega = \sum_i \delta_{x_i}$ is a typical cloud configuration for the Poisson measure $\mathbb{P}$ with constant intensity $\nu > 0$, and $W(\cdot)$ is a bounded measurable nonnegative function, compactly supported and not a.e. equal to 0. Of central i nterest is the investigation of the interaction of Brownian motion with the random obstacles. Several path measures of interest arise in this context, for instance

- Brownian motion in a Poissonian potential, described by:

$$Q_{t,\omega} = \frac{1}{S_{t,\omega}}\ \exp\Big\{ -\int_0^t V(Z_s,\omega)ds\Big\} P_0,\ \text{ (quenched measure)}, \tag{0.2}$$

with ω a $\mathbb{P}$-typical cloud configuration, $Z_\cdot$ the canonical d-dimensional Brownian motion, P_0 the Wiener measure, $S_{t,\omega}$ the normalizing constant,

$$Q_t = \frac{1}{S_t}\ \exp\Big\{ -\int_0^t V(Z_s,\omega)ds\Big\} P_0 \otimes \mathbb{P},\ \text{ (annealed measure)}\ , \tag{0.3}$$

with S_t the normalizing constant,

- Brownian crossings in a Poissonian potential, described by:

$$\hat{P}^{\lambda}_{x,\omega} = \frac{1\{H(x) < \infty\}}{e_\lambda(0,x,\omega)} \exp\Big\{ - \int_0^{H(x)} (\lambda + V(Z_s,\omega))ds \Big\} P_0, \quad \text{(quenched measure)}, \tag{0.4}$$

with ω as in (0.2), $\lambda \geq 0$, $x \in \mathbb{R}^d$, $H(x)$ the entrance time of $Z_\cdot$ in the unit ball around x, $e_\lambda(0,x,\omega)$ the normalizing constant,

$$\hat{P}^{\lambda}_{x} = \frac{1\{H(x) < \infty\}}{\bar{e}_\lambda(x)} \exp\Big\{ - \int_0^{H(x)} (\lambda + V(Z_s,\omega))ds \; g \; P_0 \otimes \mathbb{P}, \quad \text{(annealed measure)}, \tag{0.5}$$

with $\bar{e}_\lambda(x)$ the normalizing constant.

Trapping problems provide natural interpretations for these path measures. In this light, $V(x,\omega)$ can be viewed as the random rate of absorption at location x for a particle diffusing in the environment ω. Thus (0.2), (0.3) govern the so-calle d quenched and annealed behaviors of a particle conditioned to survive absorption up to a (long) time t, whereas (0.4), (0.5) govern the quenched and annealed behaviors of a particle conditioned to perform a (long) crossing without being absorbed. Ther e are other physical interpretations, and for instance (0.2) also comes as a model of "flux lines in dirty-high-temperature su perconductors", cf. Section 4.6.3 of Krug [13], or Krug-Halpin Healy [14]. In this case t represents the transvers al thickness of a material with "columnar defects", rather than time. Discrete analogues of the above path measures also aris e in the literature, see for instance Bolthausen [3], Khanin et al. [12]. It may be helpful to mention that quenched measures describe the evolution in a $\mathbb{P}$-typical environment of a particle starting at the origin, whereas for the annealed m easures the $\mathbb{P}$-integration should be viewed as the result of an ergodic average over the starting point of the particle. It i s a recurrent theme of random media that quenched and annealed behaviors can be substantially different.

I. Normalizing constants for (0.2), (0.3)

Analyzing the principal asymptotic behavior of normalizing constants is a first step in the understanding of the path measures attached to Brownian motion in a Poissonian potential.

With the help of the Feynman-Kac formula, the normalizing constants $S_{t,\omega}$ and S_t can respectively be expressed as:

$$S_{t,\omega} = u_\omega(t,0) \;\text{ and }\; S_t = \mathbb{E}[u_\omega(t,0)]\,, \tag{1.1}$$

where $u_\omega(t,x)$ is the bounded solution of

$$\begin{cases} \partial_t\, u_\omega &= \frac{1}{2}\,\Delta u_\omega - V u_\omega\,, \\ u_\omega(0,x) &= 1\,. \end{cases} \tag{1.2}$$

Their principal asymptotic behaviors as $t \to \infty$, are governed by:

$$\mathbb{P}-\text{a.s.}, \quad S_{t,\omega} = \exp\{-c(d,\nu)\, t(\log t)^{-2/d}(1+o(1))\}\ , \tag{1.3}$$

$$S_t \quad = \exp\{-\widetilde{c}(d,\nu)\, t^{\frac{d}{d+2}}(1+o(1))\}\ . \tag{1.4}$$

The constants c and $\widetilde{c}$ are "explicit", and independent of the specific choice of $W(\cdot)$ in (0.1). If $\lambda(U)$ and $|U|$ respectively denote the principal Dirichlet eigenvalue of $-\frac{1}{2}\,\Delta$ in U and the volume of U, one has:

$$c(d,\nu) = \lambda(B(0,R_0)), \quad \text{with} \quad R_0 = \Big(\frac{d}{\nu|B(0,1)|}\Big)^{1/d}, \quad \text{whereas} \tag{1.5}$$

$$\begin{aligned} \widetilde{c}(d,\nu) = \quad & \inf_{U \text{ open}} \{\nu|U| + \lambda(U)\} = \nu|B(0,\widetilde{R}_0)| + \lambda(B(0,\widetilde{R}_0)), \quad \text{with} \\ & \widetilde{R}_0 = \Big(\frac{2\lambda(B(0,1))}{d\nu\,|B(0,1)|}\Big)^{\frac{1}{d+2}}\ . \end{aligned} \tag{1.6}$$

The annealed asymptotics (1.4) goes back to Donsker-Varadhan [5], where it was obtained as an application of large deviati on theory for occupation times of Brownian motion on a torus. Both asymptotics have also been derived through the analysis of p rincipal Dirichlet eigenvalues of $-\frac{1}{2}\,\Delta + V(\cdot,\omega)$ in large boxes, and the method of enlargement of obstacl es, cf. [24], [33], [36]. Sharper versions of (1.3), (1.4) can also be found in [36].

Intuitively for the quenched asymptotics, the contribution in the Feynman-Kac formula

$$S_{t,\omega} = E_0\Big[\exp\Big\{-\int_0^t V(Z_s,\omega)ds\Big\}\Big] \tag{1.7}$$

of Brownian paths going to some obstacle-free ball of radius of order $R_0(\log t)^{1/d}$, typically occurring within distance slightly less than t from the origin, and staying there up to time t, has the principal asymptotic behavior (1.3). On the other hand for the annealed asymptotics, the contribution in the representation

$$S_t = \mathbb{E} \otimes E_0\Big[\exp\Big\{-\int_0^t V(Z_s,\omega)ds\Big\}\Big]\ , \tag{1.8}$$

of largely deviant environments, for which an obstacle-free ball of radius of order $\widetilde{R}_0\, t^{\frac{1}{d+2}}$ contains the origin, and of Brownian trajectories, which stay in the ball up to time t, has the principal behavior (1.4). Of course, under standing whether and up to what point these heuristics truly govern the quenched and annealed path measures (0.2), (0.3) is qui te another matter. As it turns out, the loose concept of *pockets of low local principal Dirichlet eigenvalue* for $-\frac{1}{2}\,\Delta + V(\cdot,\omega)$, plays an important role in the analysis of (0.2), (0.3). The predominance of atypical

"pocket s of abnormally low eigenvalues" locally describing a system is a recurrent paradigm of random media, which for instance shows up in models of intermittency, cf. Gärtner-Molchanov [8], [9], Molchanov [17], in random walks in random envi ronment, cf. [4], [19], [20], [35], or in stochastic dynamics of spin systems with random interactions, cf. [16] and references therein.

II. Pinning effect and confinement property

The large t behavior of the quenched path measure $Q_{t,\omega}$ is governed by a "competition" between the various "pockets o f low local eigenvalues", resulting in a pinning effect: the path tends to get attracted to near minima of a certain random va riational problem. The discussion of the real pinning effect would go beyond the scope of this expository article, and we restr ict here to a simplified version. We refer to [32] or [36] for the "real story". We denote by $\lambda_\omega(U)$ the principal D irichlet eigenvalue of $-\frac{1}{2}\ \Delta + V(\cdot,\omega)$ in U, and for sufficiently small $\chi > 0$, consider the random function on $\mathbb{R}^d$.

$$F_t(x,\omega) = \alpha_0(x) + t\lambda_\omega(B(x,R_t))\ , \tag{2.1}$$

with $R_t = \exp\{(\log t)^{1-\chi}\}$, a "small scale" growing slower than any positive power of t, and $\alpha_0(\cdot)$ a certain deterministic norm, the so-called quenched 0-th Lyapunov coefficient, see Section IV below, (the role of $\alpha_0(\cdot)$ is somewhat cosmetic in the simplified pinning effect we discuss here). Minimizing $F_t(\cdot,\omega)$ induces a competitio n between distance to the origin and occurrence of pockets of low local eigenvalues. One can show, cf. [32], [36], that

$$\mathbb{P}\text{-a.s., } \inf F_t(\cdot,\omega) \sim c(d,\nu)\, t(\log t)^{-2/d}, \ \text{ as } \ t \to \infty\ , \tag{2.2}$$

with $c(d,\nu)$ as in (1.5). Defining a skeleton of near minima of $F_t(\cdot,\omega)$ via

$$\mathcal{L}_{t,\omega} = \Big\{ x \in \frac{1}{\sqrt{d}}\ \mathbb{Z}^d,\ F_t(x,\omega) \le \inf F_t(\cdot,\omega) + t(\log t)^{-\chi-\frac{2}{d}} \Big\}\ , \tag{2.3}$$

it can be shown that this set "lies almost at distance t" from the origin. The (simplified) pinning effect asserts that

THEOREM: *For small* $\chi > 0$,

$$\mathbb{P}\text{-}a.s.,\ \lim_{t\to\infty}\ Q_{t,\omega}(C) = 1, \ \ where \tag{2.4}$$

$$C = \ \{Z_\cdot \ comes\ before\ time\ t\ within\ distance\ 1\ of\ some\ x \in \mathcal{L}_{t,\omega}\ from\ which\ it\ then\ does\ not\ move\ further\ away\ than\ R_t\ up\ to\ time\ t\}\ . \tag{2.5}$$

As a by-product of the proof one also has the refinement of (1.3):

$$\mathbb{P}\text{-a.s., } \log S_{t,\omega} + \inf F_t(\cdot,\omega) = o(t(\log t)^{-\chi-\frac{2}{d}})\ . \tag{2.6}$$

The true pinning effect is substantially sharper but involves certain random scales which would take too long to introduce here . In particular in the one-dimensional

case it can be shown that for $\epsilon > 0$, with $\mathbb{P} \times Q_{t,\omega}$-probability ten ding to 1 as $t \to \infty$, $Z_\cdot$ gets pinned within time $\epsilon\, t$ in scale $t(\log t)^{-3}$ within an interva l of length $2(\log t)^{2+\epsilon}$, cf. [32], [36].

Loosely speaking, in the quenched situation the particle "goes the extra mile" to find an adequate pocket of low local eigenv alue. The annealed situation is quite different and favours a "good location" for the starting point of the path which then t ends to remain "confined" there. For instance in the case of hard obstacles, i.e. for the path measure

$$Q_t = \mathbb{P} \otimes P_0[\,\cdot\,|T > t]\,, \tag{2.7}$$

with T the entrance time of $Z_\cdot$ in the obstacle set $\bigcup_i x_i + K$, $\omega = \sum_i \delta_{x_i}$ and K a fixed nonpola r compact set, one has the confinement property:

THEOREM: *For any $d \geq 1$,*

$$\lim_{t\to\infty} Q_t[\sup_{0\leq u\leq t} |Z_u| \leq 2\, t^{\frac{1}{d+2}}(\widetilde{R}_0 + \epsilon(t))] = 1\,, \tag{2.8}$$

with $\widetilde{R}_0$ as in (1.6), and $\epsilon(t)$ a suitable function tending to 0, when t tends to ∞.

Thus the path "typically lives in scale $t^{\frac{1}{d+2}}$ under Q_t". The result is considerably harder to prove when $d \geq 2$. The two-dimensional case goes back to [26]. The case of dimension $d \geq 3$ was proved by Povel [21], who used a recent version of the method of enlargement of obstacles (cf. next section), and certain isoperimetric controls of R.R. Hall [?], which play the role of the Bonnesen's inequality in the two-dimensional proof. In fact in the two-dimensional case, it was proved in [26] that

THEOREM: *($d = 2$)*

$$\begin{aligned}&\textit{There exists a measurable map } D_t(\omega),\ B(0, t^{1/4}\,(\widetilde{R}_0 + \epsilon(t)))\textit{-valued,}\\ &\textit{such that with } Q_t\textit{-probability tending to } 1,\ \textit{as } t \to \infty,\ Z_{[0,t]}\ \textit{is}\\ &\textit{included in } B(D_t, t^{1/4}(\widetilde{R}_0 + \epsilon(t)))\ \textit{and no obstacle fall in}\\ &B(D_t, t^{1/4}(\widetilde{R}_0 - \epsilon(t)))\,.\end{aligned} \tag{2.10}$$

In the case of the simple random walk on $\mathbb{Z}^2$, Bolthausen proved in [3] a version of this result using a refined version of D onsker-Varadhan's large deviation principles. It is also possible to obtain further information on the "spherical clearing" w here the process lives, cf. Schmock [23], when $d = 1$, [26], when $d = 2$, and [21], when $d \geq 3$:

$$\begin{aligned}&\text{As } t \to \infty,\ t^{-\frac{1}{d+2}} Z_{\cdot\, t^{\frac{2}{d+2}}} \text{ converges in law under } Q_t, \text{ to the}\\ &\text{mixture with weight } \psi(x)/\textstyle\int \psi \text{ of the laws of Brownian motion}\\ &\text{starting from 0 conditioned not to exit } B(x, \widetilde{R}_0), \text{ with } \psi \text{ the}\\ &\text{principal Dirichlet eigenfunction of } -\tfrac{1}{2}\,\Delta \text{ in } B(0, \widetilde{R}_0)\,.\end{aligned} \tag{2.11}$$

III. THE METHOD OF ENLARGEMENT OF OBSTACLES

As mentioned above, in many questions related to Brownian motion in a Poissonian potential, the analysis of local principal Dir ichlet eigenvalues of $-\frac{1}{2}\,\Delta + V(\cdot,\omega)$ plays in important role. Indeed these numbers control in a very qu antitative fashion the decay properties of the Dirichlet-Schrödinger semigroup. This is illustrated by the estimat e:

$$\sup_x E_x\Big[\exp\Big\{-\int_0^t V(Z_s,\omega)ds\Big\},\, T_U > t\Big] \le c(1+(\lambda_\omega(U)t)^{d/2})\, e^{-\lambda_\omega(U)t}\,, \tag{3.1}$$

with c a merely dimension dependent constant and T_U the exit time of $Z_\cdot$ from U, cf. [36]. The method of enlargement o f obstacles in particular provides an efficient way of deriving uniform controls on the numbers $\lambda_\omega(U)$ close to 0 (i .e. the bottom of the spectrum of $-\frac{1}{2}\,\Delta + V(\cdot,\omega)$ in $\mathbb{R}^d$), as U and ω vary. The rough idea is to remodel the region $V > 0$, and construct a coarse grained picture with lower combinatorial complexity than the original clo ud configuration, which for probabilistic purpose is simpler to analyze, but still has principal eigenvalues close to the origi nal objects. This remodeling of the region $V > 0$ brings into play a trichotomy of $\mathbb{R}^d$. In a first region, true obstacles a re quickly sensed by Brownian motion, and obstacles can be "enlarged" by imposing Dirichlet condition on this set. A second r egion where obstacles are insufficiently present and where enlargement of obstacles could possibly influence eigenvalues is sho wn to have little volume and thus little effect on probabilistic estimates. The third and last region receives no point of the cloud. In a sense, this parallels the trichotomy associated to any compact set K by considering the set of regular points of K, the set of irregular points of K and the complement of K.

Specifically after scaling the problem so that ϵ represents the size of the true obstacles, 1 the size of the pocket s of interest in the scaled cloud configurations (still denoted by ω), one constructs a density set $\mathcal{D}_\epsilon(\ o)$ where obstacles are enlarged and a bad set $\mathcal{B}_\epsilon(\omega)$ where obstacles are untouched, so that:

(3.2)
- i) $\mathcal{D}_\epsilon(\omega),\ \mathcal{B}_\epsilon(\omega),\ \mathbb{R}^d\backslash(\mathcal{D}_\epsilon(\omega)\cup\mathcal{B}_\epsilon(\omega))$; partition $\mathbb{R}^d$,
- ii) no point of ω falls in $\mathbb{R}^d\backslash(\mathcal{D}_\epsilon(\omega)\cup\mathcal{B}_\epsilon(\omega))$,
- iii) for each box C of size 1, the maps $\omega \to C\cap\mathcal{D}_\epsilon(\omega)$ and $\omega\to C\cap\mathcal{B}_\epsilon(\omega)$ have range of cardinality smaller then $2^{\epsilon^{-d\beta}}$, with $\beta\in(0,1)$ a fixed number.

Denoting by $V_\epsilon(\cdot,\omega)=\sum_i \epsilon^{-2}\,W(\frac{\cdot-x_i}{\epsilon})$ the scaled potential, the construct ion can be done so that for a suitable $\alpha\in(0,\beta)$, Brownian motion, when starting on $\overline{\mathcal{D}}_\epsilon(\omega)$, strongly feels the obstacles before moving at distance ϵ^α:

THEOREM A_0: *(pointwise absorption estimate). There exists $\rho_0>0$, such that*

$$\overline{\lim_{\epsilon\to 0}}\,\epsilon^{-\rho_0}\sup_{\omega,x\in\omega verline\mathcal{D}_\epsilon(\omega)} E_x\Big[\exp\Big\{-\int_0^{H_{\epsilon^\alpha}} V_\epsilon(Z_s,\omega)ds\Big\}\Big]<1,\ \ \textit{with} \tag{3.3}$$
$$H_{\epsilon^\alpha}=\inf\{s\ge 0,\ |Z_s-Z_0|\ge\epsilon^\alpha\}\,,$$

and on the other hand the bad set has small volume:

THEOREM B: *(volume estimate)*

$$\exists \kappa > 0, \quad \overline{\lim_{\epsilon \to 0}} \sup_{\text{RIPTSIZE } C \text{ BOX OF SIZE } 1, \omega} \epsilon^{-\kappa} \, |\mathcal{B}_\epsilon(\omega) \cap C| < 1 \, . \tag{3.4}$$

The construction of the trichotomy (3.3) i) relies on a type of quantitative Wiener test involving a series of capacities of a skeleton of the true obstacles at scales intermediate between ϵ^β and ϵ^α. In a sense (3.4), (3.5) parallels the Wiener test characterization of regular points of a compact set and the Kellog-Evans theorem on the smallness of the set of ir regular points of a compact set. As an application of the pointwise absorption estimates (3.3) one can in particular obtain eig envalue estimates:

THEOREM A: *(eigenvalue estimate)*

(3.5)

$$\exists \rho > 0, \ \forall M > 0, \ \lim_{\epsilon \to 0} \epsilon^{-\rho} \sup_{\omega, U} (\lambda^\epsilon psilon_\omega(U \backslash \overline{\mathcal{D}}_\epsilon(\omega)) \wedge M - \lambda^\epsilon_\omega(U) \wedge M) = 0 \, ,$$

with $\lambda^\epsilon_\omega(O) =$ *principal Dirichlet eigenvalue of* $-\frac{1}{2} \Delta + V_\epsilon(\cdot, \omega)$ *in* O.

In other words this shows that in the asymptotic regime, provided $\lambda^\epsilon_\omega(U)$ has value of order unit, an addition al Dirichlet condition on $\overline{\mathcal{D}}_\epsilon(\omega)$ does not essentially increase the principal eigenvalue.

The method of enlargement of obstacles has numerous applications to the quenched and annealed situation, cf. [36]. The method e asily applies to non-Poissonian obstacles (uniformity of controls in ω is very handy), cf. [28], to shrinking obstacles, cf. [25], see also [2], to confidence intervals on principal eigenvalues, cf. [33], see also [39]. A version of the method in the discrete setting can be found in Antal [1]. Recently L. Erdös applied in [6] a version of the method to the study o f the Lifschitz tail effect for the density of states of the magnetic Schrödinger operator with Poissonian obstacles.

IV. LYAPUNOV NORMS

The technique of Lyapunov norms has been very helpful in the investigation of "off-diagonal" properties of the path measures (0.2), (0.3), in particular in the derivation of large deviation principles governing the location of Z_t. The Lyapunov norms describe the principal exponential decay of the normalizing constants in (0.4), (0.5). In a one-dimensional setting, in the co ntext of wave propagation in random media, they can be traced back to the work of Gärtner and Freidlin, cf. Chapter 7 of Frei dlin [7].

At the heart of the method lies the fact that the functions $e_\lambda(x, y, \omega)$ satisfy an almost supermultiplicative property and still contain much information about Brownian motion in a Poissonian potential. An important role is played by certain sha pe theorems (analogous to shape theorems of first passage percolation, cf.

Kesten [11]), which construct two families of norms on $\mathbb{R}^d$, $\beta_\lambda(\cdot) \leq \alpha_\lambda(\cdot)$, $\lambda \geq 0$, the annealed and quenched Lyapunov coeffici ents:

$$\text{(4.1)} \qquad \mathbb{P}\text{-a.s. for } M > 0, \ \lim_{x\to\infty} \sup_{0\leq\lambda\leq M} \frac{1}{|x|} \, | - \log e_\lambda(0,x,\omega) - \alpha_\lambda(x)| = 0 \ ,$$

$$\text{(4.2)} \qquad \text{for } M > 0, \ \lim_{x\to\infty} \sup_{0\leq\lambda\leq M} \frac{1}{|x|} \, | - \log \, tog(\mathbb{E}[e_\lambda(0,x,\omega)]) - \beta_\lambda(x)| = 0 \ .$$

These shape theorems are quite robust and one can replace in (4.1), (4.2), $e_\lambda(0,x,\omega)$ by the λ-Green function $g_\lambda(0,x,\omega)$, or $e_\lambda(x,0,\omega)$, or $\exp\{-d_\lambda(0,x,\omega)\}$, with d_λ certain natural random distance functions (in general nongeodesic) constructed with the e_λ, cf. [36]. The Lyapunov coefficients enter several large d eviation theorems, cf. [29], [30], [31], as well as the random variational problem of the pinning effect. For instance when $arphi(t) \to \infty$,

$$\text{(4.3)} \qquad \begin{aligned} &\mathbb{P}\text{-a.s. under } Q_{t,\omega}, Z_t/\varphi(t) \text{ satisfies a large deviation principle} \\ &\text{at rate } \varphi(t), \text{ with rate function:} \\ &\text{i)} \quad \alpha_0(x), \text{ if } \varphi(t) = t(\log t)^{-2/d}, \text{ cf. [31]}, \\ &\text{ii)} \quad \alpha_0(x), \text{ if } t(\log t)^{-2/d} << \varphi << t, \text{ cf. [29]}, \\ &\text{iii)} \quad I(x) = \sup_{\lambda\geq 0} (\alpha_\lambda(x) - \lambda), \text{ if } \varphi(t) = t, \text{ cf. [29]}. \end{aligned}$$

Similar results hold under the annealed measure Q_t, when $d \geq 2$, with $t^{\frac{d}{d+2}}$ in place of $t(\log t)^{-2/d}$ and $\beta_\lambda(\cdot)$ in place of $\alpha_\lambda(\cdot)$, (the one-dimensional case is singular, cf. Povel [22]). In the discrete setting (4.3) iii) has been proved by Zerner in [40]. In fact the above strategy also applies in the context of random walks in random environments, cf. Zerner [41]. This is especially interesting since there are few mathematical results on this model.

The understanding of crossing Brownian motion in a Poissonian potential, see (0.4), (0.5), is so far rather primitive. However recently for rotationally invariant truncated Poissonian potentials, Wüthrich has been able to relate in [37], the fluctuatio n properties of $-\log e_\lambda(0,x,\omega)$ to transversal fluctuations of the path under the path measure (0.4). In a slightly d ifferent situation ("point to line" model), he was also able to obtain a result about the superdiffusive nature of transversa l fluctuations, cf. [38]. This is qualitatively similar to what happens in first passage percolation, cf. Licea-Newman-Piza [15], Newman-Piza [18].

REFERENCES

[1] Antal P.: "Enlargement of obstacles for the simple random walk". *Ann. Probab.*, 23(3):1061-1101, 1995.

[2] Bolthausen E.: "On the volume of the Wiener sausage". *Ann. Probab.*, 18:1576-1582, 1990.

[3] Bolthausen E.: "Localization of a two-dimensional random walk with an attractive path interaction". *Ann. Pro bab.*, 22:875-918, 1994.
[4] Dembo A., Peres Y. and Zeitouni O.: "Tail estimates for one-dimensional random walk in random environment". *Comm. Math. Phys.*, 181:667-683, 1996.
[5] Donsker M. and Varadhan S.R.S.: "Asymptotics for the Wiener sausage". *Comm. Pure Appl. Math.*, 28:525-565, 1 975.
[6] Erdös L.: "Lifschitz tail in a magnetic field: the non-classical regime". Preprint, 1997.
[7] Freidlin M.: "Functional integration and partial differential equations". Annals of Mathematics Studies 109, Prin ceton University Press, 1985.
[8] Gärtner J. and Molchanov S.A.: "Parabolic problems for the Anderson model I". *Comm. Math. Phys.*, 28:525-6 55, 1990.
[9] Gärtner J. and Molchanov S.A.: "Parabolic problems for the Anderson model II". *Probab. Th. Rel. Fields*, 1 11:17-55, 1998.
[10] Hall R.R.: "A quantitative isoperic inequality in n-dimensional space". *J. reine angew. Math.*, 428:161-1 76, 1992.
[11] Kesten H.: "Aspects of first passage percolation". In: Ecole d'été de Probabilités de St. Flour. Lecture Notes in Math. 1180, Springer, 125-264, 1986.
[12] Khanin K.M., Mazel A.E., S.B. Shlosman and Sinai Ya.G.: "Several results related to random walks with random pot entials". Dynkin Festschrift, M.I. Freidlin editor, Birkhäuser, 165-184, 1994.
[13] Krug J.: "Origins of scale invariance in growth processes". *Adv. in Phys.*, 46(2):139-282, 1997.
[14] Krug J. and Halpin-Healy T.: Directed polymer in the presence of columnar disorder". *J. Phys. I, France*, 3: 2179-2198, 1993.
[15] Licea C., Newman C.M. and Piza M.S.T.: "Superdiffusivity in first-passage percolation". *Prob. Theory Rel. F ields*, 106:559-591, 1996.
[16] Martinelli F.: "Lectures on spin dynamics for discrete spin models". In: Ecole d'Été de St Flour, *Lectu re Notes in Math.*, Springer, New York, 1998.
[17] Molchanov S.A.: *Lectures on random media*. Lecture Notes in Math. 1581, Ecole d'Été de St Four XXII-1992, Editor P. Bernard, Springer, New York, 1994.
[18] Newman C.M. and Piza M.S.T.: "Divergence of shape fluctuations in two dimensions". *Ann. Prob.*, 23:977-1005 , 1995.
[19] Pisztora A. and Povel T.: "Large deviation principle for random walk in a quenched random environment in the low speed regime". Preprint, 1997.
[20] Pisztora A., Povel T. and Zeitouni O.: "Precise large deviation estimates for one-dimensional random walk in ran dom environment". Preprint, 1997.
[21] Povel T.: "Confinement of Brownian motion among Poissonian obstacles in $\mathbb{R}^d$, $d \geq 3$". Preprint, 1997
[22] Povel T.: "Critical large deviations of one-dimensional Brownian motion with a drift in a Poissonian potential". *Ann. Probab.*, 25:1735-1773, 1997.
[23] Schmock U.: "Convergence of the normalized one-dimensional Wiener sausage path measure to a mixture of Brownian taboo processes". *Stochastics*, 29:171-183, 1990.
[24] Sznitman A.S.: "Lifschitz tail and Wiener sausage, I, II". *J. Funct. Anal.*, 94:223-246, 247-272, 199 0.

[25] Sznitman A.S.: "Long time asymptotics for the shrinking Wiener sausage". *Comm. Pure Appl. Math.*, 43:809 -820, 1990.
[26] Sznitman A.S.: "On the confinement property of Brownian motion among Poissonian obstacles". *Comm. Pure Appl. Math.*, 44:1137-1170, 1991.
[27] Sznitman A.S.: "Brownian asymptotics in a Poissonian environment". *Probab. Th. Rel. Fields*, 95:155-174, 1993.
[28] Sznitman A.S.: "Brownian survival among Gibbsian traps". *Ann. Probab.*, 21(1):490-509, 1993.
[29] Sznitman A.S.: "Shape theorem, Lyapounov exponents and large deviations for Brownian motion in a Poissonian potential". *Comm. Pure Appl. Math.*, 47:1655-1688, 1994.
[30] Sznitman A.S.: "Annealed Lyapunov exponents and large deviations in a Poissonian potential. I, II". *Ann. scient. Ec. Norm. Sup., 4ème série*, 28:345-370, 371-390, 1995.
[31] Sznitman A.S.: "Quenched critical large deviations for Brownian motion in a Poissonian potential". *J. F unct. Anal.*, 131(1):54-77, 1995.
[32] Sznitman A.S.: "Brownian confinement and pinning in a Poissonian potential I, II". *Probab. Th. Rel. F ields*, 105:1-30, 31-56, 1996.
[33] Sznitman A.S.: "Capacity and principal eigenvalues: The method of enlargement of obstacles revisited". *Ann. Probab.*, 25(3):1180-1209, 1997.
[34] Sznitman A.S.: "Fluctuations of principal eigenvalues and random scales". *Comm. Math. Phys.*, 189:337-363, 1997.
[35] Sznitman A.S.: "Slowdown and neutral pockets for a random walk in random environment". Preprint, 1998.
[36] Sznitman A.S.: "Brownian motion, obstacles and random media". Springer, New York, 1998.
[37] Wüthrich M.V.: "Scaling identity for crossing Brownian motion in a Poissonian potential". Preprint, 1997.
[38] Wüthrich M.V.: "Superdiffusive behaviour of two-dimensional Brownian motion in a Poissonian potential". Preprint, 1997.
[39] Yurinski V.V.: "Spectrum bottom and largest vacuity". Preprint, 1997.
[40] Zerner, M.P.W.: "Directional decay of the Green's function for a random nonnegative potential on $\mathbb{Z}^n$". it Ann. Appl. Probab., 8(1):246-280, 1998.
[41] Zerner, M.P.W.: "Lyapunov exponents and quenched large deviation for multidimensional random walk in random envir onment". Preprint, 1997.

Alain-Sol Sznitman
Departement Mathematik
ETH-Zentrum
CH-8092 Zürich
Switzerland

Within and Beyond the Reach of Brownian Innovation

Boris Tsirelson

Abstract. Given a system whose time evolution is random, we often try to describe it as a deterministic system under independent random influences. Doing so, we reduce complicated statistical correlations to a complicated but deterministic mechanism, and a stochastic but uncorrelated noise. That is the idea of innovation. The corresponding mathematics is surprisingly interesting.

1991 Mathematics Subject Classification: 60G07; 60H10, 60J65.
Keywords and Phrases: innovation, filtration, cosiness, noise.

1. The name of the game

An *innovation* is a *real-time transformation* of a *noise* into a given *random process.*

Out of the four terms, only one, "random process", is standard. The notion of a real-time transformation was introduced repeatedly, and used under various names: "lifting" (of a filtered probability space) [19, (7.1–7.3)], "Hypothèse ($\mathcal{H}$)" [7, Sect. 2.4], "extension" (of a filtered probability space) [22, Chap. 2, Def. 7.1], [8, Def. 6.1], with no name [44, 17.3.1(a)], [2, Lemma 7(c)], "morphism" (from one filtration to another) [33, Def. 1.1], "immersion" (of one filtration into another) [4], "orthogonal factor" (of a reverse filtration) [15, Sect. 2]. My favorite "real-time transformation" appeared in [33].

A noise in the discrete-time framework amounts to an independent sequence (of random variables or σ-fields), or a product (of a sequence of probability spaces). For continuous time, the classical white noise is a special case of a noise as defined in [34, Def. 1.1]; see also "factored probability spaces" [13], "measure factorizations" [36, Def. 1.2], and "product measures" (on a factorized Borel space) [36, Def. 2.4].

Innovation processes are well-known in filtering theory (see [5, Sect. 8]). A far-reaching generalization is the "innovation" introduced here. In the discrete-time framework, innovation appeared as "standard extension" (of a reverse filtration) [8, p. 885], "generating parametrization" [28, Sect. 2], [26, Def. 2.1], "substandard representation" [15, Sect. 2]. My favorite "innovation" appeared in [26].

2. Trivial cases

Let μ be a probability measure on a space $\mathcal{X}$. (Usually $\mathcal{X} = \mathbb{R}$ or $\mathbb{R}^n$, but it may be a finite set, a complete separable metric space, a standard Borel space.) Every such μ can be represented as the image of the Lebesgue measure $\mathcal{U}(0,1)$ under a measurable map $f : (0,1) \to \mathcal{X}$. Let U be a random variable distributed

uniformly on $(0,1)$ (in symbols $U \sim \mathcal{U}(0,1)$), then $f(U) \sim \mu$. Of course, μ does not determine f uniquely; $f(g(U)) \sim \mu$ for all measure preserving $g : (0,1) \to (0,1)$. So, every $\mathcal{X}$-valued random variable Y is distributed like some $X = f(U)$.

Consider a discrete time random process $Y = (Y_t)_{t\in T}$, assuming for now that T is finite, $T = \{1,\dots,n\}$; thus Y is just n random variables $Y_1,\dots,Y_n$, and its distribution is a measure μ on $\mathcal{X}^n$. Let $U_1,\dots,U_n$ be independent $\mathcal{U}(0,1)$ random variables. Choose $f_1 : (0,1) \to \mathcal{X}$ such that $f_1(U_1)$ is distributed like Y_1. For each $y_1 \in \mathcal{X}$ consider the conditional distribution of Y_2 given that $Y_1 = y_1$ (I omit trivial reservations) and choose $f_2(\cdot, y_1)$ accordingly. Introduce $X_1 = f_1(U_1)$, $X_2 = f_2(U_2, X_1)$, then the pair (X_1, X_2) is distributed like (Y_1, Y_2). Continuing the process, we get functions $f_1,\dots,f_n$ and random variables $X_1,\dots,X_n$ such that

$$\begin{aligned} &X_1 = f_1(U_1),\ X_2 = f_2(U_2, X_1),\ \dots,\ X_n = f_n(U_n, X_{n-1},\dots,X_1)\,, \\ &(X_1,\dots,X_n) \text{ is distributed like } (Y_1,\dots,Y_n)\,. \end{aligned} \tag{2.1}$$

That is the innovation: at a time $t \in T$ the process X takes on a value X_t produced by a deterministic mechanism f_t out of two sources: the past $(X_1,\dots,X_{t-1})$ of the process, and the current value U_t of a noise. Note that each U_t is used only once (formulas like $X_2 = f_2(U_2, U_1, X_1)$ are disallowed), and $U_1,\dots,U_n$ are independent. The uniform distribution of U_t is only conventional; in Sect. 4 we prefer the normal distribution. Note also the large choice available on each stage when constructing $f_1,\dots,f_n$.

Example. Let $(Y_t)_{t\in T}$ be a process with independent increments, having assumed that $\mathcal{X} = \mathbb{R}$ or another group. We may choose an innovation of the form

$$X_t = g_t(U_t) + X_{t-1}\,. \tag{2.2}$$

The simple form (2.2) seems to be decidedly preferable to (2.1) for such processes, which is a delusion, to be refuted in Sect. 3.

The distribution of $X = (X_1,\dots,X_n)$ is the given μ. Consider, however, the joint distribution of X and U. We have

$$\mathbb{E}\left(\varphi(X_1,\dots,X_n)\,\middle|\,U_1,\dots,U_t\right) = \mathbb{E}\left(\varphi(X_1,\dots X_n)\,\middle|\,X_1,\dots X_t\right) \tag{2.3}$$

for all $t = 1,\dots,n$ and all bounded Borel functions $\varphi : \mathcal{X}^n \to \mathbb{R}$. Forecasting the future of the process X, we want to know the past of X only, and not the past of U. In other words, $(X_{t+1},\dots,X_n)$ and $(U_1,\dots,U_t)$ are conditionally independent, given $(X_1,\dots,X_t)$. *

Consider the σ-field $\mathcal{F}_X(t)$ generated by $X_1,\dots,X_t$; clearly, $\mathcal{F}_X(t) \subset \mathcal{F}_U(t)$ for all t, that is, $\mathcal{F}_X \le \mathcal{F}_U$, where $\mathcal{F}_X = \big(\mathcal{F}_X(t)\big)_{t\in T}$ is the filtration generated by X. Writing (2.3) in the form $\mathbb{E}\left(\xi\,\middle|\,\mathcal{F}_U(t)\right) = \mathbb{E}\left(\xi\,\middle|\,\mathcal{F}_X(t)\right)$ for $\mathcal{F}_X(n)$-measurable ξ, note that $\mathbb{E}\left(\xi\,\middle|\,\mathcal{F}_X(t)\right)$ is the general form of an $\mathcal{F}_X$-martingale; so,

$$\mathcal{M}(\mathcal{F}_X) \subset \mathcal{M}(\mathcal{F}_U)\,, \tag{2.4}$$

* Though, (2.1) stipulates more: $(X_{t+1}, U_{t+1},\dots,X_n,U_n)$ and $(U_1,\dots,U_t)$ are conditionally independent, given $(X_1,\dots,X_t)$.

where $\mathcal{M}(\mathcal{F})$ is the set of all $\mathcal{F}$-martingales. Relation (2.4) implies $\mathcal{F}_X \le \mathcal{F}_U$, and is much stronger; try $X_2 = f_2(U_2, U_1, X_1)$ instead of $f_2(U_2, X_1)$ and you'll find (2.4) violated but $\mathcal{F}_X \le \mathcal{F}_U$ is still valid.

The following definition is formulated in terms of processes, but only their distributions are relevant. Still, $T = \{1, \dots, n\}$.

2.6 DEFINITION. A *real-time transformation* of a random process $V = (V_1, \dots, V_n)$ into another process $W = (W_1, \dots, W_n)$ is a two-component process $(V', W') = \big((V'_1, W'_1), \dots, (V'_n, W'_n)\big)$ such that V' is distributed like V, W' is like W, and for each $t = 1, \dots, n$, W'_t is equal to a function of $V'_1, \dots, V'_t$, and two vectors $(V'_1, \dots, V'_t)$ and $(W'_{t+1}, \dots, W'_n)$ are conditionally independent given $(W'_1, \dots, W'_t)$.

Reformulations via (2.3), (2.4) and generalizations for infinite T are left to the reader. Nothing new emerges for an infinite *increasing* sequence of time moments, $t \in T = \mathbb{N} = \{1, 2, 3, \dots\}$. Still, an innovation is constructed step-by-step: f_1, then f_2, and so on *ad infinitum.* The same holds for every countable ordinal number, that is, every countable linearly ordered set T that contains no infinite strictly *decreasing* sequences.

3. Decreasing sequences are highly non-trivial

The following two examples show an astonishing phenomenon: some information appears magically, from thin air; see [25, p. 156], [43, p. 136], [10] and references therein.

The first example: $X_t = \pm 1$ for $t \in \mathbb{Z}$ are i.i.d. equiprobable random signs, $U_t = X_t / X_{t-1}$; then U_t are i.i.d. equiprobable random signs, also. Thus, X is both a process with independent values, and a process with independent increments in the multiplicative group $\{-1, +1\}$. The equality $X_t = U_t X_{t-1}$ should be an innovation of the process X by the noise U. However, it is not; X contains more information than U, since U determines X only up to an overall sign. The missing information should be a kind of initial value, $X_{-\infty}$; however, any function of the germ (tail) of X at $-\infty$ is either constant almost sure, or nonmeasurable, which is the well-known *tail triviality.*

The second example is the "eternal" (stationary) Brownian motion in a circle (or any other compact Lie group). Let $\big(B(t)\big)_{t\in[0,\infty)}$ be the standard Brownian motion in $\mathbb{R}$, and α a random variable, uniform on $(0, 1)$ and independent of $\big(B(t)\big)_{t\in[0,\infty)}$. Consider the complex-valued process $X(t) = \exp\big(2\pi i\alpha + iB(t)\big)$. The process $\big(X(t)\big)_{t\in[0,\infty)}$ is stationary. Therefore, it has a unique (in distribution) extension $\big(X(t)\big)_{t\in\mathbb{R}}$, the eternal motion. Multiplicative increments $U_t = X_t / X_{t-1}$ for $t \in \mathbb{Z}$ should innovate the process $\big(X(t)\big)_{t\in\mathbb{Z}}$. However, they do not, for the same reason as in the first example: they stay invariant under transformations of the form $\big(X(t)\big)_{t\in\mathbb{R}} \mapsto \big(e^{i\varphi} X(t)\big)_{t\in\mathbb{R}}$.

About notation: ergodic people, being more light-hearted toward the time arrow than probabilists, prefer $(X'_1, X'_2, \dots)$, where $X'_1 = X_{-1}, X'_2 = X_{-2}, \dots$, to $(\dots, X_{-2}, X_{-1})$. Accordingly, dependence on the past turns into dependence on *larger* indices t [8], [16], [28], [26], [15]. I adhere to the probabilistic school, [44], [4], [9], [10], choosing $T = (-\mathbb{N}) = \{\dots, -2, -1\}$.

Every process $Y = (Y_t)_{t\in T}$ is distributed like some process X satisfying $X_t = f_t(U_t; X_{t-1}, X_{t-2}, \dots)$ for some Borel functions f_t and independent U_t. It follows that $\mathcal{M}(\mathcal{F}_X) \subset \mathcal{M}(\mathcal{F}_{X,U})$, but we need $\mathcal{M}(\mathcal{F}_X) \subset \mathcal{M}(\mathcal{F}_U)$. The two-component process (U, X) is a real-time transformation of U into X if and only if $\mathcal{F}_X \le \mathcal{F}_U$. Chaining $f_t, f_{t-1}, \dots, f_{s+1}$ we get $f_{s,t}$ such that $X_t = f_{s,t}(U_t, \dots, U_{s+1}; X_s, X_{s-1}, \dots)$. However, we need $f_{-\infty,t}$ such that $X_t = f_{-\infty,t}(U_t, U_{t-1}, \dots)$. That is possible if and only if the influence of $X_s, X_{s-1}, \dots$ on $f_{s,t}(U_t, \dots, U_{s+1}; X_s, X_{s-1}, \dots)$ disappears in the limit $s \to -\infty$. Tail triviality is necessary but not sufficient. Both examples shown above are tail trivial, and satisfy $X_t = U_t \dots U_{s+1} X_s$. Given U, the influence of X_s on X_t is strong, irrespective of s. Thus, the equality $X_t = U_t X_{t-1}$ fails to give an innovation.

Despite the strong influence of X_s on X_t, these X_s, X_t are (statistically) independent in the first example, and asymptotically independent (for $s \to -\infty$) in the second example. The strong dependence characterizes the specific way of using U_t (namely, $X_t = U_t X_{t-1}$), that is, the parametrization $(f_t)_{t\in T}$ rather than the process X itself. Is there a better parametrization for the same process? For the first example, the answer is evidently positive. Here, the conditional distribution of X_t, given the past, does not depend on the past. The parametrization $X_t = U_t X_{t-1}$ is bad because it introduces an unnecessary dependence on the past. A good parametrization is simply $X_t = U_t$, which surely is an innovation. For the second example, restricted to $t \in \mathbb{Z}$, the conditional distribution of X_t, given the past, depends on X_{t-1}. However, such distributions (corresponding to different values of X_{t-1}) overlap. A good parametrization uses the overlap for reducing dependence on the past. In continuous time, an innovation for the eternal motion is constructed [10] by inventing a coupling for processes differing in remote past. They are forced to coalesce, which never happens under the bad parametrization $X_t = U_t X_{t-1}$ of the form (2.2). That is the refutation of the delusion mentioned after (2.2).

Is there an innovation for an arbitrary tail-trivial process $(X_t)_{t\in(-\mathbb{N})}$? The answer is negative, which fact is "highly non-trivial and remarkable" [26], "deep and surprising" [15]. The first example, admitting no innovation, was discovered in the context of ergodic theory [37]. There are more examples of ergodic flavor [38], [29], [39], [28], [21], and of probabilistic flavor [8], [17], [14], [26], [4], [9]. The example of [8], furthered in [17], [14], [26], [4], is strikingly close to the sequence of i.i.d. equiprobable random signs; namely, the product measure is replaced with an equivalent (that is, mutually absolutely continuous) measure.

Some criteria for existence of an innovation, outlined in [37], [39], are elaborated in [15]. There, "substandardness" is our "existence of innovation", while "product type" is stronger, stipulating that U_t is a function of $X_t, X_{t-1}, \dots$ In such a case one says that U_t is exactly the *new* information furnished by X at t (though it depends on the chosen innovation). "Substandardness" implies "product type" provided that the conditional distribution of X_t given the past, is nonatomic [15].

4. Cosiness

Cosiness is a useful necessary condition for existence of an innovation. (*Is it also sufficient? I do not know.*) Cosiness emerged in [33, Def. 2.4] for continuous time

and in [4, Sect. 4] for discrete time, the latter with a reservation that "there is a whole range of possible variations" of the definition; one of the variations follows. Still, $T = (-\mathbb{N}) = \{\dots, -2, -1\}$, and processes are $\mathcal{X}$-valued.

4.1 DEFINITION. A random process $(X_t)_{t\in T}$ is *cosy,* if for each $\varepsilon > 0$ and each bounded Borel function $\varphi : \mathcal{X}^T \to \mathbb{R}$ there exists a two-component random process $(Y, Z) = \big((Y_t, Z_t)\big)_{t\in T}$ such that
(a) $\big((Y, Z), Y\big)$ and $\big((Y, Z), Z\big)$ are real-time transformations of (Y, Z) into X;
(b) $\mathbb{E}|\varphi(Y) - \varphi(Z)| < \varepsilon$;
(c) there exists $\delta \in (0, 1)$ such that for all bounded Borel functions $\psi, \chi : \mathcal{X}^T \to \mathbb{R}$,

$$\big(\mathbb{E}|\psi(Y)\chi(Z)|\big)^{2-\delta} \le \big(\mathbb{E}|\psi(Y)|^{2-\delta}\big)\big(\mathbb{E}|\chi(Z)|^{2-\delta}\big)\,.$$

Some comments. Condition (a) implies that each of the two processes Y, Z is distributed like X; thus, (Y, Z) is a joining of two copies of X, possessing the "real time" property $\mathcal{M}(Y) \subset \mathcal{M}(Y, Z)$, $\mathcal{M}(Z) \subset \mathcal{M}(Y, Z)$ (recall (2.4)). Condition (b) means that Y, Z are close, since φ may be one-one. Condition (c) means that Y, Z are "independent a little", since it is always satisfied for $\delta = 0$ and equivalent to independence of Y, Z for $\delta = 1$.

4.2 THEOREM. [4, Lemma 6 and Corollary 3] *A non-cosy process admits no innovation.*

The idea of a proof. Assume that X has an innovation; X is distributed like Y, $Y_t = f_{-\infty,t}(U_t, U_{t-1}, \dots)$, $U = (U_t)_{t\in T}$ being a sequence of independent $\mathcal{N}(0, 1)$ random variables. (This time we prefer the normal distribution $\mathcal{N}(0, 1)$ to $\mathcal{U}(0, 1)$.) Take another sequence $V = (V_t)_{t\in T}$ of independent $\mathcal{N}(0, 1)$ random variables such that U, V are independent. Introduce $W_t = U_t \cos\varepsilon + V_t \sin\varepsilon$, and let $Z_t = f_{-\infty,t}(W_t, W_{t-1}, \dots)$.* Condition (c) follows from the celebrated hypercontractivity theorem (pioneered by Nelson, see [24, Sect. 3])!

The first example of a non-cosy process in discrete time is given in [4, Th. 1]; it appears that the method of [8] produces non-cosy processes. It is interesting to know, whether "ergodic" examples [37], [38], [29], [39], [28], [21] are also non-cosy, or not. Another non-cosy discrete-time filtration [9] is the restriction of a continuous-time filtration to a discrete set on the time axis.

5. APPLICATIONS TO CONTINUOUS TIME

An $\mathcal{X}$-valued process $(X_t)_{t\in T}$, $T = (-\mathbb{N}) = \{\dots, -2, -1\}$, generates its filtration $\mathcal{F}_X = \big(\mathcal{F}_X(t)\big)_{t\in T}$. The family $\big(\mathcal{F}_X(2t)\big)_{t\in T}$ is also a filtration; it is generated by the $\mathcal{X}^2$-valued process $(Y_t)_{t\in T}$, $Y_t = (X_{2t-1}, X_{2t})$. If X admits an innovation, then the amalgamated process Y also does. The same applies for any infinite subset $T_1 \subset T$. If X is tail-trivial and T_1 is sparse enough, then Y admits an innovation, see [15, Th. 1.18] and references therein.

A continuous process $(X_t)_{t\in[0,\infty)}$ generates its filtration $\mathcal{F}_X = \big(\mathcal{F}_X(t)\big)_{t\in[0,\infty)}$. Choosing a sequence $(t_k)_{k\in(-\mathbb{N})}$, $t_k \in [0, \infty)$, $t_{k-1} < t_k$, $\inf t_k = 0$, we get a

* Which is anticipated in [23].

discrete-time filtration $\big(\mathcal{F}_X(t_k)\big)_{k\in(-\mathbb{N})}$, generated by the amalgamated process $(Y_k)_{k\in(-\mathbb{N})}$, $Y_k = (X_t)_{t\in[t_{k-1},t_k]}$. If Y admits no innovation, then X also admits no innovation, for any reasonable definition of continuous-time innovations. Some continuous-time problems are solved in that way.

The effect of "information from thin air" (see Sect. 3) can be reproduced by the stochastic differential equation

$$dX_t = dB_t + v\Big(t,\ \big(X_s\big)_{s\in[0,t]}\Big)\,dt \tag{5.1}$$

with a bounded drift v, if v is chosen properly. Then (5.1) fails to innovate X, which means that the equation has no strong solution. That is the "celebrated and mysterious" [25, V.3.18, p. 155] example, constructed in [32] and investigated in [5], [30], [43], [23], [10]. The eternal Brownian motion in a circle, mentioned in Sect. 3, can be obtained from X by a real-time transformation and a deterministic time change that maps $[0,\infty)$ onto $\mathbb{R}$ [10]. The same process X is a strong solution of the stochastic differential equation

$$dX_t = \sigma\Big(t,\ \big(X_s\big)_{s\in[0,t]}\Big)\,dB_t + v\Big(t,\ \big(X_s\big)_{s\in[0,t]}\Big)\,dt \tag{5.2}$$

for some $\sigma(\ldots) = \pm 1$ [10] (see also [16]). Once again, a clever parametrization is better than the straightforward parametrization.

One of the processes admitting no innovation, mentioned in Sect. 3, leads to a more ingenious drift v in (5.1); the corresponding (continuous) process X has no innovation, which means that it cannot be the strong solution of any equation of the form (5.2) [8]; see also [17], [14], [26], [4]. The drift is not bounded, but I believe that it can be made bounded. "Dreadfully complicated, their construction is almost as incredible as the existence result itself" [4]. Is it really a complicated construction? In fact, the drift is not constructed "by hands", it is chosen at random. It is a *random drift;* here "random" is interpreted like the second "random" in the phrase "random walk in a *random environment*". Thus, it is a typical drift in the same sense as a nowhere differentiable Brownian sample path is a typical function. Few parameters are adjusted by authors, such as order of magnitude, and depth of dependence on the past, both depending on time in a simple prescribed way.

There exists a pure martingale admitting no innovation [9].

6. From stochastic analysis to stochastic topology

Some continuous-time phenomena have no (evident) discrete-time counterpart. For example, Brownian motion cannot be transformed in real time into a Poisson process. A non-Gaussian stable process cannot be transformed into Brownian motion. The m-dimensional Brownian motion can be transformed into the n-dimensional Brownian motion if and only if $m \geq n$, which may be treated as the starting point of *stochastic topology,* the theory of filtration invariants of random processes.* A diffusion process with smooth nondegenerate coefficients in an n-dimensional smooth manifold is equivalent to the n-dimensional Brownian motion

* A useful classification claimed in [27, Th. 7] appeared to be not exhaustive [8, Sect. 6].

in the sense that their filtrations are isomorphic; in other words, the two processes can be connected by an *invertible* real-time transformation. What happens in presence of singularities of the topology or the coefficients? Few results are available; they are based on stochastic analysis (Itô formula, local times, ...). All negative results are based on continuous-time cosiness [33, Def. 2.4]. Brownian motion of finite or countable dimension is cosy [33, Lemma 2.5]. A cosy process cannot be transformed in real time to a non-cosy process [33, Lemma 2.6]. Therefore, all non-cosy processes are beyond the reach of Brownian innovation.

Two well-known diffusion processes in $\mathbb{R}$ are singular at the origin ($x = 0$, not $t = 0$ as in Sect. 5). The skew Brownian motion (see [20]) has a singular drift at 0, and is equivalent to the usual Brownian motion [20]. The sticky Brownian motion (see [41]) is slowed down at 0; its filtration is non-cosy [42].

Consider n rays (say, on the plane) with a single common point, the origin. There is a natural diffusion process Z_n on the union of the rays; Z_2 is the usual Brownian motion, Z_1 is the reflecting Brownian motion; $Z_3, Z_4, \ldots$ are so-called Walsh's Brownian motions [40], [3]. Such processes arise when considering small random perturbations of Hamiltonian dynamical systems [18] and some other topics [40], [3]. Processes Z_1 and Z_2 are equivalent (Lévy, Skorokhod). Nevertheless, Walsh's Brownian motions are non-cosy [33, Th. 4.13] (see also [11], [2]), which solves Problem 2 of [3].

Interestingly, stochastic topology can be of help to the classical (non-stochastic) analysis. Consider three non-intersecting domains in $\mathbb{R}^n$. If they are smoothly bounded, then points of trilateral contact are evidently rare among boundary points. It was conjectured for irregular domains, that the infimum of the three corresponding harmonic measures must vanish [6, Sect. 6], [12, Problem a]. In terms of the Martin boundary: its natural projection to the topological boundary is at most 2 to 1 almost everywhere. However, the best result of classical analysis is "at most 10 to 1" [6]. The final result "2 to 1" is achieved via stochastic topology [33, Th. 7.4]. A challenge for classical analysis!

So, some characteristic of $\mathbb{R}^n$ (or any smooth manifold) as a harmonic space, is equal to 2 irrespective of dimension, but exceeds 2 in presence of branching points. The nameless characteristic has its counterpart in stochastic topology, named *splitting multiplicity.* Introduced in [3, Def. 4.2], it was hibernating till the birth of cosiness. Every cosy process is of splitting multiplicity 2 (or 1, if it is degenerate) [2], while Walsh's Brownian motion Z_n, $n > 2$, is of splitting multiplicity n [2]. Splitting multiplicity is invariant under measure changes and time changes [2], while cosiness is not [4], [9].

7. White noise versus black noises

In discrete time we have no choice of noises for innovation; a noise is a sequence of independent random variables, each having a non-atomic distribution. In continuous time, the classical theory of processes with independent increments tells us that in general, a noise consists of a Gaussian component (a finite or countable collection of independent white noises) and a Poissonian component. The latter is useless for innovating diffusion processes. The former can innovate only

cosy processes. Thus, Walsh's Brownian motion is beyond the reach of classical innovation.

We may turn to Brownian motions (defined as continuous processes with stationary independent increments) on more general groups. In that aspect, finite-dimensional Lie groups are equivalent to $\mathbb{R}^n$. The Polish group of all unitary operators on the (separable) Hilbert space, equipped with the strong operator topology, is equivalent to (the additive group of) the Hilbert space [34, Th. 1.6]. (Interestingly, the proof involves continuous tensor products and continuous quantum measurements.) A commutative Polish group cannot give more [34, Th. 1.8].

The system of *coalescing* independent one-dimensional Brownian motions [1], [31, Sect. 2], is a limiting case of a coalescing stochastic flow. The system generates a two-parametric family of σ-fields $(\mathcal{F}_{s,t})_{s<t}$ that shares with the white noise the following property:

$$\mathcal{F}_{r,s} \otimes \mathcal{F}_{s,t} = \mathcal{F}_{r,t} \quad \text{whenever } r < s < t\,; \tag{7.1}$$

that is, $\mathcal{F}_{r,s}$ and $\mathcal{F}_{s,t}$ are independent and, taken together, generate $\mathcal{F}_{r,t}$. Nevertheless, $(\mathcal{F}_{s,t})_{s<t}$ supports no white noise (nor a Poisson process); it means that there is no Brownian motion $(B_t)_{t\in[0,\infty)}$ such that $B_t - B_s$ is $\mathcal{F}_{s,t}$-measurable for all intervals $(s,t) \subset [0,\infty)$ [35]. Thus, $(\mathcal{F}_{s,t})_{s<t}$ is a *black noise* as defined in [34, Sect. 1]. It is predictable [34, Def. 1.12] in the sense that its filtration $(\mathcal{F}_{0,t})_{t\in[0,\infty)}$ supports only continuous martingales. In fact, the filtration is Brownian! Therefore, that black noise still cannot innovate Walsh's Brownian motion.

One more example of a black noise is available [36, Sect. 5]. Does it generate a cosy filtration? I do not know.

7.2 Problem. Can a predictable noise (see [34, Defs. 1.1, 1.12]) generate a non-cosy filtration?

If the answer is positive, another problem follows.

7.3 Problem. Can Walsh's Brownian motion be innovated by *some* predictable noise?

7.4 Problem. Can a noise generate a cosy but non-Brownian filtration?*

References

[1] R. Arratia, *Coalescing Brownian motions, and the voter model on Z,* manuscript, Univ. of Southern California, 1985.

[2] M.T. Barlow, M. Émery, F.B. Knight, S. Song, M. Yor, *Autour d'un théorème de Tsirelson sur des filtrations browniennes et non browniennes,* Lect. Notes Math. (Séminaire de Probabilités XXXII), Springer, Berlin, **1686** (1998), 264–305.

[3] M. Barlow, J. Pitman, M. Yor, *On Walsh's Brownian motions,* Lect. Notes Math. (Séminaire de Probabilités XXIII), Springer, Berlin, **1372** (1989), 275–293. (MR 91a:60204)

* It was claimed in the abstract of my talk, that it never happens to Brownian motions in Polish groups. However, I withdraw the claim, since my proof was wrong. Sorry.

[4] S. Beghdadi-Sakrani, M. Émery, *On certain probabilities equivalent to coin-tossing, d'après Schachermayer,* Lect. Notes Math. (Séminaire de Probabilités XXXIII), Springer, Berlin, to appear.

[5] V.E. Beneš, *Nonexistence of strong nonanticipating solutions to stochastic DEs: implications for functional DEs, filtering, and control,* Stoch. Proc. Appl. **5**:3 (1977), 243–263. (MR 56#16788)

[6] C.J. Bishop, *A characterization of Poissonian domains,* Arkiv för Matematik **29**:1 (1991), 1–24. (MR 93a:31011)

[7] P. Brémaud, M. Yor, *Changes of filtrations and of probability measures,* Z. Wahrsch. Verw. Gebiete **45**:4 (1978), 269–295. (MR 80h:60062)

[8] L. Dubins, J. Feldman, M. Smorodinsky, B. Tsirelson, *Decreasing sequences of σ-fields and a measure change for Brownian motion,* Ann. Probab. **24**:2 (1996), 882–904. (MR 97g:60106)

[9] M. Émery, W. Schachermayer, private communication, March 1998.

[10] M. Émery, W. Schachermayer, private communication, April 1998.

[11] M. Émery, M. Yor, *Sur un théorème de Tsirelson relatif à des mouvements browniens corrélés et à la nullité de certains temps locaux,* Lect. Notes Math. (Séminaire de Probabilités XXXII), Springer, Berlin, **1686** (1998), 306–312.

[12] A. Eremenko, B. Fuglede, M. Sodin, *Harmonic measure for three disjoint domains in $\mathbb{R}^n$,* in "Linear and Complex Analysis Problem Book 3" (V.P. Havin, N.K. Nikolski, eds.) (1994), 323–325.

[13] J. Feldman, *Decomposable processes and continuous products of probability spaces,* J. Funct. Anal. **8** (1971), 1–51. (MR 44#7617)

[14] J. Feldman, *ε-close measures producing nonisomorphic filtrations,* Ann. Probab. **24**:2 (1996), 912–915. (MR 97g:60108)

[15] J. Feldman, *Decreasing sequences of measurable partitions: product type, standard, and prestandard,* Ergodic Theory and Dynamical Systems (to appear).

[16] J. Feldman, M. Smorodinsky, *Simple examples of non-generating Girsanov processes,* Lect. Notes Math. (Séminaire de Probabilités XXXI), Springer, Berlin, **1655** (1997), 247–251.

[17] J. Feldman, B. Tsirelson, *Decreasing sequences of σ-fields and a measure change for Brownian motion.II,* Ann. Probab. **24**:2 (1996), 905–911. (MR 97g:60107)

[18] M. Freidlin, Markov Processes and Differential Equations: Asymptotic Problems, Birkhäuser Verlag, Basel, 1996. (MR 97f:60150)

[19] R.K. Getoor, M.J. Sharpe, *Conformal martingales,* Invent. Math. **16** (1972), 271–308. (MR 46#4603)

[20] J.M. Harrison, L.A. Shepp, *On skew Brownian motion,* Ann. Probab. **9**:2 (1981), 309–313. (MR 82j:60144)

[21] D. Heicklen, C. Hoffman, *T, T^{-1} is not standard,* Ergodic Theory and Dynamical Systems (to appear).

[22] N. Ikeda, S. Watanabe, Stochastic Differential Equations and Diffusion Processes, second edition, North-Holland, 1989. (MR 90m:60069)

[23] J.F. Le Gall, M. Yor, *Sur l'equation stochastique de Tsirelson,* Lect. Notes Math. (Séminaire de Probabilités XVII), Springer, Berlin, **986** (1983), 81–88. (MR 86j:60132)

[24] E. Nelson, *The free Markoff field,* J. Funct. Anal. **12** (1973), 211–227. (MR 49#8556)

[25] L.C.G. Rogers, D. Williams, Diffusions, Markov Processes, and Martingales. Vol. 2: Itô Calculus, Wiley, New York, 1987. (MR 89k:60117)

[26] W. Schachermayer, *On certain probabilities equivalent to Wiener measure, d'après Dubins, Feldman, Smorodinsky and Tsirelson,* Lect. Notes Math. (Séminaire de Probabilités XXXIII), Springer, Berlin, to appear.
[27] A.V. Skorokhod, *Stochastic processes in infinite-dimensional spaces,* Proc. Internat. Cong. Math. (A.M. Gleason, ed.) 163–171. Amer. Math. Soc., Providence, RI, 1987. (In Russian.) (MR 89e:60081)
[28] M. Smorodinsky, *Processes with no standard extension,* Israel J. Math. (to appear).
[29] A.M. Stepin, *On entropy invariants of decreasing sequences of measurable partitions,* Funct. Anal. Appl. **5**:3 (1971), 237–240 (transl. from Russian).
[30] D.W. Stroock, M. Yor, *On extremal solutions of martingale problems,* Ann. Sci. École Norm. Sup. (4) **13**:1 (1980), 95–164. (MR 82b:60051)
[31] B. Tóth, W. Werner, *The true self-repelling motion,* Probab. Theory Related Fields (to appear).
[32] B.S. Tsirel'son, *An example of a stochastic differential equation having no strong solution,* Theory Probab. Appl. **20**:2 (1975), 416–418 (transl. from Russian). (MR 51#11654)
[33] B. Tsirelson, *Triple points: from non-Brownian filtrations to harmonic measures,* Geom. Funct. Anal. (GAFA) **7** (1997), 1096–1142.
[34] B. Tsirelson, *Unitary Brownian motions are linearizable,* MSRI Preprint No. 1998-027, math.PR/9806112.
[35] B. Tsirelson, *Brownian coalescence as a black noise,* manuscript in preparation.
[36] B.S. Tsirelson, A.M. Vershik, *Examples of nonlinear continuous tensor products of measure spaces and non-Fock factorizations,* Reviews in Mathematical Physics **10**:1 (1998), 81–145.
[37] A.M. Veršik, *Decreasing sequences of measurable partitions and their applications,* Soviet Math. Dokl. **11**:4 (1970), 1007–1011 (transl. from Russian). (MR 42#3258)
[38] A.M. Vershik, *Continuum of pairwise nonisomorphic diadic sequences,* Funct. Anal. Appl. **5** (1971), 182–184 (transl. from Russian). (MR 43#7593).
[39] A.M. Vershik, *Theory of decreasing sequences of measurable partitions,* St. Petersburg Math. J. **6**:4 (1995), 705–761 (transl. from Russian). (MR 96b:28018)
[40] J.B. Walsh, *A diffusion with a discontinuous local time,* Astérisque, **52–53** (1978), 37–45.
[41] J. Warren, *Branching processes, the Ray-Knight theorem, and sticky Brownian motion,* Lect. Notes Math. (Séminaire de Probabilités XXXI), Springer, Berlin, **1655** (1997), 1–15.
[42] J. Warren, *On the joining of sticky Brownian motion,* technical report of the dept. of statistics, university of Warwick, 1998.
[43] M. Yor, *Tsirel'son's equation in discrete time,* Probab. Theory Related Fields **91**:2 (1992), 135–152. (MR 93d:60104)
[44] M. Yor, Some Aspects of Brownian Motion, part II: Some Recent Martingale Problems, Birkhäuser Verlag, Basel, 1997. (MR 98e:60140)

Boris Tsirelson
School of Mathematics
Tel Aviv University
Tel Aviv 69978, Israel
email: tsirel@math.tau.ac.il

Reflecting Diffusions and Queueing Networks

R. J. Williams[1]

1 Introduction

Queueing models are of interest for analyzing congestion and delay in complex processing networks such as those occurring in computer systems, telecommunications and manufacturing (see e.g., [BG92, Ya94]). Many of these networks can process more than one class of job at a given station (so-called *multiclass* networks) and/or have complex feedback structures. Generally such models cannot be analyzed exactly and it is natural to seek more tractable approximations. In connection with this, certain diffusion processes known as *semimartingale reflecting Brownian motions (SRBMs)* [RW88] have been proposed as approximations for heavily loaded queueing networks (see e.g., [Ha88, HN93]), and there is now a substantial theory for these diffusions (see the survey in [Wi95]). However, limit theorems justifying their role as approximations have only been proved for some networks (see the overview in [Wi96]). Indeed, since a surprising example of Dai and Wang [DWa93] it has been known that these approximations are not always valid for multiclass networks with feedback. A challenging open problem has been that of establishing general conditions under which SRBM approximations for open multiclass queueing networks are valid. Recent progress on this problem and related work is summarized here.

The paper is organized as follows. In §2, the existence and uniqueness theory for SRBMs is described, including an oscillation inequality [Wi97a] which is critical to establishing tightness of normalized queueing network processes. In §3, the model used here for an open multiclass queueing network is defined. In §4, the main theorem is stated which gives general sufficient conditions for a heavy traffic limit theorem, which justifies approximating an open multiclass queueing network by a SRBM [Wi97b]. One of the key conditions involves something called *"state space collapse"*. Bramson has recently given sufficient conditions for this to hold (see [Br97b] and his article [Br98] in this volume). New heavy traffic limit theorems for two interesting collections of networks are obtained by combining the above results. The paper concludes with some open problems in §5.

2 Semimartingale Reflecting Brownian Motions

Definition of a SRBM Let J be a positive integer, $\mathbb{R}^J_+ \equiv \{x \in \mathbb{R}^J : x_j \geq 0 \text{ for } j = 1, \ldots, J\}$, $\mathcal{B}$ denote the σ-algebra of Borel subsets of $\mathbb{R}^J_+$, ν be a probability measure on $(\mathbb{R}^J_+, \mathcal{B})$, θ be a constant vector in $\mathbb{R}^J$, Γ be a $J \times J$ non-degenerate covariance matrix, and R be a $J \times J$ matrix.

[1]Research supported in part by the U.S. National Science Foundation.

DEFINITION 2.1 *A semimartingale reflecting Brownian motion (SRBM) associated with the data (θ, Γ, R, ν) is a J-dimensional process W defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ such that*

$$W = X + RY \tag{1}$$

where W, X, Y are $\{\mathcal{F}_t\}$-adapted processes such that W has continuous paths in $\mathbb{R}^J_+$, X is a J-dimensional Brownian motion with drift vector θ, covariance matrix Γ, initial distribution ν, and $\{X(t) - X(0) - \theta t, \mathcal{F}_t, t \geq 0\}$ is a martingale, and Y is a J-dimensional process such that for each $j \in \{1, \ldots, J\}$, $Y_j(0) = 0$, Y_j is continuous and non-decreasing, and $\int_0^\infty 1_{(0,\infty)}(W_j(s))dY_j(s) = 0$, i.e., Y_j can increase only when W_j is zero.

Intuitively, such a SRBM behaves in the interior of the orthant $\mathbb{R}^J_+$ like a Brownian motion with initial distribution ν, constant drift θ and covariance matrix Γ, and it is confined to $\mathbb{R}^J_+$ by "pushing" at the boundary, where for $j = 1, \ldots, J$, the allowed direction of push on the relative interior of the boundary face $F_j = \{x \in \mathbb{R}^J_+ : x_j = 0\}$ is given by the j^{th} column of the matrix R. At an intersection of faces, the allowed directions of "push" are given by the convex combinations of the push directions associated with the faces meeting there. For historical reasons, stemming from an alternative construction of the driftless process in one-dimension, the "pushing" at the boundary is called instantaneous reflection. However, it is more accurate to think of this action as deflection or regulation rather than some type of mirror reflection. The process Y is called the "pushing process" associated with W and it is related to the local time of W on the boundary. Since the state space for a SRBM is not smooth and the directions of reflection may be discontinuous at the non-smooth parts of the boundary, the general theory for diffusions with smooth boundary conditions [SV71] does not apply to SRBMs and one must develop a theory from first principles.

The above definition of a SRBM is in the spirit of weak solutions of stochastic equations. In particular, one is free to choose the filtered probability space and processes W, X, Y such that the above properties hold. Here the focus is on such weak solutions, since necessary and sufficient conditions for their existence and uniqueness are known, whereas only sufficient conditions are known for strong solutions. Furthermore, there are multiclass queueing networks (see the example due to Dai, Wang and Wang in Appendix A of [Wi97b]) whose SRBM approximants are not covered by the extant strong solution theory.

EXISTENCE AND UNIQUENESS FOR SRBMS It is straightforward to see that a necessary condition for the existence of a SRBM associated with (θ, Γ, R, ν) for each probability measure ν on $(\mathbb{R}^J_+, \mathcal{B})$ is the following: at each point on the boundary of $\mathbb{R}^J_+$ there is a positive linear combination of the "push" directions that can be used there which points into the interior of $\mathbb{R}^J_+$. This geometric description can be expressed succinctly as the following algebraic condition: the matrix R is *completely-$\mathcal{S}$* if for each principal submatrix $\tilde{R}$ of R there is a vector $\tilde{x} \geq 0$ such that $\tilde{R}\tilde{x} > 0$. (Here inequalities are to be interpreted componentwise and a principal submatrix of R is obtained by deleting all rows and columns of R

with indices in some strict (possibly empty) subset of $\{1, \ldots, J\}$.) In fact, R being completely-$\mathcal{S}$ is also sufficient for the existence and uniqueness in law of a SRBM. The following result is proved for $\nu = \delta_x$ (the unit mass at $x \in \mathbb{R}^J_+$) in [TW93] and is easily extended to all ν [Wi97a].

Theorem 2.1 *Suppose that R is completely-$\mathcal{S}$. There exists a SRBM associated with (θ, Γ, R, ν) and it is unique in law. Furthermore, the laws induced on the space of continuous paths in $\mathbb{R}^J_+$ by the SRBMs associated with $(\theta, \Gamma, R, \delta_x)$, $x \in S$, define a Feller continuous strong Markov process.*

Oscillation Inequality Solutions of a deterministic Skorokhod problem have been used to obtain strong constructions of SRBMs in some cases [DuI91, HR81]. While this Skorokhod problem will not have unique solutions for general completely-$\mathcal{S}$ matrices R [BEK91, Ma92], an oscillation inequality for a perturbed form of this problem can be used to establish tightness for suitable approximations to a SRBM. Indeed, this inequality can be used to show existence of a SRBM (using deflected random walk approximations having small inward jumps at the boundary) and the form obtained by restricting to continuous paths $x(\cdot)$ and setting $\epsilon = 0$ is used in the proof of uniqueness in law of a SRBM [TW93]. (This "continuous" case of the oscillation inequality first appeared in [BEK91].)

In the following statement of the oscillation inequality, for any $0 \le t_1 < t_2 < \infty$, $D([t_1, t_2], \mathbb{R}^J)$ denotes the set of functions $x : [t_1, t_2] \to \mathbb{R}^J$ that are right continuous on $[t_1, t_2)$ and have finite left limits on $(t_1, t_2]$ and $\mathrm{Osc}(x, [t_1, t_2]) = \sup\{|x(t) - x(s)| : t_1 \le s < t \le t_2\}$ for any $x \in D([t_1, t_2], \mathbb{R}^J)$, where $|a| = \max_{j=1}^J |a_j|$ for any $a \in \mathbb{R}^J$.

Theorem 2.2 [Wi97a] *Assume that R is completely-$\mathcal{S}$. Suppose that $\epsilon \ge 0$, $0 \le t_1 < t_2 < \infty$ and $w, x, y \in D([t_1, t_2], \mathbb{R}^J)$ are such that*

(i) *$w(t) = x(t) + Ry(t) \in \mathbb{R}^J_+$ for all $t \in [t_1, t_2]$,*

(ii) *for each $j \in \{1, \ldots, J\}$, $y_j(t_1) \ge 0$, y_j is non-decreasing, and $\int_{[t_1, t_2]} 1_{(\epsilon, \infty)}(w_j(s)) dy_j(s) = 0$.*

Then there is a constant $C > 0$, depending only on R, such that

$$\mathrm{Osc}(y, [t_1, t_2]) + \mathrm{Osc}(w, [t_1, t_2]) \le C(\mathrm{Osc}(x, [t_1, t_2]) + \epsilon). \tag{2}$$

This oscillation inequality plays a key role in establishing tightness of normalized queueing network processes approximating SRBMs (cf. §4).

Other Results and Extensions For further discussion of SRBMs, including weak versus strong solutions, conditions for recurrence, and characterization of stationary distributions, see the survey article [Wi95] and references therein. Semimartingale reflecting Brownian motions in convex polyhedrons (in contrast to the orthant) can arise as approximations to closed and capacitated queueing networks. The reader is referred to [DWi95] for sufficient conditions for the existence and uniqueness of such processes and to [DD97] for a related oscillation inequality and heavy traffic limit theorem. Semimartingale reflecting Brownian

motions in polyhedrons also arise in other applications, e.g., in economic models of monetary exchange [FL98]. Reflecting Brownian motions (RBMs) that are not semimartingales have also been proposed as approximations to some particular queueing network models (see e.g., [DuR98b, KL93]). However, the theory of existence and uniqueness for these non-semimartingale RBMs is not as complete as for SRBMs, being restricted to the two-dimensional case [VW85] or to RBMs whose geometric data is a limit of that for SRBMs [DuR98a].

3 Open Multiclass Queueing Network Model

In an open queueing network, jobs arrive from outside the system, visit a finite number of stations where they receive service, and then exit the network. The model for an open multiclass queueing network used here is a generalization of one with a first-in-first-out (FIFO) service discipline considered in [HN93]. To simplify the exposition, attention is restricted to networks that are initially empty. For a more complete specification of the model, including a treatment of networks that are initially non-empty, see [Wi97b]. The model description is broken down into assumptions concerning the network structure, primitive stochastic processes (for exogenous arrivals, service times and routing), and the service discipline.

Network Structure The model has a fixed set $\{1,\dots,J\}$ of stations with a single reliable server at each. At any given time, each job in the network belongs to one of a finite set $\mathcal{K}=\{1,\dots,K\}$ of job classes. Each class is associated with exactly one station (where the class is to receive service). The deterministic many-to-one function mapping classes to stations is specified by a $J\times K$ constituency matrix C where $C_{jk}=1$ if class k is served at station j and $C_{jk}=0$ otherwise. At a given station, jobs of different classes may be distinguished by features such as the distributions of their service times, their routing characteristics, or their order of service. Upon completing service in a class, a job changes class in Markovian fashion. Each station serves at least one class and has an infinite buffer for storing jobs awaiting service there.

Stochastic Primitives The primitive stochastic processes for the model are (E,V,Φ) where E is a K-dimensional external arrival process, V is a K-dimensional cumulative service time process, $\Phi=(\Phi^1,\Phi^2,\dots,\Phi^K)$ and Φ^k is a K-dimensional routing process for class $k\in\mathcal{K}$. More precisely, for each k and $t\geq 0$, $E_k(t)$ represents the number of exogenous arrivals to class k up to time t. It is assumed that $E_k\not\equiv 0$ for at least one k and for each such k, E_k is a renewal process derived from a sequence of positive i.i.d. interarrival times having finite mean and variance. For each class k and integer $n\geq 0$, $V_k(n)=\sum_{i=1}^n v_k(i)$ where $\{v_k(i)\}_{i=1}^\infty$ is a sequence of i.i.d. positive random variables with finite mean and variance, and $v_k(i)$ is interpreted as the service time for the i^{th} job that arrives to class k. To describe the Markovian routing, let $e_1,\dots,e_K$ denote the non-negative unit basis vectors parallel to the K coordinate axes in $\mathbb{R}^K$ and let e_0 be the K-dimensional zero vector. For each class k and integer $n\geq 0$, $\Phi^k(n)=\sum_{i=1}^n \phi^k(i)$ where $\{\phi^k(i)\}_{i=1}^\infty$ is a sequence of i.i.d. random vectors taking values in $\{e_0,e_1,\dots,e_K\}$ with $P(\phi^k(i)=e_l)=P_{kl}$, $k,l\in\mathcal{K}$, and P is a strictly

substochastic $K \times K$ matrix. The interpretation of the routing vector $\phi^k(i)$ is that the i^{th} job to depart from class k is routed next to class l if $\phi^k(i) = e_l$ and it leaves the network if $\phi^k(i) = e_0$. The strict substochasticity of P ensures that jobs eventually leave the network. The processes $E_1, \ldots, E_K, V_1, \ldots, V_K, \Phi^1, \ldots, \Phi^K$ are assumed to be mutually independent.

Service Discipline It remains to specify the order in which jobs are served at each station, i.e., the service discipline. Attention is confined to HL (head-of-the-line) service disciplines (cf. [Br97a, Wi97b]). (Other disciplines such as last-in-first-out or general processor sharing are also of interest, but the heavy traffic theory for networks with these disciplines is much less developed.) Firstly, an HL discipline is non-idling in the sense that a server is never idle when there are jobs waiting to be served at its station. In addition, jobs in each class are served on a first-in-first-out basis, i.e., service for each class is concentrated on the job at the head-of-the-line for that class. Each class receives a proportion (possibly zero) of the associated server's time, where this proportion may be random but is kept constant between changes in the arrival or departure processes, and these proportions depend in a measurable way on the "state" of the queueing network at the time of the last such change. (The "state" description includes such quantities as queue lengths, remaining service times of jobs at a station, amounts of time that jobs have been waiting in their current class, and the amount of time until the next exogenous arrival to each class cf. [Wi97b].) Common service disciplines included in the HL framework are FIFO (regardless of their class designation, jobs at a station are served in the order in which they arrived there), static priorities (classes at a station are ranked and jobs of a higher ranking class are always served before those of a lower ranking class), and HLPPS (head-of-the-line proportional processor sharing: each class at a station receives service in proportion to the number of jobs that are present in that class).

Descriptive Processes and Model Equations Let A, D be the K-dimensional processes such that $A_k(t)$ denotes the number of arrivals to, and $D_k(t)$ denotes the number of departures from, class k up to time t. The processes that are used to measure performance are a K-dimensional queue length process Z, a J-dimensional workload process W and a J-dimensional cumulative idletime process Y. For each class k, station j and time t, $Z_k(t)$ denotes the number of class k jobs that are in queue or being served at time t (the letter Z is mnemonic for the German *Zahl* or number), $W_j(t)$ denotes the amount of work for server j (measured in units of remaining service time) that is embodied in those jobs that are at station j at time t, $Y_j(t)$ denotes the total amount of time that server j has been idle up to time t.

The descriptive processes (A, D, W, Y, Z) satisfy the following equations:

$$A(t) = E(t) + \Phi(D(t)), \quad Z(t) = A(t) - D(t), \quad W(t) = CV(A(t)) - et + Y(t). \tag{3}$$

Here e is the J-dimensional vector of all ones and the k^{th} component of $\Phi(D(t))$ is to be read as $\sum_{l=1}^{K} \Phi_k^l(D_l(t))$ and the k^{th} component of $V(A(t))$ is to be read as $V_k(A_k(t))$. The equation for A indicates that the $A_k(t)$ arrivals to class k up to time t consist of $E_k(t)$ exogenous arrivals plus $\sum_{l=1}^{K} \Phi_k^l(D_l(t))$ arrivals obtained by

feedback of some of the departures that have occurred up to time t. The equation for the workload process W expresses the fact that $\sum_{k\in\mathcal{K}} C_{jk}V_k(A_k(t))$ units of work have arrived for server j in $[0,t]$ and that this has been depleted by the amount of time $t - Y_j(t)$ that server j has been active in $[0,t]$. The fact that an HL discipline is non-idling implies that $\int_0^\infty 1_{(0,\infty)}(W_j(s))dY_j(s) = 0$ for all j.

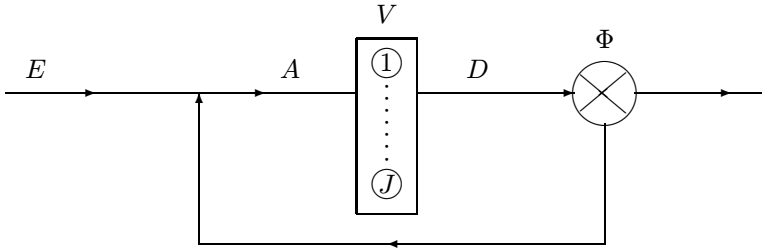


FIGURE 1: SCHEMATIC FOR AN OPEN MULTICLASS QUEUEING NETWORK

Note that the equations (3) do not give a complete description of the behavior of the queueing network. In particular one must add additional equations to provide information about the service discipline. For example, for the FIFO discipline one can add the relations: $D_k(t + W_j(t)) = A_k(t)$, for each class k and associated server j. Equations for other HL service disciplines will not be given here, since for the statement of the main theorem (Theorem 4.1), only a distillation of the service discipline is needed in the form of a $K \times J$ matrix Δ. Since this matrix is related to the heavy traffic behavior of networks, discussion of it is deferred to the next section.

HEAVY TRAFFIC The following notation is used in describing the notion of heavy traffic. Let α denote the K-dimensional long run average arrival rate vector for the exogenous arrival process E and let M denote the $K \times K$ diagonal matrix whose diagonal entries are the mean service times m_k for the classes $k \in \mathcal{K}$. Let λ be the unique solution of the "traffic flow" equation $\lambda = \alpha + P'\lambda$, i.e., $\lambda = (I - P')^{-1}\alpha$. Here $'$ denotes transpose. (To avoid degeneracies, it is assumed that $\lambda_k > 0$ for each k.) Define $\rho = CM\lambda$. The quantity λ_k is called the arrival rate for class k and ρ_j is called the traffic intensity parameter for station j. These are nominally the long run average rate at which jobs arrive to class k and the long run fraction of time that server j is busy, respectively. For single class networks, these nominal quantities represent actual long run quantities (provided $\rho_j \leq 1$ for all j). However, since the appearance of counterexamples in the early 1990s (see e.g., [LK91, RS91]), it has been known that this interpretation is not always valid for multiclass networks. Indeed, the question of whether these nominal quantities actually correspond to long run quantities is related to the stability properties of the queueing network. Rather than digressing to discuss this further here, the reader is referred to the articles on stability in [KW95], the references therein, and the article [Br98]. Here λ and ρ are simply regarded as useful parameters. Networks that are (nominally) heavily loaded or in heavy traffic are those in which ρ_j is close to one for each j. Such networks are the focus of attention in the next section.

4 Sufficient Conditions for a Heavy Traffic Limit Theorem

A Sequence of Networks Mathematically, to justify the approximation of a given heavily loaded open multiclass queueing network by a SRBM, we regard the network as being a member of a sequence of networks in which the traffic intensity vector ρ converges to e. Here, to simplify the exposition, the sequence is chosen so that only the distributions of the service times vary along the sequence where this variation is parametrized by the mean service times. (A more complex setup can be considered, allowing for more general variation of the distributions of all of the stochastic primitives along the sequence [Wi97b]. Although this implies a certain robustness of the approximation to small perturbations in the distributions of the stochastic primitives, for the purpose of stating a limit theorem that justifies the approximation of a fixed heavily loaded network, only the simpler setup described here is needed.)

Thus, we consider a sequence of networks indexed by r, which tends to infinity through a strictly increasing sequence of positive numbers. Each network in the sequence has the same basic structure as described in the previous section. Furthermore, J, K, C, E, Φ and the service discipline do not vary with r, and $v_k(i) = m_k^r u_k(i)$ where m_k^r is the mean service time for class k in the r^{th} network and $u_k(i)$ is a random variable independent of r that has mean one and finite variance. (To avoid degeneracies, it is assumed that $u_k(i)$ has positive variance for each class k. This assumption implies that the covariance matrix for the proposed SRBM approximant is non-degenerate. For other ways in which this can be achieved, see §5 of [Wi97b].) In the sequel, the superscript r is attached to all quantities that may depend on r.

Now assume the following heavy traffic conditions: as $r \to \infty$, $m_k^r \to m_k \in (0, \infty)$ for each $k \in \mathcal{K}$, such that $\gamma^r \equiv r(\rho^r - e) \to \gamma \in \mathbb{R}^J$. Define the diffusion scaled workload, cumulative idletime and queue length processes:

$$\hat{W}^r(t) = W^r(r^2 t)/r, \qquad \hat{Y}^r(t) = Y^r(r^2 t)/r, \qquad \hat{Z}^r(t) = Z^r(r^2 t)/r. \tag{4}$$

The purpose of a heavy traffic limit theorem is to justify approximating $(\hat{W}^r, \hat{Y}^r, \hat{Z}^r)$ in distribution using a SRBM.

State Space Collapse A key feature of prior limit theorems in the multiclass setting [Wh71, Pe91, Re88] has been a phenomenon called state space collapse, which states that the diffusion scaled queue length process for each class k can be approximately recovered as a multiple of the associated station's diffusion scaled workload process. Here a slightly weaker notion called multiplicative state space collapse is used. This form suffices for our purposes and seems more amenable to verification (cf. [Br97b]). Here $\|f(\cdot)\|_T = \sup_{0 \le t \le T} |f(t)|$ for any vector valued function f defined on $[0, T]$. (The notion of state space collapse is defined by omitting the denominator in (5) below.)

Definition 4.1 *Multiplicative state space collapse holds if there is a $K \times J$ matrix Δ such that for each $T \ge 0$,*

$$\frac{\|\hat{Z}^r(\cdot) - \Delta \hat{W}^r(\cdot)\|_T}{\|\hat{W}^r(\cdot)\|_T \vee 1} \to 0 \quad \textit{in probability as } r \to \infty, \tag{5}$$

where $a \vee b \equiv \max(a, b)$ for any two real numbers a, b.

Based on extant limit theorems, some conjectured forms of Δ for various service disciplines are described in [Wi97b]. In fact, one can show (see Appendix B in [Wi97b]) that a necessary condition for $\{(\hat{W}^r, \hat{Z}^r)\}$ to be C-tight under the FIFO service discipline is that (multiplicative) state space collapse holds with $\Delta = \Lambda C'$ where Λ is the $K \times K$ diagonal matrix with the entries of λ on its diagonal.

SUFFICIENT CONDITIONS FOR A HEAVY TRAFFIC LIMIT THEOREM The main content of the following theorem is that for a sequence of open multiclass queueing networks as described above (with a general HL service discipline), multiplicative state space collapse plus the natural condition that the reflection matrix R for the purported SRBM approximant is well defined and completely-$\mathcal{S}$, is sufficient for a heavy traffic limit theorem to hold. Here $\Rightarrow$ denotes convergence in distribution of processes taking values in the space of paths that are right continuous with finite left limits, where this space is endowed with the usual Skorokhod topology.

THEOREM 4.1 [Wi97b] *Suppose that multiplicative state space collapse holds and that the inverse matrix $R = (CM(I - P')^{-1}\Delta)^{-1}$ exists and is completely-$\mathcal{S}$. Then*

$$(\hat{W}^r, \hat{Y}^r, \hat{Z}^r) \Rightarrow (W^*, Y^*, Z^*) \quad \text{as } r \to \infty, \tag{6}$$

where W^* is a SRBM with data $(R\gamma, \Gamma, R, \delta_0)$ and associated pushing process Y^*, and $Z^* = \Delta W^*$. The covariance matrix Γ is a known quantity determined from C and the means and covariances of the stochastic primitives [Wi97b], and δ_0 denotes the unit mass at the origin in $\mathbb{R}_+^J$.

The proof of this theorem proceeds by showing tightness of the sequence $\{(\hat{W}^r, \hat{Y}^r, \hat{Z}^r)\}$ and uniqueness in law of any weak limit point. For the tightness, multiplicative state space collapse is combined with the oscillation inequality of Theorem 2.2. For the uniqueness of any weak limit point $(W^\dagger, Y^\dagger, Z^\dagger)$, one needs to show that $W^\dagger$ is a SRBM with associated pushing process $Y^\dagger$. In particular, the martingale property in the definition of a SRBM needs to be verified for $X^\dagger = W^\dagger - RY^\dagger$. This involves establishing a multiparameter stopping time property which is where the precise definition of a HL service discipline (including its measurable dependence on the "state") comes into play.

NEW HEAVY TRAFFIC LIMIT THEOREMS In a companion work to [Wi97b], Bramson [Br97b] (see also [Br98]) has given sufficient conditions for multiplicative state space collapse to hold. These conditions are in terms of the behavior of a balanced fluid model (a law of large numbers approximation for the sequence of heavily loaded queueing networks). In particular, using these conditions and his prior work on the fluid model behavior for FIFO Kelly type and HLPPS networks, Bramson [Br97b] has shown that multiplicative state space collapse holds for these two collections of networks. The qualifier "Kelly type" means that m_k depends only on the station j at which class k is served, i.e., the limiting mean service times are station-dependent, not class-dependent, quantities. In addition, it is known [DH93, Wi97b] that R is well defined and completely-$\mathcal{S}$ for these networks.

Combining the above results yields new heavy traffic limit theorems for these two collections of networks. In particular, the FIFO Kelly type network introduced by Dai, Wang and Wang (see Appendix A in [Wi97b]) can be approximated by a SRBM. This is particularly interesting since the continuous mapping (strong solution) approach used in most prior limit theorems cannot be applied to that example.

In independent work, Chen and Zhang [CZ97] have established a heavy traffic limit theorem for FIFO networks in which $G = CM(I - P')^{-1}P'\Lambda C'$ has spectral radius less than one. Although they do not use Theorem 4.1, they implicitly verify the conditions of that theorem for their case and avoid a continuous mapping argument in a similar manner to that in [Wi97b].

5 Open Problems

The results in [Br97b, Wi97b] reduce the problem of establishing heavy traffic limit theorems for open multiclass queueing networks with a HL service discipline to that of establishing multiplicative state space collapse through the study of balanced fluid models over long intervals of time and to verifying that the reflection matrix R is well defined and completely-$\mathcal{S}$. A compelling open problem is to identify new collections of networks that satisfy these conditions. In particular, it is natural to consider networks with static priority service disciplines (see the article [Br98] by Bramson for recent work in this direction). Another area for future investigation is heavy traffic behavior of networks with non-HL disciplines such as last-in-first-out and general processor sharing. Finally, the focus here has been on performance analysis for heavily loaded networks with a fixed structure. In some applications one may be able to vary such quantities as the service discipline or routing in a dynamic manner with the objective of optimizing some measure of performance. Again such problems frequently cannot be analyzed exactly and one may seek approximate models. An approach using approximate diffusion models has been advocated by some authors (see e.g., [HW89, KL93, Ku95]), but many open problems remain concerning justification and interpretation of such approximations in general.

References

[BEK91] Bernard, A., and El Kharroubi, A. (1991). Régulation de processus dans le premier orthant de $\mathbf{R}^n$. *Stochastics*, 34, 149–167.

[BG92] Bertsekas, D., and Gallagher, R. (1992). *Data Networks*. Prentice-Hall.

[Br97a] Bramson, M. (1997). Stability of two families of queueing networks and a discussion of fluid limits. To appear in *Queueing Syst.*

[Br97b] Bramson, M. (1997). State space collapse with application to heavy traffic limits for multiclass queueing networks. To appear in *Queueing Syst.*

[Br98] Bramson, M. (1998). State space collapse for queueing networks. *Proceedings of the International Congress of Mathematicians, Berlin, 1998,* this issue.

[CZ97] Chen, H., and Zhang, H. (1997). Diffusion approximations for some multiclass queueing networks with FIFO service disciplines. Preprint.

[DD97] Dai, J. G., and Dai, W. (1997). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. Preprint.

[DH93] Dai, J. G., and Harrison, J. M. (1993). The QNET method for two-moment analysis of closed manufacturing systems. *Ann. Appl. Prob.*, 3, 968–1012.

[DWa93] Dai, J. G., and Wang, Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Syst.*, 13, 41–46.
[DWi95] Dai, J. G. and Williams, R. J. (1995). Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Theory Probab. Appl.*, 40, 1–40.
[DuI91] Dupuis, P., and Ishii, H. (1991). On the Lipschitz continuity of the solution mapping to the Skorokhod problem. *Stochastics,* 35, 31–62.
[DuR98a] Dupuis, P., and Ramanan, K. (1998). Convex duality and the Skorokhod problem, I & II. Submitted to *Prob. Theor. Rel. Fields.*
[DuR98b] Dupuis, P., and Ramanan, K. (1998). A Skorokhod problem formulation and large deviation analysis of a processor sharing model. To appear in *Queueing Syst.*
[FL98] Flandreau, M. (1998). The burden of intervention: externalities in multilateral exchange rate arrangements. To appear in *J. Intern. Econ.*
[Ha88] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications*, W. Fleming and P.-L. Lions (eds.), Springer-Verlag, 147–186.
[HN93] Harrison, J. M., and Nguyen, V. (1993). Brownian models of multiclass queueing networks: current status and open problems. *Queueing Syst.*, 13, 5–40.
[HR81] Harrison, J. M., and Reiman, M. I. (1981). Reflected Brownian motion on an orthant. *Ann. Probab.*, 9, 302–308.
[HW89] Harrison, J. M., and Wein, L. M. (1989). Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Syst.*, 5, 265–280.
[KL93] Kelly, F. P., and Laws, C. N. (1993). Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Syst.*, 13, 47–86.
[KW95] Kelly, F. P., and Williams, R. J. (eds.) (1995). *Stochastic Networks.* Vol. 71, Springer.
[Ku95] Kushner, H. J. (1995). A control problem for a new type of public transportation system, via heavy traffic analysis. In [KW95], 139–167.
[LK91] Lu, S. H., and Kumar, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Autom. Control*, 36, 1406–1416.
[Ma92] Mandelbaum, A. (1992). The dynamic complementarity problem. Preprint.
[Pe91] Peterson, W. P. (1991). Diffusion approximations for networks of queues with multiple customer types. *Math. Oper. Res.*, 9, 90–118.
[Re84] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.* 9, 441–458.
[Re88] Reiman, M. I. (1988). A multiclass feedback queue in heavy traffic. *Adv. Appl. Prob.*, 20, 179–207.
[RW88] Reiman, M. I., and Williams, R. J. (1988–89). A boundary property of semimartingale reflecting Brownian motions. *Probab. Theory Relat. Fields,* 77, 87–97, and 80, 633.
[RS91] Rybko, A. N., and Stolyar, A. L. (1991). Ergodicity of stochastic processes describing the operation of an open queueing network. *Problemy Peredachi Informatsii*, 28, 2–26.
[SV71] Stroock, D. W., and Varadhan, S. R. S. (1971). Diffusion processes with boundary conditions. *Comm. Pure Appl. Math.*, 24, 147–225.
[TW93] Taylor, L. M., and Williams, R. J. (1993). Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probab. Theory Relat. Fields*, 96, 283–317.
[VW85] Varadhan, S. R. S., and Williams, R. J. (1985). Brownian motion in a wedge with oblique reflection. *Comm. Pure Appl. Math.*, 38, 405–443.
[Wh71] Whitt, W. (1971). Weak convergence theorems for priority queues: preemptive resume discipline. *J. Appl. Prob.*, 8, 74–94.
[Wi95] Williams, R. J. (1995). Semimartingale reflecting Brownian motions in the orthant. In [KW95], pp. 125–137.
[Wi96] Williams, R. J. (1996). On the approximation of queueing networks in heavy traffic. In *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins (eds.), Oxford University Press, Oxford, pp. 35–56.
[Wi97a] Williams, R. J. (1997). An invariance principle for semimartingale reflecting Brownian motions in an orthant. To appear in *Queueing Syst.*
[Wi97b] Williams, R. J. (1997). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. To appear in *Queueing Syst.*
[Ya94] Yao, D. D. (ed.) (1994). *Stochastic Modeling and Analysis of Manufacturing Systems.* Springer-Verlag.

Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla CA 92093-0112 USA

Section 13

Combinatorics

In case of several authors, Invited Speakers are marked with a *.

Hereditary Properties of Graphs: Asymptotic Enumeration, Global Structure, and Colouring

Béla Bollobás

1991 Mathematics Subject Classification: Primary: 05C, 05D; secondary: 51M16, 60E15

1. Introduction. In this paper we shall discuss recent developments concerning hereditary graph properties. In particular, we shall study the growth of the number of graphs with a given hereditary property; the structure of a 'typical' graph with the property; and the $\mathcal{P}$-chromatic number of a random graph $G_{n,p}$ for a fixed hereditary property $\mathcal{P}$.

A *graph property* $\mathcal{P}$ is a union of isomorphism classes of finite graphs. To avoid trivialities, we shall always assume that our properties contain infinitely many non-isomorphic graphs, but that for some n do not contain all graphs of order n. Here are some simple examples of graph properties: (i) all triangle-free graphs without 8-cycles, (ii) all graphs of chromatic number at most k, (iii) all graphs containing no induced quadrilaterals, (iv) all regular bipartite graphs, (v) all Hamiltonian graphs.

Rather than considering general properties, we frequently study hereditary properties. A property $\mathcal{P}$ is *hereditary* if it is closed under taking induced subgraphs. In other words, $\mathcal{P}$ is hereditary if $G \in \mathcal{P}$ implies that $G - x \in \mathcal{P}$ for every vertex x of G.

An important subclass of hereditary properties is the class of *monotone properties*, those that are closed under taking subgraphs. Thus $\mathcal{P}$ is monotone if $G \in \mathcal{P}$ implies that $G - x \in \mathcal{P}$ for every vertex x of G *and* $G - e \in \mathcal{P}$ for every edge e of G. Note that properties (i) and (ii) are monotone, (iii) is hereditary but not monotone, and properties (iv) and (v) are not hereditary.

The most natural way of measuring the size of a property is to take the number of elements in its finite sections. Given a property $\mathcal{P}$, write $\mathcal{P}^n$ for the set of graphs in $\mathcal{P}$ with vertex set $[n] = \{1, \dots, n\}$. Then $(|\mathcal{P}^n|)_{n=1}^{\infty}$ is, in an obvious sense, a measure of $\mathcal{P}$.

For a monotone property $\mathcal{P}$ there is another natural measure: the sequence $(e(\mathcal{P}^n))_{n=1}^{\infty}$, where $e(\mathcal{P}^n)$ is the maximal *size* (number of edges) of a graph in $\mathcal{P}^n$. For a general property $\mathcal{P}$, the sequence $(e(\mathcal{P}^n))_{n=1}^{\infty}$ may have little significance, so we have to turn to a natural extension of it. A *pregraph* is a triple $\tilde{G} = (V, \tilde{E}, \tilde{N})$, where V is a finite set, the set of *vertices*, and $\tilde{E}$ and $\tilde{N}$ are disjoint subsets of $V^{(2)}$, the set of unordered pairs of vertices; $\tilde{E}$ is the set of *edges* and $\tilde{N}$ is the

set of *non-edges* of $\tilde{G}$. A graph $G = (V, E)$ *extends* $\tilde{G}$ if $\tilde{E} \subset E \subset V^{(2)} \setminus \tilde{N}$. The *size* $e(\tilde{G})$ of a pregraph is $|V^{(2)} \setminus (\tilde{E} \cup \tilde{N})|$: the number of choices we have when extending $\tilde{G}$ to a graph. We say that a pregraph $\tilde{G}$ *belongs* to $\mathcal{P}^n$ if every graph extending $\tilde{G}$ belongs to $\mathcal{P}^n$. Then another natural measure of the size of a property $\mathcal{P}$ is the sequence $(e_n(\mathcal{P}))_{n=1}^{\infty}$, where $e_n(\mathcal{P})$ is the maximal size of a pregraph in $\mathcal{P}^n$.

It is natural to identify a graph $G = (V, E)$ with the pregraph $\tilde{G} = (V, \emptyset, V^{(2)} \setminus E)$; with this identification we find that $e(G) = e(\tilde{G})$. Hence, for a monotone property $\mathcal{P}$, the two definitions give the same value: in other words, $e(\mathcal{P}^n) = e_n(\mathcal{P})$.

Scheinerman and Zito [28] were the first to study the rate of growth of $|\mathcal{P}^n|$ for a hereditary property. They discovered that, crudely, $|\mathcal{P}^n|$ behaves in one of the following five ways: (i) for n large enough, $|\mathcal{P}^n| = 1$ or 2, (ii) it grows polynomially: for some positive integer k, $a_1 n^k \le |\mathcal{P}^n| \le a_2 n^k$ for some $a_1, a_2 > 0$, (iii) it grows exponentially: $a a_1^n \le |\mathcal{P}^n| \le a_2^n$ for some $a > 0$ and $1 < a_1 \le a_2$, (iv) it grows factorially: $a n^{a_1 n} \le |\mathcal{P}^n| \le n^{a_2 n}$ for some $a > 0$ and $0 < a_1 \le a_2$, (v) it grows superfactorially: $|\mathcal{P}^n| > n^{an}$ for every $a > 0$ and n large enough.

Here we are interested in properties whose rate of growth is not far from maximal. To measure the rate of growth of such a property $\mathcal{P}$, we replace the sequence $(|\mathcal{P}^n|)_{n=1}^{\infty}$ by the sequence $(c_n)_{n=1}^{\infty}$, where $|\mathcal{P}^n| = 2^{c_n \binom{n}{2}}$. Since $1 \le |\mathcal{P}^n| \le 2^{\binom{n}{2}}$, we have $0 \le c_n \le 1$.

We call c_n the *logarithmic density* of $\mathcal{P}^n$, and $c = \lim_{n\to\infty} c_n$ the *asymptotic logarithmic density* of $\mathcal{P}$ provided the limit exists.

Similarly, the (*normalized*) *size* of $\mathcal{P}^n$ is d_n, $0 \le d_n \le 1$, defined by $e_n(\mathcal{P}) = d_n \binom{n}{2}$.

The *asymptotic size* of $\mathcal{P}$ is $d = \lim_{n\to\infty} d_n$, provided this limit exists. Since every pregraph $\tilde{G}$ extends to $2^{e(\tilde{G})}$ graphs, we have $c_n \ge d_n$ for every property. Hence if $\mathcal{P}$ is a property with asymptotic logarithmic density c and asymptotic size d, then $c \ge d$. We shall see later that every hereditary property has an asymptotic logarithmic density c and an asymptotic size d and, in fact, they are equal.

2. MONOTONE PROPERTIES. One of the main aims of classical extremal graph theory is the study of the sequence $(e_n(\mathcal{P}))_{n=1}^{\infty}$ for various monotone graph properties. Frequently, a monotone property is given by a family $\mathcal{F}$ of forbidden subgraphs. For a family $\mathcal{F} = \{F_1, F_2, \ldots\}$ of finite graphs, let $\mathrm{Mon}(\mathcal{F})$ be the collection of all graphs containing no F_i as a subgraph. Clearly every monotone property is of the form $\mathrm{Mon}(\mathcal{F})$ for some family $\mathcal{F}$ of *forbidden subgraphs*, but one is especially interested in monotone families defined by small families of forbidden subgraphs. If there is only one forbidden subgraph F then we have a *principal* monotone property and we write $\mathrm{Mon}(F)$ instead of $\mathrm{Mon}(\{F\})$.

It has been known for over fifty years that every monotone graph property has an asymptotic size. In particular, a weak form of Turán's theorem [31] states that $d(\mathrm{Mon}(K_{r+1})) = 1 - \frac{1}{r}$ for every $r \ge 1$. The fundamental theorem of Erdős and Stone [15] extends this result to $d(\mathrm{Mon}(K_{r+1}(t))) = 1 - \frac{1}{r}$ for all $r, t \ge 1$. Here, as usual, K_n denotes a complete graph of order n and $K_r(t)$ denotes the complete r-partite graph in which each part has t vertices. An equivalent form of

the Erdős-Stone theorem is that if $\mathcal{F}$ is any family of forbidden subgraphs then $d(\mathrm{Mon}(\mathcal{F})) = 1 - \frac{1}{r}$, where $r = \min\{\chi(F) - 1 : F \in \mathcal{F}\}$ and $\chi(F)$ is the chromatic number of F.

Rather more effort is needed to prove that every monotone property has an asymptotic density. Using the method of Kleitman and Rothschild [18], Erdős, Kleitman and Rothschild [12] proved that $c(\mathrm{Mon}(K_{r+1})) = 1 - \frac{1}{r}$. This result was extended by Erdős, Frankl and Rödl [11], who proved that $c(\mathrm{Mon}(F)) = 1 - \frac{1}{r}$ for every graph F, where $r = \chi(F) - 1$. The proof of this result implies that $c(\mathrm{Mon}(\mathcal{F})) = 1 - \frac{1}{r}$ for every family $\mathcal{F}$, where r is, as before, one smaller than the minimal chromatic number of a graph in $\mathcal{F}$. In particular, $c(\mathcal{P}) = d(\mathcal{P})$ for every monotone family.

The structure of K_{r+1}-free graphs was investigated in great detail by Kolaitis, Prőmel and Rothschild [19]. Among other results, they proved that $\mathrm{Mon}(K_{r+1})$ is well approximated by the smaller property $\mathcal{N}_r$ of graphs of chromatic number at most r: not only do we have the crude result that $c(\mathrm{Mon}(K_{r+1})) = c(\mathcal{N}_r)$, but also

$$|\mathrm{Mon}(K_{r+1})^n|/|\mathcal{N}_r^n| = 1 + O(n^{-k})$$

for all $k > 0$. Furthermore, a first-order labelled $0-1$ law holds for the class of K_{r+1}-free graphs.

Before leaving monotone properties, let us note that the following somewhat surprising fact is an immediate consequence of the description of $c(\mathcal{P}) = d(\mathcal{P})$ for a monotone property. If $\mathcal{P}_1$ and $\mathcal{P}_2$ are monotone properties, and $\mathcal{P} = \mathcal{P}_1 \cap \mathcal{P}_2$, then

$$c(\mathcal{P}) = \min\{c(\mathcal{P}_1), c(\mathcal{P}_2)\}. \tag{1}$$

Thus the intersection of two monotone properties is about as large as the smaller of the two properties!

3. VOLUMES OF PROJECTIONS AND ASYMPTOTIC ENUMERATION. The existence of the asymptotic logarithmic density of a hereditary property is closely related to a family of inequalities involving volumes of projections of bodies. Our next aim is to describe this relationship.

A *body* in $\mathbf{R}^n$ is a compact convex subset of $\mathbf{R}^n$ that is the closure of its interior. Let $v_1, \ldots, v_n$ be the standard basis of $\mathbf{R}^n = \mathrm{lin}\{v_1, \ldots, v_n\}$. For a subset A of $[n]$, write K_A for the orthogonal projection of a body K onto $\mathrm{lin}\{v_j : j \in A\}$, and $|K_A|$ for the $|A|$-*dimensional volume* of K_A. In particular, $|K| = |K_{[n]}|$ is the volume of K. With $\beta(K) = (|K_A| : A \subset [n]) = (|K_A|)_{A \subset [n]} \in \mathbf{R}^{\mathcal{P}(n)} = \mathbf{R}^{2^n}$, the map $K \to \beta(K)$ can be considered to be a measure of the size of the boundary of K.

We are interested in the best possible isoperimetric inequalities involving the boundary vector $\beta(K)$ and the volume $|K|$. In other words, we would like to know for which vectors $(x_A) \in \mathbf{R}^{2^n}$ with $x_{[n]} = 1$ is there a body $K \subset \mathbf{R}^n$ of volume 1 such that $|K_A| \le x_A$ for all $A \subset [n]$. The following *box theorem* we proved with Thomason [6] gives a surprisingly simple answer to this question. A *box* B in $\mathbf{R}^n$ is a body of the form $B = \prod_{j=1}^n I_j$, where each I_j is an interval.

THEOREM 1. *For every body $K \subset \mathbf{R}^n$, there is a box $B \subset \mathbf{R}^n$ such that $|B| = |K|$ and $|B_A| \le |K_A|$ for every $A \subset [n]$.*

An immediate consequence of the box theorem is the uniform cover inequality below, extending the Loomis-Whitney inequality [20]. A sequence $(A_i)_{i=1}^m$ of subsets of $[n]$ is a *k-uniform cover* of $[n]$ if every element of $[n]$ belongs to precisely k of the sets $A_1, \ldots, A_m$. Now, if $(A_i)_{i=1}^m$ is a k-uniform cover of $[n]$, and K is a body in $\mathbf{R}^n$ then Theorem 1 implies that

$$|K|^k \leq \prod_{i=1}^{m} |K_{A_i}|. \tag{2}$$

In fact, in [6] the box theorem is deduced from the uniform cover inequality (1) by a simple compactness argument. Since the original proof, several other deductions have been suggested: Ball noted that separation theorems, and Kahn and Meshulam pointed out that properties of submodular functions, can be used to deduce the box theorem from inequality (2).

The box theorem easily implies that, as first proved by Alekseev [1], every hereditary property of graphs has an asymptotic logarithmic density.

THEOREM 2. *Let $\mathcal{P}$ be a hereditary property of graphs. Then $1 = c_1(\mathcal{P}) \geq c_2(\mathcal{P}) \geq \cdots$; in particular, the asymptotic logarithmic density $c(\mathcal{P}) = \lim_{n\to\infty} c_n(\mathcal{P})$ exists.*

It is easily seen that the arguments above apply to hereditary properties of r-uniform hypergraphs as well, *mutatis mutandis*.

4. ASYMPTOTIC ENUMERATION AND GLOBAL STRUCTURE. Given a family $\mathcal{F} = \{F_1, F_2, \ldots\}$, of finite graphs, let $\mathrm{Her}(\mathcal{F})$ be the collection of all graphs that contain no F_i as an *induced* subgraph. Clearly, every hereditary property is of the form $\mathrm{Her}(\mathcal{F})$ for some family $\mathcal{F}$ of *forbidden subgraphs*. Theorem 2 tells us that every hereditary property $\mathcal{P} = \mathrm{Her}(\mathcal{F})$ has an asymptotic logarithmic density $c(\mathcal{P})$, but gives no indication as to how one could determine $c(\mathcal{P})$ from $\mathcal{F}$. In fact, Prömel and Steger [22], [23], [24], [25] gave such a description for a principal hereditary property, i.e., for one with a single forbidden induced subgraph. They also gave approximations of principal hereditary properties by rather simple (non-principal) hereditary properties. With Thomason [7] we extended these results to general hereditary properties.

Before we can describe these results, we have to introduce some definitions. An *(r, s)-colouring* of a graph $G = (V, E)$ is a partition of the vertex set into r classes such that the first s classes induce complete graphs, and the remaining $r - s$ classes induce empty subgraphs. (Needless to say, empty classes are allowed.) Thus an $(r, 0)$-colouring of a graph is precisely a standard r-colouring. We write $\mathcal{P}_{r,s}$ for the collection of all (r, s)-colourable graphs; clearly, $\mathcal{P}_{r,s}$ is a hereditary property for all $0 \leq s \leq r$, $r \geq 1$. For example, $\mathcal{P}_{1,1}$ is the collection of all complete graphs, and $\mathcal{P}_{1,0}$ is the collection of all empty graphs. The *colouring number* $r(\mathcal{P})$ of a property $\mathcal{P}$ is

$$r(\mathcal{P}) = \max\{r : \mathcal{P}_{r,s} \subset \mathcal{P} \text{ for some } s\}.$$

Note that if $\mathcal{P} = \mathrm{Her}(\mathcal{F})$ then

$$r(\mathcal{P}) = \max\{r : \text{ for some } s \leq r, \text{ no } F \in \mathcal{F} \text{ is } (r, s)\text{-colourable}\}.$$

If $\mathcal{P} = \mathrm{Mon}(\mathcal{F})$ then $r(\mathcal{P})$ is exactly as before:

$$r(\mathcal{P}) = \min\{\chi(F) - 1 : F \in \mathcal{F}\} = \max\{r : \text{ no } F \in \mathcal{F} \text{ is } (r,0)\text{-colourable}\}.$$

The colouring number gives us a lower bound for c_n and d_n. Indeed, let $0 \le s \le r$ be such that $r = r(\mathcal{P})$ and $\mathcal{P}_{r,s} \subset \mathcal{P}$, and let $\tilde{G} = ([n], \tilde{E}, \tilde{N})$ be the pregraph obtained as follows. Partition $[n]$ into r classes as equal as possible in size, $[n] = V_1 \cup \ldots \cup V_r$, say, and let $\tilde{E}$ consist of all edges within a class V_i for $0 \le i \le s$. Since $\mathcal{P}_{r,s} \subset \mathcal{P}$, every extension of $\tilde{G}$ belongs to $\mathcal{P}$. Consequently,

$$c_n(\mathcal{P}) \ge d_n(\mathcal{P}) \ge e(\tilde{G}) / \binom{n}{2} \ge 1 - \frac{1}{r}.$$

As shown in [7], $c(\mathcal{P})$ and $d(\mathcal{P})$ exist for every hereditary property, and these inequalities are essentially best possible.

THEOREM 3. *If $\mathcal{P}$ is any hereditary property then*

$$c(\mathcal{P}) = d(\mathcal{P}) = 1 - \frac{1}{r(\mathcal{P})},$$

where $r(\mathcal{P})$ is the colouring number of $\mathcal{P}$.

The proof of this theorem is based on the three pillars of extremal graph theory: the theorems of Ramsey [26], Erdős and Stone [15], and Szemerédi [30]. One needs only the very simple case of Ramsey's theorem that the diagonal graph Ramsey function is finite: $R(k) < \infty$ for every k. On the other hand, one needs a slight extension of the Erdős-Stone theorem: for all $r, t \ge 1$ and $\epsilon > 0$ there are $\delta > 0$ and $n_0 \in \mathbf{N}$ such that if F and G are graphs with $V(F) = V(G) = [n]$, $n \ge n_0$, $e(F) \le \delta n^2$ and

$$e(G) \ge (1 - \frac{1}{r} + \epsilon)\binom{n}{2},$$

then G contains an F-avoiding $K_{r+1}(t)$. Here we say that a graph H *avoids* F if no edge of F joins two vertices of H.

The most important ingredient of the proof of Theorem 3 is *Szemerédi's uniformity lemma* [30]. Given a graph $G = (V, E)$, and subsets $A, B, \subset V$, the *density* $d(A, B)$ is defined as

$$d(A, B) = \frac{e(A, B)}{|A||B|},$$

where $e(A, B)$ is the number of A-B edges. A pair (A, B) is (ϵ, δ)-uniform if

$$|d(A', B') - d(A, B)| \le \epsilon$$

whenever $A' \subset A$, $B' \subset B$, $|A'| \ge \delta|A|$ and $|B'| \ge \delta|B|$.

Szemerédi's uniformity lemma states that for all $\epsilon, \delta, \eta > 0$ there is an $M = M(\epsilon, \delta, \eta)$ such that the vertex set of every graph G can be partitioned into at most M sets $U_1, \ldots, U_m$ of sizes differing by at most 1, such that at least $(1 - \eta)m^2$ of the (ordered) pairs (U_i, U_j) are (ϵ, δ)-uniform.

The fewer sets $U_1, U_2, \ldots$ we can take the more powerful the result is; unfortunately when $\epsilon = \delta = \eta$, all we know about $M(\epsilon, \epsilon, \epsilon)$ is that it is at most a tower of 2s of height proportional to ϵ^{-5}. As the proof of this bound seemed rather 'wasteful', for many years there had been some hope that this enormous bound could be reduced greatly. It was a great surprise when recently Gowers [17] proved the difficult result that $K(\epsilon, \delta, \eta)$ can not be less than of tower type in $1/\delta$, even when ϵ and η are kept large.

Szemerédi's uniformity lemma implies that every graph satisfying certain global conditions contains appropriate *induced* subgraphs; this is precisely how the lemma was used in the proof of Theorem 3.

The descriptions of the asymptotic logarithmic density and asymptotic size of a hereditary property provided by Theorem 3 imply that hereditary properties are much more complex than monotone ones. In particular, the simple relationship (1) fails for hereditary properties. For example, if $\mathcal{P}_1 = \mathrm{Her}(K_4)$, $\mathcal{P}_2 = \mathrm{Her}(C_7)$ and $\mathcal{P} = \mathcal{P}_1 \cap \mathcal{P}_2 = \mathrm{Her}\{K_4, C_7\}$, then $r(\mathcal{P}_1) = r(\mathcal{P}_2) = 3$ but $r(\mathcal{P}) = 2$: the intersection of two hereditary properties can be *much* smaller than either of them.

In fact, the intersection of two large hereditary properties need not even be a property in our sense: it may contain only finitely many non-isomorphic graphs. For example, if $r \geq 1$ then each of $\mathcal{P}_{r,0}$ and $\mathcal{P}_{r,r}$ has colouring number r, so that $c(\mathcal{P}_{r,0}) = c(\mathcal{P}_{r,r}) = 1 - \frac{1}{r}$, but $\mathcal{P}_{r,0} \cap \mathcal{P}_{r,r}$ consists of graphs G with $\chi(G) \leq r$ and $\chi(\bar{G}) \leq r$. In particular, $|G| \leq r^2$ for every $G \in \mathcal{P}_{r,0} \cap \mathcal{P}_{r,r}$, so $\mathcal{P}_{r,0} \cap \mathcal{P}_{r,r}$ indeed consists only of finitely many non-isomorphic graphs.

5. COLOURING RANDOM GRAPHS $G_{n,1/2}$ *with hereditary properties*. The random graph $G_{n,p}$ is a graph with vertex set $[n]$, whose edges are selected independently, with probability p. The probability space of these graphs is $\mathcal{G}(n,p)$. In particular, $\mathcal{G}(n,1/2)$ is the space of all $2^{\binom{n}{2}}$ graphs on $[n]$ with the uniform distribution.

One of the main questions left open by Erdős and Rényi when, almost forty years ago, they founded the theory of random graphs ([13], [14]; see also [5]) was the behaviour of the chromatic number of a random graph. Over 25 years later, first Shamir and Spencer [29] proved that the chromatic number of $G_{n,p}$ is highly concentrated, and then it was shown [3] that if $0 < p < 1$ is fixed and $q = 1 - p$ then

$$\chi(G_{n,p}) = (1 + o(1)) \frac{n}{2 \log_{1/q} n} \tag{3}$$

for almost every $G_{n,p}$. Substantial extensions of this result were proved by Łuczak [21], Frieze and Łuczak [16], and Alon and Krivelevich [2]. All these results use various martingale inequalities (see [4]).

For a property $\mathcal{P}$, a *$\mathcal{P}$-colouring* of a graph $G = (V, E)$ is a partition $V = V_1 \cup \ldots \cup V_k$ of the vertex set such that every class V_i induces a $\mathcal{P}$-graph: $G[V_i] \in \mathcal{P}, i = 1, \ldots, k$. The *$\mathcal{P}$-chromatic number* $\chi_{\mathcal{P}}(G)$ of a graph G is the minimal number of classes in a $\mathcal{P}$-colouring of G. Thus $\chi_{\mathcal{P}_{1,0}}(G) = \chi(G)$ and $\chi_{\mathcal{P}_{1,1}}(G) = \chi(\overline{G})$. Scheinerman [27] was the first to study the $\mathcal{P}$-chromatic number of random graphs. He noted that if $\mathcal{P}$ is a hereditary property then either $\mathcal{P}_{1,0} \subset \mathcal{P}$ or $\mathcal{P}_{1,1} \subset \mathcal{P}$ so $\chi_{\mathcal{P}}(G) \leq \max\{\chi(G), \chi(\overline{G})\}$. From this it follows that $\chi_{\mathcal{P}}(G_{n,p}) = O(n \log n)$ for every fixed $0 < p < 1$ and hereditary property $\mathcal{P}$, and it is easily seen that, in fact, $\chi_{\mathcal{P}}(G_{n,p}) = \Theta(n \log n)$.

With Thomason [8] we proved an analogue of (3) for a general hereditary property, but only in the case $p = \frac{1}{2}$.

THEOREM 4. *Let $\mathcal{P}$ be a non-trivial hereditary property of graphs, with colouring number $r = r(\mathcal{P})$. Then*

$$\chi_{\mathcal{P}}(G_{n,1/2}) = (\frac{1}{2r} + o(1))\frac{n}{\log_2 n}$$

for almost every $G_{n,1/2}$.

In fact, this result follows rather easily from (3) and from the facts that $c(\mathcal{P}) = 1 - \frac{1}{r}$ and that $\mathcal{P}_{r,s} \subset \mathcal{P}$ for some $s, 0 \le s \le r$. More precisely, $c(\mathcal{P}) = 1 - \frac{1}{r}$ implies that $\chi_{\mathcal{P}}(G_{n,1/2})$ is unlikely to be much smaller than $n/(2r \log_2 n)$, and $\mathcal{P}_{r,s} \subset \mathcal{P}$ implies that $\chi_{\mathcal{P}}(G_{n,1/2})$ is unlikely to be much larger than $n/(2r \log_2 n)$.

6. COLOURING RANDOM GRAPHS $G_{n,p}$ WITH HEREDITARY PROPERTIES. The accepted wisdom in the theory of random graphs is that whatever can be proved for the space $\mathcal{G}(n,p)$ with $p = 1/2$ can be proved for $\mathcal{G}(n,p)$ with any fixed p, $0 < p < 1$. This conventional wisdom is contradicted by the problem of determining $\chi_{\mathcal{P}}(G_{n,p})$! As we saw in Theorem 4, it is easy to determine $\chi_{\mathcal{P}}(G_{n,p})$ in the uniform case $p = 1/2$. However, for $p \neq 1/2$ not only does the proof collapse, but we are faced with a genuinely more complicated phenomenon, so that much more effort is needed to overcome the difficulties.

A lower bound for $\chi_{\mathcal{P}}(G_{n,p})$ is easily obtained from the following result, which is a consequence of the box theorem.

THEOREM 5. *Let $\mathcal{P}$ be a hereditary graph property, let $0 < p < 1$ and let the constants $e_{k,p}(\mathcal{P})$ be defined by $\mathbf{P}(G_{k,p} \in \mathcal{P}) = 2^{-e_{k,p}(\mathcal{P})\binom{k}{2}}$. Then $e_{k,p}(\mathcal{P})$ increases with k. In particular, $e_{k,p}(\mathcal{P})$ tends to a limit $e_p(\mathcal{P})$ as $k \to \infty$. Furthermore, $e_p(\mathcal{P}) > 0$ if $\mathcal{P}$ is non-trivial, i.e., if not every graph has $\mathcal{P}$.*

Theorem 5 implies that, for $\epsilon > 0$, the expected number of induced subgraphs of order k in a random graph $G_{n,p}$ having property $\mathcal{P}$ is $o(1)$ for $k \ge (2/e_p + \epsilon)\log_2 n$, and tends to infinity for $k \le (2/e_p - \epsilon)\log_2 n$. Consequently,

$$\chi_{\mathcal{P}}(G_{n,p}) \ge (e_p + o(1))n/(2\log_2 n) \tag{4}$$

almost surely.

It was conjectured in [8] that (4) is in fact an equality, as claimed by Theorem 4 for $p = 1/2$. Now, the proof of Theorem 4 is based on the fact that for $p = 1/2$ the constant $e_p(\mathcal{P})$ has a simple interpretation in terms of the values (r,s) for which $\mathcal{P}_{r,s} \subset \mathcal{P}$. However, for $p \neq 1/2$ this is no longer true: $e_p(\mathcal{P})$ cannot be characterized solely in terms of these values (r,s). For example, let $\mathcal{P}' = \mathcal{P}_{2,0}$ be the property of being bipartite, and let $\mathcal{P}''$ be the property of being 3-colourable, with two of the colour classes spanning complete bipartite graphs. Then $\mathcal{P}'$ and $\mathcal{P}''$ contain $\mathcal{P}_{1,0}$ and $\mathcal{P}_{2,0}$, and no other $\mathcal{P}_{r,s}$. Nevertheless, $e_p(\mathcal{P}') \neq e_p(\mathcal{P}'')$ for $p > 1/2$.

In spite of these difficulties, with Thomason [9] we proved the conjecture above.

THEOREM 6. *Let $\mathcal{P}$ be a hereditary graph property and let $0 < p < 1$. Let $e_p = e_p(\mathcal{P})$ be the constant defined in Theorem 5. Then*

$$\chi_{\mathcal{P}}(G_{n,p}) = (e_p + o(1))n/(2\log_2 n)$$

almost surely.

The proof of Theorem 6 makes use of Szemerédi's uniformity lemma, martingale inequalities and, above all, a careful study of the structure of a general hereditary property. The *product* $\prod_{\gamma\in\Gamma}\mathcal{P}_\gamma$ of hereditary properties $\mathcal{P}_\gamma$, $\gamma \in \Gamma$, is the class of graphs G with vertex sets $\bigcup_{\gamma\in\Gamma} V_\gamma$ such that $G[V_\gamma] \in \mathcal{P}_\gamma$ for every $\gamma \in \Gamma$. A hereditary property is *irreducible* if it is not the product of two other hereditary properties. It is easily shown that every hereditary property is the product of a finite collection of irreducible hereditary properties. Also, if $\mathcal{P} = \prod_{\gamma\in\Gamma}\mathcal{P}_\gamma$ then

$$e_p(\mathcal{P})^{-1} = \sum_{\gamma\in\Gamma} e_p(\mathcal{P}_\gamma)^{-1}.$$

Next, one can show that if Theorem 6 holds for each of the properties $\mathcal{P}_1, \dots, \mathcal{P}_k$, then it holds for $\prod_{i=1}^k \mathcal{P}_i$ as well. Consequently, it suffices to prove Theorem 6 for irreducible properties.

In fact, the heart of the proof is the assertion that Theorem 6 holds for every 'typed' property $\mathcal{P} = \mathcal{P}(\tau)$. A *type* is a labelled graph, each of whose vertices and edges is coloured black or white. Given a type τ, the property $\mathcal{P}(\tau)$ consists of those graphs G for which $V(G)$ has a partition $\bigcup_{t\in V(\tau)} V_t$ such that $G[V_t]$ is complete or empty according as t is black or white, and moreover, if the edge tu is in τ then $G[V_t, V_u]$ is a complete or empty bipartite graph according as the edge tu is black or white. The proof of the fact that Theorem 6 holds for typed properties $\mathcal{P}(\tau)$ is based on a careful analysis of the maximal number of induced edge-disjoint subgraphs of a given order having property $\mathcal{P}$ – after much work enough can be deduced so that martingale inequalities can be applied.

7. *Open problems.* Numerous open problems remain. Concerning graphs, all the discussion above is about rather 'rich' properties $\mathcal{P}$, namely those with $c(\mathcal{P}) > 0$. The case $c(\mathcal{P}) = 0$ is not understood nearly as well.

Although we know that $c(\mathcal{P}) = d(\mathcal{P})$ for every hereditary property, this is far from being the entire story. We always have

$$|\mathcal{P}^n| = 2^{c_n\binom{n}{2}} \geq 2^{e_n(\mathcal{P})} = 2^{d_n\binom{n}{2}},$$

but it would be good to decide whether $c_n = (1 + o(1))d_n$ holds as well.

More importantly, we know very little about hypergraphs. The quantities $c_n(\mathcal{P})$ and $d_n(\mathcal{P})$ are easily defined for r-graphs, and $c_n(\mathcal{P}) \geq d_n(\mathcal{P})$ for every n. Also, the box theorem implies that $c_n(\mathcal{P}) \to c(\mathcal{P})$, and one can show that $d_n(\mathcal{P}) \to d(\mathcal{P})$, but we do not know whether we always have $c(\mathcal{P}) = d(\mathcal{P})$. Nothing of importance is know about the $\mathcal{P}$-chromatic number of r-graphs: we do not even know the asymptotic $\mathcal{P}$-chromatic number of random r-graphs $G^{(r)}_{n,p}$ for $p = 1/2$.

References

[1] V.E. Alekseev, On the entropy values of hereditary classes of graphs, *Discrete Math. Appl.* 3 (1993), 191–199.

[2] N. Alon and M. Krivelevich, The concentration of the chromatic number of random graphs, *Combinatorica* 17 (1997), 303–313.

[3] B. Bollobás, The chromatic number of random graphs, *Combinatorica* 8 (1988), 49–55.

[4] B. Bollobás, Martingales, isoperimetric inequalities and random graphs, *in "Combinatorics (Eger, 1987)"* (A. Hajnal, L. Lovász and V.T. Sós, eds.), Colloq. Math. Soc. János Bolyai 52, North-Holland, Amsterdam (1988), 113–139.

[5] B. Bollobás, *Random Graphs*, Academic Press, London, 1985, xvi+447pp.

[6] B. Bollobás and A. Thomason, Projections of bodies and hereditary properties of hypergraphs, *J. London Math. Soc.* 27 (1995), 417–424.

[7] B. Bollobás and A. Thomason, Hereditary and monotone properties of graphs, in "The Mathematics of Paul Erdős II" (R.L. Graham and J. Nešetřil, eds.) *Algorithms and Combinatorics*, Vol. 14, Springer-Verlag (1997), 70–78.

[8] B. Bollobás and A. Thomason, Generalized chromatic numbers of random graphs, *Random Structures and Algorithms* 6 (1995), 353–356.

[9] B. Bollobás and A. Thomason, Colouring random graphs by hereditary properties, to appear.

[10] B. Bollobás and P. Erdős, On the structure of edge graphs, *Bull. London Math. Soc.* 5 (1973), 317–321.

[11] P. Erdős, P. Frankl and V. Rödl, The asymptotic enumeration of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent, *Graphs and Combinatorics* 2 (1986), 113–121.

[12] P. Erdős, D.J. Kleitman and B.L. Rothschild, Asymptotic enumeration of K_n-free graphs, in *International Coll. Comb.*, Atti dei Convegni Lincei (Rome) 17 (1976), 3–17.

[13] P. Erdős and A. Rényi, On random graphs I, *Publ. Math. Debrecen* 6 (1959), 290–297.

[14] P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hungar. Acad. Sci.* 5 (1961), 17–61.

[15] P. Erdős and A.H. Stone, On the structure of linear graphs, *Bull. Amer. Math. Soc.* 52 (1946), 1087–1091.

[16] A.M. Frieze and T. Łuczak, On the independence and chromatic numbers of random regular graphs, *J. Combinat. Theory (B)* 54 (1992), 123–132.

[17] W.T. Gowers, Lower bounds of tower type for Szemerédi's uniformity lemma, *Geom. Funct. Anal.* 7 (1997), 322-332.

[18] D.J. Kleitman and B.L. Rothschild, Asymptotic enumeration of partial orders on a finite set, *Trans. Amer. Math. Soc.* 205 (1975), 205–220.

[19] Ph.G. Kolaitis, H.J. Prömel and B.L. Rothschild, K_{l+1}-free graphs: asymptotic structure and a 0–1 law, *Trans. Amer. Math. Soc.* 303 (1987), 637–671.

[20] L.H. Loomis and H. Whitney, An inequality related to the isoperimetric inequality, *Bull. Amer. Math. Soc.*, 55 (1949), 961–962.

[21] T. Łuczak, The chromatic number of random graphs, *Combinatorica* 11 (1991), 45–54.

[22] H.J. Prömel and A. Steger, Excluding induced subgraphs: quadrilaterals, *Random Structures and Algorithms* 2 (1991), 55–71.

[23] H.J. Prömel and A. Steger, Excluding induced subgraphs II: extremal graphs, *Discrete Applied Mathematics* 44 (1993), 283–294.

[24] H.J. Prömel and A. Steger, Excluding induced subgraphs III: a general asymptotic, *Random Structures and Algorithms* 3 (1992), 19–31.

[25] H.J. Prömel and A. Steger, The asymptotic structure of H-free graphs, in *Graph Structure Theory* (N. Robertson and P. Seymour, eds), Contemporary Mathematics 147, Amer. Math. Soc., Providence, 1993, pp. 167-178.

[26] F.P. Ramsey, On a problem of formal logic, *Proc. London Math. Soc.* 30(2) (1929), 264–286.

[27] E.R. Scheinerman, Generalized chromatic numbers of random graphs, *SIAM J. Discrete Math.* 5 (1992), 74–80

[28] E.R. Scheinerman and J. Zito, On the size of hereditary classes of graphs, *J. Combinat. Theory (B)* 61 (1994), 16–39.

[29] E. Shamir and J. Spencer, Sharp concentration of the chromatic number on random graphs $G_{n,p}$, *Combinatorica* 7 (1987), 121–129.

[30] E. Szemerédi, Regular partitions of graphs, in *Proc. Colloque Inter. CNRS* (J.-C. Bermond, J.-C. Fournier, M. las Vergnas, D. Sotteau, eds), 1978.

[31] P. Turán, On an extremal problem in graph theory (in Hungarian), *Mat. Fiz. Lapok* 48 (1941), 436–452.

Béla Bollobás
Dept. of Mathematics,
Univ. of Memphis, USA
and
Trinity College,
Univ. of Cambridge,
England

Applications of Relaxed Submodularity

András Frank

Abstract. Combinatorial optimization problems often give rise to set-functions which satisfy the sub- or supermodular inequality only for certain pairs of subsets. Here we discuss connectivity problems and show how results on relaxed submodular functions help in solving them.

1991 Mathematics Subject Classification: 90C27, 05C40
Keywords and Phrases: combinatorial optimization, submodular functions, connectivity of graphs

1. Introduction

Let V be a finite set and $b : 2^S \to \mathbf{R} \cup \{\infty\}$ and $p : 2^S \to \mathbf{R} \cup \{-\infty\}$ two set-functions. The submodular and the supermodular inequality, respectively, for subsets $X, Y \subseteq V$ are, as follows:

$$b(X) + b(Y) \geq b(X \cap Y) + b(X \cup Y), \tag{1.1b}$$

$$p(X) + p(Y) \leq p(X \cap Y) + p(X \cup Y). \tag{1.1p}$$

Function b [respectively, p] is called *fully submodular* if (1.1b) [*fully supermodular* if (1.1p)] holds for every two subsets $X, Y \subseteq V$. (When equality holds everywhere, we speak of a modular function.) We call a function *semimodular* if it is submodular or supermodular.

Semimodular functions proved to be extremely powerful in combinatorial optimization. One intuitive explanation for this is that submodular functions may be considered as discrete counterparts of convex functions. For example, L. Lovász [L83] observed that a (natural) linear extension of an arbitrary set-function h to a real function on $\mathbf{R}_+^V$ is convex if and only if h is submodular. Another occurrence of this relationship is the discrete separation theorem [F82] asserting that

Research supported by the Hungarian National Foundation for Scientific Research Grant, OTKA T17580.

The paper was completed while the author visited the Department of Mathematics, EPFL, Lausanne, June 1998.

if an integer-valued supermodular function p is dominated by an integer-valued submodular function b, then there is an integer-valued (!) modular function m for which $p \leq m \leq b$. Recently, this kind of analogy has been developed systematically by K. Murota [M96] into a theory relating convex analysis and discrete optimization.

In applications, however, often the submodular inequality is not fulfilled by every pair of sets. Accordingly, several frameworks concerning semimodular functions have been introduced, analyzed, and applied. One fundamental property of these models is total dual integrality (TDI-ness) which ensured applicability to weighted optimization problems, as well. (See [Schrijver, 1984], for an account.) For example, C. Lucchesi and D. Younger [LY78] proved a min-max formula for the minimum number of edges of a directed graph whose contraction results in a strongly connected digraph. J. Edmonds and R. Giles [EG76], by introducing submodular flows, found an extension to a minimum cost version. Based on this ground, a polynomial time algorithm was developed in [F81] to actually find the cheapest edge set.

There have been optimization problems, however, where the minimum cardinality case was nicely treatable while the min-cost version was NP-complete. For example, making a digraph strongly connected by adding new edges is such a problem [Eswaran and Tarjan, 1976]. This type of connectivity augmentation problems gave rise recently to a new class of results concerning relaxed semimodular functions.

In this paper we outline the new frameworks, exhibit recent developments concerning submodular flows, and show applications to problems from the area of graph connectivity.

The following forms of relaxed semimodularity will be used. Let S and T be two subsets of a groundset V and b a set-function. b is *intersecting* submodular if (1.1b) holds whenever $X \cap Y \neq \emptyset$. b is *crossing* submodular if (1.1b) holds whenever $X \cap Y \neq \emptyset$ and $V - (X \cup Y) \neq \emptyset$. Intersecting and crossing supermodular funtions are defined analogously but for supermodularity we need further relaxations. Let p be a non-negative set-function. p is *ST-crossing* supermodular if (1.1p) holds whenever $p(X) > 0, p(Y) > 0$, $X \cap Y \cap T \neq \emptyset$ and $S - (X \cup Y) \neq \emptyset$. p is *T-intersecting* supermodular if (1.1p) holds whenever $p(X) > 0, p(Y) > 0$, $X \cap Y \cap T \neq \emptyset$. p is *skew supermodular* if $p(X) + p(Y) \leq \max(p(X \cap Y) + p(X \cup Y), p(X - Y) + p(Y - X))$ whenever $p(X) > 0, p(Y) > 0$. We call a set-function p *symmetric* if $p(X) = p(V - X)$ for every $X \subseteq V$. Throughout we will assume that the occurring set-functions are integer-valued.

2. Connectivity problems

In a graph or digraph G, $\lambda(u, v)$ (respectively, $\kappa(u, v)$) denotes the maximum number of edge-disjoint (openly disjoint) paths from u to v. $\lambda(u, v)$ is called the *local edge-connectivity from u to v* while the minimum of these λ-values (κ-values) is the *edge-connectivity (node-connectivity)* of G. A digraph is *k-edge- (node-) connected from root s* if $\lambda(s, v) \geq k$ ($\kappa(s, v) \geq k$) for every $v \in V$.

The problems we consider can be cast in the following general form: Create an (optimal) graph (or digraph, or hypergraph) satisfying some connectivity properties. Sometimes we are interested only in the existence of the requested object, other times finding an optimal object is also important. A connectivity property typically means that bounds are imposed on the number of edges (nodes) in cuts. "Creating" means that certain specified operations are allowed. We will consider the following operations: Given a graph or digraph, take a subgraph, take a supergraph (that is, augment the graph), orient the undirected edges, reorient some of the directed edges.

The travelling salesman problem, for example, is a special case, as it requires finding a minimum cost 2-edge-connected subgraph of n edges. Another special case is the Steiner-tree problem which seeks for cheapest subgraphs containing at least one edge from each cut separating a specified set T of terminal nodes. These well-known NP-complete problems are special cases of several other connectivity problems. On the positive side, the problem of finding a minimum cost subdigraph of a digraph that contains k edge-disjoint paths from s to t is a special min-cost flow problem and hence it is solvable in polynomial time. Here we consider other connectivity problems having a good characterization and/or a polynomial-time solution algorithm. Some of them are, as follows.

Subgraph problems

S1. Given a graph and a stable set S, find a (minimum cost) spanning tree satisfying upper and lower bound requirements for its degree of the nodes in S.

S2. Given a digraph with a root s, find a cheapest subgraph which is k-edge-(node-) connected from s.

Supergraph (=augmentation) problems

A1. Given a digraph, add a cheapest subset of new edges to get a k rooted edge-connected digraph.

A2. Given a digraph, add a minimum number of new edges to get a k-edge-(node-) connected digraph.

A3. Given a digraph and two subsets S and T of nodes, add a minimum number of new edges from S to T to get a digraph with $\lambda(s,t) \geq k$ (resp., $\kappa(s,t) \geq k$) whenever $s \in S, t \in T$.

A4. Given a hypergraph, add a minimum number of edges to obtain a k-edge-connected hypergraph.

Orientation problems

O1. Given a graph, orient the edges to get a digraph which is k-edge-connected from a root s and l-edge-connected to s. (When $k = l$, the digraph is just k-edge-connected).

O2. Given a mixed graph, orient its undirected edges so as to obtain a k-edge-connected digraph.

O3. Given a digraph with edge-costs, reorient a cheapest subset of edges to get a k-edge-connected digraph.

Problem S1 is a matroid intersection problem and therefore Edmonds' [E79] intersection theorem and algorithm apply. A solution to Problem S2 requires submodular flows, the topic of Section 3. Problem A1 may be formulated as a special case of S2, but the other augmentation problems need different techniques, to be discussed in Sections 4 and 5. All the orientation problems will be handled with the help of submodular flows.

3. Submodular flows

Let V be a ground-set and b an integer-valued set-function with $b(\emptyset) = 0$. Associate with b a polyhedron $B(b) := \{x \in R^V : x(V) = b(V), x(A) \leq b(A)$ for every $A \subseteq V\}$. When b is fully submodular, $B(b)$ is called a *base-polyhedron* (0-*base-polyhedron* in case $b(V) = 0$). For convenience, the empty set is also considered a base-polyhedron. It follows from the work of J. Edmonds [E70] that a non-empty base-polyhedron uniquely determines its defining fully submodular function. Moreover, the intersection of two base-polyhedra is integral (a version of Edmonds' polymatroid intersection theorem). Therefore it is important that weaker functions may also define base-polyhedra. For example, L. Lovász [L83] proved that if b is intersecting submodular, then $B := B(b)$ is a base-polyhedron which is non-empty if and only if $b(V) \geq \sum_i b(V_i)$ holds for every partition $\{V_1, \ldots, V_t\}$ of V. Moreover, the unique fully submodular function defining B is $b^{\downarrow}(Z) := \min(\sum_i b(Z_i) : \{Z_i\}$ a partition of $Z)$. S. Fujishige [Fu84] extended this result to crossing submodular functions. He showed that $B(b)$ is a base-polyhedron if b is crossing submodular. Moreover, $B := B(b)$ is non-empty (assuming $b(V) = 0$) if and only if $\sum_i b(Z_i) \geq 0$ and $\sum_i b(V - Z_i) \geq 0$ for every partition $\{Z_1, \ldots, Z_t\}$ of V.

What is the unique fully submodular function defining B, provided B is non-empty? We need the following notion of tree-composition of sets. The tree-composition of the ground-set V is either a partition of V or a co-partition of V (the complements of a partition of V.) Let A be a proper non-empty subset of V. Let $\{A_1, \ldots, A_k\}$ $(k \geq 1)$ be a partition of A and $\{B_1, \ldots, B_l\}$ $(l \geq 1)$ a partition of $B := V - A$. Let $U := \{a_1, \ldots, a_k, b_1, \ldots, b_l\}$ be a set of new elements and define $\varphi(v) := a_i$ if $v \in A_i$ and $:= b_j$ if $v \in B_j$. Let F be a directed tree defined on U so that every edge is of form $b_i a_j$. For every edge e of the tree, $F - e$ has two components, among which F_e denotes the one entered by e. Now a *tree-composition* of A is a family of subsets of V given in form $\{\varphi^{-1}(F_e) : e \in E(F)\}$. (A tree-composition has at most $|V| - 1$ members.)

THEOREM 3.1 [F96] *Let b be a crossing submodular function for which $b(V) = 0$ and $B := B(b)$ is non-empty. Then the unique fully submodular function $b^{\downarrow}$ defining B is given by $b^{\downarrow}(Z) = \min(b(\mathcal{F}) : \mathcal{F}$ a tree-composition of $Z)$.*

Submodular flows provide a general and powerful framework for combinatorial optimization problems. Let $\vec{G} = (V, \vec{E})$ be a directed graph. Let $f : \vec{E} \to \mathbf{Z} \cup \{-\infty\}$ and $g : \vec{E} \to \mathbf{Z} \cup \{+\infty\}$ be such that $f \leq g$. For a function $z : \vec{E} \to \mathbf{R}$ let $\varrho_z(A) := \sum(z(e) : e$ enters $A)$ and $\delta_z(A) := \sum(z(e) : e$ leaves $A)$. Let $\lambda_z(A) := \varrho_z(A) - \delta_z(A)$. Note that λ_z is modular, that is, $\lambda_z(A) = \sum_{v \in A}(\lambda_z(v))$ and

therefore we may consider λ_z as a function on V. Furthermore, let $b : 2^V \to \mathbf{Z} \cup \{\infty\}$ be a crossing submodular function with $b(V) = 0$. We call $z : \vec{E} \to \mathbf{R}$ a *submodular flow* (with respect to b) if

$$\lambda_z(A) \leq b(A) \text{ for every } A \subseteq V. \tag{3.1a}$$

Submodular flow z is *feasible* if

$$f \leq z \leq g. \tag{3.1b}$$

Submodular flows were introduced and investigated by J. Edmonds and R. Giles [EG77]. Their fundamental result asserts that the linear system (3.1) is totally dual integral, that is, the dual linear programming problem to $\max(cz : z$ satisfies (3.1)) has an integer-valued optimal solution for every integer-valued c for which the optimum exists. It follows that the primal polyhedron is also integral (i.e., every face contains an integer point) if b, f, g are integer-valued.

This result implies for example (a min-cost extension of) a theorem of C. Lucchesi and D. Younger asserting that a digraph (with no cut-edge) can be made strongly connected by reorienting at most γ edges if and only if there are no $k+1$ disjoint directed cuts. Another direct consequence of the integrality of the submodular flow polyhedron is the (weak form of an) orientation theorem of C. Nash-Williams [N60] asserting that a $2k$-edge-connected undirected graph always has a k-edge-connected orientation.

In applications, we often need criteria for feasibility which are easy to handle. An easy relationship between submodular flows and base-polyhedra enables us to formulate such a result. Namely, z is a submodular flow if and only if λ_z belongs to the base-polyhedron $B(b)$. The following was proved in [F82]. Where b is fully submodular, there exists an integer-valued feasible submodular flow if and only if $\varrho_f(A) - \delta_g(A) \leq b(A)$ holds for every $A \subseteq V$. (Note that, in the special case of $b \equiv 0$, we obtain Hoffman's circulation feasibility theorem.) When this result is combined with Theorem 3.1, one obtains the following:

THEOREM 3.2 *Let b be (A) an intersecting or (B) a crossing submodular function. There exists an integer-valued feasible submodular flow if and only if*

$$\varrho_f(A) - \delta_g(A) \leq b(\mathcal{A}) \tag{3.2}$$

holds for every $A \subseteq V$ and for every partition $\mathcal{A}$ of A in case (A) and for every tree-composition $\mathcal{A}$ of A in case (B).

The partition-type condition for *(A)* is easier to handle than the one including tree-compositions. Although there are important cases where tree-compositions cannot be avoided, in the next two special cases partition-type conditions turn out to be sufficient. As a generalization of Case *(A)* in Theorem 3.2, one has the following.

THEOREM 3.3 *Suppose that b is crossing submodular (with $b(V) = 0$) which, in addition, satisfies (1.1b) when $X \cup Y = V, X \cap Y \neq \emptyset$, and $d_{g-f}(X, Y) > 0$ hold.*

There exists an integer-valued feasible submodular flow if and only if (3.2) holds for every $A \subseteq V$ and for every partition $\mathcal{A}$ of A.

The other special case requires both partitions and co-partitions, but not tree-compositions.

THEOREM 3.4 *Suppose that b is crossing submodular (with $b(V) = 0$) satisfying*

$$\varrho_g(B) - \delta_f(B) \geq b(B) \text{ for every } B \subset V. \tag{3.3}$$

There exists an integer-valued feasible submodular flow if and only if $b(\mathcal{R}) \geq 0$ for every partition and co-partition $\mathcal{R}$ of V.

ORIENTATIONS

Connectivity orientation problems are strongly related to submodular flows. Let $G = (V, E)$ be an undirected graph and $h : 2^V \to \mathbf{Z} \cup \{-\infty\}$ a crossing G-supermodular set-function with $h(V) = h(\emptyset) = 0$, (that is, $h(X) + h(Y) \leq h(X \cap Y) + h(X \cup Y) + d_G(X, Y)$ where $d_G(X, Y)$ denotes the number of edges between $X - Y$ and $Y - X$). The connectivity orientation problem consists of finding an orientation of G so that the in-degree function $\varrho_{\vec{G}}$ of the resulting digraph $\vec{G} = (V, \vec{E})$ satisfies:

$$\varrho_{\vec{G}}(X) \geq h(X) \text{ for every } X \subseteq V. \tag{3.4}$$

Let us choose an arbitrary orientation $\vec{G}_r = (V, \vec{E}_r)$ of G whose in-degree function is denoted by $\varrho_r := \varrho_{\vec{G}_r}$. $\vec{G}_r$ will serve as a reference orientation to specify other orientations $\vec{G}$ of G. Define $b(X) := \varrho_r(X) - h(X)$. Any other orientation of G will be defined by a vector $x : \vec{E} \to \{0, 1\}$ so that $x(a) = 0$ means that we leave a alone while $x(a) = 1$ means that we reverse the orientation of a. The revised orientation of G defined this way satisfies (3.4) if and only if $\varrho_r(X) - \varrho_x(X) + \delta_x(X) \geq h(X)$ for every $X \subseteq V$. Equivalently, $\varrho_x(X) - \delta_x(X) \leq b(X)$. Clearly, the submodularity of b and the G-supermodularity of h are equivalent and hence there is a one-to-one correspondence between the good orientations of G and the $0 - 1$-valued submodular flows. Since $h \geq 0$ if and only if (3.3) holds for $f \equiv 0, g \equiv 1$, Theorem 3.4 implies:

THEOREM 3.5 [F80] *Suppose that h is non-negative and crossing G-supermodular. There exists an orientation of G satisfying (3.4) if and only if both $e_G(\mathcal{P}) \geq \sum_i h(P_i)$ and $e_G(\mathcal{P}) \geq \sum_i h(V - P_i)$ hold for every partition $\mathcal{P} = \{P_1, \ldots, P_p\}$ of V. If, in addition, h is symmetric, then it suffices to require $d_G(X) \geq 2h(X)$ for every $X \subseteq V$.*

When $h(X) \equiv k$ for $\emptyset \subset X \subset V$, we obtain Nash-Williams' weak orientation theorem. The following generalization, answering Problem O1, is also a consequence of Theorem 3.5: A graph G has an orientation which is k-edge-connected

from s and l-edge-connected to s (where $k \geq l$) if and only if $e_G(\mathcal{P}) \geq k|\mathcal{P}| + l - k$ holds for every partition $\mathcal{P}$ of V.

Using the same bridge between orientations and submodular flows, one can derive from Theorem 3.3 the following.

THEOREM 3.6 *Suppose that h is crossing G-supermodular and that h satisfies $h(A) + h(B) \leq h(A \cap B) + d_G(A, B)$ whenever $A \cup B = V, A \cap B \neq \emptyset$ and $d_G(A, B) > 0$. Then G has an orientation satisfying (3.4) if and only if $e_G(\mathcal{P}) \geq \sum_i h(P_i)$ holds for every sub-partition $\mathcal{P}$ of V.*

This result can be used to derive a (generalization) of a recent orientation theorem of Nash-Williams [N95] on the existence of a strongly connected orientation of a mixed graph that satisfies lower bound requirements on the in-degrees of nodes.

Problem O2 gives rise to crossing G-supermodular functions for which tree-compositions are needed. Let $\mathcal{A}$ be a tree-composition of a subset $A \subseteq V$ and let $j = uv$ be an edge of G. Let $e_{uv}(\mathcal{A})$ denote the number of sets in $\mathcal{A}$ entered by the directed edge with tail v and head u. Let $e_j(\mathcal{A}) := \max(e_{uv}(\mathcal{A}), e_{uv}(\mathcal{A}))$ and $e_G(\mathcal{A}) := \sum_{j \in E} e_j(\mathcal{A})$. The quantity $e_j(\mathcal{A})$ indicates the (maximally) possible contribution of an edge $j = uv$ to the sum $\sum(\varrho_{\vec{G}}(X) : X \in \mathcal{A})$ for any orientation $\vec{G}$ of G. Hence $e_G(\mathcal{A})$ measures the total of these contributions and therefore, for any orientation $\vec{G}$ of G satisfying (3.4), one has $\sum_{X \in \mathcal{A}} h(X) \leq \sum_{X \in \mathcal{A}} \varrho_{\vec{G}}(X) \leq e_G(\mathcal{A})$.

THEOREM 3.7 *Let h be a crossing G-supermodular function. G has an orientation $\vec{G}$ satisfying (3.4) if and only if $\sum_{X \in \mathcal{A}} h(X) \leq e_G(\mathcal{A})$ holds for every subset $A \subseteq V$ and for every tree-composition $\mathcal{A}$ of A.*

Let $M = (V, E + \vec{A})$ be a mixed graph and let $h(X) := k - \varrho_{\vec{A}}(X)$ for $\emptyset \subset X \subset V$. By applying Theorem 3.7 to this G and h, one obtains a characterization of mixed graphs having a k-edge-connected orientation, the problem O2.

ROOTED CONNECTIVITY

Let $G = (V, E)$ be a digraph with a special root node s and non-empty terminal set $T \subseteq V - s$ so that no edge of G enters s. Let p be a non-negative, T-intersecting supermodular function. Let $g : E \to \mathbf{Z}_+ \cup \{\infty\}$ be a non-negative upper bound on the edges of G. We assume that $\varrho_g(Z) \geq p(Z)$ for every subset $Z \subseteq V$ where $\varrho_g(Z) := \sum(g(e) : e \in E, e \text{ enters } Z)$.

THEOREM 3.8 *(a) The linear system $\{\varrho_x(Z) \geq p(Z)$ for every $Z \subset V, 0 \leq x \leq g\}$ is totally dual integral. (b) The polyhedron defined by this system is a submodular flow polyhedron.*

For the special case $T = V - s$, part *(a)* was proved in [F79] while part *(b)* in [Schrijver, 1984]. The edge-version of problem S2 could be solved via this special case. It is not difficult to observe that the proofs extend easily to the more general

case. The main advantage of this extension is that, beyond handling the edge-version of problem S2, the node-version can also be settled by using the standard node-splitting technique.

To conclude the section, we remark that there is a polynomial time algorithm to solve minimum cost submodular flow problems hence all the connectivity problems above admit polynomial time solution algorithms.

4. Covering ST-crossing supermodular functions by digraphs

We say that a digraph $G = (V, E)$ *covers* a set-function p if there are at least $p(X)$ edges entering every subset $X \subseteq V$. How many edges are needed to cover p?

THEOREM 4.1 [F94] *Let p be a crossing supermodular function and γ a positive integer. There exists a digraph $G = (V, E)$ of at most γ edges covering p if and only if $\sum(p(X) : X \in \mathcal{P}) \leq \gamma$ and $\sum(p(V - X) : X \in \mathcal{P}) \leq \gamma$ hold for every subpartition $\mathcal{P}$ of V.*

This result can be extended, as follows. Let S and T be two subsets of a ground-set V. Two subsets X, Y are called *ST-independent* if $X \cap Y \cap T = \emptyset$ or $S \subseteq X \cup Y$.

THEOREM 4.2 [FJ95] *Let $p : 2^V \to \mathbf{Z}_+$ be an ST-crossing supermodular function and γ a positive integer. There exists a digraph $G = (V, E)$ that covers p, has at most γ edges, and each edge has its tail in S and its head in T if and only if $\sum(p(X) : X \in \mathcal{P}) \leq \gamma$ holds for every family $\mathcal{P}$ of pairwise ST-independent subsets of V.*

When $S = T = V$, an ST-independent family consists of pairwise disjoint sets or of pairwise co-disjoint sets. (Two sets are *co-disjoint* if their complement is disjoint). Hence Theorem 4.1 is indeed a special case of Theorem 4.2. Theorem 4.1 may be applied to solve an extension of the edge-connectivity version of problem A2. Let $D = (V, E)$ be a directed graph and T a subset of nodes. We say that D is *k-edge-connected in T* if $\lambda(u, v) \geq k$ for every pair of nodes $u, v \in T$.

THEOREM 4.3 *It is possible to make digraph D k-edge-connected in T by adding at most γ new edges connecting elements of T if and only if $\sum_i(k - \varrho_D(X_i)) \leq \gamma$ and $\sum_i(k - \delta_D(X_i)) \leq \gamma$ holds for every family $\mathcal{F} = \{X_1, \ldots, X_t\}$ of subsets V for which $\emptyset \subset X_i \cap T \subset T$ and $\mathcal{F}|T$ is a sub-partition of T.*

We say that D is *k-edge-connected from S to T* if there are k edge-disjoint paths from every node of S to every node of T. (When $S = T$ we are back at k-edge-connectivity in T.) Theorem 4.2 gives rise to the following solution to problem A3:

THEOREM 4.4 *A digraph $D = (V, E)$ can be made k-edge-connected from S to T by adding at most γ new edges with tails in S and heads in T if and only*

if $\sum_j(k - \varrho(X_j)) \leq \gamma$ *holds for every choice of an* (S,T)*-independent family of subsets* $X_j \subseteq V$ *where* $T \cap X_j \neq \emptyset, S - X_j \neq \emptyset$ *for each* X_j*.*

There is a constructive proof of Theorem 4.1 which gives rise to a strongly polynomial algorithm to find an optimal augmentation in Theorem 4.3. The proof of Theorem 4.2 is not constructive and no combinatorial polynomial algorithm is known to construct the optimal augmentation of Theorem 4.4. It is a major open problem of the field to find one.

Another consequence of Theorem 4.2 concerns the node-connectivity version of problem A2. Given a digraph $D = (V,E)$, we say that a pair of disjoint, nonempty subsets X, Y of V is a *one-way* pair if there is no edges from X to Y. The deficiency $p_{def}(X,Y)$ of a one-way pair is defined by $k - |V - (X \cup Y)|$. Two one-way pairs (X,Y) and (A,B) are called *independent* if $X \cap A = \emptyset$ or $Y \cap B = \emptyset$.

THEOREM 4.5 *A digraph* $D = (V,E)$ *can be made* k*-node-connected by adding at most* γ *new edges if and only if* $\sum(p_{def}(X,Y) : (X,Y) \in \mathcal{F}) \leq \gamma$ *holds for every family* $\mathcal{F}$ *of pairwise independent one-way pairs.*

Are these results related to the ones mentioned in the previous section? One fundamental difference is that, while submodular flows are appropriate to handle min-cost problems, here the minimum-cost versions include NP-complete problems. For example, finding a minimum cost strongly connected augmentation of a digraph is NP-complete. However, for node-induced cost functions the node-connectivity augmentation problem turns out to be tractable. A node-induced cost of a directed edge uv is defined by $c(uv) := c_t(u) + c_h(v)$ where c_t and c_h are two cost-functions on the node set V. The better behaviour of node-induced cost-functions is based on the fact that the in-degree vectors of k-connected augmentations with γ edges span a base-polyhedron.

We conclude this section by briefly remarking that Theorem 4.2 has a surprising consequence in combinatorial geometry; a theorem of E. Győri [Gy84] asserting that every vertically convex rectilinear polygon R (bounded by horizontal and vertical segments) in the plane can be covered by γ rectangles belonging to R if and only if R does not contain more than γ pairwise independent points (where two points are called independent if they cannot be covered by one rectangle (with horizontal and vertical sides).

5. COVERING CROSSING AND SKEW SUPERMODULAR FUNCTIONS BY GRAPHS

Let p be a non-negative, symmetric, crossing supermodular function. An undirected graph is said *to cover* p if every cut $[X, V-X]$ contains at least $p(X)$ edges. What is the minimum number of edges covering p?

For a partition $\mathcal{P}$ of V, the sum $\sum(p(X) : X \in \mathcal{P})/2$ is clearly a lower bound. However, even the best such bound can be strictly smaller than the true minimum: when $p(X) \equiv 1$ for $\emptyset \subset X \subset V$ and $p(\emptyset) = p(V) = 0$, the minimum is $|V| - 1$ while the best partition bound is $|V|/2$. Hence we need a new parameter, called the dimension of p. A partition $\mathcal{F} := \{V_1, \ldots, V_h\}$ of V with $h \geq 4$ is said to be

p-full if $p(\cup\mathcal{F}') \geq 1$ for every sub-partition $\mathcal{F}', \emptyset \subset \mathcal{F}' \subset \mathcal{F}$, and $\mathcal{F}$ has a member V_l with $p(V_l) = 1$. We call the maximum size of a p-full partition the *dimension* of p and denote it by $\dim(p)$. It can easily be seen that any graph covering p must have at least $\dim(p) - 1$ edges. The content of the next result is that the minimum in question is equal to the larger of the two lower bounds.

THEOREM 5.1 [BF96] *Let $p : 2^V \to \mathbf{Z}_+$ be a symmetric, crossing supermodular function and γ a positive integer. There exists an undirected graph $G = (V, E)$ with at most γ edges covering p if and only if $\sum(p(X) : X \in \mathcal{P}) \leq 2\gamma$ holds for every partition $\mathcal{P}$ of V and $\dim(p) - 1 \leq \gamma$.*

It is an important open problem to extend this theorem to skew-supermodular functions. For even-valued functions p (that is, when $p(X)$ is even for every subset X) this was done by Z. Szigeti. The advantage of even supermodular functions is that their dimension does not play any role. To capture the difference, observe that if p_1 is identically 1 on non-empty proper sutsets of V, then a tree will be the smallest graph covering p_1, that is, the minimum number of edges is $n - 1$. If $p_2 := 2p_1$, then we do not need twice as many edges to cover p_2. Just one more edge will do as a circuit of n edges cover every cut at least twice.

THEOREM 5.2 [Sz95] *Let $p : 2^V \to \mathbf{Z}_+$ be a symmetric, even-valued, skew-supermodular function and γ a positive integer. There exists a graph $G = (V, E)$ with at most γ edges covering p if and only if $\sum(p(X) : X \in \mathcal{P}) \leq 2\gamma$ holds for every partition $\mathcal{P}$ of V.*

As a consequence of Theorem 5.1 we exhibit a result concerning hypergraph connectivity augmentation. Given a hypergraph $H' = (V, A')$, a subset $\emptyset \subset C \subset V$ is called a *component* of H' if $d_{H'}(C) = 0$ and $d_{H'}(X) > 0$ for every $\emptyset \subset X \subset C$. ($d_{H'}(X)$ denotes the number of hyperedges of H' intersecting both X and $V - X$.) For a subset $T \subset V$, we let $c_T(H')$ denote the number of components of H' having a non-empty intersection with T. H' is said to be *k-edge-connected in T* if $d_{H'}(X) \geq k$ for every subset $\emptyset \subset X \subset V$ separating T. When $T = V$ we say that H' is *k-edge-connected.*

THEOREM 5.3 *Let $H = (V, A)$ a hypergraph, T a specified subset of V, and γ a positive integer. $H = (V, A)$ can be made k-edge-connected in T by adding at most γ new graph-edges if and only if $\sum(k - d_H(X) : X \in \mathcal{P}) \leq 2\gamma$ for every sub-partition $\mathcal{P}$ of V separating T and $c_T(H') - 1 \leq \gamma$ for every hypergraph $H' = (V, A')$ arising from H by leaving out $k - 1$ hyperedges. If these conditions hold, the new edges can be chosen so as to connect elements of T.*

This result is a solution to problem A4. It extends an earlier theorem of J. Bang-Jensen and B. Jackson [BJ95] where $T = V$, which, in turn, generalizes an even earlier result of T. Watanabe and A. Nakamura [WN87] when the starting hypergraph H is itself a graph. The latter result was generalized in another direction in [F92] where, instead of global k-edge-connectivity, specified demands $r(u, v)$

were required for the augmented local edge-connectivities between every pair of nodes u and v. Since such a problem gives rise to skew-supermodular functions, Theorem 5.1 cannot be applied. However, if half-capacity edges are also allowed in the augmentation, then Theorem 5.2 can be applied. That is, one can find a graph G of minimum number of edges so that adding the edges of G with half-capacity to the starting hypergraph, the local edge-connectivities of the increased hypergraph attain a prescribed value $r(u, v)$ for every pair $\{u, v\}$ of nodes.

References

[BJ95] J. Bang-Jensen and B. Jackson, *Augmenting hypergraphs by edges of size two,* Mathematical Programming, Ser B., to appear.

[BF96] A. Benczúr and A. Frank, *Covering symmetric supermodular functions by graphs,* Mathematical Programming, Ser B., to appear.

[E70] J. Edmonds, *Submodular functions, matroids, and certain polyhedra*, in: Combinatorial Structures and their applications (R. Guy, H. Hanani, N. Sauer, and J. Schönheim, eds.) Gordon and Breach, New York, 69-87.

[E79] J. Edmonds, *Matroid intersection*, in: "Discrete Optimization", Annals of Discrete Mathematics, Vol. 4 (1979) North-Holland.

[EG77] J. Edmonds and R. Giles, *A min-max relation for submodular functions on graphs,* Annals of Discrete Mathematics 1, (1977), 185-204.

[ET76] K.P. Eswaran and R.E. Tarjan, *Augmentation problems,* SIAM J. Computing, Vol. 5, No. 4, December 1976, 653-665.

[F79] A. Frank, *Kernel systems of directed graphs*, Acta Scientiarum Mathematicarum (Szeged) 41, 1-2 (1979) 63-76.

[F80] A. Frank, *On the orientation of graphs*, J. Combinatorial Theory, Ser B. , Vol. 28, No. 3 (1980) 251-261.

[F81] A. Frank, *How to make a digraph strongly connected,* Combinatorica 1, No. 2 (1981) 145-153.

[F82] A. Frank, *An algorithm for submodular functions on graphs,* Annals of Discrete Mathematics 16 (1982) 97-120.

[F92] A. Frank, *Augmenting graphs to meet edge-connectivity requirements,* SIAM J. on Discrete Mathematics, (1992 February), Vol 5, No 1. pp. 22-53.

[F94] A. Frank, *Connectivity augmentation problems in network design,* in: Mathematical Programming: State of the Art 1994, eds., J.R. Birge and K.G. Murty), The University of Michigan, pp. 34-63.

[F96] A. Frank, *Orientations of Graphs and Submodular Flows,* Congressus Numerantium, 113 (1996) (A.J.W. Hilton, ed.) pp. 111-142.

[FJ95] A. Frank and T. Jordán, *Minimal edge-coverings of pairs of sets*, J. Combinatorial Theory, Ser. B. Vol. 65, No. 1 (1995, September) pp. 73-110.

[Fu84] S. Fujishige, *Structures of polyhedra determined by submodular functions on crossing families,* Math Programming, 29 (1984) 125-141.

[Gy84] E. Győri, *A minimax theorem on intervals,* J. Combinatorial Theory, Ser. B 37 (1984) 1-9.

[L83] L. Lovász, *Submodular functions and convexity,* in: Mathematical programming- The state of the art, (eds. A. Bachem, M. Grötschel and B. Korte) Springer 1983, 235-257.

[LY78] C.L. Lucchesi and D.H. Younger, *A minimax relation for directed graphs,* J. London Math. Soc. (2) 17 (1978) 369-374.

[M96] K. Murota, *Convexity and Steinitz's exchange property,* Advances in Mathematics, Vol.124, No.2, (1996), 272-310.

[N60] C.St.J.A. Nash-Williams, *On orientations, connectivity and odd vertex pairings in finite graphs,* Canad. J. Math. 12 (1960) 555-567.

[N95] C.St.J.A. Nash-Williams, *Strongly connected mixed graphs and connected detachments of graphs,* Journal of Combinatorial Mathematics and Combinatorial Computing 19 (1995) pp. 33-47.

[S84] A. Schrijver, *Total dual integrality from directed graphs, crossing families and sub- and supermodular functions,* in: Progress in Combinatorial Optimization,(ed. W. R. Pulleyblank) Academic Press (1984) 315-361.

[Sz96] Z. Szigeti, *Hypergraph connectivity augmentation,* Mathematical Programming, Ser B., to appear.

[WN87] T. Watanabe and A. Nakamura, *Edge-connectivity augmentation problems,* Computer and System Sciences, Vol 35, No.1, (1987) 96-144.

Dept. of Operations Research
Eötvös University
Múzeum krt. 6-8, Budapest
Hungary, H-1088
frank@cs.elte.hu

Ericsson Traffic Laboratory
Laborc u.1, Budapest
Hungary, H-1037

Ordonner le Groupe Symétrique: Pourquoi Utiliser l'Algèbre de Iwahori-Hecke ?

Alain Lascoux(*)

Abstract. The Bruhat order on the symmetric group is defined by means of subwords of reduced decompositions of permutations as products of simple transpositions. Ehresmann gave a different description by considering any permutation as a chain of sets and comparing componentwise the chains. A third method reduces the Bruhat order to the inclusion order on sets, by associating to any permutation a set of bigrassmannian permutations. This amounts to embed the symmetric group into a lattice which is distributive. The last manner to understand the Bruhat order is to use a distinguished linear basis of the Iwahori-Hecke algebra of the symmetric group, and this involves computing polynomials due to Kazhdan & Lusztig; we explicit these polynomials in the case of vexillary permutations.

1991 Mathematics Subject Classification: 05E10, 20C30
Keywords and Phrases: Symmetric group, Bruhat Order, Kazhdan-Lusztig Polynomials

1. Ordre par les sous-mots En tant que groupe de Coxeter, le groupe symétrique $\mathfrak{S}(n)$ est engendré par les transpositions simples σ_i, $i = 1 \dots n-1$, qui vérifient les relations de tresse

$$\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1} \quad \text{et} \quad \sigma_i\sigma_j = \sigma_j\sigma_i \ , \ |i-j| > 1 \tag{1.1}$$

ainsi que $\sigma_i^2 = 1$.

Une *décomposition réduite* d'une permutation μ est un mot $w^{ij\dots h} = \sigma_i\sigma_j\cdots\sigma_h$, dont le produit, de longueur minimale, est égal à μ (cette longueur est dite *longueur* $\ell(\mu)$ de μ). Par définition (cf. [Hu]), l'*ordre de Bruhat* est l'ordre induit par les sous-mots :

$$\nu \le \mu \ \Leftrightarrow \exists w^{ij\dots h} = \mu,\ w^{ij\dots h} \text{ réduit } , \exists \epsilon, \dots, \epsilon'' \in \{0,1\}, \nu = \sigma_i^{\epsilon}\,\sigma_j^{\epsilon'}\cdots\sigma_h^{\epsilon''} \tag{1.2}$$

Soit σ une transposition simple telle que $\ell(\mu\sigma) > \ell(\mu)$. Alors on a la "propriété d'échange" :

$$[1,\,\mu] = A \cup B \text{ et } [1,\,\mu\sigma] = A \cup B \cup B\sigma \tag{1.3}$$

(*) C.N.R.S.

où $[1,\,\mu] := \{\nu \in \mathfrak{S}(n),\, \nu \leq \mu\}$, $A := \{\nu : \nu \leq \mu, \nu\sigma \leq \mu\}$ et $B := [1,\,\mu] \setminus A$.

On définit, sur l'algèbre du groupe symétrique $\mathfrak{S}(n)$, des opérateurs de réordonnement $\pi_1, \ldots, \pi_{n-1}$, notés à droite

$$\mathfrak{S}(n) \ni \mu \xrightarrow{\pi_i} \begin{cases} \mu + \mu\sigma_i & \text{si } \mu_i < \mu_{i+1} \\ 0 & \text{autrement} \end{cases}$$

Il est aisé de voir que pour toute décomposition réduite de μ, l'image de 1 par un produit de π_i est la somme des éléments de l'intervalle $[1,\,\mu]$:

$$\mu = \sigma_i\sigma_j \cdots \sigma_h \text{ réduit } \Rightarrow 1\pi_i\pi_j \cdots \pi_h = \sum\nolimits_{\nu \leq \mu} \nu \tag{1.4}$$

Deux décompositions réduites de la même permutation vont donner en général des ensembles de sous-mots différents et donc ces ensembles ne sont pas des invariants de la permutation.

Une autre manière que (1.4) de corriger cette non-canonicité est de pondérer les sous-mots. Etant données n variables $x_1, \ldots, x_n$, on définit, à la suite de Yang [Ya],[Ch], une base linéaire Y_μ, $\mu \in \mathfrak{S}(n)$, de l'algèbre du groupe symétrique à coefficients rationnels en les x_i, par

$$\ell(\mu\sigma_i) > \ell(\mu) \Rightarrow Y_{\mu\sigma_i} = Y_\mu \left(\sigma_i + \frac{1}{x_{\mu_{i+1}} - x_{\mu_i}}\right) . \tag{1.5}$$

Toute décomposition réduite $w^{i\cdots h}$ de μ fournit une factorisation de Y_μ, dont le développement est une somme impliquant tous les sous-mots de $w^{i\cdots h}$. On vérifie de plus que le coefficient de ν dans Y_μ est non nul ssi $\nu \leq \mu$.

En fait, les coefficients sont des spécialisations de polynômes en deux ensembles de variables ([LLT2], [F-K]). On peut les obtenir en définissant des opérateurs sur l'anneau des polynômes vérifiant les relations de tresse [L-S4], [L-S5]. Ces opérateurs fournissent à leur tour des bases distinguées de l'anneau des polynômes en tant que module libre sur l'anneau des polynômes symétriques [BGG] et l'ordre de Bruhat joue un rôle essentiel [L-S3] (les programmes sont disponibles comme librairie Maple [Ve]). On trouvera dans [L-P] l'étude analogue de l'anneau des polynômes comme module libre sur l'anneau des polynômes symétriques en les carrés des variables, qui correspond aux groupes hyperoctahedraux.

2. ORDRE PAR PROJECTION Il existe un ordre naturel sur les sous-ensembles de $\{1, \ldots, n\}$:

$$u, v \subseteq \{1, \ldots, n\},\ u \leq v \Leftrightarrow \exists \text{ une injection croissante de } u \text{ dans } v$$

Cet ordre permet de définir les *tableaux de Young* comme étant les chaînes croissantes d'ensembles d'entiers.

Ehresmann [Eh] induit à partir de cet ordre sur les ensembles, un ordre sur les cellules de Schubert de la variété de drapeaux pour le groupe linéaire (lesquelles sont en bijection avec les permutations) :

$$\nu, \mu \in \mathfrak{S}(n),\, \nu \leq \mu \ \Leftrightarrow\ \forall i : 1 \leq i \leq n,\, \{\nu_1, \ldots, \nu_i\} \leq \{\mu_1, \ldots, \mu_i\} \tag{2.1}$$

On peut disposer les ensembles $\{\mu_1, \ldots, \mu_i\}$ dans un tableau, dit *clef* de la permutation, dont ils sont les colonnes (décroissantes). Alors deux permutations sont comparables ssi leurs clefs le sont, composante à composante. De fait, on vérifie aisément par récurrence sur la longueur que l'ordre d'Ehresmann coincïde avec l'ordre de Bruhat.

La restriction $\mu \to \{\mu_1, \ldots, \mu_i\}$ peut s'interpréter comme la projection de $\mathfrak{S}(n)$ sur $\mathfrak{S}(n)/\mathfrak{S}(i) \times \mathfrak{S}(n-i)$, où $\mathfrak{S}(i) \times \mathfrak{S}(n-i)$ est le sous-groupe de Young engendré par $\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_{n-1}$. On peut identifier les éléments de $\mathfrak{S}(n)/\mathfrak{S}(i) \times \mathfrak{S}(n-i)$ aux permutations γ (dites *grassmanniennes*) : $\gamma_1 < \cdots < \gamma_i$; $\gamma_{i+1} < \cdots < \gamma_n$, ayant une *descente* en i. La restriction de l'ordre de Bruhat à ces dernières est

$$\gamma \leq \gamma' \;\Leftrightarrow\; \gamma_1 \leq \gamma'_1, \ldots, \gamma_i \leq \gamma'_i$$

Deodhar [De] a étendu à tous les groupes de Coxeter W la définition de l'ordre de Bruhat par relèvement de l'ordre sur les W/P, P parabolique. Proctor [Pr] a généralisé aux types B, C, D la construction des clefs.

3. ORDRE PAR SOUS-ENSEMBLES Au lieu de considérer toutes les projections $\mathfrak{S}(n) \mapsto \mathfrak{S}(i) \times \mathfrak{S}(n-i)$, $i = 1, \ldots, n-1$, on peut associer à toute permutation μ l'ensemble $\mathcal{G}(\mu)$ des permutations grassmanniennes γ telles que $\gamma \leq \mu$. Le critère (2.1) se formule alors

$$\nu \leq \mu \;\Leftrightarrow\; \mathcal{G}(\nu) \subseteq \mathcal{G}(\mu) \tag{3.1}$$

Cette définition n'est pas invariante par l'involution $\mu \mapsto \mu^{-1}$, contrairement à l'ordre de Bruhat. Pour corriger cette disymétrie, on définit les permutations *bigrassmanniennes* comme étant les permutations qui sont grassmanniennes, ainsi que leurs inverses. En d'autres termes

$$\beta \text{ bigrassmannienne } \;\Leftrightarrow\; \exists! i, \exists! j \,:\, \ell(\sigma_i \mu) < \ell(\mu),\, \ell(\mu \sigma_j) < \ell(\mu)$$

(i est dit *recul* de β, et j *descente*).

Soit $\mathcal{B}(\mu)$ l'ensemble des permutations bigrassmanniennes β telles que $\beta \leq \mu$. Le critère (3.1) est équivalent à

$$\nu \leq \mu \;\Leftrightarrow\; \mathcal{B}(\nu) \subseteq \mathcal{B}(\mu) \tag{3.2}$$

En fait, on peut montrer que l'ensemble des bigrassmanniennes est optimal pour obtenir l'ordre de Bruhat par inclusion. Plus précisément, soit $\mathcal{C} \subseteq \mathfrak{S}(n)$. Pour que le morphisme $\mathfrak{S}(n) \to 2^{\mathcal{C}} : \mu \to \mathcal{C} \cap [1, \mu]$ soit un morphisme d'ordre injectif, il faut et il suffit que $\mathcal{C}$ contienne l'ensemble des bigrassmanniennes (cf. [L-S6]).

Pour les groupes de Coxeter finis, on trouvera dans [G-K] la détermination du sous-ensemble optimal codant l'ordre. Les élements de la "base de l'ordre" sont caractérisés par la propriété :

β appartient à la base ssi il existe un élément μ du groupe tel que β soit minimum dans le complémentaire de l'intervalle $[1, \mu]$.

La même construction peut être étendue aux groupes de Coxeter affines (pour une description plus classique, voir [B-B] dans le cas du type A et [Er] plus généralement).

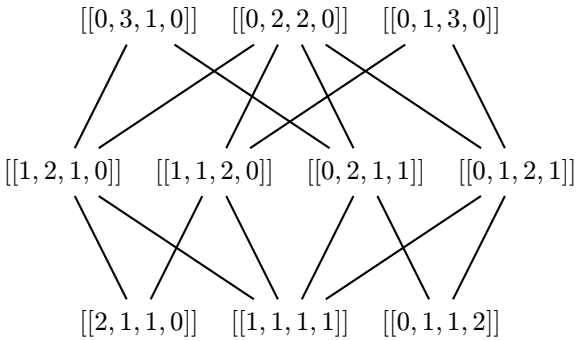
Toute permutation bigrassmannienne dans $\mathfrak{S}(n)$ est une permutation du type $[1,\dots,a,\ a+1+c,\dots,\ a+b+c,a+1,\dots,a+c,a+b+c+1,\dots,a+b+c+d]$ et donc définie par un vecteur $[[a,b,c,d]] \in N^4$, $a,d \geq 0$, $b,c \geq 1$, $a+b+c+d=n$.

La restriction de l'ordre de Bruhat aux bigrassmanniennes est

$$[[a,b,c,d]] \leq [[a',b',c',d']] \Leftrightarrow a \geq a',\ d \geq d';\ b \leq b',\ c \leq c'\ . \tag{3.3}$$

Une permutation μ est supérieure à une bigrassmannienne $\beta = [[a,b,c,d]]$ ssi l'ensemble $\{\mu_1,\dots,\ \mu_{a+b}\}$ contient au moins b valeurs $> a+c$.

Le treillis engendré par les bigrassmanniennes (en tant que sup-irréductibles; de manière équivalente, on prend l'ensemble des unions quelconques de $\mathcal{B}(\mu)$) est dit *treillis enveloppant* du groupe symétrique, ou *complétion de Mac Neille* [Bi]. Les éléments de ce treillis sont par définition en correspondance bijective avec les antichaînes de bigrassmanniennes. Par exemple, pour $\mathfrak{S}(4)$, il y a 10 bigrassmanniennes et 42 antichaînes, donc 42-24=18 éléments du treillis qui ne sont pas des permutations, l'ordre sur les bigrasmanniennes étant :



Les éléments du treillis enveloppant peuvent aussi être identifiés aux supremums (composante à composante) d'une famille quelconque de clefs. Ces supremums sont des tableaux ayant des propriétés supplémentaires de croissances diagonales, que l'on appelle *triangles monotones*, et qui sont en bijection avec les matrices à signe alternant (*alternating sign matrices*), cf. [M-R-R], [An], [Zei].

Ainsi le supremum des bigrassmanniennes $[[1,2,1,0]]$, $[[1,1,2,0]]$, $[[0,2,1,1]]$, $[[0,1,2,1]]$ se représente par

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & 1 \\ 1 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \leftrightarrow \begin{array}{ccccc} & 3 & 4 & 4 & 4 \\ & & 2 & 3 & 3 \\ & & & 1 & 2 \\ & & & & 1 \end{array} = sup \left(\begin{array}{ccccc} 3 & 3 & 3 & 4 & \\ & 2 & 2 & 3 & \\ & & 1 & 2 & \\ & & & 1 & \end{array} \text{ et } \begin{array}{ccccc} & 1 & 4 & 4 & 4 \\ & & 1 & 3 & 3 \\ & & & 1 & 2 \\ & & & & 1 \end{array} \right)$$

et c'est de plus le supremum des deux permutations $[3,2,1,4]$ et $[1,4,3,2]$ dont nous avons donné les clefs à droite.

La lecture par colonnes, de gauche à droite, des tableaux ou des matrices donne la même suite d'ensembles : 1 ou -1 en ligne i signifie que la lettre i apparaît

ou disparaît. Ainsi la matrice ci-dessus se lit : " 3 apparaît; 2 et 4 apparaîssent, 3 disparaît; 1 et 3 apparaîssent, 2 disparaît; 2 revient enfin".

On montre en outre que le treillis enveloppant du groupe symétrique est distributif [L-S6]. En d'autres termes, pour toute bigrassmannienne β, il existe une permutation η telle que $\mathfrak{S}(n)$ est l'union disjointe des deux intervalles $[1, \eta]$ et $[\beta, \omega]$, où ω est l'élément maximum de $\mathfrak{S}(n)$. Par exemple, une permutation de $\mathfrak{S}(4)$ est soit au dessus de $\beta = [1, 4, 2, 3] = [[1, 1, 2, 0]]$, soit en dessous de $[3, 2, 4, 1]$.

Codant chaque permutation par $\mathcal{B}(\mu)$, ou par le vecteur booléen $[v_\beta(\mu)]_{\beta\in\mathcal{B}}$: $v_\beta(\mu) = 1$ ou 0 selon que $\beta \leq \mu$ ou non, on dispose ainsi d'un outil purement algébrique de calcul sur les intervalles pour l'ordre de Bruhat.

4. ALGÈBRE DE HECKE Au lieu d'énumérer, on peut se proposer au contraire de chercher à caractériser la fonction génératrice $\sum_{\nu\leq\mu}(-q)^{\ell(\mu)-\ell(\nu)}\,\nu$ des éléments de l'intervalle $[1, \mu]$.

L'algèbre appropriée est cette fois-ci l'algèbre de Iwahori-Hecke $\mathcal{H}_n$ du groupe symétrique $\mathfrak{S}(n)$, définie comme l'algèbre, à coefficients dans $\mathbb{Z}[q, 1/q]$, engendrée par les $T_1, \ldots, T_{n-1}$ satisfaisant les relations de tresse

$$T_iT_{i+1}T_i = T_{i+1}T_iT_{i+1} \text{ et } T_iT_j = T_jT_i \ , \ |i-j| > 1 \tag{4.2}$$

ainsi que la relation de Hecke

$$(T_i - q)(T_i + 1/q) = 0 \ , \ i = 1, \ldots, n-1 \ . \tag{4.3}$$

Une base linéaire de $\mathcal{H}_n$ consiste en les $\{T_\mu, \mu \in \mathfrak{S}(n)\}$, définies par produits réduits de T_i. Sur $\mathcal{H}_n$, on a une involution $T_\mu \mapsto (T_{\mu^{-1}})^{-1}$, $q \mapsto 1/q$. Soit $\mathcal{L}$ le sous-module $\oplus_{\mu\in\mathfrak{S}(n)}\mathbb{Z}[q]T_\mu$ et θ la projection $\mathcal{L} \mapsto \mathcal{L}/q\mathcal{L}$. Kazhdan & Lusztig (cf. [KL1], [Lu]) ont montré que pour chaque $\mu \in \mathfrak{S}(n)$, il existe un élément unique $c_\mu \in \mathcal{L}$ qui soit invariant par l'involution et tel que $\theta(c_\mu) = \theta(T_\mu)$. Les élements c_μ, $\mu \in \mathfrak{S}(n)$ constituent donc une base linéaire de $\mathcal{H}_n$, et l'on a de plus

$$c_\mu = \sum_{\nu\leq\mu}(-q)^{\ell(\mu)-\ell(\nu)}\,P_{\nu,\mu}(q^{-2})\,T_\nu \ , \tag{4.4}$$

les $P_{\nu,\mu}$ étant des polynômes à coefficients entiers positifs, dits *Polynômes de Kazhdan & Lusztig* , qui interviennent dans de nombreuses théories [KL2], [Br].

Il est clair que $\{\,1, T_1 - q\,\}$ est la base de Kazhdan & Lusztig de $\mathcal{H}_2$. Plus généralement, posons $c_i := T_i - q$.

Tout produit $c_\mu\, c_i$ est invariant par l'involution, a pour terme dominant $T_{\mu\sigma_i}$ si $\ell(\mu\,\sigma_i) > \ell(\mu)$, et peut donc être considéré comme une approximation de $c_{\mu\sigma_i}$. On obtient $c_{\mu\sigma_i}$ en soustrayant récursivement les multiples appropriés des c_ν pour les ν tels que le coefficient de T_ν comporte un terme constant.

Par exemple, $c_{2,3,1}\cdot c_1 = (q^2+1)T_{2,1,3} + T_{3,2,1} - qT_{3,1,2} - qT_{2,3,1} + q^2T_{1,3,2} - (q^3+q)T_{1,2,3}$ et $c_{3,2,1} = c_{2,3,1}c_1 - c_{2,1,3} = T_{3,2,1} - qT_{3,1,2} - qT_{2,3,1} + q^2T_{1,3,2} + q^2T_{3,1,2} - q^3T_{1,2,3}$.

Cette récurrence élémentaire peut difficilement être mise en oeuvre dès $n > 8$ et il faut donc trouver des méthodes plus économiques qui n'imposent pas d'énumérer les éléments d'un intervalle.

En fait, pour tout i tel que $\ell(\mu\,\sigma_i) < \ell(\mu)$, alors $P_{\nu,\mu} = P_{\nu\sigma_i,\mu}$, $\forall \nu \in \mathfrak{S}(n)$, et donc, comme l'indiquent [KL1], les polynômes $P_{\nu,\mu}$ sont constants dans toute double classe $\mathfrak{S}(I)\backslash\mathfrak{S}(n)/\mathfrak{S}(J)$, où $\mathfrak{S}(I), \mathfrak{S}(J)$ sont deux sous-groupes de Young déterminés par μ ($\mathfrak{S}(I)$ est le sous groupe engendré par les σ_i, i recul, et $\mathfrak{S}(J)$ est engendré par les descentes).

L'invariance par rapport à deux sous-groupes de Young permet d'utiliser des propriétés de factorisation.

Soit $\omega = [n, \dots, 1]$, et pour $1 \le i < j \le n$, $\omega[i,j] = [1, \dots, i-1, j, \dots, i, j+1, \dots, n]$; posons pour tout entier positif $[r] := (q^r - q^{-r})/(q - q^{-1})$. Alors [DKLLST]

$$c_\omega = \sum_{\nu\in\mathfrak{S}(n)} (-q)^{\ell(\omega)-\ell(\nu)}\, T_\nu = c_{\omega[1,n-1]}\,(T_{n-1} - \tfrac{q^{n-1}}{[n-1]}) \cdots (T_2 - \tfrac{q^2}{[2]})(T_1 - \tfrac{q^1}{[1]})$$

$$= (T_1 - \tfrac{q^1}{[1]}) \cdots (T_{n-1} - \tfrac{q^{n-1}}{[n-1]})\, c_{\omega[2,n]} \tag{4.5}$$

PROPOSITION *Soit $\mu \in \mathfrak{S}(n)$; soient k l'entier tel que $\mu_k = n$ et $\nu \in \mathfrak{S}(n-1)$ obtenue par effacement de n dans μ (ν est notée $\mu\backslash n$).*

Si $n = \mu_k > \mu_{k+1} > \cdots > \mu_n$, alors

$$c_\mu = c_{\mu\backslash n}\,(T_{n-1} - \tfrac{q^{n-k}}{[n-k]}) \cdots (T_{k+1} - \tfrac{q^2}{[2]})\,(T_k - \tfrac{q^1}{[1]}) \tag{4.6}$$

Preuve L'élément de droite a pour terme dominant T_μ, et est invariant par l'involution. Par ailleurs, $c_{\omega[k,n-1]}(T_{n-1} - \frac{q^{n-k}}{[n-k]}) \cdots (T_1 - \frac{q^1}{[1]}) = c_{\omega[k,n]} = c_{\omega[k,n-1]}(T_{(n,k)} - qT_{(n,k+1)} + \cdots + (-q)^{n-1-k}T_{(n,n-1)} + (-q)^{n-k})$, somme sur toutes les transpositions de n avec i, $i \ge k$. Le produit se développe en une somme où l'on retrouve comme coefficients les polynômes de Kazhdan-Lusztig pour ν; il est donc bien égal à c_μ □

La proposition précédente, combinée aux involutions $\mu \mapsto \mu^{-1}$ et $\mu \mapsto \omega\mu\omega$, permet de factoriser totalement certains c_μ. En particulier, une permutation μ est dite *non singulière* si μ ou μ^{-1} a la propriété qu'il existe k : $\mu_k = n > \mu_{k+1} > \cdots > \mu_n$, et $\mu\backslash n$ est non singulière ($[1] \in \mathfrak{S}(1)$ est décrétée non singulière).

Dans le premier cas, $c_\mu = c_{\mu\backslash n}(T_{n-1} - \frac{q^{n-k}}{[n-k]}) \cdots (T_k - \frac{q^1}{[1]})$. Dans le deuxième, $c_\mu = (T_k - \frac{q^1}{[1]}) \cdots (T_{n-1} - \frac{q^{n-k}}{[n-k]})\, c_{[\mu^{-1}\backslash n]^{-1}}$.

COROLLAIRE *Si μ est une permutation non singulière, alors $c_\mu = \sum_{\nu\le\mu}(-q)^{\ell(\mu)-\ell(\nu)}\, T_\nu$ et c_μ factorise en un produit de facteurs $(T_i - \frac{q^j}{[j]})$. Le polynôme de Poincaré de l'intervalle $[1, \mu]$ s'obtient, à une puissance de q près, en spécialisant $T_i \mapsto -1/q$ dans chaque facteur.*

Par exemple, $\mu = [4,1,6,5,3,2]$ est non singulière; écrivant k, k^+, k^{++} pour $T_k - \frac{q^1}{[1]}$, $T_k - \frac{q^2}{[2]}$, $T_k - \frac{q^3}{[3]}$ respectivement, on obtient la suite d'égalités

$$c_{312} = 21 \mapsto c_{4132} = 23^+(21) \mapsto c_{41532} = (23^+21)4^+3 \mapsto c_{416532} = (23^+214^+3)5^{++}4^+3$$

et le polynôme de Poincaré de l'intervalle $[1, \mu]$, en la variable q^2, est égal à

$$-q^9\,[2]\,\frac{[3]}{[2]}\,[2]\,[2]\,\frac{[3]}{[2]}\,[2]\,\frac{[4]}{[3]}\,\frac{[3]}{[2]}\,[2] = (1+q^2)^2\,(1+q^2+q^4)^2\,(1+q^2+q^4+q^6)$$

Par contre, $c_{461532} = c_{416532}\, c_2$, mais le polynôme de Poincaré de l'intervalle $[123456, 461532]$ ne s'obtient pas en spécialisant T_i en $-1/q$, car $[4, 6, 1, 5, 3, 2]$ est singulière.

La factorisation du polynôme de Poincaré, dans le cas non singulier, est due à Carrell et Peterson [Ca]. La caractérisation des variétés de Schubert non singulières est donnée par Lakshmibai et Seshadri [La-Se].

La densité du nombre de permutations non singulières tend vers 0 lorsque n tend vers l'infini. En fait, [La-Sa] ont montré que

PROPOSITION *Dans $\mathfrak{S}(4)$, seules $[3, 4, 1, 2]$ et $[4, 2, 3, 1]$ sont singulières et $\mu \in \mathfrak{S}(n)$ est non-singulière ssi l'image de μ par toute projection $\mathfrak{S}(n) \to \mathfrak{S}(4)$ est différente de ces deux permutations.*

En d'autres termes, il y a deux types élémentaires de singularités. Dans ce qui suit, nous montrons que les constructions du paragraphe 3 permettent d'expliciter les polynômes $P_{\nu,\mu}$ pour toutes les permutations évitant le motif $[3, 4, 1, 2]$ (i.e. celles qui n'ont jamais $[3, 4, 1, 2]$ comme image par projection).

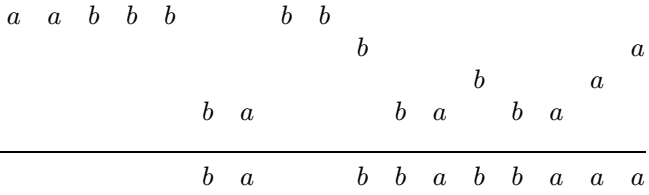
Il est commode de changer les conventions, et de noter, pour $\mu, \nu \in \mathfrak{S}(n)$, $P_\mu(\nu)(q) := P_{\nu\omega,\mu\omega}(1/q^2)$.

Lorsque μ est bigrassmannienne, alors la variété de Schubert correspondante est dite *déterminantale*, et la géométrie, ou le calcul direct, montrent que les polynômes $P_\mu(\nu)$ sont des polynômes de Gauss. Plus précisément,

LEMME *Soit $\mu = [[a, b, c, d]]$ une bigrassmannienne, et $\beta^i := [[a-i, b+i, c+i, d-i]]$, si $0 \le i \le min(a, d)$, $\beta^i := \infty$ sinon. Soit $k = min(b, c)$. Pour tout $\nu \ge \mu$, il existe un unique i, dit niveau par rapport à $\mu = \beta^0$, tel que $\nu \ge \beta^i$ et $\nu \not\ge \beta^{i+1}$, et alors $P_\mu(\nu) = \begin{bmatrix} k+i \\ i \end{bmatrix} = (1-q^{k+i}) \cdots (1-q^{k+1})/(1-q) \cdots (1-q^i)$.*

Les polynômes de Gauss correspondent à des arbres linéaires, mais plus généralement, il est facile d'associer à toute permutation grassmanienne γ un arbre ainsi qu'il est expliqué en [L-S2]. Soit en effet i la descente de γ; alors $(\gamma_1 - 1, \gamma_2 - 2, \ldots, \gamma_i - i)$ est une partition λ (croissante), et la lecture de la frontière nord-est du diagramme de Ferrers de λ donne un mot en a, b (a= pas vertical, b = pas horizonal), dont on extrait un sous-mot maximal de lettres appariées par couples successifs ba (cette opération est utilisée pour définir une action du groupe symétrique sur les mots [L-S1], ainsi qu'en théorie des graphes cristallins [K-N], [LLT1]).

Ainsi, pour $\gamma = [1, 2, 7, 12, 15, 16, 17, 3, 4, 5, 6, 8, 9, 10, 11, 13, 14]$, on a $\lambda = (0, 0, 4, 8, 10, 10, 10)$. La frontière se lit $a^2b^4ab^4ab^2a^3$. Disposant ce mot planairement de sorte à faire apparaître les appariements dans les horizontales, les lettres non appariées (à éliminer) étant dans la ligne supérieure



Ce dernier mot ($bab^2ab^2a^3$) est le parcours d'un arbre Γ (b = s'éloigner de la racine, a = s'en rapprocher). Soient $\alpha_1, \ldots, \alpha_r$ les branches terminales de Γ.

L'arbre Γ définit une fonction $f_\Gamma : \mathbb{N}^r \to \mathbb{N}[q]$ comme suit : un *étiquetage* E de Γ est un morphisme croissant de l'ensemble des arêtes de Γ dans $\mathbb{N}$ et son *poids* $p(E)$ est la somme des étiquettes. Alors (voir un exemple à la fin)

$$f_\Gamma(i_1, \dots, i_r) := \sum_{E \in \mathcal{E}(i_1, \dots, i_r)} q^{p(E)} \ , \tag{4.7}$$

somme sur toutes les étiquetages de Γ tels que les étiquettes de $\alpha_1, \dots, \alpha_r$ soient majorées respectivement par $i_1, \dots, i_r$. Le résultat essentiel de [L-S2] est

PROPOSITION *Soient γ une permutation grassmannienne, Γ l'arbre associé et r son nombre de branches terminales. Alors pour toute permutation ν, il existe des entiers $i_1, \dots, i_r$ tels que*

$$P_\gamma(\nu) = f_\Gamma(i_1, \dots, i_r) \ . \tag{4.8}$$

Zelevinsky [Zel] a donné une désingularisation explicite des variétés de Schubert indicées par des permutations grassmanniennes, qui relève la construction combinatoire précédente.

Le *code* d'une permutation $\mu \in \mathfrak{S}(n)$ est le vecteur $[c_1, \dots, c_n] \in \mathbb{N}^n$ tel que $c_i = card\big(\{j > i, \mu_j < \mu_i\}\big)$. Réordonnant le code en une partition, on peut donc associer à toute permutation μ un arbre $\Gamma(\mu)$, comme on l'a fait plus haut.

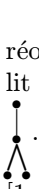
Le théorème suivant ([La]) montre que cet arbre continue à fournir tous les polynômes de Kazhdan & Lusztig dans le cas où μ est *vexillaire*, i.e. lorsqu'aucune projection de μ dans $S(4)$ n'est égale à $[2, 1, 4, 3]$.

THÉORÈME *Soient μ une permutation vexillaire, Γ l'arbre associé et $\beta_1, \dots, \beta_r$ les bigrassmanniennes maximales dans $[1, \mu]$. Alors pour tout $\nu \geq \mu$, le polynôme de Kazhdan & Lusztig est égal à*

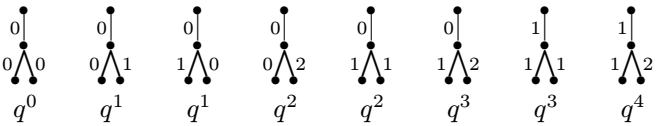
$$P_\mu(\nu) = f_\Gamma(i_1, \dots, i_r) \ , \tag{4.9}$$

où $i_1, \dots, i_r$ sont les niveaux respectifs de ν par rapport à $\beta_1, \dots, \beta_r$.

En fait, dans le cas d'une permutation vexillaire, pour tout r-uple $i_1, \dots, i_r$, tel que $\beta_1^{i_1}, \dots, \beta_r^{i_r} \neq \infty$, le supremum ζ^I de $\beta_1^{i_1}, \dots, \beta_r^{i_r}$ (calculé dans le treillis enveloppant) est une permutation, d'après [L-S6]. Sous les hypothèses du théorème, on a alors $P_\mu(\nu) = P_\mu(\zeta^I)$.

Par exemple, $\mu = [1, 5, 6, 2, 7, 3, 4, 8, 9]$ a pour code [033020000], qui se réordonne en la partition $[\dots, 0, 2, 3, 3]$. La frontière de cette dernière se lit $(\cdots aa)\, b\, b\, a\, b\, a\, a\, (bb\cdots)$ et le mot réduit *bbabaa* est le parcours de l'arbre  . Il y a deux bigrassmanniennes maximales en dessous de μ, qui sont $[1, 5, 6, 2, 3, 4, 7, 8, 9] = [[1, 2, 3, 3]] = \beta_1$ et $[1, 2, 5, 6, 7, 3, 4, 8, 9] = [[2, 3, 2, 3]] = \beta_2$. Le paramètre i_1 ne peut prendre que les valeurs 0, 1, puisque $\beta_1^1 = [[0, 3, 4, 2]]$, $\beta_1^2 = \infty$; par contre, $i_2 \in \{0, 1, 2\}$ puisque $\beta_2^1 = [[1, 4, 3, 1]]$, $\beta_2^2 = [[0, 5, 4, 0]]$, $\beta_2^3 = \infty$.

Le supremum de β_1^1 et β_2^2 est $\zeta^{1,2} = \beta_2^2 = [5,6,7,8,9,1,2,3,4]$. Toute permutation ν au dessus de cette dernière va donner le même polynôme $P_\mu(\zeta^{1,2}) = f_\Gamma(1,2) = 1 + 2q + 2q^2 + 2q^3 + q^4$, fourni par l'énumération



BIBLIOGRAPHIE

[BGG] I.N. Bernstein, I.M. Gelfand, S.I. Gelfand, *Schubert cells and the cohomology of the spaces G/P*, Russian Math. Surveys 28 (1973), 1-26.

[Bi] G. Birkoff, *Lattice theory*, 3ème ed., Am. Math. Soc., Providence (1967).

[B-B] A. Björner, F. Brenti, *Affine permutations of type A*, Electronic J. Comb. 3# R18 (1996).

[Br] F. Brenti, *A combinatorial formula for Kazhdan-Lusztig polynomials*, Invent. M. 118 (1994) 371-394.

[Ca] J. Carrell, *The Bruhat graph of a Coxeter group, a conjecture of Deodhar, and rational smoothness of Schubert varieties*, Proc. Symp. Pure M. 56 (1994)53–61.

[Ch] I.V. Cherednik, *Quantum groups as hidden symmetries of classic representation theory*, *in* Differential geometric methods in theoretical physics (A.I. Solomon ed.), World Scientific, Singapore, 1989, 47–54.

[De] V.V. Deodhar, *Some characterizations of Bruhat ordering on a Coxeter group*, Invent. Math. 39 (1977) 187–198.

[DKLLST] G. Duchamp, D. Krob, A. Lascoux, B. Leclerc, T. Scharf, J.-Y. Thibon, *Euler-Poincaré characteristic and polynomial representations of Iwahori-Hecke algebras*, Publ. RIMS Kyoto 31 (1995), 179–201.

[Eh] C. Ehresmann, *Sur la topologie de certains espaces homogènes*, Ann. Math. 35 (1934) 396–443.

[Er] H. Eriksson, *Computational and combinatorial aspects of Coxeter groups*, thèse, KTH Stockholm (1994).

[F-K] S. Fomin, A.N. Kirillov, *The Yang-Baxter equation, symmetric functions and Schubert polynomials*, Proc. FPSAC Firenze, 1993, Discr. Math. 153 (1996) 123–143.

[G-K] M. Geck, S. Kim, *Bases for the Bruhat-Chevalley order on all finite Coxeter groups*, J. of Algebra 197 (1997) 278–310.

[Hu] J.E. Humphreys, *Reflection groups and Coxeter groups*, Cambridge Studies in Adv.Math. 29, Cambridge Univ. Press (1990).

[K-L1] D. Kazhdan, G. Lusztig, *Representations of Coxeter groups and Hecke algebras* Invent. M. 53 (1979) 165–184.

[K-L2] D. Kazhdan, G. Lusztig, *Schubert varieties and Poincaré duality* Proc. Symp. Pure M. 34, A.M.S. (1980) 185–203.

[K-N] M. Kashiwara, T. Nakashima, Crystal graphs for representations of the q-analogue of classical Lie algebras, RIMS preprint 767 (1991).

[La-Sa] V. Lakshmibai, B. Sandhya, *Criterion for smoothness of Schubert varieties*, Proc.Indian Acad. Sc. M. 100 (1990) 45–52.

[La-Se] V. Lakshmibai, C.S. Seshadri, *Geometry of G/P. Singular locus of a Schubert variety*, Bull. A.M.S. 2 (1984) 363–366.
[La] A. Lascoux, *Polynômes de Kazdhan-Lusztig pour les variétés de Schubert vexillaires*, C.R. Acad. Sci. Paris, 321 (1995) 667–670.
[LLT1] A. Lascoux, B. Leclerc et J.Y. Thibon, *Crystal graphs and q-analogues of weight multiplicities for root systems of type A_n*, Letters in Math. Physics, 35 (1995) 359–374.
[LLT2] A. Lascoux, B. Leclerc et J.Y. Thibon, *Flag Varieties and the Yang-Baxter Equation*, Letters in Math. Phys. 40 (1997) 75–90.
[L-P] A. Lascoux, P. Pragacz, *Operator calculus for Q-polynomials and Schubert polynomials*, à paraître aux Adv. in Math. (1998).
[L-S1] A. Lascoux, M.P. Schützenberger, *Le monoïde plaxique, in* Non Commutative Structures, Napoli 1978, Quaderni della Ricerca, Roma 109 (1981) 129–156.
[L-S2] A. Lascoux, M.P. Schützenberger, *Polynômes de Kazdhan-Lusztig pour les grassmanniennes*, Astérisque 87-88 (1981) 249–266.
[L-S3] A. Lascoux, M.P. Schützenberger, Symmetry and Flag manifolds, *in Invariant Theory, Springer L.N.* 996 (1983) 118–144.
[L-S4] A. Lascoux, M.P. Schützenberger, *Symmetrization operators on polynomial rings*, Funk. Anal. 21 (1987) 77–78.
[L-S5] A. Lascoux, M.P. Schützenberger, *Algèbre des différences divisées*, Discrete Maths 99 (1992) 165–179.
[L-S6] A. Lascoux, M.P. Schützenberger, *Treillis et bases des groupes de Coxeter*, Electronic J. of Comb. 3 # R27 (1996).
[Lu] G. Lusztig, *Intersection Cohomology Methods in Representation Theory*, ICM Tokyo (1990) 155–174.
[MRR] W.H. Mills, D.P. Robbins, H. Rumsey, *Alternating sign matrices and descending plane partitions*, J. Comb. Th. A (1983) 340-359.
[Pr] R.A. Proctor, *Classical Bruhat orders and lexicographic shellability*, J. Alg. 77 (1982) 104–126.
[Ve] S. Veigneau, *ACE, an algebraic environment for the computer algebra system MAPLE*, `http://phalanstere.univ-mlv.fr/~ ace` (1998).
[Ya] C.N. Yang, *Some exact results for the many-body problem in one dimension with repulsive delta-function interaction*, Phys. Rev. Lett. 19 (1967), 1312–1315.
[Zei] D. Zeilberger, *Proof of the alternating sign matrix conjecture*, Electronic J. of Comb. 3 # R13 (1996).
[Zel] A. Zelevinsky, *Small resolutions of singularities of Schubert varieties* Funct. Anal. Pry. 17 (1983) 142–144.

Alain Lascoux
C.N.R.S., Institut Gaspard Monge
Université de Marne-la-Vallée,
5 Bd Descartes,
Champs sur Marne,
77454 Marne La Vallee Cedex 2
FRANCE

Mathematical Snapshots from the Computational Geometry Landscape

Jiří Matoušek

Abstract. We survey some mathematically interesting techniques and results that emerged in computational geometry in recent years.

1991 Mathematics Subject Classification: 068R99, 52C99, 52C10
Keywords and Phrases: computational geometry, combinatorial geometry, arrangement, Davenport-Schinzel sequence

We survey some mathematically interesting notions, techniques, and results that emerged in the field of computational geometry in recent years.

Computational geometry is a branch of theoretical computer science which constituted sometimes around the year 1980. It considers the design of efficient algorithms for computing with geometric objects in the Euclidean space $\mathbf{R}^d$. The objects are simple, like points, lines, spheres, etc., but there are many of them. The space dimension d is usually considered constant—many problems are studied mainly in the plane or in $\mathbf{R}^3$. As for general references, there is one fresh handbook [20] and another one pending [31]. A recent introductory textbook is [16]. Some mathematical spinoffs are nicely treated in [29].

Although this field mainly emphasizes algorithms, it has many fine purely mathematical results. I have selected a few of them for this overview quite subjectively (with many other, perhaps even nicer things omitted). Since they include the ideas of many researchers (my results being a tiny part only), it is not possible to give explicit credits to all of the contributors and to always refer to original sources (rather than surveys) in the limited space.

Combinatorial complexity of arrangements

The *arrangement* of a finite set of lines in the plane is a partition of the plane into cells of dimension 0, 1, and 2. The 0-cells (vertices) are the intersections of the lines, the 1-cells (edges) are the portions of the lines between vertices, and the 2-cells are the open convex polygons left after removing the lines from the plane. More generally, for a collection $H = \{h_1, h_2, \ldots, h_n\}$ of sets in $\mathbf{R}^d$, the arrangement of H is a decomposition of $\mathbf{R}^d$ into connected cells, where each cell is a connected component of the set of points lying in all of the sets h_i with $i \in I$ and in no h_j with $j \notin I$, for some index set $I \subseteq \{1, 2, \ldots, n\}$. In computational geometry, the most general sets considered in the role of the h_i's are usually the

so-called *surface patches*, which means $(d-1)$-dimensional closed semialgebraic sets defined by Boolean combinations of polynomial inequalities; moreover, both the number of inequalities and the degree of the polynomials are bounded by some constant.

Arrangements, especially arrangements of hyperplanes, have been investigated for a long time from various points of view. In the direction of research reflected, e.g., by the recent book [28], one is mainly interested in topological and algebraic properties of the whole arrangement. Computational geometers have mostly studied different aspects, primarily asymptotic bounds on the combinatorial complexity of various parts of arrangements,[1] and while the number n of sets in H is considered large, d is fixed (and small). Some important problems also lead to considering arrangements of less "regular" objects than hyperplanes, such as segments in the plane, triangles in space, or even pieces of complicated algebraic surfaces in $\mathbf{R}^d$. Two thorough and up-to-date surveys by Agarwal and Sharir in [31] complement our sketchy exposition here and in the next section.

The total complexity, i.e. the total number of cells, of an arrangement is quite well understood. Exact formulas are known for hyperplane arrangements, and fairly precise estimates exist for arrangements of surface patches (rough bounds for surface patches come from old papers in real-algebraic geometry by Petrov and Oleinik, Milnor, and Thom, and there are some recent refinements, such as [7]). The complexity is always at most $O(n^d)$.[2] More challenging problems concern the complexity of certain portions of the arrangements; some of them are schematically illustrated in Fig. 1.

The *zone* of a set $X \subseteq \mathbf{R}^d$ in an arrangement consists of the cells intersecting X. For hyperplane arrangements, the complexity of the zone of any hyperplane is $O(n^{d-1})$ [17]. The zone of a low-degree algebraic surface, or of an arbitrary convex surface, in a hyperplane arrangement has at most $O(n^{d-1}\log n)$ complexity [5].

The *level* k in a hyperplane arrangement consists of the $(d-1)$-dimensional cells, i.e. edges in the case of lines in $\mathbf{R}^2$, with exactly k of the hyperplanes below them (where the x_d-axis is considered vertical, say). The maximum complexity of the k-level is a tantalizing open problem even for lines in the plane; we refer to the paper by Welzl in this volume for more information.

Next, we discuss the *lower envelope* of an arrangement. Informally, this is the part of the arrangement that can be seen by an observer sitting at $(0, 0, \ldots, 0, -\infty)$. The lower envelope in an arrangement of hyperplanes is the surface of a convex polyhedron with at most n facets, whose maximum complexity, of the order $n^{\lfloor d/2 \rfloor}$, is known precisely (since McMullen's paper in 1970). This bound is trivial in the plane, but already for planar arrangements of segments, the lower envelope question is hard.

If we number the segments 1 through n and write down the numbers of the segments as they are encountered along the lower envelope from left to right, we get a

[1] If X is a set of cells in an arrangement, the *(combinatorial) complexity* of X is the number of cells of the arrangement that are contained in the closure of X. Typically, this complexity is asymptotically dominated by the number of vertices of the arrangement in the closure of X.

[2] Here and in the sequel, the constants hidden in the $O(.)$ and $\Omega(.)$ notations generally depend on d, and, in some cases, on other parameters declared fixed. For instance, here the constant also depends on the degree and formula size of the surface patches forming the arrangement.

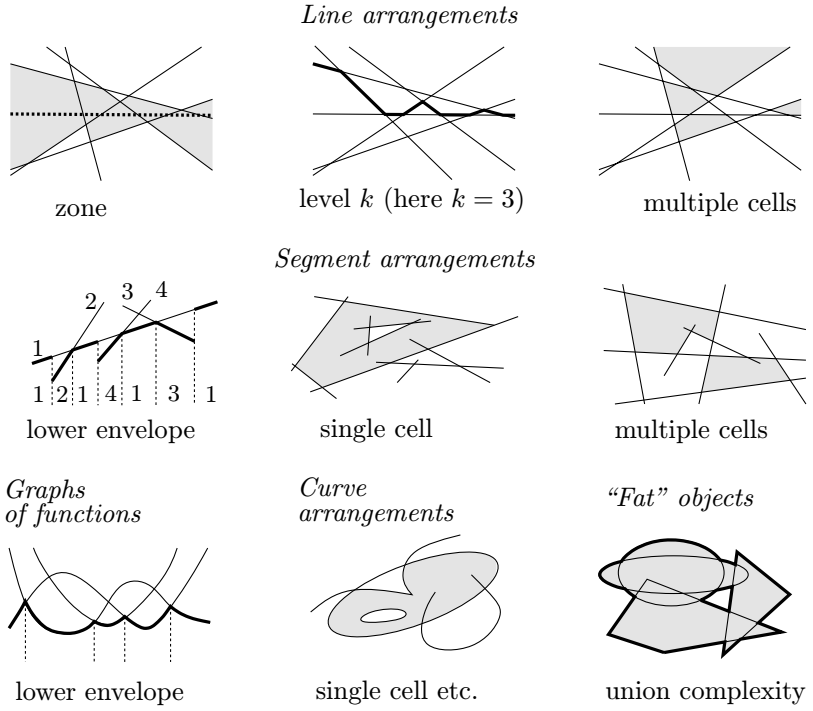


Figure 1: A bestiary of planar arrangement problems

sequence $a_1a_2a_3 \ldots a_m$, for which the following conditions hold: $a_i \in \{1, 2, \ldots, n\}$, $a_i \neq a_{i+1}$, and there is no (not necessarily contiguous) subsequence of the form $ababa$, where $a \neq b$. Any finite sequence satisfying these conditions is called a *Davenport-Schinzel sequence* (or DS-sequence for short) of order 3 over the symbols $1, 2, \ldots, n$. For DS-sequences of order s, the forbidden pattern is $abab\ldots$ with $s+2$ letters. Such sequences are obtained, e.g., from lower envelopes of x-monotone curves (i.e. graphs of univariate functions), such that any two of the curves intersect in at most s points (a typical example are graphs of degree-s polynomials). Davenport and Schinzel started investigating $\lambda_s(n)$, the maximum possible length of a DS-sequence of order s over n symbols, in 1965. Fairly precise estimates (asymptotically tight for many s's) were proved by Sharir, Hart, Agarwal, and Shor in the late 1980s (see [33] for an account). The results are remarkable: while $\lambda_1(n)$ and $\lambda_2(n)$ are easily seen to be linear, for any fixed $s \geq 3$, $\lambda_s(n)/n$ grows to infinity with $n \to \infty$, but incredibly slowly. For example, $\lambda_3(n)$ is asymptotically bounded by constant multiples of $n\alpha(n)$ from both above and below, where $\alpha(n)$ is the inverse of the Ackermann function.[3] For all practical purposes, for each fixed

[3]If we define a hierarchy of functions by $f_1(n) = 2n$ and $f_{k+1}(n) = f_k \circ f_k \circ \cdots \circ f_k(2)$ ($(n-1)$-fold composition), then the Ackermann function of n is $A(n) = f_n(n)$, and $\alpha(n) = \min\{k \geq 1 : A(k) \geq n\}$. For example, $A(4)$ is an exponential tower of 2s of height 2^{16}.

$s \geq 3$, $\lambda_s(n)$ behaves like a linear function, but it is nonlinear in a very subtle manner, and hence any proofs of the correct bounds must be quite complicated.

The maximum complexity of the lower envelope for segments is at most $\lambda_3(n) = O(n\alpha(n))$, and a construction by Wiernik and Sharir, later simplified by Shor, provides an arrangement of segments with lower envelope of complexity $\Omega(n\alpha(n))$. Thus, similar to DS-sequences, lower envelopes of segments are no laughing matter.

Before proceeding with the discussion of lower envelopes, we mention recent developments in generalized DS-sequences. In the original definition, the forbidden pattern *ababa* ... is made of two letters. Klazar, Valtr, and Adamec studied forbidden patterns consisting of more letters, such as *abccbaabc* (for a forbidden pattern with k distinct letters, an analogue of the condition $a_i \neq a_{i+1}$ for DS-sequences is that any k consecutive symbols in the sequences be all distinct). They proved that for any fixed forbidden pattern, the maximum length of a sequence in n symbols is near-linear in n, and they characterized numerous cases where a linear bound holds (see e.g. [22, 23]). One forbidden pattern of the latter type is *abcdedcbabcde* (or analogous with more letters); this result was used by Valtr [35] for solving interesting problems concerning geometric graphs. A geometric graph is a drawing of a graph in the plane with edges drawn as straight segments (possibly crossing); they have recently been studied by Pach, Katchalski, Last, Károlyi, Tóth, and others.

The main result for lower envelopes in higher dimensions is quite recent, due to Sharir and Halperin [21, 32]. For an arrangement of surface patches in $\mathbf{R}^d$, with some mild additional technical assumptions, they prove lower envelope complexity bound of $O(n^{d-1+\varepsilon})$ for any fixed $\varepsilon > 0$, which is nearly tight (there is an $\Omega(n^{d-1}\alpha(n))$ lower bound). As a sample of techniques in the area, we demonstrate this proof in the planar case. This is a ridiculous setting, since here much better results are obtained via DS-sequences, but the higher-dimensional case is too complicated to fit here.

So let us consider a set H of n x-monotone curves (such as in Fig. 1 bottom left), any two intersecting in at most s points (s fixed). Moreover, assume for convenience that no 3 curves have a common intersection. Let $L = L(H)$ be the set of vertices on the lower envelope and let $f(n)$ denote the maximum possible cardinality of L in this situation. We aim at proving $f(n) = O(n^{1+\varepsilon})$.

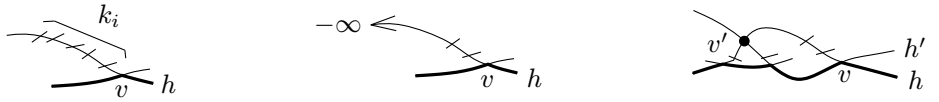
First, let k be an auxiliary parameter, $2 \leq k \leq \frac{n}{2}$, let $L^{<k}$ be the set of vertices in the arrangement of H at level smaller than k (i.e. with fewer than k curves below them), and let $f^{<k}(n)$ be the maximum possible cardinality of $L^{<k}$.

Lemma. $f^{<k}(n) = O\big(k^2 f(\lfloor n/k \rfloor)\big)$.

Here is a beautiful probabilistic argument of Clarkson and Shor [15]. Suppose that $f^{<k}(n)$ is attained for H, set $r = \lfloor n/k \rfloor$, and let $R \subset H$ be an r-element subset of H picked uniformly at random. First, we lower-bound the expected size of $L(R)$. Consider a vertex $v \in L^{<k}(H)$ at a level $j < k$. Such a v appears in $L(R)$ iff both the curves defining v fall in R and none of the j curves below v does, and so $\mathrm{Prob}\,[v \in L(R)] = \binom{n-2-j}{r-2}/\binom{n}{r}$. Calculation shows that this probability is $\Omega(k^{-2})$, and so the expected size of $L(R)$ is $\Omega(k^{-2} f^{<k}(n))$. At the same time, $|L(R)| \leq f(r)$ for all R, and the lemma follows by comparing these two bounds.

Next, we partition the set $L = L(H)$ into subsets $L_1, \ldots, L_s$, with L_i consisting of the vertices of L that are the ith leftmost intersections of their two curves. Divide $L^{<k}$ similarly, and let $f_i(n)$ and $f_i^{<k}(n)$ be the corresponding maximum possible cardinalities.

The strategy of the proof is "there shouldn't better be any vertices on the lower envelope, and if there are, someone is going to pay for it". To find out who pays for a vertex $v \in L_i$, we start walking from v to the left along the curve h passing through v and not being on the lower envelope on the left of v. We charge every vertex encountered $\frac{1}{k_i}$ units, where k_i is an integer parameter (to be fixed later). If k_i vertices are encountered without returning to the lower envelope or escaping to $-\infty$ then the charging is complete. Otherwise, if we end up at $-\infty$, we charge 1 to the curve h itself. Finally, if we are back at the lower envelope without having passed at least k_i vertices then, crucially, we must have crossed the second curve h' defining the vertex v again, at a vertex $v' \in L_{i-1}^{<k_i}$, and this v' pays 1 for v. A picture illustrates these three cases of charging:



If we do this charging for all vertices $v \in L_i$ then, altogether, each curve was charged at most 1 and each vertex of $L^{<k_i}$ was charged at most $\frac{2}{k_i}$, except possibly for vertices of $L_{i-1}^{<k}$, which could each be charged 1 extra. Since at least 1 unit was paid for each vertex of L_i, we obtain $f_i(n) \le n + \frac{2}{k_i} f^{<k_i}(n) + f_{i-1}^{<k_i}(n)$.

By substituting for $f^{<k_i}$ and $f_{i-1}^{<k_i}$ the bound from the lemma, we arrive at the system of inequalities $f_i(n) \le n + O\big(k_i f(\lfloor n/k_i \rfloor) + k_i^2 f_{i-1}(\lfloor n/k_i \rfloor)\big)$, $i = 1, 2, \ldots, s$ (where we put $f_0 = 0$), and we also have $f \le f_1 + \cdots + f_s$. If one sets $k_i = n^{\varepsilon_i}$ with $0 < \varepsilon_1 \ll \varepsilon_2 \ll \cdots \ll \varepsilon_s \ll \varepsilon$, a not too difficult calculation shows that $f(n) = O(n^{1+\varepsilon})$ as claimed. □

Bounding the maximum complexity of a single cell is usually considerably more demanding than the lower envelope question, mainly because a cell can have a complicated topology (cells in hyperplane arrangements, no more complicated than the lower envelope, are a honorable exception). In the plane, these obstacles are not too formidable, and by a reduction to DS-sequences, it can be shown that the single-cell complexity for segments is $O(n\alpha(n))$, and for pieces of algebraic curves it can be bounded by some $\lambda_s(n)$, with s depending on the maximum degree of the curves. In $\mathbf{R}^3$, a general near-tight bound of $O(n^{2+\varepsilon})$ was proved in [21]. Some more special results are known for all d, such as an $O(n^{d-1} \log n)$ bound for a single cell in an arrangement of $(d-1)$-dimensional simplices in $\mathbf{R}^d$ [6]. Very recently, Basu proved, in an unpublished manuscript, that the sum of the Betti numbers (i.e. "topological complexity") of a single cell in an arrangement of surface patches in $\mathbf{R}^d$ is $O(n^{d-1})$. This might be helpful in getting good bounds on the combinatorial complexity too.

Concerning the union of "fat" objects (Fig. 1 bottom right), let us consider n convex sets in the plane, and let us ask what is the combinatorial complexity of the complement of their union. To get a meaningful problem, we assume that

the boundaries of any two sets intersect in at most s points for some fixed $s \geq 4$ ($s = 2$ is easy). Long and skinny sets can form a grid pattern and have union complexity about n^2, but if we also require that the sets be "fat" (the ratio of the circumradius and inradius is bounded by some constant K), then a recent result of Efrat and Sharir [18] shows that the union complexity is near-linear, at most $O(n^{1+\varepsilon})$, with the constant of proportionality depending on s, K, ε ([26] gives a simpler and more precise bound for fat triangles). Various extensions to non-convex cases or to higher dimensions seem easy to conjecture but quite hard to prove.

There are still many open problems in the above-discussed areas, but what seems to be needed most at the moment is a simplification and streamlining, since building up on the existing proofs is getting more and more cumbersome.

Here is an annoying open problem concerning arrangements of n algebraic surfaces in $\mathbf{R}^d$. If the degrees of the surfaces are bounded, the complexity of the arrangement is $O(n^d)$. But the cells can be combinatorially very complicated, while for many applications, one needs to work with cells definable by constant-size formulas, the so-called *Tarski cells* (curved analogues of simplices, so to speak). Can each of the cells of the arrangement be subdivided into Tarski cells, in such a way that altogether $O(n^d)$ Tarski cells result? The best known upper bound for $d \geq 3$ is a bit larger than $O(n^{2d-3})$ [11].

Multiple cells, incidences, cuttings

Besides a single cell, also the total complexity of several cells in an arrangement has been studied, and this has interesting connections to some old combinatorial-geometric problems. Let us consider some m 2-cells in a planar arrangement of n lines (call them *marked cells*), and let us denote the maximum possible total number of vertices of these cells by $K(n, m)$. While $K(n, 1) = n$, $K(n, m)$ is considerably smaller than mn for large m.

To get a nontrivial upper bound on $K(n, m)$, we define a bipartite graph with the lines and the marked cells as vertices and with edges connecting each cell to the lines forming its sides. There cannot be 5 lines simultaneously connected to the same two cells, and the Kővári-Sós-Turán theorem in extremal graph theory implies that there are $O(m\sqrt{n} + n)$ edges; thus $K(n, m) = O(m\sqrt{n} + n)$. In particular, $K(n, \sqrt{n}) = O(n)$, (this is a result of Canham from 1969), which is obviously tight. But the bound is not tight for $n = m$, say, and the right bound is $K(n, m) = O(n^{2/3}m^{2/3} + n + m)$. This was proved by Clarkson et al. [14], using a general technique that emerged in previous work on geometric algorithms. We give the proof for $m = n$. The basic idea is this: since the bound we already have is good if there are many more lines than points, we subdivide the problem with n lines and n points into smaller subproblems, most of them with many more lines than points. The device for this subdivision is the so-called $\frac{1}{r}$-cutting.

For a parameter $r \geq 1$ and a set L of n lines in the plane, a $\frac{1}{r}$-*cutting* for L is a finite set of triangles[4] with disjoint interiors covering the plane, such that

[4]Where unbounded triangles are admitted too, i.e. a triangle means an intersection of 3 halfplanes here.

the interior of each triangle is intersected by no more than $\frac{n}{r}$ lines of L. A basic existence result says that for any L and r, a $\frac{1}{r}$-cutting exists consisting of $O(r^2)$ triangles (note that the bound is independent of n). Three proofs are known: a very elementary one [24], and two probabilistic ones which generalize to higher dimensions [12, 10].

For bounding $K(n,n)$, let L be the n considered lines, set $r = n^{1/3}$, and consider a $\frac{1}{r}$-cutting $\{\Delta_1, \ldots, \Delta_q\}$ for L, $q = O(r^2)$. Let $L_i \subset L$ be the set of lines intersecting the interior of Δ_i and suppose that there are m_i marked cells completely contained in Δ_i. The total complexity of these marked cells, over all Δ_i, is at most $\sum_{i=1}^{q} K(|L_i|, m_i) \leq \sum_{i=1}^{q} O(m_i\sqrt{n/r} + \frac{n}{r}) = O(n^{3/2}r^{-1/2} + nr)$, using the above-derived bound for $K(n,m)$ and $\sum m_i \leq n$. It remains to account for the marked cells intersecting boundaries of some of the Δ_i's. But each vertex of such a marked cell lies in the zone of a side of some Δ_i in the arrangement of L_i, and the total complexity of these zones is at most $3\sum_{i=1}^{q} O(|L_i|) = O(nr)$. Altogether we get $K(n,n) = O(n^{4/3})$. □

An easy consequence of the bound $K(n,m) = O(n^{2/3}m^{2/3} + m + n)$ is the same (and also tight) bound for the maximum number of incidences between n lines and m points in the plane. This bound for incidences was proved earlier by Szemerédi and Trotter, and the new proof via $\frac{1}{r}$-cuttings [14] was a considerable simplification. A still much simpler proof was found later by Székely [34] via geometric graphs, but so far his technique seems mainly applicable for problems in the plane, while with $\frac{1}{r}$-cuttings, various higher-dimensional problems can be handled too (see, e.g., [14, 29] or a survey by Agarwal and Sharir in [31] for more results and references).

The perhaps most challenging related problem is Erdős' question on unit distances: given n points in the plane, what is the maximum possible number of pairs of points at distance 1? By drawing a unit circle around each point, the question can be reduced to the maximum number of incidences between n points and n unit circles. Both Székely's technique and the one with $\frac{1}{r}$-cuttings yield the same $O(n^{4/3})$ bound as for line-point incidences, but while for lines this is tight, the best known lower bound for unit circles is only slightly superlinear. To decrease the upper bound for the unit-distance problem, a radically new approach seems to be needed, because the $n^{4/3}$ bound is tight for *pseudocircles*, i.e. collections of Jordan curves that combinatorially behave "like unit circles", and none of the known methods can take advantage of "true circularity" of the unit circles.

In this connection, a recent result of Elekés and Rónyai [19] should be mentioned. They characterized bivariate polynomials and rational functions that attain only $O(n)$ distinct values on $X \times Y$ for some n-element sets $X, Y \subset \mathbf{R}$. As a special case, they settled a conjecture of Purdy: if u and v are lines and $P \subset u$ and $Q \subset v$ are n-point sets such that the distance $|p - q|$ attains only $O(n)$ distinct values for $p \in P$ and $q \in Q$, then u and v must be parallel or perpendicular (provided n is large enough). The proof is in part algebraic and it strongly uses the "straightness" of the lines u and v.

RANGE SEARCHING, PARTITIONS, HEILBRONN'S PROBLEM

Let us consider the following algorithmic problem. Given an n-point set $P \subset \mathbf{R}^2$, we want to build some data structure for storing information about P, in such a way that if we get a stripe σ (bounded by two parallel lines) as a query, the number of points of P lying in σ can be determined quickly, hopefully much faster than by examining all points of P. Moreover, we insist that the space occupied by the data structure is at most proportional to n.

Questions of this type, the so-called *range searching problems*, have been studied quite intensively and in a much more general form—in higher dimensions, with different query shapes, with more space allowed, etc. (there is a survey by Agarwal in [20], and another survey is [25]). But many interesting aspects can be demonstrated on the particular problem formulated above. In this case, it is possible to answer the query in $O(\sqrt{n})$ time, and with some restriction on the type of algorithm used, this is asymptotically optimal. Ironically, while the known data structures for this problem are not very useful in practice, the underlying theory involves some of the nicest mathematics in computational geometry.

At first sight (and probably at many subsequent sights too), it is not clear how to achieve any sublinear query time. Willard discovered in 1981 that the following type of geometric construction can be used: given the point set P, partition the plane into some number r of regions, each containing roughly $\frac{n}{r}$ points of P, in such a way that no line intersects more than κ of these regions, where κ should be considerably smaller than r. How can this help with a query? We store the number of points in each of the regions. Given a query stripe σ, the boundary of σ intersects at most 2κ regions. These must be further examined, but each of the other regions can be processed in unit time using the stored point counts. The actual algorithms are more complicated but this is the basic idea.

Finding an optimal construction of such a partition took a long time. (Looking for good partitions stimulated, for instance, research in equipartitioning masses by hyperplanes—see e.g. [30]—although other approaches were used in the subsequent development.) One of the most important steps was the following result, essentially invented by Welzl, with a slight improvement in [13]: any $2n$-point set in the plane can be divided into pairs of points in such a way that any line crosses only $O(\sqrt{n})$ of the segments connecting the pairs. One almost wouldn't believe that after thousands of years of geometry, it is still possible to discover such pretty theorems about points in the plane. This was later generalized to a partition of an n-point set into r parts of size roughly $\frac{n}{r}$, with any line crossing $O(\sqrt{r})$ parts only (see [25]). Both these results are asymptotically optimal. The research in range searching also initiated a fruitful theory related to the so-called Vapnik-Chervonenkis dimension of set systems, with applications, e.g., in discrepancy theory; this is surveyed in [27].

Lower bounds for range searching were proved mainly by Chazelle; a key paper is [9]. In the proof, some integral-geometric considerations appear, and, interestingly, the lower bounds are related to a generalization of Heilbronn's problem from discrete geometry. For an n-point set $P \subset [0,1]^2$ and $3 \le k \le n$, let $a_k(P)$ denote the minimum area of the convex hull of a k-point subset of P. Heilbronn's problem

asks for determining $a_3(P)$, and although the answer is unknown, it is known that $a_3(P)$ is of much smaller order than $\frac{1}{n}$ (which is what one might perhaps expect at first). In Chazelle's proof, one needs a set P with $a_k(P) = \Omega(\frac{k}{n})$ for all $k \in [k_0, n]$, with k_0 as small as possible. He achieves this with $k_0 \approx \log n$, and this causes the presence of an $\log n$ factor in the range-searching lower bound in $\mathbf{R}^3$ which probably shouldn't be there. From Heilbronn's problem, we know that $k_0 = 3$ is impossible to reach, but perhaps it might be possible to decrease k_0 to something smaller than $\log n$, which would improve the range-searching bound. For a more recent progress in range-searching lower bounds, and some nice geometric problems, see [8].

Many other areas and results would deserve to be mentioned, such as the developments related to linear programming algorithms (see the survey [1]) which also led to a nice purely mathematical application by Amenta [3] (a short proof of a Helly-type result), or the story of *weak ε-nets*, born in computational geometry and later used by Alon and Kleitman [2] in their solution of the long-open Hadwiger-Debrunner problem in convex geometry, or an interesting question of algebraic-topological nature arising in motion planning of multiple robots [4]. But it's really time to finish.

Acknowledgment. While working in computational geometry, I met many very nice people, and a number people helped me in various ways. I would like to use this opportunity to thank them all—I will not try to list names, in order not to omit anyone. I also thank Pankaj K. Agarwal, Eva Matoušková, Micha Sharir, Ricky Pollack, and Pavel Valtr for constructive comments on a draft of the present paper.

References

[1] P. K. Agarwal and M. Sharir. Algorithmic techniques for geometric optimization. In *Computer Science Today: Recent Trends and Developments*, volume 1000 of *Lecture Notes Comput. Sci.*, pages 234–253. Springer-Verlag, 1995.

[2] N. Alon and D. Kleitman. Piercing convex sets and the Hadwiger Debrunner (p, q)-problem. *Adv. Math.*, 96(1):103–112, 1992.

[3] N. Amenta. A short proof of an interesting Helly-type theorem. *Discr. Comput. Geom.*, 15:423–427, 1996.

[4] B. Aronov, M. de Berg, F. van der Stappen, and P. Švestka. Motion planning for multiple robots. In *Proc. 14th Annu. ACM Sympos. Comput. Geom.*, 1998.

[5] B. Aronov, M. Pellegrini, and M. Sharir. On the zone of a surface in a hyperplane arrangement. *Discrete Comput. Geom.*, 9(2):177–186, 1993.

[6] B. Aronov and M. Sharir. Castles in the air revisited. *Discrete Comput. Geom.*, 12:119–150, 1994.

[7] S. Basu, R. Pollack, and M.-F. Roy. On the number of cells defined by a family of polynomials on a variety. *Mathematika*, 43:120–126, 1996.

[8] H. Brönnimann, B. Chazelle, and J. Pach. How hard is halfspace range searching. *Discrete Comput. Geom.*, 10:143–155, 1993.

[9] B. Chazelle. Lower bounds on the complexity of polytope range searching. *J. Amer. Math. Soc.*, 2:637–666, 1989.

[10] B. Chazelle. Cutting hyperplanes for divide-and-conquer. *Discrete Comput. Geom.*, 9(2):145–158, 1993.

[11] B. Chazelle, H. Edelsbrunner, L. Guibas, and M. Sharir. A singly-exponential stratification scheme for real semi-algebraic varieties and its applications. In *Proc. 16th Internat. Colloq. Automata Lang. Program.*, volume 372 of *Lecture Notes Comput. Sci.*, pages 179–192. Springer-Verlag, 1989.

[12] B. Chazelle and J. Friedman. A deterministic view of random sampling and its use in geometry. *Combinatorica*, 10(3):229–249, 1990.

[13] B. Chazelle and E. Welzl. Quasi-optimal range searching in spaces of finite VC-dimension. *Discrete Comput. Geom.*, 4:467–489, 1989.

[14] K. Clarkson, H. Edelsbrunner, L. Guibas, M. Sharir, and E. Welzl. Combinatorial complexity bounds for arrangements of curves and spheres. *Discrete Comput. Geom.*, 5:99–160, 1990.

[15] K. L. Clarkson and P. W. Shor. Applications of random sampling in computational geometry, II. *Discrete Comput. Geom.*, 4:387–421, 1989.

[16] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, 1997.

[17] H. Edelsbrunner, R. Seidel, and M. Sharir. On the zone theorem for hyperplane arrangements. *SIAM J. Comput.*, 22(2):418–429, 1993.

[18] A. Efrat and M. Sharir. On the complexity of the union of fat objects in the plane. In *Proc. 13th Annu. ACM Sympos. Comput. Geom.*, pages 104–112, 1997.

[19] Gy. Elekes and L. Rónyai. A combinatorial problem on polynomials and rational functions. *J. Comb. Thoery Ser. B*, 1998. To appear.

[20] J. E. Goodman and J. O'Rourke, editors. *Handbook of Discrete and Computational Geometry*. CRC Press LLC, Boca Raton, FL, 1997.

[21] D. Halperin and M. Sharir. Almost tight upper bounds for the single cell and zone problems in three dimensions. *Discrete Comput. Geom.*, 14:385–410, 1995.

[22] M. Klazar. A general upper bound in extremal theory of sequences. *Comment. Math. Univ. Carol.*, 33:737–746, 1992.

[23] M. Klazar and P. Valtr. Generalized Davenport–Schinzel sequences. *Combinatorica*, 14:463–476, 1994.

[24] J. Matoušek. Construction of ϵ-nets. *Discrete Comput. Geom.*, 5:427–448, 1990.

[25] J. Matoušek. Geometric range searching. *ACM Comput. Surv.*, 26:421–461, 1994.

[26] J. Matoušek, J. Pach, M. Sharir, S. Sifrony, and E. Welzl. Fat triangles determine linearly many holes. *SIAM J. Comput.*, 23:154–169, 1994.

[27] J. Matoušek. Geometric set systems. In *Proceedings of the 2nd European Congress of Mathematicians*. Birkhäuser, Basel, 1998. In press.

[28] P. Orlik and H. Terao. *Arrangements of hyperplanes*. Springer-Verlag, Berlin etc., 1991.

[29] J. Pach and P. K. Agarwal. *Combinatorial Geometry*. John Wiley & Sons, New York, NY, 1995.

[30] E. A. Ramos. Equipartition of mass distributions by hyperplanes. *Discrete Comput. Geom.*, 15:147–167, 1996.

[31] J.-R. Sack and J. Urrutia, editors. *Handbook on Computational Geometry*. North-Holland, 1998. To appear.

[32] M. Sharir. Almost tight upper bounds for lower envelopes in higher dimensions. *Discrete Comput. Geom.*, 12:327–345, 1994.

[33] M. Sharir and P. K. Agarwal. *Davenport-Schinzel Sequences and Their Geometric Applications*. Cambridge University Press, Cambridge, 1995.

[34] L. Székely. Crossing numbers and hard Erdős problems in discrete geometry. *Combinatorics, Probability, and Computing*, 6:353–358, 1997.

[35] P. Valtr. On geometric graphs with no k pairwise parallel edges. *Discrete Comput. Geom.*, 19:461–469, 1998.

Jiří Matoušek
Dept. of Applied Mathematics
Charles University
Malostranské nám. 25
118 00 Praha 1, Czech Republic

Nets, (t,s)-Sequences, and Algebraic Curves over Finite Fields with Many Rational Points

Harald Niederreiter

Abstract. The current status of the theory of (t,m,s)-nets and (t,s)-sequences is presented in a brief form, with some emphasis on the connections with algebraic geometry. Closely related work on constructions of algebraic curves over finite fields with many rational points and on improving the Gilbert-Varshamov bound in algebraic coding theory is discussed as well.

1991 Mathematics Subject Classification: 05B15, 11G20, 11K38, 11R58, 11T71, 14G15, 14H05, 94B27, 94B65.
Keywords and Phrases: quasirandom points, orthogonal arrays, algebraic curves over finite fields, rational points, algebraic-geometry codes, Gilbert-Varshamov bound.

1. Introduction and basic concepts

Nets and (t,s)-sequences are finite point sets, respectively infinite sequences, satisfying strong uniformity properties with regard to their distribution in the s-dimensional unit cube $I^s = [0,1]^s$. The general theory of these combinatorial objects was first developed in [12]. They have attracted a lot of interest in scientific computing in recent years because of their role as quasirandom points in quasi-Monte Carlo methods, e.g. for numerical integration over I^s (see [14] for the details). They also offer a great appeal for theoretical studies in view of the many links with other areas such as classical combinatorial designs, coding theory, algebra, number theory, and algebraic geometry. To set the stage, we first review some basic definitions.

Definition 1. For a given dimension $s \geq 1$ and integers $b \geq 2$ and $0 \leq t \leq m$, a (t,m,s)-*net in base* b is a point set P consisting of b^m points in I^s such that every subinterval J of I^s of the form

$$J = \prod_{i=1}^{s} [a_i b^{-d_i}, (a_i+1)b^{-d_i})$$

with integers $d_i \geq 0$ and $0 \leq a_i < b^{d_i}$ for $1 \leq i \leq s$ and with $\mathrm{Vol}(J) = b^{t-m}$ contains exactly b^t points of P.

For integers $b \geq 2$ and $m \geq 1$ and a point $\mathbf{x} \in I^s$, we obtain $[\mathbf{x}]_{b,m} \in I^s$ by truncating a b-adic expansion of each coordinate of $\mathbf{x}$ after m terms. Here

expansions with almost all digits equal to $b-1$ are allowed – thus, the truncation operates on the expansions of the coordinates of $\mathbf{x}$ and not on $\mathbf{x}$ itself. The following definition of a (t,s)-sequence is the slightly generalized version described in [20], [21] (see [14, Chapter 4] for the original narrower definition). We assume prescribed b-adic expansions on which the truncations operate.

DEFINITION 2. For a given dimension $s \geq 1$ and integers $b \geq 2$ and $t \geq 0$, a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of points in I^s is a *(t,s)-sequence in base b* if for all integers $k \geq 0$ and $m > t$ the points $[\mathbf{x}_n]_{b,m}$ with $kb^m \leq n < (k+1)b^m$ form a (t,m,s)-net in base b.

The following useful principle shows that if we can construct a (t,s)-sequence, then we can construct infinitely many nets in dimension $s+1$ (see [12, Section 5], [20, Section 6]).

LEMMA 1. *If there exists a (t,s)-sequence in base b, then for every integer $m \geq t$ there exists a $(t,m,s+1)$-net in base b.*

The aim in the construction of (t,m,s)-nets and (t,s)-sequences in base b is to make the quality parameter t as small as possible if the other parameters are fixed. Most of the known constructions of nets and (t,s)-sequences are based on the digital method which was introduced in [12, Section 6]. For the sake of brevity, we just sketch the digital method for constructing (t,m,s)-nets in base b. Select a commutative ring R with identity and of finite order $b \geq 2$. For given $m \geq 1$ and $s \geq 1$ choose a system

$$C = \left\{ \mathbf{c}_j^{(i)} \in R^m : 1 \leq i \leq s, 1 \leq j \leq m \right\}.$$

Now we get the jth b-adic digits of the ith coordinates of the points of the (t,m,s)-net by forming the inner product of $\mathbf{c}_j^{(i)}$ with all elements of R^m and then identifying elements of R with b-adic digits. The value of the quality parameter t depends on the choice of C. The resulting net is called a *digital (t,m,s)-net in base b* (or *constructed over R* if we want to emphasize R). Similarly, we speak of a *digital (t,s)-sequence in base b* (or *constructed over R* if we want to emphasize R). There is a "digital" analog of Lemma 1, i.e., a digital (t,s)-sequence yields infinitely many digital nets in dimension $s+1$ (see [20, Section 2]). For practical purposes it suffices to consider the digital method in the special case where the ring R is a finite field $\mathbf{F}_q$ of prime-power order q. Digital nets and (t,s)-sequences in an arbitrary base b can be obtained by using rings R that are direct products of finite fields (see [14, Chapter 4], [20, Section 5]).

In this paper we give a brief review of the state-of-the-art in the area of (t,m,s)-nets and (t,s)-sequences, with some emphasis on the connections with algebraic geometry. Section 2 discusses links with classical combinatorial objects such as MOLS and orthogonal arrays. Constructions of nets and (t,s)-sequences, e.g. by methods using algebraic curves over finite fields, are presented in Section 3. This leads to the discussion of algebraic curves over finite fields with many rational points in Section 4. As a by-product we obtain the applications to algebraic coding theory in Section 5, such as improvements on the Gilbert-Varshamov bound. For various aspects, more detailed expository accounts can be found in [14, Chapter 4], [21], [25], [32].

2. Connections with combinatorial designs

The fact that there are close links between nets and combinatorial designs was noticed already in [12, Section 5]. For instance, it was shown there that for $s \geq 2$ the existence of a $(0, 2, s)$-net in base b is equivalent to the existence of $s-2$ MOLS of order b. Later it was proved by Mullen and Whittle [11] that for $s \geq 2$ and any $t \geq 0$, the existence of a $(t, t+2, s)$-net in base b is equivalent to the existence of a certain set of mutually orthogonal hypercubes of order b. In the language of orthogonal arrays, there is the result in [15] that there exists a $(t, t+2, s)$-net in base b if and only if there exists an orthogonal array $\mathrm{OA}(b^{t+2}, s, b, 2)$ of index b^t.

Lawrence [6] and Mullen and Schmid [10] independently established a combinatorial equivalence between arbitrary (t, m, s)-nets in base b and suitable combinatorial designs. Depending on the language that is used, these designs can be generalized orthogonal arrays, ordered orthogonal arrays, or strongly orthogonal hypercubes. The proofs of all these combinatorial results are constructive.

These connections with combinatorial designs imply obstructions to the existence of (t, m, s)-nets for $m \geq t+2$ (nets exist trivially for $m-t = 0, 1$). Consider e.g. the following simple argument: if there exists a $(0, m, s)$-net in base b for some $m \geq 2$, then there exists a $(0, 2, s)$-net in base b, hence there are $s-2$ MOLS of order b, and so we must have $s \leq b+1$. A more general argument of this type, combined with bounds for the appropriate combinatorial designs, leads to upper bounds on s in terms of b, m, and t, under the assumption that there exists a (t, m, s)-net in base b with $m \geq t+2$. A description of this method, together with tables of bounds, can be found in [2]. More recently, this approach was further refined by Martin and Stinson [7], [8] and improved bounds were obtained. In view of Lemma 1, combinatorial obstructions to the existence of (t, m, s)-nets yield combinatorial obstructions to the existence of (t, s)-sequences, such as the following bound from [21].

Theorem 1. *Given $b \geq 2$ and $s \geq 1$, a (t, s)-sequence in base b can exist only if*

$$t \geq \frac{s}{b} - \log_b \frac{(b-1)s + b + 1}{2}.$$

3. Constructions of nets and (t, s)-sequences

The number of known construction methods for nets and (t, s)-sequences is already quite large and ideas from various areas are employed. The combinatorial approach to the construction of nets uses the equivalences between (t, m, s)-nets and suitable combinatorial designs mentioned in Section 2 and techniques of constructing such combinatorial designs. Surveys of combinatorial methods for the construction of nets are given in [2], [9]. Other important methods for the construction of nets are based on coding theory. This approach goes back to an observation in [12, Section 7] that there is a connection between the digital method over a finite field $\mathbf{F}_q$ and the construction of parity-check matrices for good linear codes over $\mathbf{F}_q$. This connection is conveniently formalized through the notion of a (d, m, s)-*system over* $\mathbf{F}_q$ introduced in [32], which is a system $\left\{\mathbf{a}_j^{(i)} \in \mathbf{F}_q^m : 1 \leq i \leq s, 1 \leq j \leq m\right\}$ of vectors such that for any integers $d_1, \ldots, d_s \geq 0$ with $\sum_{i=1}^s d_i = d$ the $\mathbf{a}_j^{(i)}, 1 \leq j \leq d_i, 1 \leq i \leq s$, are linearly independent over $\mathbf{F}_q$. Finding a digital (t, m, s)-

net constructed over $\mathbf{F}_q$ is then equivalent to finding a (d, m, s)-system over $\mathbf{F}_q$ with $d = m - t$. The surveys [2], [9] report on coding-theory methods for the construction of nets and new methods of this type can be found in [32].

Standard constructions of digital (t, s)-sequences in base b are due to Sobol' [38] for $b = 2$ and any s, to Faure [3] for prime bases $b \geq s$, and to Niederreiter [13] for any b and any s. Generalizations of these sequences are described in Tezuka [39, Chapter 6]. As a by-product, these constructions yield digital $(t, m, s+1)$-nets in base b.

An important recent development is the use of algebraic curves over finite fields (or, equivalently, of global function fields) for the construction of (t, s)-sequences. The basic idea goes back to Niederreiter [16], [17]. At present, four different construction principles using algebraic curves are available and they all rely on the digital method over $\mathbf{F}_q$. We refer to [18], [20], [21], [44] for the detailed description of these constructions and to [25], [32] for further discussions. Three of the methods, and indeed the most effective ones, are based on algebraic curves over $\mathbf{F}_q$ with many $\mathbf{F}_q$-rational points (or, equivalently, on global function fields with many rational places). Given q and a dimension $s \geq 1$, the typical procedure is to choose a smooth, projective, absolutely irreducible algebraic curve $\mathcal{C}$ over $\mathbf{F}_q$ containing at least $s + 1$ $\mathbf{F}_q$-rational points, say $P_\infty, P_1, \ldots, P_s$. The point $P_i, 1 \leq i \leq s$, is used to produce the data that are needed in the digital method (i.e., certain elements of $\mathbf{F}_q$) for generating the ith coordinates of the points of the (t, s)-sequence. These elements of $\mathbf{F}_q$ are obtained by expansions on the curve $\mathcal{C}$ in local coordinates at P_∞. The methods in [20] and [44] yield digital (t, s)-sequences constructed over $\mathbf{F}_q$ with t being the genus of $\mathcal{C}$. If we optimize these constructions, we arrive in a natural way at the following important quantity from algebraic geometry over $\mathbf{F}_q$ and at the subsequent theorem in [20].

DEFINITION 3. For given $g \geq 0$ and q, let $N_q(g)$ be the maximum number of $\mathbf{F}_q$-rational points that a smooth, projective, absolutely irreducible algebraic curve over $\mathbf{F}_q$ of genus g can have.

THEOREM 2. *For every q and s there exists a digital $(V_q(s), s)$-sequence constructed over $\mathbf{F}_q$, where $V_q(s)$ is the least value of g such that $N_q(g) \geq s + 1$.*

The behavior of $V_q(s)$ as $s \to \infty$ can be obtained from class field towers and the asymptotic theory of $N_q(g)$ (see Section 5). As stated in Section 1, we can also pass from prime-power bases q to arbitrary bases b in the digital method. Finally, this leads to the following bound (see [20, Section 5]), which in view of Theorem 1 is best possible as far as the order of magnitude in s is concerned.

THEOREM 3. *For every $b \geq 2$ and $s \geq 1$ there exists a digital (t, s)-sequence in base b with*

$$t \leq \frac{c}{\log q_1} s + 1,$$

where $c > 0$ is an absolute constant and q_1 is the least prime power in the factorization of b into pairwise coprime prime powers.

4. ALGEBRAIC CURVES WITH MANY RATIONAL POINTS

The constructions of (t, s)-sequences in Section 3 based on algebraic curves over $\mathbf{F}_q$ lead to the requirement of finding good lower bounds for the number $N_q(g)$ in Definition 3, or in other words to the problem of constructing algebraic curves

over $\mathbf{F}_q$ of given genus g with many $\mathbf{F}_q$-rational points. This problem is also of great importance in the theory of algebraic-geometry codes (see Section 5). Recent surveys of this problem, also in the equivalent language of global function fields, are given in Garcia and Stichtenoth [4], Niederreiter and Xing [26], [30], and van der Geer and van der Vlugt [42].

A well-known technique for establishing the existence of various algebraic curves over $\mathbf{F}_q$ with many $\mathbf{F}_q$-rational points is due to Serre [37] and uses methods of class field theory. This approach was continued by Auer [1] and Lauter [5]. Usually, the curves obtained by this technique are not in an explicit form. On the other hand, constructions in the function field setting that work with Artin-Schreier and Kummer extensions and with subfields of cyclotomic function fields yield explicit generators and defining equations. Such constructions can be found e.g. in [19], [21], [26], [46] for $q = 2$, in [22], [27] for $q = 3$, in [22], [23] for $q = 4$, in [22], [24], [35] for $q = 5$, in [29] for $q = 8, 16$, and in [32] for $q = 9, 27$. Explicit constructions inspired by techniques from coding theory were introduced by van der Geer and van der Vlugt [41] (see also the survey [42]).

In the function field setting, a powerful technique of obtaining global function fields with many rational places is based on Hilbert class fields. The aim is to construct unramified abelian extensions of a given global function field F in which certain selected rational places of F split completely. This method works particularly well if the divisor class number of F is large relative to the genus of F. Applications of this method can be found in [22], [24], [26], [27], [29], [30], [35], [46]. A more general approach, which contains both cyclotomic function fields and Hilbert class fields as special cases, uses the theory of narrow ray class extensions obtained from Drinfeld modules of rank 1 and was introduced in [45]. This method allows great flexibility and produces a large number of families of global function fields with many rational places. We refer to [23], [24], [26], [27], [29], [30], [31], [35], [46] for further results and examples with this method.

Table 1 contains all bounds for $N_q(g)$ available to the author for $q = 2, 3, 4, 5, 8, 9, 16, 27$ and $1 \le g \le 50$ (for $g = 0$ we trivially have $N_q(0) = q + 1$). In each entry of the table, the first number is a lower bound for $N_q(g)$ and the second an upper bound for $N_q(g)$. If only one number is given, then this is the exact value of $N_q(g)$. A program for calculating upper bounds for $N_q(g)$, which is based on Weil's explicit formula for the number of $\mathbf{F}_q$-rational points in terms of the zeta function and on the trigonometric polynomials of Oesterlé, was kindly supplied by Jean-Pierre Serre. The lower bounds in Table 1 are obtained by combining [32, Table 3] with new data in [1], [35]. We refer also to the tables of van der Geer and van der Vlugt [43] which represent the most recent result of an ongoing project to update bounds for $N_q(g)$ periodically.

5. Applications to coding theory

There is an asymptotic theory of $N_q(g)$ which has significant applications to algebraic coding theory. The basic quantity here is

$$A(q) = \limsup_{g\to\infty} \frac{N_q(g)}{g}.$$

For values of q for which $A(q)$ is larger than a known comparison function, Goppa's

construction of algebraic-geometry codes leads to improvements on the classical Gilbert-Varshamov bound for the existence of good linear codes over $\mathbf{F}_q$.

Let U_q be the set of ordered pairs $(\delta, R) \in [0,1]^2$ for which there exists a sequence of linear codes over $\mathbf{F}_q$ of increasing lengths such that δ is the limit of the relative minimum distances and R the limit of the information rates. It is known that for some continuous function α_q on $[0,1]$ we have

$$U_q = \{(\delta, R) : 0 \le R \le \alpha_q(\delta), 0 \le \delta \le 1\},$$

where $\alpha_q(0) = 1$ and $\alpha_q(\delta) = 0$ for $\delta \in [(q-1)/q, 1]$. The function α_q is unknown, and it is an important issue in algebraic coding theory to obtain good lower bounds for α_q on the interval $(0, (q-1)/q)$. The Gilbert-Varshamov bound says that

$$\alpha_q(\delta) \ge R_{GV}(q, \delta) := 1 - H_q(\delta) \quad \text{for } 0 < \delta < (q-1)/q,$$

where H_q is the q-ary entropy function. Algebraic-geometry codes lead to the bound

$$\alpha_q(\delta) \ge R_{AG}(q, \delta) := 1 - \frac{1}{A(q)} - \delta \quad \text{for } 0 \le \delta \le 1.$$

By showing that $A(q) \ge q^{1/2} - 1$ if q is a square, Tsfasman, Vlădut, and Zink [40] proved that $R_{AG}(q, \delta) > R_{GV}(q, \delta)$ if q is a sufficiently large square and δ belongs to a suitable subinterval of $[0,1]$.

For nonsquares q only weaker lower bounds for $A(q)$ are known. Serre [37] showed that $A(q)$ is at least of the order of magnitude $\log q$, and an alternative proof and an effective version of this result were recently given in [33]. In many cases the following result in [28] yields a considerable improvement: if $q = p^e$ with a prime p and an odd integer $e \ge 3$, then $A(q)$ is at least of the order of magnitude $q^{1/(2k)}$, where k is the least prime factor of e. Further discussions and refinements of this result can be found in [30], [33]. As a consequence we get the following theorem in [28] which improves on the Gilbert-Varshamov bound for sufficiently large composite nonsquares q.

THEOREM 4. *Let $m \ge 3$ be an odd integer and let r be a prime power with $r \ge 100m^3$ for odd r and $r \ge 576m^3$ for even r. Then there exists an open interval $(\delta_1, \delta_2) \subseteq (0,1)$ containing $(r^m - 1)/(2r^m - 1)$ such that*

$$R_{AG}(r^m, \delta) > R_{GV}(r^m, \delta) \quad \textit{for all } \delta \in (\delta_1, \delta_2).$$

In connection with lower bounds for $A(q)$ we mention that there is a method of Perret [36] for obtaining such lower bounds which depends, however, on a conjecture that would provide a sufficient condition for the infinitude of certain ramified class field towers. It was recently shown in [34] by a counterexample that this conjecture is wrong. Therefore, the lower bounds for $A(q)$ in Perret [36, Section III] remain unproved.

Table 1: Bounds for $N_q(g)$

$g\backslash q$	2	3	4	5	8	9	16	27
1	5	7	9	10	14	16	25	38
2	6	8	10	12	18	20	33	48
3	7	10	14	16	24	28	38	58
4	8	12	15	18	25-29	30	45-47	64-68
5	9	12-14	17-18	20-22	29-32	32-36	49-55	55-78
6	10	14-15	20	21-25	33-36	35-40	65	76-88
7	10	16-17	21-22	22-27	33-39	39-43	63-70	64-98
8	11	15-18	21-24	22-29	34-43	38-47	61-76	92-108
9	12	19	26	26-32	45-47	40-51	72-81	82-118
10	13	19-21	27-28	27-34	38-50	54-55	81-87	91-128
11	14	20-22	26-30	32-36	48-54	55-59	80-92	96-138
12	14-15	22-24	29-31	30-38	49-57	55-63	68-97	109-148
13	15	24-25	33	36-40	50-61	60-66	97-103	136-156
14	15-16	24-26	32-35	39-43	65	56-70	97-108	84-164
15	17	28	33-37	35-45	54-68	64-74	98-113	136-171
16	17-18	27-29	36-38	40-47	56-71	74-78	93-118	136-178
17	17-18	24-30	40	42-49	61-74	56-82	96-124	128-185
18	18-19	26-31	41-42	32-51	65-77	46-85	113-129	94-192
19	20	27-32	37-43	45-54	58-80	84-88	121-134	126-199
20	19-21	30-34	37-45	30-56	68-83	48-91	121-140	133-207
21	21	32-35	41-47	50-58	72-86	82-95	129-145	163-214
22	21-22	28-36	40-48	51-60	66-89	78-98	129-150	112-221
23	22-23	26-37	41-50	55-62	68-92	92-101	126-155	114-228
24	20-23	28-38	42-52	46-64	66-95	91-104	129-161	166-235
25	24	36-40	51-53	52-66	66-97	64-108	144-166	196-242
26	24-25	36-41	55	45-68	72-100	110-111	150-171	108-249
27	22-25	39-42	49-56	52-70	96-103	60-114	145-176	114-256
28	25-26	37-43	51-58	54-71	97-106	105-117	136-181	108-263
29	25-27	42-44	49-60	56-73	97-109	104-120	161-187	114-270
30	25-27	34-46	53-61	58-75	80-112	60-123	161-192	117-277
31	27-28	40-47	60-63	72-77	72-115	84-127	150-197	114-284
32	26-29	38-48	57-65	62-79	72-118	81-130	132-202	126-291
33	28-29	37-49	65-66	64-81	92-121	78-133	193-207	220-298
34	27-30	44-50	57-68	76-83	80-124	111-136	156-213	135-305
35	29-31	47-51	58-69	68-85	106-127	84-139	144-218	126-312
36	30-31	46-52	64-71	64-87	105-130	110-142	185-223	244-319
37	29-32	48-54	66-72	72-89	121-132	120-145	208-228	162-326
38	28-33	36-55	56-74	78-91	129-135	105-149	193-233	144-333
39	33	46-56	65-75	76-93	117-138	84-152	160-239	271-340
40	32-34	54-57	75-77	65-94	100-141	90-155	162-244	244-346

$g\backslash q$	2	3	4	5	8	9	16	27
41	33-35	50-58	65-78	80-96	112-144	84-158	216-249	153-353
42	33-35	39-59	66-80	60-98	129-147	90-161	209-254	280-360
43	33-36	55-60	72-81	84-100	100-150	120-164	226-259	196-367
44	33-37	42-61	68-83	60-102	129-153	90-167	162-264	153-374
45	32-37	48-62	80-84	88-104	144-156	112-170	242-268	171-381
46	34-38	55-63	81-86	75-106	129-158	138-173	243-273	162-388
47	36-38	47-65	73-87	92-108	120-161	154-177	176-277	174-395
48	34-39	55-66	77-89	82-110	126-164	163-180	184-282	325-402
49	36-40	63-67	81-90	96-111	130-167	168-183	192-286	268-409
50	40	56-68	91-92	70-113	130-170	182-186	225-291	180-416

REFERENCES

[1] R. Auer, Ray class fields of global function fields with many rational places, preprint, 1998.

[2] A.T. Clayman, K.M. Lawrence, G.L. Mullen, H. Niederreiter, and N.J.A. Sloane, Updated tables of parameters of (t, m, s)-nets, *J. Combinatorial Designs*, to appear.

[3] H. Faure, Discrépance de suites associées à un système de numération (en dimension s), *Acta Arith.* 41, 337–351 (1982).

[4] A. Garcia and H. Stichtenoth, Algebraic function fields over finite fields with many rational places, *IEEE Trans. Inform. Theory* 41, 1548–1563 (1995).

[5] K. Lauter, Ray class field constructions of curves over finite fields with many rational points, *Algorithmic Number Theory* (H. Cohen, ed.), Lecture Notes in Computer Science, Vol. 1122, pp. 187–195, Springer, Berlin, 1996.

[6] K.M. Lawrence, A combinatorial characterization of (t, m, s)-nets in base b, *J. Combinatorial Designs* 4, 275–293 (1996).

[7] W.J. Martin and D.R. Stinson, A generalized Rao bound for ordered orthogonal arrays and (t, m, s)-nets, preprint, 1997.

[8] —, —, Association schemes for ordered orthogonal arrays and (t, m, s)-nets, preprint, 1997.

[9] G.L. Mullen, A. Mahalanabis, and H. Niederreiter, Tables of (t, m, s)-net and (t, s)-sequence parameters, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* (H. Niederreiter and P.J.-S. Shiue, eds.), Lecture Notes in Statistics, Vol. 106, pp. 58–86, Springer, New York, 1995.

[10] G.L. Mullen and W.C. Schmid, An equivalence between (t, m, s)-nets and strongly orthogonal hypercubes, *J. Combinatorial Theory Ser. A* 76, 164–174 (1996).

[11] G.L. Mullen and G. Whittle, Point sets with uniformity properties and orthogonal hypercubes, *Monatsh. Math.* 113, 265–273 (1992).

[12] H. Niederreiter, Point sets and sequences with small discrepancy, *Monatsh. Math.* 104, 273–337 (1987).

[13] —, Low-discrepancy and low-dispersion sequences, *J. Number Theory* 30, 51–70 (1988).

[14] —, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.
[15] —, Orthogonal arrays and other combinatorial aspects in the theory of uniform point distributions in unit cubes, *Discrete Math.* 106/107, 361–367 (1992).
[16] —, Pseudorandom numbers and quasirandom points, *Z. angew. Math. Mech.* 73, T648–T652 (1993).
[17] —, Factorization of polynomials and some linear-algebra problems over finite fields, *Linear Algebra Appl.* 192, 301–328 (1993).
[18] H. Niederreiter and C.P. Xing, Low-discrepancy sequences obtained from algebraic function fields over finite fields, *Acta Arith.* 72, 281–298 (1995).
[19] —, —, Explicit global function fields over the binary field with many rational places, *Acta Arith.* 75, 383–396 (1996).
[20] —, —, Low-discrepancy sequences and global function fields with many rational places, *Finite Fields Appl.* 2, 241–273 (1996).
[21] —, —, Quasirandom points and global function fields, *Finite Fields and Applications* (S. Cohen and H. Niederreiter, eds.), London Math. Soc. Lecture Note Series, Vol. 233, pp. 269–296, Cambridge Univ. Press, Cambridge, 1996.
[22] —, —, Cyclotomic function fields, Hilbert class fields, and global function fields with many rational places, *Acta Arith.* 79, 59–76 (1997).
[23] —, —, Drinfeld modules of rank 1 and algebraic curves with many rational points. II, *Acta Arith.* 81, 81–100 (1997).
[24] —, —, Global function fields with many rational places over the quinary field, *Demonstratio Math.* 30, 919–930 (1997).
[25] —, —, The algebraic-geometry approach to low-discrepancy sequences, *Monte Carlo and Quasi-Monte Carlo Methods 1996* (H. Niederreiter *et al.*, eds.), Lecture Notes in Statistics, Vol. 127, pp. 139–160, Springer, New York, 1998.
[26] —, —, Algebraic curves over finite fields with many rational points, *Number Theory* (K. Győry *et al.*, eds.), pp. 423–443, de Gruyter, Berlin, 1998.
[27] —, —, Global function fields with many rational places over the ternary field, *Acta Arith.* 83, 65–86 (1998).
[28] —, —, Towers of global function fields with asymptotically many rational places and an improvement on the Gilbert-Varshamov bound, *Math. Nachr.*, to appear.
[29] —, —, Algebraic curves with many rational points over finite fields of characteristic 2, *Proc. Number Theory Conf.* (Zakopane, 1997), de Gruyter, Berlin, to appear.
[30] —, —, Global function fields with many rational places and their applications, *Proc. Finite Fields Conf.* (Waterloo, 1997), Contemporary Math., American Math. Soc., Providence, to appear.
[31] —, —, A general method of constructing global function fields with many rational places, *Algorithmic Number Theory* (J.P. Buhler, ed.), Lecture Notes in Computer Science, Vol. 1423, pp. 555–566, Springer, Berlin, 1998.
[32] —, —, Nets, (t, s)-sequences, and algebraic geometry, *Pseudo- and Quasi-Random Point Sets* (P. Hellekalek and G. Larcher, eds.), Lecture Notes in Statistics, Springer, New York, to appear.

[33] —, —, Curve sequences with asymptotically many rational points, preprint, 1997.
[34] —, —, A counterexample to Perret's conjecture on infinite class field towers for global function fields, preprint, 1998.
[35] —, —, Global function fields with many rational places over the quinary field. II, *Acta Arith.*, to appear.
[36] M. Perret, Tours ramifiées infinies de corps de classes, *J. Number Theory* 38, 300–322 (1991).
[37] J.-P. Serre, Sur le nombre des points rationnels d'une courbe algébrique sur un corps fini, *C.R. Acad. Sci. Paris Sér. I Math.* 296, 397–402 (1983).
[38] I.M. Sobol', The distribution of points in a cube and the approximate evaluation of integrals (Russian), *Zh. Vychisl. Mat. i Mat. Fiz.* 7, 784–802 (1967).
[39] S. Tezuka, *Uniform Random Numbers: Theory and Practice*, Kluwer, Boston, 1995.
[40] M.A. Tsfasman, S.G. Vlădut, and T. Zink, Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound, *Math. Nachr.* 109, 21–28 (1982).
[41] G. van der Geer and M. van der Vlugt, Curves over finite fields of characteristic 2 with many rational points, *C.R. Acad. Sci. Paris Sér. I Math.* 317, 593–597 (1993).
[42] —, —, How to construct curves over finite fields with many points, *Arithmetic Geometry* (F. Catanese, ed.), pp. 169–189, Cambridge Univ. Press, Cambridge, 1997.
[43] —, —, Tables of curves with many points, preprint, 1998.
[44] C.P. Xing and H. Niederreiter, A construction of low-discrepancy sequences using global function fields, *Acta Arith.* 73, 87–102 (1995).
[45] —, —, Modules de Drinfeld et courbes algébriques ayant beaucoup de points rationnels, *C.R. Acad. Sci. Paris Sér. I Math.* 322, 651–654 (1996).
[46] —, —, Drinfeld modules of rank 1 and algebraic curves with many rational points, *Monatsh. Math.*, to appear.

Harald Niederreiter
Institute of Information Processing
Austrian Academy of Sciences
Sonnenfelsgasse 19
A-1010 Vienna
Austria
E-mail: niederreiter@oeaw.ac.at

The Sphere Packing Problem

N. J. A. Sloane

Abstract. A brief report on recent work on the sphere-packing problem.

1991 Mathematics Subject Classification: 52C17
Keywords and Phrases: Sphere packings; lattices; quadratic forms; geometry of numbers

1 Introduction

The sphere packing problem has its roots in geometry and number theory (it is part of Hilbert's 18th problem), but is also a fundamental question in information theory. The connection is via the sampling theorem. As Shannon observes in his classic 1948 paper [37] (which ushered in the age of digital communication), if f is a signal of bandwidth W hertz, with almost all its energy concentrated in an interval of T secs, then f is accurately represented by a vector of $2WT$ samples, which may be regarded as the coordinates of a single point in $\mathbb{R}^n$, $n = 2WT$. Nearly equal signals are represented by neighboring points, so to keep the signals distinct, Shannon represents them by n-dimensional 'billiard balls', and is therefore led to ask: what is the best way to pack 'billiard balls' in n dimensions?

This talk will report on a few selected developments that have taken place since the appearance of Rogers' 1964 book on the subject, proceeding upwards in dimension from 2 to 128. The reader is referred to [16] (especially the third edition, which has 800 references covering 1988-1998) for further information, definitions and references. See also the lattice data-base [31].

2 Dimension 2

The best packing in dimension 2 is the familiar 'hexagonal lattice' packing of circles, each touching six others. The centers are the points of the root lattice A_2. The *density* Δ of this packing is the fraction of the plane occupied by the spheres: $\pi/\sqrt{12} = 0.9069\ldots$.

In general we wish to find Δ_n, the highest possible density of a packing of equal nonoverlapping spheres in $\mathbb{R}^n$, or $\Delta_n^{(L)}$, the highest density of any packing in which the centers form a lattice. It is known (Fejes Tóth, 1940) that $\Delta_2 = \Delta_2^{(L)} = \pi/\sqrt{12}$. An n-dimensional lattice Λ of determinant d and minimal nonzero squared length (or *norm*) μ has packing radius $\rho = \sqrt{\mu}/2$ and density $\Delta = V_n\rho^n/\sqrt{\det\Lambda}$, where

$V_n = \pi^{n/2}/(n/2)!$ is the volume of a unit sphere. The *center density* of a packing is $\delta = \Delta/V_n$.

We are also interested in packing points on a sphere, and especially in the 'kissing number problem': find τ_n (resp. $\tau_n^{(L)}$), the maximal number of spheres that can touch an equal sphere in $\mathbb{R}^n$ (resp. in any lattice in $\mathbb{R}^n$). It is trivial that $\tau_2 = \tau_2^{(L)} = 6$.

3 Dimension 3

In spite of much recent work ([20], [21]) Δ_3 is still unknown; nor is Δ_n known in any dimension above 2. It is conjectured that $\Delta_3 = \pi/\sqrt{18} = 0.74048\ldots$, as in the face-centered cubic (f.c.c.) lattice A_3. Muder [28] has shown that $\Delta_3 \le 0.773055\ldots$. It is worth mentioning, however, that there are packings of congruent ellipsoids with density considerably greater than $\pi/\sqrt{18}$ [3].

In two dimensions the hexagonal lattice is (a) the densest lattice packing, (b) the least dense lattice covering, and (c) is geometrically similar to its dual lattice. There is a little-known three-dimensional lattice that is similar to its dual, and, among all lattices with this property, is both the densest packing and the least dense covering. This is the m.c.c. (or *mean-centered cuboidal*) lattice [11] with Gram matrix

$$\frac{1}{2}\begin{bmatrix} 1+\sqrt{2} & 1 & 1 \\ 1 & 1+\sqrt{2} & 1-\sqrt{2} \\ 1 & 1-\sqrt{2} & 1+\sqrt{2} \end{bmatrix} .$$

In a sense this lattice is the geometric mean of the f.c.c. lattice and its dual the body-centered cubic (b.c.c.) lattice. Consider the lattice generated by the vectors $(\pm u, \pm v, 0)$ and $(0, \pm u, \pm v)$ for real numbers u and v. If the ratio u/v is respectively 1, $2^{1/2}$ or $2^{1/4}$ we obtain the f.c.c., b.c.c. and m.c.c. lattices. The m.c.c. lattice also recently arose in a different context, as the lattice corresponding to the period matrix of the hyperelliptic Riemann surface $w^2 = z^8 - 1$

4 Dimensions 4–8

Table 1 summarizes what is presently known about the sphere packing and kissing number problems in dimensions ≤ 24. Entries enclosed inside a solid line are known to be optimal, those inside a dashed line optimal among lattices.

The large box in the 'density' column refers to Blichfeldt's 1935 result that the root lattices $\mathbb{Z} \simeq A_1$, A_2, $A_3 \simeq D_3$, D_4, D_5, E_6, E_7, E_8 achieve $\Delta_n^{(L)}$ for $n \le 8$. It is remarkable that more than 60 years later $\Delta_9^{(L)}$ is still unknown.

The large box in the right-hand column refers to Watson's 1963 result that the kissing numbers of the above lattices, together with that of the laminated lattice Λ_9, achieve $\tau_n^{(L)}$ for $n \le 9$. Odlyzko and I [16, Ch. 13] and independently Levenshtein determined τ_8 and τ_{24}. The packings achieving these two bounds are unique [16, Ch. 14].

Dim.	Densest packing	Highest kissing number	
1	$\mathbb{Z} \simeq \Lambda_1$	2	
2	$A_2 \simeq \Lambda_2$	6	
3	$A_3 \simeq D_3 \simeq \Lambda_3$	12	
4	$D_4 \simeq \Lambda_4$	24	
5	$D_5 \simeq \Lambda_5$	40	
6	$E_6 \simeq \Lambda_6$	72	
7	$E_7 \simeq \Lambda_7$	126	
8	$E_8 \simeq \Lambda_8$	240	
9	Λ_9	272	(306 from P_{9a})
10	Λ_{10} (P_{10c})	336	(500 from P_{10b})
12	K_{12}	756	(840 from P_{12a})
16	$BW_{16} \simeq \Lambda_{16}$	4320	
24	Leech $\simeq \Lambda_{24}$	196560	

Table 1: Densest packings and highest kissing numbers known in low dimensions. (Parenthesized entries are nonlattice arrangements that are better than any known lattice.)

THE 'LOW DIMENSIONAL LATTICES' PROJECT Some years ago Conway and I noticed that there were several places in the literature where the results could be simplified if they were described in terms of lattices rather than quadratic forms. (It seems clearer to say 'the lattice E_8' rather than 'the quadratic form $2x_1^2+2x_2^2+4x_3^2+4x_4^2+20x_5^2+12x_6^2+4x_7^2+2x_8^2+2x_1x_2+2x_2x_3+6x_3x_4+10x_4x_5+6x_5x_6+2x_6x_7+2x_7x_8$'.) This led to a series of papers [7], [10], [13].

Integral lattices of determinant $d = 1$ ('unimodular' lattices) have been classified in dimensions ≤ 25, dimensions 24, 25 being due to Borcherds. In [16, Ch. 15] and [7, (I)] we extended this to $d \leq 25$ for various ranges of dimension.

[7, (II)] is based on the work of Dade, Plesken, Pohst and others, and describes the lattices associated with the maximal irreducible subgroups of $GL(n, \mathbb{Z})$ for $n = 1, \ldots, 9, 11, 13, 17, 19, 23$. Nebe, and Nebe and Plesken (see [29], [32]) have recently completed the enumeration of the maximal finite irreducible subgroups of $GL(n, \mathbb{Q})$ for $n \leq 31$, together with the associated lattices.

[7, (IV)] gives an improved version of the mass formula for lattices, and [7, (V)] studies when an n-dimensional integral lattice can be represented as a sublattice of $\mathbb{Z}^m$ for some $m \geq n$, or failing that, by a sublattice of $s^{-1/2}\mathbb{Z}^m$ for some integer s. [10] describes the Voronoi and Delaunay cells of all the root lattices and their duals, and [7, (VI), (VIII)] discusses how the Voronoi cell of a 3- or 4-dimensional lattice changes as the lattice is continuously varied.

[7, (VII)] determines the 'coordination sequences' of various lattices. Consider E_8, for example, and let $S(k)$ denote the number of lattice points that are k steps from the origin, where a step is a move to an adjacent sphere ($S(1)$ is the kissing

number). Then $\sum_{k=0}^{\infty} S(k)x^k = f(x)/(1-x)^8$, where $f(x) = 1 + 232x + 7228x^2 + \ldots + x^8$. Thus the coordination sequence for E_8 begins 1, 240, 9120, ... For other examples see [39]

Perfect lattices One possible approach to the determination of the densest lattices in dimensions 7 to 9 is via Voronoi's theorem that the density of Λ is a local maximum if and only if Λ is perfect and eutactic [27].

In 1975 Stacey, extending the work of several earlier authors, published a list of 33 perfect lattices in dimension 7. Unfortunately one of the 33 was omitted from her papers and her dissertation. In [7, (III)] we reconstructed the missing lattice and 'beautified' all 33, computing their automorphism groups, etc. In 1991 Jaquet-Chiffelle [22] completed this work by showing that this is indeed the full list of perfect lattices in $\mathbb{R}^7$. This provides another proof that E_7 is the densest lattice in dimension 7.

Martinet, Bergé and their students are presently attempting to classify the eight-dimensional perfect lattices, and it appears that there will be roughly 10000 of them. Whether this approach can be used to determine $\Delta_9^{(L)}$ remains to be seen!

5 Dimension 9. Laminated lattices

There is a simple construction, the 'laminating' or 'greedy' construction, that produces many of the densest lattices in dimensions up to 26. Let Λ_1 denote the even integers in $\mathbb{R}^1$, and define the n-dimensional laminated lattices Λ_n recursively by: consider all lattices of minimal norm 4 that contain some Λ_{n-1} as a sublattice, and select those of greatest density. It had been known since the 1940's that this produces the densest lattices known for $n \leq 10$. In [6] we determined *all* inequivalent laminated lattices for $n \leq 25$, and found the density of Λ_n for $n \leq 48$ (Fig. 1). A key result needed for this was the determination of the covering radius of the Leech lattice and the enumeration of the deep holes in that lattice [16, Ch. 23].

What are all the best sphere packings in low dimensions? In [13] we describe what may be *all* the best packings in dimensions $n \leq 10$, where 'best' means both having the highest density and not permitting any local improvement. In particular, we conjecture that $\Delta_n^{(L)} = \Delta_n$ for $n \leq 9$. For example, it appears that the best five-dimensional sphere packings are parameterized by the 4-colorings of $\mathbb{Z}$. We also find what we believe to be the exact numbers of 'uniform' packings among these, those in which the automorphism group acts transitively. These assertions depend on certain plausible but as yet unproved postulates.

A remarkable property of 9-dimensional packings. We also show in [13] that the laminated lattice Λ_9 has the following astonishing property. Half the spheres can be moved bodily through arbitrarily large distances without overlapping the other half, only touching them at isolated instants, the density remaining

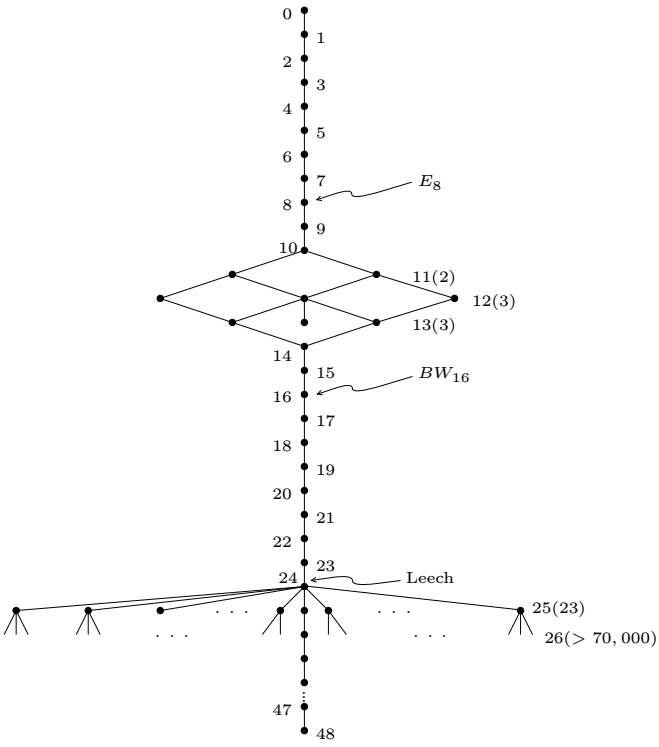


Figure 1: Inclusions among laminated lattices Λ_n.

the same at every instant. A typical packing in this family consists of the points of $D_9^{\theta+} = D_9 \cup D_9 + ((1/2)^8, \theta/2)$, for θ real. D_9^{0+} is Λ_9 and D_9^{1+} is D_9^+, the 9-dimensional diamond structure. All these packings have the same density, which we conjecture is the value of $\Delta_9 = \Delta_9^{(L)}$. Another result in [13] is that there are extraordinarily many 16-dimensional packings that are just as dense as the Barnes-Wall lattice $BW_{16} \simeq \Lambda_{16}$.

6 Dimension 10. Construction A.

In dimension 10 we encounter for the first time a nonlattice packing that is denser than all known lattices. This packing, and the nonlattice packing with the highest known kissing number in dimension 9, are easily obtained from 'Construction A' (cf. [24]). If $\mathcal{C}$ is a binary code of length n, the corresponding packing is $P(\mathcal{C}) = \{x \in \mathbb{Z}^n : x \pmod 2 \in \mathcal{C}\}$.

Consider the vectors $abcde \in (\mathbb{Z}/4\mathbb{Z})^5$ where $b, c, d \in \{+1, -1\}$, $a = c - d$, $e = b + c$, together with all their cyclic shifts, and apply the 'Gray map' $0 \to 00,\ 1 \to 01,\ 2 \to 11,\ 3 \to 10$ to obtain a binary code $\mathcal{C}_{10}$ containing 40 vectors of length 10 and minimal distance 4. This is our description [12] of a code first

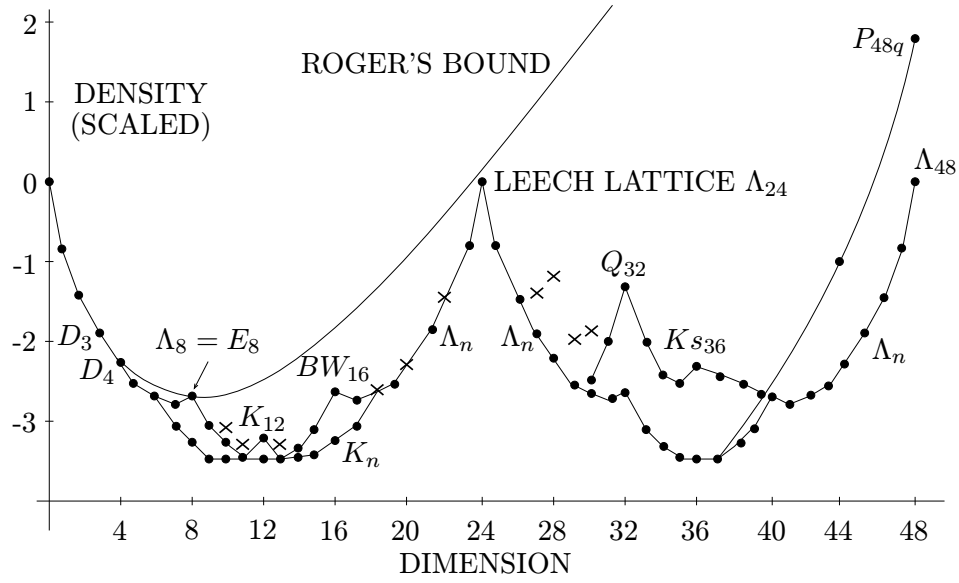


Figure 2: Densest sphere packings known in dimensions $n \le 48$.

discovered by Best. The code is unique [25]. Then $P(\mathcal{C}_{10}) = P_{10c}$ is the record 10-dimensional packing.

Figure 2 shows the density of the best packings known up to dimension 48, rescaled to make them easier to read. The vertical axis gives $\log_2 \delta + n(24-n)/96$. The figure also shows the upper bounds of Muder (for $n = 3$) and Rogers ($n \ge 4$). Lattice packings are indicated by small circles, nonlattices by crosses (however, the locations of the lattices are only approximate). The figure is dominated by the two arcs of the graph of the laminated lattices Λ_n, which touch the zero ordinate at $n = 0$, 24 (the Leech lattice) and 48. K_{12} is the Coxeter-Todd lattice.

7 Dimensions 18–22

Record nonlattice packings in dimensions 18, 20 and 22 have recently been given in [4], [14], [40]. Vardy's construction [40], 'Construction B^*', also uses binary codes. Let $\mathcal{B}$ and $\mathcal{C}$ be codes of length n such that $c\cdot(\mathbf{1}+b) = 0$ for all $b \in \mathcal{B}$, $c \in \mathcal{C}$, and set $P^*(\mathcal{B},\mathcal{C}) = \{\mathbf{0}+2b+4x, \mathbf{1}+2c+4y \,:\, b \in \mathcal{B}, c \in \mathcal{C}, x, y \in \mathbb{Z}^n, \sum x_i \text{ even}, \sum y_i \text{ odd}\}$. For example, by taking $\mathcal{B}$ to be the quadratic residue code of length 18 and $\mathcal{C}$ to be its dual, Bierbrauer and Edel [4] obtain a new record packing in $\mathbb{R}^{18}$.

8 Dimension 24. The Leech lattice

The Leech lattice Λ_{24} is a remarkably dense packing in $\mathbb{R}^{24}$ (as can be seen from Fig. 2). Here are four constructions. (i) As a laminated lattice: start in dimension 1 with the lattice $\Lambda_1 = \mathbb{Z}$ and apply the greedy algorithm (see Fig. 1). (ii) Apply

Construction A to the Golay code of length 24 to obtain a lattice L_{24}. Then Λ_{24} is spanned by $(-3/2, 1/2, \ldots, 1/2)$ and $\{x \in L_{24} : \sum x_i \equiv 0 \pmod 4\}$. (iii) Hensel lift the Golay code to an extended cyclic (and self-dual) code over $\mathbb{Z}/4\mathbb{Z}$ and apply 'Construction A mod 4' [5]. (iv) There is a unique unimodular even lattice $\mathrm{II}_{25,1}$ in Lorentzian space $\mathbb{R}^{25,1}$, consisting of the points $(x_0 x_1 \cdots x_{24} | x_{25})$ with all $x_i \in \mathbb{Z}$ or all $x_i \in \mathbb{Z}+1/2$ and satisfying $x_0 + \cdots + x_{24} - x_{25} \in 2\mathbb{Z}$. Let $w = (0\ 1 \cdots 24|70)$, a vector of zero length. Then $(w^\perp$ in $\mathrm{II}_{25,1})/w$ is Λ_{24} [16, Ch. 26].

9 Dimensions 26–31

New packings in these dimensions have been discovered by Bacher, Borcherds, Conway, Vardy, Venkov — see [16] for details.

10 Dimension 32. Modular lattices

An *N-modular* lattice [34] is an integral lattice that is similar to its dual, under a similarity that multiplies norms by N. A unimodular lattice is 1-modular. The interest in this family arises because many of the densest known lattices are N-modular: $\mathbb{Z}$, A_2, D_4, E_8, K_{12}, BW_{16}, Λ_{24}, Q_{32}, P_{48q}, ...

Quebbemann's lattice Q_{32}, for example, is 2-modular, and can be constructed from a Reed-Solomon code of length 8 over $\mathbb{F}_9$ [33], [16, Ch. 8].

Shadow theory. The concept of the shadow of a lattice or code was introduced in [8], [9] (see also [15]) and has proved to be very useful ([9] has stimulated over 50 sequels in the coding literature).

Let Λ be an n-dimensional unimodular lattice. If Λ is even then the *shadow* $S(\Lambda) = \Lambda$, otherwise $S(\Lambda) = (\Lambda_0)^* \setminus \Lambda$, where the subscript 0 denotes even sublattice. The set $2S(\Lambda) = \{2s : s \in S(\Lambda)\}$ is precisely the set of *parity vectors* for Λ, i.e. the vectors $u \in \Lambda$ such that $u \cdot x \equiv x \cdot x \pmod 2$ for all $x \in \Lambda$. Such vectors have been studied by many authors from Braun (1940) onwards, but their application to obtaining bounds on lattices seems to have been overlooked.

If the theta series of Λ is $\Theta_\Lambda(z)$ then [8] the shadow has theta series

$$\left(\frac{e^{\pi i/4}}{\sqrt{z}}\right)^n \Theta_\Lambda\left(1 - \frac{1}{z}\right) . \tag{1}$$

One of the most satisfying properties of integral lattices is the classical theorem that (a) if Λ is a unimodular lattice then Θ_Λ belongs to the graded ring $\mathbb{C}[\Theta_\mathbb{Z}, \Theta_{E_8}]$, and (b) if Λ is even then Θ belongs to $\mathbb{C}[\Theta_{E_8}, \Theta_{\Lambda_{24}}]$.

To illustrate the use of the shadow, let us prove there is no 9-dimensional unimodular lattice of minimal norm 2. If so then from (a) $\Theta_\Lambda = -\Theta_\mathbb{Z}/8 + 9\Theta_{E_8}/8 = 1 + 252q^2 + 456q^3 + \cdots$, where $q = e^{\pi i z}$. But then (1) implies $\Theta_{S(\Lambda)} = \frac{9}{4}q^{1/4} + \frac{1913}{4}q^{9/4} + \cdots$, a contradiction since $\Theta_{S(\Lambda)}$ must have integer coefficients.

In [26] we used (a), (b) to show that the minimal norm μ of an n-dimensional odd unimodular lattice satisfies

$$\mu \le \left\lceil \frac{n}{8} \right\rceil + 1 , \tag{2}$$

and for an even unimodular lattice

$$\mu \le 2\left[\frac{n}{24}\right] + 2 \ . \tag{3}$$

In [36] we used shadow theory to strengthen (2) by showing that odd lattices satisfy

$$\mu \le 2\left[\frac{n}{24}\right] + 2 \ , \tag{4}$$

except that $\mu \le 3$ when $n = 23$. In view of the similarity between (3) and (4) we propose that a lattice satisfying either bound with equality be called *extremal* (the old definition of this term was based on (2) and (3)).

Quebbemann [35] has generalized (3) to certain families of even N-modular lattices, and analogous bounds for odd N-modular lattices (using an appropriate generalization of the shadow) were given in [36]. One can then define extremal N-modular lattices.

11 Higher dimensions

Space does not permit more than a mention of the following: Kschischang and Pasupathy's lattice Ks_{36} in $\mathbb{R}^{36}$ [23]; the three extremal unimodular lattices P_{48q}, P_{48p}, P_{48n} in $\mathbb{R}^{48}$, the latter being a recent discovery of Nebe [30]; Bachoc's extremal 2-modular lattice in $\mathbb{R}^{48}$ [1]; Nebe's extremal 3-modular lattice in $\mathbb{R}^{64}$ [30]; and Bachoc and Nebe's extremal unimodular lattice in $\mathbb{R}^{80}$ [2].

The existence of the following extremal lattices is an open question: 3-modular in $\mathbb{R}^{36}$ (determinant $d = 3^{18}$, minimal norm $\mu = 8$); 2-modular in $\mathbb{R}^{64}$ ($d = 2^{32}, \mu = 10$); unimodular in $\mathbb{R}^{72}$ ($d = 1$, $\mu = 8$).

From dimensions 80 to about 4096 the densest lattices known are the Mordell-Weil lattices discovered by Elkies [19], and Shioda [38]. But we know very little about this range, as evidenced by the recent construction of record kissing numbers in dimensions 32 to 128 [17] from binary codes. In dimension 128, for example, the Mordell-Weil lattice has kissing number 218044170240 [18], whereas in our construction (which admittedly is not a lattice) some spheres touch 8812505372416 others.

It would also be desirable to have better upper bounds, especially in low dimensions (see Fig. 2). The Kabatiansky-Levenshtein bound is asymptotically better than the Rogers' bound, but not until the dimension is above about 40. We know very little about these problems!

In short, many beautiful packings have been discovered, but there are few proofs that any of them are optimal.

References

[1] C. Bachoc, *Applications of coding theory to the construction of modular lattices*, J. Combin. Theory A 78 (1997), 92–119.

[2] C. Bachoc and G. Nebe, *Extremal lattices of minimum 8 related to the Mathieu group M_{22}*, J. reine angew. Math. 494 (1998), 155–171.

[3] A. Bezdek and W. Kuperberg, *Packing Euclidean space with congruent cylinders and with congruent ellipsoids*, in *Victor Klee Festschrift*, ed. P. Gritzmann et al., Amer. Math. Soc., 1991, pp. 71–80.

[4] J. Bierbrauer and Y. Edel, *Dense sphere packings from new codes*, preprint, 1998.

[5] A. Bonnecaze, A. R. Calderbank and P. Solé, *Quaternary quadratic residue codes and unimodular lattices*, IEEE Trans. Inform. Theory 41 (1995), 366–377.

[6] J. H. Conway and N. J. A. Sloane, *Laminated lattices*, Ann. Math. 116 (1982), 593–620.

[7] J. H. Conway and N. J. A. Sloane, *Low-dimensional lattices*: Proc. Royal Soc. Ser. A. I: 418 (1988), 17–41; II: 419 (1988), 29–68; III: 418 (1988), 43–80; IV: 419 (1988), 259–286; V: 426 (1989), 211–232; VI: 436 (1991), 55–68; VII: 453 (1997), 2369–2389; VIII (in preparation).

[8] J. H. Conway and N. J. A. Sloane, *A new upper bound for the minimum of an integral lattice of determinant one*, Bull. Am. Math. Soc. 23 (1990), 383–387; 24 (1991), 479.

[9] J. H. Conway and N. J. A. Sloane, *A new upper bound for the minimal distance of self-dual codes*, IEEE Trans. Inform. Theory 36 (1990), 1319–1333.

[10] J. H. Conway and N. J. A. Sloane, *The cell structures of certain lattices*, in *Miscellanea mathematica*, ed. P. Hilton et al., Springer-Verlag, NY, 1991, pp. 71–107.

[11] J. H. Conway and N. J. A. Sloane, *On lattices equivalent to their duals*, J. Number Theory 48 (1994), 373–382.

[12] J. H. Conway and N. J. A. Sloane, *Quaternary constructions for the binary single-error-correcting codes of Julin, Best and others*, Designs, Codes, Crypt. 4 (1994), 31–42.

[13] J. H. Conway and N. J. A. Sloane, *What are all the best sphere packings in low dimensions?*, Discrete Comput. Geom. 13 (1995), 383–403.

[14] J. H. Conway and N. J. A. Sloane, *The antipode construction for sphere packings*, Invent. math. 123 (1996), 309–313.

[15] J. H. Conway and N. J. A. Sloane, *A note on unimodular lattices*, J. Number Theory (to appear).

[16] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, NY, 3rd edition, 1998.

[17] Y. Edel, E. M. Rains and N. J. A. Sloane, *On kissing numbers in dimensions 32 to 128*, Electron. J. Combin. 5 (1) (1998), paper R22.

[18] N. D. Elkies, personal communication.

[19] N. D. Elkies, *Mordell-Weil lattices in characteristic 2: I. Construction and first properties*, Internat. Math. Res. Notices (No. 8, 1994), 353–361.

[20] T. C. Hales, *Sphere packings*, Discrete Comput. Geom. I: 17 (1997), 1–51; II: 18 (1997), 135–149; III: preprint.

[21] W.-Y. Hsiang, *On the sphere packing problem and the proof of Kepler's conjecture*, Internat. J. Math. 93 (1993), 739–831; but see the review by G. Fejes Tóth, Math. Review 95g #52032, 1995.

[22] D.-O. Jaquet-Chiffelle, *Enumération complète des classes de formes parfaites en dimension 7*, Ann. Inst. Fourier 43 (1993), 21–55.

[23] F. R. Kschischang and S. Pasupathy, *Some ternary and quaternary codes and associated sphere packings*, IEEE Trans. Inform. Theory 38 (1992) 227–246.

[24] J. Leech and N. J. A. Sloane, *Sphere packing and error-correcting codes*, Canad. J. Math. 23 (1971), 718–745.

[25] S. Litsyn and A. Vardy, *The uniqueness of the Best code*, IEEE Trans. Inform. Theory 40 (1994), 1693–1698.

[26] C. L. Mallows, A. M. Odlyzko and N. J. A. Sloane, *Upper bounds for modular forms, lattices and codes*, J. Alg. 36 (1975), 68–76.

[27] J. Martinet, *Les réseaux parfaits des espaces euclidiens*, Masson, Paris, 1996.

[28] D. J. Muder, *A new bound on the local density of sphere packings*, Discrete Comput. Geom. 10 (1993), 351–375.

[29] G. Nebe, *Finite subgroups of $GL_n(\mathbb{Q})$ for $25 \leq n \leq 31$*, Comm. Alg. 24 (1996), 2341–2397.

[30] G. Nebe, *Some cyclo-quaternionic lattices*, J. Alg. 199 (1998), 472–498.

[31] G. Nebe and N. J. A. Sloane, *A Catalogue of Lattices*, published electronically at http://www.research.att.com/∼njas/lattices/.

[32] W. Plesken, *Finite rational matrix groups — a survey*, in *Proc. Conf. "The ATLAS: Ten Years After"*, to appear.

[33] H.-G. Quebbemann, *Lattices with theta-functions for $G(\sqrt{2})$ and linear codes*, J. Alg. 105 (1987), 443–450.

[34] H.-G. Quebbemann, *Modular lattices in Euclidean spaces*, J. Number Theory 54 (1995), 190–202.

[35] H.-G. Quebbemann, *Atkin-Lehner eigenforms and strongly modular lattices*, L'Enseign. Math. 43 (1997), 55–65.

[36] E. M. Rains and N. J. A. Sloane, *The shadow theory of modular and unimodular lattices*, J. Number Theory, to appear.

[37] C. E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J. 27 (1948), 379–423 and 623–656.

[38] T. Shioda, *Mordell-Weil lattices and sphere packings*, Am. J. Math. 113 (1991), 931–948.

[39] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, published electronically at http://www.research.att.com/∼njas/sequences/.

[40] A. Vardy, *A new sphere packing in 20 dimensions*, Invent. math. 121 (1995), 119–133.

N. J. A. Sloane
AT&T Labs-Research
180 Park Avenue
Florham Park NJ 07932-0971 USA
njas@research.att.com

Finite Geometries, Varieties and Codes

Joseph A. Thas

Abstract. In recent years there has been an increasing interest in finite projective spaces, and important applications to practical topics such as coding theory and design of experiments have made the field even more attractive. It is my intention to mention some important and elegant theorems, to say something about the used techniques and the relation with other fields, and to mention some open problems. First some characterizations of particular pointsets in the projective space $\mathrm{PG}(n,q), n \geq 2$, over $\mathrm{GF}(q)$ will be given, where, from the beginning, it is assumed that the pointset is contained in $\mathrm{PG}(n,q)$. A second approach is that where the object is described as an incidence structure satisfying certain properties; here the geometry is not a priori embedded in a projective space. This approach will be illustrated with some theorems on inversive planes, polar spaces and Shult spaces. Finally, there is a section on k-arcs in $\mathrm{PG}(n,q)$ and on linear Maximum Distance Separable codes, where the interplay between finite projective geometry, coding theory and algebraic geometry is particularly present. In an appendix an example of brand new research in the field is given.

1991 Mathematics Subject Classification: 51E15, 51E20, 51E21, 51E25, 51A50, 51B10, 05B05, 05B25, 94B27
Keywords and Phrases: Finite Geometries, Varieties, Codes, Designs, k-Arcs, Polar Spaces

1 Introduction and history

In recent years there has been an increasing interest in finite projective spaces (or Galois spaces), and important applications to practical topics such as coding theory and design of experiments have made the field even more attractive. Basic works on the subject are "Projective Geometries over Finite Fields", "Finite Projective Spaces of Three Dimensions" and "General Galois Geometries", the first two volumes being written by Hirschfeld [1979,1985] and the third volume by Hirschfeld and Thas [1991]; the set of three volumes was conceived as a single entity. We also mention the "Handbook of Incidence Geometry: Buildings and Foundations", edited in 1995 by Buekenhout, which covers an enormous amount of material. In 1998 the second edition of the first volume by Hirschfeld appeared; here the author writes the following on the history of finite geometry (for bibliographical details, see Hirschfeld).

"The first actual reference or near-reference on finite geometry is von Staudt's Beiträge (1856). It contains countings of real and complex points of a projective space, as if they were points over $\mathrm{GF}(q)$ and $\mathrm{GF}(q^2)$; only dimensions two and three are considered. Then Fano (1892) defined $\mathrm{PG}(n,p)$ synthetically, while more than a decade later Hessenberg (1903) did it analytically. Next, Veblen and Bussey (1906) gave the first systematic account of $\mathrm{PG}(n,q)$ for any n and q; really, it may be noted that the group $\mathrm{PGL}(n+1,q)$ of projectivities, which is implicit in the geometry, goes back to Jordan (1870). At the same time and later, Dickson was investigating modular invariants, curves and other parts of algebraic geometry over a finite field. The link with statistics was developed by Bose (1947); earlier, Fisher (1942) had produced an experimental design from a finite plane, with Yates (1935) already having made the connection with block designs".

In his investigations on graph theory, design theory and finite projective spaces, the statistician Bose mainly used pure combinatorial arguments in combination with some linear algebra. Another great pioneer in finite projective geometry was the Italian geometer Beniamino Segre. His celebrated result of 1954 stating that in the projective plane $\mathrm{PG}(2,q)$ over the Galois field $\mathrm{GF}(q)$ with q odd, every set of $q+1$ points, no three of which are collinear, is a conic, stimulated the enthusiasm of many young geometers. The work of Segre and his followers has many links with error-correcting codes and with maximum distance separable codes, in particular. Finally, the fundamental and deep work in the last four decades on polar spaces, generalized polygons, and, more generally, incidence geometry, in the first place by Tits, but also by Shult, Buekenhout, Kantor and others, gave a new dimension to finite geometry.

Here I will state some important and elegant theorems, say something about the used techniques and the relation with other fields, and mention some open problems.

2 The geometry of $\mathrm{PG}(2,q)$

First I will consider the geometry of $\mathrm{PG}(2,q)$, that is, the projective plane over the finite field $\mathrm{GF}(q)$. To begin with, it is the purpose to show how classical algebraic curves can be characterized in pure combinatorial terms. I will illustrate this with a theorem on conics and one on Hermitian curves.

A k-arc of $\mathrm{PG}(2,q)$ is a set of k points of $\mathrm{PG}(2,q)$ no three of which are collinear. Then clearly $k \leq q+2$. By Bose [1947], for q odd, $k \leq q+1$. Further, any nonsingular conic of $\mathrm{PG}(2,q)$ is a $(q+1)$-arc. It can be shown that each $(q+1)$-arc K of $\mathrm{PG}(2,q)$, q even, extends to a $(q+2)$-arc $K \cup \{x\}$ (see, e.g., Hirschfeld [1998], p.177); the point x, which is uniquely defined by K, is called the *kernel* or *nucleus* of K. The $(q+1)$-arcs of $\mathrm{PG}(2,q)$ are called *ovals*. The following celebrated theorem is due to Segre [1954].

Theorem 1. *In* $\mathrm{PG}(2,q)$, q *odd, every oval is a nonsingular conic.*

For q even, Theorem 1 is valid if and only if $q \in \{2,4\}$; see e.g., Thas [1995a].

A *Hermitian arc* or *unital* H of $\mathrm{PG}(2,q)$, with q a square, is a set of $q\sqrt{q}+1$ points of $\mathrm{PG}(2,q)$ such that any line of $\mathrm{PG}(2,q)$ intersects H in either 1 or $\sqrt{q}+1$ points. The lines intersecting H in one point are called the *tangent lines* of H. At each of its points H has a unique tangent line. Let ζ be a unitary polarity of $\mathrm{PG}(2,q)$, q a square. Then the absolute points of ζ, that is, the points x of $\mathrm{PG}(2,q)$ which lie on their image x^ζ, form a Hermitian arc. Such a Hermitian arc is called a *nonsingular Hermitian curve*. For any nonsingular Hermitian curve coordinates in $\mathrm{PG}(2,q)$ can always be chosen in such a way that it is represented by the polynomial equation

$$X_0^{\sqrt{q}+1} + X_1^{\sqrt{q}+1} + X_2^{\sqrt{q}+1} = 0.$$

In 1992 the following theorem was obtained, solving a longstanding conjecture on Hermitian curves; see Thas [1992a].

THEOREM 2. *In* $\mathrm{PG}(2,q)$, q *a square, a Hermitian arc* H *is a nonsingular Hermitian curve if and only if tangent lines of* H *at collinear points are concurrent.*

Theorems 1 and 2 are pure combinatorial characterizations of algebraic curves. Now we give a characterization, due to Hirschfeld, Storme, Thas and Voloch [1991], in terms of algebraic curves, that is, we will assume from the beginning that our pointset is an algebraic curve.

THEOREM 3. *In* $\mathrm{PG}(2,q)$, q *a square and* $q \neq 4$, *any algebraic curve of degree* $\sqrt{q}+1$, *without linear components, and with at least* $q\sqrt{q}+1$ *points in* $\mathrm{PG}(2,q)$, *must be a nonsingular Hermitian curve.*

To prove the previous theorems, classical projective geometry, finite algebraic geometry, finite field theory and counting arguments were used. A proof of a completely different nature was used to solve a conjecture from 1975 on the following easily defined pointsets in $\mathrm{PG}(2,q)$.

In $\mathrm{PG}(2,q)$ any nonempty set of k points may be described as a $(k;m)$-arc, where m ($m \neq 0$) is the greatest number of collinear points in the set. For given q and m ($m \neq 0$), k can never exceed $mq-q+m$, and a $(mq-q+m;m)$-arc is called a *maximal arc*. Equivalently, a maximal arc may be defined as a nonempty set of points meeting every line in just m points or in none at all. Trivial maximal arcs are the plane $\mathrm{PG}(2,q)$ itself ($m=q+1$), the affine plane $\mathrm{AG}(2,q)$ obtained by deleting a line L from $\mathrm{PG}(2,q)$ ($m=q$), and a single point ($m=1$). If K is a $(mq-q+m;m)$-arc of $\mathrm{PG}(2,q)$, where $m \leq q$, then it is easy to show that the set

$$K' = \{\text{lines } L \text{ of } \mathrm{PG}(2,q) \ : \ L \cap K = \emptyset\}$$

is a $(q(q-m+1)/m;q/m)$-arc (i.e., a maximal arc) of the dual plane. Hence, if the plane $\mathrm{PG}(2,q)$ contains a $(mq-q+m;m)$-arc, $m \leq q$, then it also contains a $(q(q-m+1)/m;q/m)$-arc. It follows that a necessary condition for the existence of a maximal arc, with $m \leq q$, is that m should be a factor of q.

In 1969 Denniston proves that the condition $m|q$ does suffice in the case of any plane $\mathrm{PG}(2,2^h)$, and in 1975 Thas proves that in $\mathrm{PG}(2,q)$, with $q=3^h$ and $h>1$, there are no $(2q+3;3)$-arcs and no $(q(q-2)/3;q/3)$-arcs. The longstanding conjecture that in $\mathrm{PG}(2,q)$, q odd, the only maximal arcs are the trivial ones, was proved just recently by Ball, Blokhuis and Mazzocca; see Ball, Blokhuis and Mazzocca [1997] and Ball and Blokhuis [1998].

THEOREM 4. *In* $\mathrm{PG}(2,q)$, q *odd, there is no maximal* $(qm-q+m;m)$*-arc with* $1<m<q$.

In the proof the point (x,y) of the affine plane $\mathrm{AG}(2,q)$ is identified with the element $x+\alpha y$ of $\mathrm{GF}(q^2)=\mathrm{GF}(q)(\alpha)$. Then, assuming the existence of a nontrivial maximal arc in $\mathrm{AG}(2,q), q$ odd, polynomials over $\mathrm{GF}(q^2)$ are defined the clever manipulation of which leads to a contradiction.

3 The geometry of $\mathrm{PG}(n,q)$, $n\geq 3$

If $\mathcal{V}$ is a "classical" algebraic variety in $\mathrm{PG}(n,q)$ (or one of its projections), $n\geq 3$, e.g., a quadric, a Hermitian variety, a Veronese variety, then a first approach is to characterize $\mathcal{V}$ either as a subset of $\mathrm{PG}(n,q)$ which intersects certain subspaces of $\mathrm{PG}(n,q)$ in sets with cardinalities in some range or as a subset of $\mathrm{PG}(n,q)$ whose points satisfy certain linear independence conditions. One characterization of the first type will be given here, while in Section 4 we will show how (subsets of) normal rational curves can be characterized by one simple independence condition on the points.

A *nonsingular Hermitian variety* H of $\mathrm{PG}(n,q)$, q a square and $n\geq 2$, is any subset of $\mathrm{PG}(n,q)$ which is equivalent under the group $\mathrm{PGL}(n+1,q)$ to the subset of $\mathrm{PG}(n,q)$ represented by the equation

$$X_0^{\sqrt{q}+1}+X_1^{\sqrt{q}+1}+\cdots+X_n^{\sqrt{q}+1}=0.$$

A subset K of $\mathrm{PG}(n,q)$, $n\geq 2$, is of *type* $(1,m,q+1)$ if every line of $\mathrm{PG}(n,q)$ meets it in $1,m$, or $q+1$ points. A point of K is *singular* if every line through it intersects K either in 1 or $q+1$ points. Then K is called *singular* or *nonsingular* as it has singular points or not.

THEOREM 5. *If* K *is a nonsingular set of type* $(1,m,q+1)$ *of* $\mathrm{PG}(n,q)$, *with* $3\leq m\leq q-1, n\geq 3$ *and* $q>4$, *then* K *is either a nonsingular Hermitian variety of* $\mathrm{PG}(n,q)$ *(and then* $m=\sqrt{q}+1$*) or the projection onto* $\mathrm{PG}(n,q)$ *of a nonsingular quadric* Q *of* $\mathrm{PG}(n+1,q)\supset\mathrm{PG}(n,q)$ *from a point* $x\in\mathrm{PG}(n+1,q)\setminus\mathrm{PG}(n,q)$ *other than the nucleus (or kernel) of* Q *in the case that* n *is even (and here* $m=\frac{q}{2}+1$, *so* q *is even).*

For $m\neq\frac{q}{2}+1$ the result is due to Tallini Scafati [1967] and for $m=\frac{q}{2}+1, n>3$ and part of $n=3$, to Hirschfeld and Thas [1980a, 1980b]. The missing part in the case $m=\frac{q}{2}+1$ and $n=3$ was done by Glynn [1983]. Finally the case

$(q, m) = (4, 3)$ was handled by Sherman [1983], see also Hirschfeld and Hubaut [1980] and Hirschfeld [1985]; it appears that here the sets K can be identified with the codewords of a *projective geometry code*.

A second approach is that where the object is described as an incidence structure satisfying certain properties; here the geometry is not a priori embedded in a projective space, even the finite field is in many cases a priori absent. Hence the finite projective space must be constructed.

A first example concerns circle geometries and designs. A $t-(v,k,\lambda)$ *design*, with $v > k > 1, k \geq t \geq 1, \lambda > 0$, is a set P with v elements called *points*, provided with subsets of size k called *blocks*, such that any t distinct points are contained in exactly λ blocks. A $3-(n^2+1, n+1, 1)$ design is usually called an *inversive plane* or *Möbius plane* of *order* n; here the blocks are mostly called *circles*. An *ovoid* O of $\mathrm{PG}(3,q)$, $q > 2$, is a set of q^2+1 points no three of which are collinear; an *ovoid* of $\mathrm{PG}(3,2)$ is the same as a nonsingular elliptic quadric, that is, a nonsingular quadric of $\mathrm{PG}(3,2)$ containing no lines. For properties on ovoids we refer to Hirschfeld [1985]. If O is an ovoid, then O provided with all intersections $\pi \cap O$, where π is any plane containing at least 2 (and then automatically $q+1$) points of O, is an inversive plane $\mathcal{J}(O)$ of order n. An inversive plane arising from an ovoid is called *egglike*. The following famous theorem is due to Dembowski [1964].

THEOREM 6. *Each (finite) inversive plane of even order is egglike.*

If the ovoid O is an elliptic quadric, then the inversive plane $\mathcal{J}(O)$ is called *classical* or *Miquelian*. Barlotti [1955] and, independently, Panella [1955] proved that for q odd any ovoid is an elliptic quadric. Hence for q odd any egglike inversive plane is Miquelian. For odd order no other inversive planes are known. To the contrary, in the even case Tits [1962] showed that for any $q = 2^{2e+1}$, with $e \geq 1$, there exists an ovoid which is not an elliptic quadric; these ovoids are called Tits ovoids and are related to the simple Suzuki groups $Sz(q)$. For even order no other nonclassical inversive planes than the ones associated to the Tits ovoids are known.

Let $\mathcal{J}$ be an inversive plane of order n. For any point x of $\mathcal{J}$, the points of $\mathcal{J}$ different from x, together with the circles containing x (minus x), form a $2-(n^2, n, 1)$ design, that is, an *affine plane* of order n. That affine plane is denoted by $\mathcal{J}_x$, and is called the *internal plane* or *derived plane* of $\mathcal{J}$ at x. For an egglike inversive plane $\mathcal{J}(O)$ of order q, each internal plane is Desarguesian, that is, is the affine plane $\mathrm{AG}(2,q)$. The following theorem, due to Thas [1994], solves a longstanding conjecture on circle geometries.

THEOREM 7. *Let $\mathcal{J}$ be an inversive plane of odd order n. If for at least one point x of $\mathcal{J}$ the internal plane $\mathcal{J}_x$ is Desarguesian, then $\mathcal{J}$ is Miquelian.*

The key idea in the proof of this theorem on Möbius planes is to use a fundamental result on Minkowski planes (another type of circle geometries), which in turn depends on the classification of a particular class of quasifields. As a corollary of Theorem 7 we obtain the first computer-free proof of the uniqueness (up to isomorphism) of the inversive plane of order 7.

Another beautiful illustration of this second approach is a characterization of all polar spaces of rank at least three. Here, starting from about nothing we get everything. First, let us give Tits' axioms for a polar space of rank r, with $r \geq 3$.

A *polar space* $\mathcal{S}$ of *rank* r, with $r \geq 3$, is a set P of elements called *points*, provided with distinguished subsets called *subspaces*, such that the following properties are satisfied.

(i) Any subspace, together with the subspaces it contains, is a projective space of dimension at most $r-1$.

(ii) The intersection of any family of subspaces is a subspace.

(iii) Given a subspace π of dimension $r-1$ and a point p in $P\backslash\pi$, there exists a unique subspace π' containing p such that the dimension of $\pi \cap \pi'$ is $r-2$. Also, the subspace $\pi \cap \pi'$ is the set of all points p' of π such that p and p' are contained in some subspace of dimension one.

(iv) There exist two disjoint subspaces of dimension $r-1$.

Isomorphisms between polar spaces are defined in the usual way.

Examples of finite polar spaces
(a) Let Q be a nonsingular quadric in $\mathrm{PG}(n,q)$ of rank r (that is, Q contains $(r-1)$-dimensional projective spaces, but no r-dimensional projective space), with $r \geq 3$. Then Q together with the projective spaces lying on it is a polar space of rank r.
(b) Let H be a nonsingular Hermitian variety of $\mathrm{PG}(n,q^2)$, $n \geq 5$. Then H together with the subspaces lying on it is a polar space of rank $[\frac{n+1}{2}]$ (here $[\frac{n+1}{2}]$ is the greatest integer less than or equal to $\frac{n+1}{2}$).
(c) Let ζ be a (nonsingular) symplectic polarity of $\mathrm{PG}(n,q)$, with n odd. Then $\mathrm{PG}(n,q)$ together with all absolute subspaces of ζ, is a polar space of rank $(n+1)/2$ (a projective subspace π of $\mathrm{PG}(n,q)$ is absolute for ζ if $\pi^\zeta \subseteq \pi$).

A complete classification of all polar spaces of rank at least three has been obtained by Tits [1974], building on work of Veldkamp [1959]. We state now this celebrated deep result in the finite case.

Theorem 8. *If $\mathcal{S}$ is a finite polar space of rank at least three, then $\mathcal{S}$ is isomorphic to one of (a), (b), (c).*

Polar spaces of rank 2 were also defined by Tits [1959]; these polar spaces are usually called *generalized quadrangles*. The role of generalized quadrangles in the theory of polar spaces, can be compared to the role of projective planes in the theory of projective spaces. Just as for projective planes a complete classification of all generalized quadrangles seems to be hopeless. For more on generalized quadrangles we refer to the monograph by Payne and Thas [1984] and to Thas [1995b].

Now we will describe polar spaces in an extremely simple way, just using points and lines.

A *Shult space* $\mathcal{S}$ is a nonempty set P of *points* together with distinguished subsets of cardinality at least two called *lines* such that for each line L of $\mathcal{S}$ and each point p of $P \setminus L$, the point p is collinear with either one or all points of L; here two not necessarily distinct points p_1 and p_2 are called *collinear* if there is at least one line of $\mathcal{S}$ containing these points. A Shult space is *nondegenerate* if no point of $\mathcal{S}$ is collinear with all points of $\mathcal{S}$. A *subspace* X of a Shult space $\mathcal{S}$ is a set of pairwise collinear points such that any line meeting X in more than one point is contained in X. The Shult space $\mathcal{S}$ has rank r, with $r \geq 1$, if r is the largest integer for which there is a chain

$$X_0 \subset X_1 \subset \ldots \subset X_r$$

of distinct subspaces $X_0 = \emptyset, X_1, X_2, \ldots, X_r$.

From Theorem 8 it follows that, for any finite polar space $\mathcal{S}$ of rank r, with $r \geq 3$, the pointset P together with the subspaces of dimension one is a Shult space of rank r. In fact this result also holds for infinite polar spaces. The following beautiful and extremely strong converse is due to Buekenhout and Shult [1974].

Theorem 9. *Any nondegenerate Shult space of rank r, with $r \geq 3$, all of whose lines have cardinality at least three, together with its subspaces, is a polar space of rank r.*

We remark that Buekenhout and Shult [1974] also classified all degenerate Shult spaces; in fact, the problem is reduced to the classification of the nondegenerate ones.

Finally, let us mention that further fundamental and deep work on polar spaces, point-line geometries related to buildings, and, more generally, incidence geometry, was done in the first place by Tits, and further by Buekenhout, Cohen, Cooperstein, Kantor, Shult and others; these developments gave a new dimension to finite geometry. As excellent reference we mention the "Handbook of Incidence Geometry: Buildings and Foundations", edited by Buekenhout in 1995.

4 An exemplary illustration of the interplay between Galois geometry, coding theory and algebraic geometry

Let C be a *code* of *length* k over an alphabet A of size q, with $q \in \mathbb{N}\backslash\{0,1\}$. In other words C is simply a set of (code) words where each word is an element of A^k. Having chosen m, with $2 \leq m \leq k$, we impose the following condition on C: no two words in C agree in as many as m positions. It then follows that $|C| \leq q^m$. If $|C| = q^m$, then C is called a *Maximum Distance Separable code (MDS code)*. MDS codes are exactly the codes which meet the Singleton bound; see e.g. Hill [1986]. There is a voluminous literature on the subject; we refer e.g. to the standard work by MacWilliams and Sloane [1977] and to the book by Hill [1986]. MacWilliams and Sloane introduce their chapter on MDS codes as "*one of the most fascinating in all of coding theory*".

The *Hamming distance* between two code words $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ is the number of indices i for which $x_i \neq y_i$; it is denoted by

$d(x,y)$. The *minimum Hamming distance* of C is

$$\min(d(x,y) : x,y \in C \text{ and } x \neq y)$$

and is denoted by $d(C)$. If C is an MDS code, then one easily shows that

$$d(C) = k - m + 1,$$

that is, the Singleton bound is met. One of the main problems concerning such codes is to maximize $d(C)$, and so k, for given m and q. Also, what is the structure of C in the optimal case?

Now the problem will be formulated for the case when C is linear, i.e., for the case that C is an m-dimensional subspace of the k-dimensional vector space $V(k,q)$ over $\mathrm{GF}(q)$. It goes like this. Choose any basis for C and represent it as an $m \times k$-matrix A over $\mathrm{GF}(q)$ of rank m. *Then C is MDS if and only if every set of m columns of A is linearly independent.*

Next, let us turn to particular pointsets of $\mathrm{PG}(n,q)$ introduced by Segre in 1955. A *k-arc* in $\mathrm{PG}(n,q)$, with $n \geq 2$, is a set K of k points with $k \geq n+1$ such that no $n+1$ points of K lie in a hyperplane, that is, such that any $n+1$ points are linearly independent. A k-arc K is *complete* if it is not properly contained in a larger arc. Otherwise, if $K \cup \{x\}$ is a $(k+1)$-arc for some point x of $\mathrm{PG}(n,q)$ we say that x *extends* K.

A *normal rational curve* (NRC) of $\mathrm{PG}(n,q)$, with $q > n+1$, is any set of points in $\mathrm{PG}(n,q)$ which is equivalent under the group $\mathrm{PGL}(n+1,q)$ to

$$\{(t^n, t^{n-1}, \ldots, t, 1) : t \in \mathrm{GF}(q)\} \cup \{(1,0,\ldots,0,0)\}.$$

Clearly any NRC is a $(q+1)$-arc. A NRC of $\mathrm{PG}(2,q)$ is a *nonsingular conic*; a NRC of $\mathrm{PG}(3,q)$ is a *twisted cubic*.

k-Arcs were introduced by Segre [1955], who also posed the next three fundamental problems.

(a) For given n and q what is the maximum value of k for which there exist k-arcs in $\mathrm{PG}(n,q)$?

(b) For what values of n and q, with $q > n+1$, is every $(q+1)$-arc a NRC?

(c) For given n and q, with $q > n+1$, what are the values of k for which every k-arc is contained in a $(q+1)$-arc of this space?

The famous Theorem 1 of Segre gives the answer, for q odd, to Problem (b) in the twodimensional case.

Hundreds of papers were written on k-arcs, in particular on the above problems, which are now solved for "most" values of the parameters. For example, if $q > f(n)$ with f some quadratic polynomial over $\mathbb{Q}$, then $k \leq q+1$ for $n \geq 3$, and any $(q+1)$-arc in $\mathrm{PG}(n,q)$, with $n \geq 3$ and $(n,q) \neq (3,2^h)$, is a NRC; also, by

Casse and Glynn [1982] any $(q+1)$-arc of $\mathrm{PG}(3,q)$, $q=2^h$, is equivalent under $\mathrm{PGL}(4,q)$ to

$$\{(t^{2^r+1}, t^{2^r}, t, 1) : t \in \mathrm{GF}(q)\} \cup \{(1,0,0,0)\},$$

with $(r,h)=1$.

The main tool in the proofs is that with any k-arc of $\mathrm{PG}(n,q)$ there corresponds an algebraic hypersurface in the dual space of $\mathrm{PG}(n,q)$. For $n=2$ this was proved by Segre [1967] and for $n>2$ twenty years later, by Bruen, Thas and Blokhuis [1988]. Essential also are the bounds on the number of points of an algebraic curve in $\mathrm{PG}(2,q)$, particularly the Hasse-Weil bound (see, e.g., Sections 2.9 and 2.15 of Hirschfeld [1988] for references) and the Stöhr-Voloch [1986] bound.

For surveys on k-arcs we refer to Hirschfeld and Thas [1991], Thas [1992b,1995a] and Hirschfeld and Storme [1998].

The main conjecture on k-arcs is the following.

CONJECTURE. *If K is a k-arc of $\mathrm{PG}(n,q)$, with $q \geq n+1$, then*

(a) for q even and $n \in \{2, q-2\}$ we have $k \leq q+2$,

(b) $k \leq q+1$ in all other cases.

We remark that $(q+2)$-arcs exist in $\mathrm{PG}(2,q)$ and $\mathrm{PG}(q-2,q)$, $q=2^h$ and $h \geq 2$; see Hirschfeld and Thas [1991].

As already mentioned, a linear code C of length k and dimension m, with $2 \leq m \leq k$, over $\mathrm{GF}(q)$ is MDS if and only if it is generated by the rows of an $m \times k$-matrix A over $\mathrm{GF}(q)$ for which every set of m columns is linearly independent. Now we regard the columns of A as points $p_1, p_2, \ldots, p_k$ of $\mathrm{PG}(m-1,q)$. Then, for a linear MDS code, no m of these points lie in a hyperplane, that is, for $m \geq 3$ these points form a k-arc of $\mathrm{PG}(m-1,q)$. Conversely, with any k-arc of $\mathrm{PG}(m-1,q), m \geq 3$, there corresponds a linear MDS code. Remark that the linear MDS codes of dimension two are known and are quite trivial. So we have the following theorem.

THEOREM 10. *Linear* MDS *codes of dimension at least three and k-arcs are equivalent objects.*

So each result on linear MDS codes of dimension at least three can be tranlated into a result on k-arcs, and conversely. This way a lot of new fundamental results on linear MDS codes were obtained. Many of these translated results on k-arcs were proved long before the relation with coding theory was discovered. This is indeed a beautiful example of interrelationship between pure finite geometry and coding theory.

5 APPENDIX: RECENT RESEARCH IN FINITE GEOMETRIES

In this appendix I will give an example of brand new research in the field.

Let P and B be disjoint sets, each consisting of q^2+q+1 lines of $\mathrm{PG}(n,q)$. An element L of P and an element M of B are called incident if and only if $L \cap M \neq \emptyset$. Now assume that the point-line incidence structure with pointset P, lineset B and the given incidence is a projective plane $\mathcal{P}$ of order q. Finally, we suppose that for any incident point-line pair (L, M) of $\mathcal{P}$, all points and lines of $\mathcal{P}$ incident either with L or with M are contained in a common hyperplane of $\mathrm{PG}(n,q)$. Then the author and Van Maldeghem just proved that the plane $\mathcal{P}$ is Desarguesian, that $n \in \{6,7,8\}$, that for $n=6$ q is a power of 3 and that, up to isomorphism, there is a unique example in $\mathrm{PG}(6,q)$ for any such $q=3^h$. Also, they already handled large part of the remaining cases $n=7,8$, and the complete classification normally should be finished by the beginning of the conference.

The solution of this problem is a key step in the determination of all dual classical generalized hexagons with $q+1$ points on any line, whose points are points of $\mathrm{PG}(n,q)$ and whose lines are lines of $\mathrm{PG}(n,q)$; see Thas [1995b] for the definition of generalized hexagon.

References

S. Ball and A. Blokhuis [1998]. An easier proof of the maximal arcs conjecture. *Proc. Amer. Math. Soc.*, to appear.

S. Ball, A. Blokhuis, and F. Mazzocca [1997]. Maximal arcs in desarguesian planes of odd order do not exist. *Combinatorica*, 17:31-47.

A. Barlotti [1955]. Un 'estensione del teorema di Segre-Kustaanheimo. *Boll. Un. Mat. Ital.*, 10:96-98.

R.C. Bose [1947]. Mathematical theory of the symmetrical factorial design. *Sankhyā*, 8:107-166.

A.A. Bruen, J.A. Thas and A. Blokhuis [1988]. On M.D.S. codes, arcs in $\mathrm{PG}(n,q)$ with q even, and a solution of three fundamental problems of B. Segre. *Invent. Math.*, 92:441-459.

F. Buekenhout (editor) [1995]. *Handbook of Incidence Geometry: Buildings and Foundations.* North-Holland, Amsterdam.

F. Buekenhout and E.E. Shult [1974]. On the foundations of polar geometry. *Geom. Dedicata*, 3:155-170.

L.R.A. Casse and D.G. Glynn [1982]. The solution to Beniamo Segre's problem $I_{r,q}$, $r=3, q=2^h$. *Geom. Dedicata* 13:157-164.

P. Dembowski [1964]. Möbiusebenen gerader Ordnung, *Math. Ann.*, 157: 179-205.

R.H.F. Denniston [1969]. Some maximal arcs in finite projective planes. *J. Combin. Theory*, 6:317-319.

D.G. Glynn [1983]. On the characterization of certain sets of points in finite projective geometry of dimension three. *Bull. London Math. Soc.*, 15:31-34.

R. Hill [1986]. *A First Course in Coding Theory.* Oxford University Press, Oxford.

J.W.P. Hirschfeld [1985]. *Finite Projective Spaces of Three Dimensions.* Oxford Univerity Press, Oxford.

J.W.P. Hirschfeld [1998]. *Projective Geometries over Finite Fields.* Oxford University Press, Oxford; first edition, 1979.

J.W.P. Hirschfeld and X. Hubaut [1980]. Sets of even type in PG$(3, q)$ alias the binary (85,24) projective code. *J. Combin. Theory Ser. A*, 29:101-112.

J.W.P. Hirschfeld and L. Storme [1998]. The packing problem in statistics, coding theory and finite projective spaces. *J. Statist. Plann. Inference*, to appear.

J.W.P. Hirschfeld, L. Storme, J.A. Thas, and J.F. Voloch [1991]. A characterization of Hermitian curves. *J. Geom.*, 41:72-78.

J.W.P. Hirschfeld and J.A. Thas [1980a]. The characterization of projections of quadrics over finite fields of even order. *J. London Math. Soc.*, 22:226-238.

J.W.P. Hirschfeld and J.A. Thas [1980b]. Sets of type $(1, n, q + 1)$ in PG(d, q). *Proc. London Math. Soc.*, 41:254-278.

J.W.P. Hirschfeld and J.A. Thas [1991]. *General Galois Geometries.* Oxford University Press, Oxford.

F.J.K. MacWilliams and N.J.A. Sloane [1977]. *The Theory of Error-Correcting Codes.* North-Holland, Amsterdam.

G. Panella [1955]. Caratterizzazione delle quadriche di uno spazio (tridimensionale) lineare sopra un corpo finito. *Boll. Un. Mat. Ital.*, 10:507-513.

S.E. Payne and J.A. Thas [1984]. *Finite Generalized Quadrangles.* Pitman, London.

B. Segre [1954]. Sulle ovali nei piani lineari finiti. *Atti Accad. Naz. Lincei Rend.*, 17:1-2.

B. Segre [1955]. Curve razionali normali e k-archi negli spazi finiti. *Ann. Mat. Pura Appl.*, 39:357-379.

B. Segre [1967]. Introduction to Galois Geometries (edited by J.W.P. Hirschfeld). *Atti Accad. Naz. Lincei Mem.*, 8:133-236.

B.F. Sherman [1983]. On sets with only odd secants in geometries over GF(4). *J. London Math. Soc.*, 27:539-551.

K.O. Stöhr and J.F. Voloch [1986]. Weierstrass points and curves over finite fields. *Proc. London Math. Soc.*, 52:1-19.

M. Tallini Scafati [1967]. Caratterizzazione grafica delle forme hermitiane di un $S_{r,q}$. *Rend. Mat. e Appl.*, 26:273-303.

J.A. Thas [1975]. Some results concerning $\{(q+1)(n-1); n\}$-arcs and $\{(q+1)(n-1)+1; n\}$-arcs in finite projective planes of order q. *J. Combin. Theory Ser. A*, 19:228-232.

J.A. Thas [1992a]. A combinatorial characterization of Hermitian curves. *J. Algebraic Combin.*, 1:97-102.

J.A. Thas [1992b]. M.D.S. codes and arcs in projective spaces: a survey. *Matematiche (Catania)*, 47:315-328.

J.A. Thas [1994]. The affine plane $\mathrm{AG}(2, q)$, q odd, has a unique one point extension. *Invent. Math.*, 118:133-139.

J.A. Thas [1995a]. Projective geometry over a finite field. In *Handbook of Incidence Geometry: Buildings and Foundations*, pages 295-347. North-Holland, Amsterdam.

J.A. Thas [1995b]. Generalized polygons. In *Handbook of Incidence Geometry: Buildings and Foundations*, pages 383-431. North-Holland, Amsterdam.

J. Tits [1959]. Sur la trialité et certains groupes qui s'en déduisent. *Inst. Hautes Etudes Sci. Publ. Math.,* 2:13-60.

J. Tits [1962]. Ovoïdes et groupes de Suzuki. *Arch. Math.*, 13:187-198.

J. Tits [1974]. *Buildings of Spherical Type and Finite BN-Pairs.* Lecture Notes in Math. 386, Springer, Berlin.

F.D. Veldkamp [1959]. Polar geometry I-IV. *Indag. Math.*, 21:512-551.

Joseph A. Thas
University of Ghent,
Department of Pure Mathematics
and Computer Algebra,
Krijgslaan 281,
B-9000 Gent, Belgium,
jat@cage.rug.ac.be

Multisegment Duality, Canonical Bases and Total Positivity

Andrei Zelevinsky

Abstract. We illustrate recent interactions between algebraic combinatorics, representation theory and algebraic geometry with a piecewise-linear involution called the multisegment duality.

1991 Mathematics Subject Classification: 5, 14, 20
Keywords and Phrases: multisegments, Young tableaux, polyhedral combinatorics, quivers, canonical bases, total positivity

1. Introduction. We discuss some recent interactions between representation theory, algebraic geometry and algebraic combinatorics. Classically such an interaction involves:

- finite-dimensional representations of symmetric and general linear groups;
- geometry of flag varieties and Schubert varieties;
- combinatorics of Young tableaux and related algorithms such as the Robinson-Schensted-Knuth correspondence.

More recent advances in representation theory such as Lusztig's canonical bases [14] and Kashiwara's crystal bases [10] require new geometric and combinatorial tools. On the geometric side, an important role is played by quiver representation varieties and totally positive varieties. On the combinatorial side, the objects of interest become rational polyhedral convex cones and polytopes, their lattice points, and their piecewise-linear transformations.

We illustrate this interplay with a particular piecewise-linear involution, the multisegment duality. It was introduced in [20, 21] in the context of representations of general linear groups over a p-adic field. It also naturally appears in the geometry of quiver representations, and in the study of canonical bases for quantum groups. On the combinatorial side, it is closely related to Schützenberger's involution on Young tableaux [19], as demonstrated in [4]. In this talk, we give a new combinatorial interpretation of the multisegment duality as an intertwining map between two piecewise-linear actions of the Lascoux-Schützenberger plactic monoid [12].

2. MULTISEGMENT DUALITY AND QUIVER REPRESENTATIONS. We fix a positive integer r and consider the set $\Sigma = \Sigma_r$ of pairs of integers (i,j) such that $1 \le i \le j \le r$. We regard a pair $(i,j) \in \Sigma$ as a *segment* $[i,j] := \{i, i+1, \dots, j\}$ in $[1,r]$. Note that Σ can be identified with the set of positive roots of type A_r: each segment $[i,j]$ corresponds to a root $\alpha_i + \alpha_{i+1} + \cdots + \alpha_j$, where $\alpha_1, \dots, \alpha_r$ are the simple roots of type A_r in the standard numeration. Let $\mathbb{N}\Sigma$ denote the free abelian semigroup generated by Σ. We call its elements *multisegments*; they are formal linear combinations $\mathbf{m} = \sum_{(i,j)\in\Sigma} m_{ij}[i,j]$ with nonnegative integer coefficients.

Our main object of study will be the multisegment duality involution ζ on $\mathbb{N}\Sigma$. Following [21], we define it in terms of quiver representations of the equidirected quiver of type A_r. Such a representation is a collection of finite-dimensional vector spaces $V_1, \dots, V_r$ (say over $\mathbf{C}$) and linear maps $X_k : V_k \to V_{k+1}$ for $k = 1, \dots, r-1$. Morphisms between and direct sums of representations are defined in an obvious way. As a special case of Gabriel's classification [8], isomorphism classes of these quiver representations are in natural bijection with $\mathbb{N}\Sigma$. That is, the multisegment $\mathbf{m} = \sum m_{ij}[i,j]$ corresponds to the isomorphism class $I(\mathbf{m}) = \oplus m_{ij} I_{ij}$, where the indecomposable representations I_{ij} are defined as follows: the space V_k in I_{ij} is one-dimensional for $k \in [i,j]$, and zero otherwise, and $X_k(V_k) = V_{k+1}$ for $k \in [i, j-1]$.

We also consider representations of the opposite quiver: such a representation is a collection of finite-dimensional vector spaces $V_1, \dots, V_r$ and linear maps $Y_k : V_k \to V_{k-1}$ for $k = 2, \dots, r$. The isomorphism classes of these representations are also labeled by multisegments: now a multisegment $\mathbf{m}$ corresponds to the isomorphism class $I^{\mathrm{op}}(\mathbf{m}) = \oplus m_{ij} I_{ij}^{\mathrm{op}}$, where I_{ij}^{op} is obtained by reversing arrows in I_{ij}.

Now let $(V; X)$ be a quiver representation in the isomorphism class $I(\mathbf{m})$. Let $Z(V; X)$ be the variety of opposite quiver representations $(V; Y)$ on the same collection of vector spaces V_k such that $Y_{k+1}X_k = X_{k-1}Y_k$ for $k \in [1,r]$ (with the convention that $X_0 = X_r = Y_1 = Y_{r+1} = 0$). It is easy to show that all generic representations from $Z(V; X)$ belong to the same isomorphism class (here "generic" means that, for any $(i,j) \in \Sigma$, the composition $Y_{i+1} \cdots Y_j : V_j \to V_i$ has the maximal possible rank). We define $\zeta(\mathbf{m})$ to be the multisegment corresponding to a generic representation in $Z(V; X)$; that is, the isomorphism class of this generic representation is $I^{\mathrm{op}}(\zeta(\mathbf{m}))$. The definition readily implies that the map $\zeta : \mathbb{N}\Sigma \to \mathbb{N}\Sigma$ is an involution.

3. FORMULA FOR THE MULTISEGMENT DUALITY. We now present a closed formula for ζ obtained in [11]. For $(i,j) \in \Sigma$, let T_{ij} denote the set of all maps $\nu : [1,i] \times [j,r] \to [i,j]$ such that $\nu(k,l) \le \nu(k',l')$ whenever $k \le k'$ and $l \le l'$ (in other words, ν is a morphism of posets, where $[1,i] \times [j,r]$ is supplied with the product order). For any multisegment $\mathbf{m} = \sum m_{ij}[i,j]$, we set

$$\rho_{ij}(\mathbf{m}) = \min_{\nu \in T_{ij}} \sum_{(k,l)\in[1,i]\times[j,r]} m_{\nu(k,l)+k-i,\nu(k,l)+l-j} \tag{1}$$

(with the understanding that $\rho_{ij}(\mathbf{m}) = 0$ for $(i,j) \notin \Sigma$).

THEOREM 1. *For every multisegment* $\mathbf{m}$*, the multisegment* $\zeta(\mathbf{m}) = \sum m'_{ij}[i,j]$ *is given by*

$$m'_{ij} = \rho_{ij}(\mathbf{m}) - \rho_{i-1,j}(\mathbf{m}) - \rho_{i,j+1}(\mathbf{m}) + \rho_{i-1,j+1}(\mathbf{m}) \ . \tag{2}$$

The function $\rho_{ij}(\mathbf{m})$ in (1) has the following meaning: it is the rank of the map $Y_{i+1} \cdots Y_j : V_j \to V_i$ for any quiver representation $(V;Y)$ in the isomorphism class $I^{\mathrm{op}}(\zeta(\mathbf{m}))$.

The proof of Theorem 1 in [11] is elementary, using only linear algebra and combinatorics. The main ingredients of the proof are: the "Max Flow = Min Cut" theorem from the network flow theory [7], and the result of S. Poljak describing the maximal possible rank for a given power of a matrix with a given pattern of zeros [18].

4. REPRESENTATION-THEORETIC CONNECTIONS. It was conjectured in [20, 21] that the multisegment duality describes a natural duality operation acting on irreducible smooth representations of general linear groups over p-adic fields. In [17], this conjecture was reformulated in terms of representations of affine Hecke algebras and then proved. We recall (see [17, I.2]) that the affine Hecke algebra $\mathcal{H}_n$ can be defined as the associative algebra with unit over $\mathbf{Q}(q)$ generated by the elements $S_1, \ldots, S_{n-1}, X_1^{\pm 1}, \ldots, X_n^{\pm 1}$ subject to the relations:

$$(S_i - q)(S_i + 1) = 0, \ S_i S_{i+1} S_i = S_{i+1} S_i S_{i+1},$$

$$X_j X_k = X_k X_j, \ S_i X_j = X_j S_i \, (j \neq i, i+1), \ S_i X_{i+1} S_i = q X_i.$$

As shown in [17] using the results of [20], irreducible finite-dimensional representations of $\mathcal{H}_n$ are naturally indexed by multisegments $\mathbf{m} = \sum m_{ij}[i,j]$ with $\sum_{i,j}(j+1-i)m_{ij} = n$ (here we have to take the multisegments supported on some segment $[a,b] \subset \mathbf{Z}$, not only on $[1,r]$). According to [17, Proposition I.7.3], the involution ζ corresponds to the following involution on irreducible finite-dimensional representations of $\mathcal{H}_n$: $\pi \mapsto \pi \circ \varphi$, where φ is the automorphism of $\mathcal{H}_n$ defined by

$$\varphi(S_i) = -q S_{n-i}^{-1}, \ \varphi(X_j) = X_{n+1-j}.$$

(This result was extended from GL_n to other reductive groups in [1].)

Another interpretation of ζ is in terms of quantum groups. Let $\mathbf{C}_q[N]$ be the q-deformation of the algebra of regular functions on the group N of unipotent upper triangular $(r+1) \times (r+1)$ matrices (see, e.g., [4]). We recall that $\mathbf{C}_q[N]$ is an associative algebra with unit over $\mathbf{Q}(q)$ generated by the elements $x_1, \ldots, x_r$ subject to the relations:

$$x_i x_j = x_j x_i \text{ for } |i-j| > 1,$$

$$x_i^2 x_j - (q + q^{-1}) x_i x_j x_i + x_j x_i^2 = 0 \text{ for } |i-j| = 1.$$

This algebra has a distinguished basis B, the *dual canonical basis* (it is dual to Lusztig's canonical basis constructed in [14]). It follows from the results in [14] that B is invariant under the involutive antiautomorphism $b \mapsto b^*$ of $\mathbf{C}_q[N]$ such that $x_i^* = x_i$ for all i. As shown in [4], there exists a natural labeling $\mathbf{m} \mapsto b(\mathbf{m})$ of B by multisegments such that $b(\mathbf{m})^* = b(\zeta(\mathbf{m}))$ for every $\mathbf{m} \in \Sigma$.

Recently in [13], a duality similar to the multisegment duality was introduced and studied; it involves affine Hecke algebras at roots of unity, modular representations of the groups GL_n over p-adic fields, and Kashiwara's crystal bases for affine Lie algebras.

5. MŒGLIN–WALDSPURGER RULE. We now turn to a more detailed discussion of combinatorial properties and connections of the multisegment duality ζ. We start with a recursive description of ζ given in [17].

Take any nonzero multisegment $\mathbf{m} = \sum_{(i,j)\in\Sigma} m_{ij}[i,j]$. Let k be the minimal index such that $m_{kj} \neq 0$ for some j. Define the sequence of indices $j_1, j_2, \dots, j_p$ as follows:

$$j_1 = \min\{j : m_{kj} \neq 0\},\ j_{t+1} = \min\{j : j > j_t, m_{k+t,j} \neq 0\}\ (t = 1, \dots, p-1)\ .$$

The sequence terminates when j_{p+1} does not exist: that is, when $m_{k+p,j} = 0$ for $j_p < j \leq r$. We associate to $\mathbf{m}$ the multisegment $\mathbf{m}'$ given by

$$\mathbf{m}' = \mathbf{m} + \sum_{t=1}^{p}([k+t, j_t] - [k+t-1, j_t]) \tag{3}$$

(with the convention $[i,j] = 0$ unless $1 \leq i \leq j \leq r$). The *Mœglin–Waldspurger rule* states that

$$\zeta(\mathbf{m}) = \zeta(\mathbf{m}') + [k, k+p-1]\ . \tag{4}$$

Setting $|\mathbf{m}| := \sum(j+1-i)m_{ij} \in \mathbb{N}$, we see that $|\mathbf{m}'| = |\mathbf{m}| - p < |\mathbf{m}|$ for any nonzero multisegment $\mathbf{m}$; thus (4) (combined with $\zeta(0) = 0$) indeed provides a recursive description of ζ.

6. RELATIONS WITH PLACTIC MONOID. We now give a new combinatorial interpretation of the multisegment duality as an intertwining map between two piecewise-linear actions of the Lascoux-Schützenberger plactic monoid [12]. Let Pl_r denote the plactic monoid on $r+1$ letters. By definition, Pl_r is an associative monoid with unit generated by $r+1$ elements $p_1, p_2, \dots, p_{r+1}$ subject to the relations

$$p_j p_i p_k = p_j p_k p_i,\ p_i p_k p_j = p_k p_i p_j\ (1 \leq i < j < k \leq r+1)\ ,$$

$$p_j p_i p_j = p_j^2 p_i,\ p_i p_j p_i = p_j p_i^2\ (1 \leq i < j \leq r+1)$$

(sometimes called the *Knuth relations*). As shown by A. Lascoux and M.-P. Schützenberger, this structure provides a natural algebraic framework for the study of Young tableaux and symmetric polynomials.

We now define two right actions of Pl_r on $\mathbb{N}\Sigma$, which we shall denote $(\mathbf{m}, p) \mapsto \mathbf{m}\cdot p$ and $(\mathbf{m}, p) \mapsto \mathbf{m} * p$, respectively. Given a multisegment $\mathbf{m} = \sum_{(i,j)\in\Sigma} m_{ij}[i,j]$ and an index $k \in [1, r+1]$, the multisegments $\mathbf{m}\cdot p_k$ and $\mathbf{m} * p_k$ are defined as follows.

To define $\mathbf{m}\cdot p_k$, let $j_1, j_2, \dots, j_p$ be a sequence of indices given recursively as follows:

$$j_1 = k-1,\ j_{t+1} = \min\{j : j_t < j \leq r,\ m_{tj} > 0\}\ (t = 1, \dots, p-1)\ .$$

The sequence terminates when the set under the minimum sign becomes empty: that is, when $m_{pj} = 0$ for $j_p < j \le r$. Now we set

$$\mathbf{m} \cdot p_k = \mathbf{m} + \sum_{t=1}^{p} ([t, j_t] - [t-1, j_t]) \ . \tag{5}$$

To define $\mathbf{m} * p_k$, we construct recursively two sequences of indices $c_0, c_1, \dots, c_p$ and $i_1, i_2, \dots, i_{p+1}$:

$$c_0 = r, \ i_1 = k; \ c_t = \max\left(\{c : 0 \le c < c_{t-1}, \, m_{i_t, i_t + c} > 0\} \cup \{-1\}\right) ,$$

$$i_{t+1} = \max\left(\{i : 1 \le i < i_t, \, m_{i, i+c_t} = 0\} \cup \{0\}\right) \ (t = 1, \dots, p) \ .$$

The process terminates when $i_{p+1} = 0$. Now we define

$$\mathbf{m} * p_k = \mathbf{m} + \sum_{t=1}^{p} \sum_{i_{t+1} < i \le i_t} ([i-1, i + c_t] - [i, i + c_t]) \ . \tag{6}$$

THEOREM 2. *(a) Each of the correspondences given by (5) and (6) extends by associativity to a right action of* Pl_r *on* $\mathbb{N}\Sigma$.
(b) Each of the two actions in (a) is transitive: i.e., for every two multisegments $\mathbf{m}_1$ *and* $\mathbf{m}_2$, *there exist* $p, p' \in \mathrm{Pl}_r$ *such that* $\mathbf{m}_2 = \mathbf{m}_1 \cdot p = \mathbf{m}_1 * p'$.
(c) The multisegment duality ζ *intertwines the two actions:* $\zeta(\mathbf{m} \cdot p) = \zeta(\mathbf{m}) * p$ *for any multisegment* $\mathbf{m}$ *and any* $p \in \mathrm{Pl}_r$.

In view of part (b), ζ is uniquely determined by the intertwining property (c) combined with the normalization $\zeta(0) = 0$. The following proposition, a direct consequence of the definitions, shows that the Mœglin–Waldspurger rule (4) is a special case of Theorem 2 (c).

PROPOSITION 3. *Let* $\mathbf{m}$ *be a nonzero multisegment. Suppose* k *is the minimal index such that* $m_{kj} \ne 0$ *for some* j, *and* l *is the maximal index such that* $m_{kl} \ne 0$. *Then* $\mathbf{m} \cdot (p_k p_{k-1} \cdots p_1)$ *is the multisegment* $\mathbf{m}'$ *in (3), while* $\mathbf{m} * (p_k p_{k-1} \cdots p_1) = \mathbf{m} - [k, l]$.

The idea to relate the multisegment duality with the plactic monoid was suggested to the author by M.-P. Schützenberger during the author's visit to Université de Marne-la-Vallée in May-June 1994. Theorem 2 was proved soon after, but never published.

7. SCHÜTZENBERGER INVOLUTION. Let us now explore the relation between the multisegment duality and the Schützenberger involution on Young tableaux. We need some terminology and notation related to tableaux. Let $\lambda = (\lambda_1 \ge \cdots \ge \lambda_r \ge 0)$ be a partition of length $\le r$. We identify λ with its *diagram* (denoted by the same letter)

$$\lambda = \{(i, j) \in \mathbf{Z} \times \mathbf{Z} : 1 \le i \le r, \ 1 \le j \le \lambda_i\} \ .$$

An *A_r-tableau of shape* λ is a map $\tau : \lambda \to [1, r+1]$ satisfying

$$\tau(i, j+1) \ge \tau(i, j), \ \tau(i+1, j) > \tau(i, j)$$

for all $(i, j) \in \lambda$ (with the convention that $\tau(i, j) = +\infty$ for $i > r$ or $j > \lambda_i$). The *Schützenberger involution* $\tau \mapsto \eta(\tau)$ (also known as the *evacuation involution*) is

an involution on the set of A_r-tableaux of shape λ which can be defined recursively as follows (cf. [19]).

To any A_r-tableau τ of shape λ we associate a sequence of entries $(i_1, j_1), \dots, (i_p, j_p) \in \lambda$ in the following way. We set $(i_1, j_1) = (1, 1)$ and

$$(i_{t+1}, j_{t+1}) = \begin{cases} (i_t, j_t + 1) & \text{if } \tau(i_t, j_t + 1) < \tau(i_t + 1, j_t) \ ; \\ (i_t + 1, j_t) & \text{if } \tau(i_t, j_t + 1) \geq \tau(i_t + 1, j_t) \ . \end{cases}$$

The sequence terminates at a corner point $(i_p, j_p) \in \lambda$, i.e., when none of (i_p+1, j_p) and (i_p, j_p+1) belong to λ. Now we set $\lambda' = \lambda - \{(i_p, j_p)\}$ and consider the tableau τ' of shape λ' obtained from τ by changing the values at $(i_1, j_1), \dots, (i_{p-1}, j_{p-1})$ according to $\tau'(i_t, j_t) = \tau(i_{t+1}, j_{t+1})$. The tableau $\eta(\tau)$ is defined recursively as the tableau $\eta(\tau')$ of shape λ' extended to a tableau of shape λ by setting $\eta(\tau)(i_p, j_p) = r + 2 - \tau(1, 1)$.

There are (at least) two natural ways to encode tableaux by multisegments: to each tableau $\tau : \lambda \to [1, r + 1]$ we associate two multisegments $\partial(\tau)$ and $\partial'(\tau)$ given by

$$\partial(\tau)_{ij} = \#\{s : \tau(i, s) = j + 1\}, \ \partial'(\tau)_{ij} = \#\{s : \tau(i, s) \leq j, \ \tau(i + 1, s) \geq j + 2\} \ .$$

For a given shape λ, a tableau τ is uniquely recovered from each of the multisegments $\partial(\tau)$ and $\partial'(\tau)$. More precisely, the correspondence $\tau \mapsto \partial(\tau)$ is a bijection between the set of all A_r-tableaux of shape λ and the set of multisegments $\mathbf{m}$ satisfying

$$\sum_{k=j}^{r} (m_{i,k} - m_{i+1,k+1}) \leq \lambda_i - \lambda_{i+1} \quad (1 \leq i \leq j \leq r) \ ;$$

and the multisegments $\mathbf{m} = \partial(\tau)$ and $\mathbf{m}' = \partial'(\tau)$ are related as follows:

$$m'_{ij} = \lambda_i - \lambda_{i+1} - \sum_{k=j}^{r} (m_{i,k} - m_{i+1,k+1}) \ ;$$

$$m_{ij} = \lambda_{r-j+i} - \lambda_{r-j+i+1} - \sum_{k=j}^{r} (m'_{k-j+i,k} - m'_{k-j+i,k+1}) \ .$$

The relationship between the Schützenberger involution η and the multisegment duality ζ is now given as follows.

THEOREM 4. *For every tableau τ, the multisegment $\partial'(\eta(\tau))$ is obtained from $\zeta(\partial(\tau))$ by the following permutation of indices: $\partial'(\eta(\tau))_{j-i+1,r-i+1} = \zeta(\partial(\tau))_{ij}$.*

Theorem 4 was formulated in [11] and proved in [4]; the proof uses some properties of canonical bases, and an equivalent definition of the Schützenberger involution in terms of the so-called Bender-Knuth operators (this definition is due to Gansner [9]).

8. LUSZTIG'S TRANSITION MAPS AND TOTAL POSITIVITY. We now show that the multisegment duality is a special case of Lusztig's piecewise-linear transition maps between various parametrizations of the (dual) canonical basis B. This will require some terminology.

Recall that $\Sigma = \Sigma_r$ stands for the set of all segments $[i,j] \subset [1,r]$. We say that a triple of distinct segments is *dependent* if one of these segments is the disjoint union of two remaining ones; the largest segment in a dependent triple will be called the *support* of the triple, and two remaining ones the *summands*. Let $\nu = (\nu_1, \dots, \nu_m)$ be a total ordering of Σ; here $m = r(r+1)/2$, the cardinality of Σ. We say that ν is *normal* if the support of every dependent triple of segments lies between its summands. The bijection between segments and positive roots given in Section 2 identifies normal orderings of Σ with the well-known normal orderings of positive roots; thus normal orderings are in natural bijection with reduced words for w_0, the longest permutation in the symmetric group S_{r+1} (see e.g., [3, Proposition 2.3.1]). Two examples: in the lexicographic normal ordering $\nu_{\min}$ a segment $[i,j]$ precedes $[i',j']$ if $i < i'$ or $i = i'$, $j < j'$; the reverse lexicographic normal ordering $\nu_{\max}$ is obtained from $\nu_{\min}$ by replacing each segment $[i,j]$ with $[r+1-j, r+1-i]$.

Now consider the dual canonical basis B in $\mathbf{C}_q[N]$ (see Section 4). Translating results of [15, 16] (see also [3]) into the language of segments, we see that every normal ordering ν of Σ gives rise to a bijective parametrization $b_\nu : \mathbb{N}\Sigma \to B$. (In particular, $b_{\nu_{\min}}$ is the parametrization $\mathbf{m} \to b(\mathbf{m})$ discussed in Section 4.) For any two normal orderings ν and ν', Lusztig's *transition map* between ν and ν' is a bijection $R_\nu^{\nu'}$ of $\mathbb{N}\Sigma$ onto itself given by

$$R_\nu^{\nu'} = b_{\nu'}^{-1} \circ b_\nu \ . \tag{7}$$

The multisegment duality turns out to be one of these maps (see [3, Theorem 4.2.2 and Remark 4.2.3]):

$$\zeta = R_{\nu_{\min}}^{\nu_{\max}} \ . \tag{8}$$

In [3], closed formulas for the transition maps $R_\nu^{\nu'}$ were obtained using a parallelism discovered by Lusztig [16] between canonical bases and total positivity. In particular, a new proof of Theorem 1 was obtained. We conclude with a brief discussion of the ideas and methods used in [3].

Clearly, the set of all normal orderings of Σ is closed under the following elementary moves:

2-*move.* In a normal ordering ν, interchange two consecutive (with respect to ν) segments provided they do not belong to a dependent triple.

3-*move.* Interchange the summands of a dependent triple that occupies three consecutive positions in ν.

As a consequence of the corresponding well-known property of reduced words, every two normal orderings of Σ can be obtained from each other by a sequence of 2- and 3-moves. It follows that every transition map can be expressed as a composition of "elementary" transition maps $R_\nu^{\nu'}$ for pairs (ν, ν') related by a 2- or 3-move. These elementary transition maps were computed by Lusztig in [15]. Translated into the language of multisegments they take the following form:

- if ν and ν' are related by a 2-move then $R_\nu^{\nu'}$ is the identity map;
- if ν' is obtained from ν by a 3-move

$$\cdots \alpha, \alpha \cup \beta, \beta \cdots \to \cdots \beta, \alpha \cup \beta, \alpha \cdots$$

then the only components of the multisegment $\mathbf{m}' = R_\nu^{\nu'}(\mathbf{m})$ different from the corresponding components of $\mathbf{m}$ are

$$m'_\alpha = m_\alpha + m_{\alpha\cup\beta} - \min\,(m_\alpha, m_\beta),\ m'_{\alpha\cup\beta} = \min\,(m_\alpha, m_\beta)\ , \tag{9}$$
$$m'_\beta = m_\beta + m_{\alpha\cup\beta} - \min\,(m_\alpha, m_\beta)\ .$$

The key observation now is as follows: the piecewise-linear expressions that appear in (9) can be interpreted as *rational* expressions if one uses an exotic "semifield" structure on $\mathbf{Z}$, where the usual addition plays the role of multiplication, and taking the minimum plays the role of addition. The semifield $(\mathbf{Z}, \min, +)$ is known under various names. We use the term *tropical semifield*, which we learned from M.-P. Schützenberger. A detailed study of its algebraic properties, along with numerous applications, can be found in [2].

The "rational" version of (9) takes the form

$$m'_\alpha = \frac{m_\alpha m_{\alpha\cup\beta}}{m_\alpha + m_\beta},\ m'_{\alpha\cup\beta} = m_\alpha + m_\beta,\ m'_\beta = \frac{m_\beta m_{\alpha\cup\beta}}{m_\alpha + m_\beta}\ . \tag{10}$$

We use this version to define *rational transition maps* $R_\nu^{\nu'} : \mathbf{R}_{>0}\Sigma \to \mathbf{R}_{>0}\Sigma$; here the components m_{ij} of multisegments can be any positive real numbers, and the algebraic operations in (10) are understood in the most common sense. It is not hard to show that a closed formula for some rational transition map $R_\nu^{\nu'}$ would imply such a formula for the corresponding piecewise-linear transition map, by simply translating it into the tropical language; the only caveat is that the formula in question must be *subtraction-free* because the tropical structure does not allow subtraction.

This is precisely the method used in [3]. To compute rational transition maps, we use the observation (due to Lusztig) that they have another interpretation parallel to that in (7). Namely, they describe the relationships between different parametrizations of the variety $N_{>0}$ of all totally positive unipotent upper triangular matrices (recall that a matrix $x \in N$ is *totally positive* if all the minors that do not identically vanish on N take positive values at x). We refer the reader to [3] for the details; let us only mention that the computations in [3] are based on algebraic and geometric study of totally positive varieties. This study is put into a much more general context in [5, 6].

References

[1] A.-M. Aubert, Dualité dans le groupe de Grothendieck de la catégorie des representations lisses de longueur finie d'un groupe réductif p-adique, *Trans. A.M.S.* 347 (1995), 2179–2189; Erratum: *Trans. A.M.S.* 348 (1996), 4687–4690.

[2] F. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat, *Synchronization and linearity*, Wiley, 1992.

[3] A. Berenstein, S. Fomin and A. Zelevinsky, Parametrizations of canonical bases and totally positive matrices, *Adv. in Math.* 122 (1996), 49–149.

[4] A. Berenstein and A. Zelevinsky, Canonical bases for the quantum group of type A_r and piecewise-linear combinatorics, *Duke Math. J.* 82 (1996), 473–502.

[5] A. Berenstein and A. Zelevinsky, Total positivity in Schubert varieties, *Comment. Math. Helv.* 72 (1997), 128–166.

[6] S. Fomin and A. Zelevinsky, Double Bruhat cells and total positivity, Prépublication 1998/08, IRMA, Université Louis Pasteur, Strasbourg, 1998.

[7] L.R. Ford and D. R. Fulkerson, Flows in Networks, Princeton University Press, 1962.

[8] P. Gabriel, Unzerlegbare darstellungen I, *Manuscripta Math.* 6 (1972), 71–103.

[9] E. R. Gansner, On the equality of two plane partition correspondences, *Discrete Math.* 30 (1980), 121–132.

[10] M. Kashiwara, On crystal bases of the q-analogue of universal enveloping algebras, *Duke Math. J.* 63 (1991), 465–516.

[11] H. Knight and A. Zelevinsky, Representations of quivers of type A and the multisegment duality, *Adv. in Math.* 117 (1996), 273–293.

[12] A. Lascoux and M.-P. Schützenberger, Le monoide plaxique, in: *Noncommutative structures in algebra and geometric combinatorics (Naples, 1978),* pp. 129–156, Quad. "Ricerca Sci." 109, CNR, Rome, 1981.

[13] B. Leclerc, J.-Y. Thibon and E. Vasserot, Zelevinsky's involution at roots of unity, E-print math. archive QA/9806060, June 1998.

[14] G. Lusztig, Canonical basis arising from quantised enveloping algebras, *J. Amer. Math. Soc.* 3 (1990), 447–498.

[15] G. Lusztig, *Introduction to quantum groups*, Progress in Mathematics 110, Birkhäuser, 1993.

[16] G. Lusztig, Total positivity in reductive groups, in: *Lie theory and geometry: in honor of Bertram Kostant, Progress in Mathematics* 123, Birkhäuser, 1994.

[17] C. Mœglin and J. L. Waldspurger, Sur l'involution de Zelevinski, *J. Reine Angew. Math.* 372 (1986), 136–177.

[18] S. Poljak, Maximum rank of powers of a matrix of a given pattern, *Proc. A.M.S.* 106 (1989), 1137–1144.

[19] M.-P. Schützenberger, Promotion des morphismes d'ensembles ordonnés, *Discrete Math.* 2 (1972), 73–94.

[20] A. Zelevinsky, Induced representations of reductive p-adic groups II, *Ann. Sci. E.N.S.* 13 (1980), 165–210.

[21] A. Zelevinsky, A p-adic analog of the Kazhdan-Lusztig conjecture, *Funct. Anal. Appl.* 15 (1981), 83–92.

Andrei Zelevinsky
Department of Mathematics
Northeastern University
Boston, MA 02115, USA
andrei@neu.edu

Section 14

Mathematical Aspects of Computer Science

In case of several authors, Invited Speakers are marked with a *.

Worst-Case Complexity, Average-Case Complexity and Lattice Problems

Miklós Ajtai

Abstract. There is a need both from a theoretical and from a practical point of view to create computational problems (in NP) that are hard (that is, they have no polynomial time solutions). Currently there are no methods to prove that such problemx exist at all. We may assume however as an axiom, that certain problems are hard, where the choice of the problems may have historical or theoretical motivations. These problems however are usually worst-case problems, while, e.g. for cryptographic application, we need hard average-case problems. In this paper we desrcibe two different average-case problems, and their cryptographic applications, which are at least as difficult as some well-known worst-case problems concerning lattices.

1991 Mathematics Subject Classification: lattice, worst-case, average-case, complexity, basis, shortest vector 68Q15
Keywords and Phrases: lattice, worst-case, average-case, complexity, basis, shortest vector

1. Introduction. The goal of complexity theory is to describe the necessary resources, in terms of time, memory etc. for the solutions of computational problems. For cryptographic applications it would be particularly important to know that certain problems (e.g. finding the prime factors of a large integer) cannot be solved in a reasonable amount of time. (In fact the popular RSA public-key cryptosystem is based on that assumption.) Unfortunately we do not have yet any results of this type. Still we may get some information about the (relative) hardness of such problems if we accept as an axiom the hardness of a well-known computational problem which was attacked for a long time by many mathematicians without success (that is, we accept that there is no polynomial time solution of the problem in the size of the input) and prove from this axiom the hardness of other problems or the security of a cryptographic protocol. E.g. factoring integers, finding the discrete logarithm can be such problems.

Another similar solution would be to accept as an axiom that there is a problem in NP (that is a problem where the correctness of a proposed solution can be checked in polynomial time) which has no polynomial time solution. This is the famous $P \neq NP$ conjecture. There are known problems (namely each NP-complete problem) whose hardness follows from this assumption. E.g. "find a Hamilton-cycle in a given graph" is an NP-complete problem.

Unfortunately these methods (either we choose the hardness of a famous problem or an NP-complete problem as an axiom) has an inherent limitation. The problems in both categories are so-called worst-case problems. That is, they are hard in the sense that finding a solution is assumed to be difficult only for some unknown values of the input and can be very easy for other values. E.g. there are integers whose factors are very easy to find. For cryptographic applications we have to present a hard instance of the problem, that is, a particular input where the solution cannot be found easily. The practical solution in the case of factoring e.g. is to pick the integer as the product of two random primes with some additional constraint to make sure that factoring is not made easier by the specific structure of the prime factors. That is, we use an average-case problem instead of a worst-case problem. We assume now that this problem whose input is chosen at random is difficult on the average (or for almost all choice of the input). We gave up however our original requirement namely that we use only simply stated and well-studied problems. The algorithmic theory of average-case problems are ususally only a few decades long, while the history of certain worst-case problems go back for hundreds of years. The statement of an average-case problem is also generally less clear-cut because of the many possible choices of the parameters involved in the randomization. In the case of the Hamilton cycle problem it is not even clear what would be a good randomization.

There is however a possiblitiy which unites the advantages of the two (worst-case average-case) methods. Namely we need an average-case problem which is just as difficult as a well-known worst-case problem. There are two different worst-case problems concerning short vectors for lattices which has been used recently in this way to create average-case problems which are at least as difficult as the original worst-case problems and can be used for various cryptographic purposes. It is important that individual random instances of these average-case problems can be created together with a known solution. To formulate these problems we need some basic defintions about lattices.

A lattice is a subset of the n-dimensional space $\mathbf{R}^n$ over the reals which consist of the integer linear combinations of n fixed linerly independent vectors. Such a set of vectors will be called a basis of the lattice. The history of finding short vectors in lattices goes back to the works of Gauss and Dirichlet. With the fundamental results of Minkowski about a hundred years ago the theory of lattices became a separate branch of number theory with a huge literature. Finding short vectors in a lattice (in various possible senses) was always one of the main goals of this theory. (The reader may find more information about lattices e.g. in [6] and [11]. The more modern algorithmic theory of lattices is described in [12].)

The two mentioned worst-case problems are the following:

(**P1**) *Find a basis $b_1, ..., b_n$ in the n-dimensional lattice L whose length, defined as $\max_{i=1}^{n} \|b_i\|$, is the smallest possible up to a factor of n^c, where c is constant.*

(**P2**) *Find the shortest nonzero vector in an n dimensional lattice L where the shortest vector v is unique in the sense that any other vector whose length is at most $n^c\|v\|$ is parallel to v, where c is a sufficiently large absolute constant.*

Problem (**P1**) is equivalent to the problem of finding a single vector shorter

than a given number in a class of randomly generated lattices, with a positive probability (see Ajtai [2]). Therefore finding a short vector in a random lattice is just as difficult (with high probability) as finding a short basis in the worst-case. This random construction also gives a one-way function which leads to some cryptographic tools like pseudo-random number generators. A different cryptographic application namely a collision-free hash function were given by Goldreich, Goldwasser and Halevi in [9] . Problem (**P2**) is somewhat weaker than Problem (**P1**), but it seems to be more easily applicable for cryptographic protocols, e.g. its hardness guarantees the security of a public-key cryptosystem (see Ajtai and Dwork [4]). Another completely different public-key cryptosystem based on the hardness of lattice problems (without worst-case average-case connection) was proposed by Goldreich, Goldwasser and Halevi (see [10]).

2. The construction of a random lattice. In this section we describe a way to generate random n-dimensional lattices so that, if we can find, with a positive probablitiy and in polynomial time a short vector in the random lattice Λ (where the probability is taken for the generation of Λ), then the worst-case problems (**P1**) and (**P2**) can be solved in polynomial time. (This assumption also implies that it ia possible to approximate the length of the shortest vector in an arbitrary lattice up to a polynomial factor in polynomial time. This is, again a worst-case problem.) The proofs of the results described in this section can be found in [2].

The definition of the random class. The definition of the lattices will depend on a parameter n. n can be any positive integer. (The meaning of n is the following: if it is possible to find a short vector easily in the random lattice generated with parameter n, then the n dimensional worst-case problem (**P1**) have a polynomial time solution.) The dimension of the random lattice will be somewhat larger, about $cn \log n$ for some constant c. The lattices in the random class will be subsets of $\mathbf{Z}^m$, that is, they will contain only vectors with integer coordinates. (m will be defined later as a function of n). We will fix an integer q as well (it will be also a function of n) and the lattices will be defined in a way that the fact whether a vector belongs to the lattice or not will depend only on the modulo q residue classes of its coordinates.

Assume that the positive integers n, m and q are fixed, for the moment in an arbitrary way, and $\nu = \langle u_1, ..., u_m \rangle$ where $u_1, ..., u_m \in \mathbf{Z}^n$ is an arbitrary sequence of length m from the elements of $\mathbf{Z}^n$. We define a latice $\Lambda(\nu, q)$ in the following way: $\Lambda(\nu, q)$ will consist of all sequences $\langle h_1, ..., h_m \rangle$ of integers of length m with the property: $\sum_{i=1}^{m} h_i u_i \equiv 0 \pmod{q}$.

Our definition of the random class will depend on the choice of two absolute constant c_1 and c_2. Assume that n is fixed let $m = [c_1 n \log n]$ and $q = [n^{c_2}]$. For each n we will give a single random variable λ so that $\Lambda = \Lambda(\lambda, q)$ is a lattice with dimension m. (The existence of a polynomial time algorithm which finds a short vector in Λ will imply the existence of such an algorithm which solves the mentioned problems in every lattice $L \subseteq \mathbf{R}^n$.)

First we define an "idealized" version λ' of λ, which we can define in a simpler

way. The disadvantage of λ' is that we do not know how to generate λ' together with short vector in $\Lambda(\lambda', q)$. Then we define λ (in a somewhat more complicated way) so that we can generate it together with a short vector in $\Lambda(\lambda, q)$ and we will also have that $P(\lambda \neq \lambda')$ is exponentially small. This last inequality implies that if we prove our theorem for $\Lambda(\lambda', q)$ then it will automatically hold for $\Lambda(\lambda, q)$ too.

Let $\lambda' = \langle v_1, ..., v_m \rangle$ where $v_1, ..., v_m$ are chosen independently and with uniform distribution from the set of all vectors $\langle x_1, ..., x_n \rangle$ where $x_1, ..., x_n$ are integers and $0 \leq x_i < q$. To find a short vector in the lattice $\Lambda(\lambda', q)$ is equivalent of finding a solution for a linear simultaneous Diophantine approximation problem. Dirichlet's theorem implies that if c_1 is sufficiently large with respect to c_2 then there is always a vector shorter than n. (The proof of Dirichlet's theorem is not constructive, it is based on the Pigeonhole Principle applied to a set of exponential size.)

Definition of λ. We randomize the vectors $v_1, ..., v_{m-1}$ independently and with uniform distribution on the set of all vectors $\langle x_1, ..., x_n \rangle \in \mathbf{Z}^n$, with $0 \leq x_i < q$. Independently of this randomization we also randomize a 0, 1-sequence $\delta_1, ..., \delta_{m-1}$ where the numbers δ_i are chosen independently and with uniform distribution from $\{0, 1\}$. We define v_m by $v_m \equiv -\sum_{i=1}^{m-1} \delta_i v_i \pmod{q}$ with the additional constraint that every component of v_m is an integer in the interval $[0, q-1]$. Let $\lambda = \langle v_1, ..., v_m \rangle$. (If we want to emphasize the dependence of λ on n, c_1, c_2 then we will write λ_{n,c_1,c_2}.) It is possible to prove that the distribution of λ is exponentially close to the uniform distribution in the sense that $\sum_{a \in A} |P(\lambda = a) - |A|^{-1}| \leq 2^{-cn}$, where A is the set of possible values of λ. This will imply that the random variable λ' with the given distribution can be chosen in a way that $P(\lambda' \neq \lambda)$ is exponentially small.

With this definition our theorem will be formulated in the following way: "if there is an algorithm which finds a short vector in $\Lambda(\lambda, q)$ given λ as an input, then etc." That is, we allow the algorithm whose existence is assumed in the theorem to use λ.

Definitions. 1. If v is a shortest nonzero vector in the lattice $L \subseteq \mathbf{R}^n$, and $\alpha > 1$, we say that v is α-unique if for any $w \in L$, $\|w\| \leq \alpha \|v\|$ implies that v and w are parallel.

2. If k is an integer then size(k) will denote the number of bits in the binary representation of k, (size$(0) = 1$). If $v = \langle x_1, ..., x_n \rangle \in \mathbf{Z}^n$ then size$(v) = \sum_{i=1}^n$ size(x_i). Our definition implies that for all $v \in \mathbf{Z}^n$, size$(v) \geq n$.

THEOREM . *There are absolute constants c_1, c_2, c_3 so that the following holds. Suppose that there is a probabilistic polynomial time algorithm $\mathcal{A}$ which given a value of the random variable λ_{n,c_1,c_2} as an input, with a probability of at least $1/2$ outputs a vector of $\Lambda(\lambda_{n,c_1,c_2}, [n^{c_2}])$ of length at most n. Then, there is a probabilistic algorithm $\mathcal{B}$ with the following properties. If the linearly independent vectors $a_1, ..., a_n \in \mathbf{Z}^n$ are given as an input, then $\mathcal{B}$, in time polynomial in $\sigma = \sum_{i=1}^n$ size(a_i), gives the outputs z, u, $\langle d_1, ..., d_n \rangle$ so that, with a probability of greater than $1 - 2^{-\sigma}$, the following three requirements are met:*

(1.1) if v is a shortest non-zero vector in $L(a_1, ..., a_n)$ then $z \leq \|v\| \leq n^{c_3} z$

(1.2) *if v is an n^{c_3}-unique shortest nonzero vector in $L(a_1, ..., a_n)$ then $u = v$ or $u = -v$*

(1.3) *$d_1, ..., d_n$ is a basis with $\max_{i=1}^{n} \|d_i\| \leq n^{c_3} \mathrm{bl}(L)$.*

Remarks. 1. The probability 1/2 in the assumption about $\mathcal{A}$ can be replaced by n^{-c}. This will increase the running time of $\mathcal{B}$ by a factor of at most n^c but does not affect the constants c_1, c_2 and c_3.

2. If we assume that $\mathcal{A}$ produces a vector of length at most $n^{c'}$ for some $c' > 1$ then the theorem remains true but c_1, c_2 and c_3 will depend on c'.

3. A PUBLIC-KEY CRYPTOSYSTEM. The following public-key cryptosystem was constructed by Ajtai and Dwork (see [4]). It is secure if problem (**P2**) has no polynomial time solution. Here we give only a very high level and somewhat simplified description of the cryptosystem and its mathematical background, and we refer the reader to [4] for the exact definitions and proofs.

A public cryptosystem serves an unlimited number of participants. Each of them publishes a public key and keeps a private key. The public key is available for everybody, but the private key is known only for its owner. Assume now that Alice wants to send a message to Bob, who published a public key. (We do not assume that Alice has a public or private key.) Alice, gets Bob's public key from a directory available for everybody. Then, using Bob's private key, she encodes the message and sends it to Bob through an open channel. Bob using his private key is able to decode the message, but without this private key the message cannot be decoded.

The RSA public key cryptosystem (see [14]) for example, fulfils this requirement, provided that each participant B can find a positive integer $m_B = p_B q_B$ where p_B, q_B are primes known to B, but nobody else in the knowledge of the number n_B alone is able to find the primes p_B, q_B. Since there is no known factoring algorithm which can factor in a reasonable amount of time a number n with several hundred digits we may think that the assumption is justified. Notice however that the fact that there is no such algorithm implies only that if B would be able to find the worst possible number n_B then his private key would be safe. However Bob has no way of knowing which is the "worst" number n_B. In practice the pair p_X, q_X is generated at random with some care of avoiding such pairs where the factoring of $p_B q_B$ can be easy. Therefore the assumption used in practice is that a certain (rather complicated) average-case problem is hard (with high probability).

In the cryptosystem described below the assumption is the hardness of the worst-case problem (**P2**). Still it is proved that this assumption implies that with a probability very close to one not a single message can be broken without acces to the private key.

The private key of Bob will be a sequence of equidistant $n-1$ dimensional hyperplanes in the n dimensional real space $\mathbf{R}^n$. More precisely, Bob picks a random vector u_B with uniform distribution from the n dimensional unit ball and

this vector u_B is his private key. For each integer k the set of all $x \in \mathbf{R}^n$ whose inner product with u_B is k forms a hyperplane H_k. The sequence $\langle H_i \rangle$ is the sequence of hyperplanes mentioned at the beginning. Of course Bob has only the vector u_B, we mentioned the hyperplanes only to make it easier to visualize the steps in the protocol.

The public key will be a sequence of vectors $v_1, ..., v_m$, where $m = n^{c_3}$, that Bob picks at random close to the hyperplanes. More precisely assume that Q is a large cube and U is the union of the hyperplanes H_i. Bob first picks vectors $v'_1, ..., v'_m$ independently and with uniform distribution from $Q \cap U$ (with respect to the $n-1$ dimensional Lebesgue measure). Then Bob perturbs these vectors sightly at random so that they remain close to the hyperplanes. (Their distance to the closest hyperplane remains smaller than, say, n^{-8}.) The perturbed vectors are $v_1, ..., v_m$.

It is possible to prove (assuming that (P2) has no polynomial time solution) that the sequence $v_1, .., v_m$ is computationally indistinguishable from a sequence of length m whose elements are picked independently and with uniform distribution from the cube Q. Therefore by making the public key available for anybody, Bob did not give out any information about the hyperplanes.

Knowing the public key Alice is able to generate a sequence of independent random points $x_1, ..., x_i, ... \in \mathbf{R}^n$ with identical distributions and with the following properties:

(1) with high probability x_i is very close to a hypeprlane H_k, more precisely if the distance of neighboring hyperplanes is M then there is a hyperplane H_k so that the distance of x_i and H_k is smaller than $\frac{M}{n^5}$.

(2) the distribution of x_i is computationally indistinguishable in polynomial time, form the uniform distribution on a parallelepiped $\mathcal{P}$, where $\mathcal{P}$ can be computed from the public key, so it is known to everybody.

(This last property will be a consequence of the hardness of problem (**P2**).)

For the moment we accept that Alice has a way of generating such a distribution. Assume now that Alice wants to send a single bit δ to Bob. If $\delta = 0$ then Alice picks a random point y with uniform distribution on the set $\mathcal{P}$, and sends y as the message. If $\delta = 1$ then Alice generates a random point x with the distribution of the points x_i, from the $\mathcal{P}$ and sends x as the message.

Suppose that Bob gets a message z. z is an n-dimensional vector in $\mathcal{P}$. Bob computes the inner product $\alpha = z \cdot u_B$. If α is close to an integer (say closer than $\frac{1}{n^4}$) then Bob knows z is close to a hyperplane therefore he concludes that $\delta = 1$. If the distance of α from the closest integer is greater than $\frac{1}{n^4}$, then Bob concludes that $\delta = 0$. (There is a small probability, about $\frac{1}{n^4}$, that Bob makes the wrong decision.)

Finally we sketch how can Alice generate the points x_i with the required properties. $\mathcal{P}$ will be a parallelepiped determined by n vectors from the sequence $v_1, ..., v_m$, so that the parallelepiped is relatively "fat", that is, the minimal distance between its opposite sides is not too small with respect to the length of a side of Q. (Larger then, say, n^{-2} times this length.) $\mathcal{P}$ may be the first such parallelpiped with this property or Bob can designate a parallelepiped in the public key. With a very high probability such a parallelepiped always exists. Assume

that $\mathcal{P}$ is the parallelepiped determined by the vectors $v_{i_1}, ..., v_{i_n}$

Alice takes a random $0, 1$ linear combination w of the vectors $v_1, ..., v_m$, then reduces it to the parallelepiped $\mathcal{P}$ modulo $v_{i_1}, ..., v_{i_n}$. In other words she adds an integer linear combination of the vectors $v_{i_1}, ..., v_{i_n}$ to the vector w so that the sum x is in $\mathcal{P}$. x has a distribution with the required properties.

4. The NP-hardness of the shortest vector problem. We cannot prove from any reasonable assumption from complexity theory (like $P \neq NP$) that the problems (P1) or (**P2**) are hard. Actually it is unlikely that Problem (**P2**) is NP-hard for $c > \frac{1}{2}$ since it would lead to a collapse in the computational hierarchy (see Goldreich and Goldwasser [8]). However if we drop the uniqueness requirement from the problem, that is, we want to find the shortest vector (under the Euclidean norm) then the problem is NP-hard at least for randomized reductions (see Ajtai [3]). The proof of this result uses lattices constructed from logarithms of small primes. This type of lattice construction was originally used by Adleman ([1]) to reduce factoring to the shortest vector problem (for the proof of correctness he used number theoretical conjectures about the distribution of smooth numbers.) The proof of the NP-hardness result also has a combinatorial part which is a constructive/probabilistic version of Sauer's Lemma (related to the concept VC dimension). This is the most difficult part of the proof.

The NP-hardness of the shortest vector problem was conjectured by Van Emde Boas almost twenty years ago (see [5]). He proved the analogue statement for the L_∞ norm (for deterministic reductions). The shortest vector problem in L_2 is NP-hard even in some approximate sense. Micciancio has proved recently, that the problem "find a vector which is longer than the shortest vector only by a constant factor c, where $c < n^{\frac{1}{2}}$" is NP-hard (see [13]). (The original proof in [3] gave only a factor $1+2^{-n^\epsilon}$ which was improved first by Cai and Nerurkar (see [7]) to $1+n^{-\epsilon}$.) Micciancio also proved that the NP-hardness of the shortest vector problem for deterministic reductions follows from a natural number theoretic conjecture about the existence of square-free smooth numbers in long enough intervals.

References

[1] L. Adleman, Factoring and Lattice Reduction, Manuscript, 1995.

[2] M. Ajtai, Generating Hard Instances of Lattice Problems, Proceedings 28th Annual ACM Symposium on Theory of Computing, 1996 or Electronic Colloquium on Computational Complexity, 1996, http://www.eccc.uni-trier.de/eccc/

[3] M. Ajtai, The Shortest Vector Problem is NP-hard for Randomized Reductions. Proceedings 30th Annual ACM Symposium on Theory of Computing, 1998 or Electronic Colloquium on Computational Complexity, 1997, http://www.eccc.uni-trier.de/eccc/

[4] M. Ajtai and C. Dwork, A Public-Key Cryptosystem with Worst-Case/Average-Case Equivalence, Proceedings 29th Annual ACM Symposium on Theory of Computing, 1997 or Electronic Colloquium on Computational Complexity, 1996, http://www.eccc.uni-trier.de/eccc/

[5] P. Van Emde Boas, Another NP-complete partition problem and the complexity of computing short vectors in a lattice, Tech. Report 81-04, Dept. of Mathematics, Univ. of Amsterdam, 1980.

[6] J.W.S. Cassels, *An Introduction to the Geometry of Numbers*, Springer, 1959

[7] J.-Y. Cai, A. Nerurkar, Approximating SVP to within a factor of $1 + \dim^{\epsilon}$ is NP-hard under randomized reductions, IEEE Conference on Computational Complexity (to appear), see also Electronic Colloquium on Computational Complexity, 1997, http://www.eccc.uni-trier.de/eccc/

[8] O. Goldreich, S. Goldwasser, On the Limits of Non-Approximability of Lattice Problems Electronic Colloquium on Computational Complexity, 1997, http://www.eccc.uni-trier.de/eccc/

[9] O. Goldreich, S. Goldwasser, S. Halevi, Collision-free hashing from lattice problems, Electronic Colloquium, on Computational Complexity, 1996, http://www.eccc.uni-trier.de/eccc/

[10] O. Goldreich, S. Goldwasser, S. Halevi, Public-key cryptosystems from lattice reduction problems, Electronic Colloquium on Computational Complexity, 1996, http://www.eccc.uni-trier.de/eccc/

[11] P.M. Gruber, C.G.Lekkerkerker, *Geometry of Numbers*, North-Holland, 1987

[12] M. Grötschel, L. Lovász, A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer, Algorithms and Combinatorics 2, 1988

[13] D. Micciancio, The Shortest Vector in a Lattice is Hard to Approximate within Some Constant. Electronic Colloquium, on Computational Complexity, 1996, http://www.eccc.uni-trier.de/eccc/

[14] R. Rivest, A. Shamir, L. Adelman, A Method for Obtaining Digital Signatures and Public-Key Cryptosystems, *CACM 21*(2), pp. 120–126, 1978

Miklós Ajtai
IBM Almaden Research Center, K/53
650 Harry Road
San Jose, CA 95120
USA

Games, Complexity Classes, and Approximation Algorithms

Joan Feigenbaum

Abstract. We survey recent results about game-theoretic characterizations of computational complexity classes. We also show how these results are used to prove that certain natural optimization functions are as hard to approximate closely as they are to compute exactly.

1991 Mathematics Subject Classification: 68Q15
Keywords and Phrases: Games, Complexity, Approximation Algorithms, Perfect Information, Perfect Recall

1 Introduction

Game theory provides a framework in which to model and analyze conflict and cooperation among independent decision makers. Many areas of computer science have benefitted from this framework, including artificial intelligence, distributed computing, security and privacy, and lower bounds. Games are particularly important in computational complexity, where they are used to characterize complexity classes, to understand the power and limitations of those classes, and to interpret the complete problems for those classes.

This paper surveys three sets of results in the interplay of games and complexity. First, we present several characterizations (some old, some new) of the complexity class PSPACE that show that it is extremely robustly characterized by zero-sum, perfect-information, polynomial-depth games. Next, we explain how the more recent of these characterizations of PSPACE are used to show that certain natural maximization and minimization functions, drawn from domains such as propositional logic, graph searching, graph reliability, and stochastic optimization, are as hard to approximate closely as they are to compute exactly. Finally, we present some connections between complexity classes and imperfect information games; some tight characterizations of exponential-time classes are known, but no set of imperfect-information games is as robustly identified with any complexity class as zero-sum, perfect-information, polynomial-depth games are with PSPACE.

We assume familiarity with basic computational complexity theory, especially with the complexity classes P, NP, PSPACE, EXP, and NEXP, with the notions of reduction and completeness, and with the concept of an "approximation algorithm" for an NP-hard or PSPACE-hard optimization function. Among the books

that cover this material and are accessible to all mathematically educated readers are those by Garey and Johnson [10] and Papadimitriou [16]. We also assume familiarity with elementary game theory, in particular with the notions of "perfect information" and "perfect recall." The few game-theoretic notions that we use are defined precisely in, *e.g.*, [9]

2 Alternation and Randomized Players

Chandra *et al.* [5] proved a fundamental result about the connection between games and complexity that serves as the starting point for most of the results surveyed in this paper. In the Alternating Polynomial Time computational model, there are two computationally unbounded players P_1 and P_0 and a polynomial-time referee V. There is an input string x written on a common tape readable by P_1, P_0, and V, and the goal of the computation is to determine whether x is in the language L. P_1 claims that $x \in L$, and P_0 claims that $x \notin L$. They "argue" for polynomially many rounds, and then V decides who's right. More precisely, there are two functions m and l such that, on inputs x of length n, P_1 and P_0 take turns for $m(n)$ rounds (P_1 moving in odd rounds and P_0 in even rounds), writing a string of length $l(n)$ in each round. Both $m(n)$ and $l(n)$ are polynomially bounded (abbreviated poly(n)). After the entire "game transcript" of length $m(n) \cdot l(n)$ has been written, V reads it, does a polynomial-time computation, and outputs "ACCEPT" or "REJECT," depending on whether it thinks the winner is P_1 or P_0. For (P_1, P_0, V) to be an Alternating Polynomial Time machine for the language L, it must have the property that, if $x \in L$, V always outputs ACCEPT (*i.e.*,, P_1 has a winning strategy), and, if $x \notin L$, V always outputs REJECT (*i.e.*,, P_0 has a winning strategy). The fundamental result of Chandra *et al.* [5] is that Alternating Polynomial Time is equal to PSPACE: Languages that correspond to zero-sum, perfect-information, polynomial-depth games are exactly those recognizable by Turing Machines that use polynomial space.

The fundamental correspondence between PSPACE and perfect-information games is clearly illustrated by the well-known PSPACE-complete language of true quantified Boolean formulas. Consider quantified Boolean formulas in 3CNF ("three conjunctive normal form"); that is, those of the form

$$\Phi = Q_1 x_1 Q_2 x_2 \ldots Q_n x_n \phi(x_1, x_2, \ldots, x_n),$$

where each $Q_i \in \{\exists, \forall\}$, each x_i is a Boolean variable, and ϕ is a formula in conjunctive normal form, each clause of which has exactly three literals. Let Q3SAT be the set of true quantified formulas in 3CNF. To obtain a perfect-information game, let the variables of the formula be chosen by P_1 and P_0, in order of quantification, where P_0 chooses the universally quantified variables and P_1 chooses the existentially quantified variables. By definition, the formula is true (*i.e.*, in Q3SAT) if and only if P_1 has a winning strategy for this game. The classical paper of Schaefer [18] provides many more examples of PSPACE-complete perfect-information games.

Papadimitriou [17] considers an interesting variation on the Alternating Polynomial Time model. In a "Game Against Nature," the input x is still given to P_1,

who claims that $x \in L$, to P_0, who claims that $x \notin L$, and to the polynomial-time referee V. However, P_0 is now a "random" player; in even-numbered moves of the game, P_0 just tosses fair coins to select a string of the appropriate length uniformly at random. Thus, instead of playing against a strategic opponent, P_1 is playing against "nature." If P_1 is truthful in his claim that $x \in L$, then V must accept with probability at least $1/2$. If $x \notin L$, then V must accept with probability less than $1/2$. (The probability is computed over the coin tosses of P_0.) The main result of [17] is that Games Against Nature recognize exactly the languages in PSPACE – or, at least for perfect-information, polynomial-depth games, playing against "nature" is just as hard for P_1 as playing against an evenly-matched opponent!

Babai and Moran [3] consider "Arthur-Merlin Games." These are defined in the same way as Games Against Nature, except that there must be a "gap" in acceptance probabilities: If $x \in L$, then V must accept with probability at least $2/3$, and, if $x \notin L$, then V must accept with probability at most $1/3$. One of the most highly acclaimed results in computational complexity theory, proved by Lund *et al.* [14] and Shamir [19], is that the (seemingly very stringent) requirement of this $(1/3, 2/3)$ gap does *not* change the class of languages accepted: poly(n)-round Arthur-Merlin Games also recognize exactly PSPACE.

3 Probabilistically Checkable Debate Systems

In the Alternating Polynomial Time, Games Against Nature, and Arthur-Merlin Game models, the referee reads the entire transcript of a played game before deciding the winner. In this Section, we consider models in which the referee reads only a randomly selected subset of the game transcript but can still decide the winner correctly, because the players encode their moves in a clever way that makes refereeing easy. The results obtained are the PSPACE analogue of the *probabilistically checkable proof system* theory developed for NP (see, *e.g.*, [1, 20]).

A *probabilistically checkable debate system* (PCDS) for a language L consists of a player P_1, who claims that the input x is in L, a player P_0, who claims that x is not in L, and a *probabilistic* polynomial-time referee V. The language L is in the complexity class PCD$(r(n), q(n))$ if V flips at most $O(r(n))$ coins on inputs x of length n and reads at most $O(q(n))$ bits of the game transcript produced by P_1 and P_0. On inputs $x \in L$, V always declares P_1 to be the winner, and on inputs $x \notin L$, V declares P_0 to be the winner with probability at least $2/3$. An RPCDS is a PCDS in which player P_0 follows a very simple strategy: In each even round of the game, P_0 simply chooses uniformly at random from the set of all legal moves. The class RPCD$(r(n), q(n))$ is defined by analogy with PCD$(r(n), q(n))$.

The characterizations of PSPACE presented in Section 2 are those in which $r(n) = 0$ and $q(n)$ is an arbitrary polynomial. Specifically, Alternating Polynomial Time is, by definition, PCD(0, poly(n)), and poly(n)-round Arthur-Merlin Games are RPCD(0, poly(n)).

Condon *et al.* [6, 7] study the potential tradeoff between random bits and query bits. If the referee V is allowed to flip coins, might it still be able to determine the winner of the game without reading the entire transcript? The results in [6, 7] show that, as in the PCP characterization of NP, the best possible tradeoff between

$r(n)$ and $q(n)$ is obtainable. Furthermore, this tradeoff is obtainable both when the opponents are two strategic players (a PCDS) and when they are a strategic player and a random player (an RPCDS). Specifically, it is shown in [6, 7] that

$$\text{PSPACE} = \text{PCD}(\log n, 1) = \text{RPCD}(\log n, 1).$$

One surprising aspect of these results is that, while the number of rounds of the game is poly(n), the number of bits of the game examined by the referee is $O(1)$. Thus, most of the moves of *both* players are never looked at, and yet the referee still decides the winner correctly. In order to encode games to permit such efficient refereeing, Condon *et al.* [6, 7] exploit and extend the probabilistically checkable coding techniques developed in the PCP characterization of NP [1, 20],

In conclusion, the results of [5, 6, 7, 14, 17, 19] demonstrate that the identification of PSPACE with zero-sum, perfection-information, polynomial-depth games is extremely robust. Numerous variations on the computational model of a game between two strategic players that is judged after it is played by a polynomial-time referee have been studied, *e.g.,* replacing one strategic player by a random player, putting a sharp threshold between yes-instances and no-instances precisely at acceptance probability 1/2, requiring a $(1/3, 2/3)$ gap in acceptance probability between yes-instances and no-instances, and only allowing the referee to examine a constant number of bits of the played game before making a decision. All of these variations on perfect-information games (and several combinations thereof) cause the same class of languages to be accepted, namely PSPACE. As the results surveyed below in Section 5 demonstrate, there is no complexity class known to be as robustly identifiable with a class of imperfect-information games.

4 Nonapproximability

The game-theoretic characterizations of PSPACE presented in Sections 2 and 3 can be used to prove that many optimization functions that are PSPACE-hard to compute exactly are also PSPACE-hard to approximate closely. This use of the debate-system characterizations in Section 3 was inspired by the use of the PCP$(\log n, 1)$ characterization of NP to prove nonapproximability results for NP-hard optimization functions; see Arora and Lund [2] for an overview of these results on NP.

The basic proof structure of the nonapproximability results surveyed in this Section is as follows. First, a characterization of PSPACE is used directly to show that a particular function F is hard to approximate within a certain factor; then approximability-preserving, polynomial-time reductions are given from F to other functions of interest. Note that these reductions must be constructed with some care, because the mere fact that two optimization problems are equivalent under polynomial-time reductions does not mean that they are equivalent with respect to approximability. A canonical example of a polynomial-time reduction that does not appear to preserve approximability is the one from VERTEX COVER to INDEPENDENT SET (see Section 6.1 of Garey and Johnson [10]).

Throughout this section, we say that "algorithm A approximates the function f within ratio $\epsilon(n)$," for $0 < \epsilon(n) < 1$, if, for all x in the domain of f, $\epsilon(|x|) \leq$

$A(x)/f(x) \leq 1/\epsilon(|x|)$. If $\epsilon(n) > 1$, then "algorithm A approximates the function f within ratio $\epsilon(n)$" means that $1/\epsilon(|x|) \leq A(x)/f(x) \leq \epsilon(|x|)$.

4.1 Reductions from PCD($\log n$, 1)

From the characterization PSPACE = PCD($\log n$, 1), we obtain a nonapproximability result for one optimization version of the PSPACE-complete language Q3SAT defined in Section 2. Let the variables of the formula be chosen by P_0 and P_1, in order of quantification, where P_0 chooses the universally quantified variables and P_1 chooses the existentially quantified variables. If P_1 can guarantee that k clauses of ϕ will be satisfied by the resulting assignment, regardless of what P_0 chooses, we say that k clauses of Φ are *simultaneously satisfiable*. Let MAX Q3SAT be the function that maps a quantified 3CNF formula Φ to the maximum number of simultaneously satisfiable clauses.

Theorem: There is a constant $0 < \epsilon < 1$ such that approximating MAX Q3SAT within ratio ϵ is PSPACE-hard.

Nonapproximability results for other PSPACE-hard functions can now be obtained via approximability-preserving reductions from MAX Q3SAT. The following two are given by Condon *et al.* [6]:

MAX FA-INT: The language FA-INT consists of all sets $\{A_1, A_2, \ldots, A_m\}$ of deterministic finite-state automata having the same input alphabet Σ such that there is a string w that is accepted by all of them. FA-INT plays a key role in the field of "computer-aided verification" of devices and protocols (see, *e.g.*, Kurshan [12]) and was shown to be PSPACE-complete by Kozen [11]. The PSPACE-hard function MAX FA-INT maps each set $\{A_1, A_2, \ldots, A_m\}$ to the largest integer k such that there is a string w accepted by k of the A_i's.

MAX GGEOG: Instances of the game "generalized geography" consist of pairs (G, s), where G is a directed graph and s is a distinguished start node. A marker is initially placed on s, and P_0 and P_1 alternatively play by moving the marker along an arc that goes out of the node it is currently on. Each arc can be used at most once; the first player that is unable to move loses. The language GGEOG consists of all pairs (G, s) for which P_1 has a winning strategy; GGEOG is one of the many perfect-information games shown to be PSPACE-complete by Schaefer [18]. We say that (G, s) "can be played for k rounds" if P_1 has a strategy that causes the marker to be moved along k arcs, no matter what P_0 does, even if P_1 ultimately loses. The PSPACE-hard function MAX GGEOG maps pairs (G, s) to the maximum number of rounds for which they can be played.

In fact, the lower bounds for MAX FA-INT and MAX GGEOG are stronger than the one for MAX Q3SAT: In both cases, there is a constant $\epsilon > 0$ such that approximating the function to within a factor of n^ϵ is PSPACE-hard. Additional nonapproximability results from the domains of modal logic and system specification and analysis are given, respectively, by Lincoln *et al.* [13] and by Marathe *et al.* [15].

4.2 REDUCTIONS FROM RPCD(0, poly(n))

The PSPACE-complete language SSAT is defined as follows by Papadimitriou [17]. An instance is a 3CNF formula ϕ over the set of variables $\{x_1, x_2, \ldots, x_n\}$. The instance is in SSAT if there is a choice of Boolean value for x_1 such that, for a random choice (with True and False each chosen with probability 1/2) for x_2, there is a choice for x_3, etc., for which the probability that ϕ is satisfied is at least 1/2. Think of SSAT as a game between an existential player and a random player; on odd moves i, the existential player chooses an optimal value for x_i (where "optimal" means "maximizes the probability that ϕ will be satisfied") and, on even moves i, the random player chooses a random value for x_i. Yes-instances of SSAT are those in which the existential player wins with probability at least 1/2.

The function MAX-PROB SSAT maps each SSAT instance to the probability that ϕ is satisfied if the existential player plays optimally; so yes-instances of the decision problem are those on which the value of MAX-PROB SSAT is at least 1/2. The proof that PSPACE = RPCD(0, poly(n)) (see Lund *et al.* [14] and Shamir [19]) yields the following strong nonapproximability result.

THEOREM: For any language L in PSPACE and any $\epsilon < 1$, there is a polynomial-time reduction f from L to SSAT such that

$$x \in L \Rightarrow \text{MAX-PROB SSAT}(f(x)) = 1, \text{ and}$$

$$x \notin L \Rightarrow \text{MAX-PROB SSAT}(f(x)) < 2^{-n^{\epsilon}},$$

where n is the number of variables in $f(x)$.

Condon *et al.* [7] and Papadimitriou [17] give approximability-preserving reductions from MAX-PROB SSAT to the following three functions.

MIN DMP: An instance of Dynamic Markov Process (DMP) is a set S of states and an $n \times n$ stochastic matrix P, where $n = |S|$. Associated with each state s_i is a set D_i of decisions, and each $d \in D_i$ is assigned a cost $c(d)$ and a matrix R_d. Each row of R_d must sum to 0, and each entry of $P + R_d$ must be nonnegative. The result of making decision d when the process is in state s_i is that a cost of $c(d)$ is incurred, and the probability of moving to state s_j is the $(i,j)^{th}$ entry of $P + R_d$. A strategy determines which decisions are made over time; an optimal strategy is one that minimizes the expected cost of getting from state s_1 to state s_n. The language DMP, shown to be PSPACE-complete by Papadimitriou [17], consists of tuples $(S, P, \{D_i\}, c, \{R_d\}, B)$ for which there is a strategy with expected cost at most B. The optimization function MIN DMP maps $(S, P, \{D_i\}, c, \{R_d\})$ to the expected cost of an optimal strategy.

COLORING GAMES: An instance of a coloring game (see Bodlaender [4]) consists of a graph $G = (V, E)$, an ownership function o that specifies which of P_0 and P_1 owns each vertex, a linear ordering f on the vertices, and a finite set C of colors. This instance specifies a game in which the players color the vertices in the order specified by the linear ordering. When vertex i is colored, its owner chooses a color from the set of *legal* colors, *i.e.*, those in set C that are *not* colors of the colored neighbors of i. The game ends either when all vertices are colored, or when a player cannot color the next vertex in the linear ordering f because there

are no legal colors. P_1 wins if and only if all vertices are colored at the end of the game. The length of the game is the number of colored vertices at the end of the game. In a stochastic coloring game (SCG), P_0 chooses a color uniformly at random from the set of legal colors at each stage. Two corresponding optimization problems are to maximize the following functions: MAX-PROB SCG(G, o, f, C), which is the maximum probability that P_1 wins the game (G, o, f, C), and MAX-LENGTH SCG(G, o, f, C), which is the maximum expected length of the game. Both maxima are computed over all strategies of P_1.

For each of MIN DMP and MAX-PROB SGC, there is a constant $\epsilon > 0$ such that it is PSPACE-hard to approximate the function within ratio 2^{-n^ϵ}. For MAX-LENGTH SGC, the ratio within which approximation is PSPACE-hard is $n^{-\epsilon}$, for a constant $\epsilon > 0$.

4.3 Reductions from RPCD$(\log n, 1)$

The starting point for this set of nonapproximability results is the function MAX-CLAUSE SSAT, whose value on the 3CNF formula ϕ is the expected number of clauses of ϕ that are satisfied if P_1 chooses the values of the existentially quantified variables, the other variables are assigned random values, and P_1 plays optimally with the goal of maximizing the number of satisfied clauses. Using their result that PSPACE = RPCD$(\log n, 1)$, Condon *et al.* [7] prove the following.

Theorem: There is a constant $0 < \epsilon < 1$ such that approximating MAX-CLAUSE SSAT within ratio ϵ is PSPACE-hard.

They then reduce MAX-CLAUSE SSAT to many other optimization functions, using reductions that preserve approximability. Two examples include:

MAX SGGEOG: Consider the variation of the game GGEOG defined in Section 4.2 in which P_0 plays randomly; that is, at every even-numbered move, P_0 simply chooses an unused arc out of the current node uniformly at random and moves the marker along that arc. The goal of P_1 is still to maximize the length of the game, and the function MAX SGGEOG maps an instance (G, s) to the expected length of the game that is achieved when P_1 follows an optimal strategy.

MAX-PROB DGR: The Graph Reliability problem is defined as follows by Valiant [21]: Given a directed, acyclic graph G, source and sink vertices s and t, and a failure probability $p(v, w)$ for each arc (v, w), what is the probability that there is a path from s to t consisting exclusively of arcs that have not failed? Papadimitriou [17] defines Dynamic Graph Reliability (DGR) as follows: The goal of a strategy is still to traverse the digraph from s to t. Now, however, for each vertex x and arc (v, w), there is a failure probability $p((v, w), x)$; the interpretation is that, if the current vertex is x, the probability that the arc (v, w) will fail before the next move is $p((v, w), x)$. The PSPACE-complete language DGR consists of those digraphs for which there exists a strategy for getting from s to t with probability at least $1/2$. A natural optimization function is MAX-PROB DGR, which maps a graph, vertices s and t, and a set $\{p((v, w), x)\}$ of failure probabilities to the probability of reaching t from s under an optimal strategy.

It is PSPACE-hard to approximate MAX SGGEOG within ratio $n^{-\epsilon}$, for any constant $0 < \epsilon < 1/2$, where n is the number of vertices in the graph. It is also PSPACE-hard to approximate MAX-PROB DGR within ratio $2^{-n^{\epsilon}}$, for some constant $\epsilon > 0$. See Condon *et al.* [7] for proofs of these results and for a related result about a stochastic version of the board game Mah-Jongg.

5 Imperfect Information Games

Feigenbaum *et al.* [9] develop a framework in which to generalize the connections between game classes and complexity classes. A *polynomially definable game system* (PDGS) for a language L consists of two arbitrarily powerful players P_0 and P_1 and a polynomial-time referee V. The referee may be probabilistic, but there are some interesting cases in which V does not need randomness. P_0 and P_1 and the referee V have a common input tape. On input x, P_1 claims that x is in L, P_0 claims that x is not in L, and V's job is to decide which of these two claims is true.

Each input x to a PDGS determines a *polynomially definable game* G_x as follows. The game is essentially run by the referee V. The moves in the game are relayed by the players to V. Neither player sees V's communication with the other, but V can transmit information about the current status of the game to one or both players. This reflects the fact that the players can have imperfect information to varying degrees. When the interaction is finished, V either accepts or rejects x. Because the referee is polynomial-time, G_x lasts for poly(n) moves, and each move can be written down in poly(n) bits, where $n = |x|$. The resulting game G_x clearly defines a two-person, zero-sum game tree in which the length of each path is polynomial. If V is probabilistic, then his coin tosses correspond to chance moves in the game tree.

It is essential to the PDGS framework that P_0 and P_1 use mixed strategies. (See [9, Section 1] for a discussion of why previous attempts to characterize complexity classes with imperfect-information games in which the players use pure strategies were unsatisfactory.) That is, for each possible input x, each player has a probability distribution over the space of his deterministic strategies. At the beginning of the game, the players examine x and independently choose a pure strategy using their respective probability distributions; those pure strategies are then played throughout the game. Since the game tree has exponential size, a pure strategy also has exponential size. An arbitrary mixed strategy could of course have size doubly exponential in n.

There are two ways to define acceptance of a language L by a PDGS (P_1, P_0, V). In the "exact model," yes-instances x correspond to games G_x of value at least $1/2$ and no-instances to games of value less than $1/2$:

- For all $x \in L$, there exists a mixed strategy μ_1 for P_1 such that, for all strategies μ_0 for P_0, V accepts with probability at least $1/2$.

- For all $x \notin L$, there exists a mixed strategy μ_0 for P_0 such that, for all strategies μ_1 for P_1, V accepts with probability less than $1/2$.

In the "approximate model," yes-instances x correspond to games G_x of value at least 2/3 and no-instances to games of value at most 1/3:

- For all $x \in L$, there exists a mixed strategy μ_1 for P_1 such that, for all strategies μ_0 for P_0, V accepts with probability at least 2/3.
- For all $x \notin L$, there exists a mixed strategy μ_0 for P_0 such that, for all strategies μ_1 for P_1, V accepts with probability at most 1/3.

In both models, the probability of acceptance is computed over the pure strategies of both players (if they use mixed strategies) and the coin tosses of V (if any).

The main question addressed in [9] is the relationship between the game-theoretic properties of P_0 and P_1 and the class of languages recognizable by PDGS's. One class of PDGS's studied are those in which at least one player has imperfect information (*i.e.*, those in which the referee V does not tell P_0 everything about its communication with P_1 and/or *vice versa*) but *perfect recall* (*i.e.*, P_0 and/or P_1 has enough memory to record everything they do and everything they receive from V and can use it in subsequent rounds of the protocol). Another class are those in which at least one player has imperfect recall: P_0 or P_1 or both cannot store everything they do and receive and may have to act in the i^{th} round of the game based on partial or no information about what happened in the first $i-1$ rounds.

In the results on PSPACE surveyed in Sections 2 and 3, the computational models are very special cases of PDGS's, in which the referee's role is trivial while the game is being played: V simply sends all information about P_1's current move and the entire history of the game to P_0 and *vice versa*. Therefore these results show that PDGS's in which both players have perfect information recognize exactly PSPACE, both in the exact model and in the approximate model. Feigenbaum *et al.* [9] obtain similarly tight results for PDGS's in which at least one player has imperfect recall. If P_1 has imperfect recall, but P_0 has either perfect information or perfect recall, then PDGS's accept exactly those languages recognizable in nondeterministic exponential time (the complexity class NEXP), in both the exact model and the approximate model. If P_0 is the one with imperfect recall, the class recognized in both models is coNEXP. An almost-tight characterization is obtained for PDGS's in which both players have imperfect recall (see [9] for details).

Feigenbaum *et al.* [9] also proved that, in the exact-value model, the languages accepted by PDGS's in which P_0 and P_1 both have perfect recall (but imperfect information) are exactly those languages recognizable in deterministic exponential time (the complexity class EXP). They left open the question of whether the approximate-value model is equivalent to the exact-value model when both players have perfect recall. This question was subsequently answered by Feige and Kilian [8]: One-round PDGS's with two perfect-recall players accept PSPACE; Polynomial-round PDGS's with two perfect-recall players accept EXP.

Thus, perfect-recall games seem to be fundamentally different perfect-information games and from games in which at least one player has imperfect recall; in particular, whether or not exact refereeing is equivalent to approximate

refereeing seems to depend on the number of rounds of the game. It remains open whether there are natural explanations or generalizations of these results on imperfect information games or whether these results have applications to approximability. Also open is the question of whether there are imperfect-information analogues of the Arthur-Merlin and Games-Against-Nature characterizations of PSPACE; that is, can a random player replace one of the perfect-recall players or one of the imperfect-recall players in a class of PDGS's without changing the language-recognition power of the class.

REFERENCES

[1] S. Arora, "Probabilistic Checking of Proofs and Hardness of Approximation Problems," PhD Thesis, University of California, Computer Science Division, Berkeley, 1994.

[2] S. Arora and C. Lund, "Hardness of Approximations," in APPROXIMATION ALGORITHMS FOR NP-HARD PROBLEMS, D. Hochbaum (ed.), PWS Publishing, Boston, 1997, pp. 399-446.

[3] L. Babai and S. Moran, "Arthur-Merlin Games: A Randomized Proof System and a Hierarchy of Complexity Classes," *Journal of Computer and System Sciences*, 36 (1988), pp. 254-276.

[4] H. Bodlaender, "On the Complexity of Some Coloring Games," *International Journal on Foundations of Computer Science*, 2 (1991), pp. 133-147.

[5] A. Chandra, D. Kozen, and L. Stockmeyer, "Alternation," *Journal of the Association for Computing Machinery*, 28 (1981), pp. 114-133.

[6] A. Condon, J. Feigenbaum, C. Lund, and P. Shor, "Probabilistically Checkable Debate Systems and Nonapproximability Results for PSPACE-Hard Functions," *Chicago J. Theoretical Computer Science*, vol. 1995, no. 4. `http://www.cs.uchicago.edu/publications/cjtcs/articles/-1995/4/contents.html`

[7] A. Condon, J. Feigenbaum, C. Lund, and P. Shor, "Random Debaters and the Hardness of Approximating Stochastic Functions," *SIAM Journal on Computing*, 26 (1997), pp. 369-400.

[8] U. Feige and J. Kilian, "Making Games Short," in *Proceedings of the 29th Symposium on Theory of Computing*, ACM Press, New York, 1997, pp. 506-516.

[9] J. Feigenbaum, D. Koller, and P. Shor, "A Game-Theoretic Classification of Interactive Complexity Classes," in *Proceedings of the 10th Conference on Structure in Complexity Theory*, IEEE Computer Society Press, Los Alamitos, 1995, pp. 227-237.

[10] M. Garey and D. Johnson, COMPUTERS AND INTRACTABILITY: A GUIDE TO THE THEORY OF NP-COMPLETENESS, Freeman, San Francisco, 1979.

[11] D. Kozen, "Lower Bounds for Natural Proof Systems," in *Proceedings of the 18th Symposium on Foundations of Computer Science*, IEEE Computer Society Press, Los Alamitos, 1977, pp. 254-266.

[12] R. Kurshan, Computer-Aided Verification of Coordinating Processes: The Automata-Theoretic Approach, Princeton University Press, Princeton, 1994.

[13] P. Lincoln, J. Mitchell, and A. Scedrov, "Optimization Complexity of Linear Logic Proof Games," to appear in *Theoretical Computer Science*.

[14] C. Lund, L. Fortnow, H. Karloff, and N. Nisan, "Algebraic Methods for Interactive Proof Systems," *Journal of the Association for Computing Machinery*, 39 (1992), pp. 859-868.

[15] M. Marathe, H. Hunt, R. Stearns, and V. Radhakrishnan, "Approximation Algorithms for PSPACE-Hard Hierarchically and Periodically Specified Problems," *SIAM Journal on Computing*, 27 (1998), pp. 1237-1261.

[16] C. Papadimitriou, Computational Complexity, Addison-Wesley, Reading, 1994.

[17] C. Papadimitriou, "Games Against Nature," *Journal of Computer and System Sciences*, 31 (1985), pp. 288-301.

[18] T. Schaefer, "On the Complexity of Some Two-Person, Perfect-Information Games," *Journal of Computer and System Sciences*, 16 (1978), pp. 185-225.

[19] A. Shamir, "IP = PSPACE," *Journal of the Association for Computing Machinery*, 39 (1992), pp. 869-877.

[20] M. Sudan, "Efficient Checking of Polynomials and Proofs and the Hardness of Approximation Problems," PhD Thesis, University of California, Computer Science Division, Berkeley, 1992.

[21] L. Valiant, "The Complexity of Enumeration and Reliability Problems," *SIAM Journal on Computing*, 8 (1979), pp. 410-421.

Joan Feigenbaum
AT&T Labs - Research
180 Park Avenue
Florham Park, NJ 07932-0971
USA
jf@research.att.com

On Approximating NP-Hard Optimization Problems

Johan Håstad

Abstract. We discuss the efficient approximability of NP-hard optimization problems. Although the methods apply to several problems we concentrate on the problem of satisfying the maximal number of equations in an over-determined system of linear equations. We show that over the field of two elements it is NP-hard to approximate this problem within a factor smaller than 2. The result extends to any Abelian group with the size of group replacing the constant 2.

1991 Mathematics Subject Classification: 68Q25,68Q99
Keywords and Phrases: Approximation algorithms, NP-hard optimization problems, Linear equations

1 Introduction

The basic entity in complexity theory is a computational problem which, from a mathematical point of view, is simply a function F from finite binary strings to finite binary strings. To make some functions more intuitive these finite binary strings should sometimes be interpreted as integers, graphs, or descriptions of polynomials. An important special case is given by decision problems where the range consists of only two strings, usually taken to be 0 or 1.

The basic notion of efficiently computable is defined as computable in time polynomial in the input-length. The class of polynomial time solvable decision problems is denoted by P. Establishing that a problem cannot be solved efficiently can sometimes be done but for many naturally occurring computational problems of combinatorial nature, no such bounds are known. Many such problems fall into the class NP; problems where positive answers have proofs that can be verified efficiently. The standard problem in NP is satisfiability (denoted SAT), i.e. the problem of given a Boolean formula φ over Boolean variables, is it possible to assign truth values to the variables to make φ true? The most common version of SAT, which is also the one we use here, is to assume that φ is a CNF-formula, i.e. a conjunction of disjunctions.

It is still unknown whether NP=P, although it is widely believed that this is not the case. It is even the case that much work in complexity theory, and indeed this paper, would have to be completely reevaluated if it turns out that NP=P. There is a group of problem, called the *NP-complete* problems, introduced by Cook [8], which have the property that they belong to P iff NP=P. Thus being NP-complete is strong evidence that a problem is computationally intractable and literally thousands of natural computational problems are today known to be NP-complete (for an outdated but still large list of hundreds of natural problems see [13]). SAT is the most well known NP-complete problem.

Many combinatorial optimization problems have a corresponding decision problem which is NP-complete. As an example take the traveling salesman problem of finding the shortest tour that visits a certain set of cities. The corresponding decision problem, namely that of, given K, determine if there is a tour of length K is NP-complete and thus solving the optimization problem exactly in polynomial time would mean that NP=P. Optimization problem with this property are called *NP-hard* (not NP-complete as they do not fall into the class NP as they are not decision problems). Solving NP-hard optimization problems exactly is thus hard, but in many practical circumstances it is almost as good to get an approximation of the optimum. Different NP-hard optimization problems behave very differently with respect to efficient approximation algorithms and this set of questions has lead to many interesting results.

The goal of this paper is to derive lower bounds on how well natural optimization problems can be approximated efficiently. The type of result we are interested in is a conclusion of the form "If optimization problem X and be approximated within factor c in polynomial time, then NP=P". The techniques we discuss give results for many optimization problem but we here concentrate on solving overdetermined systems of linear equations over finite Abelian groups. For this problem we are given a set of m equations in n unknowns and the task is to determine the maximal number of equations that can be simultaneously satisfied and possibly also produce an assignment that satisfies this number of equations.

An algorithm is a c-approximation algorithm if it, for every instance, finds a solution that is within a factor c of the optimal value. Thus if the best assignment satisfies o equation, such an approximation algorithm would always find an assignment that satisfies o/c equations. For linear equations over GF[2] a random assignment satisfies, on the average, half the equations. It is hence not surprising that one can efficiently find an assignment that satisfies at least half the equations. This gives a 2-approximation algorithm and the result extends to any Abelian group G to give a $size(G)$-approximation algorithm. The main result we discuss in this paper is to prove that this simple heuristic is optimal in that for any Abelian group G and any $\epsilon > 0$ it is NP-hard to $size(G) - \epsilon$-approximate the problem of linear equations over G.

The main tool for deriving such strong approximability results was introduced in the seminal paper [10]. It gives a connection to multiprover interactive proofs and let us here give an informal description of a variant of this concept. We later give some formal definitions in Section 1.1.

NP can be viewed as a proof system where a single prover P tries to convince a polynomial time verifier V that a statement is true. For concreteness let us assume that the statement is that a formula φ is satisfiable. In this case, P displays a satisfying assignment and V can easily check that it is a correct proof. This proof system is complete since every satisfiable φ admits a correct proof, and it is sound since V can never be made to accept an incorrect statement.

If φ contains n variables, V reads n bits in the above proof. Suppose we limit V to reading fewer bits where the most extreme case would be to let this number be constant independent of the the number of variables in φ. It is not hard to see that the latter is impossible unless we relax the requirements of the proof. The

proof remains a finite binary string, but we allow the verifier to make random choices. This means that given φ we should now speak of the probability that V accepts a certain proof π. Soundness is relaxed in that when φ is not satisfiable then there is some constant $s < 1$ such that for any proof π the probability that V accepts is bounded by s. A bit surprisingly it turns out that it is convenient also to relax completeness in that we only require the verifier to accept a correct proof for a correct statement with probability $c > s$ where we might have $c < 1$. Note that both completeness and soundness probabilities are taken only over V's internal random choices and hence we can improve these parameters by making several independent verifications and taking a majority decision. Naturally this increases other parameters that we want to keep small such as the number of bits of the proof that V reads.

It is an amazing fact, proved by Arora et al [1], that any NP-statement has a proof of the above type, usually called probabilistically checkable proof or simply PCP, where V only reads a constant, independent of the size of the statement being verified, number of bits of the proof and achieves soundness $s = 1/2$ and completeness $c = 1$. Apart from being an amazing proof system this gives a connection to approximation of optimization problems as follows.

Fix a formula φ and consider the PCP by Arora et al. Since everything is fixed except the proof π, we have a well defined function $acc(\pi)$, the probability that V accepts a certain proof π. Consider $\max_\pi acc(\pi)$. If φ is satisfiable then this optimum is 1, while if φ is not satisfiable then the optimum is $\leq s$. Thus, even computing this optimum approximately would enable us to decide an NP-complete question. Now by choosing the test appropriately this optimization problem can be transformed to more standard optimization problems leading to the desired inapproximability results.

1.1 Probabilistic proof systems

As discussed in the introduction we are interested in proof systems where the verifier is a probabilistic Turing machine. The simplest variant is a probabilistically checkable proof.

Definition 1.1 *A* Probabilistically Checkable Proof (PCP) *with completeness c and soundness s for a language L is given by a verifier V such that for $x \in L$ there is a proof π such that V^π outputs 1 on input x with probability at least c, and for $x \notin L$ and all π the probability that V^π outputs 1 on input x is bounded by s.*

We are interested in efficient PCPs and hence we assume that V runs in worst case polynomial time. It is also important for us to efficiently enumerate all the random choice of V and hence we need that V only makes $O(\log |x|)$ binary random choices on input x. We maintain this property without mentioning it explicitly.

We also need what is generally called a two-prover one-round interactive proof. Such a verifier has two oracles but has the limitation that it can only ask one question to each oracle and that both questions have to be produced before either of them is answered. We do not limit the answer size of the oracles. We call the two oracles P_1 and P_2.

DEFINITION 1.2 *A probabilistic polynomial time Turing machine V is a verifier in a* two-prover one-round proof system *with completeness c and soundness s for a language L if on input x it produces two strings $q_1(x)$ and $q_2(x)$, such that for $x \in L$ there are two oracles P_1 and P_2 such that the probability that V accepts $(x, P_1(q_1(x)), P_2(q_2(x)))$ is c while for $x \notin L$, for any two oracles P_1 and P_2 the probability that V accepts $(x, P_1(q_1(x)), P_2(q_2(x)))$ is bounded by s.*

In all our two-prover interactive proofs the verifier always accepts a correct proof for a correct statement, i.e. we have $c = 1$ in the above definition.

BRIEF HISTORY. The notion of PCP was introduced by Arora and Safra [2]. It was a variation of randomized oracle machines discussed by Fortnow et al [12] and transparent proofs by Babai et al [4]. Multiprover interactive proofs were introduced by Ben-Or et al [7], and all these systems are variants of interactive proofs as introduced by Goldwasser, Micali, and Rackoff [14] and Babai [3].

1.2 ESSENTIAL PREVIOUS WORK

The surprising power of interactive proofs was first established in the case of one prover [17], [20] and then for many provers [5]. After the fundamental connection with approximation was discovered [10] the parameters of the proofs improved, culminating in the result [2, 1] that it is sufficient to read a constant number of bits. Using a transformation of [18] and massaging the result slightly we arrive at the following theorem.

THEOREM 1.3 *[1] Let L be a language in NP and x be a string. There is a universal constant $c < 1$ such that, we can in time polynomial in size(x) produce a CNF formula $\varphi_{x,L}$ with exactly 3 literals in each clause such that if $x \in L$ then $\varphi_{x,L}$ is satisfiable while if $x \notin L$, any assignment satisfies at most a fraction c of the clauses of $\varphi_{x,L}$. Furthermore, we can assume that each variable appears exactly 12 times.*

Let us now turn to two-prover interactive proofs. Given a one-round protocol with soundness s and completeness 1 we can repeat it k times in sequence improving the soundness to s^k. This creates a many round protocols, whereas we need our protocols to remain one-round. This can be done by parallel repetition in that V repeats his random choices to choose k pairs of questions $(q_1^{(i)}, q_2^{(i)})_{i=1}^k$ and sends $(q_1^{(i)})_{i=1}^k$ to P_1 and $(q_2^{(i)})_{i=1}^k$ to P_2 all at once. V then receives k answers from each prover and accepts if it would have accepted in all k protocols given each individual answer. The soundness of such a protocol can be greater than s^k, but Raz [19] proved that when the answer size is small the soundness is exponentially decreasing with k.

THEOREM 1.4 *[19] For all integers d and $s < 1$, there exists $c_{d,s} < 1$ such that given a two-prover one-round proof system with soundness s and answer sizes bounded by d, then for all integers k the soundness of k protocols run in parallel is bounded above by $c_{d,s}^k$.*

1.3 Definitions for approximation algorithms

Definition 1.5 *Let O be a maximization problem. For an instance x of O let $OPT(x)$ be the optimal value. An efficient C-approximation algorithm is an algorithm that on any input x outputs a number V such that $OPT(x)/C \leq V \leq OPT(x)$ and runs in worst case polynomial time.*

2 First steps towards a good proof system

We want to construct a proof system for an arbitrary language in NP and let us start by an overview.

We start by a simple two-prover one-round protocol which is obtained more or less immediately from Theorem 1.3. We improve the soundness of this protocol by running several copies of it in parallel and using Theorem 1.4. It is possible to transform this improved two-prover protocol to a PCP simply by writing down the prover answers. The answers are, however, long and since we want to keep the number of read bits very small we write the answers in a more useful form by asking the prover to supply the value of all Boolean functions of these answers. This is the long code of the answers as defined in [6]. We now proceed to give the details.

Suppose $\varphi = C_1 \wedge C_2 \wedge \ldots \wedge C_m$, where C_j contains the variables x_{a_j}, x_{b_j} and x_{c_j}. Consider the following one-round two-prover interactive proof.

Simple two-prover protocol

1. V chooses $j \in [m]$ and $k \in \{a_j, b_j, c_j\}$ both uniformly at random. V sends j to P_1 and k to P_2.

2. V receives values for x_{a_j}, x_{b_j} and x_{c_j} from P_1 and for x_k from P_2. V accepts if the two values for x_k agree and C_j is satisfied.

We have the following proposition which can be proved by a straightforward argument which we omit.

Proposition 2.1 *If any assignment satisfies at most a fraction e of the clauses of φ, then V accepts in the simple two prover protocol with probability at most $(2+e)/3$.*

We now concentrate a protocol we called the *u-parallel 2-prover game* and which consists of u copies of this basic game. That is, the verifier picks u clauses $(C_{j_k})_{k=1}^u$ and then uniformly at random for each k he picks a random variable x_{i_k} contained in C_{j_k}. The variables of $(C_{j_k})_{k=1}^u$ are sent to P_1 while $(x_{i_k})_{k=1}^u$ are sent to P_2. The two provers respond with assignments on the queried variables and the verifier checks that the values are consistent and that the chosen clauses are satisfied. We get completeness 1 and the soundness is in the case of $\varphi_{x,L}$ of Theorem 1.3 is, by Theorem 1.4 and Proposition 2.1, bounded by c_1^u for some constant $c_1 < 1$. To fix notation, Let $U = \{x_{i_1}, x_{i_2} \ldots x_{i_u}\}$ be the set of variables sent to P_2, and W the set of variables sent to P_1.

As discussed in the introduction to this section we want to replace this two-prover interactive proof by a PCP consisting of the answers of P_1 and P_2 given in a more redundant form. We use the powerful long code introduced in [6].

DEFINITION 2.2 *The* long code *of a string x of length w is of length 2^{2^w}. The coordinates of the codeword are identified with all possible functions $f : \{0,1\}^w \mapsto \{0,1\}$ and the value of coordinate f is $f(x)$.*

Before we continue, let us fix some more notation. The written part of the PCP described above is called a *Standard Written Proof of size u* or simply SWP(u). Let $\mathcal{F}_T$ denote the set of functions on a set T and let A_T be the supposed long code of the restriction of the satisfying assignment to the set T. It is convenient to have $\{-1,1\}$ as our set of two values for Boolean functions and Boolean variables and thus exclusive-or turns into multiplication. For the supposed long code A_T we assume that $A_T(f) = -A_T(-f)$. This is achieved by, for each pair $(f,-f)$, having only one value in A_T. This value is negated if the value of $A_T(-f)$ is needed. For the tables A_W, we know that it should be a long code for some assignment that satisfies $\wedge_k C_{j_k}$ and instead of storing an entry for each $g \in \mathcal{F}_W$ we only store an entry for each function of the form $g \wedge (\wedge_k C_{j_k})$. When we want the value of a function h we access the entry for $h \wedge (\wedge_k C_{j_k})$. A SWP($u$) is correct for φ if there is an assignment x that satisfies φ and thus that $A_T(f) = f(x|_T)$ for any supposed long code A_T for any set T obtained as U or W in a run of the u-parallel 2-prover game.

We need the discrete Fourier transform defined by

$$\hat{A}_{\alpha,T} = 2^{-2^{size(T)}} \sum_f A_T(f) \prod_{x\in\alpha} f(x)$$

for any $\alpha \subseteq \{-1,1\}^T$. It is inverted by

$$A_T(f) = \sum_\alpha \hat{A}_{\alpha,T} \prod_{x\in\alpha} f(x).$$

and since $|A(f)| = 1$ for any f we have, by Parseval's identity, $\sum_\alpha \hat{A}^2_{\alpha,T} = 1$. For a set $\beta \subseteq \{-1,1\}^W$ and $U \subset W$ we let $\pi_2^U(\beta) \subseteq \{-1,1\}^U$ be those elements that have an odd number of extensions in β. This a mod 2 projection and note that it might be empty even if β is not empty.

3 LINEAR EQUATIONS

We now study the optimization problem of satisfying the maximal number of equations mod 2. For natural reasons we want to design a test for SWP(u) that accepts depending only on the exclusive-or of three bits of the proof.

TEST $L_2^\epsilon(u)$

WRITTEN PROOF. A SWP(u).

Desired property. To check that it is a correct SWP(u) for a given formula $\varphi = C_1 \wedge C_2 \ldots C_m$.
Verifier. Choose set U and W as in the u-parallel 2-prover game. Choose $f \in \mathcal{F}_U$ and $g_1 \in \mathcal{F}_W$ with the uniform probability. Choose a function $\mu \in \mathcal{F}_W$ by setting $\mu(y) = 1$ with probability $1-\epsilon$ and $\mu(y) = -1$ otherwise, independently for each $y \in \{-1,1\}^W$. Define g_2 by for each $y \in \{-1,1\}^W$, $g_2(y) = f(y|_U)g_1(y)\mu(y)$. Accept if $A_U(f)A_W(g_1)A_W(g_2) = 1$.

We need to analyze the soundness and completeness of this test.

Lemma 3.1 *The completeness of Test $L_2^\epsilon(u)$ is at least $1-\epsilon$.*

Proof: Fix a correct SWP(u) obtained from the assignment x satisfying φ We claim that V accepts unless $\mu(x|_W) = -1$ and leave the verification to the reader. ∎

Lemma 3.2 *For any $\epsilon > 0$, $\delta > 0$, suppose that the probability that the verifier of Test $L_2^\epsilon(u)$ accepts is $(1+\delta)/2$. Then there is a strategy for P_1 and P_2 in the u-parallel two prover protocol that makes the verifier of that protocol accept with probability at least $\epsilon\delta^2/2$.*

Proof: Let us first fix U and W and for notational convenience we denote the function A_U by A and the function A_W by B. We want to consider

$$E_{f,g_1,\mu}[A(f)B(g_1)B(g_2)] \tag{1}$$

since by the assumption of the lemma

$$E_{U,W,f,g_1,\mu}[A_U(f)A_W(g_1)A_W(g_2)] = \delta. \tag{2}$$

Using the Fourier expansion and moving the expected value inside (1) equals

$$\sum_{\alpha,\beta_1,\beta_2} \hat{A}_\alpha \hat{B}_{\beta_1}\hat{B}_{\beta_2} E_{f,g_1,\mu}\left[\prod_{x\in\alpha} f(x) \prod_{y\in\beta_1} g_1(y) \prod_{y\in\beta_2} (f(y|_U)g_1(y)\mu(y))\right]. \tag{3}$$

If $\beta_1 \neq \beta_2$ then since $g_1(y)$ for $y \in \beta_1 \Delta \beta_2$ is random and independent of all other variables the inner expected value in this case is 0 and thus we can disregard all terms except those with $\beta_1 = \beta_2 = \beta$. Now consider such a term and let s_x be number of $y \in \beta$ such that $y|_U = x$. Since $f(x)$ is random and independent for different x, unless for every x either $x \in \alpha$ and s_x is odd or $x \notin \alpha$ and s_x is even again the inner expected value is 0. These conditions imply that we only keep terms with $\pi_2^U(\beta) = \alpha$ and finally since $E_\mu[\prod_{y\in\beta}\mu(y)] = (1-2\epsilon)^{size(\beta)}$ we have reduced the sum (1) to

$$\sum_{\alpha} \sum_{\beta|\pi_2^U(\beta)=\alpha} \hat{A}_\alpha \hat{B}_\beta^2 (1-2\epsilon)^{size(\beta)}. \tag{4}$$

We want to prove that if the expected value of this (over random choices of U and W) is at least δ then we have a good strategy of the provers. We define good randomized strategies for P_1 and P_2.

The strategy of P_2 is first to pick a random α with and $\hat{A}_\alpha \geq \delta/2$. The probability of picking α is defined to be proportional to $\hat{A}_\alpha$ and hence by Parseval's identity it is at least $\delta\hat{A}_\alpha/2$. P_2 sends a random $x \in \alpha$. Note that α is nonempty since $A_U(f) = -A_U(-f)$ implies that $\hat{A}_\emptyset = 0$.

The strategy of P_1 is to pick a random β with probability $\hat{B}_\beta^2$ and then answer with a random $y \in \beta$.

Let us evaluate the success-rate of this strategy. By the property that $A_W(h)$ only depends on $h \wedge (\wedge_k C_{j_k})$ it is not hard to establish that every y sent by P_1 satisfies the corresponding clauses and thus we only need to look at the probability that the answers are consistent. This probability is at least $size(\beta)^{-1}$ times the probability that for the picked α and β we have $\alpha = \pi_2^U(\beta)$. The probability of picking a specific pair α and β is, provided $\hat{A}_\alpha > \delta/2$, at least $\hat{A}_\alpha\hat{B}_\beta^2\delta/2$ and thus the overall success-rate for a fixed choice of U and W is at least

$$\delta/2 \sum_{\alpha|\hat{A}_\alpha|\geq\delta/2} \sum_{\beta|\pi_2^U(\beta)=\alpha} \hat{A}_\alpha\hat{B}_\beta^2 size(\beta)^{-1}. \tag{5}$$

Comparing this sum to (4) and making some calculations one can establish that expected value over U and W is at least $\delta^2\epsilon/2$ and the proof of Lemma 3.2 is complete. ∎

Armed with the very efficient PCP given by Test $L_2^\epsilon(u)$ we can now establish the main theorem of this paper.

THEOREM 3.3 *For any $\epsilon > 0$ it is NP-hard to approximate the problem of maximizing the number of satisfied equation in a system of linear equations mod 2 within a factor $2 - \epsilon$. The result applies to systems with only 3 variables in each equation.*

PROOF: *(Sketch)* Let L be an arbitrary language in NP and given an input x, create the formula $\varphi_{x,L}$ as given in Theorem 1.3. Let δ be small positive number to be determined and consider test $\mathrm{L}_2^\delta(u)$ where u is chosen sufficiently large so that the acceptance probability in the u-parallel 2-prover game is smaller than $\delta^3/2$.

For each bit b in a SWP(u) introduce a variable x_b. To accept in the test $L_2^\delta(u)$ is equivalent to the condition

$$b_{U,f}b_{W,g_1}b_{W,g_2} = c$$

where $b_{U,f}$, b_{W,g_1} and b_{W,g_2} are the bits in the proof corresponding to $A_U(f)$, $A_W(g_1)$ and $A_W(g_2)$, respectively[1]. Write down the equation

$$x_{b_{U,f}}x_{b_{W,g_1}}x_{b_{W,g_2}} = c$$

[1]One might think that the right hand size would always be 1, but because of our convention on having one entry in A_U to represent the value on two functions this might be the case since the value corresponding to $A_U(f)$ in the proof might actually give the value of $A_U(-f)$

with a weight that is equal to the probability that the verifier chooses the tuple (U, W, f, g_1, g_2). Now each proof corresponds to an assignment to the variables x_b and the total weight of all satisfied equations is exactly the probability that this proof is accepted. This implies that if $x \in L$ this maximal weight is $1 - \delta$ while if $x \notin L$, it is, in view of Lemma 3.2 and the choice of u, at most $(1 + \delta)/2$. It is not difficult to check that we have a polynomial number of equations and an approximation algorithm with performance ratio smaller than $2 - \epsilon$ would enable us, for sufficiently small δ, to answer a NP-hard question.

As is standard, the weights can be eliminated by duplicating each equation a suitable number of times. This creates a slight degrade in the value of ϵ, but since ϵ is arbitrary anyway this can easily be compensated. We omit the details. ∎

Note that there is a meta reason that we have to introduce the error function μ and make our test have non perfect completeness. If we had perfect completeness then the equations produced in the proof of Theorem 3.3 could all be satisfied simultaneously. However, to decide if a set of linear equations have a common solution can be done in polynomial time by Gaussian elimination.

Finally, let us just state the extension to an arbitrary Abelian group.

Theorem 3.4 *For any $\epsilon > 0$ and any Abelian group G, given a system of linear equations over G, it is NP-hard to approximate the maximal number of simultaneously satisfiable equations within a factor $size(G) - \epsilon$. The result applies to systems with only 3 variables in each equation.*

4 Final remarks

As mentioned in the introduction the efficient multiprover interactive proofs give strong inapproximability results for many combinatorial optimization problems.

Independence number is to, given a graph G, find the largest set S of nodes such that no two nodes in S ares pairwise connected. It is established in [15] that it is, assuming that NP cannot be done in probabilistic polynomial time, for any $\epsilon > 0$, hard to approximate independence number within $n^{1-\epsilon}$ where n is the number of nodes G. A very related problem is that of chromatic number where we want to color the nodes in G with the minimal number of colors so that adjacent nodes get different colors. The result for independence number can be extended to chromatic number [11]. The problem of set cover is that given a number of subsets S_i of $[n]$ to find the minimal size sub-collection of the S_i that covers the entire set. This problem is, under standard complexity assumptions, hard to approximate within $(1 + o(1)) \ln n$ [9] and this result is tight. For inapproximability results on other problem, some optimal and some non-optimal we refer to the full versions of [6] and [16].

References

[1] S. Arora, C. Lund, R. Motwani, M. Sudan and M. Szegedy. Proof verification and intractability of approximation problems. Proc. of the 33rd Annual IEEE Symposium on Foundations of Computer Science, Pittsburgh, 1992, pp 14-23.

[2] S. ARORA AND S. SAFRA. Probabilistic checking of proofs: a new characterization of NP. Journal of the ACM, Vol 45, 1998, pp 70-122.
[3] L. BABAI. Trading group theory for randomness. Proc. of the 17th Annual ACM Symposium on Theory of Computation, Providence, 1985, pp 420-429.
[4] L. BABAI, L. FORTNOW, L. LEVIN, AND M. SZEGEDY. Checking computations in polynomial time. Proc. of the 23rd Annual ACM Symposium on Theory of Computation, New Orleans, 1991, pp 21-31.
[5] L. BABAI, L. FORTNOW, AND C. LUND. Non-deterministic exponential time has two-prover interactive protocols. Computational Complexity, Vol 1, 1991, pp 3-40.
[6] M. BELLARE, O. GOLDREICH AND M. SUDAN. Free Bits, PCPs and Non-Approximability-Towards tight Results. Proc. of the 36th Annual IEEE Symposium on Foundations of Computer Science, 1995, Milwaukee, pp 422-431. Full version available from ECCC, Electronic Colloquium on Computational Complexity (http://www.eccc.uni-trier.de/eccc).
[7] M. BEN-OR, S. GOLDWASSER, J. KILIAN, AND A. WIGDERSON. Multiprover interactive proofs. How to remove intractability. Proc. of the 20th Annual ACM Symposium on Theory of Computation, Chicago, 1988, pp 113-131.
[8] S. A. COOK. The complexity of Theorem Proving Procedure. Proceeding 3rd ACM Symposium on Theory of Computing, 1971, pp 151-158.
[9] U. FEIGE. A threshold of $\ln n$ for approximating set cover. Proc. of the 28th Annual ACM Symposium on Theory of Computation, Philadelphia 1996, pp 314-318.
[10] U. FEIGE, S. GOLDWASSER, L. LOVÁSZ, S. SAFRA, AND M. SZEGEDY. Interactive proofs and the hardness of approximating cliques. Journal of the ACM, Vol, 43:2, pp 268-292.
[11] U. FEIGE AND J. KILIAN. Zero-Knowledge and the chromatic number. Proc. of the 11th Annual IEEE conference on Computational Complexity, Philadelphia 1996, pp 278-287.
[12] L. FORTNOW, J. ROMPEL, AND M. SIPSER. On the power of Multi-Prover Interactive Protocols. Proc. 3rd IEEE Symposium on Structure in Complexity Theory, pp 156-161, 1988.
[13] M.R. GAREY AND D.S. JOHNSSON. Computers and Intractability. W.H. Freeman and Company, 1979.
[14] S. GOLDWASSER, S. MICALI, AND C. RACKOFF. The knowledge complexity of interactive proof systems. SIAM Journal on Computing, Vol 18, pages 186-208, 1989.
[15] J. HÅSTAD. Clique is hard to approximate within $n^{1-\epsilon}$. Proc. of the 37th Annual IEEE Symposium on Foundations of Computer Science, Burlington 1996, pp 627-636. Full version available from ECCC, Electronic Colloquium on Computational Complexity (http://www.eccc.uni-trier.de/eccc).
[16] J. HÅSTAD. Some Optimal In-approximability Results. Proc. 29th Annual ACM Symposium on Theory of Computation, 1997, pp 1-10. Full version available from ECCC, Electronic Colloquium on Computational Complexity (http://www.eccc.uni-trier.de/eccc).
[17] C. LUND, L. FORTNOW, H. KARLOFF AND N. NISAN. Algebraic methods for interactive proof systems. Journal of the ACM, Vol 39, No 2, pp 859-868.
[18] C. PAPADIMITRIOU AND M. YANNAKAKIS. Optimization, approximation and complexity classes. Journal of Computer and System Sciences, Vol 43, 1991, pp 425-440.
[19] R. RAZ. A parallel repetition theorem. Proc. of the 27th Annual ACM Symposium on Theory of Computation, Las Vegas 1995, pp 447-456.
[20] A. SHAMIR. IP=PSPACE. Journal of the ACM, Vol 39, No 2, pp 869-877.

Johan Håstad
Dept of Numerical Analysis and
Computing Science
Royal Institute of Technology
S-100 44 Stockholm
Sweden

Unsolvable Systems of Equations and Proof Complexity

Toniann Pitassi[1]

Abstract. This abstract discusses algebraic proof systems for the propositional calculus. We present recent results, current research directions, and open problems in this area.

1991 Mathematics Subject Classification: Primary 03F, 68Q, 68R05, 20C30, 14Q99
Keywords and Phrases: Systems of polynomial equations over finite fields, propositional proof complexity, lower bounds, Grobner bases.

1 Introduction

A fundamental problem in logic and computer science is understanding the efficiency of propositional proof systems. It has been known for a long time that $NP = coNP$ if and only if there exists an efficient propositional proof system, but despite 25 years of research, this problem is still not resolved. (See [21] for an excellent survey of this area; see also [2] for a more recent article focusing on open problems in proof complexity.) The intention of the present article is to discuss algebraic approaches to this problem. Our proof systems are simpler than classical proof systems, and purely algebraic. It is our hope that by studying proof complexity in this light, that new upper and lower bound techniques may emerge. This paper is a revision and update of the earlier paper ([18]); due to space considerations, we omit all proofs and focus on current research directions.

Let $C = C_1 \wedge C_2 \wedge ... \wedge C_m$ be an instance of the classical NP-compete problem, 3SAT. That is, C is a propositional formula over $\{x_1, ..., x_n\}$, in conjunctive normal form, where each C_i is a clause of size at most three. Each clause C_i can be converted into an equation, $\overline{C_i} = 1$ over F such that C is unsatisfiable if and only if $\{\overline{C}_1 = 0, ..., \overline{C}_m = 0\}$ has no 0/1 solution. The equations $Q = \{Q_1 = 0, ...Q_R = 0\}$ corresponding to C are: $\{\overline{C}_1 = 0, ..., \overline{C}_m = 0\}$, plus the equations $x^2 - x = 0$ for all variables x.

We show how to translate from the basis $\{\vee, \wedge, \neg\}$ to the basis $\{+, \times, 1\}$ over a field F. For a atomic, $t(a) = 1 - a$; $t(\neg x) = 1 - t(x)$; $t(x \vee y) = t(x)t(y)$; and lastly, $t(x \wedge y) = t(\neg(\neg x \vee \neg y) = t(x) + t(y) - t(x)t(y)$. Our translation has the property that for any truth assignment α, and any boolean formula f, f evaluates

[1]Supported by NSF NYI grant CCR-9457783, US-Israel BSF Grant 95-00238, and Grant INT-9600919/ME-103 from NSF and MŠMT (Czech Republic)

to 1 under α if and only if $t(f)$ evaluates to 0 under α. In other words, "0" represents true over the new basis. Moreover, one could further convert Q into a family of degree 2 equations by replacing each monomial xyz in Q_i by xw (where w is a new variable), and adding the extra equations $w - yz = 0$ and $w^2 - w = 0$.

The above reduction (due to Valiant [22]) shows that solving systems of degree 2 polynomial equations is NP-complete. We are interested in defining natural algebraic proofs in the case where the equations are unsolvable, and in studying the complexity of the resulting proofs. What exactly is a natural algebraic proof, and how long can such proofs be? Our starting point for defining such algebraic proof systems is Hilbert's Nullstellensatz. That is, if $Q_i(\overline{x}) = 0$ is a system of algebraic equations over F (translated from an instance of 3SAT), then the equations do not have a solution in the algebraic closure of F if and only if there exists polynomials $P_i(\overline{x})$ from $F[\overline{x}]$ such that $\sum_i P_i(\overline{x})Q_i(\overline{x}) = 1$. We can think of the polynomials P_i as a *proof* of the unsolvability of the equations Q_i. Moreover, in our scenario since Q_i includes the equations $x^2 - x = 0$ for all variables x, there exists a solution if and only if there exists a $0-1$ valued solution. This is the main property which distinguishes our investigations from earlier, classical work on the effective Nullstellensatz. ([10, 15, 5]).

Algebraic proof systems are appealing because of their simplicity and non-syntactic nature. Moreover, the question of how large a proof must be amounts to asking how many field operations are required in order to generate the constant polynomial from certain initial polynomials. Moreover these proof systems are powerful, and by studying various complexity notions (degree, monomial size, algebraic size), there are close correspondences between these systems and various classical propositional proofs.

The organization of the paper is as follows. In Section 2, we define our algebraic proof systems and various complexity measures on them. In Section 3, we state basic theorems about algebraic proofs and simulation results. In Section 4, we focus our attention on lower bounds. Lastly in Section 5, we present several open problems in this area.

2 Algebraic Proof Systems

Recall that $C = C_1 \wedge C_2 \wedge ... \wedge C_m$ is a propositional formula over $\{x_1, ..., x_n\}$, in conjunctive normal form, where each C_i is a clause of size at most three. Let Q be the corresponding system of (degree 3) polynomial equations. Here is a simple example. Let $C = (b \vee a) \wedge (\neg a \vee b) \wedge (\neg b)$. Then $Q = \{Q_1, Q_2, ..., Q_5\}$, where $Q_1 = (1-b)(1-a) = 1-a-b+ab$, $Q_2 = (a)(1-b) = a-ab$, $Q_3 = b$, $Q_4 = a^2 - a$, $Q_5 = b^2 - b$.

An *algebraic* refutation for C (over a fixed ring or field F) is an algebraic straight-line program, $S = S_1, ..., S_l$ such that each S_i is either one of the initial equations (from Q) or is obtained from previous equations by a valid rule, and where the final equation S_l is $0 = 1$. The two rules are as follows. (1) From $g_1(\overline{x}) = 0$ and $g_2(\overline{x}) = 0$, derive $ag_1(\overline{x}) + bg_2(\overline{x}) = 0$, where a, b are constants from F; (2) From $g(\overline{x}) = 0$, infer $xg(\overline{x}) = 0$ for x a variable. (Thus, a proof is merely an explicit derivation that 1 is in the ideal generated by Q.) In the above

example, a refutation is: $S_1 = Q_1$, $S_2 = Q_2$, $S_3 = Q_3$, $S_4 = S_1 + S_2 = 1 - b$, $S_5 = S_4 + S_3 = 1$. An algebraic refutation S for Q can also be put in an alternate form, $\sum_i P_i(\overline{x})Q_i(\overline{x}) = 1$.

Our algebraic proof system is *sound* since such a straight-line program is not possible to obtain if Q is solvable. The algebraic proof system is also *complete* since every unsolvable system of equations Q (derived from an unsatisfiable 3CNF formula C) has an algebraic proof. There are several proofs of completeness. One follows from (the weak form of) Hilbert's Nullstellensatz. There are also other simpler and more constructive proofs [18, 8]; one is obtained by simulating a truth-table proof and a second is by simulating a type of tableau proof.

2.1 Complexity measures

We will discuss several complexity measures on algebraic refutations. Perhaps the most natural is the *algebraic size*. This is defined to be the number of lines, l, in S. The *degree* is defined to be the maximum degree of the intermediate polynomials S_i, after simplifications. This measure has been studied quite a bit, and the name *Polynomial Calculus* (PC) is given to algebraic proofs in this form, where the S_i's are viewed as explicit sums of monomials. Another degree measure, which is called the *Nullstellensatz* (HN) degree is the maximum degree of the intermediate polynomials S_i *before* simplifications. That is, the maximum degree of the polynomial $\sum_i P_iQ_i$ in the alternate representation $\sum_i P_iQ_i = 1$.

Note that the minimal Polynomial Calculus degree of a formula f is never greater than the minimal Nullstellensatz degree of f; however, the Polynomial Calculus degree can sometimes be much smaller as is evidenced by the following example. Let IND_n denote the following system of degree 2 equations: (1) $1-x_1 = 0$; (2) $x_i(1 - x_{i+1}) = 0$ for all $1 \leq i \leq n-1$; (3) $x_n = 0$ and (4) $x_i^2 - x_i = 0$ for all $1 \leq i \leq n-1$. (These equations formalize induction: if $x_1 = 1$ and $x_n = 0$, then there must be an index i such that $x_i = 1$ and $x_{i+1} = 0$.) It is not too hard to see (by applying induction!) that these equations have a degree 2 PC refutation; on the other hand, they require degree $O(logn)$ Nullstellensatz refutatons [7].

2.2 Automatizability

An important issue in proof complexity is whether or not a given proof system can actually be used as the basis for an efficient automated theorem prover. Intuitively, it seems that the more expressive and powerful the proof system, the harder it is to perform an efficient search for a short proof. A proof system S is thus said to be *automatizable* if there exists a deterministic procedure A that takes as input a (unsatisfiable) formula f and outputs an S-proof f in time polynomial in the size of the shortest S-proof of f. In other words, if S is automatizable, then short proofs can be found efficiently.

One of the nicest features of algebraic proofs is that small degree proofs can be found quickly–in other words, small-degree proofs are automatizable. To see this in the case of small-degree Nullstellensatz proofs, note that if $\sum_i P_iQ_i = 1$ where P_i's have degree at most d, and the Q_i's have degree at most 3, then the total number of monomials on the left side is bounded by a polynomial in d and

therefore we can set up a system of linear equations (one for each monomial) and solve for the coefficient values in polynomial time. Using a modification of the Gröbner basis algorithm, [11] have shown that small-degree Polynomial Calculus proofs are also automatizable.

THEOREM. [11] For all d, n, there is an algorithm A such that for any (unsatisfiable) 3CNF formula f with underlying variables $x_1, .., x_n$, A returns a degree-d Polynomial Calculus refutation (if one exists) in time $n^{O(d)}$.

3 Relationship to classical proof systems

In this section, we will discuss the relationship between the size of algebraic proofs under the above complexity measures and the size of more standard propositional proofs.

3.1 Algebraic proofs versus Frege proofs

DEFINITION. The algebraic proof system over F is *polynomially-bounded* if there exists a constant c such that for every unsatisfiable 3CNF formula, f, there exists an algebraic proof of f of size $O(|f|^c)$ (that is, the proof is of size polynomial in the size of f).

The standard definition of a propositional proof system is as follows.
DEFINITION. Let $L \subseteq \Sigma^*$, where Σ is a finite alphabet, and Σ^* denotes all finite strings over Σ. (Typically, L encodes either the set of all tautological formulas, or the set of all unsatisfiable formulas.) Then a Cook-Reckhow proof system for L is a function $f : \Sigma^* \to L$, where f is an onto, polynomial-time computable function. A Cook-Reckhow proof system, f, is polynomially bounded if there is a polynomial $p(n)$ such that for all $y \in L$, there is an $x \in \Sigma^*$ such that $y = f(x)$ and $|x|$ (the length of x) is at most $p(|y|)$.

A key property of a Cook-Reckhow proof system is that, given an alleged proof, there is an efficient method for checking whether or not it really is a proof. For most standard, axiomatic proof systems (Extended Frege, Frege, even ZFC), there is actually a very efficient method for checking whether or not it is really a proof. This property leads to the following theorem.

THEOREM. [13] There exists a polynomially-bounded Cook-Reckhow propositional proof system if and only if $NP = coNP$.

The above theorem does not appear to hold for algebraic proofs because there is no known deterministic polynomial time algorithm to check whether or not a polynomial is identically 1, even in the case of finite fields. (In other words, there is no efficient procedure to check that it is a proof.) Nonetheless, the probabilistic polynomial-time algorithm due to Schwartz allows us to prove that if algebraic proofs are polynomially-bounded, then the polynomial hierarchy collapses.

THEOREM. [18] For any prime p, if the algebraic proof system over Z_p is polynomially-bounded, then $PH = \Sigma_2^p$.

We conjecture that the above premise also implies $NP = coNP$. It is not too hard to show that algebraic proof systems are at least as powerful as Extended Frege systems, as is evidenced by the following theorem.

THEOREM. [18] For any commutative ring R, Frege proofs (and Extended Frege proofs) can be polynomially simulated by algebraic proofs with polynomial size.

It is open whether or not the simulation holds in the reverse direction.

3.2 POLYNOMIAL CALCULUS VERSUS RESOLUTION

Resolution proofs dominate the work in automated theorem proving since they are extremely simple and can also be applied to first order theorem proving. A Resolution proof P of an (unsatisfiable) CNF formula $f = C_1 \wedge ... \wedge C_m$ is a sequence of clauses $D_1, ..., D_l$ such that: (a) each D_i is either an initial clause from f or follows from two previous clauses by the Resolution rule, and (b) the final clause D_l is the empty clause. The resolution rule derives $(A \vee B)$ from $(A \vee x)$ and $(B \vee \neg x)$, where A and B are disjunctions of literals. The size of the above Resolution proof is l; a tree-like proof has the additional property that each intermediate clause generated in the proof (not including the initial clauses) can be used at most once in the derivation–i.e., if it is used more than once it must be re-derived. Tree-like Resolution is of practical interest since most theorem provers are based on tree-like Resolution proofs. The following theorem gives a relationship between small degree Polynomial Calculus proofs and small-size Resolution proofs.

THEOREM. [11] If f has a tree-like Resolution proof of size S, then f has a degree $O(logS)$ Polynomial Calculus refutation. If f has a Resolution proof of size S, then f has a degree $O(\sqrt{nlogS})$ Polynomial Calculus refutation.

The intuition behind the above proof is as follows. Define the *width* of a Resolution proof to be the maximum clause size in the proof. The proof of the above theorem can be used to show: (1a) If f has a size S tree-like Resolution proof, then f has a width $O(logS)$ Resolution proof [9]; (1b) if f has a size S Resolution proof, then f has a width $O(\sqrt{nlogS})$ Resolution proof. And secondly, it is easy to show: (2) if f has a width d Resolution proof, then f has a degree $O(d)$ Polynomial Calculus proof.

3.3 POLYNOMIAL CALCULUS VERSUS BOUNDED-DEPTH FREGE

Bounded-depth Frege proofs are Frege proofs where the depth of each intermediate formula is bounded by a fixed constant. (See [21, 2] for motivation and details.) Bounded-depth Frege proofs are known to be strictly more powerful than Resolution, but strictly less powerful than unrestricted Frege proofs. $AC^0[p]$-Frege proofs are bounded-depth Frege proofs where the underlying connectives are: unbounded fanin AND, OR, NOT and MODp. There are no nontrivial lower bounds known at present for $AC^0[p]$-Frege proofs, and the original motivation for defining and studying small-degree algebraic proofs was to prove such lower bounds [4].

It does not seem to be possible to simulate polynomial-size $AC^0[p]$-Frege proofs by small degree Polynomial Calculus proofs (over GF_p). This is because any

single unbounded fanin OR gate would translate into a large degree polynomial. To circumvent this problem, [8] extended the Polynomial Calculus by adding new equations to the initial ones, where these new equations introduce new variables to represent or define unbounded fanin OR gates. The new equations, R, are small-degree polynomial equations in the original variables, plus the new "extension variables." The nesting level of the new equations corresponds to the depth of the unbounded fanin formulas that can be represented. Thus, loosely speaking, a degree d constant-depth Polynomial Calculus with Extension proof of f is a degree d Polynomial Calculus refutation of $0 = 1$ from the equations Q, R, where Q corresponds to the original equations defining f, and R corresponds to the new extension axioms, and such that the definitions given by R have a constant number of levels of nestings. With these definitions, [8] show that constant-depth $AC^0[p]$-Frege proofs are essentially equivalent to constant-depth Polynomial Calculus with Extension proofs.

In a different line of work, [17] show that any quasipolynomial-size $ACC^0[2]$-Frege proof can be simulated by a quasipolynomial-size, depth 3 Frege proof of a very special form: the output gate is a weak threshold gate, the middle layer consists of mod 2 gates and the input layer consists of AND gates of small fanin. Put another way, each formula in the depth 3 Frege proof is a probabilistic small-degree polynomial over GF_2. This in turn can be viewed as another generalization of small-degree Polynomial Calculus proofs.

4 Lower Bounds

In the last five years, there have been many lower bounds obtained on the degree of Nullstellensatz and Polynomial Calculus proofs of various principles. The table below summarizes the progress thus far. Of particular importance are the formulas expressing the pigeonhole principle, and the formulas expressing various counting principles.

The onto version of the propositional pigeonhole principle states that there is no 1-1, onto map from m to n, $m > n$. This can be expressed by the following equations, with underlying variables $P_{i,j}$, $i \leq m$, $j \leq n$: (1) $P_{i,1} + ... + P_{i,n} - 1 = 0$, for all $i \leq m$; (2) $P_{1,j} + ... + P_{m,j} - 1 = 0$, for all $j \leq n$; and (3) $P_{i,k}P_{j,k} = 0$, for all $i, j \leq m$, $k \leq n$. For each n, let the above set of equations be denoted by $\neg PHP^{m,n}_{onto}$. For each $m = n + 1$, there is a constant degree Nullstellensatz proof over GF_p of $\neg PHP^{m,n}_{onto}$. The proof is obtained by adding together all of the above equations in (1) and subtracting all of the above equations in (2). Each variable will cancel because it occurs once positively in (1) and once negatively in (2), and we are left with $m - n = 1$. However, for $m = nmodp$, this proof fails.

The more general version of the propositional pigeonhole principle states that there is no 1-1 map from m to n. For each $m > n$, the general pigeonhole principle can be expressed by equations (1) and (3) above, and is denoted by $\neg PHP^{m,n}$.

The mod q counting principle, Mod^q_n, states that there is no way to partition a set of size n into equivalence classes, each of size exactly q. For each n, the negation of this principle ($\neg Mod^q_n$) can be expressed by the following equations, with underlying variables X_e, $e \subseteq [1, .., m]$, $|e| = q$, $m = pn + 1$:

(1) $\sum_{e,\ i\in e} X_e - 1 = 0$, for all $i \leq m$; (2) $X_e X_f = 0$, for all e, f, $e \cap f \neq 0$.

The induction principle was explained earlier. The principle Homesitting is a variant of strong induction. The principle Graph, stands for Tseitin's graph tautologies: given a connected graph, where each vertex has a 0-1 labelling (charge) and such that the mod 2 sum of all labellings is odd, the principle states that the mod 2 sum of the edges coming into each vertex is equal to the charge of that vertex. Clearly this principle is unsatisfiable and when the underlying graphs are k-regular and have good expansion properties, the associated formula is hard to prove (as long as the field does not have characteristic 2). Subsetsum is a single equation, $m - \sum_i (c_i x_i) = 0$ and this lower bound shows that over fields of characteristic 0, there are no small Nullstellensatz degree refutations of the subset sum principle. HN means that the degree lower bound holds for Nullstellensatz; PC means that the degree bound holds in the stronger Polynomial Calculus.

By now, there are many families of formulas requiring large Nullstellensatz degree, but a lack of many explicit lower bounds for Polynomial Calculus degree. The first such lower bound for the Polynomial Calculus is the paper by Razborov [19]. In that paper, he explicitly describes the set of all polynomials derivable from the initial equations in degree d. The only other lower bound known for the Polynomial Calculus, due to Krajíček[16], uses important ideas from Ajtai [1] linking the lower bound in question to the representation theory of the symmetric group.

Formulae	Reference	Lower bound	Notes
PHP	[12]	$O(n^{1/4})$ (HN)	nearly optimal
PHP	[19]	$O(n^{1/2})$ (PC)	nearly optimal
ontoPHP	[3]	$O(n)$ (HN)	nearly optimal
IND	[7]	$O(logn)$ (HN)	nearly optimal
Homesitting	[11, 6]	$O(n^{1/2})$ (HN)	
Graph	[14]	$O(n)$ (HN)	$Char(F) \neq 2$
Modp	[4, 1]	nonconstant (HN)	
Modp	[8]	$n^{\Omega(1)}$ (HN)	
Modp	[16]	nonconstant (PC)	
Subsetsum	[8]	$O(n)$ (HN)	$Char(F) = 0$

4.1 The Design Method

In this section we review the primary method that has been used to obtain the above Nullstellensatz degree lower bounds.

Let R be any commutative ring, and let $\mathcal{Q} = \{Q_1, .., Q_m\}$ be a set of unsolvable equations of degree at most 3 over $R[x_1, ..., x_n]$, where m is $n^{O(1)}$. We want to show that there is no degree d set of polynomials $P_1, ..., P_m$ such that $\sum_i P_i Q_i = 1$. Assume for sake of contradiction that degree d P_i's do exist. Write P_i as $\sum_m a_m^i X_m$, where $m \in \{0,1\}^n$, X_m is the corresponding monomial, and a_m^i is the coefficient in front of that monomial in P_i. Because the total number of monomials

in the P_i's is bounded by $n^{O(d)}$, we can write a system of linear equations with the coefficients a^i_m as variables such that the system of linear equations has a solution if and only if such P_i's exist. In particular, the condition $\sum_i P_i Q_i = 1$ can be specified by a system of linear equations in the a^m_i's where for each nonempty monomial m of degree at most $d+3$, we have one equation specifying that the sum of all coefficients in front of this monomial must be 0, and for the empty monomial, we have one equation specifying that the sum of all coefficients in front of the empty monomial must be 1.

Now by weak duality, if we can find a linear combination of the equations such that the left-hand-side of the linear combination is 0, then there can be no solution. (Because the total sum of the right-hand-sides of the equations is 1.) Conversely, if R is a field, then we get the converse direction as well. The name *design* refers to the linear combination of the equations witnessing the fact that the equations can have no solution; because of the structure of the original Q_i's, the properties required of the linear combination can often be seen to be equivalent to the existence of a particular type of combinatorial property, and thus it is called a design.

5 Open Problems

5.1 Lower bounds for stronger proof systems

The most outstanding question is to strengthen these methods to obtain lower bounds for stronger systems, such as $AC^0[2]$-Frege proofs. A solution to this problem seems to be within reach. For this system, a candidate hard tautology is the principle Mod^n_p for p prime.

5.2 Degree lower bounds

Lower bounds and new methods for the degree of Polynomial Calculus proofs for other principles is another important problem. In particular, one can generate random 3CNF formulas with m clauses and n variables and when $m = 4.3n$, such formulas are believed to be hard to refute for all natural proof systems. An open problem is to prove linear degree lower bounds for such formulas. This would show that on average (as opposed to worst-case), unsatisfiable formulas (from this distribution) require large degree proofs.

5.3 Degree versus monomial size

What is the relationship between the minimal degree of a Nullstellensatz/Polynomial Calculus refutation and the minimal number of monomials in a refutation? This is analagous to pinning down exactly the relationship between the minimal Resolution clause width for a formula and the minimal Resolution proof size. Some weak results are known, establishing a connection between them, but they are far from tight [11, 9].

5.4 Representation theory and uniformity

Important work by Ajtai [1] exploits the uniform nature of standard unsatisfiable families of formulas to establish a close connection between Nullstellensatz degree lower bounds and representation theory of the symmetric group. These ideas were further developed by Krajíček[16] to obtain nonconstant degree lower bounds for the Polynomial Calculus. This line of research is quite promising and deserves further study.

5.5 Algebraic Theorem provers

Designing efficient theorem provers for the propositional calculus is an important practical question. To date, Resolution-based algorithms are the champion theorem provers although they are theoretically quite weak as proof systems. A recent challenger is the Polynomial Calculus and more specifically, using variants of the Gröbner basis algorithm to solve 3SAT [11]. This type of algorithm needs to be fine-tuned to the same extent as Resolution based methods and then rigorously evaluated on standard hard examples. On a more theoretical side, can the simulations of Resolution by PC be improved? Another very interesting question is whether or not Cutting Planes can be simulated by efficient PC proofs.

5.6 Natural proofs in proof complexity?

In a major blow to circuit complexity, [20] show that, subject to some plausible cryptographic conjectures, current techniques will be inadequate for obtaining super-polynomial circuit lower bounds. To this point, proof complexity has made steady progress at matching the superpolynomial lower bounds currently known in the circuit world. Unlike the circuit world, however, there is no analogue of Shannon's counting argument for size lower bounds for random functions, and there does not seem to be any inherent reason for Frege lower bounds (and similarly for superpolynoimal lower bounds for algebraic proofs) to be beyond current techniques. Is there any analogue of natural proofs in proof complexity?

References

[1] Ajtai, M., Symmetric systems of linear equations modulo p. Technical Report TR94-015, Electronic Colloquium in Computational Complexity, 1994.

[2] Beame, P., and Pitassi, T., "Propositional Proof Complexity: Past, Present and Future," To appear in the Bulletin of the EATCS, 1998.

[3] Beame, P., and Riis, S., More on the relative strength of counting principles. In *Proof Complexity and Feasible Arithmetics*, Beame and Buss eds., AMS, 1998.

[4] Beame, P., Impagliazzo, R., Krajíček, J., Pitassi, T., Pudĺak, P. (1995) Lower bounds on Hilbert's Nullstellensatz and propositional proofs. *Proceedings of the London Mathematical Society*, 73, 3, 1996, pp.1-26.

[5] Brownawell, D. (1987) Bounds for the degrees in the Nullstellensatz, *Annals of Mathematics* (Second Series), 126: 577-591.

[6] Buss, S., "Lower bounds on Nullstellensatz proofs via designs." In *Proof Complexity and Feasible Arithmetics*, Beame and Buss, eds., AMS, 1998, pp. 59-71.

[7] Buss, S., and Pitassi, T., The complexity of the induction principle. To appear in *JCSS*.

[8] Buss, S., Impagliazzo, R., Krajíček, J., Pudĺak, P., Razborov, S., Sgall, J., Proof complexity in algebraic systems and constant depth Frege systems with modular counting. *Computational Complexity,* 6, pp. 256-298, 1997.

[9] Ben Sasson, E., and Wigderson, A., Private communication, 1998.

[10] Caniglia, L., Galligo, A., and Heintz, J. Some new effectivity bounds in computational geometry. *Proc. 6th Int'l Conference on Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, Ed. T.Mora, pp. 131-151. LNCS 357, 1989.

[11] Clegg, M., Edmonds, J., and Impagliazzo, R. Using the Groebner basis algorithm to find proofs of unsatisfiability, *28th ACM STOC*, pp. 174-183, 1996.

[12] Beame, P., Cook, S., Edmonds, J., Impagliazzo, R., and Pitassi, T. The relative complexity of NP search problems, *27th ACM STOC*, pp. 303-314.

[13] Cook, S. A., and Reckhow, A. R. (1979) The relative efficiency of propositional proof systems, *J. Symbolic Logic*, $44(1)$:36-50.

[14] Grigoriev, D., Tseitin's Tautologies and lower bounds for Nullstellensatz proofs. Manuscript, 1998.

[15] Kollár, J. (1988) Sharp effective Nullstellensatz, *J. Amer. Math. Soc.*, $1(4)$:963-975.

[16] Krajíček, J., On the degree of ideal membership proofs from uniform families of polynomials over finite fields. Manuscript, 1997.

[17] Maciel, A., and Pitassi, T., On $ACC^0[p]$-Frege proofs. In *Proof Complexity and Feasible Arithmetics*, Beame and Buss eds., AMS, 1998.

[18] Pitassi, T., Algebraic propositonal proof systems. *DIMACS Series in Discrete Math. and Theoretical Computer Science,* 31, pp. 215-244, 1997.

[19] Razborov, A., Lower bounds for the polynomial calculus, To appear in *Computational Complexity.*

[20] Razborov, A., Rudich, S. Natural proofs, in *Proceedings of 26th ACM STOC*, 1994, pp. 204-213.

[21] Urquhart, A., The complexity of propositional proof systems, Survey article to appear in *Journal of Symbolic Logic.*

[22] Valiant, L. G., The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8, pp.410-421, 1979.

Toniann Pitassi
Department of Computer Science
University of Arizona
Tucson, Arizona 85719
toni@cs.arizona.edu

Probabilistic Verification of Proofs

Madhu Sudan

Abstract. Recent research in the theory of computing has led to the following intriguing result. "There exists a probabilistic verifier for proofs of mathematical assertions that looks at a proof in only a constant number of bit positions and satisfies the following properties: (Completeness) For every valid theorem there exists a proof that is always accepted. (Soundness) For invalid assertions every purported proof is rejected with some positive probability that is independent of the length of the theorem or proof." This result sheds insight into the fundamental complexity class NP and shows that it is equivalent to a seemingly smaller class of languages with efficient probabilistically checkable proofs. This result is especially significant to combinatorial optimization. For many combinatorial optimization problems it demonstrates that the task of finding even nearly-optimal solutions is computationally intractable. In this article we describe some methods used to construct such verifiers.

1991 Mathematics Subject Classification: 68Q10, 68Q15.
Keywords and Phrases: Computational complexity, Algorithms, Combinatorial optimization, Logic, Probability, Approximation.

1 Introduction

The notion of efficient verification of proofs has been a central theme in the theory of computing. The computational view of this notion abstracts the semantics of the proof system into a verification procedure or *verifier*, i.e., a polynomial time computable Boolean function described by a Turing machine. A purported theorem T and proof π are then just a sequence of bits; π proves T if the verifier accepts the pair (T, π). The purported theorem T is true if such a proof π exists. The class NP [15, 32] represents the class of all theorems with "short" proofs; and allows for very simple combinatorial descriptions of theorems and proofs. As an example, we describe the problem 3-SAT.

A 3cnf formula ϕ is described by N "clauses" $C_1, \ldots, C_N$ on n Boolean variables $x_1, \ldots, x_n$. A clause consists of up to 3 literals (i.e., a variable or its negation) and the clause is satisfied by some Boolean assignment to the variables if at least one literal is assigned a true value. The formula ϕ is said to be satisfiable, if there exists an assignment to the n variables which simultaneously satisfies all clauses. 3-SAT is the language of all satisfiable 3cnf formulae.

The NP-completeness of 3-SAT may be interpreted as follows: For any system of logic, there exists a polynomial time computable function f such given an assertion T in this system of logic and an integer n, $f(T, 1^n)$ computes a 3cnf formula that is satisfiable if and only if T has a proof of length at most n. Thus, under the equivalence class of polynomial time computation, the satisfying assignment to $f(T, 1^n)$ is a proof for the theorem T, and the statement $f(T, 1^n) \in$ 3-SAT is itself the theorem. While this method of describing theorems and proofs is equivalent to any other system of logic it has some conceptual simplicity. One formal effect that captures this simplicity is that an incorrect proof has a very local error: Such an incorrect proof is an assignment that fails to satisfy at least one clause. Hence the three bits corresponding to the assignment to the variables participating in this clause point give the explicit error in the proof. In other words every incorrect proof has a witness of the error that is at most 3 bits long. This example demonstrates some of the power of the computational view of proofs.

Over the course of the last decade a number of new computational notions of proofs have been proposed and analyzed. The common theme in these definitions is a probabilistic verifier who is allowed some small probability of making an error. One of these notions, known as a probabilistically checkable proof (PCP), is motivated by the following informally stated question: "How fast can the verifier be compared to the size of the proof?" It is easy to establish that a deterministic verifier must at the very least "look" at the whole proof. This however need not be true for probabilistic ones. The notion of "looking at a bit of the proof". can be formalized by providing the verifier with oracle access to the proof, i.e., the verifier can specify the address of a location of the proof and gets back the bit written in that location and this entire process takes only as much time as required to write the address. The number of bits of the proof that are "looked" at is now the number of oracle queries. To contrast such a verifier with the traditional verifier, one also quantifies the amount of randomness used by such a verifier. Thus we define $(r(\cdot), q(\cdot))$-restricted PCP verifier to be a probabilistic verifier with random access to a proof oracle, such that on input x of length n, the verifier tosses at most $r(n)$ coins and accesses the oracle at most $q(n)$ times, where the locations accessed are a function of the random coins. A language L is said to be in class $\mathrm{PCP}(r, q)$ if there exists an $(r(\cdot), q(\cdot))$-restricted PCP verifier V satisfying the following: If $x \in L$ there exists an oracle π such that the verifier V accepts x with probability 1 on oracle access to π. If $x \notin L$, for every π, V accepts x with probability at most $1/2$. Notice that the verifier can make mistakes when $x \notin L$. (The definition of a PCP as defined above is from [6]. Many components in this definition come from earlier works: The notion of probabilistic verifiers was first proposed in [25, 10], as part of a larger definition. The notion of oracle machine verifiers was proposed in [22]. The parameters of interest. i.e., $r(\cdot)$ and $q(\cdot)$, were implicit in [19]. A closely related definition focusing on different parameters, termed transparent proofs, was also studied by [9].)

It is immediate from the definition of PCP that $\mathrm{NP} = \cup_{c>0}\mathrm{PCP}(0, n^c)$ (the verifier is not randomized, but is allowed unlimited access to the proof). The results of [8, 9, 19] showed that by allowing the verifier small amounts of randomness, the query complexity can be reduced dramatically and in particular

NP $\subseteq \cup_{c>0}$PCP$(c \log n \log\log n, c \log n \log\log n)$. Subsequently [6, 5] showed that it is possible to restrict the verifier even more significantly to just a *constant* number of queries (independent of the theorem, the proof or the system of axioms). They also reduce randomness to strictly logarithmic in input size. Specifically, they show

Theorem 1 *$\exists q < \infty$ such that $NP = \cup_{c>0} PCP(c \log n, q)$.*

The consequences to combinatorial optimization may be described informally as follows: The notion of a PCP verifier allows one to formalize the notion of an "approximately" correct proof; and the strong results obtained above show that such a proof exists if and only if a perfectly correct proof exists. Thus the task of finding an approximately correct proof is as hard as the task of finding a perfectly correct proof. The traditional connection between proofs and optimization [15, 29, 32] now indicates that for some optimization problems (unfortunately, not necessarily natural ones) finding near-optimal solutions should be as hard as finding optimal ones. The statement can actually be formalized and made applicable even to natural optimization problems. This connection was discovered by [19] and its applicability was further extended in [5] to apply to a large number of optimization problems considered in [34].

In this article we describe some of the methods used in the construction of probabilistically checkable proofs, from a very high level. In particular, we describe some of the properties that a probabilistically checkable proof must have. We also give a hint of how such properties are effected. The primary hope is to motivate the reader to read more detailed descriptions. The concluding section includes pointers for further reading as well as to more recent work.

2 Construction and verification of PCPs

In this section we will describe from a high-level the construction of a probabilistically checkable proof. Using the completeness of 3-SAT we will assume that we restrict our attention to theorems of the form $\phi \in$ 3-SAT, where ϕ is a 3cnf formula. It will be useful to think of ϕ as a function mapping $\{0,1\}^n$ to $\{0,1\}$, using the association that 0 represents the Boolean false and 1 represents the Boolean true. $\phi(\vec{a}) = 1$ if ϕ is satisfied by the assignment $\vec{a} \in \{0,1\}^n$. We will switch between 3 possible views of a, the proof of the theorem $\phi \in$ 3-SAT. a may be thought of as a string, as a vector (over some appropriate field containing 0 and 1), or as an oracle that on query i responds with the ith coordinate of a.

Recall that our goal is to describe an alternate proof for $\phi \in$ 3-SAT. More importantly we wish to describe a new probabilistic verifier V for proofs of satisfiability of 3cnf formulae. The verifier will make "few" queries to the new proof, an oracle A, and then cast a verdict. If $\phi \notin$ 3-SAT, then no oracle satisfies the verifier with probability 1/2, while if $\phi \in$ 3-SAT, then there exists an oracle A such that V always accepts. In the latter case there exists a effective transformation T which transforms the proof $a \in \{0,1\}^n$ satisfying $\phi(a) = 1$ into the the oracle A. It is this transformation that will be our primary focus. For reasons of

space, we will focus on the weaker goal of describing a verifier V that makes only $q = q(n) = (\log n)^{O(1)}$ queries and the transformation T for such a verifier.

2.1 MOTIVATION

We start by examining some properties such a transformation T necessarily exhibits. The first interesting property exhibited by $A = T(a)$ is its redundancy. Let us view A as a string, and suppose $\tilde{A}$ is a string obtained by randomly choosing a small fraction of the bits of A and changing them (from 0 to 1 and vice versa). A PCP verifier making q queries still accepts with probability the proof $\tilde{A}$ with high probability, where this probability tends to 1 as the fraction of errors in the proof tend to 0. Furthermore it is possible to determine the acceptance probability of the verifier on string $\tilde{A}$ in polynomial time. Thus even though $\tilde{A}$ is far from A, it preserves its "meaning" (i.e., continues to prove the statement $\phi \in$ 3-SAT.) The easiest conceivable way to achieve such an effect is to insist that $\tilde{A}$ preserves the original proof. i.e., $\vec{a}$ itself (despite the fact that 1% of its bits are erroneous). This leads us to the first property of the transformation $T(\cdot)$ that we will try to achieve. T is an error-correcting code, i.e., for any two strings a_1 and a_2, $T(a_1)$ and $T(a_2)$ differ in a constant fraction of the bits.

In particular this implies that T is an expansive mapping i.e., maps $\{0,1\}^n \to \{0,1\}^N$ for $N > n$ and hence there are many strings in the range that T does not map to. Given a formula $\phi \notin$ 3-SAT and a string $\tilde{A}$, the PCP verifier has to reject the offered proof with probability at least $\epsilon > 0$ after reading just q bits in such a proof. Furthermore, when the verifier rejects the proof, it must offer an explicit error in the proof in the three bits it reads. The error described may either claim (1) $\tilde{A}$ is not describing any string in the image of T; or (2) $\tilde{A}$ may be the encoding of some string $\vec{a}$; but $\phi(\vec{a}) \neq 1$ for any such string.

To use error of the form (1) above with some string A, it must be that there exist indices $i_1, \ldots, i_q \in [N]$ such that for any string $T(\vec{a})$, the projection to the coordinates $i_1, \ldots, i_q$ does not agree with the projection of A to the same coordinates. We say that an error-correcting code T is q-locally checkable if for every string A that is not in the range of T, there exist indices $i_1, \ldots, i_q \in [N]$ with this property. It will be our goal to come up with an appropriate q-locally checkable code T, for relatively small q.

Finally, T will need to have a "semantic" part: i.e., somehow T must be dependent on ϕ, in order for it to exploit the error condition in (2) above. Summarizing, in the next sections we will describe a transformation T that is an error-correcting code, with good local checkability, that will somehow reveal the truth of the statement $\phi \in$ 3-SAT.

2.2 THE TRANSFORMATION

We start with a simple transformation which leads to some error-correction properties. (Here and later we use $[n]$ to denote the set $\{1, \ldots, n\}$.) The simplest method for adding some error-correcting feature to any information string is to encode it using the Reed-Solomon code. Specifically, to encode the information string $a_1, \ldots, a_n$ we pick a finite field F of order $\Omega(n)$ and an injective map $b : [n] \to F$.

We then pick a polynomial $P_a : F \to F$ of degree at most n such that $P_a(b(i)) = a_i$ for $i \in [n]$ and our first transformation, denoted T_1 is given by $T_1(a) = (P_a(z))_{z \in F}$ be the encoding of a. Notice that the encoding does not give us a string in $\{0,1\}^N$, but rather an element of $F^{|F|}$ that we view as a string over F. Using the elementary property that two distinct degree n polynomials can agree in at most n places, we find that this transformation is very redundant. Specifically $T_1(a_1)$ and $T_1(a_2)$ have a Hamming distance of at least $|F| - n$ when viewed as strings over F.

The above transformation has the right error-correcting property, but lacks local checkability. To get this additional property, we use the idea of encoding using multivariate polynomials. Specifically, we pick an integer m, and field F (whose size will be determined shortly), a set $H \subset F$, such that $|H|^m \geq n$ and an injective function $b : [n] \to H^m$. To encode a string $\vec{a}$, we pick an m-variate polynomial $P_a : F^m \to F$ of degree $|H|$ in each variable such that $P_a(b(i)) = a_i$ for every $i \in [n]$. (It is easy to prove that such a polynomial P_a always exists.) The total degree of such a polynomial is at most $m|H|$. The encoding of a is then simply the string $T_2(a) = (P_a(z_1, \ldots, z_m))_{z_1,\ldots,z_m \in F}$. Thus $T_2 : \{0,1\}^n \to F^{|F|^m}$ and satisfies the following distance property. For any pair of strings a_1 and a_2, $T_2(a_1)$ and $T_2(a_2)$ agree in at most $m|H|/|F|$ fraction of all indices, when viewed as strings over F. This property follows from a well-known extension of the distance property of polynomials to the multivariate case, which states that a (multivariate) polynomial of total degree d can be zero on at most $d/|F|$ fraction of the domain.

The advantage in using the multivariate polynomials is that they exhibit significantly better local checkability properties. In particular, for every function $Q : F^m \to F$ that is not a polynomial of degree d, there exist $d + 2$ points that "prove" this property. We are now ready to describe some useful choices of m, $|H|$ and $|F|$. (Incidentally, the choice of the function $b : [n] \to H^m$ does not affect the performance of the transformation T_2 in any way.) To get a good locally checkable code one would like to minimize the degree which is at most $m|H|$. However, the choice has to satisfy $|H|^m \geq n$. To ensure that $T_2(a)$ is not too long compared to a, one needs to ensure that $|F|^m$ is only polynomially larger $|H|^m$, which implies $|F|$ should be a polynomial in $|H|$. Furthermore, to get a constant distance, $|F|$ better be larger than the total degree (by at least a constant factor). One such choice of parameters is (we omit floors and ceilings in the following choices): $m = \frac{\log n}{\log \log n}$, $|H| = \log n$, and $|F| = (\log n)^2$. This creates a transformation T_2 which maps n bits to n^2 elements from a field of size $\log^2 n$, with degree and hence q-local checkability for $q \leq \log^2 n$.

We now bring in the semantic element to the error correcting code. This will take some development, so we first outline the plan for this stage. In the final construction $T(a) = T_\phi(a)$ will be a sequence of polynomials $f_0 : F^m \to F$ and $f_1, \ldots, f_k : F^{m'} \to F$, described by their value at every input in $F^{m'}$. (k, m' will be specified later.) The value of the polynomial f_i at some point $u \in F^{m'}$ will be determined by a simple formula — or "construction rule" — applied to the value of the polynomial f_{i-1} at some l places $\psi_{i,1}(u), \ldots, \psi_{i,l}(u)$. (Again, l will be determined shortly.) The polynomial f_0 will be $T_2(a)$. The rules will be constructed so that f_k is identically zero if and only if a satisfies ϕ.

To get such a sequence, we start by "arithmetizing" the notion of a 3-SAT

formula and the notion of satisfiability of a clause. Recall that a variable is specified by an index in $[n]$ and thus a literal can be specified as an element of $[n] \times \{0,1\}$. A clause is a triple of literals and thus an element of $[n] \times [n] \times [n] \times \{0,1\}^3$. A 3cnf formula ϕ can thus be described by a function ϕ' mapping $[n] \times [n] \times [n] \times \{0,1\}^3$ to $\{0,1\}$. $\phi'(i,j,k,b_1,b_2,b_3) = 1$ iff the clause with literals (i,b_1), (j,b_2) and (k,b_3) occurs in ϕ. Using the function $b : [n] \to H^m$, we can identify a clause with an element of H^{3m+3}. We say that a polynomial $\hat{\phi} : F^{3m+3} \to F$ of degree less than $|H|$ in each variable is the arithmetization of ϕ if $\phi'(i,j,k,b_1,b_2,b_3) = \hat{\phi}(b(i),b(j),b(k),b_1,b_2,b_3)$ for any $i,j,k \in [n]$ and $b_1,b_2,b_3 \in \{0,1\}$. We will fix $\hat{\phi}(u) = 0$ for all other $u \in H^{3m+3}$. As claimed earlier, it can be shown that such a polynomial $\hat{\phi}$ does exist and that it can be computed in polynomial time from ϕ.

We now move on to the task of arithmetizing the notion of satisfiability. Given a clause C on literals (i,b_1) (j,b_2) and (k,b_3), and an assignment $a_1,\ldots,a_n$ that has been transformed by the transformation T_2 into the polynomial $f_0 : F^m \to F$, notice that the formula $(f_0(b(i)) - b_1) \cdot (f_0(b(j)) - b_2) \cdot (f_0(b(k)) - b_3)$ is 0 if and only if the clause C is satisfied. This leads us to the definition of the polynomial $f_1 : F^{3m+3} \to F$ to be

$$f_1(u,v,w,b_1,b_2,b_3) = \hat{\phi}(u,v,w,b_1,b_2,b_3)(f_0(u) - b_1)(f_0(v) - b_2)(f_0(w) - b_3) \quad (1)$$

where $u,v,w \in F^m$ and $b_1,b_2,b_3 \in F$. By construction it is clear that f_1 is a polynomial on $m' = 3m+3$ variables having degree at most $2|H|$ in each variable; and furthermore is identically zero on the domain $H^{m'}$ if and only if ϕ is satisfied by the assignment a.

This is close in spirit to what we desire. In what follows, we will develop a sequence of polynomials which will move the condition on f_1 being zero on the domain H^m to the condition that $f_{m'+1}$ being zero on $F^{m'}$. This will be achieved inductively: specifically we will define $f_i : F^{m'} \to F$ to be such that f_{i+1} is zero on the domain $F^i \times H^{m-i}$ if and only if f_i is zero on the domain $F^{i-1} \times H^{m-i+1}$. This is achieved by the following rule.

$$f_{i+1}(r_1,\ldots,r_i;z_{i+1},\ldots,z_{m'}) = \sum_{j=1}^{|H|} r_i^j \cdot f_i(r_1,\ldots,r_{i-1};\zeta_j;z_{i+1},\ldots,z_{m'}), \quad (2)$$

where $\zeta_1,\ldots,\zeta_{|H|}$ is any enumeration of the elements of H. It is easy to argue that the polynomials f_{i+1} satisfies the desired property by an inductive argument on i. This concludes the transformation T, that we summarize as follows: given ϕ, a, we pick a field F, a subset H, an integer m, an injective map $b : [n] \to H^m$ and an enumeration $\zeta_1,\ldots,\zeta_{|H|}$ of the elements of H. We then let $f_0 = T_2(a)$, f_1 be as defined by (1), $f_2,\ldots,f_{m'+1}$ be as defined by (2) and let $T(a) = (f_0,\ldots,f_{m'+1})$.

2.3 The verification

We now describe the verifier for the transformation T. The verifier will be given oracles for functions $f_0,\ldots,f_{m'+1}$ and needs to verify that (a) The oracles f_i, $i \in \{0,m'+1\}$ describe polynomials of the correct degree. (b) For every $i \in [m'+1]$, the

polynomial f_i is constructed from p_{i-1} using (1) or (2) (as appropriate). (c) The polynomial $f_{m'+1}$ is identically zero. Assuming that the functions $f_0, \ldots, f_{m'+1}$ are indeed polynomials of the correct degree, (b) and (c) can be verified very easily, probabilistically. To verify (c) the verifier queries the oracle $f_{m'+1}$ at a randomly chosen input $u \in F^{m'}$. By the distance property of polynomials, if $f_{m'+1}$ is not identically zero then $f_{m'+1}(u) \neq 0$ with high probability. To check (b) the verifier checks that the appropriate rule (1) or (2) holds for the oracles f_{i+1} and f_i for randomly chosen u. If the polynomial f_{i+1} is not identical to the polynomial obtained by applying the rule to f_i, then this difference will be witnessed by the point u with high probability. (Once again the distance property of polynomial is being used here.)

Thus the entire verification process reduces to the task of checking condition (a). A test for this condition is termed a "low-degree test" and has been a subject of active investigation recently. Specifically a low-degree test probabilistically queries q locations in an oracle Q and behaves as follows: If Q is a polynomial of total degree d, then the test accepts with probability 1. If the test accepts with probability $1-\delta$, then there is a polynomial $P : F^w \to F$ such that Q and P agree in all but at most ϵ fraction of the inputs, where ϵ, δ are parameters associated with the test.

To test that Q is a polynomial of degree at most d, one exploits the geometry of the space F^m as follows: For $u, v \in F^m$ and $t \in F$, let $l_{u,v}(t) = u + tv$ and let the line through u with slope v, denoted $l_{u,v}$, be the parameterized set of points $\{l_{u,v}(t) | t \in F\}$. It is immediate that for any polynomial $P : F^m \to F$ of degree at most and a line $l = l_{u,v}$, the function $P|_l : F \to F$ given by $P|_l(t) = P(l_{u,v}(t))$ is a univariate polynomial of degree at most d. Based on this observation a low-degree test was proposed in [37]: "Pick u, v uniformly and independently at random from F^m and verify that the points $\{(t, Q(l_{u,v}(t)) | t \in F\}$ are described by a univariate polynomial of degree at most d." It is clear that the tester makes $|F|$ queries to the oracle for Q and accepts all degree d polynomials. The "converse" is harder to prove and we will not attempt to hint at the proof here. A sequence of results [37, 6, 5, 36, 7] concludes showing that this test works for every $\epsilon < \delta < 1$, provided $|F|$ is polynomially larger that $d/(1-\delta)$.

We are still not done, since the low-degree test does not guarantee that the oracle Q is always equal to a low-degree polynomial, but only close to one. To patch this problem, we again resort to the error-correcting nature of polynomials; by using a probabilistic (and highly-efficient) error-correcting algorithm C for low-degree polynomials, due to [11]. C will have oracle access to some function $Q : F^w \to F$ and behave as follows on input $u \in F^w$: If Q is a degree d polynomial, it will return $Q(u)$. If Q is ϵ-close to a degree d polynomial P, it will return $P(u)$ or "error" with high probability (over its internal coin tosses). Again C uses the property of lines in F^m. "Given $\vec{u} \in F^w$ C picks at random $\vec{v} \in F^w$ and considers the function $q(t) = Q|_{l_{\vec{u},\vec{v}}}(t)$. If q is a polynomial in t of degree at most d, then outputs $q(0)$ else outputs error." A simple probabilistic argument shows that C has the desired properties for every $\epsilon < 1$, provided F is large enough.

We are now ready to specify the complete PCP verifier for verifying $\phi \in$ 3-SAT. The verifier has access to the oracles $f_0, \ldots, f_{m'+1}$. It performs a low-

degree test on every oracle f_i, $i \in \{0, \ldots, m'+1\}$. If all low-degree tests pass, it then picks a random $u \in F^{m'}$ and verifies that for every i that the oracles $C^{f_{i+1}}$ satisfies the appropriate rule (1 or 2) w.r.t C^{f_i}. (Notice that we are now working with the oracles C^{f_i} rather than f_i. This is the right choice, since C^{f_i} is a polynomial — not merely close to one.) Finally it checks that $C^{f_{m'+1}}(u) = 0$. If all checks pass, then it accepts the proof, else it rejects the proof. Thus for the choices of m, $|F|$ etc. as made above, the construction yields a verifier making a total of $O(\log^c n)$ queries to all the oracles, for some absolute constant c. A formalization of the arguments above yields (modulo the analysis of the low-degree test) that the verifier accepts incorrect proofs with probability $o(1)$, as $n \to \infty$. Thus we conclude:

THEOREM 2 *$NP \subseteq PCP(O(\log n), \log^c n)$.*

NOTES The result from Theorem 2 is essentially due to [8, 9], though the randomness efficiency was not reduced to $O(\log n)$ until the work of [6]. To get the full effect of Theorem 1 a number of new ideas are required. A central theme is a paradigm to compose proof systems, developed by [6]. In addition [5] present two new PCP constructions to prove Theorem 1. The interested reader may read the original papers for further details. Additional details may be found in [1, 38].

Subsequently there has been a significant amount of work improving the constant q of Theorem 1. This quest was initiated in [13] and further pursued in [20, 14, 35, 12]. Recently, a surprisingly sharp result, essentially showing $q = 3$, has been obtained by [27] (see also [23] for a variant of this result). This work introduces novel techniques to analyze the soundness of verifiers and while the result does rely on some prior work, may be read completely independently.

The consequences to optimization problems have also improved significantly since the initial works of [19, 5]. In particular, a number of new optimization problems have been related to PCPs and sharp results obtained in [33, 3, 21, 18, 26, 39]. Detailed surveys of such connections are available in [4, 17]. The connections have also motivated some new systematic study of combinatorial optimization problems — see [16, 31, 30].

The renewed interest in the approximability of optimization problems has also resulted, surprisingly, in a new spurt in algorithmic results. Particularly striking results in this direction are [24, 2]. Some of these algorithmic results, in particular [28, 40], are needed to analyze the tightness of the new PCP constructions of [27].

REFERENCES

[1] S. ARORA. *Probabilistic Checking of Proofs and Hardness of Approximation Problems.* PhD thesis, U.C. Berkeley, 1994. Available from `http://www.cs.princeton.edu/~arora`.

[2] S. ARORA. Polynomial-time approximation schemes for Euclidean TSP and other geometric problems. *Proc.* 37th *Symposium on Foundations of Computer Science*, IEEE, 1996.

[3] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *J. Computer and System Sciences*, 54(2):317-331, April 1997.

[4] S. Arora and C. Lund. Hardness of approximations. In *Approximation Algorithms for NP-hard problems*, D. Hochbaum, ed. PWS Publishing, 1996.

[5] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximations. To appear *J. ACM*, 45(3), 1998.

[6] S. Arora and S. Safra. Probabilistic checking of proofs: a new characterization of NP. *J. ACM*, 45(1):70-122, 1998.

[7] S. Arora and M. Sudan. Improved low degree testing and its applications. *Proc.* 29th *Annual Symposium on Theory of Computing*, ACM, 1997.

[8] L. Babai, L. Fortnow, and C. Lund. Non-deterministic exponential time has two-prover interactive protocols. *Computational Complexity*, 1:3-40, 1991.

[9] L. Babai, L. Fortnow, L. Levin, and M. Szegedy. Checking computations in polylogarithmic time. *Proc.* 23rd *Annual Symposium on Theory of Computing*, ACM, 1991.

[10] L. Babai and S. Moran. Arthur-Merlin games: a randomized proof system, and a hierarchy of complexity classes. *J. Computer and System Sciences*, 36(2):254-276, 1988.

[11] D. Beaver and J. Feigenbaum. Hiding instances in multioracle queries. Proc. *Symposium on Theoretical Aspects of Computer Science*, 1990.

[12] M. Bellare, O. Goldreich and M. Sudan. Free bits, PCPs and non-approximability — towards tight results. *SIAM J. Computing,* 27(3):804-915, 1998.

[13] M. Bellare, S. Goldwasser, C. Lund, and A. Russell. Efficient probabilistically checkable proofs. *Proc.* 25th *Annual Symposium on Theory of Computing*, ACM, 1993.

[14] M. Bellare and M. Sudan. Improved non-approximability results. *Proc.* 26th *Annual Symposium on Theory of Computing*, ACM, 1994.

[15] S. Cook. The complexity of theorem-proving procedures. *Proc.* 3rd *Annual Symposium on Theory of Computing*, ACM, 1971.

[16] N. Creignou. A dichotomy theorem for maximum generalized satisfiability problems. *J. Computer and System Sciences*, 51(3):511-522, 1995.

[17] P. Crescenzi and V. Kann, A compendium of NP optimization problems. Technical Report, Dipartimento di Scienze dell'Informazione, Università di Roma "La Sapienza", SI/RR-95/02, 1995. Available from `http://www.nada.kth.se/ viggo/problemlist/compendium.html`.

[18] U. Feige. A threshold of ln n for Set Cover. *Proc.* 28th *Annual Symposium on Theory of Computing*, ACM, 1996.

[19] U. Feige, S. Goldwasser, L. Lovasz, S. Safra, and M. Szegedy. Interactive proofs and the hardness of approximating cliques. *J. ACM*, 43(2):268-292, 1996.

[20] U. Feige and J. Kilian. Two prover protocols — Low error at affordable rates. *Proc.* 26th *Annual Symposium on Theory of Computing*, ACM, 1994.

[21] U. Feige and J. Kilian. Zero knowledge and chromatic number. *Proc.* 11th *Annual Conference on Structure in Complexity Theory*, IEEE, 1996.

[22] L. Fortnow, J. Rompel, and M. Sipser. On the power of multi-prover interactive protocols. *Theoretical Computer Science*, 134(2):545-557, 1994.

[23] V. GURUSWAMI, D. LEWIN, M. SUDAN AND L. TREVISAN. A tight characterization of NP with 3 query PCPs. *ECCC* Tech. Report TR98-034, 1998. Available from `http://www.eccc.uni-trier.de/eccc/`.

[24] M. GOEMANS AND D. WILLIAMSON. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115-1145, 1995.

[25] S. GOLDWASSER, S. MICALI, AND C. RACKOFF. The knowledge complexity of interactive proof-systems. *SIAM J. Computing*, 18(1):186-208, 1989.

[26] J. HÅSTAD. Clique is hard to approximate within $n^{1-\epsilon}$. *Proc.* 37th *Symposium on Foundations of Computer Science*, IEEE, 1996.

[27] J. HÅSTAD. Some optimal inapproximability results. *Proc.* 29th *Annual Symposium on Theory of Computing*, ACM, 1997.

[28] H. KARLOFF AND U. ZWICK. A 7/8-approximation algorithm for MAX 3SAT? *Proc.* 38th *Symposium on Foundations of Computer Science*, IEEE, 1997.

[29] R. KARP. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, Advances in Computing Research, pp. 85-103. Plenum Press, 1972.

[30] S. KHANNA, M. SUDAN AND L. TREVISAN. Constraint satisfaction: The approximability of minimization problems. *Proc.* 12th *Annual Conference on Structure in Complexity Theory*, IEEE, 1997.

[31] S. KHANNA, M. SUDAN, AND D. P. WILLIAMSON. A complete classification of the approximability of maximization problems derived from Boolean constraint satisfaction. *Proc.* 29th *Annual Symposium on Theory of Computing*, ACM, 1997.

[32] L. LEVIN. Universal'nyĭe perebornyĭe zadachi (Universal search problems : in Russian). *Problemy Peredachi Informatsii*, 9(3):265-266, 1973.

[33] C. LUND AND M. YANNAKAKIS. On the hardness of approximating minimization problems. *J. ACM*, 41(5):960-981, September 1994.

[34] C. PAPADIMITRIOU AND M. YANNAKAKIS. Optimization, approximation and complexity classes. *J. Computer and System Sciences* 43(3):425-440, 1991.

[35] R. RAZ. A parallel repetition theorem. *SIAM J. Computing*, 27(3):763-803, 1998.

[36] R. RAZ AND S. SAFRA. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. *Proc.* 29th *Annual Symposium on Theory of Computing*, ACM, 1997.

[37] R. RUBINFELD AND M. SUDAN. Robust characterizations of polynomials with applications to program testing. *SIAM J. Computing* 25(2):252-271, 1996.

[38] M. SUDAN. *Efficient Checking of Polynomials and Proofs and the Hardness of Approximation Problems.* ACM Distinguished Theses, Lecture Notes in Computer Science, no. 1001, Springer, 1996.

[39] L. TREVISAN. When Hamming meets Euclid: The approximability of geometric TSP and MST. *Proc.* 29th *Annual Symposium on Theory of Computing*, ACM, 1997.

[40] U. ZWICK. Approximation algorithms for constraint satisfaction problems involving at most three variables per constraint. *Proc.* 9th *Annual Symposium on Discrete Algorithms*, ACM-SIAM, 1998.

Madhu Sudan
LCS, MIT, 545 Technology Square
Cambridge, MA 02139, U.S.A.
email: madhu@lcs.mit.edu

Halving Point Sets

Artur Andrzejak and Emo Welzl

Abstract. Given n points in $\mathbb{R}^d$, a hyperplane is called halving if it has at most $\lfloor n/2 \rfloor$ points on either side. How many partitions of a point set (into the points on one side, on the hyperplane, and on the other side) by halving hyperplanes can be realized by an n-point set in $\mathbb{R}^d$?

1991 Mathematics Subject Classification: 52C10, 52B05, 52B55, 68R05, 68Q25, 68U05
Keywords and Phrases: combinatorial geometry, computational geometry, k-sets, k-levels, probabilistic method, matroid optimization, oriented matroids, Upper Bound Theorem.

Consider the following algorithmic problem first. Given n points in $\mathbb{R}^d$, we want to find a hyperplane that minimizes the sum of Euclidean distances to these n points. A glimpse of reflection tells us that an optimal hyperplane cannot have a majority ($\lfloor n/2 \rfloor + 1$ or more) of the points on either side; otherwise a parallel motion towards this side will improve its quality [YKII, KM]. A hyperplane with at most $\lfloor n/2 \rfloor$ points on either side is called *halving*. How many partitions of a point set (into the points on one side, on the hyperplane, and on the other side) by halving hyperplanes can be realized by an n-point set in $\mathbb{R}^d$? The notions and results mentioned below are closely related to this question. Emphasis in the presentation is on techniques that may be useful elsewhere, and on interconnections to other topics in discrete geometry and algorithms. A more complete treatment is in preparation [AW].

Halving edges and a crossing lemma

Let P be a set of n points in the plane, n even, and no three points on a line. A *halving edge* is an undirected edge between two points, such that the connecting line has the same number of points on either side. Around 1970 L. Lovász [Lo] and P. Erdős et al. [ELSS] were the first to investigate the geometric graph of halving edges of a point set, and proved that there cannot be more than $O(n^{3/2})$ such edges. Except for a small improvement to $O(n^{3/2}/\log^* n)$ [PSS], there was no progress on the problem until T. Dey [De] recently gave an upper bound of $O(n^{4/3})$. He shows that the graph of halving edges cannot have more than $O(n^2)$ pairs of crossing edges. Then he employs a *crossing lemma* (due to M. Ajtai et al. [ACNS] and T. Leighton [Le]), which has a number of other applications: A geometric

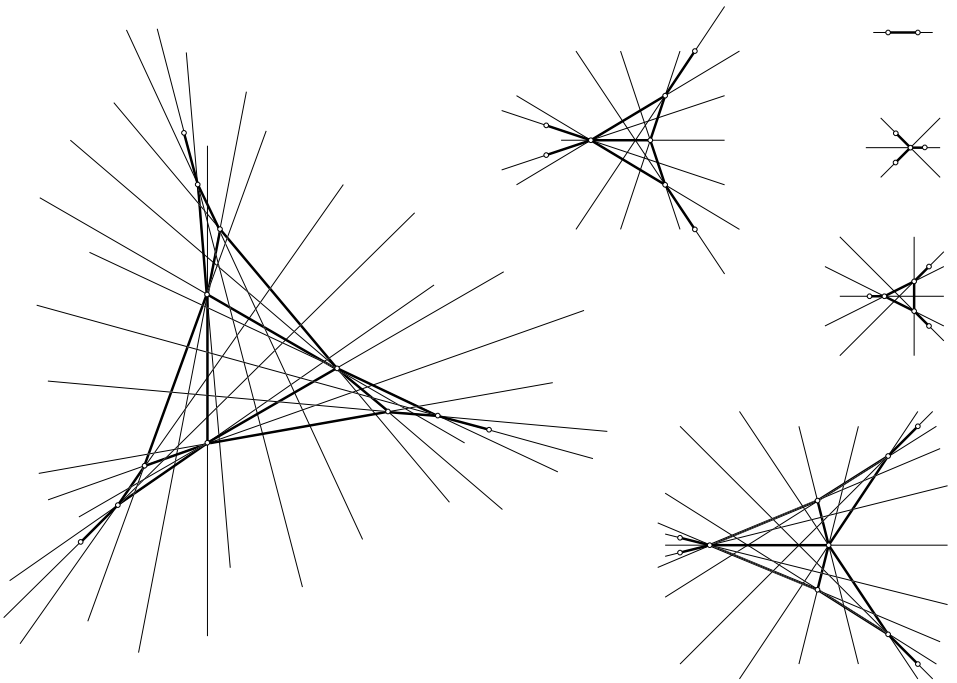


Figure 1: Graphs of halving edges. The configurations maximize the number of halving edges for the given number of points [AAHSW]. Note that, in general, graphs of halving edges are not plane!

graph with n vertices and c pairs of crossing edges has at most $O(\max(n, \sqrt[3]{cn^2}))$ edges.

A variation of Dey's proof ([AAHSW]) goes via the following identity.

LEMMA 1

$$C + \sum_{p \in P} \binom{(\deg p + 1)/2}{2} = \binom{n/2}{2}$$

where $\deg p$ *is the number of halving edges incident to* p *(this number is always odd), and* C *is the number of pairwise crossings of halving edges.*

The lemma shows that the number of pairwise crossings in a graph of halving edges is bounded by $\binom{n/2}{2} < n^2/4$. We will now prove the implication on the number, m, of halving edges of P. Recall that a geometric graph without crossings of edges has at most $3n - 6$ edges. Now we choose a random induced subgraph G_x of the graph of halving edges of P by taking each point with probability[1] $x = 2/\sqrt[3]{n}$, independently from the other points. Let P_x be the resulting point set, let m_x be the number of halving edges of P with both endpoints in P_x, and let C_x be the

[1]Here we have to assume that $n \geq 8$.

number of pairwise crossings among such edges. We know that $m_x - C_x \leq 3n_x - 6$, since all crossings in G_x can be removed by deletion of C_x edges. The expected value for n_x is xn, for m_x it is $x^2 m$, and for C_x it is $x^4 C$. Hence, due to linearity of expectation, $x^2 m - x^4 C \leq 3xn - 6$, which gives[2]

$$m \leq x^2 C + 3n/x - 6/x^2 \leq \frac{4}{\sqrt[3]{n^2}} \frac{n^2}{4} + 3n \frac{\sqrt[3]{n}}{2} = \frac{5}{2} n^{4/3} \ .$$

For a proof of Lemma 1, we first observe that the identity holds if P is the set of vertices of a regular n-gon. Then the halving edges are connecting the antipodal vertices of the polygon. We have $n/2$ halving edges, $\deg p = 1$ for all points, and any pair of halving edges crosses. An alternative example is given by the vertices of a regular $(n-1)$-gon together with its center. Then the halving edges connect this center with the other points, with no crossing of halving edges. In a second and final step one verifies that the identity remains valid under continuous motion of a point set. We will not go through this argument, but we mention here a lemma due to L. Lovász [Lo], which is essential for this argument and for most proofs in this context.

LEMMA 2 *Let line ℓ contain a unique point p in P. Assume there are x halving edges incident to p emanating into the side of ℓ which contains less points from P than the other side of*[3] *ℓ. Then there are $x+1$ halving edges emanating into the other side of ℓ.*

The lemma can be proven by rotating a line λ about point p starting in position ℓ until it coincides with ℓ again. The halving edges incident to p are encountered in alternation on the large and small side of ℓ, starting and ending on the large side.

It is remarkable, that the graph of halving edges is the *unique* graph that satisfies Lemma 2, i.e., it completely characterizes the graph of halving edges of a point set. Simple implications of the lemma are that the number of halving edges incident to a point in P is always odd, and that there is exactly one halving edge incident to each extreme point of P. Moreover, we have the following implication, which, in fact, is equivalent to Lemma 2.

COROLLARY 1 *Let ℓ be a line disjoint from P with x points from P on one side and y points on the other side, $x + y = n$. Then ℓ crosses* $\min(x, y)$ *halving edges of P.*

The corresponding problem of bounding the number of *halving triangles* of n points in $\mathbb{R}^3$, n odd, has also been investigated in a sequence of papers with a currently best bound of $O(n^{8/3})$ due to T. Dey and H. Edelsbrunner [DE]. Building blocks of the proof are a probabilistic argument similar to the one given above, and a counterpart of Corollary 1: No line crosses more than $n^2/8$ halving triangles.

While the bound in $\mathbb{R}^3$ still allows for a simple proof, the situation gets more involved in dimensions 4 and higher, where the best bounds due to P. Agarwal et al. [AACS] are based on a colored version of Tverberg's Theorem [Tv] by R. T. Živaljević and S. T. Vrećica [ZV].

[2] The general bound of $O(\max(n, \sqrt[3]{cn^2}))$ in the crossing lemma [ACNS, Le] is obtained with $x = \min(1, \sqrt[3]{n/c})$. The best known constant in the asymptotic bound can be found in [PT].

[3] This side is unique, since ℓ contains a point, and $|P|$ is even.

k-LEVELS AND PARAMETRIC MATROID OPTIMIZATION.

Let H be a set of n non-vertical lines in $\mathbb{R}^2$. For $0 \leq k \leq n-1$, the k-level of the arrangement of H is the set of all points which have at most k lines below and at most $n-k-1$ above. Clearly, points on the k-level must lie on at least one line. Moreover, the k-level can be easily seen to be an x-monotone polygonal curve from $-\infty$ to $+\infty$, since it intersects every vertical line in exactly one point.

We will now show how the halving edges of a planar point set P, $|P|$ even, correspond to vertices of the $(n/2-1)$- and $(n/2)$-level of some line arrangement. To this end we consider the mapping $p = (a,b) \mapsto p^* : y = ax + b$ from points to non-vertical lines, and the mapping $h : y = kx + d \mapsto h^* = (-k, d)$ from non-vertical lines to points. This mapping preserves incidences and relative position: p lies on h iff p^* contains h^*, and p lies above h iff h^* lies below p^*. Set $P^* = \{p^* | p \in P\}$. Now a pair of points p and q is connected by a halving edge iff the intersection[4] of p^* and q^* lies both on the $(n/2-1)$- and the $(n/2)$-level of the arrangement of P^*.

The results in [De] imply an upper bound of $O(n\sqrt[3]{k+1}\,)$ on the number of vertices on the k-level. k-levels have a number of applications in the analysis of algorithmic problems in geometry. We briefly outline here a connection where the methods for analyzing k-levels proved useful.

A *matroid* of rank k consists of a set of n elements and a non-empty family of k-element subsets, called *bases*. The family of bases is required to fulfill the basis exchange axiom: for two bases B_1, B_2 and an element $x \in B_1 \setminus B_2$ we can always find $y \in B_2 \setminus B_1$ such that $(B_1 \setminus \{x\}) \cup \{y\}$ is again a basis.

Typical examples of matroids are the set of edges of a graph with its spanning trees as bases, or a set of vectors with its bases. If we equip the elements e with weights $w(e)$, we can ask for the minimal weight basis (i.e., the basis with minimal sum of weights). The matroid property ensures that the greedy method finds such an optimal basis.

Assume now that the weights are linear functions $w(e) = k_e\lambda + d_e$ depending on some real number λ [Gu, KI]. While λ ranges from $-\infty$ to ∞, we obtain a sequence of minimal weight bases. How long can this sequence be?

By plotting the weights of the elements along the λ-axis, we obtain an arrangement of n lines. The changes of the minimal weight basis occur at vertices of this arrangement. In the special case of a uniform matroid, i.e., where each set of k elements forms a basis, the changes of minimal weight basis occur at the vertices of the $(k-1)$-level of the line arrangement. N. Katoh was the first to notice this connection.

For general rank k matroids it is known [Ep] that the length of the minimal base sequence is bounded by the total number of vertices of k convex polygons whose edges do not overlap and are drawn from n lines. T. Dey [De] has shown an upper bound $\mathrm{O}(nk^{1/3}+n^{2/3}k^{2/3})$ (which is $\mathrm{O}(nk^{1/3})$ for $k \leq n$) on this quantity by a modification of his proof for the complexity of a k-level. This bound is optimal, due to a lower bound $\Omega(nk^{1/3})$ obtained by D. Eppstein [Ep].

[4]This intersection may vanish to infinity, if the halving edge is vertical.

For graphs the problem looks at the number of different minimal spanning trees for edge weights parameterized by some linear fuction in a parameter λ. The best lower bound for this quantity is $\Omega(n\alpha(k))$ [Ep], where n is the number of edges, $k+1$ is the number of vertices, and α is a slowly growing inverse of the Ackermann function. The known upper bound is the same as for general matroids.

Lower bounds and oriented matroids

The upper bounds mentioned may be far from optimal. In the plane several constructions of n-point sets with $\Omega(n \log n)$ halving edges are known [ELSS, EW, EVW]. If we consider the corresponding problem for oriented matroids (cf. [BLSWZ]) of rank 3 (or pseudoline arrangements in the dual), then there is an unpublished lower bound of $n2^{\Omega(\sqrt{\log n})}$ due to M. Klawe, M. Paterson, and N. Pippenger, inspired by a connection to sorting networks (cf. [AW]). It is open whether this construction is realizable (stretchable) or not.

j-Facets and the Upper Bound Theorem

The following notion generalizes halving edges and triangles. Let P be a set of $n > d$ points in $\mathbb{R}^d$ in general position, i.e., no $d+1$ points on a common hyperplane. A *j-facet of P* is an oriented $(d-1)$-simplex spanned by d points in P that has exactly j points from P on the positive side of its affine hull. The 0-facets correspond to the facets of the convex hull of P. Hence, the Upper Bound Theorem due to P. McMullen [McM] (cf. [Zi]) gives us a tight upper bound on the number of 0-facets, which is attained by the vertices of cyclic polytopes: $2\binom{n-\lfloor d/2\rfloor-1}{\lfloor d/2\rfloor}$ for d odd, and $2\binom{n-\lfloor d/2\rfloor}{\lfloor d/2\rfloor} - \binom{n-\lfloor d/2\rfloor-1}{\lfloor d/2\rfloor}$ for d even. Below we will use the fact that these expressions are upper bounded by $2\binom{n}{\lfloor d/2\rfloor}$. For d fixed, K. L. Clarkson and P. W. Shor [CS] derive an asymptotically tight bound of $O(n^{\lfloor d/2\rfloor}(j+1)^{\lceil d/2\rceil})$ for the number of $(\leq j)$-facets (i.e., i-facets with $0 \leq i \leq j$) by an argument along the following lines.

We use g_j for the number of j-facets of P and G_j for the number of $(\leq j)$-facets, i.e., $G_j = \sum_{i=0}^{j} g_i$. Now fix some j, $0 \leq j \leq n-d$ and x, $0 < x \leq 1$. We take a random sample P_x of P by selecting each point in P with probability x, independently from the other points. Let $n_x = |P_x|$ and let F_x be the number of 0-facets of P_x.

On the one hand, the Upper Bound Theorem implies $F_x \leq 2\binom{n_x}{\lfloor d/2\rfloor}$ and so

$$\mathrm{E}(F_x) \leq 2\binom{n}{\lfloor d/2\rfloor} x^{\lfloor d/2\rfloor} , \tag{1}$$

since $\mathrm{E}(\binom{X}{i}) = \binom{N}{i}x^i$ for a random variable X following the binomial distribution of N Bernoulli trials with success probability x. On the other hand, an i-facet of P appears as a 0-facet of P_x with probability $x^d(1-x)^i$ – we have to select the d points that determine the i-facet, but none of the i points on its positive side.

Hence,

$$\mathrm{E}(F_x) = \sum_{i=0}^{n-d} x^d(1-x)^i g_i \geq x^d(1-x)^j \sum_{i=0}^{j} g_i = x^d(1-x)^j G_j \ . \qquad (2)$$

Combining (1) and (2), we have $G_j \leq 2(1-x)^{-j}\binom{n}{\lfloor d/2\rfloor}x^{-\lceil d/2\rceil}$. By setting $x = \lceil d/2\rceil/(j+\lceil d/2\rceil)$,

$$G_j \leq 2\binom{n}{\lfloor d/2\rfloor}\frac{(j+\lceil d/2\rceil)^{j+\lceil d/2\rceil}}{j^j\lceil d/2\rceil^{\lceil d/2\rceil}} \leq 2\left(\frac{e}{\lceil d/2\rceil}\right)^{\lceil d/2\rceil}\binom{n}{\lfloor d/2\rfloor}(j+\lceil d/2\rceil)^{\lceil d/2\rceil}$$

and the claimed asymptotic bound follows.

Except for dimensions 2 and 3, no exact upper bounds for the number of $(\leq j)$-facets are known. In particular, it is not known whether the *exact* maximum is attained for sets in convex position or not. It is still possible that the exact maximum can be obtained for points on the moment curve, where the number of $(\leq j)$-facets can be easily counted.

We summarize the known bounds for the number of j-facets.

PROPOSITION 1 *Let P be a set of $n > d$ points in $\mathbb{R}^d$ in general position, i.e., no $d+1$ points on a common hyperplane. Let $0 \leq j \leq n-d$.*
(0) *There is a constant $\varepsilon_d > 0$ dependent on d only, such that*

$$g_j = O(n^{\lfloor d/2\rfloor}(j+1)^{\lceil d/2\rceil-\varepsilon_d})$$

[AACS]. *There are point sets with $g_{\lfloor (n-d)/2\rfloor} = \Omega(n^{d-1}\log n)$* [Ed].

$$G_j = O(n^{\lfloor d/2\rfloor}(j+1)^{\lceil d/2\rceil})$$

which, for d fixed, is asymptotically tight for points on the moment curve [CS].
(1) *If $d = 2$ then*

$$g_j = O(n\sqrt[3]{j+1}\,)$$

[De]. $G_j \leq n(j+1)$ *for $j < n/2-1$* [AG, Pe]*, which is tight for points in convex position.*
(2) *If $d = 3$ then*

$$g_j = O(n(j+1)^{5/3})$$

[AACS].

$$G_j \leq (j+1)(j+2)n - 2(j+1)(j+2)(j+3)/3$$

for $j \leq n/4-2$, which is tight if P is in convex position [AAHSW].

AND k-SETS?

We have met halving edges and triangles, k-levels and j-facets, but if the reader inspects the references, she will repeatedly encounter the term 'k-set.' In fact, many people think of the problem in the following setting (although proofs and applications go via the notions we have discussed above):

Let P be a set of n points in $\mathbb{R}^d$. A subset S of P is called *k-set*, if $|S| = k$ and S can be separated from $P \setminus S$ by a hyperplane. The maximum possible number of k-sets of n-point sets in $\mathbb{R}^d$ is related to the maximum possible number of k-facets, although the connection is somewhat subtle [AAHSW, AW].

References

[AACS] Pankaj K. Agarwal, Boris Aronov, Timothy M. Chan, and Micha Sharir, On levels in arrangements of lines, segments, planes, and triangles, *Discrete Comput Geom* (1998), to appear

[ACNS] Miklós Ajtai, Vasek Chvátal, Monty M. Newborn, and Endre Szemerédi, Crossing-free subgraphs, *Ann Discrete Math* 12 (1982) 9–12

[AG] Noga Alon and Ervin Győri, The number of small semispaces of a finite set of points in the plane, *J Combin Theory Ser A* 41 (1986) 154–157

[AAHSW] Artur Andrzejak, Boris Aronov, Sariel Har-Peled, Raimund Seidel, and Emo Welzl, Results on k-sets and j-facets via continuous motion, *in* "Proc 14th Ann ACM Symp on Comput Geom" (1998) 192-199

[AW] Artur Andrzejak and Emo Welzl, k-Sets and j-facets – A tour of discrete geometry, in preparation (1998)

[BLSWZ] Anders Björner, Michel Las Vergnas, Bernd Sturmfels, Neil White, and Günter Ziegler, *Oriented Matroids*, Cambridge University Press, Cambridge UK (1993)

[CS] Kenneth L. Clarkson and Peter W. Shor, Applications of random sampling in computational geometry, II, *Discrete Comput Geom* 4 (1989) 387–421

[De] Tamal K. Dey, Improved bounds for planar k-sets and related problems, *Discrete Comput Geom* 19 (1998) 373-382

[DE] Tamal K. Dey and Herbert Edelsbrunner, Counting triangle crossings and halving planes, *Discrete Comput Geom* 12 (1994) 281–289

[Ed] Herbert Edelsbrunner, *Algorithms in Combinatorial Geometry*, Springer Verlag, Berlin, Heidelberg, New York (1987)

[EVW] Herbert Edelsbrunner, Pavel Valtr, and Emo Welzl, Cutting dense point sets in half, *Discrete Comput Geom* 17 (1997) 243–255

[EW] Herbert Edelsbrunner and Emo Welzl, On the number of line separations of a finite set in the plane, *J Combin Theory Ser A* 38 (1985) 15–29

[ELSS] Paul Erdős, László Lovász, A. Simmons, and Ernst G. Straus, Dissection graphs of planar point sets, *in* A Survey of Combinatorial Theory (Eds. Jagdish N. Srivastava et al.), North Holland Publishing Company (1973) 139–149

[Ep] David Eppstein, Geometric lower bounds for parametric matroid optimization, *in* "Proc 27th Symp Theory Comput" (1995) 662–671

[Gu] Dan Gusfield, Bounds for the parametric spanning tree problem, *in* "Proc Humboldt Conf on Graph Theory, Combinatorics, and Computing", Utilitas Mathematica (1979) 173–183

[KI] Naoki Katoh and Toshihide Ibaraki, On the total number of pivots required for certain parametric problems, Tech. Report Working Paper 71, Inst. Econ. Res., Kobe Univ. Commerce (1983)

[KM] Nikolai M. Korneenko and Horst Martini, Hyperplane approximations and related topics, *in* "New Trends in Discrete and Computational Geometry" (János Pach, Ed.), *Algorithms and Combinatorics* 10 (1993) 135–161

[Le] Tom Leighton, *Complexity Issues in VLSI*, Foundation of Computing Series, MIT Press, Cambridge (1983)

[Lo] László Lovász, On the number of halving lines, *Ann Universitatis Scientarium Budapest, Eötvös, Sectio Mathematica* 14 (1971) 107–108

[McM] Peter McMullen, The maximum number of faces of a convex polytope, *Mathematika* 17 (1971) 179–184

[PSS] János Pach, William Steiger, and Endre Szemerédi, An upper bound on the number of planar k-sets, *Discrete Comput Geom* 7 (1992) 109–123

[PT] János Pach and Géza Toth, Graphs drawn with few crossings per edge, *Lecture Notes in Comput Sci* 1190 (1997) 345–354

[Pe] G. W. Peck, On k-sets in the plane, *Discrete Math* 56 (1985) 73–74

[Sh] Micha Sharir, k-Sets and random hulls, *Combinatorica* 13 (1993) 483–495

[Tv] Helge Tverberg, A generalization of Radon's Theorem, *J London Math Soc* 41 (1966) 123–128

[YKII] Peter Yamamoto, Kenji Kato, Keiko Imai, and Hiroshi Imai, Algorithms for vertical and orthogonal L_1 linear approximation of points, *in* "Proc 4th Ann ACM Symp on Comput Geom" (1988) 352–361

[Zi] Günter Ziegler, *Lectures on Polytopes*, Graduate Texts in Mathematics, Spinger-Verlag (1995)

[ZV] Rade T. Živaljević and Sinisa T. Vrećica, The colored Tverberg's problem and complexes of injective functions, *J Combin Theory Ser A* 61 (1992) 309–318

Artur Andrzejak
Institute of
Theoretical Computer Science
ETH Zürich
CH-8092 Zürich
Switzerland
artur@inf.ethz.ch

Emo Welzl
Institute of
Theoretical Computer Science
ETH Zürich
CH-8092 Zürich
Switzerland
emo@inf.ethz.ch

Section 15

Numerical Analysis and Scientific Computing

In case of several authors, Invited Speakers are marked with a *.

On Multiresolution Methods in Numerical Analysis

Gregory Beylkin

Abstract. As a way to emphasize several distinct features of the multiresolution methods based on wavelets, we describe connections between the multiresolution LU decomposition, multigrid and multiresolution reduction/homogenization for self-adjoint, strictly elliptic operators. We point out that the multiresolution LU decomposition resembles a direct multigrid method (without W-cycles) and that the algorithm scales properly in higher dimensions.

Also, the exponential of these operators is sparse where sparsity is defined as that for a finite but arbitrary precision. We describe time evolution schemes for advection-diffusion equations, in particular the Navier-Stokes equation, based on using sparse operator-valued coefficients. We point out a significant improvement in the stability of such schemes.

1991 Mathematics Subject Classification: 65M55, 65M99, 65F05, 65F50, 65R20, 35J, 76D05
Keywords and Phrases: multigrid methods, fast multipole method, wavelet bases, multiresolution analysis, multiresolution LU decomposition, time evolution schemes, exponential of operators, advection-diffusion equations

1 Introduction

Multiresolution methods have a fairly long history in numerical analysis, going back to the introduction of multigrid methods [10], [18] and even earlier [22]. A renewed interest in multiresolution methods was generated recently by the development of wavelet bases and other bases with controlled time-frequency localization [23], [20], [13], [19], [12], [2], [1], etc.. The introduction of these new tools allows us to relate numerical analysis with harmonic analysis and signal processing by the fundamental need of an efficient representation of operators and functions.

It is useful to compare the wavelet approach with the multigrid method (MG) and the Fast Multipole Method (FMM). For most problems the wavelet approach, FMM, and MG provide the same asymptotic complexity. The differences are typically in the "constants" of the complexity estimates. These differences will, most likely, diminish in the future.

A typical MG is a fast iterative solver based on a hierarchical subdivision. Hierarchical subdivision is also used in FMM which was initially proposed for computing potential interactions [21], [17]. This algorithm requires order N operations to compute all the sums

$$p_j = \sum_{i \neq j} \frac{q_i q_j}{|x_i - x_j|}, \quad \text{where} \quad x_i \in \mathbf{R}^3 \quad i, j = 1, \ldots, N, \tag{1}$$

and the number of operations is independent of the configuration of charges. In the FMM, the reduction of the complexity of computing the sums in (1) from order N^2 to $-N \log \epsilon$, where ϵ is the desired accuracy, is achieved by approximating the far field effect of a cloud of charges located in a box by the effect of a single multipole at the center of the box.

Although both MG and FMM have been extended well beyond their original applications, neither of these methods use the notion of bases in their development and, specifically, orthonormal bases[1]. On the conceptual level using bases makes it easier to consider efficient representations of functions and operators that handle smooth, oscillatory, and scaling behavior.

In particular, to emphasize several distinct features of the wavelet approach, we consider two topics. First, we describe connections between the multiresolution LU decomposition, MG, and multiresolution reduction/homogenization for self-adjoint, strictly elliptic operators. Second, we describe the effects of computing the exponential of such operators on numerical properties of time evolution schemes for advection-diffusion equations.

The essence of the first topic is that multiresolution LU decomposition (the usual LU decomposition interlaced with projections) is equivalent to the direct MG, i.e., a MG without W-cycles. The reason for the absence of W-cycles is that on every scale we construct equations for the orthogonal projection of the *true* solution. Once these equations are solved, there is no need to return to a coarser scale to correct the solution (which is the role of W-cycles in MG). Moreover, equations obtained in this manner on coarser scales are of interest by themselves, since they can be interpreted as "homogenized" or reduced equations, leading to (numerical) multiresolution reduction and homogenization.

The essence of the second topic is that we can drastically improve properties of time evolution schemes for advection-diffusion equations by using the exponential of operators. As it turns out, for self-adjoint, strictly elliptic operators $\mathcal{L}$ the exponential $\exp(-t\mathcal{L})$ is sparse in wavelet bases (for a finite but arbitrary precision) for all $t \geq 0$. This observation makes the construction of $\exp(-t\mathcal{L})$ feasible in two and three spatial dimensions. Given a proper choice of basis and several additional algorithms, we are led to adaptive numerical schemes for the solution of advection-diffusion equations [8].

[1] We note that the representation of functions via their values and via coefficients in an expansion are closely related. In fact if one uses interpolating bases functions then there is a way to simplify this relation (see [3]).

2 Multiresolution Direct Solvers

Direct solvers are not used for problems in multiple dimensions since the standard LU decomposition will fill most of the matrix and, thus, render the method inefficient. This is even without considering additional difficulties due to the high condition numbers typical in these problems. It turns out that both difficulties can be overcome for self-adjoint, strictly elliptic operators by using wavelet bases and multiresolution LU decomposition [7], [16].

As usual, we consider multiresolution analysis (MRA), a chain of subspaces

$$\ldots \subset \mathbf{V}_2 \subset \mathbf{V}_1 \subset \mathbf{V}_0 \subset \mathbf{V}_{-1} \subset \mathbf{V}_{-2} \subset \ldots$$

such that

$$\bigcap_j \mathbf{V}_j = \{0\} \text{ and } \overline{\bigcup_j \mathbf{V}_j} = \mathbf{L}^2(\mathbf{R}^d).$$

Let the subspace $\mathbf{V}_j$ be spanned by an orthonormal basis formed by the tensor product of scaling functions $\{\phi_k^j(\cdot) = 2^{-j/2}\phi(2^{-j}\cdot - k)\}_{k\in\mathbf{Z}}$, where ϕ satisfies the two-scale difference equation (see e.g. [13] for details). Let us denote by $\mathbf{W}_j$ the orthogonal complement of $\mathbf{V}_j$ in $\mathbf{V}_{j-1}$, $\mathbf{V}_{j-1} = \mathbf{V}j \oplus \mathbf{W}j$. We use $\mathbf{P}_j$ and $\mathbf{Q}_j$ to denote the projection operators onto $\mathbf{V}_j$ and $\mathbf{W}_j$. If $x \in \mathbf{V}_j$, we write $s_x = \mathbf{P}_{j+1}x$ and $d_x = \mathbf{Q}_{j+1}x$, where $s_x \in \mathbf{V}_{j+1}$ and $d_x \in \mathbf{W}_{j+1}$.

Given a bounded linear operator $\mathbf{S}$ on $\mathbf{L}^2(\mathbf{R}^d)$, let us consider its projection $\mathbf{S}_j$ on $\mathbf{V}_j$, $\mathbf{S}_j = \mathbf{P}_j\mathbf{S}\mathbf{P}_j$ and represent the operator $\mathbf{S}_j$ as a (possibly infinite) matrix in that basis. With a slight abuse of notation, we will use the same symbol $\mathbf{S}_j$ to represent both the operator and its matrix. Since $\mathbf{V}_j = \mathbf{V}_{j+1} \oplus \mathbf{W}_{j+1}$, we may also write $\mathbf{S}_j : \mathbf{V}_j \to \mathbf{V}_j$ in a block form

$$\mathbf{S}_j = \begin{pmatrix} \mathbf{A}_{\mathbf{S}_j} & \mathbf{B}_{\mathbf{S}_j} \\ \mathbf{C}_{\mathbf{S}_j} & \mathbf{T}_{\mathbf{S}_j} \end{pmatrix} : \mathbf{V}_{j+1} \oplus \mathbf{W}_{j+1} \to \mathbf{V}_{j+1} \oplus \mathbf{W}_{j+1}, \tag{2}$$

where $\mathbf{A}_{\mathbf{S}_j} = \mathbf{Q}_{j+1}\mathbf{S}_j\mathbf{Q}_{j+1}$, $\mathbf{B}_{\mathbf{S}_j} = \mathbf{Q}_{j+1}\mathbf{S}_j\mathbf{P}_{j+1}$, $\mathbf{C}_{\mathbf{S}_j} = \mathbf{P}_{j+1}\mathbf{S}_j\mathbf{Q}_{j+1}$, and $\mathbf{T}_{\mathbf{S}_j} = \mathbf{S}_{j+1} = \mathbf{P}_{j+1}\mathbf{S}_j\mathbf{P}_{j+1}$. Each of the operators may be considered as a matrix and in the matrix form the transition from $\mathbf{S}_j$ in (2) to $\begin{pmatrix} \mathbf{A}_{\mathbf{S}_j} & \mathbf{B}_{\mathbf{S}_j} \\ \mathbf{C}_{\mathbf{S}_j} & \mathbf{T}_{\mathbf{S}_j} \end{pmatrix}$ requires application of the wavelet transform. We refer to $\mathbf{A}_{\mathbf{S}_j}$, $\mathbf{B}_{\mathbf{S}_j}$, $\mathbf{C}_{\mathbf{S}_j}$ and $\mathbf{T}_{\mathbf{S}_j}$ as the $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{T}$ *blocks* of $\mathbf{S}_j$.

Consider a bounded linear operator $\mathbf{S}_j : \mathbf{V}_j \to \mathbf{V}_j$ and a linear equation

$$\mathbf{S}_j x = f, \tag{3}$$

which we may write as

$$\begin{pmatrix} \mathbf{A}_{\mathbf{S}_j} & \mathbf{B}_{\mathbf{S}_j} \\ \mathbf{C}_{\mathbf{S}_j} & \mathbf{T}_{\mathbf{S}_j} \end{pmatrix} \begin{pmatrix} d_x \\ s_x \end{pmatrix} = \begin{pmatrix} d_f \\ s_f \end{pmatrix}. \tag{4}$$

Formally eliminating d_x from (4) by substituting $d_x = \mathbf{A}_{\mathbf{S}_j}^{-1}(d_f - \mathbf{B}_{\mathbf{S}_j}s_x)$ (Gaussian elimination) yields

$$(\mathbf{T}_{\mathbf{S}_j} - \mathbf{C}_{\mathbf{S}_j}\mathbf{A}_{\mathbf{S}_j}^{-1}\mathbf{B}_{\mathbf{S}_j})s_x = s_f - \mathbf{C}_{\mathbf{S}_j}\mathbf{A}_{\mathbf{S}_j}^{-1}d_f. \tag{5}$$

We call (5) the *reduced equation*, and the operator

$$\mathbf{R}_{\mathbf{S}_j} = \mathbf{T}_{\mathbf{S}_j} - \mathbf{C}_{\mathbf{S}_j}\mathbf{A}_{\mathbf{S}_j}^{-1}\mathbf{B}_{\mathbf{S}_j} \tag{6}$$

the *one-step reduction* of the operator $\mathbf{S}_j$. The right-hand side of (6) is also known the Schur complement of the block-matrix $\begin{pmatrix} \mathbf{A}_{\mathbf{S}_j} & \mathbf{B}_{\mathbf{S}_j} \\ \mathbf{C}_{\mathbf{S}_j} & \mathbf{T}_{\mathbf{S}_j} \end{pmatrix}$.

Note that the solution s_x of the reduced equation is exactly $\mathbf{P}_{j+1}x$, the projection of the solution of the original equation in $\mathbf{V}_{j+1}$. The solution of the reduced equation is the same on the subspace $\mathbf{V}_{j+1}$ as the solution of the original equation (3). Once we have obtained the reduced equation, it may be reduced again to produce an equation on $\mathbf{V}_{j+2}$. Likewise, we may reduce n times to produce an equation on $\mathbf{V}_{j+n}$ the solution of which is the projection of the solution of (3) on $\mathbf{V}_{j+n}$. We note that in the finite-dimensional case, the reduced equation (5) has $1/2^d$ as many unknowns as the original equation (3). Reduction, therefore, preserves the coarse-scale behavior of solutions while reducing the number of unknowns.

The critical questions are: (i) can we control the sparsity (for any finite but arbitrary precision) of the matrix $\mathbf{C}_{\mathbf{S}_j}\mathbf{A}_{\mathbf{S}_j}^{-1}\mathbf{B}_{\mathbf{S}_j}$? and, (ii) can we repeat the reduction step for $\mathbf{R}_{\mathbf{S}_j}$? In MG literature the Schur complement appears in a number of papers but these questions were not answered. In [7] and [16] these questions were answered affirmatively for a finite number of reduction steps. The key property that makes this affirmative answer possible is the vanishing moments property of the basis functions.

The sparsity (for any finite but arbitrary precision) of the multiresolution LU factorization does not depend on dimension. This is in a sharp contrast with the usual practice, where LU factorization is not recommended as an efficient approach in problems of dimension two or higher. For example, if we consider the Poisson equation, then LU decomposition is not considered as a practical option since the fill-ins will yield dense LU factors.

A close examination of the algorithm in [16] reveals a striking resemblance of the multiresolution LU decomposition coupled with the multiresolution forward and backward substitution to a MG technique. The important difference, however, is that there are no W-cycles.

As described above, reduction is an algebraic procedure carried out on matrices over a finite number of scales. It relies on the explicit hierarchy of scales provided by the MRA to algebraically eliminate the fine-scale variables, leaving only the coarse-scale variables and can be cast as a multiresolution reduction procedure for the corresponding ODEs and PDEs [11]. The classical homogenization of partial differential equations is the process of finding "effective" coefficients. In classical homogenization, the fine scale is associated with a small parameter, and the limit is considered as this small parameter goes to zero. In dimension one a connection has been established [15],[14] between multiresolution reduction and classical homogenization (see e.g. [4]). It is important to point out that reduction approximately preserves small eigenvalues of elliptic operators, and the accuracy of this approximation depends on the order of the wavelets [7].

3 Sparsity of Exponential Operators

If $\mathcal{L}$ is a self-adjoint, strictly elliptic operator then the operator $e^{\mathcal{L}t}$ is sparse in wavelet bases (for a finite but arbitrary precision) for all $t \geq 0$. This observation has a significant effect on the methods for solving PDEs.

Let us consider a class of advection-diffusion equations of the form

$$u_t = \mathcal{L}u + \mathcal{N}(u), \qquad x \in \Omega \subset \mathbf{R}^d, \tag{7}$$

where $u = u(x,t)$, $x \in \mathbf{R}^d$, $d = 1,2,3$ and $t \in [0,T]$ with the initial conditions,

$$u(x,0) = u_0(x), \qquad x \in \Omega, \tag{8}$$

and the linear boundary conditions

$$\mathcal{B}u(x,t) = 0, \qquad x \in \partial\Omega, \quad t \in [0,T]. \tag{9}$$

In (7) $\mathcal{L}$ represents the linear and $\mathcal{N}(\cdot)$ the nonlinear terms of the equation, respectively.

Using the semigroup approach we rewrite the partial differential equation (7) as a nonlinear integral equation in time,

$$u(x,t) = e^{(t-t_0)\mathcal{L}}u_0(x) + \int_{t_0}^{t} e^{(t-\tau)\mathcal{L}}\mathcal{N}(u(x,\tau))\, d\tau, \tag{10}$$

and describe a new class of time-evolution schemes based on its discretization. A distinctive feature of these new schemes is exact evaluation of the contribution of the linear part. Namely, if the nonlinear part is zero, then the scheme reduces to the evaluation of the exponential function of the operator (or matrix) $\mathcal{L}$ representing the linear part.

We note that the incompressible Navier-Stokes equations can be written in the form (7). Let us start with the usual form of the Navier-Stokes equations for $x \in \Omega \subset \mathbf{R}^3$,

$$\mathbf{u}_t = \nu\Delta\mathbf{u} - (u_1\partial_1 + u_2\partial_2 + u_3\partial_3)\,\mathbf{u} - \nabla p, \tag{11}$$

$$\partial_1 u_1 + \partial_2 u_2 + \partial_3 u_3 = 0, \tag{12}$$

$$\mathbf{u}(x,0) = \mathbf{u_0}, \tag{13}$$

where p denotes the pressure and $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$, $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ and $\partial_k = \frac{\partial}{\partial x_k}$. In addition, we impose the boundary condition

$$\mathbf{u}(x,t) = 0 \quad x \in \partial\Omega, \quad t \in [0,T], \tag{14}$$

Let us introduce the Riesz transforms which are defined in the Fourier domain as

$$\widehat{(R_j f)}(\xi) = \frac{\xi_j}{|\xi|}\widehat{f}(\xi), \quad j = 1,2,3, \tag{15}$$

where $\widehat{f}$ denotes the Fourier transform of the function f. It is not difficult to show that the projection operator on the divergence free functions (the Leray projection) may be written with the help of the Riesz transforms,

$$\mathbf{P} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} - \begin{pmatrix} R_1^2 & R_1R_2 & R_1R_3 \\ R_2R_1 & R_2^2 & R_2R_3 \\ R_3R_1 & R_3R_2 & R_3^2 \end{pmatrix}. \tag{16}$$

Applying the divergence operator to (11), we obtain $-\Delta p = \sum_{k,l=1}^{3} \partial_k \partial_l u_k u_l$ and an expression for pressure in terms of the Riesz transforms, $p = -\sum_{k,l=1}^{3} R_k R_l (u_k u_l)$. Substituting the expression for the pressure into (11) and taking into consideration that the Riesz transforms commute with derivatives and, moreover, $R_k \partial_l = R_l \partial_k$, we obtain

$$\mathbf{u}_t = \nu \Delta \mathbf{u} - \mathbf{P}(\sum_{m=1}^{3} u_m \partial_m \mathbf{u}), \tag{17}$$

instead of (11) and (12). Equations (17) are now in the form (7), where $\mathcal{L} = \nu\Delta$ and $\mathcal{N}(\mathbf{u}) = -\mathbf{P}(\sum_{m=1}^{3} u_m \partial_m \mathbf{u})$. The transformation from (11) and (12) to (17) is well known and appears in a variety of forms in the literature. Here we followed a derivation presented by Yves Meyer at Summer School at Luminy in 1997.

The apparent problem with (17) for use in numerical computations is that the Riesz transforms are integral operators (which makes (17) into an integro-differential equation). Let us point out that the presence of the Riesz transforms does not create serious difficulties if we represent operators $R_j, j = 1, 2, 3$ in a wavelet basis with a sufficient number of vanishing moments (for a given accuracy). The reason is that these operators are nearly local on wavelets, and thus, have a sparse representation. This approximate locality follows directly from the vanishing moments property. Vanishing moments imply that the Fourier transform of the wavelet and its several first derivatives vanish at zero, and therefore, the discontinuity of the symbol of the Riesz transform at zero has almost no effect. The precise statements about such operators can be found in [6] and [5].

Finally, in rewriting (17) as $\mathbf{u}_t = \mathcal{L}\mathbf{u} + \mathcal{N}(\mathbf{u})$, we incorporate the boundary conditions into the operator $\mathcal{L}$. For example, $\mathbf{u} = \mathcal{L}^{-1}\mathbf{v}$ means that u solves $\mathcal{L}\mathbf{u} = \mathbf{v}$ with the boundary conditions $\mathcal{B}u = 0$. Similarly, $u(x,t) = \mathrm{e}^{\mathcal{L}t}u_0(x)$ means that u solves $u_t = \mathcal{L}u$, $u(x,0) = u_0(x)$ and $\mathcal{B}u(x,t) = 0$.

Computing and applying the exponential or other functions of operators in the usual manner typically requires evaluating dense matrices and is highly inefficient unless there is a fast transform that diagonalizes the operator. For example, if $\mathcal{L}$ is a circulant matrix, then computing functions of operators can be accomplished using the FFT. It is clear that in this case the need of the FFT for diagonalization prevents one from extending this approach to the case of variable coefficients.

In the wavelet system of coordinates computing the exponential of self-adjoint, strictly elliptic operators always results in sparse matrices, and therefore, using the exponential of operators for numerical purposes is an efficient option [8].

Further development of the approach of [8] can be found in [9], where issues of stability of time-discretization schemes with exact treatment of the linear part

(ELP) schemes are considered. The ELP schemes are shown to have distinctly different stability properties as compared with the usual implicit-explicit schemes. The stability properties of traditional time-discretization schemes for advection-diffusion equations are controlled by the linear term and, typically, these equations require implicit treatment to avoid choosing an unreasonably small time step. As it is shown in [9], using an explicit ELP scheme, it is possible to achieve stability usually associated with implicit predictor-corrector schemes.

If an implicit ELP scheme is used, as it is done in [8], an approximation is used only for the nonlinear term. This changes the behavior of the corrector step of implicit schemes. The corrector step iterations of usual implicit schemes for advection-diffusion equations involve either both linear and nonlinear terms or only the linear term. Due to the high condition number of the matrix representing the linear (diffusion) term, convergence of the fixed point iteration requires a very small time step, making the fixed point iteration impractical. Implicit ELP schemes do not involve the linear term and, typically, the fixed point iteration is sufficient as in [8].

We would like to note, that (10) in effect reduces the problem to an ODE–type setup, and for that reason, a variety of methods can be used for its solution. We present operator valued coefficients of multistep ELP schemes and our main point is that these coefficients can be represented by sparse matrices and applied in an efficient manner.

Let us consider the function $u(x,t)$ at the discrete moments of time $t_n = t_0 + n\Delta t$, where Δt is the time step so that $u_n \equiv u(x,t_n)$ and $N_n \equiv \mathcal{N}(u(x,t_n))$. Discretizing (10) yields

$$u_{n+1} = e^{l\mathcal{L}\Delta t} u_{n+1-l} + \Delta t \left(\gamma N_{n+1} + \sum_{m=0}^{M-1} \beta_m N_{n-m} \right), \tag{18}$$

where $M+1$ is the number of time levels involved in the discretization, and $l \le M$. The expression in parenthesis in (18) may be viewed as the numerical quadrature for the integral in (10). The coefficients γ and β_m are functions of $\mathcal{L}\Delta t$. In what follows we restrict our considerations to the case $l = 1$. We observe that the algorithm is explicit if $\gamma = 0$ and it is implicit otherwise. Typically, for a given M, the order of accuracy is M for an explicit scheme and $M+1$ for an implicit scheme due to one more degree of freedom, γ.

For $l = 1$ we provide Tables 1 and 2 for $M = 1, 2, 3$ with expressions for the coefficients of the implicit ($\gamma \neq 0$) and the explicit ($\gamma = 0$) schemes in terms of $\mathrm{Q}_k = \mathrm{Q}_k(\mathcal{L}\Delta t)$, where

$$\mathrm{Q}_k(\mathcal{L}\Delta t) = \frac{\mathrm{e}^{\mathcal{L}\Delta t} - \mathrm{E}_k(\mathcal{L}\Delta t)}{(\mathcal{L}\Delta t)^k}, \tag{19}$$

and

$$\mathrm{E}_k(\mathcal{L}\Delta t)) = \sum_{l=0}^{k-1} \frac{(\mathcal{L}\Delta t)^l}{l!} \tag{20}$$

M	γ	β_0	β_1	β_2
1	Q_2	$Q_1 - Q_2$	0	0
2	$\frac{1}{2}Q_2 + Q_3$	$Q_1 - 2Q_3$	$Q_3 - \frac{1}{2}Q_2$	0
3	$\frac{1}{3}Q_2 + Q_3 + Q_4$	$Q_1 + \frac{1}{2}Q_2 - 2Q_3 - 3Q_4$	$Q_3 - Q_2 + 3Q_4$	$\frac{1}{6}Q_2 - Q_4$

Table 1: Coefficients of implicit ELP schemes for $l = 1$, where $Q_k = Q_k(\mathcal{L}\Delta t))$.

M	β_0	β_1	β_2
1	Q_1	0	0
2	$Q_1 + Q_2$	$-Q_2$	0
3	$Q_1 + 3Q_2/2 + Q_3$	$-2(Q_2 + Q_3)$	$Q_2/2 + Q_3$

Table 2: Coefficients of explicit ELP schemes for $l = 1$, where $Q_k = Q_k(\mathcal{L}\Delta t)$.

In Tables 1 and 2 we have presented examples of the so-called "bare" coefficients. Modified coefficients [8] differ in high order terms: these terms do not affect the order of accuracy but do affect the stability properties. Modified coefficients depend on a particular form of the nonlinear term.

Let us describe a method to compute operators Q_0, Q_1, Q_2, ... without computing $(\mathcal{L}\Delta t)^{-1}$. In computing the exponential, Q_0, we use the scaling and squaring method which is based on the identity

$$Q_0(2x) = (Q_0(x))^2 . \tag{21}$$

First we compute $Q_0(\mathcal{L}\Delta t 2^{-l})$ for some l chosen so that the largest singular value of $\mathcal{L}\Delta t 2^{-l}$ is less than one. This computation is performed using the Taylor expansion. Using (21), the resulting matrix is then squared l times to obtain the final answer. In all of these computations it is necessary (and possible) to remove small matrix elements to maintain sparsity, and at the same time, maintain a predetermined accuracy.

A similar algorithm may be used for computing $Q_j(\mathcal{L}\Delta t)$, $j = 1, 2, \ldots$ for any finite j. Let us illustrate this approach by considering $j = 1, 2$. It is easy to verify that

$$Q_1(2x) = \tfrac{1}{2}\left(Q_0(x)Q_1(x) + Q_1(x)\right), \tag{22}$$

$$Q_2(2x) = \tfrac{1}{4}\left(Q_1(x)Q_1(x) + 2Q_2(x)\right). \tag{23}$$

Thus, a modified scaling and squaring method for computing operator-valued quadrature coefficients for ELP schemes starts by the computation of $Q_0(\mathcal{L}\Delta t 2^{-l})$, $Q_1(\mathcal{L}\Delta t 2^{-l})$ and $Q_2(\mathcal{L}\Delta t 2^{-l})$ for some l selected so that the largest singular value of all three operators is less than one. For these evaluations we use the Taylor expansion. We then proceed by using identities in (21), (22) and (23) l times to compute the operators for the required value of the argument.

As an example consider Burgers' equation

$$u_t + uu_x = \nu u_{xx}, \qquad 0 \le x \le 1, \quad t \ge 0, \tag{24}$$

for $\nu > 0$, together with an initial condition,

$$u(x,0) = u_0(x), \qquad 0 \le x \le 1, \tag{25}$$

and periodic boundary conditions $u(0,t) = u(1,t)$. Burgers' equation is the simplest example of a nonlinear partial differential equation incorporating both linear diffusion and nonlinear advection. In [8] a spatially adaptive approach is used to compute solutions of Burgers' equation via

$$u_{n+1} = \mathrm{Q}_0(\mathcal{L}\Delta t)u_n - \frac{\Delta t}{2}\mathrm{Q}_1(\mathcal{L}\Delta t)\left[u_n\partial_x u_{n+1} + u_{n+1}\partial_x u_n\right]. \tag{26}$$

We refer to [9] for the analysis of stability of ELP schemes.

4 Conclusions

The wavelet based algorithms described above are quite efficient in dimension one. Although algorithms described above scale properly with size in all dimensions, establishing ways of reducing operation counts remains an important task in dimensions two and three. This is an area of the ongoing research and the progress will be reported elsewhere.

References

[1] B. Alpert. A Class of Bases in l^2 for the Sparse Representation of Integral Operators. *SIAM J. Math. Anal*, 24(1):246–262, 1993.

[2] B. Alpert, G. Beylkin, R. R. Coifman, and V. Rokhlin. Wavelet-like bases for the fast solution of second-kind integral equations. *SIAM Journal of Scientific and Statistical Computing*, 14(1):159–174, 1993.

[3] B. Alpert, G. Beylkin, D. Gines, and L. Vozovoi. Toward adaptive solution of partial differential equations in multiwavelet bases. 1998. in progress.

[4] A. Bensoussan, J.L. Lions, and G. Papanicolaou. *Asymptotic Analysis for Periodic Structures.* North-Holland Pub. Co., New York, 1978.

[5] G. Beylkin. On the representation of operators in bases of compactly supported wavelets. *SIAM J. Numer. Anal.*, 29(6):1716–1740, 1992.

[6] G. Beylkin, R. R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comm. Pure and Appl. Math.*, 44:141–183, 1991.

[7] G. Beylkin and N. Coult. A multiresolution strategy for reduction of elliptic PDE's and eigenvalue problems. *Applied and Computational Harmonic Analysis*, 5:129–155, 1998.

[8] G. Beylkin and J.M. Keiser. On the adaptive numerical solution of nonlinear partial differential equations in wavelet bases. *J. Comp. Phys.*, 132:233–259, 1997.

[9] G. Beylkin, J.M. Keiser, and L. Vozovoi. A new class of stable time discretization schemes for the solution of nonlinear PDEs. PAM Report 347, 1998. submitted to JCP.

[10] A. Brandt. Multi-level adaptive solutions to boundary value problems. *Math. Comp.*, 31:333–390, 1977.

[11] M. E. Brewster and G. Beylkin. A Multiresolution Strategy for Numerical Homogenization. *ACHA*, 2:327–349, 1995.

[12] R. R. Coifman and Y. Meyer. Nouvelles bases orthogonales. *C.R. Acad. Sci., Paris*, 1990.

[13] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure and Appl. Math.*, 41:909–996, 1988.

[14] M. Dorobantu and B. Engquist. Wavelet-based numerical homogenization. *SIAM J. Numer. Anal.*, 35(2):540–559, 1998.

[15] A. C. Gilbert. A comparison of multiresolution and classical one-dimensional homogenization schemes. *to appear in Appl. and Comp. Harmonic Analysis.*

[16] D.L. Gines, G. Beylkin, and J. Dunn. LU factorization of non-standard forms and direct multiresolution solvers. *Applied and Computational Harmonic Analysis*, 5:156–201, 1998.

[17] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comp. Phys.*, 73(1):325–348, 1987.

[18] W. Hackbusch. On multi-grid method applied to difference equations. *Computing*, 20:291–306, 1978.

[19] H. S. Malvar. Lapped Transforms for Efficient Transform/Subband Coding. *IEEE Trans. Acoust., Speech, Signal Processing*, 38(6):969–978, 1990.

[20] Y. Meyer. Principe d'incertitude, bases hilbertiennes et algébres d'opérateurs. In *Séminaire Bourbaki,*, page 662. Société Mathématique de France, 1985-86. Astérisque.

[21] V. Rokhlin. Rapid solution of integral equations of classical potential theory. *J. Comp. Phys.*, 60(2), 1985.

[22] R.P.Fedorenko. The speed of convergence of one iterative process. *USSR Comp. Math. and Math. Physics*, 4:227–235, 1964.

[23] J. O. Stromberg. A Modified Franklin System and Higher-Order Spline Systems on $\mathbf{R}^n$ as Unconditional Bases for Hardy Spaces. In *Conference in harmonic analysis in honor of Antoni Zygmund, Wadworth math. series*, pages 475–493, 1983.

Gregory Beylkin
Dept. of Applied Mathematics
University of Colorado at Boulder
Boulder, CO 80309-0526, USA

Uniform Asymptotics for Orthogonal Polynomials

P. Deift, T. Kriecherbauer,
K. T-R McLaughlin, S. Venakides and X. Zhou

Abstract. We consider asymptotics of orthogonal polynomials with respect to a weight $e^{-Q(x)}dx$ on $\mathbb{R}$, where either $Q(x)$ is a polynomial of even order with positive leading coefficient, or $Q(x) = NV(x)$, where $V(x)$ is real analytic on $\mathbb{R}$ and grows sufficiently rapidly as $|x| \to \infty$. We formulate the orthogonal polynomial problem as a Riemann-Hilbert problem following the work of Fokas, Its and Kitaev. We employ the steepest descent-type method for Riemann-Hilbert problems introduced by Deift and Zhou, and further developed by Deift, Venakides and Zhou, in order to obtain uniform Plancherel-Rotach-type asymptotics in the entire complex plane, as well as asymptotic formulae for the zeros, the leading coefficients and the recurrence coefficients of the orthogonal polynomials. These asymptotics are also used to prove various universality conjectures in the theory of random matrices.

1991 Mathematics Subject Classification: 33D45, 60F99, 15A52, 45E05.
Keywords and Phrases: orthogonal polynomials, asymptotics, random matrix theory, universality.

Let $w(x)dx = e^{-Q(x)}dx$ be a measure on the real line. Denote by $\pi_n(x, Q) = \pi_n(x) = x^n + \dots$ the n-th monic orthogonal polynomial with respect to the measure, and by $p_n(x, Q) = p_n(x) = \gamma_n \pi_n(x)$, $\gamma_n > 0$, the normalized n-th orthogonal polynomial, or simply the n-th orthogonal polynomial, i.e.

$$\int_{\mathbb{R}} p_n(x) p_m(x) e^{-Q(x)} dx = \delta_{n,m} \qquad , n, m \in \mathbb{N}. \tag{1}$$

Furthermore, denote by $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ the coefficients of the associated three term recurrence relation, namely, $xp_n(x) = b_n p_{n+1}(x) + a_n p_n(x) + b_{n-1} p_{n-1}(x)$, $n \in \mathbb{N}$, and denote by $x_{1,n} > x_{2,n} > \dots > x_{n,n}$ the roots of p_n.

In [8], the authors considered the case where

$$Q(x) = \sum_{k=0}^{2m} q_k x^k, \qquad q_{2m} > 0, \quad m > 0, \tag{2}$$

is a polynomial of even degree with a positive leading coefficient, and in [7] the case where

$$\begin{aligned} &Q(x) = NV(x), \qquad V(x) \quad \text{is real analytic on } \mathbb{R}, \\ &\text{and} \qquad \frac{\mathrm{V(x)}}{\log(\mathrm{x}^2+1)} \to \infty \qquad \text{as} \quad \mathrm{x} \to \infty. \end{aligned} \tag{3}$$

In [8], the authors are concerned with the asymptotics as $n \to \infty$ of the leading coefficient γ_n, the recurrence coefficient a_n, b_n and the zeros x_{jn}, as well as Plancherel-Rotach-type asymptotics for the orthogonal polynomials p_n, i.e asymptotics for $p_n(zc_n+d_n)$ uniformly for all $z \in \mathbb{C}$, where c_n, d_n are certain quantities related to the so-called Mhaskar-Rahmanov-Saff numbers (see (5) below). The name "Plancherel-Rotach" refers to [17] in which the authors prove asymptotics of this type for the classical case of Hermite polynomials. In [7], the authors are concerned with the asymptotics of γ_n, a_n, b_n and $p_n(z; NV)$ in the case $c^{-1}N \leq n \leq cN$ for some $c > 1$, as $N \to \infty$. These asymptotics are crucial ingredients in proving a variety of universality conjectures in random matrix theory (see [7]).

Due to the page restrictions in these Proceedings, we limit our considerations to a description of the results in [8]. Plancherel-Rotach-type asymptotics for polynomial orthogonal with respect to exponential weights of the above type, play a central role in various questions of weighted approximation on the line (see e.g. [15]). In order to prove our results we use a reformulation of the orthogonal polynomial problem as a Riemann-Hilbert problem, due to Fokas, Its and Kitaev [13, 14] (see below). This Riemann-Hilbert problem is then analyzed in turn asymptotically using the non-commutative steepest-descent method introduced by Deift and Zhou in [11], and further developed in [12] and [9], and placed eventually in a general form by Deift, Venakides and Zhou in [10]. In [8], and particularly in [7], a basic role is played by the results on the equilibrium measure (see below) obtained by Deift, Kriecherbauer and Ken McLaughlin in [5]. In this paper we will only have the opportunity to give a very rough sketch of the steepest descent method: full details can be found in [8]. For the case of varying weights $e^{-NV(x)}dx$, we must, alas, refer the reader to [7], for both a detailed description of the results as well as their proofs, and the connection to random matrix theory. The methods in [7] are similar to those in [8], but require additional technical considerations. In the special case where V is an even quartic polynomial, the results in [7] should be compared with the results of Bleher and Its [2], who were the first to use the steepest-descent method in [11] to study the asymptotics of orthogonal polynomials via a Riemann-Hilbert problem. Some of the results in [7] and in [8] were announced in [6].

There is a vast literature on asymptotic questions for orthogonal polynomials. The list of researchers who have made important contributions close to the results of [7] and [8], includes, in addition to Plancherel and Rotach, and Bleher and Its, Bauldry, Chen, Criscuolo, Della Vechia, Geronimo, Ismail, Lubinsky, Magnus, Maskar, Mastroiani, Mate, Nevai, Rahmanov, Saff, Sheen, Totik and Van Assche, but there are many others. Again, we do not have the opportunity to describe their work in any detail. Fortunately there is an excellent review [15]: also, a detailed description of the work of the above authors related to the present paper is given in [8].

Henceforth we will assume that the potential $Q(x)$ is of the form (2). The statement of our results involves the *n-th Mhaskar-Rahmanov-Saff numbers* (in

short: MRS-numbers [16], [18]) α_n, β_n which can be determined from the equations

$$\frac{1}{2\pi}\int_{\alpha_n}^{\beta_n}\frac{Q'(t)(t-\alpha_n)}{\sqrt{(\beta_n-t)(t-\alpha_n)}}dt = n, \qquad \frac{1}{2\pi}\int_{\alpha_n}^{\beta_n}\frac{Q'(t)(\beta_n-t)}{\sqrt{(\beta_n-t)(t-\alpha_n)}}dt = -n, \quad (4)$$

and in particular the interval $[\alpha_n, \beta_n]$ whose width and midpoint are given by

$$c_n := \frac{\beta_n - \alpha_n}{2}, \qquad d_n := \frac{\beta_n + \alpha_n}{2}. \quad (5)$$

For the weights under consideration it is straightforward to prove the existence of the MRS-numbers for sufficiently large n. Moreover, they can be expressed in a power series in $n^{-\frac{1}{2m}}$. We obtain

$$c_n = n^{\frac{1}{2m}}\sum_{l=0}^{\infty} c^{(l)} n^{-\frac{l}{2m}}, \qquad d_n = \sum_{l=0}^{\infty} d^{(l)} n^{-\frac{l}{2m}}, \quad (6)$$

where the coefficients $c^{(l)}, d^{(l)}$ can be computed explicitly. From now on we will assume that n is sufficiently large for (6) to hold.

Statement of Results

To simplify the analysis, we normalize the interval $[\alpha_n, \beta_n]$ to be $[-1, 1]$ by making the linear change of variable

$$\lambda_n : \mathbb{C} \to \mathbb{C} : z \mapsto c_n z + d_n, \quad (7)$$

which takes the interval $[-1, 1]$ onto $[\alpha_n, \beta_n]$, and we work with the function

$$V_n(z) := \frac{1}{n} Q(\lambda_n(z)). \quad (8)$$

The function V_n is again a polynomial of degree $2m$ with leading coefficient $(mA_m)^{-1} > 0$, whereas all other coefficients tend to zero as n tends to ∞.

We present our results in terms of the well-known equilibrium measure μ_n (see e.g. [19]) with respect to V_n which is defined as the unique minimizer in $\mathcal{M}_1(\mathbb{R}) = \{\text{probability measures on } \mathbb{R}\}$ of the functional

$$I^{V_n} : \mathcal{M}_1(\mathbb{R}) \to (-\infty, \infty] : \mu \mapsto \int_{\mathbb{R}^2} \log|x-y|^{-1} d\mu(x) d\mu(y) + \int_{\mathbb{R}} V_n(x) d\mu(x). \quad (9)$$

The equilibrium measure and the corresponding variational problem emerge naturally in our asymptotic analysis of the Riemann-Hilbert problem. The minimizing measure is given by

$$d\mu_n(x) = \frac{1}{2\pi}\sqrt{1-x^2}\, h_n(x) \mathbf{1}_{[-1,1]}(x) dx, \quad (10)$$

where $\mathbf{1}_{[-1,1]}$ denotes the indicator function of the set $[-1, 1]$ and h_n is a polynomial of degree $2m-2$,

$$h_n(x) = \sum_{k=0}^{2m-2} h_{n,k} x^k, \quad (11)$$

and the coefficients $h_{n,k}$ can be expanded in an explicitly computable power series in $n^{-\frac{1}{2m}}$.

Finally, to state our first theorem, we define

$$l_n := \frac{1}{\pi}\int_{-1}^{1}\sqrt{1-t^2}h_n(t)\log|t|dt - V_n(0), \tag{12}$$

which also has an explicitly computable power series in $n^{-\frac{1}{2m}}$.

ASYMPTOTICS OF THE LEADING AND RECURRENCE COEFFICIENTS OF THE ORTHOGONAL POLYNOMIALS p_n

THEOREM 13. *In the above notation we have*

$$\gamma_n\sqrt{\pi c_n^{2n+1}e^{nl_n}} = 1 - \frac{1}{n}\left(\frac{4h_n(1)-3h_n'(1)}{48h_n(1)^2} + \frac{4h_n(-1)+3h_n'(-1)}{48h_n(-1)^2}\right) + \mathcal{O}\left(\frac{1}{n^2}\right),$$
$$\frac{b_{n-1}}{c_n} = \frac{1}{2} + \mathcal{O}\left(\frac{1}{n^2}\right), \quad a_n = d_n + \frac{c_n}{2n}\left(\frac{1}{h_n(1)} - \frac{1}{h_n(-1)} + \mathcal{O}\left(\frac{1}{n}\right)\right). \tag{14}$$

In all three cases there are explicit integral formulae for the error terms which all have an asymptotic expansion in $n^{-\frac{1}{2m}}$, e.g. $\mathcal{O}\left(\frac{1}{n}\right) = \frac{1}{n}\left(\kappa_0 + \kappa_1 n^{-\frac{1}{2m}} + \dots\right)$. The coefficients of these expansions can be computed via the calculus of residues by purely algebraic means.

Next we will state the Plancherel-Rotach type asymptotics of the orthogonal polynomials p_n, i.e. the limiting behavior of the rescaled n-th orthogonal polynomial $p_n(\lambda_n(z))$, as n tends to infinity and $z \in \mathbb{C}$ remains fixed. We will give the leading order behavior and produce error bounds which are uniform in the entire complex plane $\mathbb{C}$.

NOTATION: In the following, $(\cdot)^\alpha$, $\alpha \in \mathbb{R}$, denotes the principal branch of the α^{th} root. On the other hand, we will reserve the notation $\sqrt{a}$ for nonnegative numbers a, and we always take $\sqrt{a}$ nonnegative: thus $\sqrt{1-x^2}$, $-1 \le x \le 1$ in (10) is positive.

PLANCHEREL-ROTACH ASYMPTOTICS

We state our second theorem in terms of the function

$$\psi_n : \mathbb{C}\setminus((-\infty,-1]\cup[1,\infty)) \to \mathbb{C} : z \mapsto \frac{1}{2\pi}(1-z)^{1/2}(1+z)^{1/2}h_n(z). \tag{15}$$

The function ψ_n is an analytic extension of the density of μ_n on $(-1,1)$ to $\mathbb{C}\setminus((-\infty,-1]\cup[1,\infty))$ and is thus closely linked to the equilibrium measure (cf. (10)). We show that there exist analytic functions f_n, $\tilde{f}_n$ in a neighborhood of 1, respectively -1, satisfying

$$
\begin{aligned}
(-f_n(z))^{3/2} &= -n\frac{3\pi}{2}\int_1^z \psi_n(y)dy, \quad \text{for } |z-1| \text{ small}, \quad z \notin [1,\infty). \\
(\tilde{f}_n(z))^{3/2} &= n\frac{3\pi}{2}\int_{-1}^z \psi_n(y)dy, \quad \text{for } |z+1| \text{ small}, \quad z \notin (-\infty,-1].
\end{aligned} \tag{16}
$$

As $p_n(z) = \overline{p_n(\bar{z})}$, it is sufficient to describe the asymptotics of $p_n(c_n z + d_n)$ in the closed upper half plane $\overline{\mathbb{C}}_+$. Depending on a small parameter δ, we divide $\overline{\mathbb{C}}_+$ into six *closed* regions, as shown in Figure 17 below. We only describe the asymptotics in $A_\delta, C_{1,\delta}, C_{2,\delta}$ and B_δ. The asymptotics in $D_{j,\delta}$, $j = 1, 2$, is of a similar form to that in $C_{j,\delta}$, $j = 1, 2$ respectively, with $\tilde{f}_n$ replacing f_n. Let $Ai(z)$ denote the Airy function [1, 10.4].

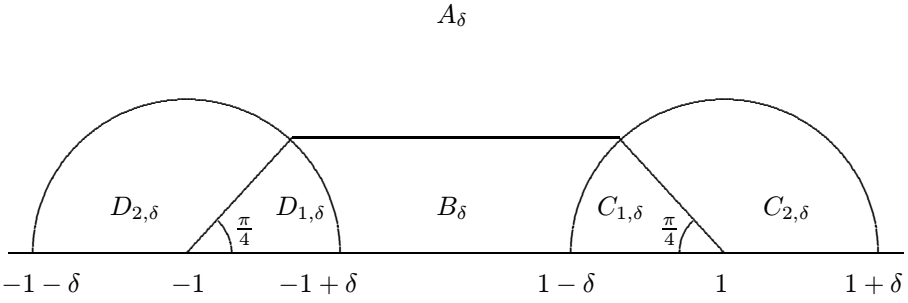


Figure 17. Different asymptotic regions for $p_n(c_n z + d_n)$ in $\overline{\mathbb{C}_\pm}$.

THEOREM 18. *There exists a δ_0 such that for all $0 < \delta \le \delta_0$ the following holds (see Figure 17):*

(i) For $z \in A_\delta$:

$$
\begin{aligned}
p_n(c_n z + d_n)e^{-\frac{1}{2}Q(c_n z+d_n)} &= \sqrt{\frac{1}{4\pi c_n}}\exp\left(-n\pi i\int_1^z \psi_n(y)dy\right) \\
&\quad \times \left(\frac{(z-1)^{1/4}}{(z+1)^{1/4}} + \frac{(z+1)^{1/4}}{(z-1)^{1/4}}\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right).
\end{aligned} \tag{19}
$$

(ii) For $z \in B_\delta$:

$$
\begin{aligned}
p_n(c_n z + d_n)e^{-\frac{1}{2}Q(c_n z+d_n)} &= \sqrt{\frac{2}{\pi c_n}}(1-z)^{-1/4}(1+z)^{-1/4} \\
&\quad \times \left\{\cos\left(n\pi\int_1^z \psi_n(y)dy + \frac{1}{2}\arcsin z\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\right. \\
&\quad \left. + \sin\left(n\pi\int_1^z \psi_n(y)dy + \frac{1}{2}\arcsin z\right)\mathcal{O}\left(\frac{1}{n}\right)\right\}.
\end{aligned} \tag{20}
$$

(iii) For $z \in C_{1,\delta}$:

$$p_n(c_n z + d_n) e^{-\frac{1}{2}Q(c_n z + d_n)} \tag{21}$$

$$= \sqrt{\frac{1}{c_n}} \left\{ \left(\frac{(z+1)^{1/4}}{(z-1)^{1/4}} \left(f_n(z)\right)^{1/4} Ai\left(f_n(z)\right) \right) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \right.$$

$$\left. - \left(\frac{(z-1)^{1/4}}{(z+1)^{1/4}} \left(f_n(z)\right)^{-1/4} Ai'\left(f_n(z)\right) \right) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \right\}. \tag{22}$$

(iv) For $z \in C_{2,\delta}$:

$$p_n(c_n z + d_n) e^{-\frac{1}{2}Q(c_n z + d_n)} = \sqrt{\frac{1}{c_n}} \left\{ \frac{(z+1)^{1/4}}{(z-1)^{1/4}} \left(f_n(z)\right)^{1/4} Ai(f_n(z)) \right.$$

$$\left. - \quad \frac{(z-1)^{1/4}}{(z+1)^{1/4}} \left(f_n(z)\right)^{-1/4} Ai'(f_n(z)) \right\} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right). \tag{23}$$

All the error terms are uniform for $\delta \in$ compact subsets of $(0, \delta_0]$ and for $z \in X_\delta$, where $X \in \{A, B, C_1, C_2\}$. There are integral formulae for the error terms from which one can extract an explicit asymptotic expansion in $n^{-\frac{1}{2m}}$.

REMARKS:

(a) Some of the expressions in Theorem 18 are not well defined for all $z \in \mathbb{R}$ (see e.g. $(z-1)^{1/4}$, $\int_1^z \psi_n(y)dy$). In these cases we always take the limiting expressions as z is approached from the upper half-plane.

(b) The function arcsin is defined as the inverse function of

$$\sin : \{z \in \mathbb{C} : |Re(z)| < \frac{\pi}{2}\} \to \mathbb{C} \setminus ((-\infty, -1] \cup [1, \infty)).$$

ASYMPTOTIC LOCATION OF THE ZEROS

In order to state our result on the location of the zeros, we denote the zeros of the Airy function Ai by $0 > -\iota_1 > -\iota_2 > \dots$. Recall that the all the zeros of Ai lie in $(-\infty, 0)$, so that there exists a largest zero $-\iota_1 < 0$. Furthermore, note that $[-1, 1] \ni x \mapsto \int_x^1 \psi_n(t)dt \in [0, 1]$ is bijective and we define its inverse function to be $\zeta_n : [0, 1] \mapsto [-1, 1]$.

THEOREM 24. *The zeros $x_{1,n} > x_{2,n} > \dots > x_{n,n}$ of the n-th orthogonal polynomials p_n satisfy the following asymptotic formulae:*

(i) Fix $k \in \mathbb{N}$. Then

$$\frac{x_{k,n} - d_n}{c_n} = 1 - \left(\frac{2}{h_n(1)^2}\right)^{1/3} \frac{\iota_k}{n^{2/3}} + \mathcal{O}\left(\frac{1}{n}\right), \qquad \text{as } n \to \infty, \tag{25}$$

and

$$\frac{x_{n-k,n} - d_n}{c_n} = -1 + \left(\frac{2}{h_n(-1)^2}\right)^{1/3} \frac{\iota_k}{n^{2/3}} + \mathcal{O}\left(\frac{1}{n}\right), \qquad \text{as } n \to \infty. \tag{26}$$

(ii) There exist constants $k_0, C > 0$, such that for all $k_0 \le k \le n - k_0$ the following holds:

$$\frac{x_{k,n} - d_n}{c_n} \in \left(\zeta_n\left(\frac{6k-1}{6n}\right), \zeta_n\left(\frac{6k-5}{6n}\right)\right). \tag{27}$$

$$\left|\frac{x_{k,n} - d_n}{c_n} - \zeta_n\left(\frac{6k-3}{6n} + \frac{1}{2\pi n}\arcsin\left(\zeta_n\left(k/n\right)\right)\right)\right| \le \frac{C}{n^2[\alpha(1-\alpha)]^{4/3}}, \tag{28}$$

where $\alpha := k/n$.

(iii) There exists a constant $C_1 > 0$ such that

$$\frac{1}{C_1} < \frac{x_{k,n} - x_{k+1,n}}{c_n[nk(n-k)]^{1/3}} < C_1 \quad \text{for all } 1 \le k \le n-1. \tag{29}$$

REMARKS:

(a) Using the asymptotic expansion for the error terms in Theorem 18 one can of course approximate the k-th zero $x_{k,n}$ of the orthogonal polynomial p_n to arbitrary accuracy.

(b) Note that the error term in (28) is at most of order $\mathcal{O}(n^{-2/3})$. Furthermore it is obvious that for any compact subset K of $(0,1)$, there exists a constant C_K, such that the error term in (28) is bounded by C_K/n^2, as long as $\alpha = k/n \in K$.

As noted earlier, our approach to the asymptotic problem for orthogonal polynomials, is based on the reformulation of the orthogonal polynomial problem as a Riemann-Hilbert problem due to Fokas, Its and Kitaev (see [13], [14]: a specialized version appeared also in [4]).

A general reference for Riemann-Hilbert problems is, for example, [3]. Let Σ be an oriented contour in $\mathbb{C}$.

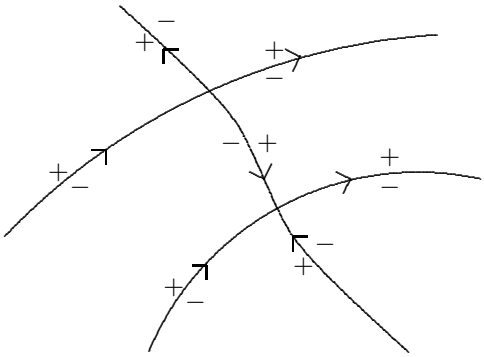


Figure 30.

As indicated in the Figure, the (+)-side (resp, (−)-side) of the contour lies to left (resp, right) as one moves along the contour in the direction of the orientation. Let v be a given map from Σ to $Gl(k, \mathbb{C})$. We say that $m = m(z)$ is a solution of the Riemann-Hilbert problem (Σ, v) if

- $m(z)$ is analytic in $\mathbb{C} - \Sigma$,

- $m_+(z) = m_-(z)v(z)\ ,\ z \in \Sigma,$

where $m_\pm(z) = \lim_{z' \to zz' \in (\pm)-side} m(z')$. The matrix v is called the *jump matrix* for the Riemann-Hilbert problem. If in addition

- $m(z) \to I$ as $z \to \infty$,

we say that the the Riemann-Hilbert problem is normalized at infinity.

THEOREM 31. ([13, 14]) *Let $w : \mathbb{R} \to \mathbb{R}_+$ denote a function with the property that $w(s)s^k$ belongs to the Sobolev space $H^1(\mathbb{R})$ for all $k \in \mathbb{N}$. Suppose furthermore that n is a positive integer. Then the Riemann-Hilbert problem on $\Sigma = \mathbb{R}$, oriented from $-\infty$ to $+\infty$,*

$$\begin{aligned} Y : \mathbb{C} \setminus \mathbb{R} \to \mathbb{C}^{2\times 2} \quad &\text{is analytic}\ , \quad Y_+(s) = Y_-(s) \begin{pmatrix} 1 & w(s) \\ 0 & 1 \end{pmatrix} \text{ for } s \in \mathbb{R}, \\ Y(z) \begin{pmatrix} z^{-n} & 0 \\ 0 & z^n \end{pmatrix} &= I + \mathcal{O}\left(\frac{1}{|z|}\right), \qquad \text{as } |z| \to \infty, \end{aligned} \tag{32}$$

has a unique solution, given by

$$Y(z) = \begin{pmatrix} \pi_n(z) & \int_{\mathbb{R}} \frac{\pi_n(s)w(s)}{s-z} \frac{ds}{2\pi i} \\ -2\pi i \gamma_{n-1}^2 \pi_{n-1}(z) & \int_{\mathbb{R}} \frac{-\gamma_{n-1}^2 \pi_{n-1}(s)w(s)}{s-z} ds \end{pmatrix}, \tag{33}$$

where π_n denotes the n-th monic orthogonal polynomial with respect to the measure $w(x)dx$ on $\mathbb{R}$ and $\gamma_n > 0$ denotes the leading coefficient of the n-th orthogonal polynomial $p_n = \gamma_n \pi_n$. Furthermore, there exist Y_1, $Y_2 \in \mathbb{C}^{2\times 2}$ such that

$$\begin{aligned} &Y(z) \begin{pmatrix} z^{-n} & 0 \\ 0 & z^n \end{pmatrix} = I + \frac{Y_1}{z} + \frac{Y_2}{z^2} + \mathcal{O}\left(\frac{1}{|z|^3}\right), \qquad \text{as } |z| \to \infty, \\ &\text{and} \quad \gamma_{n-1} = \sqrt{(Y_1)_{21}/-2\pi i}, \qquad \gamma_n = 1/\sqrt{-2\pi i (Y_1)_{12}}, \\ &\qquad a_n = (Y_1)_{11} + (Y_2)_{12}/(Y_1)_{12}, \qquad b_{n-1} = \sqrt{(Y_1)_{21}(Y_1)_{12}}, \end{aligned} \tag{34}$$

where a_n, b_n are the recurrence coefficients associated to the orthogonal polynomials p_n.

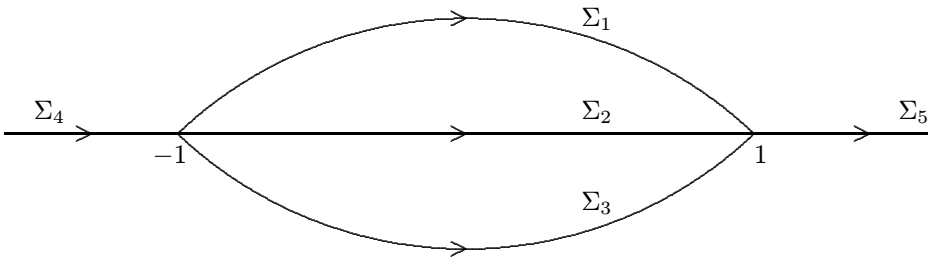


Figure 35. The contour Σ_S.

We are interested in the case where $w(x) = e^{-Q(x)}$, and $Q(x)$ satisfies (2). Theorem 31 converts the problem of computing the asymptotics of $\gamma_n, a_n, b_n, \ldots$ into a problem of computing the asymptotics of the Riemann-Hilbert problem (32) as $n \to \infty$. As indicated, this is achieved by using the steepest descent method for Riemann-Hilbert problems introduced in [11], and further developed in [12]. We conclude with a brief sketch of the method, which involves a sequence of transformations of the Riemann-Hilbert problem:

(i) RESCALING: $Y \to U_n(z) \equiv \begin{pmatrix} c_n^{-n} & 0 \\ 0 & c_n^n \end{pmatrix} Y(c_n z + d_n)$, where c_n, d_n are related to the MRS-numbers as in (5).

(ii) introduction of the "g-FUNCTION" which is the analog for the Riemann-Hilbert problem of the phase function of linear WKB theory: $U \to T(z) \equiv e^{-nl\sigma_3/2} U(z) e^{-n(g(z)-l/2)\sigma_3}$ where σ_3 is the Pauli matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ and $l = l_n$ is given in (12). The function $g(z)$ is analytic in $\mathbb{C} \setminus \mathbb{R}$, has asymptotics $g(z) \sim \log z$ as $z \to \infty$ and is uniquely determined as in [10] by requiring that $g_\pm(z) \equiv \lim_{\epsilon \to 0^+} g(z \pm i\epsilon)$ satisfy certain equalities and inequalities ("Phase Conditions") on $\mathbb{R}$. A simple computation shows that $T(z)$ is the solution of the following Riemann-Hilbert problem, normalized at infinity:

- $T(z)$ is analytic in $\mathbb{C} \setminus \mathbb{R}$,
- $T_+(z) = T_-(z) \begin{pmatrix} e^{-n(g_+(z)-g_-(z))} & e^{n(g_+(z)+g_-(z)-V_n(z)-l)} \\ 0 & e^{n(g_+(z)-g_-(z))} \end{pmatrix}$ for $z \in \mathbb{R}$,
- $T(z) = I + O(\frac{1}{|z|})$ as $z \to \infty$.

(iii) involves a FACTORIZATION of the jump matrix and a DEFORMATION of the contour: $T \to S$. The 2×2 matrix function $S = S(z)$ solves a Riemann-Hilbert problem on a contour of type Σ_S as in Figure 35. Now the Phase Conditions in (ii) are chosen PRECISELY to ensure that the jump matrix v_S for S on $\Sigma_1, \Sigma_3, \Sigma_4$ and Σ_5, converges exponentially to the identity matrix as $n \to \infty$, whereas $v_S = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ on $\Sigma_2 = [-1, 1]$. Thus as $n \to \infty$, we expect that S converges to the solution of the simple Riemann-Hilbert problem ($\Sigma_2 = [-1, 1], v = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$), which may be solved in turn in terms of elementary radicals.

The final step (iv) involves the construction, following [12], of a PARAMETRIX for S at the points of self-intersection $\{-1, 1\}$ of Σ_S: $S \to R$. Although $v_S \to I$ on $\Sigma_1 \cup \Sigma_3 \cup \Sigma_4 \cup \Sigma_5$, the convergence is not uniform and is slower and slower near 1 and -1. This is the central analytical difficulty in the method, and requires delicate consideration. The parametrix for S is chosen so that R solves a Riemann-Hilbert problem on an extended contour $\Sigma_R \supset \Sigma_S$ with a jump matrix v_R satisfying $\|v_R - I\|_{L^\infty(\Sigma_R)} \to 0$ as $n \to \infty$. By standard Riemann-Hilbert methods, R can then be solved in terms of a Neumann series, and retracing the steps $R \to S \to T \to U \to Y$, we obtain the asymptotics for $\gamma_n, a_n, b_n, x_{kn}$ and $p_n(c_n z + d_n)$ as advertised in Theorem 13, 18 and 24.

Finally we note that it is a remarkable piece of luck that the phase condition in (ii) above can be expressed simply in terms of the equilibrium measure $d\mu_n$ corresponding to $V_n(z)$ as in (9), (10) above. Indeed, if we set $g(z) = \int \log(z - x) d\mu_n$,

then it turns out that the Euler-Lagrange variational equations for μ_n, the minimizing measure in (9), are EQUIVALENT to the desired phase condition on g. In this way we construct the g-function in terms of the equilibrium measure.

Acknowledgements. The authors would like to thank Alexander Its and Peter Sarnak for many useful conversations on orthogonal polynomials. The authors would also like to thank Alexander Its and Pavel Bleher for making available to them their preprint [2] at an early stage. Percy Deift was supported in part by NSF grant # DMS-9500867. Thomas Kriecherbauer was supported in part by DFG grant # Kr 1673/2-1. Kenneth T-R McLaughlin was supported in part by NSF postdoctoral fellowship grant # DMS-9508946. Stephanos Venakides was supported in part by NSF grant # DMS-9500623 and ARO grant # DAAH 04-96-1-0157. Xin Zhou was supported in part by NSF grant # DMS-9706644.

References

[1] M. Abramowitz, and I.A. Stegun. *Handbook of Mathematical Functions.* Dover Publications, New York, 1968.

[2] P. Bleher and A. Its. Asymptotics of Orthogonal Polynomials and Universality in Matrix Models. Preprint, 1996.

[3] K. Clancey and I. Gohberg. *Factorization of matrix functions and singular integral operators.* Operator Theory, 3, Birkhäuser Verlag, Basel, (1981).

[4] P. Deift, S. Kamvissis, T. Kriecherbauer, and X. Zhou. The Toda Rarefaction Problem. *Comm. Pure Appl. Math.*, 49, 35-83, (1996).

[5] P. Deift, T. Kriecherbauer, and K. T-R McLaughlin. New Results on the Equilibrium Measure for Logarithmic Potentials in the Presence of an External Field. *To appear, J. Approx. Theory.*

[6] P. Deift, T. Kriecherbauer, K. T-R McLaughlin, S. Venakides and X. Zhou. Asymptotics for Polynomials Orthogonal with Respect to Varying Exponential Weights. *International Mathematics Research Notices* 16, 759-782, (1997).

[7] P. Deift, T. Kriecherbauer, K. T-R McLaughlin, S. Venakides and X. Zhou. *Uniform Asymptotics for Polynomials Orthogonal with respect to Varying Exponential Weights and Applications to Universality Questions in Random Matrix Theory.* Preprint (1998).

[8] P. Deift, T. Kriecherbauer, K. T-R McLaughlin, S. Venakides and X. Zhou. *Strong Asymptotics for Orthogonal Polynomials with respect to Varying Exponential Weights.* Preprint (1998).

[9] P. Deift, S. Venakides, and X. Zhou. The collisionless Shock Region for the Long-time Behavior of Solutions of the KdV Equation. *Comm. Pure and Appl. Math.*, 47, 199-206 (1994).

[10] P. Deift, S. Venakides, and X. Zhou. New Results in Small Dispersion KdV by an Extension of the Steepest Descent Method for Riemann - Hilbert problems. *Intl. Math. Res. Notes*, No.6, 285-299 (1996).

[11] P. Deift and X. Zhou. A steepest descent method for oscillatory Riemann - Hilbert problems. Asymptotics for the mKdV equation, *Ann. of Math.* 137, 295-370, (1993).

[12] P. Deift and X. Zhou. Asymptotics for the Painlevé II equation. *Comm. Pure and Appl. Math.* 48, 277-337, (1995).

[13] A.S. Fokas, A.R. Its and A.V.Kitaev. Isomonodromic approach in the theory of two-dimensional quantum gravity. *Usp. Matem. Nauk*, 45,135-136, (1990) (In Russian).

[14] A.S.Fokas, A.R. Its and A.V. Kitaev. Discrete Painlevé equations and their appearance in quantum gravity. *Comm. Math. Phys.*, 142, 313-344, (1991).

[15] D. S. Lubinsky. An Update on Orthogonal Polynomials and Weighted Approximation on the Real Line. *Acta Applicandae Mathematicae*, 33, 121-164, (1993).

[16] H.N. Mhaskar and E.B. Saff. Extremal problems for polynomials with exponential weights. *Trans. Amer. Math. Soc.*, 285, 203-234, (1984).

[17] M. Plancherel, and W. Rotach. Sur les valeurs asymptotiques des polynomes d'Hermite $H_n(x) = (-1)^n e^{x^2/2} d^n(e^{-x^2/2})/dx^n$. *Commentarii Mathematici Helvetici*, 1, 227-254, (1929).

[18] E.A. Rachmanov. On asymptotic properties of polynomials orthogonal on the real axis. *Mat. Sb.*, 119, 163-203 (1982). English Transl.: *Math. USSR - Sb.*, 47, (1984).

[19] E.B. Saff and V. Totik. *Logarithmic Potentials with External Fields*. Springer-Verlag, New York, 1997.

P. Deift
Courant Institute of Mathematical Sciences, New York, New York, USA
deift@math1.cims.nyu.edu

T. Kriecherbauer
University of Munich, Munich, Germany, and Courant Institute of Mathematical Sciences, New York, New York, USA
tkriech@rz.mathematik.uni-muenchen.de

K. T-R McLaughlin
Princeton University, Princeton, New Jersey, USA, and University of Arizona, Tucson, Arizona, USA
mcl@math.arizona.edu

S. Venakides and X. Zhou
Duke University, Durham, North Carolina, USA
ven@math.duke.edu
zhou@math.duke.edu

Wavelet Based Numerical Homogenization

Bjorn Engquist

Abstract. In analytic homogenization, a differential equation and its solution with multiple scales are replaced by an approximating equation and its corresponding smoother solution with fewer scales. The scales related to the shortest wavelengths are eliminated. We shall start from a discretization of the original differential equation, which includes all the scales. The solution and the difference operator will be represented in a wavelet basis and the homogenized discrete operator will correspond to a particular form of an approximative projection onto the coarser scales. We shall show that this new operator inherits many of the properties of the original discrete operator, including sparseness. Some numerical examples will be presented and comparisons with the analytic homogenization process will be given. We shall also discuss direct coarse grid approximation.

1. Introduction. Homogenization is a classical analytical way to approximate the effect of some classes of periodic or stochastic oscillations. The problem is often formulated as follows. Consider a set of operators L_ϵ, indexed by the small parameter ϵ, and a right hand side f. Find the *homogenized operator* $\bar{L}$ defined by

$$L_\epsilon u_\epsilon = f, \qquad \lim_{\epsilon \to 0} u_\epsilon = \bar{u}, \qquad \bar{L}\bar{u} = f. \tag{1}$$

In certain cases the convergence above and existence of the homogenized operator can be proved, [3].

In the d-dimensional elliptic case, let $A(y) \in \mathbf{R}^{d\times d}$ be one-periodic in each of its arguments and let I_d denote the unit square. It can then be shown, [3], that

$$L_\epsilon = -\nabla \cdot \left(A\left(\frac{x}{\epsilon}\right) \nabla \right), \quad \bar{L} = -\nabla \cdot (H\nabla), \quad H = \int_{I_d} A(y) - A(y) D\chi(y) dy, \tag{2}$$

where $D\chi$ is the Jacobian of the vector valued function $\chi(y) \in \mathbf{R}^d$, whose components χ_k are given by solving the so called cell problems

$$\nabla \cdot (A(y)\nabla\chi_k) = \sum_{i=1}^{d} \frac{\partial}{\partial y_i} a_{ik}(y), \qquad k = 1, \ldots, d, \tag{3}$$

with periodic boundary conditions for χ_k. Note that H is a constant matrix. See [9] for a direct numerical application of this analytic formalism.

In this paper we present a general procedure for constructing numerical subgrid models to be used on a coarse grid where the smallest scales are not resolved. As in analytic homogenization the subgrid phenomina can be oscillations. The wave length ϵ in the oscillations may be smaller than the the typical grid step size h. The objective is to find models that accurately reproduce the effect of subgrid scales and that in some sense are similar to the original differential operator as is the case in analytic homogenization. The starting point is a finite-dimensional approximation, $Lu = f$, of a differential equation where L approximates the differential operator and u the solution. The operator L can be written on the form

$$L = P(\Delta, A, h, \epsilon), \tag{4}$$

where Δ is a collection of difference operators, A are discretized variable coefficients, typically diagonal matrices, h represents the grid size, and ϵ the smallest scale of significance in the problem.

We shall first briefly discuss the possibility of directly discretizing (4) on a coarse grid, $h > \epsilon$. In general, for finite difference and finite element methods, a reasonable number of grid points or elements are required per wave length of the oscillation, $h << \epsilon$. Phase and group velocity errors will otherwise be $\mathcal{O}(1)$.

For a special type of problems and numerical methods it is, however, possible to prove convergence in a weak sense even if the oscillations are not resolved on the computational grid. These types of techniques are studied in [10], [11] and commented on in section 2.

For the wavelet based homogenization technique we start with a resolved discretization, $h << \epsilon$, and a coarse grid approximation. The specific scale ϵ does not play a role any longer and is dropped in the notation.

We seek a finite dimensional operator $\tilde{L}$ and a right hand side $\bar{f}$ with the following properties. First, $\tilde{L}\tilde{u} = \bar{f}$ and $\tilde{u}$ is a projection of u onto a lower dimensional subspace. Second, $\tilde{L}$ can be written on the same form as L,

$$\tilde{L} = P(\Delta, H, \bar{h}), \tag{5}$$

but with $\bar{h} >> h$ and the structure of H close inheting essential properties from the structure of A, typically diagonal dominance and *sparsity*. The sparsity of the discrete operator is important and corresponds to $\tilde{L}$ being an approximation of a differential operator. We interpret H as the subgrid model of A. If A corresponds to a material coefficient, H can be seen as the effective material coefficient. The procedure outlined above resembles that of the analytic *homogenization* technique used for the continuous case, see section 4. In view of this, we will call $\tilde{L}$ the homogenized operator. See Bensoussan et al., [3], for a thorough presentation of classical homogenization.

Our method is based on multiresolution analysis with wavelet projections and approximation of the discrete operator. Although it can be used with any type of discretization, it is algebraic and, in the present form, only deals with linear systems of equations. The great advantage of this procedure to derive subgrid models is its generality. It can be used on any system of differential equations and does not require separation into the distinct $\mathcal{O}(\epsilon)$ and $\mathcal{O}(1)$ scales or periodic

coefficients. It can also be used to test if it is physically reasonable to represent fine scale effects on a coarse grid with a local operator.

This work was initially presented in Dorobantu and Engquist [8], Andersson, Engquist, Ledfelt and Runborg [1], and based on ideas from Brewster and Beylkin, [5]. See also [13] for analysis in the one-dimensional case. Moreover, there are similarities with numerical homogenization based on techniques from algebraic multigrid, [15,16] and from the use of special purpose finite element methods, [14].

2. DIRECT DISCRETIZATION. Let us first consider the simple approach of using a coarse grid even if not all scales of the original differential equation are clearly resolved. For solutions which are highly oscillatory relative to the grid discretization, numerical techniques without phase velocity errors are needed. In [10], [11] particle scheme or method of characteristics approximations of hyperbolic partial differential equations are analyzed. For a restricted class of schemes it is possible to prove convergence, or weak convergence, in L^p of the numerical approximation to the analytic solution as $h \to 0$ essentially independent of ϵ. *Convergence essentially independent of* ϵ means that a set of ratios of h/ϵ with arbitrary small Lebesque measure must be excluded to avoid resonance, [10], [11].

One simple but typical example for which a rigorous theory is possible is the method of characteristics for the Carleman equations,

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} + u^2 - v^2 = 0 \\ &\frac{\partial v}{\partial t} - \frac{\partial v}{\partial x} + v^2 - u^2 = 0 \\ &u(x,0) = a(x, x/\epsilon) \\ &v(x,0) = b(x, x/\epsilon) \end{aligned} \tag{6a}$$

$$\begin{aligned} &a(x,y), b(x,y), \text{ 1-periodic in } y \\ &u(x_j, t_n) \sim u_j^n \\ &x_j = j\Delta x, \ t_n = n\Delta t, \ \Delta t = \Delta x, \end{aligned}$$

$$\begin{aligned} u_j^{n+1} &= u_{j-1}^n + \Delta t((v_{j-1}^n)^2 - (u_{j-1}^n)^2), \\ v_j^{n+1} &= v_{j+1}^n + \Delta t((u_{j+1}^n)^2 - (v_{j+1}^n)^2), \\ u_j^0 &= a(x_j, x_j/\epsilon) \\ v_j^0 &= b(x_j, x_j/\epsilon) \end{aligned} \tag{6b}$$

The homogenization theory of Tartar [17] applies to the differential equations (6a) and is also used in the convergence proof. The local truncation errors are large for $h > \epsilon$ and a cancelation of the errors must be established. The theorem gives strong convergence in L_∞ essentially independent of ϵ as $h \to 0$.

The wavelet based type of homogenization was derived in order to handle wider classes of differential equations.

3. WAVELET BASED HOMOGENIZATION. Given the full discrete solution operator on a fine grid we wish to find an operator of lower dimension that extracts only the coarse scales of the solution. Let V_j and W_j refer to the usual scaling and wavelet spaces, see e.g. [7]. Then, for a solution in $V_{j+1} = V_j \oplus W_j$, the coarse scale is represented by V_j, and we are thus interested in the operator that yields the solution's projection onto V_j.

Consider the equation

$$L_{j+1}U = F, \qquad U, F \in V_{j+1}, \tag{7}$$

originating from a discretization of a differential equation, where U, in the Haar case, is identified as a piecewise constant approximation. We introduce the orthogonal transformation

$$\mathcal{W}_j : V_{j+1} \to W_j \times V_j, \quad \mathcal{W}_j U \equiv \begin{bmatrix} U_h \\ U_l \end{bmatrix} \quad U_h \in W_j, j \quad U_l \in V_j, \tag{8}$$

and note that the linear operator $\mathcal{W}_j L_{j+1} \mathcal{W}_j^T$ can be decomposed into four operators $L_{j+1} = A_j + B_j + C_j + L_j$, acting between the subspaces V_j and W_j, and such that (7) becomes

$$\begin{bmatrix} A_j & B_j \\ C_j & L_j \end{bmatrix} \begin{bmatrix} U_h \\ U_l \end{bmatrix} = \begin{bmatrix} F_h \\ F_l \end{bmatrix}, \quad U_h, F_h \in W_j, \quad U_l, F_l \in V_j. \tag{9}$$

when we apply $\mathcal{W}_j$ from the left. Block Gaussian elimination now gives an equation for U_l, the coarse part of the solution,

$$\bar{L}_j U_l = \bar{F}_j, \quad \bar{L}_j = L_j - C_j A_j^{-1} R_j, \quad \bar{F}_j = F_l - C_j A_j^{-1} F_h. \tag{10}$$

Hence, our new "coarse grid operator" $\bar{L}_j$ is the Schur complement of $\mathcal{W}_j L_{j+1} \mathcal{W}_j^T$. We also get the homogenized right hand side, $\bar{F}_j$.

For higher dimensions, a standard tensor product extension of the multiresolution analysis allows us to use essentially the same derivation as above to obtain coarse grid operators.

We should note that in general $\bar{L}_j$ will not be sparse even if L_{j+1} is. For the method to be efficient we must be able to approximate $\bar{L}_j$ with a *sparse* matrix $\tilde{L}_j$. This is possible in many important cases. The fact that $\bar{L}_j$ is approximately sparse is fundamental. The finite dimensional operator $\tilde{L}_j$ is our numerically homogenized operator.

The homogenization procedure can be applied recursively on $\bar{L}_j$ to get $\bar{L}_{j-1}$ and so on. This can easily be verified when L_{j+1} is symmetric positive definite. Furthermore, the condition number will not deteriorate. From [2], Chapter 3 with $L_{j+1} = L_{j+1}^T$ and

$$c_1 \|U\|^2 \leq \langle L_{j+1} U, U \rangle \leq c_2 \|U\|^2, \qquad \forall U \in \mathbf{R}^{2^{j+1}} \tag{11}$$

we have the same constants c_1, c_2,

$$c_1\|V\|^2 \le \langle L_j V, V\rangle \le c_2\|V\|^2, \quad \forall V \in \mathbf{R}^{2^j}, \tag{12}$$

where $\bar{L}_j$ is defined by (10) and $\langle u, v\rangle = \sum_k \bar{u}_k v_k$. For the first step in the process an improvement in the condition number can often be estimated from

$$\begin{aligned}\langle \bar{L}_j V, V\rangle &= \langle (L_j - B_j^T \Lambda_j^{-1} B_j)V, V\rangle = \langle L_j V, V\rangle - \langle A_j^{-1} B_j V, B_j V\rangle \\ &\le \langle L_j V, V\rangle. \end{aligned} \tag{13}$$

When the operator L_{j+1} is derived from a finite difference, finite element or finite volume discretization, it is sparse and of a certain structure. In one dimension it might for instance be tridiagonal. However, as remarked above, the matrix $\bar{L}_j$ is not sparse since A_j^{-1} is usually dense. Computing all components of $\bar{L}_j$ would be inefficient. Fortunately, $\bar{L}_j$ will be diagonal dominant in many important cases. For instance, in [8] we proved that for a class of elliptic problems the matrix elements of $\bar{L}_j$ decay exponentially away from the diagonal. We are then able to find a sparse matrix that is a close approximation of $\bar{L}_j$. In general the sparse approximation property follows from the analysis of Calderon-Zygmund operators in Beylkin, Coifman and Rokhlin, [4].

One simple way approximate $\bar{L}_j$ is to set all components outside a prescribed bandwidth equal to zero. Let us define *truncation* of M to bandwidth ν as

$$\mathrm{trunc}(M, \nu)_{ij} = \begin{cases} M_{ij}, & \text{if } 2|i-j| \le \nu - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

There are natural extensions to multi dimensions. This procedure was introduced in [4] and used in [8]. We propose that $\bar{L}_j$ be projected onto banded form in a more effective manner. Let $\{v_j\}_{j=1}^{\nu}$ be a set of linearly independent vectors in $\mathbf{R}^{2^j}$. We define the *band projection*, $\mathrm{band}(M, \nu)$, of a matrix M as the projection of M onto the subspace of matrices with bandwidth ν such that

$$Mx = \mathrm{band}(M, \nu)x, \qquad \forall x \in \mathrm{span}\{v_1, v_2, \ldots, v_\nu\}. \tag{15}$$

In our setting M will usually operate on vectors representing smooth functions, for instance solutions to elliptic equations, and a natural choice for the v_j vectors in one dimension are thus the first ν polynomials,

$$v_j = \{1^{j-1}, 2^{j-1}, \ldots, N^{j-1}\}^T. \tag{16}$$

For the case $\nu = 1$ we should remark that we get the standard "masslumping" of a matrix.

This technique is similar to the probing technique used by Chan et al, [6]. In that case the vectors v_j are sums of unit vectors. Other probing techniques have been suggested by Axelsson, Pohlman and Wittum, see e.g. Chapter 8 in [2]. In some cases the band projection technique only gives improvements for small values

of ν, see figure 1. Numerical evidence indicate that for small values of ν, the band projection technique is quite efficient.

In figure 1 the results from different types of compressions are given for a simple test example. The differential equation is

$$\frac{d}{dx}a(x)\frac{d}{dx}u(x) = 1, \quad 0 \le x \le 1, \tag{17}$$
$$u(0) = 0, \ \frac{d}{dx}u(1) = 0,$$

and it is approximated by centered differences. The variable coefficient $a(x) > 0$ is a uniformly distributed random function. It is clear that the divergence form of the operator with an explicit H-matrix is important. Reducing the number of non-zero elements in H is more efficient than in $\bar{L}_j$.

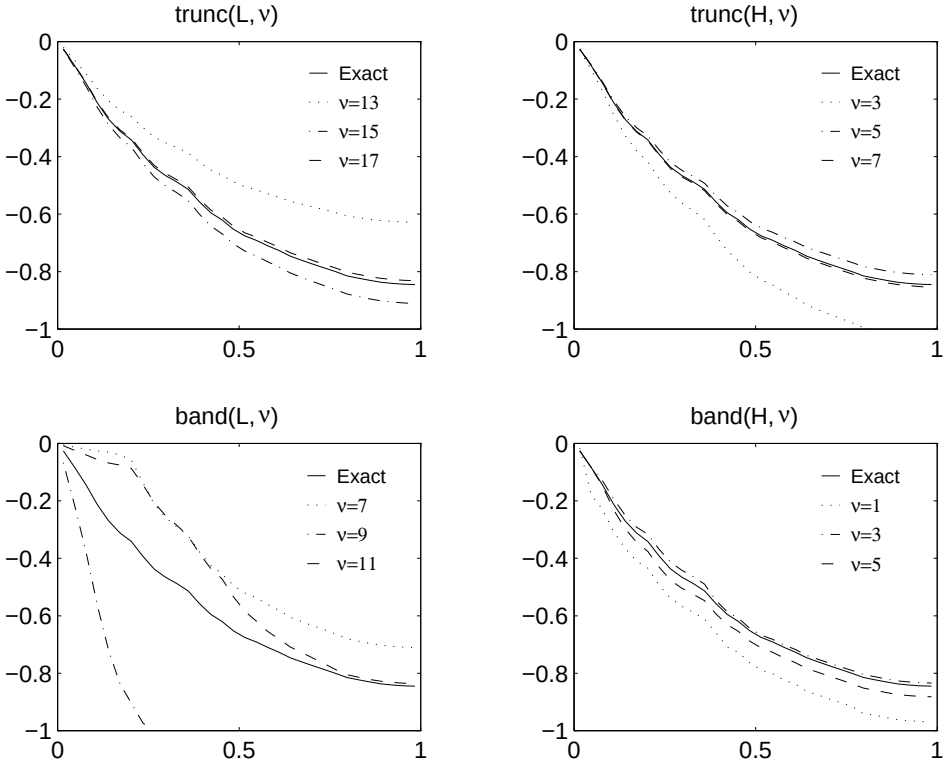


Figure 1: The effect of different types of projections of L_j in the homogenized approximations of (17).

For simple elliptic problems with periodic coefficients it is possible to prove rigorous error estimates for the projection techniques, [1], [8]. The theorems typically state that a finite approximation error δ can be achieved by an operator $\tilde{L}_j$, the matrix of which has only a low number (m) of non-zero elements in each row. The number m depends on δ but at most as $\log h$ on h.

4. COMPARISONS AND EXAMPLES. There is a striking relationship between the analytically homogenized operator in (2) and the Schur complement $\bar{L}_j$ in (10). The first term in both the expressions represent averaged operators, L_j in a discrete sense and $\int_{I_d} A$ is an integral sense. In both formulations a correction term for the high frequency interaction is subtracted from the average. Furthermore, in the correction term χ is the solution of an elliptic equation and A_j^{-1} is an analogous discrete positive definite solution operator.

In the classical analytical setting (2) homogenization gives the asymptotic expansion

$$u_c(x) = \bar{u}(x) + \epsilon u_1(x, x/\epsilon) + \mathcal{O}(\epsilon^2) \tag{18}$$

for the solution, see [3]. The techniques described in [14], [15] and [16] give numerically homogenized operators with coarse grid solutions which directly samples (18). The oscillatory term $\epsilon u_1(x, x/\epsilon)$ is contained in the solution. In the wavelet homogenization the solution is a projection of u_ϵ onto a coarse scale V_j space. The influence of the u_1 term, which oscillates with mean zero [4], is then significantly reduced in orders of ϵ, depending on the type of wavelet basis. Another advantage of using wavelets is the favorable compression ratio for A_j^{-1}, [4].

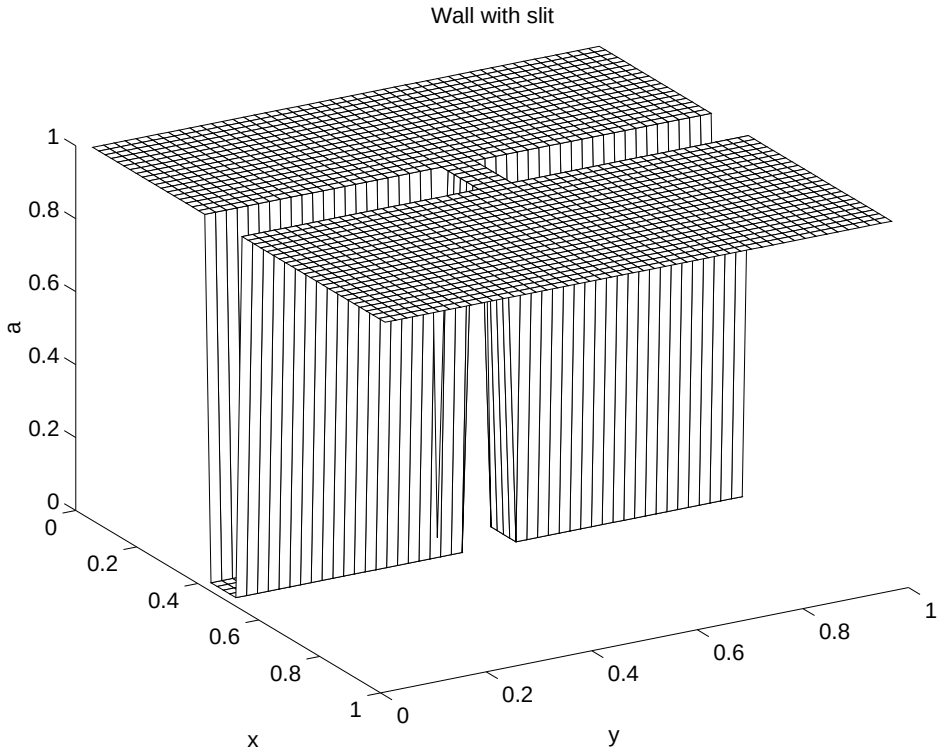


Figure 2: coefficient $a(x)$ in the Helmholtz equation

A number of problems have been studied numerically. Some are simple test cases in one space dimension for the analysis of different properties of the wavelet

based homogenization technique. One example was given in section 3 above. See also [1], [7] and [8]. Others originate from more realistic simulations of e.g. wave propagation in fiber optics and flow in porous media, [12].

A practical application is subcell modelling in the form of the coarsening of a mesh refinement. Sometimes mesh refinements are necessary in order to resolve small geometric details. The refined mesh increases the computational complexity. The numerical homogenization can produce an operator on a uniform coarser grid which inherits the correct resolution from the finer grid. In figure 2 and 3 an example of such a grid coarsening is given.

A standard centered finite difference approximation of the Helmholtz equation,

$$-\nabla a(x)\nabla u - k^2 u = 0, \quad u : \mathbf{R}^2 \to \mathbf{R}, \tag{19}$$

is projected onto a coarser wavelet space. The figure displays different levels of homogenizations, and it is clear that the quality of the

The application comes from the study of wave propagation through a slit for the analysis of electromagnetic compatibility. The figure 4 gives the structure of the matrix corresponding to $\tilde{L}_j$.

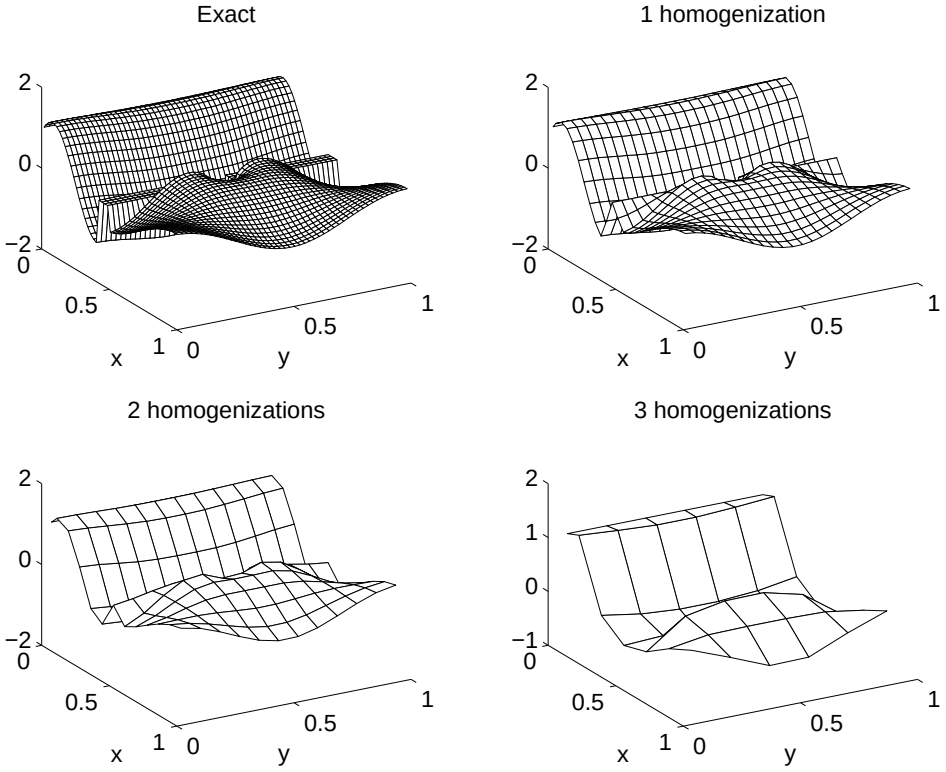


Figure 3: Discrete solution of the Helmholtz equation for different levels of numerical homogenization

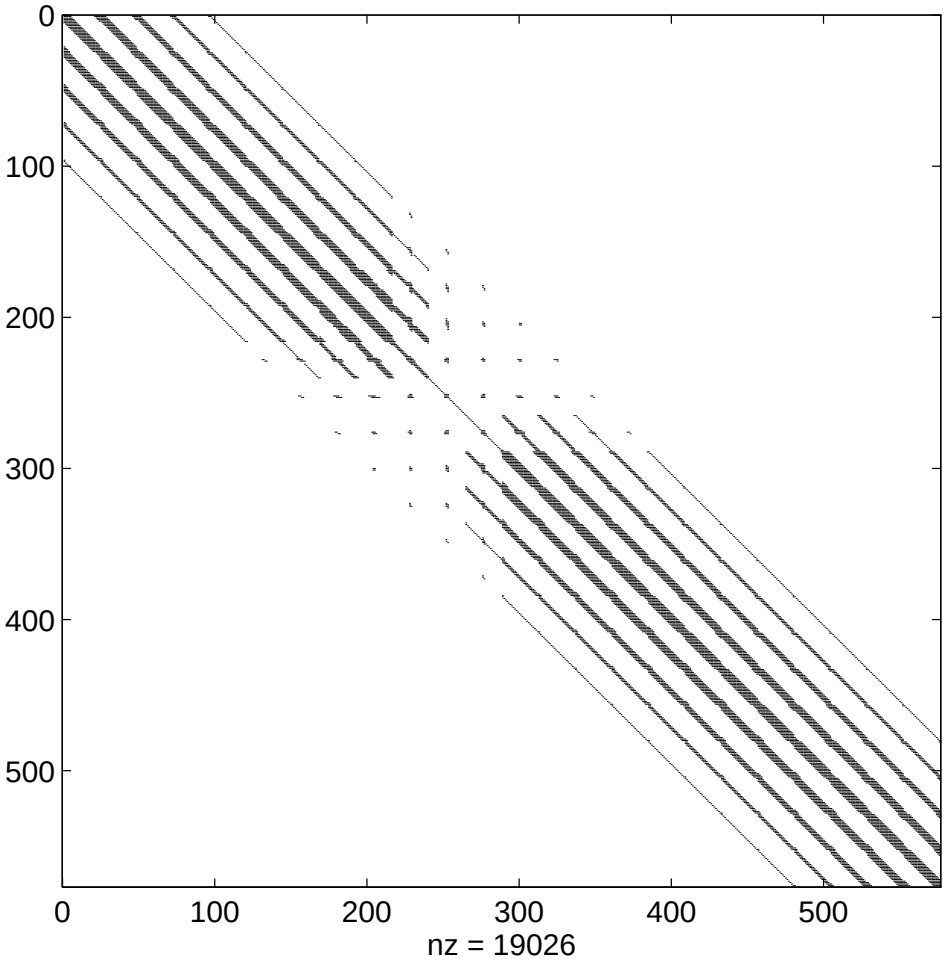


Figure 4: Non-zero elements in the matrix representation of the projected $\bar{L}_j$ corresponding to the solution given in figure 3.

References

[1] U. Andersson, B. Engquist, G. Ledfelt and O. Runborg. A contribution to wavelet-based subgrid modeling. To appear.

[2] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, 1991.

[3] A. Bensoussan, J.-L. Lions, and G. Papanicolau. *Asymptotic Analysis for Periodic Structures.* North-Holland Publ. Comp., The Netherlands, 1978.

[4] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.*, 44:141-183, 1991.

[5] M. Brewster and G. Beylkin. A multiresolution strategy for numerical homogenization. *Applied and Computational Harmonic Analysis*, 2:327-349, 1995.

[6] T. Chan and T. Mathew. The interface probing technique in domain decomposition. *SIAM J. Matrix Anal. Appl.*, 13(1):212-238, January 1992.

[7] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, 1991.

[8] M. Dorobantu and B. Engquist. Wavelet-based numerical homogenization. *SIAM J. on Num. Anal.*, 35(2):540-559, April 1998.
[9] L. Durlofsky. Numerical calculation of equivalent grid block permeability tensors for heterogenuous porous media. *Water Resour. Res.*, 27:699-708, 1991.
[10] B. Engquist and T. Hou. Particle method approximation of oscillatory solutions to hyperbolic differential equations. *SIAM J. Num. Analysis*, 26:289-319, 1989.
[11] B. Engquist and J.-G. Liu. Numerical methods for oscillatory solutions to hyperbolic problems. *Comm. Pure Appl. Math.*, 46:1-36, 1993.
[12] B. Engquist, P.-O. Persson and O. Runborg. Simulation of optical resonance filters using wavelet based homogenization. To appear.
[13] A.C. Gilbert. A comparison of multiresolution and classical one-dimensional homogenization schemes. *Appl. Comput. Harmon. Anal.*, 5(1):1-35, 1998.
[14] T.Y. Hou and X.H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169-189, 1997.
[15] S. Knapek. Matrix-dependent multigrid-homogenization for diffusion problems. *SIAM J. Sci. Stat. Comp.*, 1998. To appear.
[16] N. Neuss. *Homogenisierung und Mehrgitter.* PhD thesis, Fakultät Mathematik Universität Heidelberg, 1995.
[17] L. Tartar, Etudes des oscillations dans les equations aux derivées partielle non linéaires. Lecture Notes in Physics, 195:384-412, Springer-Verlag, 1984.

Bjorn Engquist
Department of Mathematics, UCLA,
Los Angeles, CA 90095-1555 USA
and
NADA, KTH, 10044 Stockholm, Sweden

A Study of Bifurcation of Kolmogorov Flows with an Emphasis on the Singular Limit

Hisashi Okamoto[1]

Abstract.

We consider a family of stationary Navier-Stokes flows in 2D flat tori. The flow is driven by an outer force which is of the form $(\sin y, 0)$. Varying the Reynolds number and the aspect ratio of the torus, we numerically compute bifurcating solutions by a path-continuation method. Folds and cusps are obtained in the range where the Reynolds number is < 100. Some solutions are computed up until the Reynolds number becomes 10,000. Asymptotic properties as the Reynolds number tends to infinity are discussed. Also given is an analysis as the aspect ratio of the torus tends to zero.

1991 Mathematics Subject Classification: Primary 76D30; Secondary 76C05, 35Q30, 35Q35.
Keywords and Phrases: Kolmogorov flows in 2D tori, incompressible fluid, bifurcation, singular perturbation, internal layer, inviscid limit.

1 Introduction

The Navier-Stokes equations have attracted very much attention of both mathematicians and physicists; accordingly scientific papers on them are almost innumerable. Nonetheless, many difficult problems remain to be analyzed; this is especially true when the Reynolds number is large (see [7] and [4]). One of the purposes of the present paper is to point out that something new can be found even if we restrict ourselves to steady-states.

[1]Supported by the Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan, # 09304023, # 09554003

We compute numerically a family of stationary motions of incompressible viscous fluid, which is governed by the following Navier-Stokes equations:

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = \frac{1}{R}\triangle u - \frac{\partial p}{\partial x} + \frac{1}{R}\sin y, \tag{1}$$

$$u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = \frac{1}{R}\triangle v - \frac{\partial p}{\partial y}, \tag{2}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{3}$$

where R is the Reynolds number, (u, v) the velocity, and p the pressure. Those equations are satisfied in $(x, y) \in [-\pi/\alpha, \pi/\alpha] \times [-\pi, \pi]$, with periodic boundary condition. Namely we consider the Navier-Stokes flows in a two-dimensional flat torus and α is its aspect ratio. See [5] and [11].

If the driving force is $(\sin ky, 0)$ with $k \geq 2$, then interesting phenomena are already known for nonstationary motions as well as steady-states (see [2], and the references in [11] for instance). In our problem, which comes from Iudovich [5], the flow is driven by an outer force $(R^{-1}\sin y, 0)$. This simplifies the problem greatly. We would like to refer the reader to [3] and [11], where motives of investigation and historical comments are found. The purpose of the present paper is to report that we can observe many interesting phenomena when we change the aspect ratio α as well as the Reynolds number R. Varying α and R, we numerically compute bifurcating solutions ([11]). Folds and cusps are obtained in the range where $R < 100$. Some solutions are computed until R becomes 10,000 ([13]). We hope that such a list of solutions serves as raw materials for future study of the Navier-Stokes equations. In particular, we would like to obtain a bifurcation diagram which is global in the sense that solutions of all the parameters (α, R) are listed in the diagram. Such global bifurcation diagrams are computed in many one-dimensional systems, notably reaction-diffusion systems ([10]). However, global diagram for the Navier-Stokes equations are substantially more difficult to obtain and we are forced to be content with a partial answer, which we are going to present in the present paper. The following study of Kolmogorov flows is motivated by A. Majda's pioneering works on incompressible fluid motions (see, e.g., [7]) and Nishiura's analysis of reaction-diffusion systems [10].

From our numerical computation, we can guess some interesting asymptotic behavior as the Reynolds number tends to infinity. Since the Navier-Stokes equations are defined in a 2D torus, there can not be a boundary layer. However, we can observe some internal layers, which we will explain in the next section. In order to analyze those internal layers, we apply in section 3 a singular perturbation method in the range where $(\alpha, R) \approx (1, \infty)$. The solution obtained by the perturbation method shows a good agreement with the numerical solution. Analysis when $\alpha \to 0$ is given in section 4.

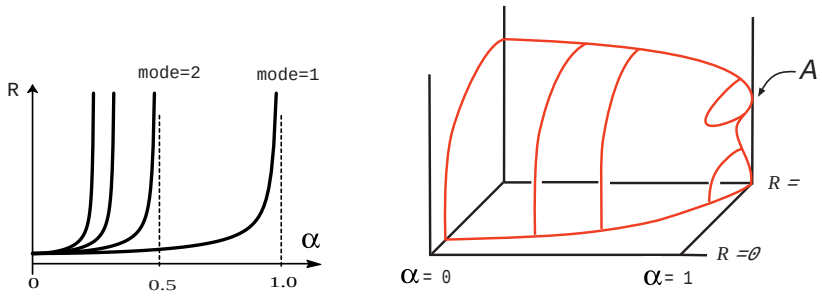


Figure 1: Neutral curves of mode n ($n = 1, 2, \cdots$) (left). Nontrivial solutions bifurcate from the points on the neutral curves. The curve of mode n starts from $(\alpha, R) = (0, \sqrt{2})$ and ends at $(\alpha, R) = (1/n, \infty)$. Schematic bifurcation diagram of solutions of mode 1 (right). The point A represents $(\alpha, R, \psi) = (1.0, \infty, -(\cos x + \cos y)/2)$. Only the upper half of the bifurcating solutions are drawn.

2 GLOBAL PICTURE OF SOLUTIONS AND INVISCID LIMIT

We first recall some numerical facts reported in [11]. We discretize (1)–(3) by the spectral method. The resulting nonlinear equations are solved by the path-continuation method (see [6], for instance). One easily notice that $(u, v, p) = (\sin y, 0, 0)$ solves the equations and the boundary conditions for all $R > 0$ and all $\alpha > 0$. We call it a trivial solution. It is Meshalkin and Sinai [9] which proves that any bifurcation from the trivial solution occurs by steady-states. Namely, the Hopf bifurcation from the trivial solution is prohibited. Iudovich [5] showed that there are bifurcation from the trivial solution if $0 < \alpha < 1$ and that there is none if $1 \leq \alpha < \infty$. See Figure 1 (left). Bifurcating solutions are classified by a positive number called a mode. Roughly speaking, solutions of mode n contains n pairs of eddies in the rectangle $(-\pi/\alpha, \pi/\alpha) \times (-\pi, \pi)$. See [11]. When $0 < \alpha < 0.966\cdots$, then with R as a bifurcation parameter, there exists a pitchfork of bifurcating solutions. There is no secondary bifurcation in the class of those solutions satisfying $\psi(x, y) = \psi(-x, -y)$. When $0.966\cdots < \alpha < 1$, the branch of nontrivial solutions possesses two turning points (= limit points) but still there is no secondary bifurcation in the same function class ([11]). We recently re-computed the stability of those solutions in the function space where we do not assume any symmetry. We have found that the solutions are stable even in this general setting. Therefore the nontrivial solutions on the pitchfork is stable however large the Reynolds number may be. The global view of the solutions of mode one is given in Figure 1 (right). See [11] for detail. This suggests that a possibility that the global attractor for $1/2 < \alpha < 1$ is of one-dimension *however large the Reynolds number may be*. Such low-dimensionality is reported in a different context in Afendikov and Babenko [1] and Chen and Price [2].

The pitchfork bifurcations are supercritical for all $\alpha \in (0,1)$. Namely, the nontrivial solutions in a neighborhood of the bifurcation point lie in the right hand side, where the Reynolds number is greater than the critical Reynolds number. This was shown numerically in [11] but a mathematical proof was not available there, although the supercriticality for sufficiently small α is proved by Afendikov and Babenko [1], and independently by M. Yamada. See [11]. Recently Matsuda and Miyatake [8] gave a proof of supercriticality when α is close to one.

We now consider the asymptotic behavior of the solutions as $R \to \infty$. When R increases with a fixed α, the nontrivial solution seems to converge on a certain function. The numerical experiments in [13] suggests that the vorticity as a limit of $R \to \infty$ is at most C^1. See Figure 2, which shows that $(-\triangle)^{-3/2}\psi$ (ψ is the stream function) loses smoothness along certain curves. We call these curves internal layers. The layer yields an energy spectrum of k^{-r}, where k is the wave number and r is between -7 and -4, depending on the aspect ratio. See [13] for detail. We remark that our "singularity" is much weaker than those found in [7].

Since the knowledge of solutions with large R may help us understand the turbulent motion of fluid, it would be of practical importance to mathematically analyze an asymptotic behavior of steady-states as $R \to \infty$. In the present case, asymptotic analysis seems to be very difficult for a general α. First of all, we encounter the following problem: The Euler equations, which are obtained from (1)–(3) by setting $R = \infty$, possess an infinite number of solutions. In fact, they have a continuum of steady-states. On the other hand, the Navier-Stokes equations have finitely many steady-states for a fixed $R < \infty$. Therefore the vast majority of the stationary Euler flows are disconnected with the Navier-Stokes flows. Hence we would like to know how the Euler flows which are connected with the Navier-Stokes flows are distinguished from those which are disconnected. Some partial answers are given in [12, 13] but we do not know the real mechanism for it. We will show in the next section that a certain heuristic analysis is possible for those solutions which lie in the neighborhood of the point A in Figure 1 (right).

3 Asymptotic analysis as $(R, \alpha) \to (\infty, 1)$

The equations (1)–(3) are written by the stream function as follows (see [11]):

$$\frac{\triangle^2\psi + \cos y}{R} + \psi_x \triangle \psi_y - \psi_y \triangle \psi_x = 0, \tag{4}$$

where the subscript means differentiation. Substituting $x = x'/\alpha, y = y'$ and dropping the primes, we obtain

$$0 = \left(\alpha^2\partial_x^2 + \partial_y^2\right)^2 \psi + \cos y + R\alpha \left[\psi_x \left(\alpha^2\partial_x^2 + \partial_y^2\right)\psi_y - \psi_y \left(\alpha^2\partial_x^2 + \partial_y^2\right)\psi_x\right], \tag{5}$$

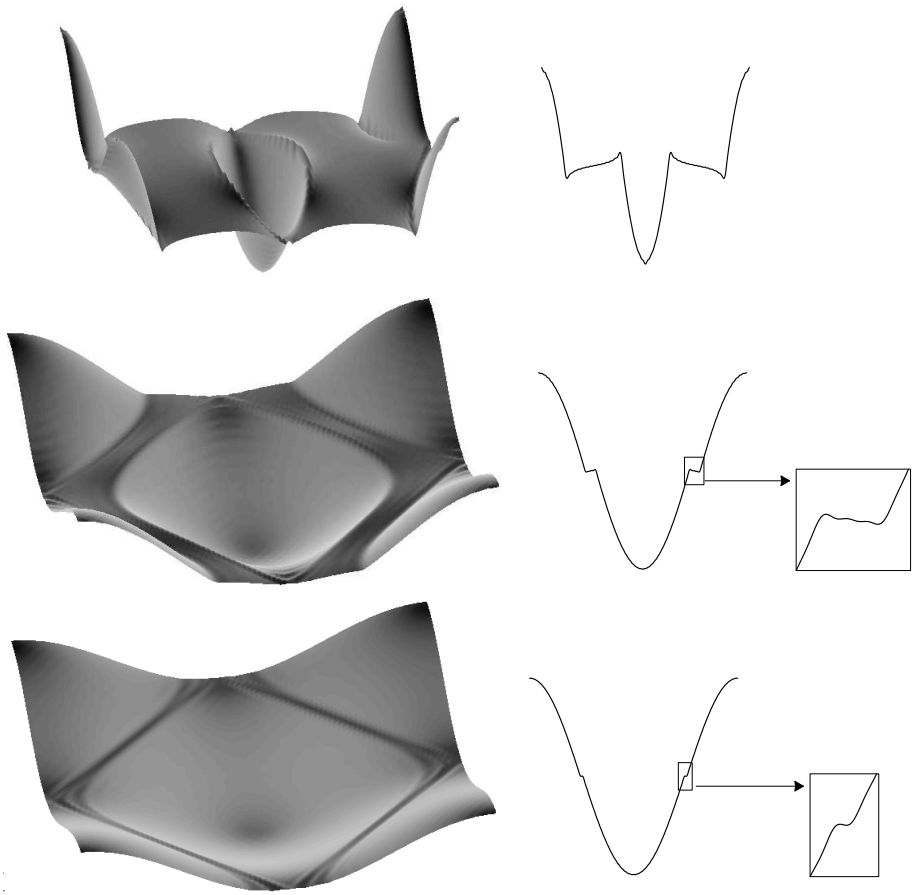


Figure 2: Graphs of $(-\triangle)^{-3/2}\psi$. Bird's-eye views (left) and slices of the graphs along the line $y = \alpha x$ (right). α is 0.7 (top), 0.984 (center), and 0.999 (bottom), respectively. The equations are discretized by the Fourier-Galerkin method. The resulting nonlinear equations, which contains 544 to more than 5,000 independent variables depending on the Reynolds number R, are solved by the path-continuation method.

which is satisfied in $-\pi < x < \pi, -\pi < y < \pi$. Let $\delta = 1/(R\alpha)$ and $\gamma = 1 - \alpha^2$. Then, by defining $J(f, g) = f_x g_y - f_y g_x$, we obtain

$$0 = \delta \left(\triangle^2 \psi + \cos y - 2\gamma \triangle \psi_{xx} + \gamma^2 \psi_{xxxx} \right) + J(\psi, \triangle \psi) - \gamma J(\psi, \psi_{xx}). \qquad (6)$$

We now consider those solutions which are close to the point A in Figure 1 (right). We expand $\gamma \in \mathbf{R}$ and ψ as follows:

$$\psi = \sum_{j,k=0}^{\infty} \epsilon^j \delta^k \psi^{j,k}(x, y), \qquad \gamma = \sum_{j,k=0}^{\infty} \epsilon^j \delta^k \gamma(j, k), \qquad (7)$$

where ϵ is an artificial parameter. It is taken along the vertical tangent of surface of solution set at $(\alpha, R) = (1, \infty)$. See Figure 1(right) and Figure 3 (left). $\gamma(0, 0) = 0$ is assumed so as to comply with the numerical results. Substituting (7) into (6), we compute coefficients of $\epsilon^j \delta^k$. Then we first obtain $J(\psi^{0,0}, \triangle \psi^{0,0}) = 0$. We already know from the numerical results in [11, 13] that $\psi^{0,0} = -(\cos y \pm \cos x)/2$. Since both cases are dealt with in the same way, we choose $\psi^{0,0} = -(\cos y + \cos x)/2$. From the coefficient of $\epsilon^0 \delta^1$, we obtain

$$\triangle^2 \psi^{0,0} + \cos y + J(\psi^{0,0}, \triangle \psi^{0,1}) + J(\psi^{0,1}, \triangle \psi^{0,0}) - \gamma(0, 1) J(\psi^{0,0}, \psi^{0,0}_{xx}) = 0,$$

which is written as

$$\frac{\cos y - \cos x}{2} - \gamma(0, 1) \frac{\sin x \sin y}{4} + \frac{1}{2} (\sin x \partial_y - \sin y \partial_x)(I + \triangle) \psi^{0,1} = 0.$$

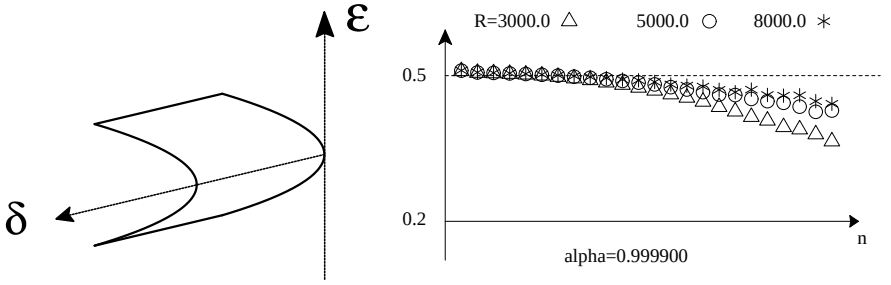


Figure 3: Two coordinates near the turning pint A (left).
Plots of $\alpha R(-1)^n n(1 - 2n^2) a(n, n)$. $\alpha = 0.9999$ and $R = 3000, 5000, 8000$. (right)

Let us define the operator K by $K = \frac{1}{2} (\sin x \partial_y - \sin y \partial_x)$. Note that

$$K \left(\log \left| \frac{\cos \frac{x-y}{2}}{\cos \frac{x+y}{2}} \right| \right) = \frac{1}{2} (\cos y - \cos x) \qquad \text{and} \qquad K(\cos x - \cos y) = \sin x \sin y.$$

Note also that a function $u = u(x, y)$ satisfies $Ku = 0$ if and only if there exists a function of one variable f such that $u(x, y) = f(\cos x + \cos y)$. These facts lead us to

$$\log\left|\frac{\cos\frac{x-y}{2}}{\cos\frac{x+y}{2}}\right| + \frac{1}{4}\gamma(0,1)(\cos y - \cos x) + (I + \triangle)\psi^{0,1} + f(\cos x + \cos y) = 0$$

with some function f. Multiplying this equation with $\cos x - \cos y$, we integrate it on $[-\pi, \pi]^2$. Then we obtain $\gamma(0, 1) = 0$, whence

$$\log\left|\frac{\cos\frac{x-y}{2}}{\cos\frac{x+y}{2}}\right| + (I + \triangle)\psi^{0,1} + f(\cos x + \cos y) = 0. \tag{8}$$

If we further assume that $f \equiv 0$, then we obtain

$$\psi^{0,1}(x, y) = -\sum_{n=1}^{\infty} \frac{2(-1)^{n-1}}{n(1-2n^2)} \sin nx \sin ny + c_1 \cos x + c_2 \cos y, \tag{9}$$

where c_j's are constant.

Because of the limitation of the paper size we do not compute other coefficients in the present paper. Even with our incomplete computation, we can guess a certain interesting asymptotic behavior as $(\alpha, R) \to (1, \infty)$. In fact,

$$\begin{aligned} \psi &= -\frac{\cos x + \cos y}{2} + \epsilon\psi^{1,0} \\ &\quad + \frac{1}{\alpha R}\sum_{n=1}^{\infty} \frac{2(-1)^n}{n(1-2n^2)} \sin nx \sin ny + \text{ smooth function} + \cdots \end{aligned} \tag{10}$$

shows that the nonsmooth function appearing in (9) is a dominant factor for a large (but not too large) wave number range when $(R, \alpha) \to (\infty, 1)$. Figure 3 (right) shows the plot of $\alpha R n(1 - 2n^2)a(n, n)$. This figure indicates that

$$\psi = -\frac{\cos x + \cos y}{2} + \frac{2}{\alpha R}\sum_{n=1}^{\infty} \frac{(-1)^n}{n(1-2n^2)} \sin nx \sin ny + \cdots \tag{11}$$

is a good approximation to the solutions on the turning points in Figure 1 (right) in an intermediate wave number space.

4 KOLMOGOROV FLOWS OF SMALL ASPECT RATIO

The purpose of the present section is to consider the asymptotic behavior of the solutions of (5) as $\alpha \to 0$ with a fixed R. The stationary solution of

the Navier-Stokes equations are expanded into the Fourier series: $\psi(x,y) = \sum_{(m,n)\neq(0,0)} a(m,n)\exp(i\alpha mx + iny)$. Then the Fourier coefficients satisfy the following equations:

$$\frac{1}{R}\left(\alpha^2 j^2 + k^2\right)^2 a(j,k) + \frac{1}{2R}\delta_{k,\pm 1}\delta_{j,0} - \sum_{p=-\infty}^{+\infty}\sum_{q=-\infty}^{+\infty} \alpha a(p,q)a(j-p,k-q)(kp-qj)(\alpha^2 jp + kq) = 0, \quad (12)$$

where Kronecker's delta is used. In particular we obtain

$$a(j,0) + R\sum_{p,q} a(p,q)a(j-p,-q)\frac{pq}{\alpha j^2} = 0. \quad (13)$$

This suggests the following asymptotic relation:

$$a(j,0) = O(\alpha^{-1}) \qquad \text{as} \qquad \alpha \to 0,$$

which we assume from now on. Also we assume that

$$a(j,k) = O(1) \qquad \text{as} \qquad \alpha \to 0 \qquad (\ k \neq 0).$$

These asymptotic relations are compatible with our numerical experiment, which we can not show because of the page limitation. We now define $b(j,k)$ as follows: $b(j,k) = \lim_{\alpha\to 0} a(j,k)$ for $k \neq 0$ and $b(j,0) = \lim_{\alpha\to 0} \alpha a(j,0)$. Then, we have the following equations:

$$\frac{1}{R}k^4 b(j,k) + \frac{1}{2R}\delta_{k,\pm 1}\delta_{j,0} - \sum_{p\neq j} b(p,k)k^3(p-j)b(j-p,0) = 0, \quad (k \neq 0) \quad (14)$$

$$b(j,0) + R\sum_{p\neq 0, q\neq 0} b(p,q)b(j-p,-q)pqj^{-2} = 0. \quad (15)$$

After some computation, with the aid of symmetry $b(-j,k) = b(j,-k) = (-1)^{j+k-1}b(j,k)$, we can prove that $b(j,k) = 0$ if $|k| > 1$. Then we can rewrite (14) by means of $\{b(j,1)\}_{j=-\infty}^{+\infty}$ only:

$$b(j,1) + \frac{1}{2}\delta_{j,0} - R^2 \sum_{p-j:\text{odd}} \sum_{s\neq 0} b(p,1)\frac{2s(-1)^{s-1}}{j-p} b(s,1)b(j-p-s,1) = 0. \quad (16)$$

This equation is supplemented by the equation (15) with $q = \pm 1$.

We have seen that the Navier-Stokes equations (12), which are written by the two-dimensional array $\{a(m,n)\}$, are reduced to a system of nonlinear equations of a one-dimensional array $\{b(j,1)\}$. We have thus achieved a substantial reduction. However, the reduced equation contains a new difficulty. In fact, the equation (16)

has a trivial solution such that $b(0,1) = -1/2, \quad b(j,1) = 0 \quad (j \neq 0)$. Linearizing (16) at this trivial solution, we can easily see that the system (16) possesses one and only one bifurcation point, which degenerates with *infinite multiplicity*. Consequently, the set of solutions of (16) near the trivial solution would look like Figure 4 (left). Note, however, that this figure is based on a naive guess from the linear analysis and the truth may well be different.

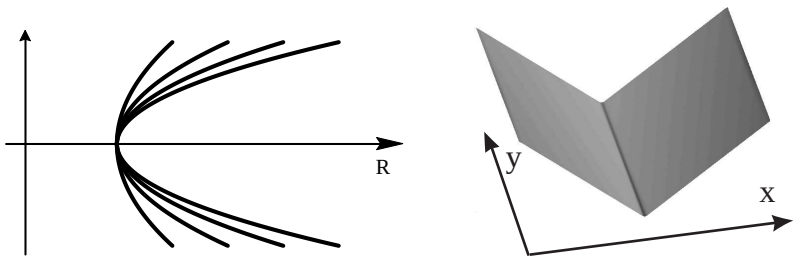


Figure 4: Infinite number of pitchforks at $(\alpha, R) = (0, \sqrt{2})$ (left). Graph of the stream function of mode 1. $\alpha = 0.02, R = 100.0$. Since x ranges from -50π to 50π, it is rescaled to a scale similar to y (right).

Now let us come back to the equation (4). Figure 4 (right) is the graph of the numerical stream function when α is small. This and the Fourier analysis above suggest that

$$\psi \sim \frac{f(\alpha x)}{\alpha} + g(\alpha x)\cos y + h(\alpha x)\sin y + O(\alpha) \qquad \text{as } \alpha \to 0, \tag{17}$$

where f, g, and h are functions of one variable. Figure 4 (right) shows that $f(\xi) = \mu(R)\left(|\xi| - \frac{\pi}{2}\right)$, where $\mu(R)$ is a constant depending on R. Substituting (17) into (5), we obtain

$$g(\xi) = \frac{-1}{1 + R^2(f'(\xi))^2}, \qquad h(\xi) = \frac{Rf'(\xi)}{1 + R^2(f'(\xi))^2}.$$

Thus, the solutions are calculated up to order $O(1)$ as $\alpha \to 0$. However, further expansion accompanies a substantial difficulty.

The analysis of the singular perturbation problems of this and the preceding sections will be left to the future work.

REFERENCES

[1] A.L. Afendikov and K.I. Babenko, Bifurcations in the presence of a symmetry group and loss of stability of some plane flows of a viscous fluid, Soviet Math. Dokl., vol. 33 (1986), pp. 742–747.

[2] D. Armbruster et al., Symmetry and dynamics for 2-D Navier-Stokes flows, Physica D, vol. 95 (1996), pp. 81–93. M.S. Jolly, Bifurcation computations on an approximate inertial manifold for the 2D Navier-Stokes equations, ibid, vol. 63 (1993), pp. 8-20. Z.-M. Chen and W.G. Price, Long-time behavior of Navier-Stokes flow on a two-dimensional torus excited by an external sinusoidal force, J. Stat. Phys., vol. 86 (1997), pp. 301–335.

[3] V. I. Arnold, Kolmogorov's hydrodynamic attractors, Proc. R. Soc. Lond. A, vol. 434 (1991), pp. 19–22.

[4] P. Constantin, Some mathematical problems of fluid mechanics, in Proceedings if ICM 1994, ed. S. G. Chatterji, Birkhäuser (1995), pp. 1086–1095.

[5] V. I. Iudovich, Example of the generation of a secondary stationary or periodic flow when there is loss of stability of the laminar flow of a viscous incompressible fluid, J. Appl. Math. Mech., vol. 29 (1965), pp. 527–544.

[6] H.B. Keller, Lectures on Numerical Methods in Bifurcation Theory (Tata Institute of Fundamental Research No. 79), Springer Verlag (1987).

[7] A.J. Majda, The interaction of nonlinear analysis and modern applied mathematics, Proceedings of ICM 1990, ed. I. Satake, Springer Verlag, (1991), pp. 175–191.

[8] M. Matsuda and S. Miyatake, Bifurcation curves on Kolmogorov flow, preprint.

[9] L.D. Meshalkin and Y.G. Sinai, Investigation of the stability of a stationary solution of a system of equations for the plane movement of an incompressible viscous liquid, J. Appl. Math. Mech., vol. 25 (1962), pp. 1700–1705.

[10] Y. Nishiura, Global structure of bifurcating solutions of some reaction-diffusion systems, SIAM J. Math. Anal., vol. 13 (1982), pp. 555–593. Y. Nishiura and H. Fujii, Stability of singularly perturbed solutions to systems of reaction-diffusion equations, ibid., vol. 18 (1987), pp. 1726–1770.

[11] H. Okamoto and M. Shōji, Bifurcation diagrams in the problem of incompressible viscous fluid flows in 2D tori, Japan J. Indus. Appl. Math., vol. 10 (1993), pp. 191–218.

[12] H. Okamoto, A variational problem arising in the two dimensional Navier-Stokes equations with vanishing viscosity, Appl. Math. Lett., vol. 7 (1994), pp. 29–33.

[13] H. Okamoto, Nearly singular two-dimensional Kolmogorov flows for large Reynolds number, J. Dynamics and Diff. Eqns., vol. 8 (1996), pp. 203–220.

Hisashi Okamoto
Research Institute for Mathematical Sciences,
Kyoto University, Kyoto, 606-8502 Japan,
okamoto@kurims.kyoto-u.ac.jp

Computation with Wavelets in Higher Dimensions

Jan-Olov Strömberg[1]

Abstract.

In dimension d, a lattice grid of size N has N^d points. The representation of a function by, for instance, splines or the so-called non-standard wavelets with error ϵ would require $O(\epsilon^{-ad})$ lattice point values (resp. wavelet coefficients), for some positive a depending on the spline order (resp. the properties of the wavelet). Unless d is very small, we easily will get a data set that is larger than a computer in practice can handle, even for very moderate choices of N or ϵ.

I will discuss how to organize the wavelets so that functions can be represented with $O((\log(1/\epsilon))^{a(d-1)}\epsilon^{-a})$ coefficients. Using wavelet packets, the number of coefficients may be further reduced.

1991 Mathematics Subject Classification: Primary Secondary
Keywords and Phrases: Wavelets, high dimension .

1 Introduction

Although we live in a three-dimensional space it is often useful to consider spaces with much higher dimension. For example, describing the positions in a system with P particles we may use a $3P$ dimensional space. Although we in theory can work very high dimesnsion, it is very limited what we can do in practical numerical computations. The are properties of the geometry in very high dimension that may be surprising.. For example, consider a geometric object as simple as a cube in $\mathbf{R}^d$. A cube with side length as small as a finger nail may still contain a sets as large as the earth on a three-dimensional subspaces, provided d is large enough.

The fundamental issue of analysis in high dimensions involves the approximation to prescribed accuracy of transformations of high dimensional data. Approximating functions with a grid of size N in dimension d means that we have the N^d grid points with data. With the limited amount of data we can handle in

[1]Professor at University of Tromsø, Norway until 1998. Supported by Norwegian Research Counsil

practice, this imposes strong restrictions both on N and on d. The current state of approximation theory is essentially useless in dimensions larger than 10.

One may think that this problem would be solved with faster and faster computers. But there are limitations how fast computers can be.. Let us illustrate this by the following example of "ultimate massive parallel super computer": (I have taken physical constants from a standard physics handbook.)

Let the number or parallel processors be as many as the estimate of total number of protons in universe, let clock cycle speed on each processor defined by the time to travel the distance of a nuclear radius at the speed of light, and finally let running time be as long as estimated live time of universe. Totally this will be about 10^{120} cycles. This correspond to a the number of grid points N^d with $N = 256$ and with $d \approx 50$, For a systems or P particles ($d = 3P$), this means that P can not be larger than 17. Let f function be in the unit cube in $\mathbf{R}^{3P}$ with first order derivatives bounded by 1. Then f may approximated in this grid with accuracy $1/10$ (when $P = 17$). In reality, it seems to be beyond our reach to handle $P > 3$ or maybe even $P > 2$ particles.

The theory for numerical computation in high dimensions is in a premature state, but the approach of Jones; Davis and Semmmes([5]), is the first indication that a powerful theory for high dimensions exists.

In the rest of this paper, I will discuss some ideas which in practice are useful only in rather low dimensions (≤ 10).

In recent years wavelet methods have appeared as useful tools for reducing complexity in numerical computations. By expanding functions in wavelet coefficients one has been able to compress the data to be handled. For example, consider Singular Integral Operator on functions on $\mathbf{R}$ bounded on L^2 and with kernel $K(x, y)$. Assume the kernel satisfies the standard decay properties away from the diagonal:

$$|\partial_x^\alpha \partial_y^\beta K(x, y)| \leq C|x - y|^{-d-\alpha-\beta} \tag{1}$$

for $\alpha, \beta| \leq m$. Representing the operator with a $N \times N$ matrix, we need to use essentially all N^2 matrix elements, even if the elements far away from the diagonal are very small - the total contribution of all the matrix elements on distance $\approx 2^j$ from diagonal will not decay by j. In the famous paper by Beylkin, Coifman, and Rokhlin([2], the authors has shown how the Singular Integral Operators can be expressed in a wavelet basis with a matrix, where the elements decays much faster away from the diagonal. In fact, if the accuracy level is ϵ one may use a diagonal band limited matrix with bandwidth proportional to $\epsilon^{\frac{1}{m}}$. Here m is the order of the wavelets. Thus one need only to use $NO(\epsilon^{-\frac{1}{m}})$ non-zero matrix elements.

In dimension $d > 1$ one has often used the so called non-standard tensor extension , using tensor product combinations of the one-dimensional wavelets and and its scale functions, with all factors of the same scale. Let $M = N^d$, the number of grid points in which the functions arthe represented. Instead of using all M^2 elements to represent the Singular Integral Operator, one need only use a matrix with non-zero element limited to a band around the diagonal ($x = y$), For accuracy level ϵ one need to use $MO(\epsilon^{-\frac{d}{m}})$ non-zero matrix elements. When d is large and m small the number of terms $O(\epsilon^{-\frac{d}{m}})$ increases very fast as ϵ gets smaller.

Even in as low dimension as $d = 3{,}4$ or 5 we feel the restriction on how small ϵ may be. We shall see than under certain circumstances, with good control of the mixed variation of f, the exponential dependence of d in $O((1\epsilon)^{\frac{d}{m}})$ may be replaced with the somewhat better, but still exponential expression; $O(\log \frac{1}{\{}\epsilon^{\frac{1}{m}})^{d-1}\epsilon^{-\frac{1}{m}})$. The error in the approximation, in sup norm, is of magnitude $O(\log \frac{1}{\{}\epsilon^{\frac{1}{m}})^{d-1}\epsilon)$. More exactly, if the smoothness condition on f is stated with the expression $\alpha_1 + \cdot + \alpha_d <= m$ replaced by $\max \alpha_i \leq m$. Here $\alpha = (\alpha_1, .., \alpha_d)$ is the multi-index for the derivative $D^\alpha f$. This is a smoothness condition that is especially suitable to use for functions which are tensor products as $f_1(x_1) \cdots f_d(x_d)$, or for functions that behaves almost like such tensor products.

The ideas ,which are presented here ,comes from some very trivial observations I did trying to work with wavelets in dimension $d > 2$: First, as said above, we very easily get a terrible amount of data.

Second, the full tensor extension of the wavelets to higher dimensions seems to give better compression of the data than the, now classical, non-standard tensor wavelet extension.

This is certainly not a new observation. There have, for example, been arguments for using the full tensor wavelet expansion on R^2 in image compression. I have also seen a mixed tensor wavelet representation for Operators on functions on $\mathbf{R}^d$: The non-standard wavelet tensor basis on $\mathbf{R}^d$ was extended to a basis on $\mathbf{R}^d \times \mathbf{R}^d$ by a full tensor extension of the d dimensional wavelet basis.

I have, so far, only made rather trivial estimates using ideas with full wavelet tensor extension. My Ph.D student Øyvind Bjørkås has done some more detailed studies in his Cand. Scient. Thesis. ([3]).The implementations would be longer future projects Some of the ideas presented in here were communicated to R.R. Coifman, who in a joint paper with D.L. Donoho ([4] has used them in the setting of stochastic variables and their distribution functions. A tensor wavelets expansion in 3-dimension has been used by Averbuch, Israeli and Vozovoi ([1]) to implement a fast PDE solver.

We will, in this presentation, only consider the case $m = 1$. In this case the wavelet functions are the classical Haar functions. However, it is not difficult to generalize the corresponding statements to general $m > 0$. One may also, without much difficulty, generalize to the situation with fractional smoothness conditions (as multi-Lipschitz conditions).

2 PRELIMINARIES

Let χ be characteristic function of interval $[0, 1]$, and the Haar function

$$H(x) = \begin{array}{rll} -1 & \text{when} & 0 \leq x \leq \frac{1}{2} \\ 1 & \text{when} & \frac{1}{2} < x \leq 1 \end{array}$$

The family of Haar functions in dimension $d = 1$ is defined by

$$H_k j(x) = 2^{j/2} H(2^j x - k)$$

We also define the corresponding set of so called scale functions by

$$\chi_k j(x) = 2^j \chi(2^j x - k)$$

With this notation the set of Haar functions H_{kj}, $0 \le k < 2^j, j \ge 0$. together with function χ_{00} is an orthonormal basis on the unit interval $[0,1]$ in $\mathbf{R}$ In this paper I prefer to work the the L^∞ normalized Haar functions $h_{j,k} = 2^{-j/2} H_{jk}$ and the L^1 normalized dual functions $\tilde{h}_{jk} = 2^{j/2} H_{jk}$. We also write $\psi = 2^{-j/2}\chi_{jk}$ and $\tilde{\psi}_{jk} = 2^{j/2}\chi_{jk}$ The expansion of a function f on $[0,1]$ with the Haar wavelets then may be written

$$f =< f, \tilde{\psi} > \psi + \sum_{jk} < f, \tilde{h}_{jk} > h_{jk}.$$

Clearly, $| < f, \tilde{\psi} > | \le |||f||_\infty$. Also, if there is a constant $A > 0$ such that

$$|f(x) - f(y)| \le A|x - y| \tag{2}$$

then

$$| < f, \tilde{h}_{jk} > | \le \frac{1}{4} A 2^{-j}. \tag{3}$$

Note that the factor 2^{-j} is equal to the length of the supporting interval of $\tilde{h}_{jk}$

3 The non-standard tensor wavelet extension

I non-standard tensor extension, all the 2_d combinations of the wavelets h_{jk} and the scale functionswhere for each scale j are used exept for the tensor product where all factors are scale functions. The latter tensor product is unly used on coarses scale. Estimating the wavelets coefficients we have the the worst cases tensor products with only on wavelet factor and thus only one directions where we have the estimate decreasing with j as above . The number of functions needed for accuracy level ϵ is $(1\epsilon)^d$. We will later compare this with the full wavelet tensor expansion (See figure 1)

4 Mixed variation

With good control of the mixed variation ,we will get better estimates for the wavelet coefficients in the full tensor wavelet expansion. To define what we mean with mixed variation, we need some definition. Let R be a rectangle in $\mathbf{R}^d$ of dimension s, $0 \le s \le d$, which is parallel to the axes. Let $Corner(R)$ be the set of corners of R. For $p = (x_1, \ldots, x_d) \in Corner(R)$ we associate the number δ_p equal to $+1$ or -1. We do this by setting $p = 1$ at the point p for which $x_1 + \cdot + x_d$ is maximal and by changing the sign of the value of p as me move along each of the edges (of dimension 1) of R. The difference operator $\triangle_R$ is defined bye

$$\triangle_R f = \sum_{p \in Corner(R)} \delta_p f(p)$$

(In the case $s = 0$ R is a point and $p = R$ and $\triangle_R f = f(p)$.)

DEFINITION 1 *A function f on the unit cube I^d in $\mathbf{R}^d$ is of bonded mixed variation of order $m = 1$ and with constant A if*

$$|\triangle_R f| \leq A(|R|_s) \tag{4}$$

for each rectangles R parallel to the axes and of dimension $s, 0 \leq s \leq d,$. We use notation $|R|_s$ for the s dimensional volume of R. (In the case $s = 0$ R is a point and $|R|_s = 1$.)

Let $M(\underline{x}_1, \ldots, \underline{x}_m) = \underline{x}_1 + \cdots + \underline{x}_m$ be a mapping $\mathbf{R}^{md}$ to $\mathbf{R}^d$

DEFINITION 2 *The function f onf I^d has bounded mixed variation higher order m with constant A if for the function $F(\underline{x}) = f(M(\underline{x})$ we have*

$$|\triangle_R F| \leq A|R|_s \tag{5}$$

for each sub-rectangle R in $\mathbf{R}^{md}$ of of dimension $s, 0 \leq s \leq md$ contained in I^d

As a direct consequence of the classical mean value theorem, the condition 4 holds for any function f satisfying the mixed derivative condition $\partial^\alpha f/\partial x^\alpha$ satisfying

$$|\partial^\alpha f/\partial x^\alpha| \leq A \tag{6}$$

for all multi index $\alpha = (\alpha_1, \ldots, \alpha_d)$ with

$$\max \alpha_i \leq m \tag{7}$$

5 THE FULL TENSOR WAVELET EXPANSION

In the full tensor wavelet expansion we use tensor products

$$\eta_J = \eta_{j_1,k_1} \otimes \cdots \otimes \eta_{j_d,k_d},$$

where $\eta_R = \eta \in \{h, \chi\}$ with $0 =< k_i < 2^{j_i}$. Here the scale index in the i-direction $j_i \geq 0$ when $\eta = h$ and $j_i = 0$ when $\eta = \chi$. R indicates the supporting rectangle of the this wavelet function. For the full tensor wavelet extension of the one dimensional wavelet we get the estimate of the of the coefficients related to the volume of the supporting rectangle. In case of the Haar wavelet extension we have, more precisely,

$$| < f, \tilde{\eta}_R > | \leq (1/4)^s A|R|_d \tag{8}$$

where s is the number of h factors in the tensor product η_R. This is much better estimate than the estimate for the non-standard extension of the wavlets where a coefficients in the worst case are related to the side length of the supporting cube.

6 THE MULTI SCALE SPACE GRID

In dimension one we have a sequence of nested spaces V_j which we may think of as points along a line. In the Haar case V_j is the space of functions partially constant on intervals of length 2^{-j}. In dimension $d \geq 1$ we will consider the spaces

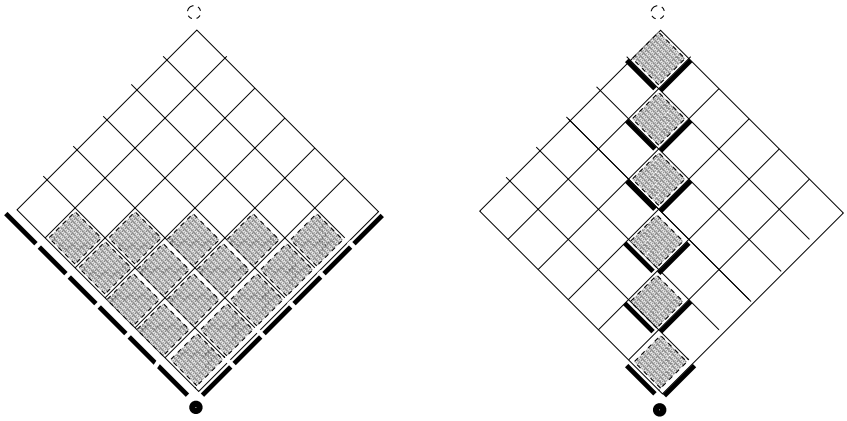


Figure 1: (*left:*) Approximation with full tensor expansion, (*right:*) Approximation with non-standard tensor expansion

$V_j = V_{j_1,\ldots j_d} = V_{j_1} \otimes \cdots \otimes V_{j_d}$ as points $j = (j_1, \ldots, j_d)$ in a d dimensional integer grid. Let P_j be the projection to V_j. We evaluate a function f at the space point j as the function $P_j f$. As above may define the mixed different $\triangle_R$ for any axes parallel rectangle R on this grid (with dimension $s, 0 \leq s \leq d$). One can show that $\triangle_R$ is a projection. In this space grid we identify rectangles R as the space corresponding to projection $\triangle_R$ Now, we may make simple algebraic rules of how any rectangle R and its lower dimensional boundaries are related. We may also add together collections of rectangles R_i and express the sum in terms of their union and its boundaries. We introduce the variable $J = j_1 + \cdots + j_d$ and turn out multi-scale grid with the J-axis (not drawn) pointing vertically upwards:

Let Q be the whole Multi Scale Space Grid as a cube in $\mathbf{R}^d$ The space corresponding to the top point of Q is $V_{n,\ldots,n}$, while the bottom point is the space $V_{0,\ldots,0}$ In the full tensor expansion of the wavelets the top point is decomposed as a direct sum of all spaces corresponding to all the small s-dimensional cubes, $0 \leq s \leq d$, lying on those s - dimensional boundary cubes of Q, which contains the bottom point. These spaces are indicated on the left part of figure 1 as all filled squares,all bold line segment and finally the bottom point.

7 Approximation with full tensor wavelets expansions

The main observation is, that the a priori estimate of the wavelet coefficient for a the subspace in this decomposition is essentially proportional to

$$2^{-J}$$

or the J - coordinate of the position of this subspace. On the other hand the number of bases element in the subspace is proportional to

$$2^J$$

This means that it strategic to approximate a functions in the top space by using the projection with all subspaces with J coordinate under some level, such as $J \leq n$. We get

THEOREM 1 *Let $\epsilon > 0$. Then there is a set S with $O((\log \frac{1}{\epsilon})^{d-1}\epsilon^{-1})$ Haar full tensor wavelets functions such that any function f be a function in the unit cube in $\mathbf{R}^d$ satisfying condition 7 may be approximated by*

$$\sum_{h \in S} < f, \tilde{h} > h$$

with accuracy in sup norm $O((\log \frac{1}{\epsilon})^{d-1}\epsilon)$ (in L^2 norm $O((\log \frac{1}{\epsilon})^{fracd-12}\epsilon)$). The value of f at a single point may be computed in $O((\log \frac{1}{\epsilon})^d$ steps.

8 A SPARSE SET OF RECTANGLES

The approximation in the theorem above is done with subspaces with space grid coordinate $J \leq N$. All those subspaces are in the span of the subspaces V_j with the coordinate $J = n$. This means in the Haar case we that the mean f_{R_0} may be computed from the mean values f_R, R dyadic rectangle (with some accuracy).

THEOREM 2 *Let f be a function on the unit cube in $\mathbf{R}^d$ satisfying condition 7. Let R_0 be a dyadic sub rectangle with volume $|R_0|_d \leq 2^{-n}$. and let $\mathcal{R}_k$ be the set of dyadic rectangle R in the unit cube with volume $|R|_d = 2^{-k}$ and let f_R the mean value of f over R Then, with errorr $O(n^{d-1}2^{-n})$ we have*

$$f_{R_0} \approx \sum_{k=0}^{d-1} (-1)^k \begin{pmatrix} d-1 \\ k \end{pmatrix} \sum_{\substack{R \in \mathcal{R}_{n-k} \\ R \supset R_0}} f_R$$

9 A SPARSE SUBSET OF GRID POINTS

We would not have much practical use of the full tensor approximation with with the sparse set of $O((\log N)^{d-1}N)$ wavelet coefficients if, in order to compute them, we need all N^d samples points of the function. However, it is not very difficult to show that that these wavelet coefficients may be calculated from a sparse set of sample values of the function.

Let G_N be the set of grid points in the unit cube in R^d where the grid size is $1/N$ Let S_N be the subset of G_N consisting of all corners of dyadic rectangles with volumes $\geq 1/N$. The number of points in S_N is $O((\log N)^{d-1}N)$.

THEOREM 3 *Let f be a function satisfying 5. Given the value of f at S_N one may also compute the the sparse set of $(O((\log N)^{d-1}N)$ coefficients $< f, \tilde{h}_{\underline{jk}} >$ with an error bounded by $O((\log N)^{d-1}\frac{1}{7}N)$ (with $\tilde{h}_{\underline{jk}}$ L^1- normalized) This can be done in $O((\log N)^{d-1}N)$ steps.*

There is also a fast algorithm to recover the values of functions at any grid point in G_N

THEOREM 4 *Let f be a function satisfying 5. Given the value of f at all points in S_N one may compute the values of f at any point $p \in G_N$ with error bounded by $O((\log N)^{d-1}\frac{1}{N})$. The complexity of such a computation is $O((\log N)^{d-1})$.*

We do not have room here to include any proofs of this.

10 TENSOR WAVELETS ON SINGULAR INTEGRAL OPERATORS

Wavelets was used with great success, by Beylkin, Coifman and Rokhlin with great success to reduce complexity for the computation of Singular Integral Operators . The operator T is assumed to be bounded on $L^2(\mathbf{R}^d)$ and its kernel $K(x,y)$ satisfies the usual smoothness conditions 1 away away from the diagonal. Let us consider the problem as to compute the inner product

$$< f, Tg >=< f \otimes g, K > .$$

We may think of K represented by a matrix ($2d$ dimensional tensor) with N^{2d} elements, $N = 2^n$. Beylkin, Coifman and Rokhlin represent this inner product by use the non-standard tensor bases extension of the wavelets (of order m) on $\mathbf{R}^{2d}$. In this bases the Kernel is represented by a matrix, which may by approximated with accuracy level ϵ to a matrix with finite diagonal bandwidth containing $N^d O(\epsilon^{\frac{d}{m}})$ non-zero elements. We see that estimate for the number of non-zero elements in this matrix is not very good when d is large. Using a hybrid of non-standard and full tensor basis extension it is possible to improve their result:

THEOREM 5 *There is a hybrid tensor wavelet extension basis on the unit cube I^{2d} in $\mathbf{R}^{2d}$, in which the kernel K is represented with accuracy level ϵ by a matrix with $N^d O(1/\epsilon)$ non-zero elements. The coefficients of K in this basis may be computed in $O(N^{2d})$ steps. The wavelet coefficients of $f \otimes g$ in this basis are simple products of coefficients taken from on set of coefficients for f and one set set of coefficients for g, Each of these two sets is order $O(N^d)$ and may be computed in $O(N^d)$ steps.*

We will not give the proof here.

11 THE HAAR PACKETS

A C^∞ function may be approximated much better with wavelet packets than by wavelets. We will for simplicity only consider wavelet packets of the Haar functions for approximating functions f satisfying

$$\frac{d^k|f}{dx^k}(x)| \leq 1 \text{ for all } k \geq 0 \tag{9}$$

Then we have

THEOREM 6 *Let $\epsilon > 0$ then there is as set of*

$$M_\epsilon = O\left(\exp\left(c\sqrt{\log(\frac{1}{\epsilon})}\right)\right)$$

Haar wavelet packet functions $\{W_k\}_{k=1}^{M_\epsilon}$ on the interval $[0,1]$, such that any function f as above is approximated with error less than ϵ by

$$f \approx \sum_{k=1}^{M_\epsilon} < f, W_k > W_k.$$

The smoothness condition 9 on the function f is very strong. However, one may get some rather similar estimates with much weaker condition on f. The wavelet packe tree starting from the top consists of nodes, which are connected by branches of low-pass and high-pass filters. At the bottom of the tree, each space is of dimensions one. Let $j_1, j_2, \ldots, j_s$ be a sequence of positve integers which indicates the levels of the branshes, where we have passed through the high-pass filter to reach a node. Then we get, by iteration of the mean value theorem, the a priori estimate

$$2^{-(j_1+\cdots+j_s)}2^{-2s}.$$

(We assume have normalized the filters so that the Low-pass filter does not increase the norm.) Let $2^{-n} = \epsilon$. To get the Theorem above we mainly hove to solve the following problem in combinatorics: Estimate the number of finite sequences of integers $(j_1, j2, \cdots j_s)$ with

$$0 < j_1 < j_2 < \cdots < j_s \leq n$$

such that

$$j_1 + j_2 + \cdots + j_s \leq n.$$

Approximating with the Haar packets on the unit cube i $\mathbf{R}^d$ leads to a similar combinatorics problem with sets where up to d indices j_k may assume the same integer values. One will get the estimate

$$M = O\left(\exp\left(c\sqrt{d\log(\frac{1}{\epsilon})}\right)\right)$$

The problem with the approximating Singular Integral Operator kernels with Haar packets will also lead to a combinatoric problem. As in the compression of the kernel K by Haar wavelets above one my use Whitney decomposition also in this case. We have not analized this in detailed yet. However, one would probably get something like, that the kernel K, may be approximated at accuracy level ϵ using a set of

$$N^d O\left(\exp\left(c\sqrt{d\log(\frac{1}{\epsilon})}\right)\right),$$

Haar Packet functions.

REFERENCES

[1] Averbush, A., Israeli, M. an Vozovoi, L. . A fast Poisson solver of arbitrary order accuracy in rectangular regions.*SIAM J. Sci. Comput. 19 (1998), no. 3, 933–952.*

[2] G. Beylkin, G., Coifman, R. and Rokhlin, V, Fast wavelet transforms and numerical algorithms. *Comm. Pure. Appl. Math., 1991.*

[3] Bjørkås, Ø. Numerical Analysis of Singular Integral Opteratos with Wavelets. Cand. Scient Thesis at University of Tromsø, 1997.

[4] Coifman, R. and Donoho, D. To appear

[5] David, G, and Semmes,S *Fractured fractals an broken dreams.*Self-similar geometry through metric and measure. The Clarendon Press, Oxford University Press,NewYork, 1997, pp. 212

Jan-Olov Strömberg
Department of Mathematics
Royal Institute of Technology
Stockholm
Sweden
janolov@math.kth.se

Schwarz–Christoffel Mapping in the Computer Era

Lloyd N. Trefethen and Tobin A. Driscoll

Abstract. Thanks to powerful algorithms and computers, Schwarz–Christoffel mapping is a practical reality. With the ability to compute have come new mathematical ideas. The state of the art is surveyed.

1991 Mathematics Subject Classification: 30C30, 31A05
Keywords and Phrases: conformal mapping, Schwarz–Christoffel formula

1. Introduction. In the past twenty years, because of new algorithms and new computers, Schwarz–Christoffel conformal mapping of polygons has matured to a technology that can be used at the touch of a button. Many authors have contributed to this progress, including Däppen, Davis, Dias, Elcrat, Floryan, Henrici, Hoekstra, Howell, Hu, Reppe, Zemach, and ourselves. The principal SC software tools are the Fortran package SCPACK [15] and its more capable Matlab successor, the Schwarz–Christoffel Toolbox [3]. It is now a routine matter to compute an SC map involving a dozen vertices to ten digits of accuracy in a few seconds on a workstation.

With the power to compute has come the ability to explore. The obvious kind of problem—"here is a polygon; map it onto a disk or a half-plane"—is rarely what one encounters in practice. Instead it is variations on the idea of SC mapping that arise. In this article, after briefly mentioning the algorithmic developments that have made SC mapping possible, we describe four of these variations: oblique derivative problems on polygons; ideal free-streamline flows; the CRDT (cross-ratio Delaunay triangulation) algorithm; and Green's functions for symmetric multiply connected domains.

2. Numerical Algorithms. Independently around 1869, Schwarz and Christoffel derived their famous formula,

$$f(z) \;=\; A + B \int_0^z \prod_{k=1}^{n} (1 - \zeta/z_k)^{-\beta_k}\, d\zeta. \tag{1}$$

They proved that any conformal map $f(z)$ of the unit disk or the upper-half plane onto a polygon P with n vertices can be written in the form (1) for some constants A, B, $\{z_k\}$, and $\{\beta_k\}$ [9,10]. The exponents $\{\beta_k\}$ are determined by the angles at the vertices of P (exterior turning angles divided by π), but the other parameters are unknown. Translation, scaling, and rotation are accomplished by A and B, and the *prevertices* $\{z_k\}$, unknown a priori, are the preimages on the unit circle or the real axis of the vertices of P.

Three computational hurdles arise in implementing (1): finding the unknown parameters, evaluating the integrals, and computing the inverse map. Analytically, one can do very little. The hurdles must be crossed numerically, and the history here is long, in part because this is a topic that every engineer has heard of and

may be tempted to solve from scratch. The list of names above includes only some of the more prominent contributors. Two of the earliest contributions in the list were those of Davis [1] and Reppe [13].

We will not give details, but mention a few of the algorithmic ideas that are the basis of SCPACK and the SC Toolbox. Schwarz–Christoffel integrals can be evaluated by an automatic process of *compound Gauss–Jacobi quadrature:* Gauss–Jacobi formulas handle the singularities at endpoints, and adaptive subdivision of intervals combats the phenomenon of exponential distortions that is universal in conformal mapping (see §5). The unknown parameters can be found by solving a system of nonlinear equations that assert that the side lengths of P are correct. By a change of variables, the ordering conditions among the prevertices can be eliminated to make this system of equations unconstrained, whereupon it can be treated by quasi-Newton iteration or by more specialized techniques. Finally, f^{-1} can be evaluated by a Newton iteration (the derivative f' is just the integrand of (1)), made robust as necessary by the generation of initial guesses via numerical solution of an ordinary differential equation.

We urge readers to download a copy of the SC Toolbox [3] and give all this a try. Begin by typing `scgui` to explore the graphical user interface, but remember that everything can also be done by inline Matlab commands. For example,

```
plot(diskmap(polygon([2 2+i 1+i 1+2i 2i 0])))
```

generates a plot of a conformal map of the unit disk onto an L-shaped polygon in five seconds on the first author's workstation. Changing `diskmap` to `extermap` maps the exterior of the same polygon. The command

```
plot(rectmap(polygon([2 2+i 1+i 1+2i 2i 0]),[1 4 5 6]))
```

maps the interior onto a rectangle, whose length-to-width ratio is necessarily the conformal modulus of the L-shaped region, $\sqrt{3}$. Similar SC Toolbox commands construct maps from a half-plane or an infinite strip and onto generalized polygons with slits, vertices at infinity, or overlapping regions.

3. OBLIQUE DERIVATIVE PROBLEMS ON POLYGONS. The following problem, considered by Trefethen and Williams [18], arises in queuing theory and in the study of the classical Hall effect in electronics. On each side Γ_j of a polygon P, an angle θ_j is specified. We seek a non-constant harmonic function u in P, i.e., a solution to Laplace's equation $\Delta u = 0$, that satisfies the condition $du/ds = 0$ along the direction at angle θ_j to the boundary.

Such problems can be solved by conformal mapping as follows. Suppose a function f is found that maps P conformally onto a second polygon Q, whose side lengths are unspecified, with the property that $f(\Gamma_j)$ is oriented at the angle θ_j from the vertical. Then all the boundary directions in Q line up vertically, and therefore $u(z) = \mathrm{Re} f(z)$ is the function required. Theorems 1–3 of [18] establish that this procedure generates all solutions to the oblique derivative problem.

Figure 1 shows an example of an oblique derivative problem solved in this manner. Though only pictures are presented, it is an easy matter to extract numbers from such a computation accurate to ten or more digits.

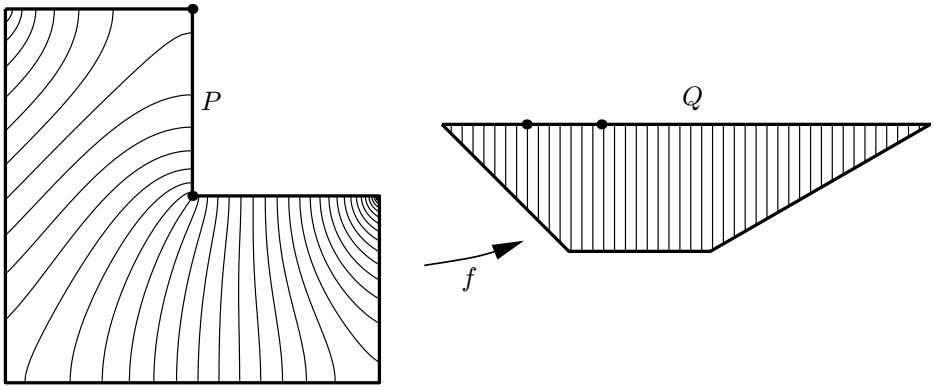


Figure 1. Solution of an oblique derivative problem by conformal mapping. On the L-shaped problem domain P, a bounded harmonic function u is sought with $du/ds = 0$ at angle $\pi/4$ on the left edge, $\pi/3$ on the right edge, and $\pi/2$ (the usual Neumann condition) elsewhere. The required function is $u(z) = \mathrm{Re} f(z)$, where $f(z)$ is a conformal map of P to a trapezoid Q with sides oriented at the prescribed angles from the vertical, and all solutions are of this form, differing only in shift and scale. The preimages of vertical lines in the trapezoid are curves $u(z) = \text{const}$. The two dots show the conformal images of the two vertices in the L that map into degenerate vertices of the trapezoid.

4. Ideal Free-Streamline Flows. A longstanding topic of fluid mechanics is the study of jets, wakes, and cavities, all of which may involve a surface, or in two dimensions a line, across which the flow properties change. In 1868, Helmholtz and Kirchhoff introduced the theory of free streamlines for such problems and proposed the use of complex analysis to find solutions in 2D. Other early contributors were Planck, Joukowski, Réthy, Levi-Civita, Greenhill, and von Mises, and later generations saw major extensions and survey publications by Birkhoff and Zarantonello, Gilbarg, Gurevich, Monakhov, and Wu, among others [12,20].

The classical theory of 2D free-streamline flows has two limitations. The first is that in many cases it omits important aspects of the physics. The second is that the flows in question can be computed analytically for only the simplest geometries. Here, however, it turns out that by a modification of the SC idea, effective algorithms can be devised that are exactly analogous to those used for SC mapping. If one is careful about the physics, as in various papers over the years by J. Keller and by Vanden-Broeck, among others, the results can reveal a great deal about certain flows.

One version of a classical 2D free-streamline flow problem goes like this. A semi-infinite inviscid, incompressible fluid flows in the absence of gravity above a polygonal wall that ends at a point, after which the fluid continues on into free space. Beyond the separation point, the flow is bounded by a curved free streamline on which the boundary condition (because of Bernoulli's equation) is that the magnitude of the velocity must be constant—say, 1.

The solution can be found as follows. Let z be the spatial variable; the boundary of the problem domain in the z-plane is partly unknown, because of the free streamline. Let w be the velocity potential, inhabiting the upper half-plane, with the origin corresponding to the point of separation. Let $\zeta = dw/dz$ be the hodograph or conjugate-velocity variable. The problem is to find an analytic function $\zeta(w)$ such that $\arg\zeta(w)$ takes prescribed piecewise constant values along $(-\infty, 0]$ and has constant modulus $|\zeta(w)| = 1$ along $[0, \infty)$.

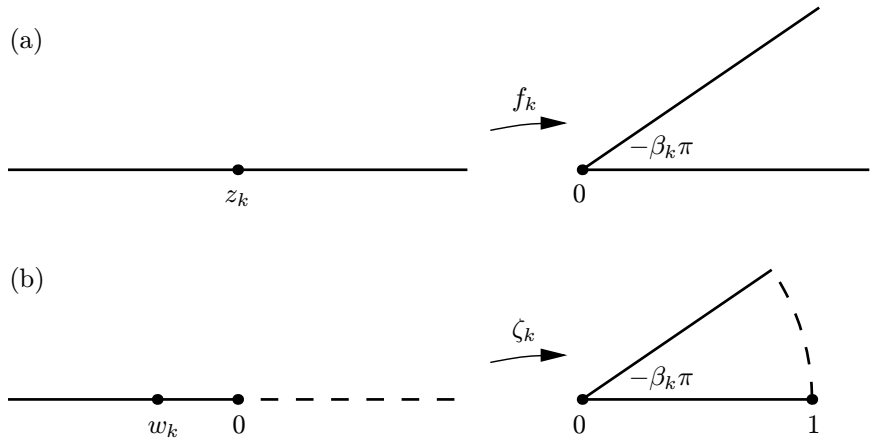


FIGURE 2. Comparison of the ideas underlying Schwarz–Christoffel and free-streamline mapping. (a) For an SC map $f(z)$, the derivative $f'(z)$ has piecewise constant argument for $z \in (-\infty, \infty)$, so it can be written as a product of elementary maps of the upper half-plane onto infinite wedges. (b) For a free-streamline map, $\zeta'(w)$ has piecewise constant argument for $w \in (-\infty, 0]$ and constant modulus for $w \in [0, \infty)$, so it can be written as a product of maps of the upper half-plane onto a circular-arc wedges.

The hodograph (ζ) domain is bounded by straight segments and a circular arc, and the classical approach would be to take the logarithm to reduce it to a polygon, which could then be mapped by the SC formula. For all but very simple geometries, however, this method is unworkable because of unknown ordering of prevertices, for the topology of the polygon is not known in advance. Instead, general polygonal boundaries can be handled by the method suggested in Fig. 2, developed by Monakhov [12] and Elcrat and Trefethen [6]. The key idea is to employ a modification of the SC integral (1) in which each term in the product in the integrand is an elementary map onto a bounded circular-arc wedge rather than an unbounded wedge. A numerical example is presented in Figure 3; for more, see [2] and [6].

The method just described for free-streamline flows is a general one, permitting the ready computation, with slight variations, of flows in a wide variety of geometries. Once such a general tool is in hand it is an easy matter to reproduce computations from the past, including those of Kirchhoff (flat plate, 1869),

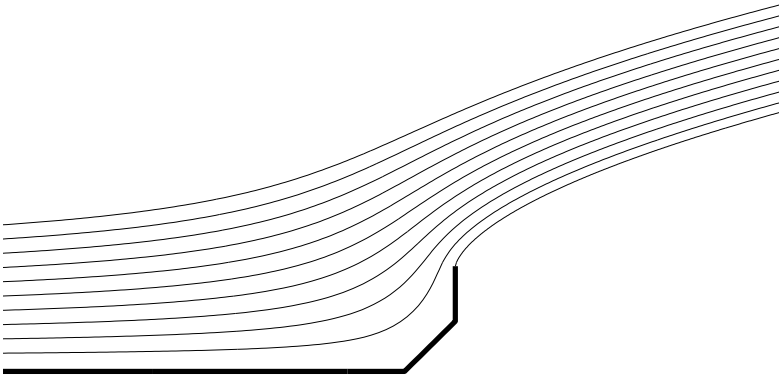


Figure 3. Example of a free-streamline flow computed by the method of Fig. 2. The free streamline is the curve that separates from the tip of the solid boundary. The equal spacing of the curves on the right reveals that the constant-speed condition has been satisfied.

Rayleigh (inclined plate, 1876), Bobyleff (symmetric wedge, 1881), Michell (slot, 1890), von Mises (funnel, 1917), Chaplygin and Lavrentiev (plate with separation, 1933), Keller ("teapot effect", 1957), Lin (asymmetrical wedge, 1961), Wu and Wang (symmetric 4-piece wedge, 1964), and Elcrat (plate with spoiler, 1982).

5. The CRDT (Cross-Ratio Delaunay Triangulation) Algorithm. The methods described above are based on what might be called standard SC mapping technology, in which the realization of (1) is achieved by standard "best practice" methods of numerical analysis. It has been recognized since around 1980, however, that such methods fail for highly elongated regions. Suppose, for illustration, that the unit disk is mapped onto a rectangle of aspect ratio L, with 0 mapping to the center of the rectangle. Then the prevertices along the unit circle lie in two pairs separated by intervals that shrink in proportion to $\exp(-\pi L/2)$. For $L = 30$, for example, adjacent prevertices are separated by only about 10^{-21}. Thus conformal maps are subject to exponential distortions, a phenomenon known as *crowding,* and in floating point arithmetic, the result is that highly elongated regions cannot be treated by standard methods.

One solution, due to Howell and Trefethen [11], is to dispense with the disk or half-plane and map directly onto a highly elongated domain such as an infinite strip or a long rectangle. Both options are included in the SC Toolbox. For many problems arising in practice, this solves the crowding problem by weakening the effect from exponential to algebraic.

Domains that are elongated in multiple directions, however, require more radically new approaches, and one such, also included in the SC Toolobx, has recently been developed by Driscoll and Vavasis [4]. The idea behind the CRDT algorithm is that no matter how distorted a conformal map may be globally, any portion of it can be made locally well-behaved by some Möbius transformation $(az + b)/(cz + d)$. By composing Möbius transformations, it ought to be possible to represent arbitrarily great distortions by compositions of well-behaved maps.

The CRDT algorithm begins by constructing a Delaunay triangulation of the target polygon P, generally after introducing extra degenerate vertices so that the triangles will be not too slender. The task then is to find the correspondence function between prevertices on the unit circle and vertices on the boundary of P. The Möbius idea is used by formulating the correspondence condition not globally but four vertices at a time, corresponding to two adjacent triangles. A convenient, Möbius-invariant description of the unknown prevertices is furnished by the *cross-ratios* of these 4-tuples, defined by

$$\rho(z_1, z_2, z_3, z_4) \;=\; \frac{(z_4 - z_1)(z_2 - z_3)}{(z_3 - z_4)(z_1 - z_2)}, \tag{2}$$

which is negative and real when $z_1, \ldots, z_4$ lie on counterclockwise order on the unit circle. The CRDT algorithm formulates a system of $n - 3$ nonlinear equations in which the independent variables are the negatives of the logarithms of the cross-ratios of 4-tuples of prevertices on the unit circle, and the dependent variables are the deviations from their correct values of the absolute values of the logarithms of the cross-ratios of 4-tuples of vertices on P. This system of equations is observed to be very well behaved and easily solved by iteration.

The algorithm sounds complicated, and its elements of computational geometry certainly give it a flavor different from other algorithms in this field, but it has proved remarkably effective. It makes possible the mapping of regions that would have been regarded as impossible a few years ago, except in multiple precision arithmetic. Figure 4 gives an example.

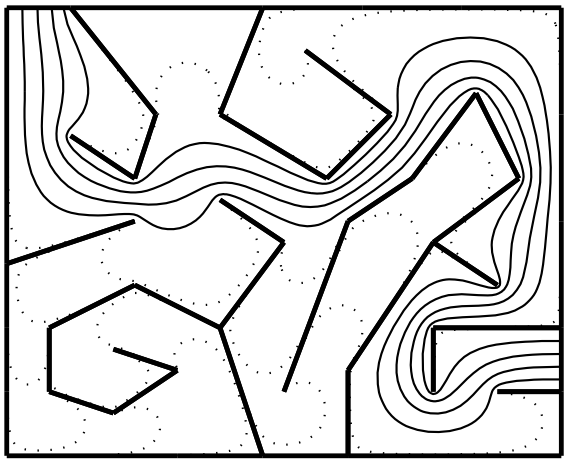


FIGURE 4. Conformal map of "Emma's maze" onto a rectangle of aspect ratio 18.2, computed by the CRDT algorithm. The solid curves are the conformal images of straight lines in the rectangle; the dotted curves are the same, but correspond to lines in the rectangle exponentially close to the sides $(10^{-2}, 10^{-4}, 10^{-6}, \ldots, 10^{-16})$. Because of exponential distortions, the numerical computation of maps like this is far beyond the capabilities of ordinary algorithms.

6. Green's Functions for Multiply Connected Domains. Our final SC variation represents recent work by Embree and Trefethen [6] based on ideas going back at least to Widom [19]. It is a longstanding dream to generalize SC maps to multiply connected domains, but except in the doubly-connected case treated in successive works by Henrici, Däppen, and Hu, no general results of much practical value are available in this line. Consider, however, the special case of a region P in the complex plane consisting of polygons $P_1, \ldots, P_K$ that are symmetrically located along the real axis, illustrated in Fig. 5 for an example with $K = 3$. (In an important special case, the polygons degenerate to intervals along the real axis.) To be specific, suppose we seek the *Green's function* for P, the function $u(z)$ harmonic outside P with boundary values $u(z) = 0$ on the boundary of P and $u(z) \sim \log|z|$ as $z \to \infty$.

Such a Green's function can be computed by SC mapping. The crucial observation is that the upper half of the problem domain, with segments of the real axis inserted as necessary to provide a complete boundary, is simply-connected. Let $g(z)$ be a conformal map of this region onto a semi-infinite strip with vertices πi, 0, and ∞ and $K - 1$ slits of indeterminate height and length along the complex interval $[0, \pi i]$ (Fig. 5(b)). The semi-infinite segments of the real axis map to the sides $[0, \infty)$ and $[\pi i, \infty)$ of the strip, and the bounded segments between the polygons P_j map to the horizontal slits; the boundaries of the polygons themselves map into segments of $[0, \pi i]$. The Green's function required is now given by

$$u(z) = \mathrm{Re}\, g(z). \tag{3}$$

A second conformal transformation may provide further insight. If $f(z) = \exp(g(z))$, then f maps the upper half of the problem domain onto the exterior of the unit disk with protruding spikes (Fig. 5(c)). In terms of this new map, the Green's function is given by

$$u(z) \;=\; \log|f(z)|. \tag{4}$$

This Green's function algorithm has several special features from the SC mapping point of view, both of which arise in certain other applications too. One is that the positions of the slits (spikes) in the semi-infinite strip (exterior of the disk) are unknown a priori, and must be determined as part of the mapping process. Doing so is easy, since this part of the parameter problem enters linearly, as is often the case with slits in SC mapping; but we can view this aspect of the calculation as prototypical of more complicated *generalized parameter problems* that arise in various applications such as inverse problems [16].

The second special feature of this SC computation is that although the Green's function $u(z)$ is single-valued, the conformal maps involved in getting it, if viewed in the large, are multiple-valued. Specifically, consider the function $f(z)$. A priori, it maps the upper half of Fig. 5(a) to the upper half of Fig. 5(c), and reflection in the semi-infinite segments of the real axis completes this to a single-valued map of all of (a) to all of (c). Reflection in the bounded segments of the real axis between the polygons that correspond to the spikes in Fig. 5(c), however, is equally justified mathematically. After one such reflection, further reflections become possible, and

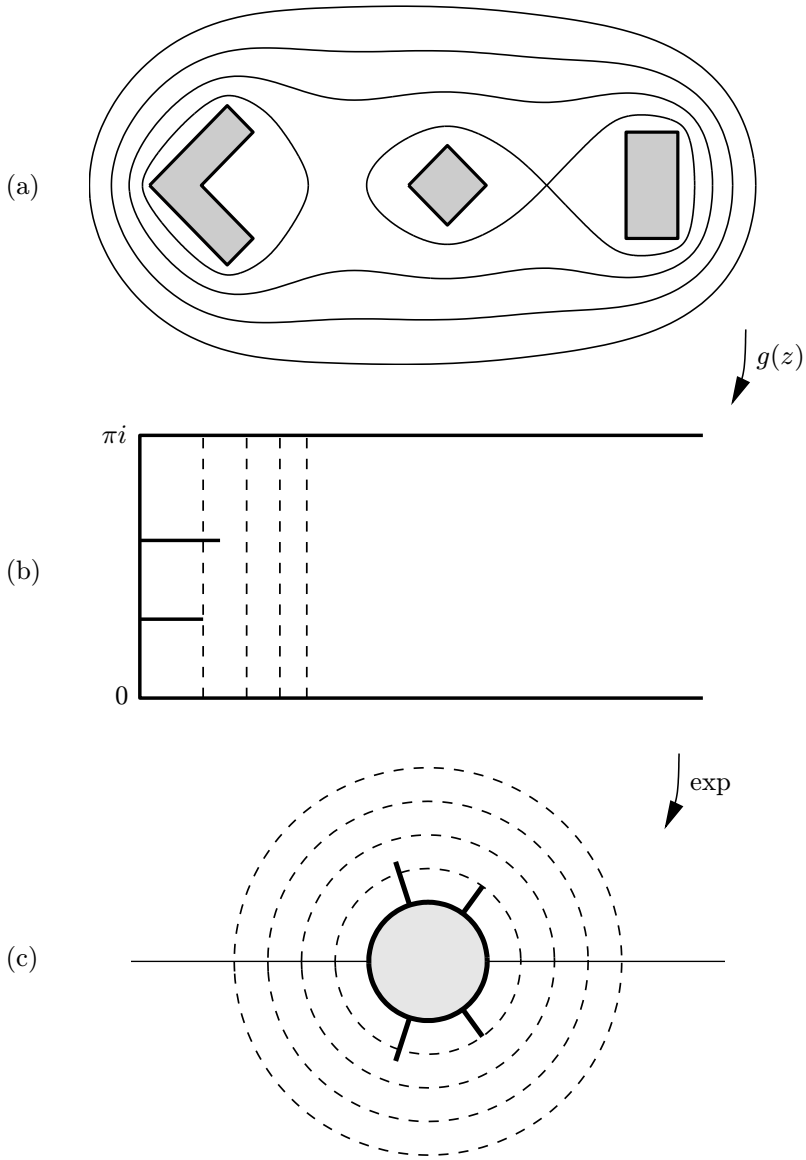


FIGURE 5. Computation of the Green's function $u(z)$ for a multiply connected domain. (a) Final result, showing four level curves $u(z) = \text{const.}$; the innermost is the critical one at which two connected components first touch. (b) To compute u, the portion of the problem domain in the upper half-plane is transplanted by an SC map $g(z)$ to a semi-infinite slit strip; the upper halves of the level curves of (a) are the preimages of vertical lines in the strip (dashed). (c) The exponential function transplants the slit strip to the upper half of the exterior of a disk with spikes, and the level curves of (a) become preimages of concentric circles.

so on, and the ultimate result is that $f(z)$ maps the multiply connected region onto a Riemann surface with (in general) infinitely many sheets [19]. Fortunately, these complications do not matter for the application of computing the Green's function.

Green's functions for multiply connected regions have applications to problems of polynomial approximation. When the polygons reduce to intervals we have a problem in digital filtering [14], and the general case relates to beautiful theorems of Fuchs in the 1970s [8]. Suppose in Fig. 5, for example, that three distinct entire functions f_1, f_2, f_3 are defined and polynomials $p_n(z)$ of increasing degrees n are sought that approximate f_j on P_j in the uniform norm. Then apart from a small algebraic factor, the approximation errors decrease at the rate $\exp(-\beta n)$, where β is the length of the shortest slit in Fig. 5(b). The heights of the slits give asymptotic information about the proportions of interpolation points on each set P_j, with the images of roots of unity in Fig. 5(c) on the boundaries of P_j providing near-optimal interpolation points, and the optimal approximants have the property that they converge precisely inside the critical level curve of Fig. 5(a).

7. Further Variants and Applications. We have touched upon only a few developments in Schwarz–Christoffel mapping in the computer era. Among the variants not mentioned are periodic domains, fractals, circular polygons, curved boundaries, gearlike domains, and polygonal Riemann surfaces. Among the applications not mentioned are Faber polynomials, matrix iterations, the KdV equation, electrical resistances and capacitances, magnetostatics, and vortex methods in fluid mechanics. References for some of these problems can be found in [17].

We would like to put in writing our view of the best applications of Schwarz–Christoffel transformations. Many people have the idea that SC methods may be useful for general geometric purposes such as grid generation for finite differencing, perhaps even for domains with curved boundaries approximated by polygons. Our opinion is that whereas such applications are of course sometimes effective, the real excitement of SC mapping lies elsewhere. What is special about the SC formula is that it solves a certain precisely defined problem exactly, delivering a semi-analytic solution dependent only on a small number of parameters. We favor applications where this semi-analytic solution solves the problem of ultimate interest exactly, for in such circumstances, SC methods far outperform general-purpose tools such as adaptive finite elements.

References

[1] R. T. Davis, Numerical methods for coordinate generation based on Schwarz–Christoffel transformations, *Proc. 4th AIAA Comp. Fluid Dyn. Conf.,* 1979.

[2] F. Dias, A. R. Elcrat and L. N. Trefethen, Ideal jet flow in two dimensions, *J. Fluid Mech. 185* (1987), 275–288.

[3] T. A. Driscoll, Algorithm 765: A MATLAB toolbox for Schwarz–Christoffel mapping, *ACM Trans. Math. Softw. 22* (1996), 168–186. Software available at `http://amath.colorado.edu/appm/faculty/tad/research/sc.html`.

[4] T. A. Driscoll and S. A. Vavasis, Numerical conformal mapping using cross-ratios and Delaunay triangulation, *SIAM J. Sci. Comput.*, to appear.

[5] A. R. Elcrat and L. N. Trefethen, Classical free-streamline flow over a polygonal obstacle, *J. Comput. Appl. Math.* 14 (1986), 256–265.

[6] M. Embree and L. N. Trefethen, Green's functions for multiply connected domains via Schwarz–Christoffel mapping, manuscript in preparation.

[7] J. M. Floryan and C. Zemach, Schwarz–Christoffel mappings: A general approach, *J. Comput. Phys.* 72 (1987), 347–371.

[8] W. H. J. Fuchs, On the degree of Chebyshev approximation on sets with several components, *Izv. Akad. Nauk. Armyan. SSR* 13 (1978), 396–404.

[9] D. Gaier, *Konstruktive Methoden der Konformen Abbildung,* Springer, Berlin, 1964.

[10] P. Henrici, *Applied and Computational Complex Analysis, v. 3*, Wiley, New York, 1986.

[11] L. H. Howell and L. N. Trefethen, A modified Schwarz–Christoffel transformation for elongated regions, *SIAM J. Sci. Stat. Comput.* 11 (1990), 928–949.

[12] V. N. Monakhov, *Boundary-Value Problems with Free Boundaries for Elliptic Systems of Equations,* Trans. of Math. Monographs, vol. 57, Amer. Math. Soc., 1983.

[13] K. Reppe, Berechnung von Magnetfeldern mit Hilfe der konformen Abbildung durch numerische Integration der Abbildungsfunktion von Schwarz–Christoffel, *Siemen Forsch. u. Entwick. Ber.* 8 (1979), 190–195.

[14] J. Shen, G. Strang, and A. Wathen, The potential theory of several intervals and its applications, manuscript in preparation.

[15] L. N. Trefethen, Numerical computation of the Schwarz–Christoffel transformation, *SIAM J. Sci. Stat. Comput. 1* (1980), 82–102.

[16] L. N. Trefethen, Analysis and design of polygonal resistors by conformal mapping, *J. Appl. Math. Phys.* 335 (1984), 692–704.

[17] L. N. Trefethen, Schwarz–Christoffel mapping in the 1980s, TR 93-1381, Dept. of Comp. Sci., Cornell U., 1993.

[18] L. N. Trefethen and R. J. Williams, Conformal mapping solution of Laplace's equation on a polygon with oblique derivative boundary conditions, *J. Comp. Appl. Math. 14* (1986), 227–249.

[19] H. Widom, Extremal polynomials associated with a system of curves in the complex plane, *Adv. Math.* 3 (1969), 127–232.

[20] T. Y. Wu, Cavity and wake flows, *Annual Reviews in Fluid Mech.* (1972), 243–84.

Lloyd N. Trefethen
Oxford U. Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD, UK
LNT@comlab.ox.ac.uk

Tobin A. Driscoll
Dept. of Applied Mathematics
University of Colorado
Boulder, CO 80309-0526, USA
driscoll@na-net.ornl.gov

Section 16

Applications

In case of several authors, Invited Speakers are marked with a *.

The Minimum-Entropy Algorithm and Related Methods for Calibrating Asset-Pricing Models

Marco Avellaneda

Abstract. We describe an algorithm for calibrating asset-pricing models based on minimizing the relative entropy between probabilities. The algorithm determines a probability measure on path-space which minimizes the Kullback information with respect to a given prior and satisfies a finite number of moment constraints which correspond to fitting prices. It admits, generically, a unique, stable, solution that depends smoothly on the input prices. We study the sensitivities of the model values of contingent claims to variations in the input prices. We find that hedge ratios can be interpreted as "risk-neutral" regression coefficients of the contingent claim's payoff on the set of payoffs of the input instruments. We also show that the minimum-entropy algorithm is a special case of a general class of algorithms for calibrating asset-pricing models based on stochastic control and convex optimization. As an illustration, we use minimum-entropy to construct a smooth curve of instantaneous forward rates from US LIBOR data and to study the corresponding sensitivities of fixed-income securities to variations in input prices.

1 Introduction

Despite its practical importance, model calibration has received little attention in Mathematical Finance. Calibrating an asset-pricing model means specifying a probability distribution for the underlying state-variables in such a way that the model reproduces, by taking discounted expectations, the current market prices of a set of reference securities. The reference securities, or inputs, characterize the market under consideration. The most common models of this kind are yield-curve models, used for managing portfolios of fixed-income securities.[1] Other, less ubiquitous, examples are the so-called local volatility models used for managing option portfolios.[2]

[1] In this case, it is customary to vary the swap rates or bond yields corresponding to standard maturities by one basis point and to compute the corresponding dollar change in the portfolio value. These sensitivities are the so-called "DV01"s (dollar value of one basis point) used to quantify interest-rate exposure.

[2] Also known as "smile models".

In many cases of interest, the calibration problem is equivalent to a classical problem in statistics: the determination of a probability distribution from a finite set of moments. The "moments" correspond to the discounted expectations of the cash-flows of the reference instruments. It is well-known, however, that such problems are ill-posed: there can be many solutions or, sometimes, no solution at all. In financial-economic terms, this signifies that prices may not be consistent with *any* risk-neutral probability (and hence that an arbitrage exists) or, more likely, that there exist several risk-neutral probabilities consistent with the current prices due to market incompleteness. Selecting a probability is tantamount to "completing the market", in the sense that Arrow-Debreu prices are assigned to all future states. Thus, any calibration procedure involves making subjective choices. Taking into account available econometric information and stylized facts about the market reduces (partially) the ill-posedness of the model selection problem. Intuitively, a calibrated model which is "near" our prior beliefs and market knowledge is more desirable than one that is "far away" from the prior. [3]

In this paper, we study an algorithm which consists in choosing the risk-neutral probability that minimizes the *relative entropy*, or *Kullback-Leibler entropy* with respect to a subjective prior. This approach was pioneered in statistics by Jaynes (1996) and others; see McLaughlin (1984), Cover and Thomas (1991). An appealing feature of the method is that it takes into account the *a priori* (econometric) information available. This information is modeled by the prior probability, which can be viewed as a "first step" towards adjusting the model to econometric data but not necessarily to current prices. The entropy minimization algorithm provides a way of reconciling the prior with the information contained in current market prices.

Buchen and Kelly(1996) and Gulko(1995, 1996) used entropy minimization for calibrating one-period asset pricing models; see also Jackwerth and Rubinstein (1996) and Platen and Rebolledo (1996). In a previous article, Avellaneda, Friedman, Holmes and Samperi (1997) applied the minimum relative entropy method to the calibration of volatility surfaces in the context of commodity option pricing. In the present paper, following Buchen and Kelly and Avellaneda *et al*, we use Lagrange multipliers to model the price constraints. However, we go one step further in the analysis and study also the sensitivities of the model with respect to the input prices. For this purpose, we use the matrix of second derivatives with respect to the Lagrange multipliers computed at the critical point.

The paper is organized as follows: In Section 2, we consider a one-period model. Under mild no-degeneracy assumptions, we show that if there exists a probability with finite relative entropy, then the calibration problem has a unique solution. We establish also that the price-sensitivities of contingent claims depend smoothly on the input prices. The calibrated model has a remarkable property: the *deltas* (price-sensitivities) and the *betas* (regression coefficients of the cash-flow of a contingent claim on the space generated by the cash-flows of the input instruments)

[3] For example, practitioners tend to favor models in which interest rates are mean-reverting and oscillate about some asymptotic distribution. Processes that have unit roots and can reach very large values with large probabilities are discarded and appear to fail to pass simple statistical tests.

are, in fact, equal. More precisely, let Π denote the model price of a contingent claim which has a discounted payoff h. Let us denote by $G_i,\ i = 1, 2, ..., N$ the discounted cash-flows of the reference instruments, and by $C_1, ... C_N$ their prices. Then, we have

$$\frac{\partial \Pi}{\partial C_i} \ = \ \sum_{j=1}^{N} K_{ij} \ \mathrm{Cov}\{G_j, \, h\}$$

where

$$K \ = \ H^{-1} \ , \quad H_{ij} \ = \ \mathrm{Cov}\{G_i, \, G_j\}$$

and Cov represents the covariance operator under the risk-neutral (calibrated) measure. It is well-known that the right-hand side of the first equation corresponds to the value of the coefficient β_i in the linear regression model

$$h \ = \ \alpha \ + \ \sum_{i=1}^{N} \beta_i \, G_i \ + \ \epsilon$$

where ϵ has mean zero and is uncorrelated with the cash-flows $\{G_j\}$ under the risk-neutral measure. This property of the minimum-entropy algorithm suggests that it has econometric relevance. [4]

Sections 3 and 4 are devoted to inter-temporal asset-pricing models, where we formulate the algorithm in terms of partial differential equations. The algorithm involves solving a Hamilton-Jacobi-Bellman partial differential equation of "quasi-linear" type[5] and minimizing the value of the solution at one point in terms of a finite set of Lagrange multipliers. The gradient of the objective function corresponds to a coupled system of linearized equations.

In Section 5, we show that the algorithm can be formulated as a constrained stochastic control problem. This suggests that there are many generalizations of the "pure" entropy algorithm that can be made by changing the form of the cost function. Specifically, minimizing relative entropy is equivalent to minimizing the L_2 norm of the risk-premia $m_i(t)$, *i.e.*

$$\mathrm{E}^P \left\{ \int_0^{T_{max}} \sum_{i=1}^{\nu} \ m_i(t)^2 \, dt \right\}$$

where T_{max} is the time-horizon and ν is the number of factors. In practice, it is computationally advantageous to consider functionals of the form

$$\mathrm{E}^P \left\{ \int_0^{T_{max}} e^{-\int_0^t r(s)\, ds} \sum_{i=1}^{\nu} \ m_i(t)^2 \, dt \right\} \ ,$$

[4] Calibration via relative entropy minimization is, in a certain sense, the non-parametric counterpart of the maximum-likelihood estimation method; cf Jaynes (1996).

[5] This means that the nonlinearity appears in the gradient terms.

because this reduces the dimensionality of the computation, while preserving at the same time the essential features of the algorithm. [6]

In Section 6, we use the algorithm to construct smooth forward rate curves from US LIBOR data (FRAs and swap rates). We pay particular attention to the sensitivities with respect to input swap rates, an issue that remains somewhat controversial among practitioners. Hedges tend to be model-dependent and therefore a certain amount of risk is taken when choosing different forward rate curves. The issue is whether smooth curves, which give rise to "non-local" hedges[7], are preferable to discontinuous forward rate curves, such as the ones obtained by the bootstrapping method. The latter method tends to give rise to "local" hedges in which the sensitivities are essentially limited to the nearest maturities. It is our hope that the minimum-entropy method can compete favorably and perhaps even improve on some of the other methods used to generate smooth forward-rate curves, in the sense that the resulting sensitivities are acceptable from a practical viewpoint. These issues will be investigated in a separate paper.

2 Relative entropy minimization with moment constraints

We consider the problem of determining a probability density function $f(X)$ for a real-valued random variable X satisfying

$$\int G_i(X)\, f(X)\, dX \;=\; C_i\,, \quad 1 \;\leq\; i \;\leq\; N\,, \tag{1}$$

where $G_1(X)$, ...$G_N(X)$ are given functions and C_1, ...C_N are given numbers. [8] Financially, X represents a state-variable describing the economy; $G_i(X)$ and C_i represent, respectively, the cash-flows and prices of a set of traded securities.

Buchen and Kelly proposed, in the context of option pricing, to choose the density $f(X)$ that minimizes the functional

$$H(f|f_0) \;=\; \int f\, log\left(\frac{f}{f_0}\right)\, dX\,, \tag{2}$$

where $f_0(X)$ is a prior probability density function. The expression $H(f|f_0)$ is known as the Kullback-Leibler entropy or relative entropy of f with respect to f_0. It represents the "information distance" between $f(X)$ and $f_0(X)$. [9]

It is well-known (Cover and Thomas) that if there exists a probability density function f satisfying the constraints (1) and such that $H(f|f_0)$ is finite, the solution of the constrained entropy minimization problem exists and can be found by the method of Lagrange multipliers. Namely, we solve

[6] The advantage of passing from minimum-entropy to a more general control formulation was also shown in Avellaneda *et. al.*, where the technique was used to "regularize" the relative entropy of two mutually singular diffusions.

[7] By this we mean hedges that imply correlations between bonds with distant maturities.

[8] Henceforth, we say that a probability satisfying the constraints (1) is *calibrated*. It is implicitly assumed that the functions $G_i(X)$ are such that all integrals considered are well-defined.

[9] The relative entropy is not symmetric with respect to the variables f and f_0, so it is not a distance in the mathematical sense of the word. Nevertheless, it measures the "deviation" of f from f_0 (Cover and Thomas).

$$\inf_{\lambda_i} \sup_{f} \left[-H(f|f_0) \; + \; \sum_i \lambda_i \left(\int G_i \, f \, dX \; - \; C_i \right) \right] . \tag{3}$$

Let us first fix λ and seek the density that maximizes this "augmented Lagrangian". An elementary calculation of the first-order optimality conditions (Cover and Thomas) shows that for each λ, the optimal probability density function is given by

$$f_\lambda(X) \;\; = \;\; \frac{1}{Z(\lambda)} \, f_0(X) \, e^{\sum_i \lambda_i \, G_i(X)} \tag{4}$$

where $Z(\lambda)$ is the normalization factor

$$Z(\lambda) \;\; = \;\; \int f_0 \; e^{\sum_i \lambda_i \, G_i} \, dX.$$

Substituting expression (3a) into (2), it follows that the optimization over the Lagrange multipliers is equivalent to minimizing the function

$$\log \left(Z(\lambda) \right) \; - \; \sum_i \lambda_i \, C_i \; , \tag{5}$$

over all values of $\lambda \; = \; (\lambda_1, \, ... \, \lambda_N)$. The first-order conditions for a minimum are

$$\frac{1}{Z(\lambda)} \, \frac{\partial Z(\lambda)}{\partial \lambda_i} \; = \; C_i \; .$$

This shows, in view of (4), that if λ is a critical point of (5) then f_λ is calibrated.

The stability of the solution, i.e. the continuous dependence of f_λ on input prices, follows from convex duality. To see this, notice first that

$$\begin{aligned} \left(log \left(Z(\lambda) \right) \right)_{\lambda_i, \, \lambda_j} \;\; &= \;\; \frac{Z_{\lambda_i \; \lambda_j}}{Z} \; - \; \frac{Z_{\lambda_i} \, Z_{\lambda_j}}{Z^2} \\ &= \;\; \mathrm{Cov}_{f_\lambda} \left[G_i(X), \; G_j(X) \right] \; \equiv \; H_{ij} \; . \end{aligned}$$

Since covariance matrices are non-negative definite, $log(Z(\lambda)) \; - \; \lambda \cdot C$ is convex. It also follows from this characterization that $log(Z(\lambda))$ is strictly convex if the N payoff functions are linearly independent. [10]

Let λ^* be the value of the Lagrange multipliers that minimizes the objective function $log \left[Z(\lambda) \right] \; - \; \lambda \, C$. To assess the sensitivity of the calibrated probability

[10] As a rule, redundancies within the class of input securities should be avoided when fitting prices. They lead to instabilities, since the input prices must satisfy linear relation exactly (i.e. with infinite precision) in order to avoid mispricing these instruments with the model.

f_{λ^*} to input prices, consider a new contingent claim with payoff $h(X)$ (the "target payoff"). Let $\Pi(\lambda) = \mathrm{E}^{f_\lambda}(h(X))$. Then, we have

$$\begin{aligned}\frac{\partial \Pi(\lambda^*)}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j}\,\frac{\int f_0\, e^{\lambda\cdot G}\, h\, dX}{\int f_0\, e^{\lambda\cdot G}\, dX} \\ &= \mathrm{E}^{f_\lambda}(h(X)\, G_j(X)) \;-\; \mathrm{E}^{f_\lambda}(h(X))\;\mathrm{E}^{f_\lambda}(G_j(X)) \\ &= \mathrm{Cov}_{f_{\lambda^*}}(\,h(X),\; G_j(X))\;.\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial \Pi(\lambda^*)}{\partial C_i} &= \sum_j \left(\frac{\partial \Pi(\lambda)}{\partial \lambda_j}\right)_{\lambda\,=\,\lambda^*} \frac{\partial \lambda_j^*}{\partial C_i} \\ &= \sum_j \left(\frac{\partial \Pi(\lambda^*)}{\partial \lambda_j}\right)_{\lambda\,=\,\lambda^*} \left(H^{-1}\right)_{ij} \\ &= \sum_j \mathrm{Cov}_{f_{\lambda^*}}(\,h(X),\; G_j(X))\,\left(H^{-1}\right)_{ij}\;. \qquad (6)\end{aligned}$$

Here, in deriving the second equation, we made use of the well-known duality relations (Rockafellar (1970))

$$\frac{\partial C_i}{\partial \lambda_j^*} \;=\; H_{ij} \qquad , \qquad \frac{\partial \lambda_j^*}{\partial C_i} \;=\; \left(H^{-1}\right)_{ij}\;.$$

It follows from equations (4) and (6) that $\Pi = \Pi(C_1, \ldots C_N)$ is infinitely differentiable as a function of $C_1, \ldots C_N$. In particular the sensitivities $\frac{\partial \Pi}{\partial C_i}$ vary continuously with the input prices.

Formula (6) admits a simple interpretation. Consider the linear regression model

$$h(X) \;=\; \alpha \;+\; \sum_{i=1}^{N} \beta_i\, G_i(X) \;+\; \epsilon\;,$$

where we assume that ϵ is a random variable with mean zero uncorrelated with $G_i(X)$ $i = 1, .. N$ under the the risk-neutral measure. The coefficients β_i which minimize the variance of the residual $h - \alpha - \sum_i \beta_i G_i$ are given by

$$\beta_i \;=\; \sum_j \left(H^{-1}\right)_{ij} \mathrm{Cov}_{f_{\lambda^*}}(h(X), G_j(X)) \;=\; \frac{\partial \Pi}{\partial C_i}\;, \quad 1 \;\le\; i \;\le\; N\;.$$

We summarize the results of this section in

PROPOSITION 1. *(a) The minimum-relative-entropy method reduces the class of candidate solutions of the moment problem to an N-parameter exponential family $f_\lambda(X)$ given by (4).*

Assume that the input payoffs $G_1(X), \ldots G_N(X)$ are linearly independent. Then:

(b) If there exists a calibrated density $f(X)$ such that $H(f|f_0) < \infty$, the solution of the constrained entropy-minimization problem is unique.

(c) The sensitivities of contingent-claim prices to variations in input prices are equal to the linear regression coefficients of the target payoff on the input payoffs under the calibrated measure.

3 INTER-TEMPORAL MODELS

We consider a classical continuous-time economy, represented by a state-vector $\mathbf{X}(t) = (X_1(t), \ldots, X_\nu(t))$ which follows a diffusion process under the prior probability measure:

$$dX_i(t) = \sum_{j=1}^{\nu} \sigma_{ij}^{(0)}\, dZ_j(t) + \mu_i^{(0)}\, dt \qquad 1 \leq i \leq \nu\,. \tag{7}$$

Here $(Z_1, \ldots Z_\nu)$ are independent Brownian motions and $\sigma_{ij}^{(0)}$ and $\mu^{(0)}$ are functions of X and t.

We assume that there are N traded securities, with prices $C_1, \ldots C_N$. Our goal is to find a risk-neutral probability measure P consistent with these prices based on the principle of minimum relative entropy with respect to the prior (denoted by P_0).

The price constraints can be written in the form on N equations

$$C_i = \mathrm{E}^P \left\{ \sum_{k=1}^{n_i} e^{-\int^{T_{ik}} r(s)\, ds}\, G_{ik}(\mathrm{X}(T_{ik})) \right\}, \qquad 1 \leq i \leq N\,, \tag{8}$$

where $\{T_{ik}\}_{k=1}^{n_i}$ are the cash-flow dates of the i^{th} security and $\{G_{ik}(\mathrm{X})\}_{j=1}^{n_i}$ represent the corresponding cash-flows. We assume that the latter are bounded, continuous functions of X. The process $r(s) = r(\mathrm{X}(s), s)$ represents the short-term (continuously compounded) interest rate. Notice that in (8) the expectation value is taken with respect to a calibrated (risk-neutral) measure P which, in general, is not equal to P_0.

We follow the approach of the previous section. First, we consider the Kullback-Leibler relative entropy of P with respect to P_0 in the diffusion setting. For this purpose, it is convenient to define a finite time horizon $0 < t < T_{max}$, (where $T_{max} \geq \max_{ik} T_{ik}$). The relative entropy of P with respect to P_0 is given by

$$H(P|P_0) = \mathrm{E}^P \left\{ \log \left(\frac{dP}{dP_0} \right)_{T_{max}} \right\},$$

where $\left(\frac{dP}{dP_0}\right)_{T_{max}}$ is the Radon-Nykodym derivative of P with respect to P_0 over the time-horizon T_{max} . [11]

Next, we consider the augmented Lagrangian associated with the constraints (8) (compare with (3))

$$-\mathrm{E}^P\left\{\log\left(\frac{dP}{dP_0}\right)_{T_{max}}\right\}+\sum_{i=1}^{N}\lambda_i\left(\sum_j \mathrm{E}^P\left\{e^{-\int_0^{T_{ij}} r(s)ds} G_{ij}(\mathrm{X}(T_{ij}))\right\} - C_i\right). \quad (9)$$

The solution of the inf-sup problem is identical to the one outlined in the previous section. Accordingly, we define the normalization factor (cf. (4))

$$Z(\lambda) \;=\; \mathrm{E}^{P_0}\left\{\exp\left(\sum_{ij}\lambda_i\, e^{-\int_0^{T_{ij}} r(s)ds} G_{ij}(\mathrm{X}(T_{ij}))\right)\right\} . \quad (10)$$

Further, by mimicking equation (4), we obtain a parametric family of measures $\{P_\lambda\}_\lambda$ defined by their Radon-Nykodym derivatives with respect P_0:

$$\frac{dP_\lambda}{dP_0} \;=\; \frac{1}{Z(\lambda)}\cdot\exp\left(\sum_{ij}\lambda_i\, e^{-\int_0^{T_{ij}} r(s)ds} G_{ij}(\mathrm{X}(T_{ij}))\right) . \quad (11)$$

Elementary calculus of variations shows that for any fixed vector $\lambda = (\lambda_1, \lambda_2, ...\lambda_N)$, the measure P_λ realizes the supremum of the Lagrangian (9) over all probability measures. As expected, the supremum is given by

$$log[\,Z(\lambda)\,] \;-\; \sum_{i=1}^{N}\lambda_i\, C_i \;.$$

If λ is a critical point, we have

$$C_i \;=\; \frac{Z_{\lambda_i}}{Z} \;=\; \mathrm{E}^{P_\lambda}\left\{\sum_{j=1}^{n_i} e^{-\int_0^{T_{ij}} r(s)ds} G_{ij}(\mathrm{X}(T_{ij}))\right\} \qquad 1 \le i \le N \;.$$

Therefore, the corresponding measure P_λ is calibrated to the input prices.

[11] In particular, the relative entropy is infinite if P is not absolutely continuous with respect to P_0.

Define the discounted cash-flows

$$\Gamma_i \;=\; \sum_{j=1}^{n_i} e^{-\int_0^{T_{ij}} r(s)ds} G_{ij}(\mathrm{X}(T_{ij}))\;,\quad 1 \;\leq\; i \;\leq\; N\;.$$

As in the previous section, we can interpret the Hessian of $log(Z(\lambda)) - \lambda\, C$ as a covariance matrix, *viz.*,

$$\frac{\partial^2}{\partial\lambda_i\;\partial\lambda_j}\;(log(Z(\lambda)) - \lambda\,) \;=\; \mathrm{Cov}^P\,[\Gamma_i,\;\Gamma_j]\;.$$

Similarly, if $h(X_T)$ is the payoff of a security maturing at time $T \;\leq\; T_{max}$, we have

$$\frac{\partial}{\partial\lambda_j}\mathrm{E}^P\left\{e^{-\int_0^T r_s\,ds} h(X_T)\right\} \;=\; \mathrm{Cov}^P\left[\Gamma_j,\; e^{-\int_0^T r_s\,ds} h(X_T)\right]\;.$$

Like in the previous section, we conclude that

PROPOSITION 2. *(a) Relative entropy minimization is equivalent assuming that the probability measure belongs to an N-parameter exponential family given by (11).*

(b) If the input payoffs are linearly independent, there is at most one calibrated measure that minimizes relative entropy.

(c) The model prices and sensitivities of contingent claims depend continuously on input prices.

(d) The sensitivities with respect to input prices can be interpreted as the linear regression coefficients of the target discounted cash-flows on the space generated by the discounted cash-flows of the input instruments.

4 PDE FORMULATION

Let $\mathrm{L}^{(0)}$ represent the infinitesimal generator of the semi-group corresponding to the prior P_0 *i.e.*,[12]

$$\mathrm{L}^{(0)}\,\phi \;=\; \frac{1}{2}\sum_{ij} a_{ij}\,\phi_{X_i\,X_j} \;+\; \sum_i \mu_i^{(0)}\,\phi_{X_i} \tag{12}$$

where

$$a_{ij} \;=\; \sum_{p=1}^{\nu} \sigma_{ip}^{(0)}\,\sigma_{jp}^{(0)}\;.$$

[12]We use the notation $\phi_{X_i} \;=\; \frac{\partial\phi}{\partial X_i}$ for partial derivatives.

It follows from (10) and standard diffusion theory that the normalization factor is given by

$$Z(\lambda) \;=\; U(\mathrm{X}(0),\, 1,\, 0\,;\lambda)\;, \tag{13}$$

where $U(\mathrm{X},\, Y,\, t\,;\lambda)$ is the solution of the Cauchy problem

$$U_t \;+\; \mathrm{L}^{(0)}\, U \;-\; r\, Y\, U_Y \;=\; 0 \qquad t \;\neq\; T_{ij} \tag{14}$$

with the boundary conditions at cash-flow dates $t = T_{ij}$

$$U(\mathrm{X},\, Y,\, T_{ij} - 0\,;\lambda) \;=\; U(\mathrm{X},\, Y,\, T_{ij} + 0\,;\lambda)\cdot \exp\left(\sum_i \lambda_i\, G_{ij}(\mathrm{X})\, Y\right)\;. \tag{15}$$

To derive (14), we introduced the auxiliary state-variable $Y_t \;=\; e^{-\int_0^t r(s)\, ds}$ and the $\nu+1$-dimensional process $(\mathrm{X}_t,\, Y_t)$ which is a Markov process with an infinitesimal generator given by the left-hand side of (14).

From (14) we can derive partial differential equations satisfied by $log(Z(\lambda))$ and its gradient with respect to λ. Accordingly, we obtain

$$log(Z(\lambda)) \;=\; W(\mathrm{X}(0),\, 1,\, 0\,;\lambda)\;, \qquad \frac{Z_{\lambda_i}}{Z} \;=\; V^{(i)}(\mathrm{X}(0),\, 1,\, 0\,;\lambda)$$

where W satisfies the PDE

$$W_t + \mathrm{L}^{(0)} W + \frac{1}{2}\sum_{ij=1}^{N} a_{ij}\, W_{X_i}\, W_{X_j} \;-\; r\, Y\, W_Y \;=\sum_{ij} \lambda_i\, G_{ij}(\mathrm{X})\, Y\, \delta(t - T_{ij})\;. \tag{16}$$

The PDE for $V^{(l)}$ is obtained by differentiating (16) with respect to λ_l, *viz.*

$$V_t^{(l)} \;+\; \mathrm{L}^{(0)} V^{(l)} \;+\; \sum_{ij=1}^{N} a_{ij}\, W_{X_i}\, V^{(l)}_{X_j} \;-\; r\, Y\, V^{(l)}_Y \;=$$

$$\sum_{j=i}^{n_i} G_{ij}(\mathrm{X})\, Y\, \delta(t - T_{ij})\;. \tag{17}$$

From this last equation, we deduce the following characterization of the calibrated measure.

PROPOSITION 3. *The calibrated measure which minimizes the relative entropy corresponds to the diffusion process*

$$dX_i \;=\; \sum_{j=1}^{\nu} \sigma_{ij}^{(0)}\; dZ_j \;+\; \left(\;\mu_i^{(0)} \;+\; \sum_{j=1}^{\nu} \sigma_{ij}^{(0)}\, m_j\;\right)\; dt$$

with

$$m_i \;=\; \sum_{i=1}^{N} \sigma_{ij}^{(0)}\, W_{X_j}\ , \tag{18}$$

where W is computed with λ at the critical point.

5 MODIFIED ENTROPIES AND THE OPTIMAL CONTROL FORMULATION.

It is useful to view the entropy minimization algorithm as a stochastic optimal control problem with constraints. We recall the following result (Platen and Rebolledo): PROPOSITION 4. *The class of diffusion measures P which have finite relative entropy with respect to P_0 consists of Ito processes*

$$dX_i(t) \;=\; \sum_{j=1}^{\nu} \sigma_{ij}^{(0)}\, dZ_j(t) \;+\; \mu_i\, dt$$

with

$$\mu_i \;=\; \mu_i^{(0)} \;+\; \sum_{j} \sigma_{ij}^{(0)}\, m_j\ ,$$

where, $m_j \quad 1 \le j \le \nu$ are square-integrable. Moreover, the relative entropy of P with respect to P_0 (viewed as measures in path-space with the time horizon $0 < t < T_{max} = \max_{ik} T_{ik}$) is given by

$$H(P|P_0) \;=\; \frac{1}{2}\,\mathrm{E}^P \left\{ \int_0^{T_{max}} \sum_{j=1}^{\nu} m_j(t)^2\, dt \right\}\ . \tag{19}$$

Thus, minimizing the KL entropy is equivalent to selecting the risk-neutral measure in such a way that the vector of risk-premia has the smallest mean-square norm (cf. Platen and Rebolledo (1996), Samperi(1997)).

Using (19) we rewrite the augmented Lagrangian (9) as

$$-\mathrm{E}^P \left[\int_0^{T_{max}} \sum_{j=1}^{\nu} m_j^2(t)\, dt \right] \;+\; \sum_{i=1}^{N} \lambda_i\, \mathrm{E}^P \left[\sum_{j=1}^{n_i} e^{-\int_0^{T_{ij}} r(s)ds}\, G_{ij}(\mathrm{X}(T_{ij})) \right] \tag{20}$$

The advantage of the stochastic control formulation is that it can be generalized considerably. In fact, we can replace the function $\sum_j m_j^2(t)$ by more general functions of the form $\eta(_t,\, m_1(t),\, m_2(t),\, ...m_\nu(t))$, which are strictly convex in $m_i(t)$.

The class of functionals of the form

$$H_{mod}(P|P_0) \;=\; \frac{1}{2}\,\mathrm{E}\left\{ \int_0^{T_{max}} e^{-\int_0^t r(s)\,ds}\, \eta(m(t))\;\, dt \right\}\,, \tag{21}$$

where $\eta(m)$ is a deterministic and strictly convex is of particular importance. In this case, $H_{mod}(P|P_0)$ can be seen as a "running cost" with respect to the choice of parameters which penalizes deviations from the prior.

Notice that the definition of entropy in (19) is independent of the interest rate. One important advantage of discounting the local entropy by the interest rate is *dimension reduction*: we can dispense of the auxiliary state variable Y. In fact, the HJB equation corresponding to the modified entropy (21) is

$$W_t \;+\; \mathrm{L}^{(0)}W \;+\; \eta^*\left(\sigma^{(0)}\cdot W_X\right) \;-\; r\,W \;=$$

$$\sum_{ij} \lambda_i\, G_{ij}(\mathrm{X})\,\delta(t-T_{ij})\;, \tag{22}$$

where η^* is the Legendre transform of η (Rockafellar). The function W plays the role of $log(Z(\lambda)$ in the "pure entropy" framework. Note, however, that in the special case $\eta(t,\,m) \;=\; \frac{1}{2}\sum\limits_j m_j^2$ we have $\eta = \eta^*$. The corresponding Bellman equation is

$$W_t \;+\; \mathrm{L}^{(0)}W \;+\; \frac{1}{2}\sum a_{ij}\,W_{X_i}\,W_{X_j} \;-\; r\,W \;=$$

$$\sum_{ij} \lambda_i\, G_{ij}(\mathrm{X})\,\delta(t-T_{ij})\;, \tag{23}$$

In the rest of this section we assume this particular form for the modified entropy. Following the steps outlined in §2, the algorithm consists of minimizing

$$W\,(\mathrm{X}(0),\,0\,;\;\lambda_1,\;\ldots\,\lambda_N) - \sum_{i=0}^{M} \lambda_i\, C_i\;,$$

over λ. This is done with a gradient-based optimization algorithm such as L-BFGS (Zhu, Boyd, Lu and Nocedal (1994)). The gradient is computed by solving the N linearized equations:

$$V_t^{(l)} \;+\; \mathrm{L}^{(0)}\,V^{(l)} \;+\; \sum_{ij} a_{ij}\,W_{X_i}\,V_{X_j}^{(l)} \;-\; r\,V^{(l)} \;=$$

$$\sum_{j=0}^{n_l} G_{lj}(\mathrm{X})\,\delta(t-T_{lj}) \tag{24}$$

Notice that the first-order conditions for the minimum in λ are

$$V^{(l)}\,(\mathrm{X}(0),\,0\,;\;\lambda_1,\;\ldots\,\lambda_N)\;-\;\lambda_i\,C_l\;=\;0\,,\qquad 1\;\leq\;l\;\leq\;N\;\;.$$

Formally, these equations imply that the corresponding probability measure is calibrated, since

$$V^{(l)}\,(\mathrm{X}(0),\,0\,;\;\lambda_1,\;\ldots\,\lambda_N)\;=\;\mathrm{E}^P\left\{\sum_{k=1}^{n_l}\;e^{-\int^{T_{lk}}r(s)\,ds}\,G_{lk}(\mathrm{X}(T_{lk})\right\}\;.$$

Here P is the diffusion process with drift $\mu^{(0)}+W_X{\cdot}\sigma^{(0)}$, where W is calculated at the optimal values of the Lagrange multipliers. We refer to the diffusion measure implied by solving equation (23) as P_λ, a slight abuse of notation. The optimal control formulation has the same mathematical structure (i.e. convexity λ) as the "pure" entropy problem. To study the dependence on the inputs, we consider the Hessian of $W(\lambda)$. Differentiating equations (24) with respect to λ, we find that the Hessian matrix

$$H^{(lm)}\;=\;\frac{\partial^2 W}{\partial\lambda_l\partial\lambda_m}$$

satisfies

$$H_t^{(lm)}\,+\,L\,H^{(lm)}\,+\sum_{ij}\,a_{ij}\,W_{X_i}\,H_{X_j}^{(lm)}\,+\sum_{ij}\,a_{ij}\,V_{X_i}^{(l)}\,V_{X_j}^{(m)}\;-\;r\,H^{(lm)}\;=0\,.\quad(25)$$

In particular, we have

$$H^{(lm)}(\mathrm{X}(0),\,0;\,\lambda^*)\;=\;\mathrm{E}^P\left\{\int_0^{T_{max}}\;e^{\;-\int_0^t r(s)\,ds}\;\sum_{i\,j\,=1}^{M}\,a_{ij}\,V_{X_i}^{(l)}\,V_{X_j}^{(m)}dt\right\}\;.\qquad(26)$$

Unlike the case of "pure" entropy, the Hessian does not admit a simple interpretation in terms of linear regression coefficients. Nevertheless, we can express the difference between the Hessian and the covariance matrix of the discounted input cash-flows as an expectation. More precisely, we have

$$\mathrm{Cov}^{P_\lambda}\,\left(\Gamma^{(l)},\;\Gamma^{(m)}\,\right)\;=\;\mathrm{E}^{P_\lambda}\left\{\int_0^{T_{max}}\;e^{\;-2\int_0^t r(s)\,ds}\;\sum_{ij=1}^{\nu}\,a_{ij}\,V_{X_i}^{(l)}\,V_{X_j}^{(m)}\;dt\right\}\,,\quad(27)$$

which differs from (26) in the fact that the stochastic discount factor is squared. Therefore, we conclude that

$$H^{(lm)} \;=\; \mathrm{Cov}^{P_\lambda}\left(\Gamma^{(l)},\, \Gamma^{(m)}\right) \;+$$

$$\mathrm{E}^{P_\lambda}\left\{ \int_0^{T_{max}} \left(e^{-\int_0^t r(s)\,ds} \;-\; (e^{-2\int_0^t r(s)\,ds} \right) \sum_{ij=1}^{\nu} a_{ij}\, V_{X_i}^{(l)}\, V_{X_j}^{(m)}\; dt \right\} . \qquad (28)$$

In particular, this shows that if the instruments are not linearly dependent with P_0-probability 1, the Hessian matrix is positive definite.[13] [14] Barring trivial redundancies, the argument establishes that there is at most one λ that minimizes the objective function.

Finally, we analyze the sensitivities of model prices to input prices.

Given a contingent claim with a payoff $h(X_T)$ due date T, $(T < T_{max})$, let Π and $\Pi^{(l)}$ denote, respectively, the model price and the sensitivity of this price with respect to λ_l.

The functions Π and $\Pi^{(l)}$ are readily computed by solving the system of equations

$$\Pi_t \;+\; \mathrm{L}^{(0)}\,\Pi \;+\; \sum_{ij} a_{ij}\, W_{X_i}\, \Pi_{X_j} \;-\; r\,\Pi \;= \delta(t-T)\, h(X)\;, \qquad (29)$$

and

$$\Pi_t^{(l)} \;+\; \mathrm{L}^{(0)}\,\Pi^{(l)} \;+\; \sum_{ij} a_{ij}\, W_{X_i}\, \Pi_{X_j}^{(l)} \;+\; \sum_{ij} a_{ij}\, \Pi_{X_i}\, V_{X_j}^{(l)} \;-\; r\,\Pi^{(l)} \;= 0\;. \quad (30)$$

It follows from this that the $\Pi^{(l)} \;=\; \Pi^{(l)}(\mathrm{X}(0),\, 0)$ satisfies

$$\Pi^{(l)} \;=\; \mathrm{E}^{P_\lambda}\left\{ \int_0^{T_{max}} e^{-\int_0^t r(s)\,ds} \sum_{ij=1}^{\nu} a_{ij}\, V_{X_i}^{(l)}\, \Pi_{X_j}\; dt \right\}$$

$$\;=\; \mathrm{Cov}^{P_\lambda}\left[e^{-\int_0^T r(s)\,ds}\, h(\mathrm{X}_T),\;\; \Gamma^{(l)} \right] \;+$$

[13] This property also follows directly from equation (25). The strict positivity of the Hessian holds for any strictly convex modified entropy function $\eta(\mathrm{M},\, t)$, provided that the inputs are not linearly dependent.

[14] For example, the following set of inputs is linearly dependent, or redundant: (i) a one-year swap resetting quarterly, and (ii) four 3-month forward-rate agreements starting at the swap reset dates. This constitutes a redundancy because the swap can be replicated exactly with the FRAs.

$$\mathrm{E}^{P_\lambda}\left\{\int_0^{T_{max}}\left(e^{-\int_0^t r(s)\,ds} - e^{-2\int_0^t r(s)\,ds}\right)\sum_{ij=1}^{\nu} a_{ij}\,V_{X_i}^{(l)}\,\Pi_{X_j}\,dt\right\}. \qquad (31)$$

As in §2, we can compute the sensitivities of Π with respect to the input prices C_1, ..., C_N using the inverse Hessian and the sensitivities with respect to λ. Accordingly, we have

$$\begin{aligned}\frac{\partial\Pi}{\partial C_m} &= \sum_{l=1}^{N}\frac{\partial\Pi}{\partial\lambda_l}\frac{\partial\lambda_l}{\partial C_m}\\ &= \sum_{l=1}^{N}\Pi^{(l)}\,\left(H^{-1}\right)_{lm} \qquad (32)\end{aligned}$$

where H^{-1} is the inverse of H.

6 Forward-rate modeling and hedging portfolios of interest rate swaps

To illustrate the minimum-entropy algorithm, we calibrate a one-factor interest-rate model to the prices of standard instruments in the US LIBOR market.

We consider a set of input instruments consisting of forward-rate agreements (FRAs) and swaps with standard maturities. Using the algorithm, we compute a probability measure on the process driving the short-term rate which has the property that all the input instruments are priced correctly by the model by discounting cash-flows. Since we do not use options to calibrate the model, we view the algorithm as a way of generating a *curve of instantaneous forward rates* from the discrete dataset. In other words, we are primarily concerned with the modeling of "straight" debt instruments and not the study of the volatility of the forward rate curve. The curve is generated by the formula

$$\begin{aligned}f(T) &= -\frac{\partial}{\partial T}\log P(T)\\ &= -\frac{\partial}{\partial T}\log \mathrm{E}^{P}\left\{e^{-\int_0^T r_t\,dt}\right\}.\end{aligned}$$

where $f(T)$ and $P(T)$ represent the instantaneous forward rate and the discount factor (present value of a dollar) associated with the maturity date T. The instantaneous forward-rate curve allows us to price arbitrary fixed-income securities without optionality. Hedge-ratios for different instruments are derived from the sensitivities of the curve to input prices.

We consider a prior distribution for the short-term interest rate

$$\frac{dr_t}{r_t} \;=\; \sigma\, dZ_t \;+\; \mu_t^{(0)}\, dt\;, \tag{33}$$

where σ is constant and $\mu_t^{(0)}$ is given. For simplicity, we take $\mu_t^{(0)} \;\equiv\; 0$ under the prior, which, as we shall see, corresponds essentially to a prior belief of a flat forward-rate curve. [15]

Given the considerations of the previous sections, the family of candidate probability measures for the short rates has the form (33) where $\mu^{(0)}$ is replaced by an unknown drift μ_t.

The modified entropy functional (21) with $\eta \;=\; \frac{1}{2}m^2$ is

$$\begin{aligned} H_{mod}(P\,|\,P_0) \;&=\; \frac{1}{2\sigma^2}\,\mathrm{E}\left\{ \int_0^{T_{max}} e^{-\int_0^t r_s\,ds} \left(\mu_t \;-\; \mu_t^{(0)}\right)^2\, dt \right\} \\ &=\; \frac{1}{2\,\sigma^2}\,\mathrm{E}\left\{ \int_0^{T_{max}} e^{-\int_0^t r_s\,ds} \mu_t^2\, dt \right\}\;. \end{aligned} \tag{34}$$

We calibrated this model to a data-set extracted from the US LIBOR market in late November 1997, consisting of FRAs and swap rates; cf. Table 1. The futures data corresponds to a series of 3-month Eurodollar contracts from January 1998 to December 2002. Forward-rates were computed from futures prices using an empirical convexity adjustment, which is displayed on the left of the futures price. Swap rates were computed from Treasury bond yields adding the corresponding credit spread, also displayed on the right of the yield. [16] Accordingly, the 3-month forward rate four months from today is computed as follows:

$$\begin{aligned} \text{forward rate} \;&=\; \text{futures-implied rate} \;-\; \text{conv. adjustment} \\ &=\; (100 - 94.20) \;-\; 0.12 \\ &=\; 5.68\ \% \end{aligned}$$

The 6-year swap rate was taken to be

$$\begin{aligned} \text{swap rate} \;&=\; \text{Treasury yield} \;+\; \text{spread} \\ &=\; 5.8150 \;+\; 0.3975 \\ &=\; 6.2125\ \% \end{aligned}$$

[15] Of course, we could have chosen any other drift for prior probability on short rates– this constitutes the "subjective" portion of the method. The significance of different priors will be clarified below.

[16] We shall not be concerned here about how convexity adjustments were generated or about the computation of the spread between swaps and Treasurys.

Table 1: Data for US LIBOR Market

ED futures / FRAs			Bonds / Swaps		
04m	94.20	0.0012	06y	5.8150	0.3975
10m	94.14	0.0023	07y	5.8236	0.4150
13m	94.08	0.0030	10y	5.8470	0.4475
16m	93.98	0.0044	12y	5.8683	0.4700
19m	93.98	0.0092	15y	5.9002	0.4800
22m	93.94	0.0131	20y	5.9535	0.4750
25m	93.91	0.0176	30y	6.0600	0.3750
28m	93.85	0.0234			
31m	93.87	0.0232			
34m	93.85	0.0371			
37m	93.83	0.0447			
40m	93.77	0.0522			
43m	93.79	0.0637			
46m	93.77	0.0730			
49m	93.75	0.0830			

In implementing the calibration algorithm for these instruments, we assumed that the discounted cash-flows of the FRAs per dollar notional are given by

$$\Gamma_f \;=\; e^{-\int_0^T r_t\,dt} \;-\; e^{-\int_0^{T+0.25} r_t\,dt} \left(1 \;+\; \frac{FRA\times\,0.25}{100}\right)$$

where FRA is the 3-month forward rate (expressed in percentages) corresponding to the maturity T. The discounted cash-flows of a semi-annual vanilla interest swap with N cash-flow dates is

$$\Gamma_s \;=\; 1 \;-\; \sum_{n=1}^{N} e^{-\int_0^{0.5\,n} r_t\,dt} \left(\frac{SWAP\times\,0.5}{100}\right) \;-\; e^{-\int_0^{0.5\,N} r_t\,dt}\;,$$

where $SWAP$ is the swap rate and where we assumed that the floating leg of the swap is valued at par.

In both cases (FRAs, swaps) we assumed that, under the risk-neutral probability, we have

$$\mathrm{E}^P\{\,\Gamma_i\,\} \;=\; 0 \quad i \;=\; f, s\;.$$

These equations represent the constraints for calibration in this context. We have therefore 22 constraints: 15 for the FRAs and 7 for the swaps. The entropy-minimization was implemented by solving the partial differential equations (23), (24), (25) using a finite-difference scheme (trinomial lattice) and using L-BFGS to find

the minimum of the augmented Lagrangian. We assumed a discretization of 12 periods per year.

Figure 1 shows the corresponding forward rate curve which derives from the data. We assumed a value of $\sigma = .10$ in this calculation. We noticed that the sensitivity of the curve to σ is negligible for $\sigma \leq 10\%$. The hedging properties of the model can be quantified by analyzing the sensitivities of the prices of par bonds with N years to maturity, for $N = 1, 2, 3 \ldots 30$. These results are exhibited in the bar graphs diplayed hereafter. Each chart considers a par swap with a give maturity. The bars on the graph represent the sensitivity of the price of the instrument with respect to the prices of the input securities. Notice, in particular that the maturities that correspond to an input security consist of a single column. Intermediate maturities (not represented in the input instruments) give rise to multiple bars that decay as we move away from the corresponding maturity.

Finally, we point out that the volatility parameter σ in this model has an interesting interpretation. Heuristically speaking, the construction of the forward rate curve can be viewed as a problem in interpolation from a discrete set of data. Since the problem is ill-posed, various regularizations have been proposed at the level of forward-curve building, without having recourse to an underlying probability model. These regularizations typically penalize oscillations in the curve by means of penalization functions of the form

$$\int_0^{T_{max}} \eta(f(t),\, f'(t),\, f''(t),\, t)\, dt$$

that are typically minimized subject to the constraints and to a choice of function space for $f(t)$.

It is easy to see that, in the limit $\sigma \ll 1$, the minimum-entropy calibration algorithm is associated with a special choice of the above functional, namely,

$$\int_0^{T_{max}} e^{-\int_0^t f(s)\, ds} \left(\frac{f'(t)}{f(t)} - \mu^{(0)} \right)^2 dt\,. \tag{35}$$

This can be seen from the results of Section 5 and by letting σ formally tend to zero in the entropy functional

$$\mathrm{E}^P \left\{ \int_0^{T_{max}} m^2(t)\, dt \right\} = \frac{1}{\sigma^2}\, \mathrm{E}^P \left\{ \int_0^{T_{max}} \left(\mu(t) - \mu^{(0)}(t) \right)^2 dt \right\}\,.$$

This result corresponds mathematically to the relation between the "viscous" solution of the penalized problem associated with (35) and the stochastic control problem discussed in Section 5. Form a numerical point of view, we can therefore view the minimum-entropy algorithm as an "articifial viscosity" method for minimizing the functional (35) subject to the price constraints.

References

Avellaneda, M., Friedman, C., Holmes, R. and Samperi, D. (1997), Calibrating volatility surfaces via relative entropy minimization, *Applied Mathematical Finance*, March issue

Buchen, P. W. and Kelly, M.(1996), The maximum entropy distribution of an asset inferred from option prices, *Journal of Financial and Quantitative Analysis*, 31 (1).

Cover, J. and Thomas, J. A. (1991), *Elements of Information Theory*, New York: John Wiley & Sons

Gulko, L. (1995), The Entropy Theory of Option Pricing , Yale University Working Paper; (1996) The entropy theory of bond pricing, *ibid.*

Jackwerth, J. C. and Rubinstein, M. (1996), Recovering probability distributions from contemporaneous security prices, *Journal of Finance*, 69(3), 771-818

Jaynes, E.T., (1996), Probability theory: the logic of science, Unpublished manuscript, Washington University.

McLaughlin, D. W. (1984), *Inverse problems*, Providence RI: SIAM and AMS, Volume 14

Platen, E., Rebolledo, R. (1996), Principles for modelling financial markets, Advances in Applied Probabilitys 33(3) 601-613

Rockafellar, R.T. (1970), *Convex Analysis*, Princeton, NJ: Princeton University Press.

Samperi, D.(1997), *Inverse problems and entropy in Finance*, Ph.D. Thesis, New York University

Zhu, C., Boyd, R.H., Lu, P. and Nocedal, J. (1994),*L-BFGS-B: FORTRAN Subroutines for Large-Scale Bound-Constrained Optimization*, Northwestern University, Department of Electrical Engineering

Marco Avellaneda
Courant Institute
of Mathematical Sciences
New York University,
251 Mercer Street
New York, NY, 10012, USA

The Tree of Life and Other Affine Buildings

Andreas Dress, Werner Terhalle

In this note, we discuss some mathematics which has proven to be of use in the analysis of molecular evolution – and, actually, was discovered in this context (cf. [D]).

According to evolutionary theory, the spectrum of present-day species (or biomolecules) arose from their common ancestors according to a well-defined scheme of bi-(or multi-)furcation steps. The task of phylogenetic analysis as defined by E. Haeckel is to unravel that scheme by comparing systematically all data available regarding present and extinct species. This task has been simplified enormously in recent years through the availability of molecular sequence data, first used for that purpose by W. Fitch and E. Margoliash in their landmark paper from 1967 dealing with Cytochrome C sequences [FM]. The basic idea in that field is that species (or molecules) which appear to be closely related should have diverged more recently than species which appear to be less closely related.

A standard formalization is to measure relatedness by a metric defined on the set of species (or molecules) in question. The task then is to construct an ($\mathbb{R}$-)tree which represents the metric (and hence the bifurcation scheme) as closely as possible. Below, we discuss necessary and sufficient conditions for the existence of such a tree that represents the metric *exactly*, as well as some constructions which lead to that tree if those conditions are fulfilled, and to more or also less treelike structures if not. Remarkably, the theory we developed in this context allowed also to view affine buildings (which in the rank 1 case are $\mathbb{R}$-trees) from a new perspective.

Here are some basic definitions and results:

Definition 1: *Given a non-empty set E, an integer $m \geq 2$, and a map*

$$v : E^m \to \{-\infty\} \cup \mathbb{R},$$

the pair (E, v) is called a valuated matroid of rank m *if the following properties hold:*

(VM0) for every $e \in E$, there exist some $e_2, \ldots, e_m \in E$ such that

$$v(e, e_2, \ldots, e_m) \neq -\infty,$$

(VM1) v is totally symmetric,

(VM2) for $e_1, \ldots, e_m \in E$ with $\#\{e_1, \ldots, e_m\} < m$, one has

$$v(e_1, \ldots, e_m) = -\infty,$$

(VM3) for all $e_1, \ldots, e_m, f_1, \ldots, f_m \in E$, *one has*

$$\begin{aligned} & v(e_1, \ldots, e_m) + v(f_1, \ldots, f_m) \\ \leq \quad & \max_{1 \leq i \leq m} \{v(f_1, e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_m) + v(e_i, f_2, \ldots, f_m)\}. \end{aligned}$$

Condition (VM3) is also called the VALUATED EXCHANGE PROPERTY.

If $\{b_1, \ldots, b_m\} \subseteq E$ *satisfies* $v(b_1, \ldots, b_m) \neq -\infty$, *then* $\{b_1, \ldots, b_m\}$ *is called a* BASE *of the valuated matroid* (E, v).

Note that (VM3) implies the bases exchange property of ordinary matroids for the set $B_{(E,v)}$ of bases of (E, v).

Here is a "generic" example:
Let K be a field with a non-archimedean valuation $w : K \to \{-\infty\} \cup \mathbb{R}$, that is a map satisfying the conditions

$$w(x) = \infty \iff x = 0,$$

$$w(x \cdot y) = w(x) + w(y),$$

and

$$w(x + y) \leq \max\{w(x), w(y)\}$$

for all $x, y \in K$; then – in view of the GRASSMANN-PLÜCKER identity

$$\begin{aligned} & \det(e_1, \ldots, e_m) \cdot \det(f_1, \ldots, f_m) \\ = \quad & \sum_{i=1}^{m} \det(e_1, \ldots, e_{i-1}, f_1, e_{i+1}, \ldots, e_m) \cdot \det(e_i, f_2, \ldots, f_m) \end{aligned}$$

$(e_1, \ldots, e_m, f_1, \ldots, f_m \in K^m)$ – the pair $(K^m \setminus \{0\}, w \circ \det)$ is a valuated matroid of rank m.

DEFINITION 2: *Given a valuated matroid* (E, v) *of rank* m, *we put*

$$T_{(E,v)} := \{p : E \to \mathbb{R} \mid \forall e \in E : p(e) = \max_{e_2, \ldots, e_m \in E} \{v(e, e_2, \ldots, e_m) - \sum_{i=2}^{m} p(e_i)\}\}.$$

$T_{(E,v)}$ *is also called the* tight span *of* (E, v) *or its* T-CONSTRUCTION.

The following proposition details this set of maps:

PROPOSITION 1: *Let* $H := \{(t_1, \ldots, t_m) \in \mathbb{R}^m \mid \sum_{i=1}^{m} t_i = 0\}$. *Then, for every base* $\{b_1, \ldots, b_m\} \in B_{(E,v)}$ *of a valuated matroid* (E, v) *of rank* m, *the map* $\Phi_{b_1, \ldots, b_m} : H \to \mathbb{R}^E$ *which maps each* $(t_1, \ldots, t_m) \in H$ *to the map*

$$E \to \mathbb{R} : e \mapsto \max_{1 \leq i \leq m} \{v(e, b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_m) + t_i\} - \tfrac{m-1}{m} v(b_1, \ldots, b_m)$$

is an injective map into $T_{(E,v)}$.

Furthermore, one has

$$T_{(E,v)} = \bigcup_{\{b_1,\ldots,b_m\}\in B_{(E,v)}} \Phi_{b_1,\ldots,b_m}(H).$$

Thus, $T_{(E,v)}$ is a union of (images of) affine hyperplanes of dimension $m-1$, called the *apartments* in $T_{(E,v)}$.

These apartments intersect as follows:

PROPOSITION 2:

1) *Given two bases $B, B' \subseteq E$ of a valuated matroid (E,v) of rank m, with suitable orderings of their elements as $B = \{b_1,\ldots,b_m\}$ and $B' = \{b'_1,\ldots,b'_m\}$, resp., one has*

$$\Phi_{b_1,\ldots,b_m}(H) \cap \Phi_{b'_1,\ldots,b'_m}(H) = \bigcap_{i=0}^{m} \Phi_{b_1,\ldots,b_i,b'_{i+1},\ldots,b'_m}(H).$$

2) *Given a base $\{b_1,\ldots,b_m\} \in B_{(E,v)}$, an element $b_0 \in E \setminus \{b_1,\ldots,b_m\}$, and a subset $I \subseteq \{1,\ldots,m\}$ so that $\{b_0, b_1,\ldots,b_{i-1},b_{i+1},\ldots,b_m\}$ is a base if and only if $i \in I$, then one has*

$$\begin{aligned} &\Phi^{-1}(\Phi_{b_1,\ldots,b_m}(H) \cap \Phi_{b_0,b_1,\ldots,b_{i-1},b_{i+1},\ldots,b_m}(H)) \\ = \;& \{(t_1,\ldots,t_m) \in H \;\Big|\; t_i + v(b_0,b_1,\ldots,b_{i-1},b_{i+1},\ldots,b_m) = \\ &\qquad \max_{j\in I}\{t_j + v(b_0,b_1,\ldots,b_{j-1},b_{j+1},\ldots,b_m)\}\} \end{aligned}$$

for every $i \in I$.

We return to our generic example mentioned above, that is, to the valuated matroid $(E := K^m \setminus \{0\}, v = w \circ \det)$, with K a field with a non-archimedean valuation $w : K \twoheadrightarrow \{-\infty\} \cup \mathbb{Z}$. By $\Gamma_{\mathbb{Z}}$, we denote the group of all affine maps from H to itself consisting of a translation by an integer vector and a permutation of coordinates, that is,

$$\begin{aligned} \Gamma_{\mathbb{Z}} := \{\gamma : H \to H \,\big|\, &(t_1,\ldots,t_m) \mapsto (t_{\sigma(1)} + a_1,\ldots,t_{\sigma(m)} + a_m) \text{ for some} \\ &(a_1,\ldots,a_m) \in H \cap \mathbb{Z}^m \text{ and some } \sigma \in S_m\}. \end{aligned}$$

Every subset

$$C = \{\Phi_{b_1,\ldots,b_m} \circ \gamma(t_1,\ldots,t_m) \;\Big|\; (t_1,\ldots,t_m) \in H \text{ with } t_1 \le t_2 \le \ldots \le t_m \le t_1 + 1\}$$

with $\{b_1,\ldots,b_m\} \in B_{(E,v)}$ some base and $\gamma \in \Gamma_{\mathbb{Z}}$ is called a CHAMBER of $T_{(E,v)}$; in case $\{b_1,\ldots,b_m\}$ is the canonical base of the vector space K^m and γ equals id_H, the resulting chamber C_0 is called the FUNDAMENTAL CHAMBER, while the

apartment $A_0 = \Phi_{b_1,\ldots,b_m}(H)$ for the canonical base is called the FUNDAMENTAL APARTMENT.
If a map p in $T_{(E,v)}$ satisfies $p(e) \equiv \frac{i}{m} \mod 1$ for some $i \in \{0,\ldots,m-1\}$ and every $e \in E$, then p is called a VERTEX of $T_{(E,v)}$ (OF TYPE i).
It is easy to see that the general linear group $GL_m(K)$ acts transitively on the set of vertices of $T_{(E,v)}$ via its group action defined on $T_{(E,v)}$ by

$$\begin{aligned} GL_m(K) \times T_{(E,v)} &\to T_{(E,v)} : \\ (X,p) &\mapsto (E \to \mathbb{R} : e \mapsto p(X^{-1}e) + \tfrac{1}{m} w \circ \det(X)) \end{aligned}$$

$(X \in GL_m(K), p \in T_{(E,v)}, e \in E)$.
This action induces a transitive action of the group $SL_m(K)$ on the set of apartments as well as on the set of chambers of $T_{(E,v)}$; since the stabilizers of these actions give rise to a BN-pair in the sense of building theory, one has

THEOREM 1: *For the valuated matroid $(E = K^m \setminus \{0\}, v = w \circ \det)$ with K a field with a non-archimedean valuation $w : K \twoheadrightarrow \{-\infty\} \cup \mathbb{Z}$, the T-construction $T_{(E,v)}$ is a geometrical realization of the affine building defined for the group $GL_m(K)$.*

Now, we come back to the general case of an arbitrary valuated matroid (E,v) of rank m.

LEMMA 1: *For every $p \in T_{(E,v)}$, the map*

$$\begin{aligned} d_p : E \times E &\to \mathbb{R} \\ (e,f) &\mapsto e^{\sup\{v(e,f,e_3,\ldots,e_m) - p(e) - p(f) - \sum_{i=3}^{m} p(e_i) \mid e_3,\ldots,e_m \in E\}} \end{aligned}$$

(with $e^{-\infty} := 0$) is a (pseudo-ultra-)metric on E.
In addition, for any two maps p and q in $T_{(E,v)}$, the metrics d_p and d_q are topologically equivalent.

DEFINITION 3: *A valuated matroid (E,v) of rank m is called* COMPLETE *if, for some (or equivalently: for every) $p \in T_{(E,v)}$, the metric space (E,d_p) is complete.*

Up to "projective equivalence" and identifying "parallel elements" (we refer to [DT1] for details), one has

THEOREM 2: *Every valuated matroid has an (essentially unique) completion.*

(In fact, one only has to complete (E,d_p) to a metric space $(\hat{E},\hat{d})$ and then to define $\hat{v} : \hat{E}^m \to \{-\infty\} \cup \mathbb{R}$ as the continuous extension of v.)

Concerning the T-construction, one has the following result:

THEOREM 3: *Let $(\hat{E},\hat{v})$ be a completion of the valuated matroid (E,v) with $\hat{E} \supseteq E$. Then the restriction map from $T_{(\hat{E},\hat{v})} \subseteq \mathbb{R}^{\hat{E}}$ to $\mathbb{R}^E$, mapping every $p \in T_{(\hat{E},\hat{v})}$ to $p\,\big|_E$, is a bijection into $T_{(E,v)}$.*

From now on, we assume for simplicity (E,v) to be a complete valuated matroid of rank m.

DEFINITION 4: *An* END *of $T_{(E,v)}$ is a map ε from $T_{(E,v)}$ to $\mathbb{R}$ satisfying*

(E1) for every base $\{b_1, \ldots, b_m\}$*, there exist some* $r \in \{1, \ldots, m\}$*, some affine map* $\gamma : H \to H$ *with a coordinate permutation as linear component, and some* $c \in \mathbb{R}$ *such that, for every* $(t_1, \ldots, t_m) \in H$*, the equation*

$$\varepsilon \circ \Phi_{b_1,\ldots,b_m} \circ \gamma(t_1, \ldots, t_m) = \max_{1 \le i \le r} t_i + c$$

holds;

(E2) there exist some base $\{b_1, \ldots, b_m\}$ *and some* $c \in \mathbb{R}$ *such that, for every* $(t_1, \ldots, t_m) \in H$*,*

$$\varepsilon \circ \Phi_{b_1,\ldots,b_m}(t_1, \ldots, t_m) = t_1 + c.$$

The set of all ends of $T_{(E,v)}$ *will be denoted by* $\mathcal{E}_{T_{(E,v)}}$*.*

With this definition, one has

PROPOSITION 3: *For every* $e \in E$*, the map*

$$\begin{aligned} \varepsilon_e : T_{(E,v)} &\to \mathbb{R} \\ p &\mapsto p(e) \end{aligned}$$

is an end of $T_{(E,v)}$*.*
And, for every $\varepsilon \in \mathcal{E}_{T_{(E,v)}}$*, there exist some* $e \in E$ *and some* $c \in \mathbb{R}$ *such that* $\varepsilon = \varepsilon_e + c$*.*

And one has

THEOREM 4: *If one defines a map* w *from the set* $\mathcal{E}^m_{T_{(E,v)}}$ *of* m*-tupels of ends of* (E, v) *to* $\{-\infty\} \cup \mathbb{R}$ *by*

$$w(\varepsilon_1, \ldots, \varepsilon_m) := \inf_{p \in T_{(E,v)}} \sum_{i=1}^{m} \varepsilon_i(p)$$

for $\varepsilon_1, \ldots, \varepsilon_m \in \mathcal{E}_{T_{(E,v)}}$*, then one has*

$$w(\varepsilon_{e_1}, \ldots, \varepsilon_{e_m}) = v(e_1, \ldots, e_m)$$

for all $e_1, \ldots, e_m \in E$*. That is,* $(\mathcal{E}_{T_{(E,v)}}, w)$ *is a complete valuated matroid of rank* m *which – up to "parallel elements" – is isomorphic to* (E, v)*.*

We now restrict ourselves to the case that the rank m equals 2. Here, $T_{(E,v)}$ is a *path-infinite* $\mathbb{R}$-tree, that is an $\mathbb{R}$-tree being the union of isometric images of the real line – namely the apartments from above: for any two $p, q \in T_{(E,v)}$, there exists some base $\{b_1, b_2\}$ such that $p, q \in \Phi_{b_1,b_2}(H)$, say $p = \Phi_{b_1,b_2}((s, -s))$ and $q = \Phi_{b_1,b_2}((t, -t))$ for some $s, t \in \mathbb{R}$; then putting $d(p, q) := |s - t|$ leads to a (well-defined) metric on $T_{(E,v)}$ having the desired property.
And the ends of $T_{(E,v)}$ in our sense correspond to its ends in the way ends are defined for $\mathbb{R}$-trees, that is, they correspond to (equivalence classes of) isometric embeddings of real halflines into $T_{(E,v)}$.

An example which we found particularly intriguing is the following one: Let E denote the set of subsets of $\mathbb{R}$ which are bounded from above, and for $e, f \in E$, let $v(e,f) := \sup(e \triangle f)$ be the supremum of their symmetric difference. Then it is easy to see that (E, v) is a valuated matroid of rank 2. The corresponding $\mathbb{R}$-tree has the particular property that omitting any point leads to the same "number" of connected components, and this number equals $\#\mathfrak{P}(\mathbb{R})$, the cardinality of the powerset of $\mathbb{R}$.

Now, it is well-known that, for the metric d of an $\mathbb{R}$-tree T, the so-called *four-point condition*

$$d(x,y) + d(z,w) \leq \max \left\{ \begin{array}{c} d(x,z) + d(y,w), \\ d(x,w) + d(y,z) \end{array} \right\}$$

holds for all $x, y, z, w \in T$. But this four-point condition is literally the exchange property (VM3) in the rank 2 case! Of course, one has $d(x,x) = 0$ instead of $d(x,x) = -\infty$ (cf. (VM2)).

This observation led us to the definition of matroidal trees:

DEFINITION 5: *A* MATROIDAL TREE *or, for short,* MATREE, *is a pair (X,u) consisting of a non-empty set X together with a map $u : X \times X \to \{-\infty\} \cup \mathbb{R}$ satisfying the following three conditions:*

(MT0) for every $x \in X$, there exists some $y \in X$ with $u(x,y) \neq -\infty$,

(MT1) u is symmetric,

(MT2) for all $x_1, x_2, y_1, y_2 \in X$, one has

$$u(x_1,x_2) + u(y_1,y_2) \leq \max \left\{ \begin{array}{c} u(y_1,x_2) + u(x_1,y_2), \\ u(y_1,x_1) + u(x_2,y_2) \end{array} \right\}$$

(and no restriction on the diagonal corresponding to (VM2)).

Note that, for every matree (X,u), the restriction $u \,\big|_{\{x \in X \mid u(x,x)=0\}^2}$ is a (pseudo)metric.

Now, let's have a look at the set

$$\begin{aligned} H_{(X,u)} := \{ f : X \to \{-\infty\} \cup \mathbb{R} \;\big|\; & f(x) + u(y,z) \leq \\ & \max \left\{ \begin{array}{c} f(y) + u(x,z), \\ f(z) + u(x,y) \end{array} \right\} \text{ for all } x,y,z \in X, f \not\equiv -\infty \}, \end{aligned}$$

the set of all *one-point extensions* of a matree (X,u) (containing at least all maps

$$\begin{array}{c} h_a : X \to \{-\infty\} \cup \mathbb{R} : \\ x \mapsto u(a,x) \end{array}$$

for $a \in X$).
If one wants to make a new matree $(H_{(X,u)}, w)$ from this set, and one wants the map $w : H_{(X,u)} \times H_{(X,u)} \to \{-\infty\} \cup \mathbb{R}$ to satisfy $w(h_x, h_y) = u(x,y)$ for all $x, y \in X$

(in order to have an "homomorphism" $X \to H_{(X,u)} : x \mapsto h_x$), and, slightly more general, $w(f, h_x) = f(x)$ for every $f \in H_{(X,u)}$ and every $x \in X$, then w necessarily has to satify

$$w(f,g) + u(x,y) \leq \max \left\{ \begin{array}{c} f(x) + g(y), \\ f(y) + g(x) \end{array} \right\}$$

for all $f, g \in H_{(X,u)}$ and all $x, y \in X$.
And, indeed, one has

THEOREM 5: *If, for a matree (X, u) and for $H_{(X,u)}$ as above, one defines*

$$w := w_{(X,u)} : H_{(X,u)} \times H_{(X,u)} \to \{-\infty\} \cup \mathbb{R}$$
$$(f,g) \mapsto \inf_{x,y \in X} \left\{ \max \left\{ \begin{array}{c} f(x) + g(y), \\ f(y) + g(x) \end{array} \right\} - u(x,y) \right\}$$

(with the convention $(-\infty) - (-\infty) := +\infty$), then $(H_{(X,u)}, w)$ is again a matree. In addition, for every $f \in H_{(X,u)}$ and every $x \in X$, one has

$$w(f, h_x) = f(x)$$

– in particular, one has $w(h_x, h_y) = u(x, y)$ for all $x, y \in X$.

The matree $(H_{(X,u)}, w)$ can be seen as a "hull" of (X, u), as one has

THEOREM 6: *If, for $F \in H_{(H_{(X,u)}, w_{(X,u)})}$, one defines*

$$\varphi(F) : X \to \{-\infty\} \cup \mathbb{R}$$
$$x \mapsto F(h_x),$$

and, for $f \in H_{(X,u)}$,

$$\psi(f) : H_{(X,u)} \to \{-\infty\} \cup \mathbb{R}$$
$$g \mapsto w_{(X,u)}(f, g),$$

then φ is a bijective map from $H_{(H_{(X,u)}, w_{(X,u)})}$ to $H_{(X,u)}$, and ψ is a bijective map in the other direction; both maps are inverse to each other; and for all $f, g \in H_{(X,u)}$, one has

$$w_{(H_{(X,u)}, w_{(X,u)})}(\psi(f), \psi(g)) = w(f, g).$$

Thus, $(H_{(H_{(X,u)}, w_{(X,u)})}, w_{(H_{(X,u)}, w_{(X,u)})})$ and $(H_{(X,u)}, w_{(X,u)})$ are canonically isomorphic matrees.

In addition, one has

THEOREM 7: *$H_{(X,u)}$ is the smallest set of maps $X \to \{-\infty\} \cup \mathbb{R}$ that a) contains $\{h_x \mid x \in X\}$ and b) is closed under addition of constants, under suprema, and under limites.*
More precisely: for every $f \in H_{(X,u)}$, one of the following three possibilities hold:

(i) there exist some $x \in X$ and some $c \in \mathbb{R}$ such that $f = h_x + c$,

(ii) there exist some $x, y \in X$ and some $b, c \in \mathbb{R}$ such that

$$f = \max\{h_x + d, h_y + c\},$$

(iii) there exist sequences $(x_n)_{n\in\mathbb{N}}$ in X and $(c_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ such that

$$f = \lim_{n\to\infty}(h_{x_n} + c_n).$$

Essential for the study of matrees is the following

FUNDAMENTAL LEMMA: *Let (X, u) be a matree; for $x, y \in X$ with $u(x, y) \neq -\infty$, put*

$$s_{x,y} := \frac{1}{2}(u(y, y) - u(x, y)) \in \{-\infty\} \cup \mathbb{R},$$
$$s^{x,y} := \frac{1}{2}(u(x, y) - u(x, x)) \in \mathbb{R} \cup \{+\infty\},$$

and $I(x, y) := [s_{x,y}, s^{x,y}] \cap \mathbb{R}$; for $t \in \mathbb{R}$, define $h^t \in H_{(X,u)}$ by

$$h^t := \max\{h_x + t, h_y - t\} - \frac{1}{2}xy.$$

Then the map $I(x, y) \to H_{(X,u)} : t \mapsto h^t$ is a surjective isometry onto the set

$$\{f \in H_{(X,u)} \mid w(h_x, h_y) = w(f, h_x) + w(f, h_y) \text{ and } w(f, f) = 0\}$$

– with isometry meaning that $w(h^s, h^t) = |s - t|$ holds for all $s, t \in I(x, y)$.

COROLLARY: *The set*

$$\{f \in H_{(X,u)} \mid w(f, f) = 0\}$$

is connected; hence – since the restriction of w to it is a metric satisfying the four-point condition – it is an $\mathbb{R}$-tree relative to the restriction of w (cf. [D]).

We want to close this section by a short discussion on the relationship between $H_{(E,v)}$ and $T_{(E,v)}$ for a valuated matroid (E, v) of rank 2.
For this, let

$$T'_{(E,v)} := \{p : E \to \mathbb{R} \mid p(e) = \sup_{f\in E}\{v(e, f) - p(f)\} \text{ for every } e \in E\}$$

– note the "sup" instead of "max" as for $T_{(E,v)}$; and define the canonical metric d on $T'_{(E,v)}$ by $d(p, q) := \sup_{e\in E} |p(e) - q(e)|$. It is easy to see that $T'_{(E,v)}$ is the completion of $T_{(E,v)}$ relative to this metric.
One should remark that $T'_{(E,v)}$ is the set of all minimal elements in the polytope

$$P_{(E,v)} := \{p : E \to \mathbb{R} \mid p(e) + p(f) \geq v(e, f) \text{ for all } e, f \in E\}$$

relative to the order $p \leq q :\iff p(e) \leq q(e)$ for every $e \in E$.

Coming back to the comparison of $H_{(E,v)}$ with $T'_{(E,v)}$, the following holds: The maps $p \in T'_{(E,v)}$ are exactly those maps in $H_{(E,v)}$ satisfying

$$w_{(E,v)}(p,p) = 0,$$

and one has

$$d = w_{(X,u)} \mid_{\{p \in H_{(E,v)} \mid w_{(E,v)}(p,p)=0\}^2}.$$

Slightly more general, one has

$$\{p \in H_{(E,v)} \mid w_{(E,v)}(p,p) \neq -\infty\} = \{p + c \mid p \in T'_{(E,v)}, c \in \mathbb{R}\}.$$

And the maps $p \in H_{(E,v)}$ satisfying $w_{(E,v)}(p,p) = -\infty$ correspond to the ends of the $\mathbb{R}$-tree $T_{(E,v)}$.

Based on these considerations, an algorithm for analyzing distance data and for constructing phylogenetic trees if those data fit exactly into trees and *phylogenetic networks* based on the T-construction if the data do not fit into a tree has been developed jointly with D. Huson and others which is available via `http://bibiserv.techfak.uni-bielefeld.de/splits/` where also further references can be found.

References

[BD1] H.-J. Bandelt, A. W. M. Dress: *Reconstructing the Shape of a Tree from Observed Dissimilarity Data*, Advances in Applied Mathematics 7 (1986), 309-343.

[BD2] H.-J. Bandelt, A. W. M. Dress: *A canonical decomposition theory for metrics on a finite set*, Advances in Mathematics 92 (1992), 47-105.

[BD3] H.-J. Bandelt, A. W. M. Dress: *Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data*, Molecular Phylogenetics and Evolution 1, No. 3 (1992), 242-252.

[D] A. W. M. Dress: *Trees, Tight Extensions of Metric Spaces, and the Cohomological Dimension of Certain Groups: A Note on Combinatorial Properties of Metric Spaces*, Advances in Mathematics 53 (1984), 321-402.

[DDH] A. W. M. Dress, J. Dopazo, A. v. Haeseler: *Split decomposition: a new technique to analyse viral evolution*, PNAS 90 (1993), 10320-10324.

[DEW1] A. W. M. Dress, M. Eigen, R. Winkler-Oswatitsch: *Statistical geometry in sequence space: A method of quantitative comparative sequence analysis*, Proc. Natl. Acad. Sci. USA 85 (1988), 5913-5917.

[DEW2] A. W. M. Dress, M. Eigen, R. Winkler-Oswatitsch: *How old is the genetic code?*, Science, 244 (1989), 673-679.

[DMT] A. W. M. Dress, V. L. Moulton, W. F. Terhalle: *T-Theory — An Overview*, European Journal of Combinatorics 17 (1996) 161-175.

[DT1] A. W. M. Dress, W. F. Terhalle: *A Combinatorial Approach to $\wp$-adic Geometry, Part I: The Process of Completion*, Geometriae Dedicata 46 (1993), 127-148.

[DT2] A. W. M. Dress, W. F. Terhalle: *The Real Tree*, Advances in Mathematics 120 (1996) 283-301.

[DW] A. W. M. Dress, R. Wetzel: *The Human Organism – a Place to Thrive for the Immuno-Deficiency Virus*, in: New Approaches in Classification and Data Analysis, ed. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy, Springer-Verlag, 1994, 636-643.

[FM] W. M. Fitch, E. Margoliash: *The Construction of Phylogenetic Trees — A Generally Applicable Method Utilizing Estimates of the Mutation Distance Obtained from Cytochrome c Sequences*, Science 155 (1967), 279-284.

[T1] W. F. Terhalle: *Ein kombinatorischer Zugang zu $\wp$-adischer Geometrie: Bewertete Matroide, Bäume und Gebäude*, Dissertation, Universität Bielefeld, 1992.

[T2] W. F. Terhalle: *Coordinatizing $\mathbb{R}$-Trees in Terms of Universal c-Trees*, Annals of Combinatorics 1 (1997), 183-196.

[T3] W. F. Terhalle: *$\mathbb{R}$-Trees and Symmetric Differences of Sets*, European Journal of Combinatorics 18 (1997), 825-833.

[T4] W. F. Terhalle: *Matroidal Trees. A Unifying Theory of Treelike Spaces and Their Ends*, Habilitationsschrift, Universität Bielefeld, submitted.

[WWW1] The www page THE TREE OF LIFE (http://phylogeny.arizona.edu/tree/phylogeny.html) provides uptodate information regarding our present knowledge of (or believe in) the detailed branching structure of species evolution.

[WWW2] The www page SPLITSTREE 2 (http://bibiserv.techfak.uni-bielefeld.de/splits) allows to use or to download the SplitsTree algorithm via the net and to view what that algorithm does to an appropriately specified data input.

[WWW2] The www page COMPUTATIONAL MOLECULAR EVOLUTION (http://dexter.gnets.ncsu.edu/lab/moleevol.html) provides links to (almost) all www pages relevant in this field, from data banks to algorithms (incl. SplitsTree 2) to journals.

Andreas Dress, Werner Terhalle
FSP Mathematisierung
Universität Bielefeld
Postfach 10 01 31
D-33501 Bielefeld
dress@mathematik.uni-bielefeld.de
terhalle@mathematik.uni-bielefeld.de

A New Version of the Fast Gauss Transform

Leslie Greengard and Xiaobai Sun

Abstract. The evaluation of the sum of N Gaussians at M points in space arises as a computational task in diffusion, fluid dynamics, finance, and, more generally, in mollification. The work required for direct evaluation grows like the product NM, rendering large-scale calculations impractical. We present an improved version of the fast Gauss transform [L. Greengard and J. Strain, *SIAM J. Sci. Stat. Comput.* 12, 79 (1991)], which evaluates the sum of N Gaussians at M arbitrarily distributed points in $O(N+M)$ work, where the constant of proportionality depends only on the precision required. The new scheme is based on a diagonal form for translating Hermite expansions and is significantly faster than previous versions.

1991 Mathematics Subject Classification: 65R10 , 44A35, 35K05
Keywords and Phrases: diffusion, fast algorithms, Gauss transform

1 Introduction

Many problems in mathematics and its applications involve the *Gauss transform*

$$G_\delta f(x) = (\pi\delta)^{-d/2} \int_\Gamma e^{-|x-y|^2/\delta} f(y)\, dy \qquad (\delta > 0) \tag{1}$$

of a function f, where Γ is some subset of $\mathbf{R}^d$. This is, of course, the exact solution to the Cauchy problem

$$\begin{aligned} u_t(x,t) &= \Delta u(x,t), \qquad t > 0 \\ u(x,0) &= f(x), \qquad x \in \mathbf{R}^d \end{aligned}$$

at time $t = \delta/4$ and corresponds to a *mollification* of the function f. Similar transforms occur in solving initial/boundary value problems for the heat equation by means of potential theory [3, 10, 11] and in nonparametric statistics [4, 17].

In the present paper, we will focus our attention on the *discrete Gauss transform*

$$G(x) = \sum_{j=1}^{N} q_j\, e^{-|x-s_j|^2/\delta}\ , \tag{2}$$

where the coefficients q_j and "source" locations s_j are given, and we wish to evaluate the expression (2) at a large number of "target" points x_j.

If the number of target points is denoted by M, we can define the rectangular transform matrix $\mathbf{G}$ by the formula

$$\mathbf{G}_{ij} = e^{-|x_i - s_j|^2/\delta}. \tag{3}$$

Direct application of this matrix to the vector $q = (q_1, \ldots, q_N)^T$. requires $O(NM)$ work, which makes large scale calculations prohibitively expensive.

To overcome this obstacle, Greengard and Strain developed a fast Gauss transform [9], which requires only $O(N + M)$ work, with a constant prefactor which depends on the physical dimension d and the desired precision. The amount of memory required is also proportional to $N + M$, so that the algorithm is asymptotically optimal in terms of both work and storage. In this scheme, the sources and targets can be placed anywhere; methods based on the fast Fourier transform (FFT), by contrast, are restricted to a regular grid and require $O(N \log N)$ operations. For the case where the variance δ is not constant:

$$G(x) = \sum_{j=1}^{N} q_j \, e^{-|x - s_j|^2/\delta_j}, \tag{4}$$

a generalization of the fast Gauss transform has been developed by Strain [18], but we will consider only the simpler case (2) here.

The fast Gauss transform is an *analysis-based* fast algorithm. Like the closely related fast multipole methods for the Laplace and Helmholtz equations [1, 5, 7, 13, 8, 14, 15, 16], it achieves a speedup in computation by using approximation theory to attain a specified, albeit arbitrarily high, precision. The FFT, on the other hand, is exact in exact arithmetic. It is an *algebra-based* fast algorithm which uses symmetry properties to reduce the computational work.

2 The original fast Gauss transform

The starting point for the fast algorithms of [9, 18] is the generating function for Hermite polynomials [2, 12]

$$e^{2xs - s^2} = \sum_{n=0}^{\infty} \frac{s^n}{n!} H_n(x) \ ,$$

where

$$H_n(x) = (-1)^n e^{x^2} D^n e^{-x^2} \qquad\qquad x \in \mathbf{R}$$

and $D = d/dx$. A small amount of algebra leads to the expansion

$$e^{-(x-s)^2/\delta} \quad = \quad \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{s - s_0}{\sqrt{\delta}} \right)^n h_n \left(\frac{x - s_0}{\sqrt{\delta}} \right),$$

where

$$h_n(x) = (-1)^n D^n e^{-x^2}$$

and s_0 is an arbitrary point.

This formula describes the Gaussian field $e^{-(x-s)^2/\delta}$ at the target x due to the source at s as an Hermite expansion centered at s_0. The higher dimensional analog of (5) is obtained using multi-index notation. Let x and s lie in d-dimensional Euclidean space $\mathbf{R}^d$, and consider the Gaussian

$$e^{-|x-s|^2} = e^{-(x_1-s_1)^2-\ldots-(x_d-s_d)^2}.$$

For any multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ and any $x \in \mathbf{R}^d$, we define

$$|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_d$$

$$\alpha! = \alpha_1!\alpha_2! \ldots \alpha_d!$$

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}$$

$$D^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_d^{\alpha_d}$$

where ∂_i is differentiation with respect to the ith coordinate in $\mathbf{R}^d$. If p is an integer, we say $\alpha \geq p$ if $\alpha_i \geq p$ for $1 \leq i \leq d$.

The multidimensional Hermite polynomials and Hermite functions are defined by

$$H_\alpha(x) = H_{\alpha_1}(x_1) \ldots H_{\alpha_d}(x_d)$$

$$h_\alpha(x) = e^{-|x|^2} H_\alpha(x) = h_{\alpha_1}(x_1) \ldots h_{\alpha_d}(x_d) \tag{5}$$

where $|x|^2 = x_1^2 + \ldots + x_d^2$. The Hermite expansion of a Gaussian in $\mathbf{R}^d$ is then simply

$$e^{-|x-s|^2} = \sum_{\alpha \geq 0} \frac{(x-s_0)^\alpha}{\alpha!} h_\alpha(s-s_0)\,. \tag{6}$$

Lemma 2.1 ([9], 1991) *Let N_B sources s_j lie in a box B with center s_B and side length $\sqrt{\delta}$. Then the Gaussian field due to the sources in B,*

$$G(x) = \sum_{j=1}^{N_B} q_j\, e^{-|x-s_j|^2/\delta}, \tag{7}$$

is equal to a single Hermite expansion about s_B:

$$G(x) = \sum_{\alpha \geq 0} A_\alpha\, h_\alpha \left(\frac{x - s_B}{\sqrt{\delta}} \right).$$

The coefficients A_α are given by

$$A_\alpha = \frac{1}{\alpha!} \sum_{j=1}^{N_B} q_j \left(\frac{s_j - s_B}{\sqrt{\delta}} \right)^\alpha. \tag{8}$$

The error $E_H(p)$ due to truncating the series after p^d terms satisfies the bound:

$$|E_H(p)| = |\sum_{\alpha \geq p} A_\alpha h_\alpha \left(\frac{x - s_B}{\sqrt{\delta}} \right)| \leq 2.75^d\, Q_B \left(\frac{1}{p!} \right)^{d/2} \left(\frac{1}{2} \right)^{(p+1)d/2} \tag{9}$$

where $Q_B = \sum |q_j|$.

LEMMA 2.2 ([9], 1991) *Let N_B sources s_j lie in a box B with center s_B and side length $\sqrt{\delta}$ and let x be a target point in a box C with center x_C. Then the corresponding Hermite expansion*

$$G(x) = \sum_{\alpha \geq 0} A_\alpha \, h_\alpha \left(\frac{x - s_B}{\sqrt{\delta}} \right) .$$

can be expanded as a Taylor series of the form

$$G(x) = \sum_{\beta \geq 0} B_\beta \left(\frac{x - x_C}{\sqrt{\delta}} \right)^\beta .$$

The coefficients B_β are given by

$$B_\beta = \frac{(-1)^{|\beta|}}{\beta!} \sum_{\alpha \geq 0} A_\alpha \, h_{\alpha+\beta} \left(\frac{s_B - x_C}{\sqrt{\delta}} \right). \tag{10}$$

The error $E_T(p)$ due to truncating the series after p^d terms satisfies the bound:

$$|E_T(p)| = | \sum_{\beta \geq p} B_\beta \left(\frac{x - x_C}{\sqrt{\delta}} \right)^\beta | \leq 2.75^d \, Q_B \left(\frac{1}{p!} \right)^{d/2} \left(\frac{1}{2} \right)^{(p+1)d/2} \tag{11}$$

These are the only tools required to construct a simple fast algorithm for the evaluation of

$$G(x_i) = \sum_{j=1}^{N} q_j e^{-|x_i - s_j|^2/\delta} \tag{12}$$

for $1 \leq i \leq M$, using $O(M + N)$ work. By shifting the origin and rescaling δ if necessary, we can assume (as a convenient normalization) that the sources s_j and targets x_i all lie in the unit box $B_0 = [0, 1]^d$.

ALGORITHM

STEP 1. Subdivide B_0 into smaller boxes with sides of length $\sqrt{\delta}$ parallel to the axes. Assign each source s_j to the box B in which it lies and each target x_i to the box C in which it lies. The source boxes B and the target boxes C may, of course, be the same.

STEP 2. Given ϵ, use Lemma 2.1 to create an Hermite expansion for each source box B with p^d terms satisfying:

$$\begin{aligned} G(x) &= \sum_B \sum_{s_j \in B} q_j e^{-|x - s_j|^2/\delta} \\ &= \sum_B \sum_{\alpha \leq p} A_\alpha(B) h_\alpha \left(\frac{x - s_B}{\sqrt{\delta}} \right) + O(\epsilon) \end{aligned}$$

where

$$A_\alpha(B) = \frac{1}{\alpha!} \sum_{s_j \in B} q_j \left(\frac{s_j - s_B}{\sqrt{\delta}} \right)^\alpha . \tag{13}$$

The amount of work required for this step is of the order $p^d N$.

Consider now a fixed target box C. For each $x_j \in C$, we need to evaluate the total field due to sources in all boxes of type B. Because of the exponential decay of the Gaussian field, however, it is easy to verify that, if we include only the sources in the nearest $(2r+1)^d$ boxes, we incur an error bounded by Qe^{-r^2}, where $Q = \sum_{j=1}^N |q_j|$. Given a desired precision ϵ, we can always choose r so that this truncation error is bounded by $Q\epsilon$. With $r = 4$, for example, we get single precision accuracy ($\epsilon = 10^{-7}$) and with $r = 6$, we get double precision ($\epsilon = 10^{-14}$). We denote the nearest $(2r+1)^d$ boxes as the *interaction region* for box C, denoted by *IR(C)*.

Step 3. For each target box C, use Lemma 2.2 to transform all Hermite expansions in source boxes within the *interaction region* into a single Taylor expansion. Thus, we approximate $G(x)$ in C by

$$\begin{aligned} G(x) &= \sum_B \sum_{s_j \in B} q_j e^{-|x-s_j|^2/\delta} \\ &= \sum_{\beta \le p} C_\beta \left(\frac{x - x_C}{\sqrt{\delta}} \right)^\beta + O(\epsilon) \end{aligned}$$

where

$$C_\beta = \frac{(-1)^{|\beta|}}{\beta!} \sum_{B \in IR(C)} \sum_{\alpha \le p} A_\alpha(B)\, h_{\alpha+\beta} \left(\frac{s_B - x_C}{\sqrt{\delta}} \right) , \tag{14}$$

and the coefficients $A_\alpha(B)$ are given by (13). Because of the product form (5) of $h_{\alpha+\beta}$, the computation of the p^d coefficients C_β involves only $O(\,d\,p^{d+1})$ operations for each box B. Therefore, a total of $O((2r+1)^d\,d\,p^{d+1})$ work per target box C is required. Finally, evaluating the appropriate Taylor series for each target x_i requires $O(p^d M)$ work. Hence this algorithm has net CPU requirements of the order

$$O((2r+1)^d\,d\,p^{d+1} N_{box}) + O(p^d N) + O(p^d M) ,$$

where the number of boxes N_{box} is bounded by $min(\delta^{-d/2}, N+M)$. The work is cleanly decoupled into three parts; $O(p^d N)$ to form Hermite expansions, $O(p^d M)$ to evaluate Taylor series, and a constant term depending on the number of box-box interactions and the cost of transforming Hermite expansions into Taylor series.

Remark: A proper implementation of the fast Gauss transform is a bit more complex. For example, if a box contains only a few sources, it is more efficient to compute their influence directly than to use expansions.

Suppose now that the source boxes are denoted by $B_1, B_2, \ldots, B_S$, that the target boxes are denoted by $C_1, C_2, \ldots, C_T$, that N_j sources lie in box B_j, that M_j targets lie in box C_j, and that the points are ordered so that

$$\begin{aligned} \{s_1, \ldots, s_{N_1}\} &\subset B_1 \\ \{s_{N_1+1}, \ldots, s_{N_1+N_2}\} &\subset B_2 \\ &\cdots \\ \{s_{N-N_S+1}, \ldots, s_N\} &\subset B_S, \\ \{x_1, \ldots, x_{M_1}\} &\subset C_1 \\ \{x_{M_1+1}, \ldots, x_{M_1+M_2}\} &\subset C_2 \\ &\cdots \\ \{x_{M-M_T+1}, \ldots, x_M\} &\subset C_T. \end{aligned}$$

Then the approximation $\mathbf{G}_\epsilon$ to the discrete Gauss transform matrix (3) can be written in the factored form

$$\mathbf{G}_\epsilon = \mathbf{D} \cdot \mathbf{E} \cdot \mathbf{F}. \tag{15}$$

Here, $\mathbf{F}$ is a block diagonal matrix of dimension $S \times S$. The jth diagonal block $\mathbf{F}(j) \in \mathbf{R}^{p^d \times N_j}$ satisfies

$$\mathbf{F}(j)_{n,m} = \frac{1}{\alpha_n!} \left(\frac{s_m - s_{B_j}}{\sqrt{\delta}} \right)^{\alpha_n},$$

where s_{B_j} is the center of box B_j and the p^d Hermite expansion coefficients are ordered in some fashion from $n = 0, \ldots, p^d$. $\mathbf{D}$ is very similar. It is a block diagonal matrix of dimension $T \times T$, with the jth diagonal block $\mathbf{D}(j) \in \mathbf{R}^{M_j \times p^d}$ satisfying

$$\mathbf{D}(j)_{n,m} = \frac{(-1)^{\beta_m}}{\beta_m!} \left(\frac{x_n - x_{C_j}}{\sqrt{\delta}} \right)^{\beta_m}$$

where x_{C_j} is the center of box C_j, and the p^d Taylor expansion coefficients are ordered from $n = 0, \ldots, p^d$ in the same fashion as the Hermite series. Note that, if the sources and targets coincide, then $\mathbf{D}$ is the transpose of $\mathbf{F}$.

The mapping $\mathbf{E}$ is a sparse block matrix of dimension $T \times S$, with up to $(2r+1)^d$ nonzero entries per row. The nonzero entries $\mathbf{E}(ij)$ are matrices of dimension $p^d \times p^d$, corresponding to a conversion of the Hermite series for box S_j into a Taylor series for box T_i, assuming S_j is in the *interaction region* $IR(T_i)$. The matrix entries are dense.

$$\mathbf{E}(ij)_{nm} = h_{\alpha_n+\beta_m} \left(\frac{s_{B_j} - x_{C_i}}{\sqrt{\delta}} \right).$$

Given this notation, Step 2 of the fast Gauss transform described above corresponds to multiplying the vector $\{q_1, q_2, \ldots, q_N\}$ by $\mathbf{F}$. Step 3 of the fast Gauss transform corresponds to multiplying the output of Step 2 by $\mathbf{E}$ to create all the Taylor expansions. The result is then multiplied by $\mathbf{D}$ to evaluate the Taylor series at all target locations.

Remark: The factorization (15) reveals the structure of G_ϵ. When δ is large enough, only one box is created and the rank of G_ϵ is bounded by p^d (the order of the factor $\mathbf{E}$). When δ is very small, the dimensions of $\mathbf{E}$ grow, but it becomes sparse and structured.

3 Diagonal form for translation operators

Our new version of the fast Gauss transform is based on replacing Hermite and Taylor expansions with an expansion in terms of exponentials (plane waves). The starting point is the Fourier relation

$$e^{-|x-s|^2/\delta} = \left(\frac{1}{2\sqrt{\pi}}\right)^d \int_{\mathbf{R}^d} e^{-|k|^2/4} e^{ik\cdot(x-s)/\sqrt{\delta}}\, dk \tag{16}$$

which is easily seen to satisfy the estimate

$$\left| e^{-\frac{|x-s|^2}{\delta}} - \left(\frac{1}{2\sqrt{\pi}}\right)^d \int_{|k|\le K} e^{-\frac{|k|^2}{4}} e^{i\frac{k\cdot(x-s)}{\sqrt{\delta}}}\, dk \right| \le \begin{cases} e^{-\frac{K^2}{4}} & \text{for } d = 1, 2 \\ K e^{-\frac{K^2}{4}} & \text{for } d = 3. \end{cases}$$

Setting $K = 7.5$, the truncation error from ignoring high frequency contributions is approximately 10^{-7}. Ssetting $K = 12$, the truncation error is approximately 10^{-14}. It still remains to discretize the Fourier integral in (16) within the range determined by K. The trapezoidal rule is particularly appropriate here since it is rapidly convergent for functions which have decayed at the boundary. Note, however, that the integrand is more and more oscillatory as $x - s$ grows. Fortunately, we only need accurate quadrature when s is within the interaction region of x, so that $|x-s|/\sqrt{\delta} \le 5$ for seven digit precision and $|x-s|/\sqrt{\delta} \le 7$ for fourteen digit precision. It is easy to verify that $p = 12$ equispaced modes in the interval $[0, 7.5]$ are sufficient to reduce the quadrature error to 10^{-7} when $|x-s|/\sqrt{\delta} \le 5$ and that $p = 24$ equispaced modes in the interval $[0, 12]$ are sufficient to reduce the quadrature error to 10^{-14} when $|x-s|/\sqrt{\delta} \le 7$.

Thus, for a source box B with center s_B, we replace the Hermite series of Lemma 2.1 with

$$\begin{aligned} G(x) &= \sum_{s_j \in B} q_j e^{-|x-s_j|^2/\delta} \\ &= \sum_{\beta \le p} C_\beta e^{i\frac{K\beta\cdot(x-s_B)}{p\sqrt{\delta}}} + O(\epsilon), \end{aligned}$$

where

$$C_\beta = \left(\frac{K}{2p\sqrt{\pi}}\right)^d e^{-|\beta|^2 K^2/(4p^2)} \sum_{j=1}^{N_B} q_j\, e^{-i\frac{K\beta\cdot(s_j - s_B)}{p\sqrt{\delta}}}.$$

There are two reasons to prefer this form. First, the translation operator described in Lemma 2.2 becomes diagonal.

COROLLARY 3.1 *Let N_B sources s_j lie in a box B with center s_B and side length $\sqrt{\delta}$ and let x be a target point in a box C with center x_C. Then the plane wave expansion*

$$G(x) = \sum_{\beta \leq p} C_\beta e^{i \frac{K\beta \cdot (x_i - s_B)}{p\sqrt{\delta}}} + O(\epsilon),$$

can be expanded about x_C as

$$G(x) = \sum_{\beta \leq p} D_\beta e^{i \frac{K\beta \cdot (x_i - x_C)}{p\sqrt{\delta}}} + O(\epsilon).$$

The coefficients D_β are given by

$$D_\beta = C_\beta \, e^{i \frac{K\beta \cdot (x_C - s_B)}{p\sqrt{\delta}}}. \tag{17}$$

In terms of matrix factorization, we have

$$\mathbf{G}_\epsilon = \mathbf{D}' \cdot \mathbf{E}' \cdot \mathbf{F}'. \tag{18}$$

In this formulation, the diagonal blocks of $\mathbf{F}'$ and $\mathbf{D}'$ are given by

$$\mathbf{F}'(j)_{n,m} = \left(\frac{K}{2p\sqrt{\pi}} \right)^{d/2} e^{\frac{-|\beta_n| K}{2p}} \, e^{-i \frac{K\beta_n \cdot (s_m - s_{B_j})}{p\sqrt{\delta}}}.$$

$$\mathbf{D}'(j)_{n,m} = \left(\frac{K}{2p\sqrt{\pi}} \right)^{d/2} e^{\frac{-|\beta_n| K}{2p}} \, e^{i \frac{K\beta_m \cdot (x_n - x_{C_j})}{p\sqrt{\delta}}}.$$

As in the original algorithm, note that if the sources and targets coincide, then $\mathbf{D}'$ is the adjoint of $\mathbf{F}'$. The nonzero entries $\mathbf{E}'(ij)$ are now *diagonal* matrices of dimension $p^d \times p^d$, with entries defined in (17). The net cost of all translations per target box is reduced from $O((2r+1)^d \, d \, p^{d+1})$ work to $O((2r+1)^d \, p^d)$ work.

The second (and more important) reason to prefer the new form is that the number of translations can be dramatically reduced. We describe the modification to the algorithm in the one-dimensional case. For this, imagine that we are sweeping across all boxes from left to right and that, at present, a target box C_j has accumulated all plane wave expansions from source boxes within its interaction region (Fig. 1(a)). The net expansion can be shifted to the center of C_{j+1} using Corollary 3.1. By adding in the contribution from the box marked by $+$ and subtracting the contribution from the box marked by $-$, we have the correct plane wave expansion for box C_{j+1} (Fig. 1(b)). Thus, $(2r+1)$ translations are replaced by three. In d-dimensions, the cost $O((2r+1)^d \, p^d)$ work can be reduced to $O(3d \, p^d)$, by sweeping across each dimension separately.

4 CONCLUSIONS

We have presented a new version of the fast Gauss transform, which uses plane wave expansions to diagonalize the translation of information between boxes. The

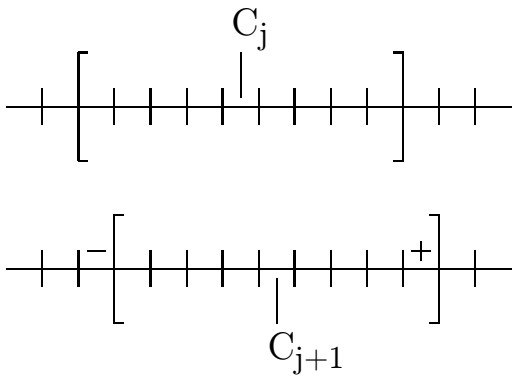


Figure 1: After shifting the expansion from box C_j to box C_{j+1}, one needs only to subtract the contribution from the box marked $-$ and add the contribution from the box marked $+$. (The interaction regions are indicated by the square brackets).

approach is similar to the new diagonal forms used in fast multipole methods for the Laplace and Helmholtz equations [8, 6, 13]. When the present improvements have been incorporated into existing fast Gauss transform codes, the resulting scheme should provide a powerful kernel for one, two and three-dimensional calculations.

References

[1] J. Carrier, L. Greengard, and V. Rokhlin, *A Fast Adaptive Multipole Algorithm for Particle Simulations*, Siam J. Sci. Stat. Comput., 9 (1988), pp. 669–686.

[2] H. Dym and H. P. McKean, *Fourier Series and Integrals*, Academic Press, San Diego, 1972.

[3] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, New Jersey, 1964.

[4] S. Geman and C. Hwang, *Nonparametric Maximum Likelihood Estimation by the Method of Sieves*, Ann. Statist., 10 (1982), pp. 401–414.

[5] L. Greengard, *Fast algorithms for classical physics*, Science, 265 (1994), pp. 909-xxx.

[6] L. Greengard, J. Huang, V. Rokhlin and S. Wandzura, *Accelerating fast multipole methods for low frequency scattering*, CMCL Report 1998-003, New York University.

[7] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[8] L. Greengard and V. Rokhlin, *A new version of the fast multipole method for the Laplace equation in three dimensions*, Acta Numerica, 6 (1987), pp. 229–269.

[9] L. Greengard and J. Strain, *The Fast Gauss Transform*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 79–94.

[10] L. Greengard and J. Strain, *A Fast Algorithm for the Evaluation of Heat Potentials*, Comm. on Pure and Appl. Math., 43 (1990), pp. 949–963.

[11] R. B. Guenther and J. W. Lee, *Partial Differential Equations of Mathematical Physics and Integral Equations,* Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[12] E. Hille, *A Class of Reciprocal Functions*, Ann. Math., 27 (1926), pp. 427–464.

[13] T. Hrycak and V. Rokhlin, *An improved fast multipole algorithm for potential fields*, Department of Computer Science Research Report 1089, Yale University, 1995.

[14] V. Rokhlin, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys. 60 (1985), pp. 187–207.

[15] V. Rokhlin, *Rapid solution of integral equations of scattering theory in two dimensions*, J. Comput. Phys., 86 (1990), pp. 414–439.

[16] V. Rokhlin, *Diagonal forms of translation operators for the Helmholtz equation in three dimensions*, Appl. and Comput. Harmonic Analysis 1 (1993), pp. 82–93.

[17] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[18] J. Strain, *The Fast Gauss Transform with variable scales*, SIAM J. Sci. Stat. Comput. 12 (1991), pp. 1131–1139.

Leslie Greengard
Courant Institute
New York University
New York, NY 10012-1110 USA
greengard@cims.nyu.edu

Xiaobai Sun
Department of Computer Science
Duke University
Durham, NC 27708-0129 USA
xiaobai@cs.duke.edu

Strategies for Seeing

Ulf Grenander

Abstract. We shall study the mathematical basis for computer vision using ideas from pattern theory. Starting from some general principles for vision several strategies for seeing will be derived and implemented by computer code. Using the code computer experiments have been carried out in order to examine the performance of the resulting inference engines for vision.

1991 Mathematics Subject Classification: 62P
Keywords and Phrases: pattern theory, computer vision

The mathematics of vision is not well understood. The human visual system is an awesome inference engine of unparalleled power, but its working remains a mystery in spite of great advances in the study of vision in recent years: much is known about its detailed functioning on the physiological level but theories proposed about its overall logical architecture are still tentative.

After the appearence of David Marr's seminal work, Marr (1982), many researchers in vision have adopted his view that seeing should be treated as a computational activity, where 'computational' is understood in a wide sense, more general than von Neumann architecture or Turing machines. We adhere to this view although we do not insist on his feed-forward paradigm. Therefore we believe that there should be a mathematical theory of vision underlying the visual computing and that machine vision would be aided by such a theory.

Another difference to Marr's approach is that we shall emphasize the primacy of analysis of the environment: this is needed for the understanding of the 'why' and 'how' of the algorithms that are realized through the sensory processing. An early proponent of this research strategy was Gibson with his 'ecological psychology', Gibson (1979).

To analyze the environment, the scene ensemble to be encountered by the visual system, we shall apply ideas from pattern theory and will use methods from this discipline as presented in Grenander (1993). A similar approach to vision, but oriented toward human rather than machine vision, has been outlined in Mumford (1994), (1996).

The vision strategies will be reductionist in the sense that they will be *derived* from general and mathematically articulated prinicples in contrast to being based on ad hoc devices. To achieve this the starting point will be the mathematical representation of the image algebra of the likely scenes. Different representations will lead to different strategies for seeing. Several strategies have been derived and

implemented computationally. We do not attribute much significance to the algorithms themselves, since they are based on quite simple minded representations, but more to the way they are derived from first principles.

1. Principles for vision.

1. *To be able to see it is necessary to know what one is looking for.* In other words, the system must be equipped with knowledge about scenes that are likely to be encountered and be based on an explicitly formulated purpose. It is therefore the first task for the system designer to express such knowledge in a form that is sufficiently precise for the software development. In a biological system such knowledge may have been created and stored during evolution, but we shall only be concerned with computer vision in the following. The system must also possess the ability to handle scenes it is not expecting, send warning signals and be honest enough to admit ignorance in doubtful situations.

2. *Different scene types and different sensors will require different strategies of vision.* To ask for a universal vision system, a system that is able to see and interpret anything, any electro-magnetic radiation emanating from completely arbitrary scenes, is a hopeless task. Instead of searching for such a chimera we shall narrow down and specify the ensemble of scenes that the system is intended for. We do not believe there is any universal representation valid for all scene/sensor combinations. Therefore the representations must be tailored to the particular scene types.

3. *Knowledge about the image esemble should be represented by logical structures formulated so precisely that they can serve as a basis for computing.* The representations shall be *compositional* in the sense that scenes are built from geometric objects, generators, that are combined together according to rules that may be deterministic or stochastic. They shall be *transformational* in that generators are themselves obtained from prototypes,*templates*, that are modified by transformations that play the role of generalizations.

It is clearly impossible to store all expected scenes in memory: this is avoided by the compositional/transformational scheme. Compare Chomskyan linguistics.

4. *The transformations shall form groups, arranged in a cascade that starts with solid, often low-dimensional, transformations and ends with diffeomorphisms.* The cascade will typically begin with translation, rotation, and perhaps scaling groups, whose semi-direct product forms a low dimensional group S_{solid}, but greater flexibility is needed to get enough generative power to deal with complex image ensembles of high variability and that will be supplied by the full diffeomorphic group S_{diff} or one of its high-dimensional sub-groups. The idea of group cascades has been examined in Matejic (1996). To represent abnormal variability it may be necessary to extend the transformations by giving up the group property, but this will not be explored here.

5. *The occurence of templates in the scene is controlled by probabilities, and deformations of the templates will be controlled by other probability measures on the groups; these measures evaluate how likely are the occurrences of various transformation of the templates.* Consider the set $C = S_{diff}/S_{solid}$ of right (or left) cosets of the sub-group S_{solid} in the full group S_{diff}.

The elements in C represent shape changes while S_{rigid} describes the less drastic transformations that moves sets around etc. The cosets can carry vital information while the elements of S_{solid} often play the role of nuisance parameters in the statistical sense of this term.

6. *The mechanism T that maps a scene into sensory entities shall be explicitly defined.* In general we shall let $\mathcal{T}$, the range of the T's, consist of arrays, not necessarily rectangular, with scalar entries and of fixed shape.

7. *The T transformation can be controlled by the system.* This allows the system to concentrate its attention on a detail of the scene, to direct its sensor(s) to point in a new direction or vary the focal length. In animal vision this corresponds to focussing the *fovea* and it also enables the vision system to function at different scales.

8. *The control of T is governed by an attention function A that attributes different weights to different parts of the observed image $I^{\mathcal{D}}$.* We should think of A as a real valued function of sub-images of $I^{\mathcal{D}}$ that takes real values, $A : 2^{I^{\mathcal{D}}} \to \mathbf{R}$. The attention function formalizes the purpose(s) of the vision engine.

9. *The saccadic search will be controlled by covariants w.r.t. the solid groups.* This will suggest plausible candidates for the generators that make up the true scene.

10. *For fixed T visual understanding will be attempted by an inference engine that selects plausible generators and elements from the groups that deform these generators.* In this way local decisions are made sequentially, forming, accepting or rejecting hypotheses. The selection may be deterministic, say maximizing some estimation criterion, or have random elements, as in simulating a posterior distribution. The visual understanding shall result in a structured description of the scene that can be used for decision making.

11. *The saccadic search is intended to reduce global inference problems to local ones.* The saccads should give rough estimates of the true group elements; the estimates will then be refined by applying the local group operations applying the diffeomorpic deformations.

12. *Once a ROI (Region Of Interest) has been analyzed the attention function is examined again to find other possible ROIs.* If the saccads result in more than one ROI they are all analyzed in the same way until the attention function points to no more ROI.

13. *The noise in the system is represented by a stochastic process N operating on the array outputted by the sensor transformation T.* Note that this randomness

is essentially different from the one governing the variability of the scenes. The latter is inherent in the vision setup, while the first can differ from sensor to sensor.

2. Mathematical Formalization.

2.1. Let us now express the principles mathematically and concretize to specific choices of assumptions that have been used in a series of computer experiments. But first let us explain what we mean by a solid group, or rather solid group action, in a general context, Consider a configuration of generators g_i coupled by the connector σ

$$c \;=\; \sigma(g_1, g_2, \ldots g_n) \;=\; \cup_k c^k$$

where the sub-configurations c^k are the connected (w.r.t. the neighborhood system induced by the graph σ) components of c. Then we shall define a general solid transformation to be of the form

$$c \;\mapsto\; \cup_k s^k c^k; s^k \in S_{solid}$$

so that each connected component is transformed separately and with the same group element for all the generators in the component and with the semi-direct product $S_{solid} = SL(d) \propto \mathbf{R}(d)$ of the special linear group with the translation group in d dimensions.

To formalize principles 1 - 3 let the generators form a space partitioned into the subsets G^α

$$G \;=\; \cup_\alpha G^\alpha$$

where α denotes the object type.

Principle 4 will be realized by choosing some of the sub-groups of S_{solid} and S_{diffeo}.

The purpose of the cascade is to allow large deformations, which is not possible with the single group elastic model, but without the large computing effort needed for the fluids model, see Christensen, Rabbit, Miller (1993).

Principle 5 will be implemented by introducing probability measure on the groups which is straightforward for the solid ones since they are low-dimensional. For S_{diffeo} (discretized approximation) on the other hand we induce a probability measure via the stochastic difference equation

$$(Ls)(x) \;=\; e(x); x \in X$$

for the displacement field $s(x) = (s_1(x), s_2(x))$ and $e(x)$ is a stochastic field; the group action is $x \mapsto x + s(x)$. Let us choose basis functions for S_{diffeo} (discretized to a lattice $\mathbf{Z}_{l_1 \times l_2}$) as the eigen functions of L as in Grenander (1993), p. 523,

$$\phi_{\mu\nu}(x) \;=\; sin(\frac{\pi x_1 \mu}{l_1}) sin(\frac{\pi x_2 \nu}{l_2}); x = (x_1, x_2) \in [1, l_1] \times [1, l_2]$$

with $\mu, \nu = 1, 2, \ldots r$, where the choice of r depends on the resolution of the sensor. Then we can expand the displacement fields

$$s_1(x) = \sum_{\mu=1}^{r} \sum_{\nu=1}^{r} t_{1\mu\nu} \phi_{\mu\nu}(x)$$

$$s_2(x) = \sum_{\mu=1}^{r} \sum_{\nu=1}^{r} t_{2\mu\nu} \phi_{\mu\nu}(x)$$

and we combine the Fourier coefficients into two matrices

$$t_1 = (t_{1\mu\nu}; \mu, \nu = 1, 2, \ldots r)$$

$$t_2 = (t_{2\mu\nu}; \mu, \nu = 1, 2, \ldots r)$$

We shall assume that for each generator index α the set G^α can be generated by applying S_{diffeo} to a single template g^α_{temp} so that

$$G^\alpha = S_{diffeo} g^\alpha_{temp}$$

In pattern theoretic terminology G^α then forms a pattern, actually a *finest pattern*, see Grenander (1993) p. 55-56. Then (G^α, S_{diffeo}) forms a homogeneous space.

For principle 6 we shall allow the range $\mathcal{T}$ of the T-transformation to be quite different from the scene that is being captured. For example, the output of a radar with a cross array of antennas will consist of two vectors with complex entries, superficially completely different from the target/background configuration. Or, the sinogram in a CAT scan which is quite different from the organ scanned.

Principles 7,8 will be realized by attention functions that will formalize the purpose of the system. For example, it could give great weight to regions close to the sensor, or to regions with high optical activity, or to objects of particular shape or texture. The function will generate saccadic movement of the fovea and/or the sensor(s).

For principle 9 we shall use classical covariants. Say that the intensity functions $I(\cdot)$ are continuous with compact support. For example, dealing with the translation group in the plane we use the 2-vector valued covariant

$$\phi^1(I) = m = 1/J \int\int (x_1, x_2) I(x_1, x_2) dx_1 dx_2$$

with

$$J = \int\int I(x_1, x_2) dx_1 dx_2$$

For $SO(2)$ we calculate the moment matrix

$$R = 1/J \int\int (x - m)(x - m)^T I(x) dx_1 dx_2$$

and diagonalize it $R = O^T DO$ and put

$$\phi^2(I) \;=\; O$$

Note however that this definition needs a further qualification in order to be unique. First, we should choose the orthogonal matrix O so that $det(O) > 0$ since we are dealing with the *special* orthogonal group $\mathbf{SO}(2)$. Second, we should select O so that its first column equals the eigen vector of R corresponding to the largest eigen vallue. The sign of the eigen vector is arbitrary so that this leads to an ambiguity that must be kept in mind when developing the code. Third, if the two eigen values coincide, typical for symmetric objects, we get more ambiguity and the covariant must be augmented with further information.

For the uniform scaling group in the plane we can use the scalar covariant

$$\phi^3(I) \;=\; 1/J \int\int \|x\| I(x) dx_1 dx_2$$

The use of saccadic search has split the global inference problem into several local ones in which we can let the inference engine look just for a local optimum, principles 10, 11.

Of course the whole group can push templates outside the total (bounded) region $\mathbf{Z}_{(l_1,l_2)}$, so that search should be limited to the latter region unless the sensor is re-directed to some other region. The saccadic search will lead to one ROI after another, point 12, until the remaining attention values are suffiently small; then the inference engine stops and outputs a structured description of the scene.

For principle 13 let us assume that the noise process of the system forms a stationary process in the plane, for example the Gaussian one with the non-singular covariance operator Cov. Then the likelihood function will be proportional to

$$L \;=\; exp - \frac{1}{2\sigma^2} \| I^{\mathcal{D}} \;-\; TsI_{temp} \|^2_{Cov^{-1}}$$

with the norm associated with the kernel Cov^{-1}. Introducing the positive definite square root M

$$M \;=\; +\sqrt{Cov^{-1}}$$

we can write the likelihood function in terms of the standard l_2-norm

$$L \;=\; exp - \frac{1}{2\sigma^2} \| MI^{\mathcal{D}} \;-\; MTsI_{temp} \|^2 =$$

$$= exp - E_{likelihood}$$

where $E_{likelihood}$ is the likelihood energy.

In a similar fashion we are led to prior probability measures on each group in the cascade. For the S_{diffeo}, for example, we have used the expression

$$E_{prior}(s) \;=\; 1/2\sigma^2 \sum_{\mu=1}^{l_1} \sum_{\nu=1}^{l_2} [l_1 \mu^2 t^2_{1\mu\nu} \;+\; l_2 \nu^2 t^2_{2\mu\nu}]$$

involving the t-matrices introduced earlier and with some scaling constant σ^2. We can then apply Markov Chain Monte Carlo to simulate the probability measure on one of the groups in the cascade and solve the SDE

$$ds(t) \;=\; -grad[E_{prior}(s) \;+\; E_{likelihood}(s)]dt \;+\; d(W(t)$$

in terms of the d-dimensional Wiener process $W(t)$ and continue iterating until the algorithmic time parameter t is so large that approximate statistical equilibrium has been reached. The previous propagated template, say $I^{\alpha}_{temp}(k,x)$, is then further propagated

$$I^{\alpha}_{temp}(k,x) \to I^{\alpha}_{temp}(k+1,x) = I^{\alpha}_{temp}(k, s^*_k x)$$

where s^*_k is the resulting group element from the SDE.

We now do this for each group in the cascade, successively propagating the template, see Matejic (1997). The resulting propagated template then induces the output of the vision engine under the adopted strategy for seeing.

3. Experiments.

Based on the above principles three strategies for seeing have been developed. Due to space limitations it is not possible to describe them in detail here; the reader is referred to Grenander (1998) where the strategies are fully described and their code is attached. The algorithms are so complex that it is difficult to predict their behavior. For this reason extensive experimentation has been carried out in order to find their strengths and weaknesses.

Here a few remarks will have to suffice. The first strategy was considering objects as sets in the plane and the attention funcion was then just measuring the optical activity in sub-sets. The observations were degraded by deformations of the generators, additive noise as well as clutter. Additive noise was easily handled by this algorithm, while clutter confused the algorithm and occassionally led it to make the wrong decision; this occurred even for moderate amounts of clutter. Obscuration was well handled if the overlap of generators was not too large but otherwise mistakes were made sometimes.

To handle obscuration better a second strategy was developed where the generators were closed simple curves in the plane, the boundaries of the sets. The attention function was designed to measure the (estimated) lengths of boundaries in subsets. Again additive noise caused no problem for the recognition algorithm. This strategy was less confused by obscuration than the first one, but it was quite sensitive to clutter, apparently because of the differential-geometric nature of the attention function.

A third strategy was constructed for a dynamic situation with moving generators. The attention function measured the amount of change in sub-sets from one frame to the next. This strategy was not very sensitive to clutter although it sometimes made the algorithm answer "do not understand the scene".

We draw the following conclusions from the experiments. The experiments have been carried out under controlled laboratory conditions, and since the inference algorithms are optimal modulo given assumptions, the observed weaknesses of the engines cannot be blamed on the construction of the algorithms. Instead they are essential to the visual set up and point to the need for a careful formulation of the purpose to be realized.
(i) Additive noise in not much of a problem but clutter is. In order to build effective strategies in the future one should *develop a better understanding of how clutter can represented mathematically.*
(ii) The purpose of a vision engine must be clearly articulated with *attention functions that combine several properties of the image*, not just a single one as in the three experiments.
(iii) Related to (ii) is the *need for incorporating cues in the observed "image"* $I^{\mathcal{D}}$: in addition to the image itself relevant facts known to the operator of the inference engine should also be included.
(iv) The vision engines should be *integrated systems* for multi-sensor, multi-target, multi-purpose situations with parallel implementations.

Work is under way to implement (i) - (iv).

REFERENCES.

G.E. Christensen, R.D. Rabbitt, M.I. Miller (1993): A Deformable Neuroanatomy Textbook Based on Viscous Fluid Mechanics, in Proc. 27th Annual Conf. Information Sci. and Systems, J.Prince and T. Runolfsson eds., pp, 211-216.

J.J. Gibson (1979): The Ecological Approach to Visual Perception, Houghton Mifflin.

U. Grenander (1993): General Pattern Theory, Oxford University Press.

U.Grenander (1998): Strategies for Seeing, Brown University Tech. Rep.

L. Matejic (1998): Group Cascades for Representing Biological Variability in Medical Images, to appear in the Quart. Appl. Math.

D.Mumford (1994): Neuronal Architectures for Pattern-Theoretic Problems, in Large-Scale Neuronal Theories of the Brain, C.Koch and J.Davis eds., MIT Press.

D.Mumford(1996): The Statistical Description of Visual Signals, in ICIAM95 , K.Kirchgassner, O. Mahrenholtz, R. Mennicken eds., Akademie Verlag.

Ulf Grenander
Brown University Box F
Providence RI 02912
USA

Canonical Models in Mathematical Neuroscience

Frank Hoppensteadt and Eugene Izhikevich

Abstract. Our approach to mathematical neuroscience is not to consider a single model but to consider a large family of neural models. We study the family by converting every member to a simpler model, which is referred to as being canonical. There are many examples of canonical models [7]. Most of them are derived for families of neural systems near thresholds; that is, near transitions between the rest state and the state of repetitive spiking. The canonical model approach enables us to study frequency and timing aspects of networks of neurons using frequency domain methods [6]. We use canonical (phase) models to demonstrate our theory of FM interactions in the brain: Populations of cortical oscillators self-organize by frequencies [6]; same-frequency sub-population of oscillators can interact in the sense that a change in phase deviation in one will be felt by the others in the sub-population [7]; and oscillators operating at different frequencies do not interact in this way. In our theory, sub-networks are identified by the firing frequency of their constituents. Network elements can change their sub-population membership by changing their frequency, much like tuning to a new station on an FM radio. Also discussed here are mechanisms for changing frequencies obtained in our recent work using similar models to study spatial patterns of theta and gamma rhythm phase locking in the hippocampus.

1991 Mathematics Subject Classification: Primary 11E16; Secondary 11D09, 11E04, 15A63.
Keywords and Phrases: Neuroscience, canonical models, phase locked loops.

A promising approach to mathematical neuroscience is to consider not a single neural model but a large family of such models. A reasonable way to study such a family is to convert every member to a simpler model by a continuous (possibly non-invertible) change of variables. We refer to such a simple model as being canonical for the family [7]. We present here a few examples of such families and their canonical models.

1 Neural Excitability

Most neurons are at rest, but they can fire repeatedly when stimulated. If the emerging firing pattern has very low frequency, then the neuron is said to exhibit

Class 1 neural excitability [5]. If it starts with a high frequency, it is said to exhibit Class 2 excitability.

The transition from rest to oscillatory firing as the stimulus is increased is a bifurcation. A typical bifurcation corresponding to Class 1 excitability is the *saddle-node on limit cycle* (SNLC) bifurcation. The family of all neural systems having this bifurcation has the canonical model

$$\theta' = (1 + \cos\theta) + (1 - \cos\theta)\lambda \,, \qquad \theta \in S^1 \,, \tag{1}$$

where λ is the bifurcation parameter that characterizes the stimulus [4, 7].

A typical bifurcation corresponding to Class 2 excitability is the *supercritical Andronov-Hopf* (AH) bifurcation. The family of all neural systems having this bifurcation has the canonical model

$$z' = (\lambda + i)z - z|z|^2 \,, \qquad z \in C \,, \tag{2}$$

which is a topological normal form for the bifurcation. Notice that (2) is local in the sense that a continuous change of variables that converts a dynamical system into (2) is defined is some small neighborhood of equilibrium. The canonical model (1) is not local in this sense. Many other canonical models for neuroscience applications are derived in [7].

2 FM Interactions in Phase Models

Rhythmic behavior is ubiquitous in nature and especially in the brain. Since we do not know (and probably will never know) the exact equations describing any neural system we consider a family of brain models of the following general form

$$x_i' = f_i(x_i) + \varepsilon g_i(x_1, \ldots, x_n, \varepsilon) \,, \qquad x_i \in R^m \,, \tag{3}$$

where each x_i describes activity of the ith neural element (neurons, cortical columns, etc.), and the dimensionless parameter $\varepsilon \geq 0$ measures the strength of connections. Many neuro-physiological experiments suggest that ε is small; see discussion in [7].

When each neural element exhibits oscillatory activity; that is, when each subsystem $x_i' = f_i(x_i)$ in (3) has a limit cycle attractor, then the weakly connected system (3) can be transformed into the canonical (phase) model

$$\theta_i' = \Omega_i + \varepsilon h_i(\theta_1, \ldots, \theta_n, \varepsilon) \,, \qquad \theta_i \in S^1 \,, \tag{4}$$

by a continuous change of variables. Here $\Omega_i > 0$ is the frequency, and θ_i is the phase of the ith oscillating element.

The phase model (4) can be simplified further depending on the presence of resonances between the frequencies $\Omega_1, \ldots, \Omega_n$. For example, when the frequencies are non-resonant and some other technical conditions are satisfied, each connection function h_i can be transformed into a constant. This implies that such oscillators do not interact even though there are synaptic connections between them; i.e., even though the functions g_i are non-constant in (3). A detailed analysis [7, 8] shows

that the interaction between oscillators is most effective when their frequencies are nearly identical, less effective when the frequencies are nearly low-order resonant, and practically non-effective otherwise.

Since this result was obtained for the canonical model (4), it can be applied to an arbitrary neural system of the form (3) regardless of the details of the mathematical equations. This universality suggests a far-reaching biological principle: The existence of synaptic connections between two neurons or two cortical columns does not guarantee that the they interact. To interact they must establish a certain low-order resonant relation between their frequencies. We say that interactions are frequency modulated (FM) in this case.

We see that an entire network can be partitioned into relatively independent ensembles of neurons processing information on different frequencies (channels). Each neuron can change its membership simply be changing its frequency. Thus, the entire brain can reconfigure itself by changing the frequency of oscillations of its units without changing the efficacy of synaptic connections (the wiring).

Finally, we notice that when the frequencies are chosen appropriately, the neural elements interact through modulation of the timing of their spikes. Therefore, the brain might employ FM radio principles: The frequency of neural rhythmic activity does not encode any information other than identifying the channel of communication; the information is carried by phases.

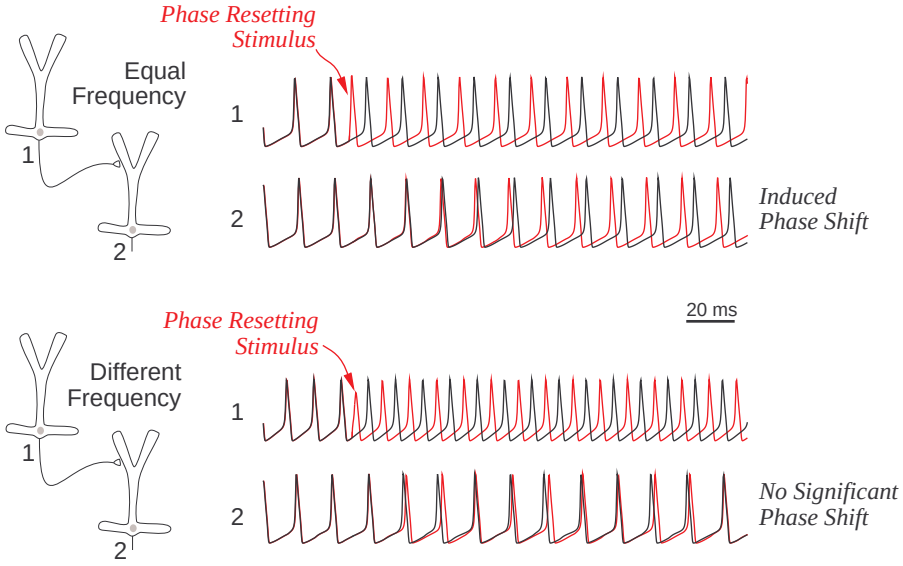


Figure 1: Temporal integration of a periodic input depends on the frequency of the input. *Upper part:* Neurons have identical frequencies. If a brief strong stimulus is applied to neuron 1 to change its phase, then neuron 2 can "feel" the change by acquiring a phase shift. *Lower part:* Neurons have different frequencies (close to the resonance 4 : 5.) The post-synaptic neuron is relatively insensitive to

the phase of the pre-synaptic one. (These simulations are based on space-clamped Hodgkin-Huxley equations.)

3 The Hippocampus

Similar methods are used to study the hippocampus and its role in information processing [1, 2]. In this, the three dimensional structure of the CA1, CA3 and DG regions of the hippocampus and their inputs from the medial septum and the entorhinal cortex are modeled by lumping the continuum model into discrete segments. These segments do not necessarily correspond to anatomical features of the hippocampus; they result from standard mathematical analysis. The model is

$$\dot{x}_j = \gamma + \cos x_j + (1 - \cos x_j)(\cos \phi_j(t) + \cos \psi_j(t) + \sum_{i=1}^{N} C_{i,j} V(x_i))$$

where

- γ is the gamma-rhythm frequency ($\approx 40Hz$).
- x_j is the phase of the j^{th} segment.
- ψ_j is the phase deviation of the input to the j^{th} segment from the entorhinal cortex. This is taken to be a theta-rhythm ($\approx 5Hz$) having phase deviations increasing along the array of sites from the right, so $\psi_j(t) = \omega t + j\Delta + \Phi$ where Δ is the propagation time of stimulation from one segment to the next.
- ϕ_j is the phase deviation of the input from the medial septum to segment j: $\phi_j(t) = \omega t + (N - j)\Delta$
- Φ indicates the difference in timing between the two inputs.

This system is depicted in Figure 2.

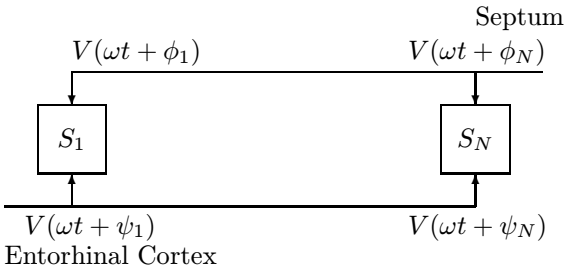


FIGURE 2: A segment model comprising N identical segments that have inputs from the Septum and from the Entorhinal Cortex, that have a fixed wave form (V), a fixed frequency (ω) and a phase deviation (ϕ_j or ψ_j). The phase differences along the line are $\phi_j - \psi_j$ for $j = 1, \cdots, N$.

The value Φ is the key control variable, and we show in [1] that as Φ increases through 2π various patterns of phase locking to the theta rhythm occur

in the model; the other segments oscillate at near the gamma rhythm. Thus, Φ is encoded in a spatial pattern of theta-rhythm activity. Figure 3 shows typical power-spectrum densities resulting from a simulation of 64 segments. Note that there is for the choice of Φ used here an interval of segments that are locked at the theta rhythm while the remaining segments oscillate at or near the gamma rhythm. Changing Φ changes the pattern of theta-rhythm oscillations. So, the firing frequency of individual cells can be changed by external forcing (here Φ) that is applied uniformly to the entire network.

Each of these phase variables has an asymptotic limit of the form $x_j \to \rho_j t + \phi_j(t)$ where ρ_j is the asymptotic frequency (rotation number) and $\phi_j(t)$ is the asymptotic phase deviation. This result is the basis of the rotation vector method which is discussed later.

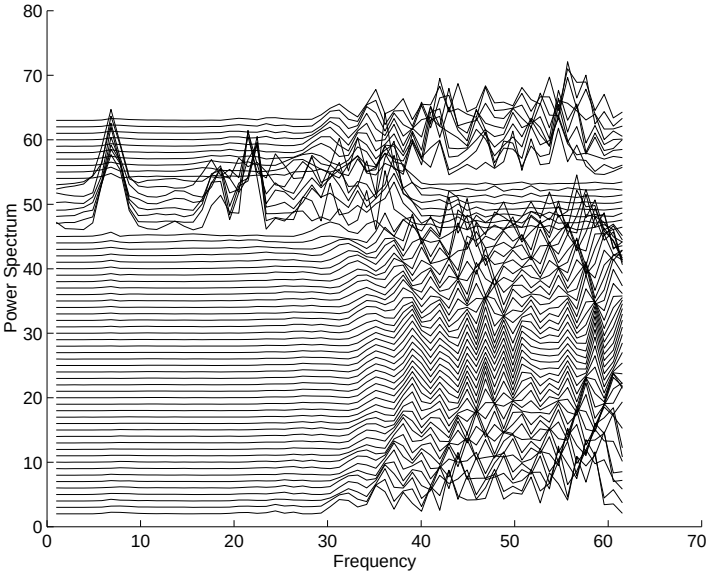


FIGURE 3: Power spectrum of 64 segments. There is an interval of segments having frequency $\approx 5Hz$, and the rest are near $40Hz$.

A sequence of input phases Φ_k can be memorized by adaptive connections within the structure. Lateral connections along the longitudinal axis of the hippocampus are modeled as before (Equation (??)), but now the connection strengths C change in response to correlation between pre- and post- synaptic activities:

$$\Omega \dot{C}_{i,j} + C_{i,j} = K \sin x_i \sin x_j$$

where K is a mixer gain and Ω is a time constant for a synapse. The matrix C accumulates memory traces, and it forms a slowly changing record of memorized states and transitions between them. In particular, this matrix can learn a sequence of input control variables $\Phi_1, \Phi_2, \cdots, \Phi_M$, and the resulting matrix has a left to right structure that can be used to recall this sequence. This is short term memory in the circuit.

In addition, this matrix C serves as the basis for studying recall of information in the presence of random noise. Its structure reflects the connections that correspond to memorized stimuli and to the transition from one memory to the next. This form can be abstracted into a Markov chain, and it can be studied by methods for Markov chains in random environments [1, 9].

4 Discussion

We will never know a complete model of any brain structure, no matter how small. However, a powerful aspect of mathematics is that (usually) quite simple models can accurately describe aspects of broad ranges of physical and biological systems. In particular, the approach we have developed for mathematical neuroscience is based on canonical models [7]. Care must be taken in interpreting and applying results obtained using canonical models, but a principal goal of this work is to suggest experiments and alternate ways of interpreting experimental data. Some outcomes of this approach are the use of VCONs to process voltage recordings from electrodes in behaving animals and the use of Markov chains to describe navigation by behaving rats.

Patterns of phase locking in networks of VCONs can be determined using the rotation vector method [6]: The vector $\vec{x}$ describes the phases in an entire network. If this population is in synchrony, then the phases have the form $\vec{x} \to \omega \mathbf{1} t + \vec{\phi}(t)$ where ω is the common frequency, $\mathbf{1}$ is the vector of all ones, and the phase deviations $\vec{\phi}$ are less significant in the sense that $\vec{\phi}(t)/t \to 0$ as $t \to \infty$. In FM radio, ω identifies the sending station and $\vec{\phi}$ carries the signal. We have shown here how two cells that are in synchrony can interact by demonstrating that a change in the timing of one will induce a change in the timing of the receiver. We propose that this is a fundamental mechanism for propagating and processing information in the brain. Using this approach, we can derive a system of equations for the phase deviations $\vec{\phi}$, and results of Liapunov and Malkin can be combined with singular perturbation methods to determine energy surfaces that govern the dynamics of $\vec{\phi}$ [6].

The illustrations from our hippocampus model suggest that there are many possible mechanisms for cells to change their firing frequency; for example, as described here through external oscillatory inputs or through chemical modification by hormones, neurotransmitters, etc.

The systems approach described here is based on canonical models, and it brings out possibilities for FM interactions and communications in brain structures by describing how a network can process such complex data in parallel.

References

[1] R. Borisyuk, F. C. Hoppensteadt (1998), Memorizing and recalling spatial-temporal patterns in an oscillator model of the hippocampus, Biosystems, in press.

[2] R. Borisyuk, F. C. Hoppensteadt, O. S. Vinogradova, Computational models of theta rhythm generators, submitted.

[3] G. Bus'zaki, R. Llinas, W. Singer, A. Berthoz, U. Christen, Eds., (1994) Temporal Coding in the Brain, Springer-Verlag, Heidelberg.

[4] G.B. Ermentrout and N. Kopell (1986) Parabolic Bursting in an Excitable System Coupled With a Slow Oscillation. SIAM Journal on Applied Mathematics 46:233–253.

[5] A. L. Hodgkin (1948), The local electric changes associated with repetitive action in non-medulated axon, J. Physiol. 107:165-181.

[6] F. C. Hoppensteadt (1997), An Introduction to the Mathematics of Neurons: Modeling in the Frequency Domain, Second Ed., Camb. U. Press, New York.

[7] F. C. Hoppensteadt, E. M. Izhikevich (1997), Weakly Connected Neural Networks, Springer-Verlag, New York.

[8] F.C. Hoppensteadt, E. M. Izhikevich (1998), Thalamo-Cortical Interactions Modeled by Weakly Connected Oscillators: Can Brain Use FM Radio Principles? Biosystems, in press.

[9] F. C. Hoppensteadt, H. S. Salehi, A. V. Skorokhod, Dynamical Systems with Random Perturbations, in preparation.

[10] P. Horowitz, W. Hill (1989) The Art of Electronics, Camb. U. Press, New York.

System Science
and Engineering Research Center
Departments of Mathematics
and of Electrical Engineering
Arizona State University
Tempe AZ 85287-7606
USA
fchoppen@asu.edu

Numerical Study of Free Interface Problems Using Boundary Integral Methods

Thomas Yizhao Hou

Abstract. Numerical study of fluid interfaces is a difficult task due to the presence of high frequency numerical instabilities. Small perturbations even at the round-off error level may experience rapid growth. This makes it very difficult to distinguish the numerical instability from the physical one. Here, we perform a careful numerical stability analysis for both the spatial and time discretization. We found that there is a compatibility condition between the numerical discretizations of the singular integral operators and of the Lagrangian derivative operator. Violation of this compatibility condition will lead to numerical instability. We completely eliminate the numerical instability by enforcing this discrete compatibility condition. The resulting scheme is shown to be stable and convergent in both two and three dimensions. The improved method enables us to perform a careful numerical study of the stabilizing effect of surface tension for fluid interfaces. Several interesting phenomena have been observed. Numerical results will be presented.

1991 Mathematics Subject Classification: 65M12, 76B15.
Keywords and Phrases: Free boundary, numerical stability, surface tension, topological singularity.

1 Introduction.

Many physically interesting problems involve propagation of free interfaces. Water waves, boundaries between immiscible fluids, vortex sheets, Hele-Shaw cells, thin-film growth, crystal growth and solidification are some of the better known examples. Numerical simulations for interfacial flows play an increasingly important role in understanding the complex interfacial dynamics, pattern formations, and interfacial instabilities. Many numerical methods have been developed to study these interfacial problems, including phase field models, volume-of-fluid methods, level set methods, front tracking methods, and boundary integral/element methods. Here we will focus on boundary integral methods.

Numerical study of fluid interfaces is a difficult task due to the presence of high frequency numerical instabilities [6, 7]. Small perturbations even at the

round-off error level may experience rapid growth. This makes it very difficult to distinguish the numerical instability from the physical one. In our study, we first establish the well-posedness of the linearized motion far from equilibrium. This involves careful analysis of singular integral operators defined on free interfaces [2, 11]. The continuous well-posedness analysis provides a critical guideline for our numerical analysis of the discrete system. We found that there is a corresponding compatibility condition between the discrete singular operators and the discrete derivative operator. Violation of this compatibility condition will lead to numerical instability. We completely eliminate the numerical instability by introducing an effective filtering. The amount of filtering is determined by enforcing this discrete compatibility condition. The resulting scheme is shown to be stable and convergent [3]. The corresponding 3-D problem is considerably more difficult since the singular operators have non-removable branch point singularities and there is no spectrally accurate discretization. A new stabilizing technique is introduced to overcome this difficulty [12, 13]. This technique is very general and effective. It also applies to non-periodic problems with rigid boundaries.

The improved method enables us to perform a careful numerical study of the stabilizing effect of surface tension for fluid interfaces. Water waves with small surface tension are shown to form singular capillary waves dynamically. The mechanism for generating such capillary waves is revealed and the zero surface tension limit is investigated [5]. In another study, surface tension is shown to regularize the early curvature singularity induced by the Rayleigh-Taylor instability in an unstably stratified two-fluid interface. However, a pinching singularity is observed in the late stage of the roll-up. The interface forms a trapped bubble and self-intersects in finite time [4, 9].

2 Stability of Boundary Integral Methods for 2-D Water Waves

In this section, we consider the stability of boundary integral methods for 2-D water waves. The result can be generalized to two-fluid interfaces and Hele-Shaw flows [4, 8]. Consider a 2-D incompressible, inviscid and irrotational fluid below a free interface. We assume the interface is 2π-periodic in the horizontal direction and parametrize the interface by a complex variable, $z(\alpha,t) = x(\alpha,t) + iy(\alpha,t)$, where α is a Lagrangian parameter along the interface. We use the usual convention of choosing the tangential velocity to be that of the fluid. The first boundary integral method for water waves was proposed by Longuet-Higgins and Cokelet [15] who used a single layer representation. Here we will use a double layer representation introduced by Baker-Meiron-Orszag [1]. Following [1], we obtain a system of evolution equations as follows:

$$\bar{z}_t = \frac{1}{4\pi i}\int_{-\pi}^{\pi} \gamma(\alpha')\cot(\frac{z(\alpha)-z(\alpha')}{2})d\alpha' + \frac{\gamma(\alpha)}{2z_\alpha(\alpha)} \equiv u - iv, \tag{1}$$

$$\phi_t = \frac{1}{2}(u^2 + v^2) - gy\ , \tag{2}$$

$$\phi_\alpha = \frac{\gamma}{2} + \mathrm{Re}\left(\frac{z_\alpha}{4\pi i}\int_{-\pi}^{\pi} \gamma(\alpha')\cot(\frac{z(\alpha)-z(\alpha')}{2})d\alpha'\right)\ , \tag{3}$$

where ϕ is the potential, γ is the vortex sheet strength, $\bar{z}$ is the complex conjugate of z. Equations (1)-(3) completely determine the motion of the system. The advantage of using the double layer representation is that the Fredholm integral equation of second kind has a global convergent Neumann series [1]. Thus γ can be solved by fixed point iteration.

The boundary integral formulation of water waves is naturally suited for numerical computation. There are many ways one can discretize the boundary integral equations, depending on how we choose to discretize the singular integrals and the derivatives. These choices affect critically the accuracy and stability of the numerical method. Straightforward numerical discretizations of (1)-(3) may lead to rapid growth in the high wavenumbers. In order to avoid numerical instability, a certain compatibility between the choice of quadrature rule for the singular integral *and* the discrete derivatives must be satisfied. This compatibility ensures that a delicate balance of terms on the continuous level is preserved on the discrete level. Violation of this compatibility will lead to numerical instability.

Let $z_j(t)$ be the numerical approximation of $z(\alpha_j, t)$, where $\alpha_j = jh$, $h = 2\pi/N$. $\phi_j(t), \gamma_j(t)$ are defined similarly. To approximate the velocity integral, we use the alternating trapezoidal rule:

$$\int_{-\pi}^{\pi} \gamma(\alpha') \cot\left(\frac{z(\alpha_j) - z(\alpha')}{2}\right) d\alpha' \simeq \sum_{\substack{j=-N/2+1 \\ (j-i) \text{ odd}}}^{N/2} \gamma_k \cot\left(\frac{z_j - z_k}{2}\right) 2h \,. \tag{4}$$

The advantage of using this alternating trapezoidal quadrature is that the approximation is spectrally accurate. We denote by D_h the discrete derivative operator. In general, we have $\widehat{(D_h)}_k = ik\rho(kh)$ for some nonnegative even function ρ. The specific form of $\rho(\xi)$ depends on the approximation. For example, we have $\rho_c(kh) = 3\sin(kh)/(kh(2+\cos(kh)))$ for the cubic spline approximation, and $\rho(kh) = 1$ for a pseudo-spectral derivative.

Now we can present our numerical algorithm for the water wave equations (1)-(3) as follows:

$$\frac{d\bar{z}_j}{dt} = \frac{1}{4\pi i} \sum_{(k-j)\text{odd}} \gamma_k \cot\left(\frac{z_j^{(\rho)} - z_k^{(\rho)}}{2}\right) 2h + \frac{\gamma_j}{2D_h z_j} \equiv u_j - iv_j, \tag{5}$$

$$\frac{d\phi_j}{dt} = \frac{1}{2}(u_j^2 + v_j^2) - g y_j \,, \tag{6}$$

$$D_h \phi_j = \frac{\gamma_j}{2} + Re\left(\frac{D_h z_j}{4\pi i} \sum_{(k-j)\text{odd}} \gamma_k \cot\left(\frac{z_j^{(\rho)} - z_k^{(\rho)}}{2}\right) 2h \right) \,, \tag{7}$$

where $z^{(\rho)}$ is a Fourier filtering defined as $\widehat{(z^{(\rho)})}_k = \hat{z}_k \rho(kh)$. The Fourier filtering $z^{(\rho)}$ in (5) and (7) is to balance the high frequency errors introduced by D_h. This will become apparent in the discussion of stability below.

THEOREM 1. *Assume that the water wave problem is well-posed and has a smooth solution in C^{m+2} ($m \geq 3$) up to time T. Then if D_h corresponds to a r-th*

order derivative approximation, we have for $0 < h \le h_0(T)$

$$\|z(t) - z(\cdot,t)\|_{l^2} \le C(T)h^r . \tag{8}$$

Similar convergent results hold for ϕ_j *and* γ_j. *Here* $\|z\|_{l^2}^2 = \sum_{j=1}^N |z_j|^2 h$.

2.1 Discussion of stability analysis

Here we discuss some of the main ingredients in the stability analysis of the scheme given by (5)-(7). We will mainly focus on the linear stability. Once linear stability is established, nonlinear stability can be obtained relatively easily by using the smallness of the error and an induction argument. The reader is referred to [3] for details.

To analyze linear stability, we first derive evolution equations for the errors $\dot{z}_j(t) \equiv z_j(t) - z(\alpha_j, t)$, etc., and try to estimate their growth in time. If we take the difference between the sum in (5) for the discrete velocity and the corresponding sum for the exact solution, the linear terms in $\dot{z}_j$, $\dot{\gamma}_j$ for the difference are

$$\frac{h}{\pi i} \sum_{(k-j)\text{odd}} \frac{\dot{\gamma}_k}{z(\alpha_j)^{(\rho)} - z(\alpha_k)^{(\rho)}} - \frac{h}{\pi i} \sum_{(k-j)\text{odd}} \frac{\gamma(\alpha_k)(\dot{z}_j^{(\rho)} - \dot{z}_k^{(\rho)})}{(z(\alpha_j)^{(\rho)} - z(\alpha_k)^{(\rho)})^2}, \tag{9}$$

where we have expanded the periodic sum, with k now unbounded. To identify the most singular terms, we use the Taylor expansion to obtain the most singular symbols

$$\frac{1}{z(\alpha_j) - z(\alpha_k)} = \frac{1}{z_\alpha(\alpha_j)(\alpha_j - \alpha_k)} + f(\alpha_j, \alpha_k) ,$$

where f is a smooth function. Thus, the most important contribution to the first term in (9) is $(2iz_\alpha)^{-1} H_h \dot{\gamma}_j$, where H_h is the discrete Hilbert transform

$$H_h(\dot{\gamma}_j) \equiv \frac{1}{\pi} \sum_{(k-j)\text{odd}} \frac{\dot{\gamma}_k}{\alpha_j - \alpha_k} 2h . \tag{10}$$

Similarly, the most important contribution to the second term in (9) is $-\gamma(2iz_\alpha^2)^{-1}\Lambda_h(z_j^{(\rho)})$, where Λ_h is defined as follows:

$$\Lambda_h(\dot{f}_j) \equiv \frac{1}{\pi} \sum_{(k-j)\text{odd}} \frac{\dot{f}_j - \dot{f}_k}{(\alpha_j - \alpha_k)^2} 2h . \tag{11}$$

Let H and Λ be the corresponding continuous operators for H_h and Λ_h respectively, i.e. replacing the discrete sums by the continuous integrals. In the continuous level, it is easy to show by integration by parts that

$$\Lambda(f) = H(D_\alpha f) , \tag{12}$$

where D_α is the continuous derivative operator. It turns out that in order to maintain numerical stability of the boundary integral method, the quadrature rule for the singular integral and the discrete derivative operator D_h must satisfy a compatibility condition similar to (12). That is, given a quadrature rule, which defines a corresponding discrete operators H_h and Λ_h, and a discrete derivative D_h, they must satisfy the following compatibility condition:

$$\Lambda_h(\dot{z}_i) \;=\; H_h D_h(\dot{z}_i)\;, \tag{13}$$

for $\dot{z}$ satisfying $\widehat{\dot{z}}_0 \;=\; \widehat{\dot{z}}_{N/2} \;= 0$. If (13) is violated, it would cause a mismatch of a singular operator of the form $(\Lambda_h - H_h D_h)(\dot{z})$ in the error equations. This will generate numerical instability.

By performing appropriate Fourier filtering in the approximations of the velocity integral, we can ensure a variant of the compatibility condition (13) is satisfied,

$$\Lambda_h(\dot{z}_j^{(\rho)}) \;=\; H_h D_h(\dot{z}_j)\;. \tag{14}$$

This can be verified from the spectrum properties of H_h and Λ_h and the definition of the ρ filtering. This modified compatibility condition is sufficient to ensure stability of our modified boundary integral method. This explains why we need to filter z in (5) and (7) when we approximate the velocity integral. The modified algorithm also allows use of non-spectral derivative operators.

By using properties of the discrete Hilbert transform: (i) $H_h^2 = -I$, (ii) $\Lambda_h(z^{(\rho)}) = H_h D_h(z)$, (iii) the commutator, $[H_h, f]$, is a smoothing operator, i.e. $[H_h, f](\dot{z}^{(\rho)}) = A_{-1}(\dot{z})$ for smooth f, we can derive an error equation for $\dot{z}_j$ which is similar to the continuum counterpart in the linear well-posedness study [2, 3]

$$\frac{d\dot{z}_j}{dt} = z_\alpha^{-1}(I - iH_h)D_h\dot{F} + A_0(\dot{z}) + A_{-1}(\dot{\phi}) + O(h^r),$$

where $\dot{F} = \dot{\phi} - u\dot{x} - v\dot{y}$, A_0 is a bounded operator from l^p to l^p, and A_{-1} is a smoothing operator of order one, i.e. $D_h A_{-1} = A_0$ and $A_{-1} D_h = A_0$. The leading order error equation suggests that we project the error equation into the local tangential and normal coordinate system. In this local coordinate, the stability property of the error equations becomes apparent. Let $\dot{z}^N, \dot{z}^T$ be the normal and tangential components of $\dot{z}$, with respect to the underlying curve $z(\alpha)$, $\mathbf{N}$ being the outward unit normal, and $\dot{\delta} = \dot{z}^T + H_h \dot{z}^N$. We obtain after some simplification

$$\dot{\delta}_t \;=\; A_{-1}(\dot{F}) + A_0(\dot{z}), \tag{15}$$

$$\dot{z}_t^N \;=\; \frac{1}{|z_\alpha|} H_h D_h \dot{F} + A_{-1}(\dot{F}) + A_0(\dot{z}), \tag{16}$$

$$\dot{F}_t \;=\; -c(\alpha,t)\dot{z}^N + A_{-1}(\dot{z}), \qquad c(\alpha,t) = (u_t, v_t + g)\cdot \mathbf{N}, \tag{17}$$

where equation (17) is obtained by performing error analysis on Bernoulli's equation and using the Euler equations. In this form it is clear that only the normal component of $\dot{z}$ is important. This is consistent with the physical property of interfacial dynamics. Now it is a trivial matter to establish an energy estimate for the error equations. Note that $H_h D_h$ is a positive operator with a Fourier symbol $\rho(kh)|k|$. The discretization is stable if the water wave problem is well-posed, i.e. the sign condition, $c(\alpha,t) > 0$, is satisfied. We refer to [3] for details.

3 Generalization to 3-D Water Waves.

Numerical stability of 3-D boundary integral methods is much more difficult. Let $\mathbf{z}(\alpha_1, \alpha_2)$ be a parametrization of a 3-D surface. Recall that the 3-D free space Green function for the Laplace equation is given by $G(\mathbf{z}) = -1/(4\pi|\mathbf{z}|)$. The corresponding velocity integral is given by

$$\frac{d\mathbf{z}}{dt} = \int \Omega(\alpha') \times \nabla_{\mathbf{z}'} G(\mathbf{z}(\alpha) - \mathbf{z}(\alpha')) d\alpha' + w_{loc}(\alpha) \ ,$$

where $\Omega = \mu_{\alpha_1} \mathbf{z}_{\alpha_2} - \mu_{\alpha_2} \mathbf{z}_{\alpha_1}$, $w_{loc} = \Omega(\alpha) \times (\mathbf{z}_{\alpha_1} \times \mathbf{z}_{\alpha_2}) / |\mathbf{z}_{\alpha_1} \times \mathbf{z}_{\alpha_2}|^2$ is the local velocity, and μ is the dipole strength.

One of the main difficulties for 3-D boundary integral methods is that the velocity integral has a branch point singularity which is not removable by desingularization. Also, unlike the 2-D case, we cannot express the leading order contribution of the singular operator as an integral operator defined on a flat surface. Now, the leading order singular operators depend on the free surface and have variable coefficients (assuming the tangent vectors are orthogonal for simplicity):

$$\begin{aligned} H_l(f) &= \frac{1}{2\pi} \int \frac{(\alpha_l - \alpha_l') f(\alpha') d\alpha'}{(|\mathbf{z}_{\alpha_1}(\alpha)|^2 (\alpha_1 - \alpha_1')^2 + |\mathbf{z}_{\alpha_2}(\alpha)|^2 (\alpha_2 - \alpha_2')^2)^{3/2}} \ , \quad l = 1, 2, \\ \Lambda(f) &= \frac{1}{2\pi} \int \frac{(f(\alpha) - f(\alpha')) d\alpha'}{(|\mathbf{z}_{\alpha_1}(\alpha)|^2 (\alpha_1 - \alpha_1')^2 + |\mathbf{z}_{\alpha_2}(\alpha)|^2 (\alpha_2 - \alpha_2')^2)^{3/2}}. \end{aligned}$$

As in 2-D, there are certain compatibility conditions among singular operators and the derivative operator. For example, we have $\Lambda = H_1 D_1 + H_2 D_2$. Stability of the boundary integral method requires a similar compatibility condition to hold:

$$\Lambda_h(z) = (H_1^h D_1^h + H_2^h D_2^h) z,$$

which, unfortunately, is generically violated by almost all discretizations. Although this compatibility condition can be imposed by applying a Fourier filtering as in the 2-D case, such filtering can no longer be evaluated efficiently by Fast Fourier Transform (FFT) since the singular operator, H_l^h or Λ_h, is not a convolution operator. The kernel depends on a variable coefficient.

In additional to the above compatibility condition, there are several other compatibility conditions that need to be satisfied for 3-D surfaces. Since there are no spectrally accurate approximations to the singular integrals in 3-D, it is almost impossible to enforce all the other compatibility conditions by using Fourier filtering alone.

To overcome this difficulty, we introduce a new stabilizing method without using the Fourier filtering. This new technique can be illustrated more clearly for the 2-D point vortex method [12]. Let us illustrate how we enforce the compatibility condition $\Lambda_h = H_h D_h$ indirectly by adding a stabilizing term. The modified point vortex method approximation for 2-D water waves is given by

$$\frac{d\bar{z}_j}{dt} = \frac{1}{2\pi i} \sum_{k \neq j} \frac{\gamma_k h}{z_j - z_k} + \frac{\gamma_j}{2 D_h z_j} + C_j^I,$$

where

$$C_j^I = \frac{\gamma_j}{2i(D_h z_j)^2}(\Lambda_h - H_h D_h) z_j.$$

This method is clearly consistent since the point vortex method gives a first order approximation to the singular integral: $(H_h D_h - \Lambda_h) z(\alpha_j) = O(h)$. Let $\dot{z}_j = z_j - z(\alpha_j)$, $\dot{\gamma}_j = \gamma_j - \gamma(\alpha_j)$ be the errors in z_j and γ_j. Let $E_i = \frac{1}{2\pi i} \sum_{j \neq i} \frac{\gamma_j}{z_i - z_j} h$, and define $\dot{E}_i = E_i - E(\alpha_i)$, $\dot{C}_i^I = C_i^I - C^I(\alpha_i)$. Using the same argument as before, we can show that the linear variation in E_i is given by

$$\dot{E}_i = \frac{1}{2iz_\alpha(\alpha_i)} H_h(\dot{\gamma}_i) - \frac{\gamma(\alpha_i)}{2iz_\alpha(\alpha_i)^2} \Lambda_h \dot{z}_i + A_0(\dot{z}_i) + A_{-1}(\dot{\gamma}_i) \ .$$

Similarly, we have

$$\dot{C}_i^I = \frac{\gamma(\alpha_i)}{2iz_\alpha(\alpha_i)^2}(\Lambda_h - H_h D_h)\dot{z}_i + A_0(\dot{z}_i) + A_{-1}(\dot{\gamma}_i),$$

where we have used the fact that $(\Lambda_h - H_h D_h) z(\alpha_i) = O(h)$. Now combining $\dot{E}_i$ with $\dot{C}_i^I$, we obtain

$$\begin{aligned} \dot{E}_i + \dot{C}_i^I &= \frac{1}{2iz_\alpha(\alpha_i)} H_h(\dot{\gamma}_i) - \frac{\gamma(\alpha_i)}{2iz_\alpha(\alpha_i)^2} \Lambda_h \dot{z}_i \\ &\quad + \frac{\gamma(\alpha_i)}{2iz_\alpha(\alpha_i)^2}(\Lambda_h - H_h D_h)\dot{z}_i + A_0(\dot{z}_i) + A_{-1}(\dot{\gamma}_i) \\ &= \frac{1}{2iz_\alpha(\alpha_i)} H_h(\dot{\gamma}_i) - \frac{\gamma(\alpha_i)}{2iz_\alpha(\alpha_i)^2} H_h D_h \dot{z}_i + A_0(\dot{z}_i) + A_{-1}(\dot{\gamma}_i). \end{aligned}$$

Note that the two $\Lambda_h \dot{z}_i$ terms cancel each other in the above equation, and only the $H_h D_h \dot{z}_i$ term survives in place of $\Lambda_h \dot{z}_i$. This in effect enforces the compatibility condition $\Lambda_h = H_h D_h$. This stabilizing technique is very general, and it applies to 3-D water waves. For 3-D water waves, we have four more compatibility conditions that need to be satisfied. We need to handle each one of them by adding a corresponding stabilizing term just as we outlined above. This will give a stable discretization for 3-D water waves. Moreover, by using a generalized arclength frame which enforces $|\mathbf{z}_{\alpha_1}|^2 = \lambda_1(t)|\mathbf{z}_{\alpha_2}|^2$ and $(\mathbf{z}_{\alpha_1}, \mathbf{z}_{\alpha_2}) = \lambda_2(t)|\mathbf{z}_{\alpha_2}|^2$, these correction terms can be evaluated efficiently by FFT, see [13].

4 Stabilizing Effect of Surface Tension

Surface tension plays an important role in understanding fluid phenomena such as pattern formation in Hele-Shaw cells, the motion of capillary waves on free surfaces, and the formation of fluid droplets. On the other hand, surface tension also introduces high order spatial derivatives into the interface motion through local curvature which couples to the interface equation in a nonlinear and nonlocal manner. These terms induce strong stability constraints on the time step if an explicit time integration method is used. These stability constraints are generally

time dependent, and become more severe by the differential clustering of points along the interface.

Hou, Lowengrub, and Shelley [8] proposed to remove the stiffness of surface tension for 2-D fluid interfaces by using the Small Scale Decomposition technique and reformulating the problem in the tangent angle θ and arclength metric s_α. Curvature has a very simple expression in these variables, $\kappa = \theta_\alpha / s_\alpha$. One important observation is that the stiffness only enters at small scales. The leading order contribution of these singular operators at small scales can be expressed in terms of the Hilbert transform, which is diagonalizable using Fourier Transform. By treating the leading order terms implicitly, but treating the lower order terms explicitly, we obtain a semi-implicit discretization which can be inverted efficiently using FFT. This reformulation greatly improves the time step stability constraint. Many interfacial problems that were previously not amenable are now solvable using this method. This idea has been subsequently generalized to 3-D filaments by Hou, Klapper, and Si who use curvature and arclength metric as the new dynamic variables [10]. Applications to Hele-Shaw flows, 3-D vortex filaments, and the Kirchhoff rod model for protein folding all give very impressive results.

In the following, we would like to present two numerical calculations using our numerical methods. In Fig. 1, we show that water waves with small surface tension generate singular capillary waves dynamically. Our study shows that the dynamic generation of capillary waves is a result of the competition between convection and dispersion. The capillary waves originate near the crest in a neighborhood where both the curvature and its derivative are maximum. For fixed but small surface tension, the maximum of curvature increases in time and the interface develops oscillatory capillary waves in the forward front of the crest. The minimum distance between adjacent capillary crests appears to approach zero, suggesting the formation of trapped bubbles as observed in Koga's experiments of breaking waves [14]. On the other hand, for a fixed time, as the surface tension coefficient τ is reduced, both the capillary wavelength and its amplitude decreases nonlinearly. The interface converges strongly to the zero surface tension profile [5].

We study the stabilizing effect of surface tension for an unstably stratified two-fluid interface in Fig. 2. This problem was first investigated by Pullin in [16]. Due to the numerical instability, Pullin's calculations were not conclusive. Using our improved method, we do not observe any numerical instability and we are able to perform well-resolved calculations to study the stabilizing effect of surface tension. Our study shows that surface tension indeed regularizes the early curvature singularity induced by the Rayleigh-Taylor instability. The interface rolls up into two spirals as time evolves. Note that the tips of the fingers broaden as they continue to roll, and that the interface bends towards the tip of the fingers. At around $t = 1.785$, the interface forms a trapped bubble and self-intersects. The minimum distance between the neck of the bubble is approximately 5×10^{-4} [4]. This process of bubble formation through self-intersection of a fluid interface has been observed in [8, 9] for a vortex sheet. In both cases, we found a convincing evidence that the minimum distance between the neck of the bubble scales like $(t_c - t)^{2/3}$, providing a partial agreement with the self-similar scaling.

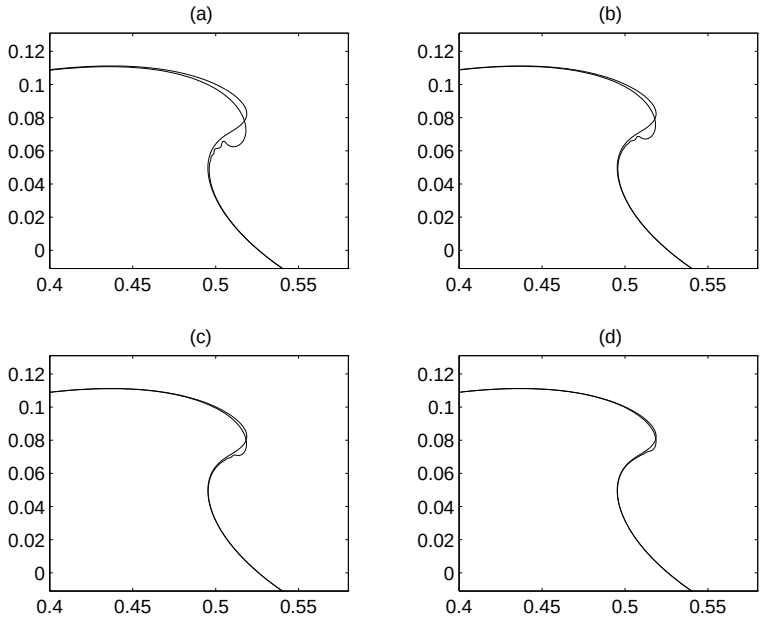


Figure 1: Comparison of the zero surface tension interface profile with the corresponding ones for decreasing surface tension τ at $t = 0.45$, $N = 2048$. (a)$\tau = 2.5 \times 10^{-4}$. (b)$\tau = 1.25 \times 10^{-4}$. (c)$\tau = 6.25 \times 10^{-5}$. (d)$\tau = 3.125 \times 10^{-5}$.

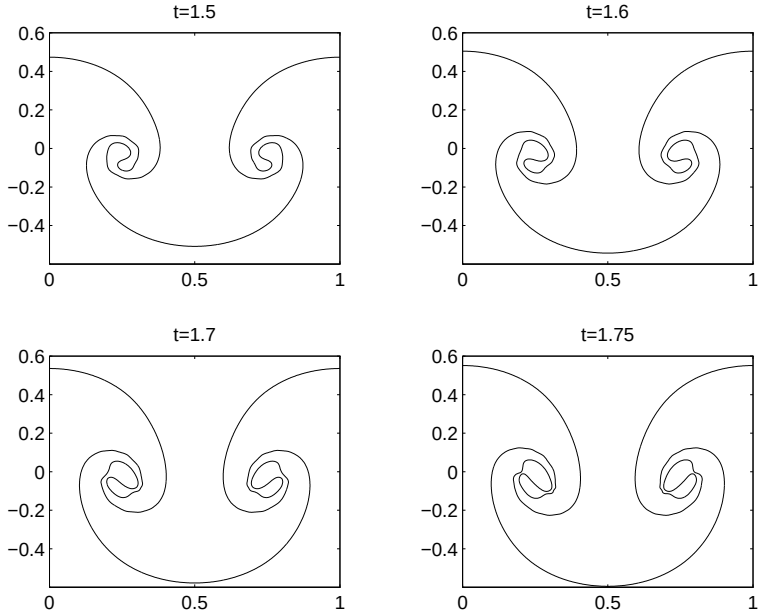


Figure 2: Rayleigh-Taylor instability: Atwood ratio $A = -0.1$, surface tension $\tau = 0.005$, $N = 2048$ and $\Delta t = 1.25 \times 10^{-4}$.

REFERENCES

[1] G. Baker, D. Meiron, and S. Orszag, *Generalized vortex methods for free-surface flow problems*, J. Fluid Mech. *123*, 477-501 (1982).
[2] J. T. Beale, T. Y. Hou and J. Lowengrub, *Growth Rates for Linearized Motion of Fluid Interfaces Away from Equilibrium,* Comm. Pure Appl. Math, *46*, 1269-1301 (1993).
[3] J. T. Beale, T. Y. Hou and J. Lowengrub, *Convergence of a Boundary Integral Method for Water Waves,* SIAM J. Numer. Anal., *33*, 1797-1843 (1996).
[4] H. Ceniceros and T. Y. Hou, *Convergence of a Non-Stiff Boundary Integral Method for Interfacial Flows with Surface Tension,* Math. Comp., *67*, 137-182 (1998).
[5] H. Ceniceros and T. Y. Hou, *Dynamic Generation of Capillary Waves*, submitted to Phys. Fluid. A, 1998.
[6] J. W. Dold, *An Efficient Surface-Integral Algorithm Applied to Unsteady Gravity Waves,* J. Comput. Phys., *103*, 90-115 (1992).
[7] T. Y. Hou, *Numerical Solutions to Free Boundary Problems,* Acta Numerica, 1995, pp. 335-415.
[8] T. Y. Hou, J. Lowengrub and M. Shelley, *Removing the Stiffness from Interfacial Flows with Surface Tension,* J. Comput. Phys, *114*, 312-338 (1994).
[9] T. Y. Hou, J. Lowengrub and M. Shelley, *The Long-Time Motion of Vortex Sheets with Surface Tension,* Phys. of Fluid, A, *9*, 1933-1954 (1997).
[10] T. Y. Hou, I. Klapper, and H. Si, *Removing the Stiffness of Three Dimensional Interfacial Flows with Surface Tension,* to appear in J. Comput. Phys., 1998.
[11] T. Y. Hou, Z.-H. Teng, and P. Zhang, *Well-Posedness of Linearized Motion for 3-D Water Waves Far From Equilibrium,* Comm. in PDE's, *21*, 1551-1586 (1996).
[12] T. Y. Hou and P. Zhang, *Stability of the Point Vortex Method for 2-D Water Waves,* submitted to SIAM J. Numer. Anal., 1998.
[13] T. Y. Hou and P. Zhang, *Stability of the Point Vortex Method for 3-D Water Waves,* submitted to SIAM J. Numer. Anal., 1998.
[14] M. Koga, *Bubble entrainment in breaking wind waves,* Tellus, *34*, 481-489 (1982).
[15] M. S. Longuet-Higgins and E. D. Cokelet, *The deformation of steep surface waves on water, I. A numerical method of computation,* Proc. Roy. Soc. London A, *350*, 1-26 (1976).
[16] D.I. Pullin, *Numerical studies of surface tension effects in nonlinear Kelvin-Helmholtz and Rayleigh-Taylor instability*, J. Fluid Mech., *119*, 507-532 (1982).

Applied Mathematics, 217-50
California Institute of Technology
Pasadena, CA 91125
USA
Email: hou@ama.caltech.edu

Travelling Water-Waves, as a Paradigm for Bifurcations in Reversible Infinite Dimensional "Dynamical" Systems

Gérard Iooss

Abstract. We first show a typical bifurcation study for a finite dimensional reversible system, near a symmetric equilibrium taken at 0. We state the results on known small bounded solutions: periodic, quasi-periodic, homoclinic to 0, and homoclinics to periodic solutions. The main tool for such a study is center manifold reduction and normal form theory, in presence of reversibility. This allows to prove persistence of large class of reversible (symmetric) solutions under higher order terms, not considered in the normal form. We then present water-wave problems, where we look for 2D travelling waves in a potential flow. In case of finite depth layers, the problem of finding small bounded solutions, is shown to be reducible to a finite dimensional center manifold, on which the system reduces to a reversible ODE. Bounded solutions of this ODE lead to various kinds of travelling waves which are discussed.

If the bottom layer has infinite depth, which appears to be the most physically realistic case, concerning the validity of results in the parameter set, the mathematical problem is more difficult. We don't know how to reduce it to a finite dimensional one, due to the occurence of a continuous spectrum (of the linearized operator) crossing the imaginary axis. We give some hints, on how to attack this difficulty, specially for periodic and homoclinic solutions which have now a *polynomial decay* at infinity .

1991 Mathematics Subject Classification: 58F39, 58F14, 76B15, 34A47, 76B25

1 Bifurcations of reversible systems near a symmetric equilibrium

1.1 Basic tools

Let us first consider a finite dimensional vector field of the form

$$\frac{dU}{dx} = F(U) \tag{1}$$

where $U(x)$ lies in $\mathbb{R}^n$, we say that system (1) is *reversible* if there exists a linear symmetry S, satisfying $S^2 = \mathbb{I}$, such that $SF = -F \circ S$. This implies, in particular, that if $x \longmapsto U(x)$ is solution of (1), then $x \longmapsto SU(-x)$ is also solution. Assume

in addition that $F(0) = 0$ and that F is $C^k, k \geq 2$, and define the derivative at the origin: $L = DF(0)$. It is clear that $SL = -LS$, which implies that the set of eigenvalues of L is symmetric with respect to both axis in $\mathbb{C}$. In what follows, we are specially interested in solutions of (1) which *stay in a neighborhood of* 0 *for* $x \in \mathbb{R}$. The main tool for understanding such solutions is a center manifold reduction theorem [19] (see [28] for a complete and pedagogic proof):

THEOREM 1 (CENTER MANIFOLD THEOREM) *Assume that the spectrum of L is composed with a part σ_0 on the imaginary axis and another part σ_h lying at a positive distance from the imaginary axis. Let us denote respectively by E_0 and E_h the subspaces invariant under L, corresponding to this splitting of the set of eigenvalues of L. Then, there exists a function $\Psi \in C^k(E_0, E_h), \Psi(0) = 0, D\Psi(0) = 0$, and a neighborhood $\mathcal{U}$ of 0 in $\mathbb{R}^n$, such that the manifold*

$$M_0 = \{X + \Psi(X) | X \in E_0\} \subset \mathcal{U}$$

has the following properties

(i) M_0 is locally invariant under (1);

(ii) M_0 contains all solutions of (1) staying in $\mathcal{U}$ for all $x \in \mathbb{R}$;

(iii) Ψ commutes with the symmetry S : $S\Psi = \Psi \circ S_0$ (we denote by S_0 the restriction of S on the space E_0).

The part (iii) of the above theorem is not in [19] but results easily from the proof of the theorem, as it is also true for any linear unitary operator commuting with F.

We are in fact interested in *Bifurcations of solutions* lying in a neighborhood of 0, i.e. in a structural change of these solutions when some parameter varies. To fix ideas, we now consider systems of the form

$$\frac{dU}{dx} = F(\mu, U), \quad F(0,0) = 0 \tag{2}$$

where μ is a real parameter, F being smooth with respect to both arguments, and $F(0, \cdot)$ satisfies the same assumptions as F in (1). Then, it is nearly straightforward that there is a neighborhood of 0 in $\mathbb{R} \times \mathbb{R}^n$ for (μ, U) for which a family of center manifolds M_μ exist, of the form

$$U = X + \Psi(\mu, U), X \in E_0, \Psi(0,0) = 0, D_X\Psi(0,0) = 0. \tag{3}$$

The interest of this result rests in particular in the "uniform" validity for μ in a neighborhood of 0. Indeed, in case 0 stays an equilibrium of (2) when μ varies, the eigenvalues of $D_U F(\mu, \cdot)$ may escape from the imaginary axis, and we might be tempted to apply the classical invariant manifold theorem for hyperbolic situations. This would lead to a domain of validity much smaller than the one given by the present theorem. Of course we "pay" this by the non uniqueness of such center manifolds, and the fact that the more regularity we wish, the smaller is the existence domain for M_μ.

The reduced system on M_μ is written

$$\frac{dX}{dx} = f(\mu, X) \text{ in } E_0 \tag{4}$$

and is *still reversible*, provided that the representation S_0 of S on E_0 is not trivial. Moreover, $f(0,0) = 0$ and $D_X f(0,0) = L_0 = D_U F(0,0)|_{E_0}$ has all its eigenvalues of zero real part.

Now, a very powerfull tool for studying the reduced system (4) is *normal form* theory. This technique consists in making near the origin, a change of variables close to identity and polynomial in X, which modifies the form of (4) in simplifying its Taylor expansion up to a fixed order (the degree of the polynomial). We then expect to recognize more easily relevant solutions of our system on a "simplified" f. Normal form theory goes back to Poincaré and Birkhoff and was more recently developed in particular by V.Arnold [2], Belitskii [4], Cushman & Sanders [6] and Elphick et al [9]. In the context of a system like (4) where all eigenvalues of L_0 lie on the imaginary axis, we use the following global characterization result of [9], (see also [12]):

THEOREM 2 (NORMAL FORM THEOREM) *For any $p \leq k$, there is a neighborhood $\widetilde{\mathcal{U}}_p$ of 0 in $\mathbb{R}\times E_0$ and there are polynomials $\Phi(\mu,\cdot)$ and $N(\mu,\cdot)$ $E_0 \to E_0$, of degree p, with coefficients smooth in μ, such that $\Phi(0,0) = N(0,0) = 0, D_X\Phi(0,0) = D_X N(0,0) = 0$ and such that for $(\mu, X) \in \widetilde{\mathcal{U}}_p$ the change of variable*

$$X = \widetilde{X} + \Phi(\mu, \widetilde{X})$$

transforms (4) into the following system which has the same regularity in $(\mu, \widetilde{X})$:

$$\frac{d\widetilde{X}}{dx} = L_0\widetilde{X} + N(\mu, \widetilde{X}) + R(\mu, \widetilde{X}) \tag{5}$$

where N is characterized by

$$N(\mu, e^{L_0^* x}X) = e^{L_0^* x}N(\mu, X), \forall x \in \mathbb{R}, \forall X \in E_0, \forall \mu \text{ near } 0,$$

and $R(\mu, \widetilde{X}) = o(||X||^p)$. In addition, (5) inherits the symmetries of (2).

This theorem provides an additional symmetry to nonlinear "simplified" terms, this symmetry only resulting from the linearized operator! The proof which includes the parameter dependance of the polynomial coefficients and the optimal estimate on the rest R, is quite technical (see hint in [9], and [12]).

1.2 STUDY OF SOME REVERSIBLE NORMAL FORMS

Let us restrict our attention to systems such that 0 stays solution of (2) for $\mu \neq 0$. This eliminates some cases which are not of interest here. Now, because of reversibility, we know that the eigenvalues of $D_U F(\mu, 0)$ are symmetric with respect to both axis, hence theorem 1 indicates that bifurcation situations may occur at

least when some eigenvalues meet (by pairs) the imaginary axis. The simplest case is when L_0 has only a double 0 eigenvalue on the imaginary axis. This leads to a 2 dimensional center manifold for the study of small bounded solutions of (4). We give below some details only on the next most important cases, i.e. when

(i) L_0 has only a double 0 and a pair of simple pure imaginary eigenvalues on the imaginary axis;

(ii) L_0 has only a pair of double pure imaginary eigenvalues on the imaginary axis.

Notice that case (ii) was introduced by Y.Rocard (see chapter I.14 of [27]) when he presents the instability "par confusion de fréquences propres", which occurs in the phenomenon of the fluttering of a wing (submitted to the aerodynamic forcing of a big wind), particularly dangerous for planes and for long suspended bridges.

1.2.1 Case (i)

Here the center manifold is four dimensional. Let us denote by $\pm iq$ the pair of simple eigenvalues and decompose $X = A\xi_0 + B\xi_1 + C\zeta + \overline{C\zeta}$, where (A, B) are real amplitudes, C a complex one and $L_0\xi_0 = 0, L_0\xi_1 = \xi_0, L_0\zeta = iq\zeta$.

Then we need to know how the reversiblity symmetry S_0 acts on $(A, B, C, \overline{C})$. There are two theoretical possibilities, depending on whether $S\xi_0 = \xi_0$ or $-\xi_0$. In most physical problems we have the first case, so $S_0 : (A, B, C, \overline{C}) \to (A, -B, \overline{C}, C)$ and after parameter dependent rescaling, the normal form, truncated at quadratic order, reads

$$\begin{cases} \frac{dA}{dx} = B \\ \frac{dB}{dx} = \mu A + A^2 + c|C|^2, \\ \frac{dC}{dx} = iC(q + d_1\mu + d_2 A), \end{cases} \tag{6}$$

where $c = \pm 1$ and (real) coefficients d_j can be explicitely computed (see [14] for a proof of (6) and the computation of principal part of coefficients on a specific physical problem). This system is *integrable*, with the two first integrals

$$K = |C|^2, H = B^2 - (2/3)A^3 - \mu A^2 - 2cKA, \tag{7}$$

and we show at figure 1 the various graphs of functions

$$f_{\mu,K,H}(A) = (2/3)A^3 + \mu A^2 + 2cKA + H$$

depending on (K, H), for $\mu > 0$. In this case, we have, in addition to the trivial equilibrium, another "conjugate" equilibrium, and several types of periodic solutions, quasi-periodic solutions (interior of the triangular region in (K, H) plane, and homoclinic solutions, one homoclinic to 0, and all others homoclinic to one of the periodic solutions.

We represent on figure 2, in the (A, B) plane all bounded solutions for $cK < \mu^2/4$. Notice that the homoclinic solution to A_+ corresponds here to a solution

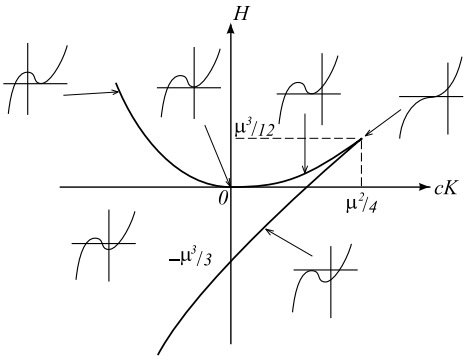


Figure 1: case (i). Graphs of $f_{\mu,K,H}(A)$ for $\mu > 0$.

homoclinic to a periodic solution since $K \neq 0$. Notice that $A_+ \sim -cK/\mu$ when $|K| \lll |\mu|$, meaning that oscillations at ∞ are very small in this case. For $K = 0$ this corresponds to a *solution homoclinic to 0*, even though the stable and unstable manifolds of 0 are only one dimensional (in the 4 dim space!). We shall see in next section that this solution does not exist in general for the full system (5), even though one may compute its expansion in powers of the bifurcation parameter μ up to any order.

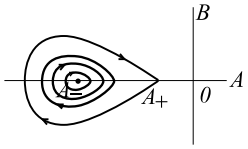


Figure 2: case (i). Bounded solutions of (6) for various values of H in the (A, B) plane, for $\mu > 0, cK < \mu^2/4$. $A_\pm = 1/2(-\mu \pm \sqrt{\mu^2 - 4cK})$.

An analogous study holds for $\mu < 0$ using $f_{\mu,H,K}(A) = -f_{-\mu,-H,K}(-A))$.

1.2.2 CASE (II)

Here the center manifold is again four dimensional. Let us denote by $\pm iq$ the pair of double eigenvalues at criticallity, and define by (A, B) the complex amplitudes corresponding respectively to the eigenmode and to the generalized eigenmode. This case is often denoted by "1:1 reversible resonance". We can always assume that the reversibility symmetry S_0 acts as: $(A, B) \longmapsto (\overline{A}, -\overline{B})$. The normal form,

at any order, reads (see a proof in [12]):

$$\frac{dA}{dx} = iqA + B + iAP[\mu, |A|^2, i/2(A\overline{B} - \overline{A}B)], \tag{8}$$
$$\frac{dB}{dx} = iqB + iBP[\mu, |A|^2, i/2(A\overline{B} - \overline{A}B)] + AQ[\mu, |A|^2, i/2(A\overline{B} - \overline{A}B)],$$

where $P(\mu, \cdot, \cdot)$ and $Q(\mu, \cdot, \cdot)$ are real polynomials. Let us define more precisely the coefficients of Q, for the cubic normal form [N of degree 3 in (5)]: $Q(\mu, u, v) = \mu + q_2 u + q_3 v$, where q_2 may be taken as ± 1, after a parameter dependent rescaling. This means that for $\mu > 0$ the eigenvalues are at a distance $\sqrt{\mu}$ from the imaginary axis, while, for $\mu < 0$, they sit on the imaginary axis. The explicit computation of the principal parts of coefficients of polynomials P and Q is made for instance in [7] on a specific physical example. The vector field (8) is integrable, with the two following first integrals:

$$K = i/2(A\overline{B} - \overline{A}B), \quad H = |B|^2 - \int_0^{|A|^2} Q[\mu, u, K]du.$$

It is then possible to describe all small bounded solutions of (8), and to discuss the various types of solutions in the (K, H) plane, for $\mu > 0$, or $\mu < 0$ (see [17]). We obtain families of periodic and quasi-periodic solutions and, for $\mu > 0, q_2 < 0$ a "circle" of solutions homoclinic to 0, for $H = K = 0$, due to the $SO(2)$ invariance of the normal form, while for $\mu < 0, q_2 > 0$ we have a family (curve in the (H, K) plane) of "circles" of solutions homoclinic to periodic solutions (as in case (i)) (the amplitude is here minimum at $x = 0$).

1.3 Typical persistence results

In section 1.2, we investigated the normal forms, i.e. equation (5) with no remaining term R, and we obtained various type of solutions that we would like to be persistent for the complete problem (5). The problem consists now in proving persistence results. In summary, the *persistence of periodic solutions* of the normal form can in general be performed, through an adaptation of the Lyapunov-Schmidt technique (see [14],[22]). The *persistence of quasi-periodic solutions* is much more delicate, and can only be performed in a subset of the (H, K) plane, where these solutions exist for the normal form. Typically, it is proved for case (i), that for any fixed μ, quasi-periodic solutions exist on a subset of the interior of the triangular region of figure 1, which is locally the cartesian product of a curve with a Cantor set (see a complete proof in [16] for case (ii), and see [14] for case (i), both applied to specific examples in fluid mechanics). The persistence of solutions homoclinic to periodic solutions, provided that they are not too small, needs some technicality, see for instance [14] for case (i) and [17] for case (ii). The same results holds for solutions homoclinic to 0 in case (ii). In fact one can prove the persistence of two symmetric (reversible) solutions (instead of a a full circle of solutions), using a transversality argument (intersection of the stable manifold of a periodic orbit (or of the fixed point in case (ii) for $\mu < 0$) with the subspace of symmetric

points), after controlling the size of the perturbation due to R, which applies for $x \in [0, +\infty)$.

Now, for the normal form of case (i), there is a family of orbits homoclinic to periodic solutions whose amplitudes can be chosen arbitrarily small. It can be proved (see Lombardi [22] for a complete proof) that there are two families of reversible solutions homoclinic to a periodic solution whose size may be chosen arbitrary, until a (non zero) *exponentially small size* $\mu^{-1}e^{-c/\sqrt{\mu}}$ (smaller than any power of the bifurcation parameter μ). The method used by Lombardi consists in a complete justification of a matching asymptotic expansion method of the solution which is extended in a strip of the complex plane, where the singularity in the complex plane originates from one of the homoclinic solution of the truncated normal form (6). Moreover, despite of the fact that a solution homoclinic to 0 exists for the normal form (6), this is not true in general for the full system (5) (see [23]), even though one can compute an asymptotic expansion up to any order of such an homoclinic (non existing) "solution"! This non obvious result says in particular that we cannot avoid small oscillations at infinity in this case.

2 Application to the water wave problem

Let us consider the case of one layer (thickness h) of an inviscid fluid, the flow is assumed potential, under the influence of gravity g and surface tension T acting at the free surface (see left of figure 3). We are interested in steady waves of permanent form, i.e. travelling waves with constant velocity c. Formulating the problem in a moving reference frame, our solutions are steady in time, and we intend to consider the unbounded horizontal coordinate ξ as a "time". Let us denote by ρ the fluid density, then we choose c as the velocity scale and $l = T/\rho c^2$ as the length scale. The important dimensionless parameters occuring in the equations are $\lambda = ghc^{-2}$, $b = T(\rho hc^2)^{-1} = l/h$.

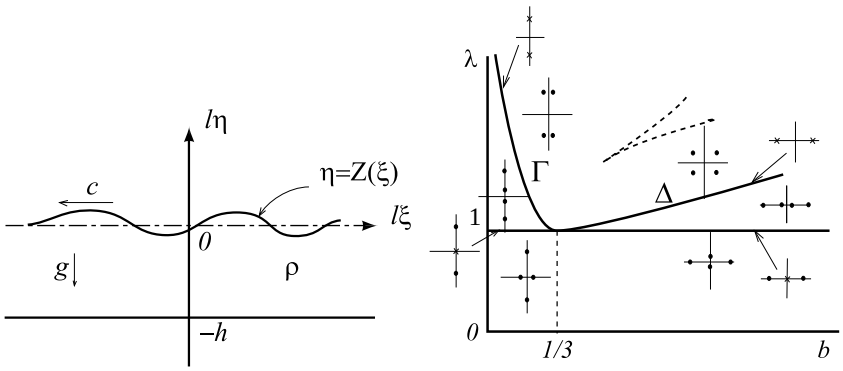


Figure 3: Left: geometric configuration of the water-wave problem. Right: positions of the 4 critical eigenvalues of L_μ in function of $\mu = (b, \lambda)$.

A nice formulation of this problem uses a change of coordinates introduced by Levi-Civita [21]. He uses the coordinates (x, y) defined by the complex potential $w(\xi + i\eta) = x + iy$ and unknown are α and β defined by $w'(\xi + i\eta) = e^{-i(\alpha+i\beta)}$ (complex velocity). The free surface is given by $y = 0$, the rigid bottom by $y = -1/b$. The physical free surface is given by $\eta = Z(\xi) = \widetilde{Z}(x) = \int_{-1/b}^{0}(e^{-\beta}\cos\alpha - 1)dy$. In our formulation, the unknown is $[U(x)](y) = (\alpha_0(x), \alpha(x, y), \beta(x, y))^t$ and the system has the form

$$\frac{dU}{dx} = F(\mu, U) = \left\{ \begin{array}{c} \sinh\beta_0 + \lambda b e^{-\beta_0}\int_{-1/b}^{0}(e^{-\beta}\cos\alpha - 1)dy \\ \frac{\partial\beta}{\partial y}, \\ -\frac{\partial\alpha}{\partial y} \end{array} \right\} -1/b < y < 0. \tag{9}$$

where $\mu = (b, \lambda)$, and equation (9) has to be understood in the space $\mathbb{H} = \mathbb{R} \times \{L^1(-1/b, 0)\}^2$, and $U(x)$ lies in $\mathbb{D} = \mathbb{R} \times \{W^{1,1}(-1/b, 0)\}^2 \cap \{\alpha_0 = \alpha|_{y=0}, \alpha|_{y=-1/b} = 0\}$, where we denote by β_0 the trace $\beta|_{y=0}$ and by $W^{1,1}(-1/b, 0)$ the space of integrable functions with an integrable first derivative on the interval $(-1/b, 0)$. A solution of our water-wave problem is any $U \in \mathcal{C}^0(\mathbb{D}) \cap \mathcal{C}^1(\mathbb{H})$ which is solution of (9), where (e.g.) $\mathcal{C}^0$ means continuous and bounded for $x \in \mathbb{R}$.

It is clear that $U = 0$ is a particular solution of (9), which corresponds to the flat free surface state. A very important property of (9) is its *reversibility*: indeed let us define the symmetry S: $SU = (-\alpha_0, -\alpha, \beta)^t$, then it is easy to see that the linear operator S *anticomutes with* $F(\mu, \cdot)$. This reflects the invariance under reflexion symmetry $\xi \to -\xi$ of our original problem.

REMARK 3 *There is a large class of water-wave problems which can be treated in a similar way: one may consider several layers of non miscible perfect fluids, and consider cases with or without surface (or interface) tension (see [11] for these formulations).*

Since we are interested in solutions near 0, it is natural to study the problem obtained after linearization near 0. We then define the linear operator $L_\mu = D_U F(\mu, 0)$, unbounded and closed in $\mathbb{H}$. In all problems, for layers with finite depth , it can be shown that the spectrum of L_μ which is *symmetric with respect to both axis* of the complex plane because of reversibility, is only composed of isolated eigenvalues of finite multiplicities, accumulating only at infinity. More precisely, denoting by ik these eigenvalues (not necessary pure imaginary), then one has the classical "dispersion relation" for solving the eigenvalues, under the form of a complex equation $f(\mu, k) = 0$. For problem (9), we obtain the following dispersion relation:

$$(\lambda b + k^2)k^{-1}\sinh k/b - \cosh k/b = 0, \quad \text{for } k \neq 0. \tag{10}$$

There is no more than 4 eigenvalues on (or close to) the imaginary axis, the rest of them being located in a sector $(ik \in \mathbb{C}; |k_r| < p|k_i| + r)$ of the complex plane (see right side of figure 3). There is a codimension 2 case when $(b, \lambda) = (1/3, 1)$, where 0 is a quadruple eigenvalue. The roots of the dispersion equation give the

poles of the resolvent operator $(ik\mathbb{I} - L_\mu)^{-1}$. In addition, we obtain an estimate of the form

$$||(ik\mathbb{I} - L_\mu)^{-1}||_{\mathcal{L}(\mathbb{H})} \leq c/|k|, \tag{11}$$

where $c > 0$ is fixed and for large enough $|k|$, and where $\mathcal{L}(\mathbb{H})$ is the space of bounded linear operators in $\mathbb{H}$. The choice of the basic space $\mathbb{H}$ should be appropriate for finding the good estimate (11) of the resolvent, this is a little delicate for problems with several layers and no surface tension (see [11]). This estimate is essential in our method of reduction to a center manifold.

For the study of the nonlinear problem (9) the idea is now to use a *center manifold* reduction like in section 1, which leads to *an ordinary differential equation* of dimension at most 4 in the present problem. Let us assume that, for μ near μ_0, the eigenvalues of L_μ are contained either in a small vertical strip of width tending towards 0 for $\mu \to \mu_0$, or at a distance of order 1 from the imaginary axis, then the estimate (11) allows us to find such a center manifold as in finite dimensional case (see [20], [24], [29]). Roughly speaking, all "small" bounded continuous solutions taking values in $\mathbb{D}$, of the system (9) for values of μ near μ_0, lie on an invariant manifold M_μ which is smooth (however loosing the C^∞ regularity) and which exists in a neighborhood of 0 independent of μ (depending on the required smoothness). The dimension of M_μ is equal to the sum of dimensions of invariants subspaces belonging to pure imaginary eigenvalues occuring for the critical value μ_0 of the parameter. In addition, the trace of system (9) on M_μ is also *reversible under the restriction* S_0 of the symmetry S. It results, in particular that the study we made at sections 1.2 and 1.3 applies here (after a suitable choice of the bifurcation parameter). The situation near the set $\lambda = 1, b < 1/3$ was studied first in [1] and [25] in an uncomplete way. Here case (i) applies (not too close to the codimension 2 point, since we would need to use another normal form there (see [10] for such a study). Denoting by $\mu = \lambda - 1$, it is shown in [14] that we are in situation of figure 1, with $c > 0$. The study of unavoidable exponentially small oscillations at infinity was first studied directly on the water wave problem in [3] and [26], and as a general property for a large class of problems in [22]. The generic non existence of solitary waves in this case follows from [23]. Now, the study made for case (ii) applies near the curve Γ of figure 3 (right). For problem (9) it is shown that coefficient q_2 *is negative*. For other water wave problems with more than one layer, this coefficient may change of sign, which leads to new types of solutions near this singular case. In the present problem, we then have for (b, λ) slightly above the curve Γ, the *bifurcation of two reversible solitary waves, with exponentially damping oscillations at infinity* [13].

REMARK 4 *Such reversible bifurcations in function of 2 parameters also appear in various physical problems. A very nice example is in the study of localized stuctures for long (assumed infinitely long) rubber rods subject to end tension and moment! The basic state is the straight rod. The study of eigenvalues of the linearized operator lead to a picture analogue to figure 3 (right). In particular, the two homoclinic orbits above, become four because of an extra symmetry of the problem, and are physically important in the study of buckling of such rods (see [5]).*

3 Physical relevance - infinite depth problem

A common point for the various water wave problems, is that when the bottom layer thickness grows ($b \to 0$ in (10)), there is an *accumulation of eigenvalues on the whole real axis*, and at the limit, as we choose a space $\mathbb{D}$ where we replace $1/b$ by ∞ and suppress the boundary condition at $y = -1/b$, all real eigenvalues disappears, leaving the place to the *entire real axis forming the essential spectrum*: for σ real $\neq 0$ the operator $(\sigma\mathbb{I} - L_\mu)$ is not Fredholm [18]: it is one-to-one, but its range is not closed and its closure has a non zero finite codimension (see [15],[11]).

At this point we should emphasize that the physical relevance of the center manifold reduction for the finite depth problem is linked with the distance of the rest of (non critical) eigenvalues from the imaginary axis. So, the validity of the bifurcation analysis is becoming empty when the thickness of the layer increases. To fix ideas, let us give some physical numerical values for air-water free surface waves. The point $(b, \lambda) = (1/3, 1)$ then corresponds to a thickness $h = 0.48\ cm$, and a velocity of waves $c = 21.6\ cm/s$. This means that a layer with thickness more than few centimeters leads to a spectrum with real eigenvalues very close to 0, so the analysis which might be done (as in previous section) near the curve Γ (right of figure 3) for λb near $1/4$ on the upper branch) would be valid only in a very tiny neighborhood of this curve, and *this analysis would have no physical interest*. We need to study the *worse limiting case, which is here the infinite depth case*, and physical cases are in fact considered as regular perturbations of this limiting case. We shall see below that this has dramatic consequences on the mathematical analysis!

For the limiting problem the dispersion relation (10) has at most 4 roots. There is a pair of two pure imaginary double eigenvalues for $\lambda b = 1/4$. The remaining of the spectrum of L_μ is formed by the full real line, hence it crosses the imaginary axis at 0, and we *cannot use the center manifold reduction*. However, we still have the resolvent estimate (11), due to a good choice of space $\mathbb{H}$. In particular this type of results is also valid for problems with several layers, one being of infinite depth, with an additional eigenvalue in 0 (embedded in the essential spectrum), when there is no surface tension at one of the free surfaces [11].

3.1 Normal forms in infinite dimensions

Since we cannot reduce our problems to finite dimensional ODE's, and since we still would like to believe that eigenvalues near the imaginary axis are ruling the bounded solutions, this is a motivation for developing a theory of normal forms in separating the finite dimensional critical space, from the rest (the "hyperbolic" part of the spectrum, including 0). This leads to "partial normal forms", where there are *coupling terms,* specially in the infinite dimensional part of the system (see [15],[8]). For developing this theory, there are some technical difficulties, specially for problems with more than one layer and no surface tension at some free surface. A first difficulty is due to cases where 0 is an eigenvalue embedded in the essential spectrum: for extracting it from the spectrum, we use the *explicit form of the resolvent operator near the real axis*, to explicitely obtain the continuous

linear form which can be used for the projection on the eigenspace belonging to 0. A second difficulty is that in space $\mathbb{H}$ the linear operator has not an "easy" (even formal) adjoint. This adjoint and some of its eigenvectors are usually necessary for expressing projections on the critical finite dim space. Fortunately, in our problems, we use the explicit form of the resolvent operator near the (for example double) eigenvalues, to make explicit the projection commuting with the linear operator (see [18]).

3.2 Typical results

Since we have not yet a center manifold reduction process to a finite ODE, the method we use now, needs to give *a priori* the type of solution, we are looking for. This is a major difference with the cases we had before, for finite depth layers. For periodic solutions, we use an adaptation of Lyapunov-Schmidt method, except that the presence of 0 in the spectrum gives some trouble (resonant terms). It appears that we can formulate all these problems, such that there is no such resonant term for *reversible solutions* (symmetric under S). As a result, there are *as many periodic solutions as in the finite depth problem* [11]. For solutions homoclinic to 0 (*solitary waves*), for example in our one layer problem, we first derive the infinite dimensional normal form, then we inverse the infinite dimensional part of the system, using Fourier transform. Indeed, the linearized Fourier transform uses the above resolvent operator, where we eliminated, via a suitable projection, the poles given by eigenvalues sitting on the imaginary axis. The fact that the resolvent operator is not analytic near 0 (there is a jump of the resolvent in crossing the real axis [15]), leads to the fact that this "hyperbolic part" of the solution *decays polynomially at infinity.* Putting this solution into the four dimensional part of the system, we can solve as before except that the decay of solutions is now polynomial (as $1/x^2$), instead of exponential. The principal part of the solution (of order $(\lambda b - 1/4)^{1/2}$) at finite distance still comes from the four dimensional truncated normal form, but its decays faster at infinity than the other part of the solution, which makes this queue part predominant at infinity. This is the main difference with the finite depth case, where the principal part coming from the normal form is valid for all values of x (see [15] for the proofs related with problem (9)).

As a conclusion, let us just say that I present here a specific type of physical problems which motivate some developments of existing mathematical theories. It also gives motivation for finding a new tool, probably very difficult to produce, like a center manifold reduction in cases when a continuous part of the spectrum crosses the imaginary axis. This is another illustration of the fact that progresses in mathematics may come from non academic questions raised naturally from discussions and collaboration with other disciplines.

REFERENCES

[1] C.J.Amick, K.Kirchgässner. Arch. rat. Mech. Anal. 105, 1-49, 1989.

[2] V.Arnold. Geometrical methods in the theory of ordinary differential equations. Springer, New York, 1983

[3] J.T.Beale. Comm. Pure Appl. Math. 44, 211-257, 1991.

[4] G.Belitskiĭ. Trans. Mosc. Math. Soc., 40, 1-39, 1981.

[5] A.Champneys, J.Thompson. Proc. Roy. Soc. Lond. A, 452, 2467-2491, 1996.

[6] R.Cushman & J.Sanders. J.Dyn. and Stab. of Systems, 2, 4, 235-246, 1988.

[7] F.Dias, G.Iooss. Physica D 65, 399-423, 1993.

[8] F.Dias, G.Iooss. Eur. J. Mech. B/Fluids 15, 3, 367-393, 1996.

[9] C.Elphick, E.Tirapegui, M.Brachet, P.Coullet, G.Iooss. Physica D29, 95-127, 1987

[10] G.Iooss. Fields Inst. Com. 4, 201-217, 1995.

[11] G.Iooss. Preprint INLN 1998.

[12] G.Iooss, M.Adelmeyer. Topics in bifurcation theory and Applications. Adv. series in Nonlinear Dynamics, 3, World Sci.1992.

[13] G.Iooss, K.Kirchgässner. Note C.R.Acad. Sci. Paris 311, I, 265-268, 1990.

[14] G.Iooss, K.Kirchgässner. Proc. Roy. Soc. Edinburgh, 122A, 267-299, 1992.

[15] G.Iooss, P.Kirrmann. Arch. Rat. Mech. Anal. 136, 1-19, 1996.

[16] G.Iooss, J.Los. Nonlinearity 3, 851-871, 1990.

[17] G.Iooss, M.C.Pérouème. J.Diff. Equ. 102, 62-88, 1993.

[18] T.Kato. Perturbation theory for linear operators. Springer Verlag, 1966.

[19] A.Kelley. J.Diff. equ. 3, 546-570, 1967.

[20] K.Kirchgässner. J.Diff. Equ. 45, 113-127, 1982.

[21] T.Levi-Civita. Math. Annalen, 93, 264-314, 1925.

[22] E.Lombardi. Arch. Rat. Mech. Anal. 137, 227-304, 1997.

[23] E.Lombardi J.Dyn. Diff. Equ. (to appear).

[24] A.Mielke. Math. Meth. Appl. Sci. 10, 51-66, 1988.

[25] S.M.Sun. J. Math. Anal. Appl. 156, 471-504, 1991.

[26] S.M.Sun, M.C.Shen. J.Math. Anal. Appl. 172, 533-566, 1993.

[27] Y.Rocard. Dynamique générale des vibrations. 4ème édition. Masson, 1971.

[28] A.Vanderbauwhede. Dynamics Reported 2, 89-169, 1989.

[29] A.Vanderbauwhede, G.Iooss. Dynamics reported, 1 new series, 125-163, 1992.

Gérard Iooss
Institut Universitaire de France
INLN UMR CNRS-UNSA 6618
1361 route des Lucioles
F-06560 Valbonne
e-mail: iooss@inln.cnrs.fr

Exact Relations for Composites: Towards a Complete Solution

Yury Grabovsky and Graeme W. Milton[1]

Abstract. Typically, the electrical and elastic properties of composite materials are strongly microstructure dependent. So it comes as a nice surprise to come across exact formulae for (or linking) effective tensor elements that are universally valid no matter what the microstructure. Here we present a systematic theory of exact relations embracing the known exact relations and establishing new ones. The search for exact relations is reduced to a search for tensor subspaces satisfying certain algebraic conditions. One new exact relation is for the effective shear modulus of a class of three-dimensional polycrystalline materials.

1991 Mathematics Subject Classification: Primary 35B27, 73B27, 73S10; Secondary 73B40, 49J45
Keywords and Phrases: Composites, homogenization, polycrystals, exact relations

Introduction

Take a metal rod. We can bend it, twist it, stretch it, vibrate it or use it as a conduit for the flow of electrons or heat. It looks just like a homogeneous material with behavior governed by bulk and shear elastic moduli and electrical and thermal conductivities. However if we break the metal rod there is a surprise! One can see that the surface of the break is rough, comprised of individual crystalline grains sparkling in the light. Similarly foam rubber behaves like a highly compressible homogeneous elastic material, even though its pore structure is quite complicated. Homogenization theory provides a rigorous mathematical basis for the observation that materials with microstructure can effectively behave like homogeneous materials on a macroscopic scale. A typical result is the following. To ensure ellipticity of the equations let us suppose we are given positive constants α and $\beta > \alpha$ and

[1]Both authors are supported by the National Science Foundation through grants DMS9402763, DMS9629692, DMS9704813 and DMS-9803748.

a periodic conductivity tensor field $\boldsymbol{\sigma}(\boldsymbol{x})$ taking values in the set $\mathcal{M}_c$ comprising of all $d \times d$ symmetric matrices $\boldsymbol{\sigma}$ satisfying

$$\alpha \boldsymbol{v} \cdot \boldsymbol{v} \leq \boldsymbol{v} \cdot \boldsymbol{\sigma} \boldsymbol{v} \leq \beta \boldsymbol{v} \cdot \boldsymbol{v}, \tag{1}$$

for all vectors $\boldsymbol{v}$. Then with $\boldsymbol{\sigma}_\epsilon(\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{x}/\epsilon)$ the electrical potential $\phi_\epsilon(\boldsymbol{x})$ which solves the Dirichlet problem

$$\nabla \cdot \boldsymbol{\sigma}_\epsilon(\boldsymbol{x}) \nabla \phi_\epsilon(\boldsymbol{x}) = f(\boldsymbol{x}) \ \text{ within } \ \Omega, \quad \phi_\epsilon(\boldsymbol{x}) = \psi(\boldsymbol{x}) \ \text{ on } \partial\Omega, \tag{2}$$

converges as $\epsilon \to 0$ (i.e. as the length scale of the periodicity of $\boldsymbol{\sigma}_\epsilon(\boldsymbol{x})$ shrinks to zero) to the potential ϕ_0 which solves

$$\nabla \cdot \boldsymbol{\sigma}_* \nabla \boldsymbol{\phi}_0(\boldsymbol{x}) = f(\boldsymbol{x}) \ \text{ within } \ \Omega, \quad \phi_0(\boldsymbol{x}) = \psi(\boldsymbol{x}) \ \text{ on } \partial\Omega, \tag{3}$$

where the effective conductivity tensor $\boldsymbol{\sigma}_*$ is in $\mathcal{M}_c$ and only depends on $\boldsymbol{\sigma}(\boldsymbol{x})$ and not upon the choice of Ω, the source term $f(\boldsymbol{x})$, nor upon the potential $\psi(\boldsymbol{x})$ prescribed at the boundary. The effective conductivity tensor $\boldsymbol{\sigma}_*$ is obtained by solving the following *cell-problem*. One looks for periodic vector fields $\boldsymbol{j}(\boldsymbol{x})$ and $\boldsymbol{e}(\boldsymbol{x})$, representing the current and electric fields, which satisfy

$$\boldsymbol{j}(\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{x}) \boldsymbol{e}(\boldsymbol{x}), \qquad \nabla \cdot \boldsymbol{j} = 0, \qquad \nabla \times \boldsymbol{e} = 0. \tag{4}$$

The relation $\langle \boldsymbol{j} \rangle = \boldsymbol{\sigma}_* \langle \boldsymbol{e} \rangle$ between the average current and electric fields serves to define $\boldsymbol{\sigma}_*$. Here, as elsewhere, the angular brackets will be used to denote volume averages over the unit cell of periodicity. Homogenization results extend to fields $\boldsymbol{\sigma}_\epsilon(\boldsymbol{x})$ taking values in $\mathcal{M}_c$ which are locally periodic, or random and stationary, or simply arbitrary: see Bensoussan, et. al. (1978), Zhikov, et. al. (1994), and Murat and Tartar (1997) and references therein.

Similar results hold for elasticity. Given positive constants α and $\beta > \alpha$ and a periodic elasticity tensor field $\boldsymbol{\mathcal{C}}(\boldsymbol{x})$ taking values in the set $\mathcal{M}_e$ comprised of all elasticity tensors $\boldsymbol{\mathcal{C}}$ satisfying

$$\alpha \boldsymbol{A} \cdot \boldsymbol{A} \leq \boldsymbol{A} \cdot \boldsymbol{\mathcal{C}} \boldsymbol{A} \leq \beta \boldsymbol{A} \cdot \boldsymbol{A}, \tag{5}$$

for all symmetric $d \times d$ matrices $\boldsymbol{A}$, there is an associated effective elasticity tensor $\boldsymbol{\mathcal{C}}_*$ in $\mathcal{M}_e$. It is obtained by looking for periodic symmetric matrix valued fields $\boldsymbol{\tau}(\boldsymbol{x})$ and $\boldsymbol{\epsilon}(\boldsymbol{x})$, representing the stress and strain fields, which satisfy

$$\boldsymbol{\tau}(\boldsymbol{x}) = \boldsymbol{\mathcal{C}}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{x}), \qquad \nabla \cdot \boldsymbol{\tau} = 0, \quad \boldsymbol{\epsilon} = [\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T]/2, \tag{6}$$

in which $\boldsymbol{u}(\boldsymbol{x})$ represents the displacement field. The relation $\langle \boldsymbol{\tau} \rangle = \boldsymbol{\mathcal{C}}_* \langle \boldsymbol{\epsilon} \rangle$ between the average stress and strain fields serves to define $\boldsymbol{\mathcal{C}}_*$.

A key problem, of considerable technological importance, is to determine the effective tensors $\boldsymbol{\sigma}_*$ and $\boldsymbol{\mathcal{C}}_*$ governing the behaviour on the macroscopic scale. For a long while it was the dream of many experimentalists and theorists alike that there should be some universally applicable "mixing formula" giving the effective tensors as some sort of average of the tensors of the crystalline grains or

constituent materials. However the reality is that the details of the microgeometry can sometimes play an influential role in determining the overall properties, particularly when the crystalline grains have highly anisotropic behavior or when there is a large contrast in the properties of the constituent materials. Consider, for example, a two-phase composite where one phase is rigid and the second phase is compressible. The question of whether the composite as a whole is rigid or compressible is not solely determined by the volume fractions occupied by the phases, but depends on whether the rigid phase has a connected component spanning the material or consists of isolated inclusions embedded in the compressible phase.

So we have to temper the dream. Instead of seeking a universally applicable "mixing formula" one can ask whether certain combinations of effective tensor elements can be microstructure independent. Indeed they can. Sometimes these exact relations are easy to deduce and sometimes they are not at all obvious. Such exact relations provide useful benchmarks for testing approximation schemes and numerical calculations of effective tensors. Grabovsky (1998) recognized that there should be some general theory of exact relations. Utilizing the fact that an exact relation must hold for laminate materials he derived restrictive constraints on the form that an exact relation can take. This reduced the search for candidate exact relation to an algebraic question that was analysed by Grabovsky and Sage (1998). Here we give sufficient conditions for an exact relation to hold for all composite microgeometries, and not just laminates. At present the general theory of exact relations is still not complete. There is a gap between the known necessary conditions and the known sufficient conditions for an exact relation to hold. In addition the associated algebraic questions have only begun to be investigated. Before proceeding to the general theory let us first look at some examples: see also the recent review of Milton (1997).

Examples of some elementary exact relations

An example of a relation which is easy to deduce is the following. Lurie, Cherkaev and Fedorov (1984) noticed that if the elasticity tensor field $\boldsymbol{\mathcal{C}}(\boldsymbol{x})$ is such that there exist non-zero symmetric tensors $\boldsymbol{V}$ and $\boldsymbol{W}$ with $\boldsymbol{\mathcal{C}}(\boldsymbol{x})\boldsymbol{V} = \boldsymbol{W}$ for all $\boldsymbol{x}$ then the effective tensor $\boldsymbol{\mathcal{C}}_*$ must satisfy $\boldsymbol{\mathcal{C}}_*\boldsymbol{V} = \boldsymbol{W}$. The reason is simply that the elastic equations are solved with a constant strain $\boldsymbol{\epsilon}(\boldsymbol{x}) = \boldsymbol{V}$ and a constant stress $\boldsymbol{\tau}(\boldsymbol{x}) = \boldsymbol{W}$ and the effective tensor, by definition, relates the averages of these two fields. In particular, consider a single phase polycrystalline material, where the crystalline phase has cubic symmetry. Each individual crystal responds isotropically to hydrostatic compression, and we can take $\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{W} = d\kappa_0\boldsymbol{I}$ where d is the spatial dimension (2 or 3) and κ_0 is the bulk modulus of the pure crystal. The result implies that the effective bulk modulus κ_* of the polycrstal is κ_0 (Hill, 1952). Another way of expressing this exact relation is to introduce the manifold

$$\mathcal{M} = \mathcal{M}(\boldsymbol{V}, \boldsymbol{W}) = \{\boldsymbol{\mathcal{C}} \in \mathcal{M}_e \mid \boldsymbol{\mathcal{C}}\boldsymbol{V} = \boldsymbol{W}\}, \tag{7}$$

of elasticity tensors. The exact relation says that if $\boldsymbol{\mathcal{C}}(\boldsymbol{x}) \in \mathcal{M}$ for all $\boldsymbol{x}$ then $\boldsymbol{\mathcal{C}}_* \in \mathcal{M}$. In other words the manifold $\mathcal{M}$ is stable under homogenization. It defines an exact relation because it has no interior. Many other important exact

relations derive from uniform field arguments: see Dvorak and Benveniste (1997) and references therein.

The classic example of a non-trivial exact relation is for two-dimensional conductivity (or equivalently for three-dimensional conductivity with microstructure independent of one coordinate). When $d = 2$ the equations (4) can be written in the equivalent form

$$\boldsymbol{j}'(\boldsymbol{x}) = \boldsymbol{\sigma}'(\boldsymbol{x})\boldsymbol{e}'(\boldsymbol{x}), \quad \nabla \cdot \boldsymbol{j}'(\boldsymbol{x}) = 0, \quad \nabla \times \boldsymbol{e}'(\boldsymbol{x}) = 0, \tag{8}$$

where

$$\boldsymbol{j}'(\boldsymbol{x}) \equiv c\boldsymbol{R}_\perp \boldsymbol{e}(\boldsymbol{x}), \quad \boldsymbol{e}'(\boldsymbol{x}) \equiv \boldsymbol{R}_\perp \boldsymbol{j}(\boldsymbol{x}), \quad \boldsymbol{\sigma}'(\boldsymbol{x}) \equiv c\boldsymbol{R}_\perp [\boldsymbol{\sigma}(\boldsymbol{x})]^{-1} \boldsymbol{R}_\perp^T. \tag{9}$$

in which c is a constant and $\boldsymbol{R}_\perp$ is the matrix for a 90° rotation. In other words the fields $\boldsymbol{j}'(\boldsymbol{x})$ and $\boldsymbol{e}'(\boldsymbol{x})$ solve the conductivity equations in a medium with conductivity $\boldsymbol{\sigma}'(\boldsymbol{x})$. Moreover by looking at the relations satisfied by the average fields one sees that the effective conductivity tensor $\boldsymbol{\sigma}'_*$ associated with $\boldsymbol{\sigma}'(\boldsymbol{x})$ and the effective conductivity tensor $\boldsymbol{\sigma}'_*$ associated with $\boldsymbol{\sigma}(\boldsymbol{x})$ are linked by the relation

$$\boldsymbol{\sigma}'_* = c\boldsymbol{R}_\perp (\boldsymbol{\sigma}_*)^{-1} \boldsymbol{R}_\perp^T, \tag{10}$$

[see Keller (1964), Dykhne (1970) and Mendelson (1975)]. Now suppose the conductivity tensor field is such that its determinant is independent of $\boldsymbol{x}$, i.e. $\det\boldsymbol{\sigma}(\boldsymbol{x}) = \Delta$. With $c = \Delta$ we have $\boldsymbol{\sigma}'(\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{x})$ implying $\boldsymbol{\sigma}'_* = \boldsymbol{\sigma}_*$. From (10) one concludes that $\det\boldsymbol{\sigma}_* = \Delta$. In other words the manifold

$$\mathcal{M} = \mathcal{M}(\Delta) = \{\boldsymbol{\sigma} \in \mathcal{M}_c \mid \det\boldsymbol{\sigma} = \Delta\} \tag{11}$$

is stable under homogenization (Lurie and Cherkaev, 1981). Again it defines an exact relation because it has no interior. An important application of this result is to a single phase polycrystalline material where the crystalline phase has a conductivity tensor with determinant Δ. If the polycrystal has an isotropic conductivity tensor the exact relation implies the result of Dykhne (1970) that $\boldsymbol{\sigma}_* = \sqrt{\Delta}\boldsymbol{I}$.

An equation satisfied by the polarization field

For simplicity, let us consider the conductivity problem and take as our reference conductivity tensor a matrix $\boldsymbol{\sigma}_0$ in $\mathcal{M}_c$. Affiliated with $\boldsymbol{\sigma}_0$ is a non-local operator $\boldsymbol{\Gamma}$ defined as follows. Given any periodic vector-valued field $\boldsymbol{p}(\boldsymbol{x})$ we say that $\boldsymbol{e}' = \boldsymbol{\Gamma}\boldsymbol{p}$ if $\boldsymbol{e}'$ is curl-free with $\langle \boldsymbol{e}' \rangle = 0$ and $\boldsymbol{p} - \boldsymbol{\sigma}_0 \boldsymbol{e}'$ is divergence-free. Equivalently, we have

$$\begin{aligned} \widehat{\boldsymbol{e}}'(\boldsymbol{k}) &= \boldsymbol{\Gamma}(\boldsymbol{k})\widehat{\boldsymbol{p}}(\boldsymbol{k}) \quad \text{for } \boldsymbol{k} \neq 0, \\ &= 0 \quad \text{when } \boldsymbol{k} = 0, \end{aligned} \tag{12}$$

where $\widehat{\boldsymbol{e}}'(\boldsymbol{k})$ and $\widehat{\boldsymbol{p}}(\boldsymbol{k})$ are the Fourier coefficients of $\boldsymbol{e}'(\boldsymbol{x})$ and $\boldsymbol{p}(\boldsymbol{x})$ and

$$\boldsymbol{\Gamma}(\boldsymbol{k}) = \frac{\boldsymbol{k} \otimes \boldsymbol{k}}{\boldsymbol{k} \cdot \boldsymbol{\sigma}_0 \boldsymbol{k}}. \tag{13}$$

Now suppose we take a polarization field $\boldsymbol{p}(\boldsymbol{x}) = (\boldsymbol{\sigma}(\boldsymbol{x}) - \boldsymbol{\sigma}_0)\boldsymbol{e}(\boldsymbol{x})$ where $\boldsymbol{e}(\boldsymbol{x})$ solves the conductivity equations. It is analogous to the polarization field introduced in dielectric problems. From the definition (13) of the operator $\boldsymbol{\Gamma}$ we see immediately that it solves the equation

$$[\boldsymbol{I} + (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)\boldsymbol{\Gamma}]\boldsymbol{p} = (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)\langle \boldsymbol{e}\rangle \quad \text{and } \langle \boldsymbol{p}\rangle = (\boldsymbol{\sigma}_* - \boldsymbol{\sigma}_0)\langle \boldsymbol{e}\rangle. \tag{14}$$

For investigating exact relations it proves convenient to use another form of these equations. We choose a fixed matrix $\boldsymbol{M}$, define the fractional linear transformation

$$W_{\boldsymbol{M}}(\boldsymbol{\sigma}) = [\boldsymbol{I} + (\boldsymbol{\sigma} - \boldsymbol{\sigma}_0)\boldsymbol{M}]^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_0), \tag{15}$$

(in which we allow for $\boldsymbol{\sigma} - \boldsymbol{\sigma}_0$ to be singular) and rewrite (14) as

$$[\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}]\boldsymbol{p} = \boldsymbol{K}\boldsymbol{v}, \quad \langle \boldsymbol{p}\rangle = \boldsymbol{K}_*\boldsymbol{v}, \tag{16}$$

where

$$\boldsymbol{K}(\boldsymbol{x}) = W_{\boldsymbol{M}}(\boldsymbol{\sigma}(\boldsymbol{x})), \quad \boldsymbol{K}_* = W_{\boldsymbol{M}}(\boldsymbol{\sigma}_*), \quad \boldsymbol{v} = \langle \boldsymbol{e}\rangle + \boldsymbol{M}\langle \boldsymbol{p}\rangle, \tag{17}$$

and $\boldsymbol{A}$ is the non-local operator defined by its action, $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{M}(\boldsymbol{p} - \langle \boldsymbol{p}\rangle) - \boldsymbol{\Gamma}\boldsymbol{p}$. The formula (16) involves the operator $\boldsymbol{K}\boldsymbol{A}$. If $\boldsymbol{q} = \boldsymbol{K}\boldsymbol{A}\boldsymbol{p}$ we have

$$\boldsymbol{q}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\neq 0} e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\boldsymbol{K}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{k})\widehat{\boldsymbol{p}}(\boldsymbol{k}), \quad \text{where } \boldsymbol{A}(\boldsymbol{k}) = \boldsymbol{M} - \boldsymbol{\Gamma}(\boldsymbol{k}), \tag{18}$$

and $\widehat{\boldsymbol{p}}(\boldsymbol{k})$ is the Fourier component of $\boldsymbol{p}(\boldsymbol{x})$.

NECESSARY CONDITIONS FOR AN EXACT RELATION

Since exact relations hold for all microstructures they must in particular hold for laminate microstructures for which the tensors and hence the fields only have variations in one direction, $\boldsymbol{n}$. This simple consideration turns out to impose very stringent constraints. Consider the conductivity problem. Let us take $\boldsymbol{M} = \boldsymbol{\Gamma}(\boldsymbol{n})$ and let $W_{\boldsymbol{n}}(\boldsymbol{\sigma})$ denote the transformation $W_{\boldsymbol{M}}(\boldsymbol{\sigma})$. When $\boldsymbol{K}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{n}\cdot\boldsymbol{x})$ (16) is easily seen to have the solution $\boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x})\boldsymbol{v}$ and $\boldsymbol{K}_* = \langle \boldsymbol{K}\rangle$ because $\boldsymbol{A}$ annihilates any field which only has oscillations in the direction $\boldsymbol{n}$. [Milton (1990) and Zhikov (1991) give related derivations of the formula $\boldsymbol{K}_* = \langle \boldsymbol{K}\rangle$: see also Backus (1962) and Tartar (1976) for other linear lamination formulae.] Since $\boldsymbol{K}_*$ is just a linear average of $\boldsymbol{K}(\boldsymbol{x})$ any set of conductivity tensors which is stable under homogenization, and hence lamination, must have a convex image under the transformation $W_{\boldsymbol{n}}$. In particular if a manifold $\mathcal{M}$ defines an exact relation, and $\boldsymbol{\sigma}_0 \in \mathcal{M}$ then $W_{\boldsymbol{n}}(\mathcal{M})$ must be convex and contain the origin. But $\mathcal{M}$ and hence $W_{\boldsymbol{n}}(\mathcal{M})$ have no interior, and a convex set with no interior must lie in a hyperplane. It follows that $W_{\boldsymbol{n}}(\mathcal{M})$ must lie in a hyperplane passing through the origin, i.e. in a subspace $\mathcal{K} = \mathcal{K}_{\boldsymbol{n}}$. Moreover, since $\mathcal{M}$ must be stable under lamination in all directions the set $W_{\boldsymbol{m}}(W_{\boldsymbol{n}}^{-1}(\mathcal{K}))$ must be a subspace for each choice of unit vector $\boldsymbol{m}$. Now given some tensor $\boldsymbol{K} \in \mathcal{K}$ and expanding $W_{\boldsymbol{m}}(W_{\boldsymbol{n}}^{-1}(\epsilon\boldsymbol{K}))$ in powers of ϵ gives

$$\begin{aligned} W_{\boldsymbol{m}}(W_{\boldsymbol{n}}^{-1}(\epsilon\boldsymbol{K})) &= \epsilon\boldsymbol{K}\{\boldsymbol{I} - [\boldsymbol{\Gamma}(\boldsymbol{n}) - \boldsymbol{\Gamma}(\boldsymbol{m})]\epsilon\boldsymbol{K}\}^{-1} \\ &= \epsilon\boldsymbol{K} + \epsilon^2\boldsymbol{K}\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K} + \epsilon^3\boldsymbol{K}\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K}\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K} + \ldots, \end{aligned} \tag{19}$$

where $\boldsymbol{A}(\boldsymbol{m})$ is given by (18) with $\boldsymbol{M} = \boldsymbol{\Gamma}(\boldsymbol{n})$. Since the linear term is $\epsilon\boldsymbol{K}$ the hyperplane $W_{\boldsymbol{m}}(W_{\boldsymbol{n}}^{-1}(\mathcal{K}))$ must in fact be $\mathcal{K}$ itself, i.e. $\mathcal{K}$ does not depend on $\boldsymbol{n}$. From an examination of the quadratic term we then see that

$$\boldsymbol{K}\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K} \in \mathcal{K} \quad \text{for all } \boldsymbol{m} \text{ and for all } \boldsymbol{K} \in \mathcal{K}. \tag{20}$$

Higher order terms in the expansion do not yield any additional constraints. Indeed substitution of $\boldsymbol{K} = \boldsymbol{K}_1 + \boldsymbol{K}_2$ in (20), where $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ both lie in $\mathcal{K}$, yields the corollary,

$$\boldsymbol{K}_1\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K}_2 + \boldsymbol{K}_2\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K}_1 \in \mathcal{K} \quad \text{for all } \boldsymbol{K}_1, \boldsymbol{K}_2 \in \mathcal{K}. \tag{21}$$

Applying this with $\boldsymbol{K}_1 = \boldsymbol{K}$ and $\boldsymbol{K}_2 = \boldsymbol{K}\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K}$ shows that the cubic term lies in the space $\mathcal{K}$. Similarly all the remaining higher order terms must also lie in $\mathcal{K}$ once (20) is satisfied. Therefore the condition (20) is both necessary and sufficient to ensure the stability under lamination of the set of all conductivity tensors in $\mathcal{M}_c \cap W_{\boldsymbol{n}}^{-1}(\mathcal{K})$.

For example, consider two-dimensional conductivity and take $\boldsymbol{\sigma}_0 = \sigma_0\boldsymbol{I}$. Then $\boldsymbol{A}(\boldsymbol{m}) = (\boldsymbol{n}\otimes\boldsymbol{n} - \boldsymbol{m}\otimes\boldsymbol{m})/\sigma_0$ is a trace-free 2×2 symmetric matrix. Now trace free 2×2 symmetric matrices have the property that the product of any three such matrices is also trace free and symmetric. So (20) will be satisfied when $\mathcal{K}$ is the space of trace free 2×2 symmetric matrices. Then $W_{\boldsymbol{n}}^{-1}(\mathcal{K})$ consists of 2×2 symmetric matrices $\boldsymbol{\sigma}_*$ such that $\mathrm{Tr}[(\sigma_0\boldsymbol{I} - \boldsymbol{\sigma}_*)^{-1}] = 1/\sigma_0$. Equivalently, it consists of matrices $\boldsymbol{\sigma}_*$ such that $\det\boldsymbol{\sigma}_* = \sigma_0^2$. This confirms that the manifold (11) is stable under lamination.

The preceeding analysis extends easily to the elasticity problem (and also to piezoelectric, thermoelectric, thermoelastic, pyroelectric and related coupled problems). Candidate exact relations are found by searching for subspaces $\mathcal{K}$ of fourth-order tensors $\mathcal{K}$ satisfying (20) where $\boldsymbol{A}(\boldsymbol{m}) = \boldsymbol{\Gamma}(\boldsymbol{n}) - \boldsymbol{\Gamma}(\boldsymbol{m})$ and $\boldsymbol{\Gamma}(\boldsymbol{k})$ is a fourth-order tensor dependent upon the choice of a reference elasticity tensor $\boldsymbol{\mathcal{C}}_0 \in \mathcal{M}_e$. In particular, for three-dimensional elasticity, if $\boldsymbol{\mathcal{C}}_0$ is elastically isotropic with bulk modulus κ_0 and shear modulus μ_0, $\boldsymbol{\Gamma}(\boldsymbol{k})$ has cartesian elements

$$\begin{aligned}\{\boldsymbol{\Gamma}(\boldsymbol{k})\}_{ij\ell m} &= \frac{1}{4\mu_0}\Big(k_i\delta_{j\ell}k_m + k_i\delta_{jm}k_\ell + k_j\delta_{i\ell}k_m + k_j\delta_{im}k_\ell - 4k_ik_jk_\ell k_m\Big)\\ &\quad + \frac{3k_ik_jk_\ell k_m}{3\kappa_0 + 4\mu_0}.\end{aligned} \tag{22}$$

Once such a subspace $\mathcal{K}$ is found the canditate exact relation is the set

$$\mathcal{M} = \mathcal{M}_e \cap W_{\boldsymbol{n}}^{-1}(\mathcal{K}), \tag{23}$$

where $W_{\boldsymbol{n}}^{-1}$ is the inverse of the transformation

$$W_{\boldsymbol{n}}(\boldsymbol{\mathcal{C}}) = [\boldsymbol{\mathcal{I}} + (\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{C}}_0)\boldsymbol{\Gamma}(\boldsymbol{n})]^{-1}(\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{C}}_0). \tag{24}$$

Using a related procedure Grabovsky and Sage (1998) found as a canditate exact relation, stable under lamination, the manifold $\mathcal{M} = \mathcal{M}(\mu_0)$ consisting of all elasticity tensors in $\mathcal{M}_e$ expressible in the form

$$\boldsymbol{\mathcal{C}} = 2\mu_0(\boldsymbol{\mathcal{I}} - \boldsymbol{I}\otimes\boldsymbol{I}) + \boldsymbol{D}\otimes\boldsymbol{D}, \tag{25}$$

for some choice of symmetric second-order tensor $\boldsymbol{D}$, in which $\boldsymbol{\mathcal{I}}$ is the fourth-order identity tensor. We will establish that this manifold $\mathcal{M}$ does in fact define an exact relation valid for all composites and not just laminates. For planar elasticity the analogous exact relation was proved by Grabovsky and Milton (1998).

SUFFICIENT CONDITIONS FOR AN EXACT RELATION

We would like to show that the manifold $\mathcal{M}$ of elasticity tensors defined by (23) is stable under homogenization and not just lamination, i.e. to ensure that any composite with elasticity tensor $\boldsymbol{\mathcal{C}}(\boldsymbol{x}) \in \mathcal{M}$ always has an effective elasticity tensor $\boldsymbol{\mathcal{C}}_* \in \mathcal{M}$. Here we will prove it is sufficient that there exist a larger space of fourth-order tensors $\overline{\mathcal{K}}$ (not necessarily self-adjoint) such that

$$\boldsymbol{K}_1\boldsymbol{A}(\boldsymbol{m})\boldsymbol{K}_2 \in \overline{\mathcal{K}} \quad \text{for all } \boldsymbol{m} \text{ and for all } \boldsymbol{K}_1, \boldsymbol{K}_2 \in \overline{\mathcal{K}}, \tag{26}$$

and such that $\mathcal{K}$ equals the subspace of all self-adjoint tensors in $\overline{\mathcal{K}}$.

To avoid confusion let us first return to the setting of the conductivity problem. To find $\boldsymbol{K}_*$ and hence $\boldsymbol{\sigma}_*$ we need to solve (16) for a set of d different values $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots \boldsymbol{v}_d$ of $\boldsymbol{v}$. Associated with each value $\boldsymbol{v}_i$ of $\boldsymbol{v}$ is a corresponding polarization field $\boldsymbol{p}_i(\boldsymbol{x})$. Let $\boldsymbol{V}$ and $\boldsymbol{P}(\boldsymbol{x})$ be the $d \times d$ matrices with the vectors $\boldsymbol{v}_i$ and $\boldsymbol{p}_i(\boldsymbol{x})$, $i = 1, 2, \ldots, d$, as columns. [Similar matrix valued fields were introduced by Murat and Tartar (1985).] Taking $\boldsymbol{V} = \boldsymbol{I}$ the set of equations (16) for $\boldsymbol{K}_*$ and the d polarization fields can be rewritten as

$$[\boldsymbol{I} - \boldsymbol{K}\boldsymbol{A}]\boldsymbol{P} = \boldsymbol{K}, \quad \langle \boldsymbol{P} \rangle = \boldsymbol{K}_*, \tag{27}$$

where now the field $\boldsymbol{Q} = \boldsymbol{K}\boldsymbol{A}\boldsymbol{P}$ is given by

$$\boldsymbol{Q}(\boldsymbol{x}) = \sum_{\boldsymbol{k} \neq 0} e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \boldsymbol{K}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{k})\widehat{\boldsymbol{P}}(\boldsymbol{k}), \tag{28}$$

in which $\widehat{\boldsymbol{P}}(\boldsymbol{k})$ is the Fourier component of $\boldsymbol{P}(\boldsymbol{x})$, and $\boldsymbol{K}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{k})$ acts on $\widehat{\boldsymbol{P}}(\boldsymbol{k})$ by matrix multiplication. The extension of this analysis to elasticity is mathematically straight-forward, but physically intriguing since in the elasticity setting $\boldsymbol{P}(\boldsymbol{x})$ is taken as a fourth-order tensor field.

Provided $\boldsymbol{K}(\boldsymbol{x})$ is sufficiently small for all $\boldsymbol{x}$, i.e. $\boldsymbol{\sigma}(\boldsymbol{x})$ is close to $\boldsymbol{\sigma}_0$, the solution to (27) is given by the perturbation expansion

$$\boldsymbol{P}(\boldsymbol{x}) = \sum_{j=0}^{\infty} \boldsymbol{P}_j(\boldsymbol{x}) \quad \text{where } \boldsymbol{P}_j = (\boldsymbol{K}\boldsymbol{A})^j\boldsymbol{K}. \tag{29}$$

Now let us suppose $\boldsymbol{K}(\boldsymbol{x})$ takes values in a tensor subspace $\overline{\mathcal{K}}$ satisfying (26). Our objective is to prove that each field $\boldsymbol{P}_j(\boldsymbol{x})$ in the perturbation expansion also takes values in $\overline{\mathcal{K}}$. Certainly the first term $\boldsymbol{P}_0(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x})$ does. Also if for some $j \geq 0$ the field $\boldsymbol{P}_j$ takes values in $\overline{\mathcal{K}}$ then its Fourier coefficients also take values in $\overline{\mathcal{K}}$ and (28) together with (26) implies that $\boldsymbol{P}_{j+1} = \boldsymbol{K}\boldsymbol{A}\boldsymbol{P}_j$ also lies in $\overline{\mathcal{K}}$. By induction it follows that every term in the expansion takes values in $\overline{\mathcal{K}}$. Provided

the perturbation expansion converges this implies that $\langle \boldsymbol{P} \rangle = \boldsymbol{K}_*$ lies in $\overline{\mathcal{K}}$. Even if the perturbation expansion does not converge, analytic continuation arguments imply the exact relation still holds provided $\boldsymbol{\sigma}(\boldsymbol{x}) \in \mathcal{M}_c$ for all $\boldsymbol{x}$, as will be shown in a forthcoming paper.

The effective shear modulus of a family of polycrystals

To illustrate the power of this method of generating exact relations, let us consider three-dimensional elasticity and prove that the manifold $\mathcal{M}$ consisting of all elasticity tensors in $\mathcal{M}_e$ expressible in the form (25) for some choice of $\boldsymbol{D}$ defines an exact relation. We take $\boldsymbol{\mathcal{C}}_0$ to be an arbitrary isotropic elastcity tensor with bulk modulus κ_0 and shear modulus μ_0. The associated tensor $\boldsymbol{\Gamma}(\boldsymbol{n})$, given by (22) has the property that $\mathrm{Tr}[\boldsymbol{\Gamma}(\boldsymbol{m})\boldsymbol{I}]$ is independent of $\boldsymbol{m}$ implying that with

$$\boldsymbol{M} = \boldsymbol{I} \otimes \boldsymbol{I}/3(3\kappa_0 + 4\mu_0), \tag{30}$$

we have

$$\mathrm{Tr}\{[\boldsymbol{M} - \boldsymbol{\Gamma}(\boldsymbol{m})]\boldsymbol{I}\} = 0 \quad \text{for all } \boldsymbol{m}. \tag{31}$$

Now consider the subspace $\overline{\mathcal{K}}$ consisting of all fourth order tensors $\boldsymbol{K}$ expressible in the form $\boldsymbol{K} = \boldsymbol{I} \otimes \boldsymbol{B} + \boldsymbol{B}' \otimes \boldsymbol{I}$ for some choice of symmetric matrices $\boldsymbol{B}$ and $\boldsymbol{B}'$. Now given symmetric matrices $\boldsymbol{B}_1$, $\boldsymbol{B}_1'$, $\boldsymbol{B}_2$ and $\boldsymbol{B}_2'$ (31) implies there exist symmetric matrices $\boldsymbol{B}_3$ and $\boldsymbol{B}_3'$ such that

$$[\boldsymbol{I} \otimes \boldsymbol{B}_1 + \boldsymbol{B}_1' \otimes \boldsymbol{I}]\boldsymbol{A}(\boldsymbol{m})[\boldsymbol{I} \otimes \boldsymbol{B}_2 + \boldsymbol{B}_2' \otimes \boldsymbol{I}] = \boldsymbol{I} \otimes \boldsymbol{B}_3 + \boldsymbol{B}_3' \otimes \boldsymbol{I}. \tag{32}$$

Therefore the subspace $\overline{\mathcal{K}}$ satisfies the desired property (26). The subspace $\mathcal{K}$ of self-adjoint fourth-order tensors within $\overline{\mathcal{K}}$ is six-dimensional consisting of all tensors of the form $\boldsymbol{K} = \boldsymbol{I} \otimes \boldsymbol{B} + \boldsymbol{B} \otimes \boldsymbol{I}$, where $\boldsymbol{B}$ is a symmetric matrix. When $\boldsymbol{K} = \boldsymbol{I} \otimes \boldsymbol{B} + \boldsymbol{B} \otimes \boldsymbol{I}$ and $3\kappa_0 + 4\mu_0 - 2\mathrm{Tr}\boldsymbol{B} > 0$ algebraic manipulation shows that

$$\boldsymbol{\mathcal{C}} = W_{\boldsymbol{M}}^{-1}(\boldsymbol{K}) = 2\mu_0(\boldsymbol{\mathcal{I}} - \boldsymbol{I} \otimes \boldsymbol{I}) + \boldsymbol{D} \otimes \boldsymbol{D}, \tag{33}$$

with

$$\boldsymbol{D} = [3\kappa_0 + 4\mu_0 - \mathrm{Tr}\boldsymbol{B})\boldsymbol{I} + 3\boldsymbol{B}]/\sqrt{3(3\kappa_0 + 4\mu_0 - 2\mathrm{Tr}\boldsymbol{B})}. \tag{34}$$

The manifold $\mathcal{M}$ associated with $\mathcal{K}$ therefore consists of all tensors $\boldsymbol{\mathcal{C}} \in \mathcal{M}_e$ expressible in the form (25), and is stable under homogenization.

As an example, consider a three-dimensional elastic polycrystal where the elasticity tensor takes the form

$$\boldsymbol{\mathcal{C}}(\boldsymbol{x}) = \boldsymbol{R}(\boldsymbol{x})\boldsymbol{R}(\boldsymbol{x})\boldsymbol{\mathcal{C}}_0\boldsymbol{R}^T(\boldsymbol{x})\boldsymbol{R}^T(\boldsymbol{x}), \tag{35}$$

where $\boldsymbol{R}(\boldsymbol{x})$ is a rotation matrix, giving the orientation of the crystal at each point $\boldsymbol{x}$ and $\boldsymbol{\mathcal{C}}_0$ is the elasticity tensor of a single crystal which we assume has the form

$$\boldsymbol{\mathcal{C}}_0 = 2\mu_0(\boldsymbol{\mathcal{I}} - \boldsymbol{I} \otimes \boldsymbol{I}) + \boldsymbol{D}_0 \otimes \boldsymbol{D}_0, \quad \text{where } [\mathrm{Tr}(\boldsymbol{D}_0)]^2 - 2\mathrm{Tr}(\boldsymbol{D}_0^2) > 4\mu_0 > 0, \tag{36}$$

in which the latter condition ensures that $\boldsymbol{\mathcal{C}}_0$ is positive definite. The elasticity tensor field $\boldsymbol{\mathcal{C}}(\boldsymbol{x})$ is of the required form (25) with $\boldsymbol{D}(\boldsymbol{x}) = \boldsymbol{R}(\boldsymbol{x})\boldsymbol{D}_0\boldsymbol{R}^T(\boldsymbol{x})$ and therefore the effective tensor $\boldsymbol{\mathcal{C}}_*$ of the polycrystal must lie on the manifold $\mathcal{M}$ for some $\beta > \alpha > 0$. In particular if $\boldsymbol{\mathcal{C}}_*$ is isotropic then its shear modulus is μ_0, independent of the polycrystal microgeometry. For planar elasticity the analogous result was proved by Avellaneda et. al. (1996).

SOME INTERESTING EXACT RELATIONS FOR COUPLED FIELD PROBLEMS

We are left with the algebraic problem of characterizing which tensor subspaces satisfy the conditions (20) or (26). One might wonder if there is perhaps some easy characterization. For elasticity and conductivity in two or three dimensions all possible rotationally invariant exact relations have now been found [see Grabovsky (1998), Grabovsky and Sage (1988) and references therein] but in a more general context the following example shows that the task is not so simple.

Consider a coupled field problem where there are there are m divergence free fields $\boldsymbol{j}_1(\boldsymbol{x}), \boldsymbol{j}_2(\boldsymbol{x}), ..., \boldsymbol{j}_m(\boldsymbol{x})$ and m curl free fields $\boldsymbol{e}_1(\boldsymbol{x}), \boldsymbol{e}_2(\boldsymbol{x}), ..., \boldsymbol{e}_m(\boldsymbol{x})$ which are linked through the constitutive relation

$$j_{i\alpha}(\boldsymbol{x}) = \sum_{j=1}^{d} \sum_{\beta=1}^{m} L_{i\alpha j\beta}(\boldsymbol{x}) e_{j\beta}(\boldsymbol{x}), \tag{37}$$

where α and β are field indices while i and j are space indices. Milgrom and Shtrikman (1989) have obtained some very useful exact relations for coupled field problems. Rather than rederiving these let us look for exact relations with $\boldsymbol{M} = 0$ and a reference tensor $\boldsymbol{L}_0$ which is the identity tensor $\boldsymbol{I}$. The associated tensor $\boldsymbol{A}(\boldsymbol{m}) = \boldsymbol{M} - \boldsymbol{\Gamma}(\boldsymbol{m})$ has elements $A_{i\alpha j\beta} = -\delta_{\alpha\beta} m_i m_j$. Now take $\mathcal{R}$ to be a r-dimensional subspace of $m \times m$ matrices and let $\mathcal{S}$ denote the d^2-dimensional space of $d \times d$ matrices, and consider the rd^2-dimensional subspace $\overline{\mathcal{K}}$ spanned by all tensors $\boldsymbol{K}$ which are tensor products of matrices $\boldsymbol{R} \in \mathcal{R}$ and matrices $\boldsymbol{S} \in \mathcal{S}$, i.e. which have elements $K_{i\alpha j\beta} = R_{\alpha\beta} S_{ij}$. Given a tensor $\boldsymbol{K}_1$ which is the tensor product of $\boldsymbol{R}_1 \in \mathcal{R}$ and $\boldsymbol{S}_1 \in \mathcal{S}$ and a tensor $\boldsymbol{K}_2$ which is the tensor product of $\boldsymbol{R}_2 \in \mathcal{R}$ and $\boldsymbol{S}_2 \in \mathcal{S}$, the product $\boldsymbol{K}_1 \boldsymbol{A}(\boldsymbol{m}) \boldsymbol{K}_2$ will certainly be in $\overline{\mathcal{K}}$ provided $\boldsymbol{R}_1 \boldsymbol{R}_2 \in \mathcal{R}$. Moreover if this holds for all $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathcal{R}$ then $\overline{\mathcal{K}}$ defines an exact relation because it is spanned by matrices of the same form as $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$. This observation allows us to generate countless exact relations. The condition on $\mathcal{R}$ just says that it is closed under multiplication, i.e. that it forms an algebra. Unfortunately there is no known way of characterizing which subspaces of matrices form an algebra for general m, and this hints of the difficulties involved in trying to obtain a complete characterization of exact relations. Since $\boldsymbol{M} = 0$ the manifold $\mathcal{M}$ consists of an appropriately bounded coercive subset of tensors of the form $\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{K}$ where $\boldsymbol{K} \in \overline{\mathcal{K}}$. The case where $m = 2$ and $\mathcal{R}$ is the set of all 2×2 matrices of the form $\boldsymbol{R} = a\boldsymbol{I} + b\boldsymbol{R}_\perp$ (which is clearly closed under multiplication) corresponds to tensor fields $\boldsymbol{L}(\boldsymbol{x})$ for which the consitutive relation can be rewritten in the equivalent form of a complex equation

$$\boldsymbol{j}_1(\boldsymbol{x}) + i\boldsymbol{j}_2(\boldsymbol{x}) = (\boldsymbol{A}(\boldsymbol{x}) + i\boldsymbol{B}(\boldsymbol{x}))(\boldsymbol{e}_1(\boldsymbol{x}) + i\boldsymbol{e}_2(\boldsymbol{x})). \tag{38}$$

The effective tensor $\boldsymbol{L}_*$ will have an associated complex form $\boldsymbol{A}_* + i\boldsymbol{B}_*$.

REFERENCES

Avellaneda, M., Cherkaev, A.V., Gibiansky, L.V., Milton, G.W., and Rudelson, M. 1996 *J. Mech. Phys. Solids* 44, 1179-1218.

Backus, G.E. 1962 *J. Geophys. Res.* 67, 4427-4440.

Bensoussan, A., Lions, J.L., and Papanicolaou, G. 1978 *Asymptotic analysis for periodic structures. Studies in mathematics and its applications*, 5, North-Holand.

Dvorak, G.J., and Benveniste, Y. 1997 In *Theoretical and Applied Mechanics 1996*, pp. 217-237, ed. by T.Tatsumi, E.Watanabe, and T.Kambe, Elsevier.

Dykhne, A.M. 1970 *Zh. Eksp. Teor. Fiz.* 59 , 110-115. [*Soviet Physics JETP* 32 (1971), 63-65.]

Grabovsky, Y., and Milton, G.W. 1998 *Proc. Roy. Soc. Edin. A* 128, 283-299.

Grabovsky, Y., 1998 *Arch. Rat. Mech. Anal.* to appear.

Grabovsky, Y., and Sage, D.S. 1998 *Arch. Rat. Mech. Anal.* to appear.

Hill, R. 1952 *Proc. Phys. Soc. Lond. A* 65, 349-354.

Keller, J.B. 1964 *J. Math. Phys.* 5, 548-549.

Lurie, K.A., and Cherkaev, A.V. 1981 *Dokl. Akad. Nauk.* 259, 328-331.

Lurie, K.A., Cherkaev, A.V., and Fedorov, A.V. 1984 *J. Opt. Theory Appl.* 42, 247-281.

Mendelson, K.S. 1975 *J. Appl. Phys.* 46, 4740-4741.

Milgrom, M. and Shtrikman, S. 1989 *Phys. Rev. A* 40, 1568-1575.

Milton, G.W. 1997 In *Theoretical and Applied Mechanics 1996*, pp. 443-459, ed. by T.Tatsumi, E.Watanabe, and T.Kambe, Elsevier.

Murat, F. and Tartar, L. 1985 In *Les méthodes de l'homogénéisation: théorie et applications en physique, Coll. de la Dir. des Études et Recherches de Électricité de France*, Eyrolles, Paris, 319-370.

Murat, F., and Tartar, L. 1997 In *Topics in the mathematical modelling of composite materials*, ed. by A. Cherkaev and R. Kohn, *Progress in nonlinear differential equations and their applications*, 31, pp. 21-43, Birkhauser.

Tartar, L. 1976 In *Computer Methods in Applied Sciences and Engineering*, R. Glowinski, J.-L. Lions eds., Springer-Verlag Lecture Notes in Mathematics 704, pp.136-212, Springer-Verlag.

Zhikov (Jikov), V.V., Kozlov, S.M., and Oleinik, O.A. 1994 *Homogenization of differential operators and integral functionals*, Springer-Verlag.

Yury Grabovsky
Department of Mathematics
The University of Utah
Salt Lake City, Utah 84112, U.S.A.
yuri@math.utah.edu

Graeme W. Milton
Department of Mathematics
The University of Utah
Salt Lake City, Utah 84112, U.S.A.
milton@math.utah.edu

Optimal Dynamic Instability of Microtubules

Charles S. Peskin[1]

Abstract. Microtubules are polymers that play many important structural and functional roles within biological cells, including the separation of newly replicated chromosomes into the daughter cells during cell division. In order to catch the chromosomes that they must transport, microtubules grow out of the centrosome in each of the daughter cells. For any particular microtubule, epochs of steady growth are punctuated by episodes of rapid decay; this is known as dynamic instability. It allows for multiple attempts on the part of each microtubule to hit the small target at the center of each chromosome known as the kinetochore, where the microtubule can attach and apply traction to the chromosome. The optimal design of dynamic instability is the subject of this paper.

1991 Mathematics Subject Classification: 92C05, 92C40, 92C45
Keywords and Phrases: dynamic instability, microtubule, chromosome

1 Introduction: How to Catch and Transport a Chromosome

During cell division, the newly replicated chromosomes are pulled into the daughter cells by microtubules. This activity is organized by the centrosomes, one in each of the daughter cells, which form the two poles of the familiar mitotic spindle. Microtubules are polymers, made of protein subunits known as tubulin, that grow radially outward from the centrosomes. Dynamic instability, discovered by Mitchison and Kirschner [1], is a phenomenon concerning the assembly and disassembly of microtubules. Specifically, the individual steps of addition and removal of tubulin subunits to and from the end of a given microtubule, although random, are far from independent. Indeed, the microtubule acts like a two-state device, with a steadily growing state and a rapidly decaying state. Transitions between these states occur much more rarely than the individual steps of addition or removal of subunits.

As has been emphasized by Hill and Chen [2], dynamic instability drastically alters the statistical properties of microtubules, in comparison to the properties that would be expected on the basis of independent addition and removal of subunits. In this paper, we shall continue the exploration of this theme, from a

[1]Supported by the National Science Foundation (USA) under DMS/FD 92-20719. Thanks also to George Oster and David McQueen for their help in connection with this work.

somewhat different perspective, that of optimal design. Specifically, we shall state and solve an optimization problem that explains why dynamic instability is needed and determines certain relationships between the rate constants that characterize the assembly and disassembly of microtubules.

Mathematical theories and computer simulations of the dynamic instability of microtubules may be divided into two broad categories. First, there are the theories that simplify the microtubule by treating it as a one-dimensional polymer. Among such works are [2, 3] and also the present paper adopts this simplified point of view, which is amenable to analysis. Another possibility is to take into account the two-dimensional tubular lattice in which the subunits of the microtubule are actually arranged. This has been done in [4, 5, 6, 9]. The two-dimensional lattice models have been studied by Monte-Carlo simulation.

For a detailed review of the role of microtubules in chromosome transport, including but not limited to dynamic instability, see [7]. Dynamic instability makes possible the trial-and-error process that leads to chromosome capture by microtubules. Following capture, traction on the chromosome is generated by depolymerization of the microtubules [8].

2 Polymerization and Depolymerization of Microtubules

A typical microtubule consists of 13 protofilaments, each of which runs in a straight line, parallel to the axis of the microtubule. We shall simplify the description of the microtubule by regarding it as a one-dimensional polymer; this polymer may be thought of as representing any one of the 13 protofilaments, even though this ignores significant interactions between neighboring protofilaments, interactions which tend to coordinate their assembly and disassembly.

The subunits of a microtubule are tubulin dimers, here denoted by the symbol T. Each such tubulin dimer has two possible states, denoted T.GDP or T.GTP according to whether a guanosine diphosphate (GDP) or a guanosine triphosphate (GTP) molecule is bound to the tubulin dimer. Following the lateral cap hypothesis of Bayley et al. [5], in the simplified form appropriate to our one-dimensional model, we assume that only T.GTP can be added to a microtubule, and only T.GDP can exist in the interior of a protofilament (i.e., not at its end). Note that these rules allow the terminal subunit to be either T.GDP or T.GTP. This one bit of information will determine whether the model microtubule is in a polymerizing mode (with T.GTP at the tip), or in a depolymerizing mode (with T.GDP at the tip).

In case the terminal subunit is T.GTP, then the following polymerization reaction, driven by GTP hydrolysis (GTP $\rightarrow$ GDP + Pi, where Pi denotes inorganic phosphate) can occur

$$(\text{T.GDP})_{n-1}(\text{T.GTP}) + (\text{T.GTP}) \rightarrow (\text{T.GDP})_n(\text{T.GTP}) + \text{Pi} \qquad (1)$$

Note that the protofilament has grown by the addition of one tubulin dimer and that it still has a T.GTP subunit at its tip, so the process described by Eq.1 may be repeated indefinitely.

If, on the other hand, the terminal subunit is T.GDP, then this subunit can spontaneously dissociate:

$$(\text{T.GDP})_n \rightarrow (\text{T.GDP})_{n-1} + (\text{T.GDP}) \tag{2}$$

and this depolymerization process, too, may be repeated indefinitely (until the microtubule has shrunk to zero length).

Conversion in either direction between the polymerizing mode (Eq.1) and the depolymerizing mode (Eq.2) may occur through the following reversible reaction, which is supposed to be rare (i.e., slow) in comparison to the reactions described by Eqs.1 and 2, above:

$$(\text{T.GDP})_{n-1} + (\text{T.GTP}) \leftrightarrow (\text{T.GDP})_{n-1}(\text{T.GTP}) \tag{3}$$

The forward reaction in Eq.3 switches the protofilament from the depolymerizing to the polymerizing mode, and the reverse reaction accomplishes the opposite. There are other possible ways to switch modes (involving phosphorylation or dephosphorylation of the terminal T.GDP or T.GTP, respectively), but we shall adhere to Eq.3 as the switching mechanism throughout this paper.

The following diagram summarizes the kinetic scheme for the assembly and disassembly of microtubules that is used in this paper:

$$\begin{array}{ccccccccccccc}
 & & \alpha & & & \alpha & & \alpha & & & \alpha & & \\
\mathrm{A}_1 & \rightarrow & \mathrm{A}_2 & \rightarrow & \cdots & \rightarrow & \mathrm{A}_n & \rightarrow & \mathrm{A}_{n+1} & \rightarrow & \cdots & & \\
\alpha' \uparrow\downarrow \beta' & & \alpha' \uparrow\downarrow \beta' & & & & \alpha' \uparrow\downarrow \beta' & & \alpha' \uparrow\downarrow \beta' & & & & \\
\mathrm{B}_0 & \leftarrow & \mathrm{B}_1 & \leftarrow & \cdots & \leftarrow & \mathrm{B}_{n-1} & \leftarrow & \mathrm{B}_n & \leftarrow & \cdots & & \\
 & & \beta & & & \beta & & \beta & & & \beta & &
\end{array} \tag{4}$$

In this diagram, the symbol A denotes the polymerizing mode and B denotes the depolymerizing mode. The subscript on A or B denotes the total number of subunits in the polymer. Thus, $\mathrm{A}_n = (\text{T.GDP})_{n-1}(\text{T.GTP})$, $n \geq 1$; and $\mathrm{B}_n = (\text{T.GDP})_n$, $n \geq 0$. Note that B_0 is the fixed anchor, or seed, located within the centrosome, from which the microtubule grows. Implicit in the kinetic scheme Eq.4 is the assumption that this seed has the same properties as a T.GDP molecule.

The rate constants, with dimensions of inverse time, that appear in the foregoing scheme, are defined as follows: α = rate constant for fast (Eq.1) polymerization; β = rate constant for fast (Eq.2) depolymerization; α' = rate constant for slow (Eq.3) polymerization; β' = rate constant for slow (Eq.3) depolymerization. Note that the polymerizing rate constants are proportional to the concentration of T.GTP in solution: $\alpha = a\,[\text{T.GTP}]$; $\alpha' = a'\,[\text{T.GTP}]$, but that the depolymerizing rate constants are independent of concentration.

Implicit in our whole discussion of the microtubule as a two-state device, with a polymerizing state and a depolymerizing state, are the inequalities $\beta' < \alpha$ and $\alpha' < \beta$, so that the microtubule takes many steps of polymerization with rate constant α before losing its T.GTP cap, and many steps of depolymerization with rate constant β before regaining that cap.

For the sake of comparison, however, it is also of interest to consider the special case $\alpha = \alpha'$, $\beta = \beta'$. This corresponds to the situation in which polymerization

and depolymerization proceed without regard to the distinction between T.GTP and T.GDP, and there is no phenomenon of dynamic instability. We shall try to understand why nature does *not* proceed in this simple manner.

The differential equations describing an ensemble of protofilaments can now be written down by inspection of the kinetic scheme (Eq.4). Let $p_n(t)$ be the probability of finding the system in state A_n at time t, and let $q_n(t)$ be the corresponding probability for the state B_n. Then

$$\frac{dq_0}{dt} = \beta' p_1 + \beta q_1 - \alpha' q_0 \tag{5}$$

$$\frac{dp_1}{dt} = \alpha' q_0 - (\alpha + \beta') p_1 \tag{6}$$

and for $n \geq 1$

$$\frac{dq_n}{dt} = \beta' p_{n+1} + \beta q_{n+1} - (\alpha' + \beta) q_n \tag{7}$$

$$\frac{dp_{n+1}}{dt} = \alpha p_n + \alpha' q_n - (\alpha + \beta') p_{n+1} \tag{8}$$

Finally, the q_n and p_n are normalized according to

$$\sum_{n=0}^{\infty} q_n + \sum_{n=1}^{\infty} p_n = 1 \tag{9}$$

It follows from Eqs.5-8 that

$$\sum_{k=0}^{n-1} \left(\frac{dq_k}{dt} + \frac{dp_{k+1}}{dt} \right) = -(\alpha p_n - \beta q_n) \tag{10}$$

for $n \geq 1$. This will be useful in constructing a steady-state solution of Eqs.5-9.

3 Steady-State Solution [3]

In the steady state ($dp_n/dt = dq_n/dt = 0$ for all n), Eq.10 becomes $\alpha p_n = \beta q_n$. Thus, we may set $u_n = \alpha p_n = \beta q_n$, $n \geq 1$. It then follows from the steady-state form of Eq.7 or 8 that $u_{n+1} = r u_n$, $n \geq 1$, where $r = (1 + \alpha'/\beta)/(1 + \beta'/\alpha)$. Thus, we have a normalizable solution if and only if $r < 1$, which (since all of the rate constants are positive) is equivalent to

$$0 < (\beta\beta' - \alpha\alpha') \tag{11}$$

This means that depolymerization is dominant over polymerization. If the inequality Eq.11 is not satisfied, then the microtubule just grows forever and there is no steady state. From now on we shall assume that this important inequality is indeed satisfied.

According to the foregoing, the u_n form a geometric sequence for $n \geq 1$. It is then straightforward to express all of the p_n and q_n in terms of u_1, and to determine u_1 with the help of the normalization condition, Eq.9, thus completing

the steady-state solution. We omit the details, but just give the following useful result:

Let N be the random variable which is the number of subunits in a protofilament, and let E[ ] denote the expected value (ensemble average) of the enclosed quantity. Then

$$n_p = \mathrm{E}[N|N>0] = \frac{\sum_{n=1}^{\infty}(p_n+q_n)n}{\sum_{n=1}^{\infty}(p_n+q_n)} = \frac{\mathrm{E}[N]}{1-q_0} = \frac{\beta(\alpha+\beta')}{\beta\beta'-\alpha\alpha'} \tag{12}$$

Note that n_p measures the average length (in subunits) of microtubules by averaging only over actual microtubules, i.e., by *not* including microtubules of zero length in the average.

We now compare two special cases. First suppose $\alpha = \alpha'$ and $\beta = \beta'$. This is the above-mentioned case in which the kinetics are indifferent to the distinction between T.GDP and T.GTP. In this case, we find $n_p = 1/(1-\alpha/\beta)$ Clearly, to achieve a microtubule of any significant length (e.g., $n_p = 100$) in this situation, (α/β) must be very close to 1. On the other hand, it is also required that $(\alpha/\beta) < 1$, or the steady-state solution does not exist. This implies that the parameters must be poised on the edge of disaster in order for the system to function!

Now consider instead the limiting case $\beta \to \infty$, with α, α', and β' all finite. In this limit, $n_p \to (\alpha/\beta') + 1$, the steady-state solution always exists, and we can make the microtubules as long as we like by choosing (α/β') large. This is much better! The limiting case $\beta \to \infty$ has other virtues as well. These will appear below.

4 Mean and Variance of the Cycle Time

In order to participate in chromosome transport, a microtubule must first grow until it hits the kinetochore of a chromosome. This being an unlikely event, repeated trials are needed. To the extent that microtubules grow in straight lines, a new trial cannot be said to begin until the microtubule shrinks all the way down to zero length and then starts to grow again, at a possibly different angle. Thus, an important random variable is the *cycle time*, S_c, which we define as the elapsed time between successive departures from the state B_0, in which the microtubule has zero length, see Eq.4.

The cycle time S_c has two components:

$$S_c = S_0 + S_p \tag{13}$$

where S_0 is the waiting time in state B_0, and S_p is the time elapsed between a given *departure* from B_0 and the subsequent *arrival* at B_0. Since S_0 and S_p are independent random variables, we have $\tau_c = \mathrm{E}[S_c] = \mathrm{E}[S_0] + \mathrm{E}[S_p]$ and $v_c = \mathrm{Var}[S_c] = \mathrm{Var}[S_0] + \mathrm{Var}[S_p]$, where Var[ ] denotes the variance of the enclosed random variable.

On general principles concerning chemical reactions, we know that the waiting time S_0 in the state B_0 is exponentially distributed with mean $1/\alpha'$. It follows that $\mathrm{E}[S_0] = 1/\alpha'$ and $\mathrm{Var}[S_0] = (1/\alpha')^2$ Thus, to evaluate τ_c and v_c, we just

need to know the mean and variance of the random variable S_p. These are found by assuming that the system starts ($t = 0$) in the state A_1 and by treating B_0 as an absorbing state. The reaction scheme is the same as Eq.4, except that the transition $B_0 \to A_1$ is omitted:

$$\begin{array}{ccccccccccccc} & & \alpha & & & \alpha & & \alpha & & \alpha & & \alpha & & \\ A_1 & \to & & A_2 & \to & \cdots & \to & & A_n & \to & A_{n+1} & \to & \cdots & \\ \downarrow \beta' & & \alpha' \uparrow\downarrow \beta' & & & & & \alpha' \uparrow\downarrow \beta' & & & \alpha' \uparrow\downarrow \beta' & & & \\ B_0 & \leftarrow & & B_1 & \leftarrow & \cdots & \leftarrow & & B_{n-1} & \leftarrow & & B_n & \leftarrow & \cdots \\ & & \beta & & & \beta & & \beta & & \beta & & \beta & & \end{array} \tag{14}$$

The differential equations are

$$\frac{dq_0}{dt} = \beta' p_1 + \beta q_1 \tag{15}$$

$$\frac{dp_1}{dt} = -(\alpha + \beta') p_1 \tag{16}$$

and for $n \geq 1$, we have, as before, Eqs.7 and 8. Finally, the initial conditions are $p_1(0) = 1$ with all of the other p_n and all of the q_n equal to zero at $t = 0$.

If we can solve this initial-value problem, then we shall have the probability density function of S_p, denoted $\rho_p(t)$, which is given by

$$\rho_p(t) = \frac{dq_0}{dt} = \beta q_1(t) + \beta' p_1(t) \tag{17}$$

The initial-value problem stated above can indeed be solved in terms of Laplace transforms. Instead of inverting the Laplace transform to find $\rho_p(t)$, however, we shall be content with finding the mean and variance of S_p, which can be evaluated directly from the Laplace transform itself. Specifically, if we define $\hat{\rho}_p(\lambda) = \int_0^\infty \rho_p(t) \exp(-\lambda t) dt$ (and similarly for all other functions of t), then we have

$$\mathrm{E}[S_p] = \int_0^\infty t \rho_p(t) dt = -\frac{d\hat{\rho}_p}{d\lambda}(0) \tag{18}$$

and similarly,

$$\mathrm{Var}[S_p] = \mathrm{E}[S_p^2] - (\mathrm{E}[S_p])^2 = \frac{d^2\hat{\rho}_p}{d\lambda^2}(0) - \left(\frac{d\hat{\rho}_p}{d\lambda}(0)\right)^2 \tag{19}$$

In terms of the transformed variables $\hat{q}_n$ and $\hat{p}_n$, the initial value problem becomes

$$\lambda \hat{q}_0 - \beta \hat{q}_1 - \beta' \hat{p}_1 = 0 \tag{20}$$

$$\lambda \hat{p}_1 + (\alpha + \beta') \hat{p}_1 = 1 \tag{21}$$

and for $n \geq 1$:

$$(\lambda + \alpha' + \beta) \hat{q}_n - \beta' \hat{p}_{n+1} - \beta \hat{q}_{n+1} = 0 \tag{22}$$

$$(\lambda + \alpha + \beta') \hat{p}_{n+1} - \alpha \hat{p}_n - \alpha' \hat{q}_n = 0 \tag{23}$$

Now Eq.21 gives $\hat{p}_1$ directly, and we look for a solution of Eqs.22-23 of the following form

$$\hat{p}_n(\lambda) = \hat{p}_1(\lambda) z^{n-1}, n \geq 1 \tag{24}$$

$$\hat{q}_n(\lambda) = \hat{q}_1(\lambda) z^{n-1}, n \geq 1 \tag{25}$$

where we must require $|z| < 1$. With this assumed form, Eqs.22 and 23 reduce to the homogeneous 2×2 system

$$\begin{pmatrix} \lambda_1 z - \alpha & -\alpha' \\ -\beta' z & \lambda_2 - \beta z \end{pmatrix} \begin{pmatrix} \hat{p}_1(\lambda) \\ \hat{q}_1(\lambda) \end{pmatrix} = 0 \tag{26}$$

where $\lambda_1 = \lambda + \alpha + \beta'$ and $\lambda_2 = \lambda + \alpha' + \beta$. Of course, z is chosen so that Eq.26 has nontrivial solutions and $|z| < 1$. Then $\hat{q}_1$ can be found from Eq.26, since $\hat{p}_1$ is already known from Eq.21. The details are left as a (lengthy) exercise for the reader. The results, after adding the expectations of S_p and S_0, and similarly after adding their variances, are as follows:

$$\tau_c = \mathrm{E}[S_c] = \frac{1}{\alpha'} + \frac{\alpha + \beta}{\beta\beta' - \alpha\alpha'} \tag{27}$$

$$\begin{aligned} v_c &= \mathrm{Var}[S_c] \\ &= \left(\frac{1}{\alpha'}\right)^2 + \frac{2\beta}{(\alpha + \beta')(\beta\beta' - \alpha\alpha')} \left(1 + \frac{\alpha\beta'(\alpha + \beta' + \alpha' + \beta)^2}{(\beta\beta' - \alpha\alpha')^2}\right) \\ &\quad - \left(\frac{\alpha + \beta}{\beta\beta' - \alpha\alpha'}\right)^2 \end{aligned} \tag{28}$$

5 Optimal Design of Dynamic Instability

We are now ready to state the optimization problem that is the main subject of this paper. In order that the stochastic process of dynamic instability should proceed as regularly as possible, let us choose α, α', β, and β' to minimize the variance v_c of the cycle time, subject to given values of the mean cycle time τ_c and the mean length n_p of nonzero length microtubules.

Since there are 4 variables and 2 constraints, it should be possible to reduce the number of independent variables to 2. A convenient choice of independent variables is α' and β. From the constraints, Eqs.12 and 27, we find

$$\alpha = \frac{\beta(n_p - 1)}{Q} \tag{29}$$

$$\beta' = \frac{\alpha' n_p + \beta}{Q} \tag{30}$$

where

$$Q = (\alpha'\tau_c - 1)(n_p + \frac{\beta}{\alpha'}) - \alpha'\tau_c(n_p - 1) \tag{31}$$

We can use these results to express v_c as a function of α' and β only (though of course it will also contain as parameters the given constants n_p and τ_c):

$$\begin{aligned} v_c(\alpha', \beta; n_p, \tau_c) &= \left(\frac{1}{\alpha'}\right)^2 + \tau_c^2 \left(1 + 2(n_p - 1)\frac{\alpha'}{\beta}\right) - 2\frac{\tau_c}{\alpha'} \\ &+ \left(\frac{1}{\alpha' + \beta}\right)\left(\left(\frac{1}{\alpha'}\right)\left(2n_p - 1 + \frac{\beta}{\alpha'}\right) - 4(n_p - 1)\tau_c\right) \quad (32) \end{aligned}$$

Our goal is to minimize v_c with respect to α' and β. Let us first consider

$$\frac{\partial v_c}{\partial \beta} = -\frac{2(n_p - 1)}{\alpha'(\alpha' + \beta)^2}\left(\left(\frac{\alpha' + \beta}{\beta}\right)^2 (\alpha'\tau_c)^2 - 2(\alpha'\tau_c) + 1\right) \quad (33)$$

Since $(\alpha' + \beta)/\beta > 1$, and since $n_p > 1$, it is evident that $\partial v_c/\partial\beta < 0$, and there can be no minimum at finite β We can, however, look for a minimum at $\beta = \infty$. Letting $\beta \to \infty$, we find

$$v_c(\alpha', \infty; n_p, \tau_c) = \left(\frac{1}{\alpha'}\right)^2 + \left(\tau_c - \frac{1}{\alpha'}\right)^2 \quad (34)$$

which is minimized by setting

$$\alpha' = \frac{2}{\tau_c} \quad (35)$$

To complete the solution, we need only find α and β'. Taking the limit $\beta \to \infty$ in Eqs.29-31, we find

$$\alpha = \frac{n_p - 1}{\tau_c - \frac{1}{\alpha'}} = \frac{2}{\tau_c}(n_p - 1) \quad (36)$$

$$\beta' = \frac{1}{\tau_c - \frac{1}{\alpha'}} = \frac{2}{\tau_c} \quad (37)$$

Thus, in summary, the optimal solution is given by $\alpha' = \beta' = 2/\tau_c$; $\alpha = (n_p - 1)(2/\tau_c)$; and $\beta = \infty$. The variance in the mean cycle time obtained in this way is given by $v_c^{\min} = \tau_c^2/2$, which is half that of an exponentially distributed random variable with the same mean. Note that $v_c^{\min}$ is independent of n_p.

To appreciate better the optimal solution, let us contrast it with the case obtained by setting $\alpha' = \alpha$ and $\beta' = \beta$. As discussed above, this means that there is no distinction between the T.GDP and the T.GTP subunit. Under these (degenerate) circumstances, we have, after some algebra,

$$v_c = \left(\frac{1}{\alpha}\right)^2 + \frac{2\beta}{(\beta - \alpha)^3} - \left(\frac{1}{\beta - \alpha}\right)^2 = \left(\frac{\tau_c}{n_p}\right)^2 \left(1 + (2n_p - 1)(n_p - 1)^2\right) \quad (38)$$

which is asymptotic to $2n_p\tau_c^2$ as $n_p \to \infty$.

Thus, in the absence of a mechanism that distinguishes T.GDP from T.GTP, we find that the variance of the cycle time is a *large* multiple of the square of the mean cycle time, instead of being fixed at $\tau_c^2/2$ as in the optimal solution. Such

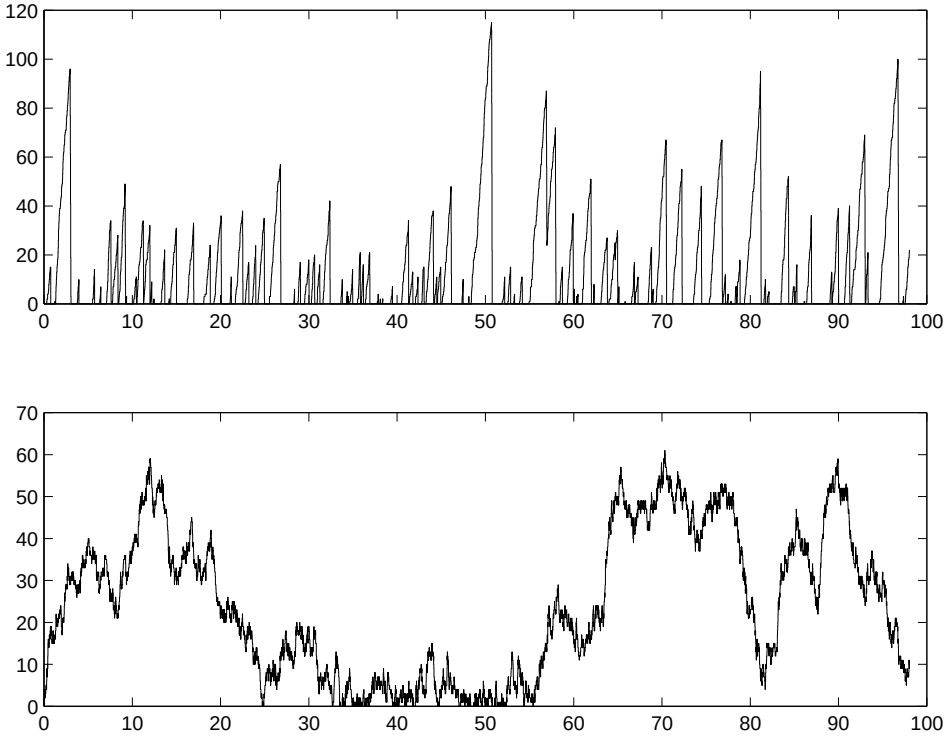


Figure 1: Near-optimal dynamic instability (above) and *no* dynamic instability (below); sample trajectories obtained by Monte-Carlo simulation. The horizontal axis measures time in units of τ_c, and the vertical axis is polymer length expressed in terms of the number of subunits. Note the extreme difference in statistical character of the trajectories, even though both have the same mean polymer length $n_p = 25$ and the same mean cycle time $\tau_c = 1$. The (nearly) optimal case has many cycles of comparable duration, whereas the degenerate case has a few long cycles and a great many cycles that are much too short to be effective. This difference in statistics, which is already quite dramatic, can be tremendously accentuated by increasing the mean polymer length n_p.

a large multiple indicates a long-tailed distribution of cycle times. Under these conditions, a microtubule that missed its target (and that would be most of them, after the first try) might spend a long time wandering up and down in length before shrinking to zero length to try again. In the case of the optimal solution, though, the cycle time is rather tightly controlled, and its variance is independent of the mean length of the microtubules. The length can therefore be made large without paying a price in terms of the variability of the cycle time. The degenerate case and a near-optimal case (finite but large β) are further contrasted in Figure 1.

References

[1] Mitchison T and Kirschner M: Dynamic instability of microtubule growth. Nature 312:237-242, 1984

[2] Hill TL and Chen Y: Phase changes at the end of a microtubule with a GTP cap. PNAS 81:5772-5776, 1984

[3] Hill TL: Introductory analysis of the GTP-cap phase change kinetics at the end of a microtubule. PNAS 81:6728-6732, 1894

[4] Chen Y and Hill TL: Monte Carlo study of the GTP cap in a five-start helix model of a microtubule. PNAS 82:1131-1135, 1985

[5] Bayley PM, Schilstra MJ, and Martin SR: Microtubule dynamic instability: numerical simulation of microtubule transition properties using a Lateral Cap model. J Cell Science 95:33-48, 1990

[6] Martin SR, Schilstra MJ, and Bayley PM: Dynamic Instability of Microtubules: Monte Carlo Simulation and Application to Different Types of Microtubule Lattice. Biophysical J 65:578-596, 1993

[7] Inoué S and Salmon ED: Force Generation by Microtubule Assembly/Disassembly in Mitosis and Related Movements. Molecular Biology of the Cell 6:1619-1640, 1995

[8] Peskin CS and Oster GF: Force production by depolymerizing microtubules: load-velocity curves and run-pause statistics. Biophysical J 69:2268-2276, 1995

[9] Tao YC and Peskin CS: Simulating the role of microtubules in depolymerization-driven transport – A Monte Carlo approach. Biophysical J. (in press)

Charles S. Peskin
Courant Institute
New York University
251 Mercer Street
New York, NY 10012, USA
peskin@cims.nyu.edu

Section 17

Control Theory and Optimization

In case of several authors, Invited Speakers are marked with a *.

On the Solution of Traveling Salesman Problems

David Applegate, Robert Bixby,
Vašek Chvátal and William Cook[1]

Abstract. Following the theoretical studies of J.B. Robinson and H.W. Kuhn in the late 1940s and the early 1950s, G.B. Dantzig, R. Fulkerson, and S.M. Johnson demonstrated in 1954 that large instances of the TSP could be solved by linear programming. Their approach remains the only known tool for solving TSP instances with more than several hundred cities; over the years, it has evolved further through the work of M. Grötschel, S. Hong, M. Jünger, P. Miliotis, D. Naddef, M. Padberg, W.R. Pulleyblank, G. Reinelt, G. Rinaldi, and others. We enumerate some of its refinements that led to the solution of a 13,509-city instance.

1991 Mathematics Subject Classification: 90C10 90C27
Keywords and Phrases: traveling salesman; cutting planes

The *traveling salesman problem*, or TSP for short, is easy to state: given a finite number of "cities" along with the cost of travel between each pair of them, find the cheapest way of visiting all of the cities and returning to your starting point. The travel costs are symmetric in the sense that traveling from city X to city Y costs just as much as traveling from Y to X; the "way of visiting all the cities" is simply the order in which the cities are visited. The simplicity of this problem, coupled with its apparent intractability, makes it an ideal platform for exploring new algorithmic ideas. Surveys of work on the TSP can be found in Bellmore and Nemhauser [1968], Lawler, Lenstra, Rinnooy Kan, and Shmoys [1985], Reinelt [1994], and Jünger, Reinelt, and Rinaldi [1995].

The origins of the TSP are obscure. In the 1920's, the mathematician and economist Karl Menger publicized it among his colleagues in Vienna. In the 1930's, the problem reappeared in the mathematical circles of Princeton. In the 1940's, it was studied by statisticians (Mahalanobis [1940], Jessen [1942]) in connection with an agricultural application and the mathematician Merrill Flood popularized it among his colleagues at the RAND Corporation. Eventually, the TSP gained notoriety as the prototype of a hard problem in combinatorial optimization.

A breakthrough came when Dantzig, Fulkerson, and Johnson [1954] published a description of a method for solving the TSP and illustrated the power of this method by solving an instance with 49 cities, an impressive size at that time. Riding the wave of excitement over the numerous applications of the simplex

[1]Supported by ONR Grant N00014-98-1-0014.

method (designed by Dantzig in 1947) and following the studies of Robinson [1949] and Kuhn [1955], Dantzig, Fulkerson, and Johnson attacked the salesman with linear programming as follows.

Each TSP instance with n cities can be specified by a vector c with $n(n-1)/2$ components, whose values are the travel costs; each tour through the n cities can be represented as its incidence vector with $n(n-1)/2$ components; if $\mathcal{S}$ denotes the set of the incidence vectors of all the tours, then the problem is to

$$\text{minimize } c^T x \text{ subject to } x \in \mathcal{S}. \tag{1}$$

Like the man searching for his lost wallet not in the dark alley where he actually dropped it, but under a street lamp where he can see, Dantzig, Fulkerson and Johnson begin not with the problem they *want to* solve, but with a related problem they *can* solve,

$$\text{minimize } c^T x \text{ subject to } Ax \leq b \tag{2}$$

with some suitably chosen system $Ax \leq b$ of linear inequalities satisfied by all x in $\mathcal{S}$: solving linear programming problems such as (2) is precisely what the simplex method is for. Since (2) is a *relaxation* of (1) in the sense that every feasible solution of (1) is a feasible solution of (2), the optimal value of (2) provides a lower bound on the optimal value of (1).

The ground-breaking idea of Dantzig, Fulkerson, and Johnson was that solving (2) can help with solving (1) in a far more substantial way than just by providing a lower bound: having determined that the wallet is not under the street lamp, one can pick the street lamp up and bring it a little closer to the place where the wallet was lost. If (2) has an optimal solution and if the polyhedron $\{x : Ax \leq b\}$ has an extreme point, then the simplex method finds an optimal solution x^* of (2) such that x^* is an extreme point of $\{x : Ax \leq b\}$; in particular, if x^* is not a member of $\mathcal{S}$, then some linear inequality is satisfied by all the points in $\mathcal{S}$ and violated by x^*. Such an inequality is called a *cutting plane* or simply a *cut*. Having found cuts, one can add them to the system $Ax \leq b$, solve the resulting tighter relaxation by the simplex method, and iterate this process until one arrives at a linear programming relaxation of (1) and its optimal solution x^* such that $x^* \in \mathcal{S}$.

The influence of this work reached far beyond the narrow confines of the TSP: the *cutting-plane method* can be used to attack any problem (1) such that $\mathcal{S}$ is a finite subset of $\mathbf{R}^m$ and an efficient algorithm to recognize points of $\mathcal{S}$ is available. Many problems in *combinatorial optimization* have this form: in the maximum clique problem, $\mathcal{S}$ consists of the incidence vectors of all cliques in the input graph; in the maximum cut problem, $\mathcal{S}$ consists of the incidence vectors of all edge-cuts in the input graph; and so on. Applications of the cutting-plane method to these problems stimulated the development of the flourishing field of polyhedral combinatorics. Another important class of problems (1) are the *integer linear programming* problems, where $\mathcal{S}$ is specified as the set of all integer solutions of a prescribed system of linear inequalities. For this class, Gomory [1958] designed efficient procedures to generate cutting planes in a way that guarantees the cutting-plane method's termination.

The efficiency of the cutting-plane method is a different matter. Where the TSP is concerned, there are reasons to believe that the method may require

prohibitively large amounts of time even on certain reasonably small instances: R.M. Karp, E.L. Lawler, and R.E. Tarjan (see Karp [1972]) proved that the decision version of the TSP is an $\mathcal{NP}$-complete problem.

When $\mathcal{S}$ consists of all tours through a set of cities V, Dantzig, Fulkerson, and Johnson let the initial polyhedron consist of all vectors x, with components subscripted by edges of the complete graph on V, that satisfy

$$0 \leq x_e \leq 1 \qquad \text{for all edges } e, \tag{3}$$

$$\sum(x_e : v \in e) = 2 \qquad \text{for all cities } v. \tag{4}$$

In solving the 49-city problem, Dantzig, Fulkerson, and Johnson tightened this initial LP relaxation first by a number of *subtour inequalities*,

$$\sum(x_e : e \cap S \neq \emptyset, e - S \neq \emptyset) \geq 2 \quad \text{with } S \subset V,\ S \neq \emptyset,\ S \neq V, \tag{5}$$

and then by two additional cuts, after which x^* became the incidence vector of a tour; to show that these two inequalities are satisfied by incidence vectors of all tours, Dantzig, Fulkerson, and Johnson used ad hoc combinatorial arguments.

When an LP relaxation of a TSP instance includes all constraints (3), (4), a nonempty set of cuts can be found easily whenever $x^* \notin \mathcal{S}$: on the one hand, if x^* is not an integer vector, then Gomory's procedures find a nonempty set of cuts; on the other hand, if x^* is an integer vector, then it is the incidence vector of the edge-set of a disconnected graph and each connected component of this graph yields a subtour cut. This scheme is used, with embellishments, in the computer code of Martin [1966], which seems to be the first computer code for solving the TSP. Eventually, subtour inequalities became a staple of TSP cuts but, when new ways of finding TSP cuts emerged, Gomory cuts fell into disuse as TSP cuts.

1 Finding cuts

Hypergraph cuts

Given a subset S of V and given an x satisfying (3), (4), we write

$$\eta(S, x) = \sum(x_e : e \cap S \neq \emptyset, e - S \neq \emptyset) - 2.$$

A *hypergraph* is an ordered pair $(V, \mathcal{F})$ such that V is a finite set and $\mathcal{F}$ is a family of (not necessarily distinct) subsets of V; elements of V are called the *vertices* of the hypergraph and the elements of $\mathcal{F}$ are called the *edges* of the hypergraph. Given a hypergraph $(V, \mathcal{F})$ denoted $\mathcal{H}$, we write $\mathcal{H} \circ x = \sum(\eta(S, x) \ : \ S \in \mathcal{F})$ and we let $\mu(\mathcal{H})$ stand for the minimum of $\mathcal{H} \circ x$ taken over the incidence vectors of tours through V. Every linear inequality satisfied by all the incidence vectors of tours through V is the sum of a linear combination of equations (4) and a *hypergraph inequality*,

$$\mathcal{H} \circ x \geq t$$

with $t \leq \mu(\mathcal{H})$. Subtour inequalities are the simplest instances of hypergraph inequalities; one class of more complex instances is as follows.

The *intersection graph* of a hypergraph $(V, \mathcal{F})$ is the graph with vertex-set $\mathcal{F}$ and with two vertices adjacent if and only if these two members of $\mathcal{F}$ intersect. A *clique tree* is any hypergraph $\mathcal{H}$ such that

- the intersection graph of $\mathcal{H}$ is a tree

and such that the edge-set of $\mathcal{H}$ can be partitioned into a set of "handles" and a set of "teeth" with the following properties:

- there is at least one handle,
- the handles are pairwise disjoint,
- the teeth are pairwise disjoint,
- the number of teeth that each handle intersects is odd and at least three,
- each tooth includes a point that belongs to no handle.

Grötschel and Pulleyblank [1986] introduced this notion and proved that, for every clique-tree $\mathcal{H}$ with s teeth, the incidence vector x of any tour through V satisfies

$$\mathcal{H} \circ x \geq s - 1. \tag{6}$$

Let us give a short proof of this theorem here. Consider a clique tree with handles $H_1, \ldots, H_r$ and teeth $T_1, \ldots, T_s$; let t_j denote the number of handles that intersect tooth T_j and let h_i denote the number of teeth that intersect handle H_i; write

$$c_{ij} = \begin{cases} 1 & \text{if the tour includes an edge from } H_i \cap T_j \text{ to } T_j - H_i, \\ 0 & \text{otherwise.} \end{cases}$$

Since the teeth are pairwise disjoint, we have $\eta(H_i, x) \geq \sum_j c_{ij} - 2$; by definition, we have $\sum_j c_{ij} \leq h_i$; since $\eta(H_i, x)$ is even and h_i is odd, we conclude that

$$\eta(H_i, x) \geq 2 \sum_{j=1}^{s} c_{ij} - h_i - 1. \tag{7}$$

The restriction of the tour on a tooth T_j consists of $1 + \eta(T_j, x)/2$ segments; one of these segments passes through the point of T_j that belongs to no handle; since the handles are pairwise disjoint, each i such that $H_i \cap T_j \neq \emptyset$ and $c_{ij} = 0$ adds a new segment; we conclude that

$$\eta(T_j, x) \geq 2(t_j - \sum_{i=1}^{r} c_{ij}). \tag{8}$$

From (7) and (8), we obtain $\mathcal{H} \circ x \geq 2 \sum_j t_j - \sum_i h_i - r = \sum_j t_j - r$; since the intersection graph of $\mathcal{H}$ is a tree, we have $\sum_j t_j = r + s - 1$ and (6) follows.

Clique-trees with precisely one handle are called *combs* and the corresponding inequalities (6) are called *comb inequalities*.

FACET-INDUCING CUTS AND THE TEMPLATE PARADIGM

Some cuts are better than others. The ultimate measure of quality of a cut is its contribution to reducing the total running time of the cutting-plane method.

It is well known (Grötschel and Padberg [1975], Maurras [1975]) that the affine hull of the set $\mathcal{S}$ of all tours through V consists of all solutions x of (4); it follows that every cut is the sum of

- a linear combination of equations (4) and
- a nonnegative combination of linear inequalities that induce facets of the convex hull of $\mathcal{S}$.

Appealing to this fact, one may argue for preferring facet-inducing cuts to all others. This point of view suggests a two-phase paradigm for finding TSP cuts:

(i) describe linear inequalities that induce facets of the convex hull of $\mathcal{S}$,

(ii) for each template obtained in phase (i), design an efficient algorithm that, given an x^*, finds a cut matching that template, if such a cut exists.

Algorithms designed in phase (ii) are called *exact separation algorithms*; algorithms that attempt to find a cut matching the template, and may fail even if such a cut exists, are called *heuristic separation algorithms*.

The template paradigm was championed by Grötschel and Padberg [1979a, 1979b] and by Padberg and Hong [1980]. As for its phase (i), Grötschel and Padberg [1979a, 1979b] proved that both subtour inequalities and comb inequalities induce facets of the convex hull of $\mathcal{S}$; Grötschel and Pulleyblank [1986] proved that clique tree inequalities induce facets of the convex hull of $\mathcal{S}$; Naddef and Rinaldi [1998] proved that *path inequalities* (another generalization of comb inequalities, introduced by Cornuéjols, Fonlupt, and Naddef [1985]) induce facets of the convex hull of $\mathcal{S}$.

A polynomial-time exact separation algorithm for subtour inequalities was pointed out by Hong [1972]. It uses the observation that the problem of minimizing $\eta(S, x^*)$ subject to $S \subset V$, $S \neq \emptyset$, $S \neq V$ reduces to $|V|-1$ instances of the problem

$$\text{minimize } \eta(S, x^*) \text{ subject to } S \subset V,\ s \in S,\ t \notin S \tag{9}$$

with s fixed and t ranging through the remaining cities; it relies on the fact that (9) can be solved in polynomial time by variations on the max-flow min-cut theme of Ford and Fulkerson [1962]. The appeal of this scheme for actual computations is much enhanced when the input size is first reduced by "shrinking procedures" designed by Crowder and Padberg [1980] and by Padberg and Rinaldi [1990]; these procedures alone, without the subsequent max-flow min-cut computations, constitute fast heuristic separation algorithms for subtour inequalities.

A comb with each tooth having exactly two vertices is called a *blossom*. Padberg and Rao [1982] designed a polynomial-time exact separation algorithm for blossom inequalities. Their algorithm is an important tool in the computer codes of Grötschel and Holland [1991] and Padberg and Rinaldi [1991]: besides delivering blossom cuts, it is also used in heuristic separation algorithms for the more general comb inequalities. (The idea is to select sets S such that $\eta(S, x^*) = 0$ and to shrink each of these sets into a single vertex: blossom inequalities over the shrunken image of V yield comb inequalities over the original V.)

Other heuristic separation algorithms for comb inequalities, and for 2-handled clique tree inequalities, were designed by Padberg and Rinaldi [1991]; guided by the structure of the graph with vertex-set V and edge-set $\{e : 0 < x^*_e < 1\}$, they attempt to build the desired hypergraph in a greedy fashion. Heuristic separation

algorithms for path inequalities and other templates of TSP cuts were designed by Clochard and Naddef [1993] and by Christof and Reinelt [1995].

We have written a computer code for the TSP that follows in part the template paradigm. Our separation algorithms are

• an exact separation algorithm for subtour cuts that consists of Padberg-Rinaldi shrinking followed by repeated calls of the push-relabel method, as implemented by Cherkassky and Goldberg [1997], to solve max-flow min-cut problems,

• the Padberg-Rao exact separation algorithm for blossom cuts,

• the Grötschel-Holland and Padberg-Rinaldi heuristics for comb cuts,

• a greedy heuristic of the Clochard-Naddef kind for certain path cuts,

and heuristic separation algorithms that we have designed. Three of them that turned out to be important in solving the more difficult instances are as follows.

• Like most TSP codes, ours maintains the best tour $\overline{x}$ that we know of. One may suspect that, as both x^* and $\overline{x}$ approximate an optimal tour, sets S with $\eta(S, x^*) < 0$ are likely to satisfy $\eta(S, \overline{x}) < 2$, and so constitute single segments of the tour $\overline{x}$; our computational experience confirms this suspicion. We have designed an algorithm that, given x^* and $\overline{x}$, returns a family of segments $S_v (v \in V)$ of $\overline{x}$ such that each S_v minimizes $\eta(S, x^*)$ over all segments S that begin at v; its running time is in $O(m \log |V|)$, where m is the number of positive components of x^*. (We have used this algorithm not only in solving TSP instances, but also in computing lower bounds for TSP instances with up to 500,000 cities.)

• Having collected a family $\mathcal{F}$ of sets S such that $\eta(S, x^*) < 2$, we search for combs with handle H and teeth T_1, T_2, T_3 such that $H, T_1, T_2, T_3 \in \mathcal{F}$ and such that x^* violates the corresponding comb inequality. The search is guided by the observation that the desired $\{H, T_1, T_2, T_3\}$, as well as $\{H, T_1, T_2, V - T_3\}$, is a minimal family without the *consecutive ones property*; as an oracle for testing the consecutive ones property, we use *PQ-trees*, an efficient data structure designed by Booth and Lueker [1976].

• One way of showing that comb inequalities are satisfied by all tours is related to the framework for describing Gomory cuts propounded by Chvátal [1973]. We decided to turn the argument into an algorithm and search for comb cuts by solving certain systems of linear congruences mod 2. Our implementation of this plan uses PQ-trees once again, this time as a compact device for storing families of sets S such that $\eta(S, x^*) = 0$: variables in our system of linear congruences are in a one-to-one correspondence with Q-nodes of our PQ-tree. (Later on, Adam Letchford pointed out to us how our algorithm could be adjusted to search for the more general path cuts.)

BEYOND THE TEMPLATE PARADIGM

There are routine and well known algorithms that, given a finite subset $\mathcal{S}$ of some $\mathbf{R}^m$ and given a point x^* in $\mathbf{R}^m$, either express x^* as a convex combination of points in $\mathcal{S}$ or find a linear inequality that is satisfied by all points of $\mathcal{S}$ and violated by x^*. Using these algorithms directly to find cuts would be insane, since their running time is prohibitively long when m is large; using them in conjunction

with the trick of first projecting $\mathcal{S}$ and x^* into a lower-dimensional space was a crucial ingredient in our solution of a 13,509-city TSP instance.

Given x^*, we choose many different linear mappings $\phi : \mathbf{R}^m \to \mathbf{R}^d$; for each of our choices of ϕ, we express x^* as a convex combination of points in $\phi(\mathcal{S})$ or find a linear inequality $a^T\xi \geq b$ that is satisfied by all points ξ of $\phi(\mathcal{S})$ and $a^T\phi(x^*) < b$; in the latter case, inequality $a^T\phi(x) \geq b$ is a cut. This scheme is feasible as long as d is reasonably small: large size of $\phi(\mathcal{S})$ presents no difficulty provided that $\phi(\mathcal{S})$ can be accessed by an efficient oracle which, given any vector c in $\mathbf{R}^d$, returns an element ξ of $\phi(\mathcal{S})$ that maximizes $c^T\xi$.

True, it may happen that $\phi(x^*)$ belongs to the convex hull of $\phi(\mathcal{S})$ even though x^* lies outside the convex hull of $\mathcal{S}$; however, this is not always the case, and $\phi(\mathcal{S})$ is easier to handle than $\mathcal{S}$. Going a step further in this spirit adds flexibility to the method: for $\phi(\mathcal{S})$, we may substitute any $\mathcal{T}$ such that $\phi(\mathcal{S}) \subseteq \mathcal{T}$. True, it may happen that $\phi(x^*)$ belongs to the convex hull of $\mathcal{T}$ even though it lies outside the convex hull of $\phi(\mathcal{S})$; however, this is not always the case, and $\mathcal{T}$ may be easier to handle than $\phi(\mathcal{S})$.

Success of this method depends on making choices of ϕ and $\mathcal{T}$ in such a way that $\phi(x^*)$ has a reasonable chance of lying outside the convex hull of $\mathcal{T}$ and yet $\mathcal{T}$ is reasonably easy to handle. In the special case where $\mathcal{S}$ consists of all tours through a set V, our computer code makes each choice of ϕ by choosing a partition of V into nonempty sets $V_0, V_1, \ldots, V_k$. The corresponding ϕ is defined by shrinking each of these sets into a single point: the component of $\phi(x)$ that is indexed by i and j $(0 \leq i < j \leq k)$ has value $\sum(x_e : e \cap V_i \neq \emptyset, e \cap V_j \neq \emptyset)$. Our $\mathcal{T}$ consists of all nonnegative integer vectors ξ with components indexed by edges of the complete graph with vertex-set $\{0, 1, \ldots, k\}$ such that

- the graph with vertex-set $\{0, 1, \ldots, k\}$ and edge-set $\{e : \xi_e > 0\}$ is connected,
- $\sum(\xi_e : v \in e)$ is even whenever $v \in \{0, 1, \ldots, k\}$.

(Cornuéjols, Fonlupt, and Naddef [1985] call the problem of minimizing a prescribed linear function over this $\mathcal{T}$ the *graphical traveling salesman problem*.) We let k range between 8 and 30; our choices of $V_0, V_1, \ldots, V_k$ are guided by the structure of x^*; in particular, $\eta(V_j, x^*) = 0$ for all $j = 1, 2, \ldots, k$.

We do not know how useful this approach might prove in finding cuts for other problems (1); possibly its success in our experience with the TSP comes at least in part from the peculiar nature of the TSP; let us elaborate. The algorithm that we use to deal with $\mathcal{T}$ and $\phi(x^*)$ either expresses x^* as a convex combination of points in $\mathcal{T}$ or finds an inequality $a^T\xi \geq b$ that induces a facet of the convex hull of $\mathcal{T}$ and is violated by $\phi(x^*)$. In the latter case, we transform $a^T\xi \geq b$ into a hypergraph inequality $\mathcal{H} \circ \xi \geq t$ before substituting $\phi(x)$ for ξ; in our experience, these hypergraph inequalities are often (but not always) *tight triangular*; a conjecture implicit in the work of Naddef and Rinaldi [1992] suggests that, under this condition, inequality $\mathcal{H} \circ \phi(x) \geq t$ induces a facet of the convex hull of $\mathcal{S}$.

Another algorithm for finding TSP cuts that strays off the beaten path of the template paradigm, but starts from the Naddef-Rinaldi notion of tight triangular inequalities, has been designed by Carr [1998].

ALTERATIONS WHILE YOU WAIT

Watching our computer code run, we have observed that optimal solutions x^* of the successive LP relaxations often react to each new cut we add by shifting the defect prohibited by this cut to an area just beyond the cut's control. The remedy is obvious: we respond to each slight adjustment of $x^\star$ with a slight adjustment of our hypergraph cuts.

Given a hypergraph $\mathcal{H}$ with edges $E_1, \ldots, E_m$, we set

$$\alpha(I, \mathcal{H}) = \bigcap_{i \in I} E_i - \bigcup_{i \notin I} E_i$$

for each subset I of $\{1, \ldots, m\}$; we refer to each $\alpha(I, \mathcal{H})$ as an *atom* of $\mathcal{H}$; we write $\mathcal{H} \sqsubseteq \mathcal{H}'$ to signify that $\mathcal{H}$ and $\mathcal{H}'$ are hypergraphs with the same set of vertices and the same number of edges, and such that $\alpha(I, \mathcal{H}') \neq \emptyset$ whenever $\alpha(I, \mathcal{H}) \neq \emptyset$. It can be shown that $\mathcal{H} \sqsubseteq \mathcal{H}'$ implies $\mu(\mathcal{H}') \geq \mu(\mathcal{H})$. By *tightening* a hypergraph $\mathcal{H}$ with respect to a vector x^*, we mean a swift attempt to modify $\mathcal{H}$ in such a way that the resulting hypergraph, $\mathcal{H}'$, satisfies

- $\mathcal{H} \sqsubseteq \mathcal{H}'$ and $\mathcal{H}' \circ x^* < \mathcal{H} \circ x^*$.

We tighten $\mathcal{H}$ by a greedy algorithm that moves single vertices from one atom to another if such a move decreases $\mathcal{H} \circ x^*$ (or, with some restrictions, if the move at least does not increase $\mathcal{H} \circ x^*$). Some of these permissible moves are more appealing than others; all of them are kept in a priority queue, which is updated after each move is made. We make extensive use of tightening in our computer code. Every cut that we find is tightened before it is added to the LP relaxation. We also periodically run through all constraints of the LP relaxation and tighten each of them.

We use one additional technique for adjusting comb inequalities. Let us refer to a comb with some of its teeth removed as a *generalized comb*; let us say that a tooth of a generalized comb is *big* if its size is at least three; for every generalized comb $\mathcal{H}_0$, let $\Delta(\mathcal{H}_0, x^*)$ denote the minimum of $\mathcal{H} \circ x^* - \mu(\mathcal{H})$ over all combs $\mathcal{H}$ such that $\mathcal{H}$ and $\mathcal{H}_0$ have the same handle and all big teeth of $\mathcal{H}$ are teeth of $\mathcal{H}_0$. We have designed a dynamic programming algorithm that, given a generalized comb $\mathcal{H}_0$, finds either

- a comb $\mathcal{H}$ such that all big teeth of $\mathcal{H}$ are teeth of $\mathcal{H}_0$ and, if $\Delta(\mathcal{H}_0) \leq 0$, then $\mathcal{H} \circ x^* - \mu(\mathcal{H}) \leq \Delta(\mathcal{H}_0)$

or else a subtour inequality violated by x^*. We refer to this algorithm as *teething*, and we apply it to comb constraints in the LP relaxation.

2 THE BRANCH-AND-CUT METHOD

Progress of the cutting-plane method towards solving a particular problem instance is often estimated by the increase in the optimal value of its LP relaxation; as more and more cuts are added, these increases tend to get smaller and smaller. When they become too small, the sensible thing is to *branch*: having partitioned the set $\mathcal{S}$ of tours into sets $\mathcal{S}_1, \mathcal{S}_2$, apply the cutting-plane method first to one of the subproblems

$$\text{minimize } c^T x \text{ subject to } x \in \mathcal{S}_i$$

and then to the other. At some later time, one or both of these subproblems can be split into sub-subproblems, and so on. In the resulting binary tree of subproblems, each leaf has been either solved by the cutting-plane method without recourse to branching or else found irrelevant when the optimal value of its LP relaxation turned out to be at least as large as the cost of a previously known tour. The standard way of splitting a problem into subproblems is

$$\mathcal{S}_1 = \{x \in \mathcal{S} : x_e = 0\}, \quad \mathcal{S}_2 = \{x \in \mathcal{S} : x_e = 1\} \tag{10}$$

for a suitably chosen edge e; Clochard and Naddef [1993] advocated

$$\mathcal{S}_1 = \{x \in \mathcal{S} : \eta(S, x) = 0\}, \quad \mathcal{S}_2 = \{x \in \mathcal{S} : \eta(S, x) \geq 2\} \tag{11}$$

for a suitably chosen subset S of V. Our computer code chooses the most appealing of all options (10), (11); in our experience with the larger TSP instances, this policy reduces the size of the tree of subproblems.

Every subproblem in the tree has the form

$$\text{minimize } c^T x \text{ subject to } x \in \mathcal{S},\ Cx \leq d$$

for some system $Cx \leq d$ of linear inequalities. When this subproblem is attacked by the cutting-plane method, the initial LP relaxation is

$$\text{minimize } c^T x \text{ subject to } Ax \leq b,\ Cx \leq d$$

with $\mathcal{S} \subseteq \{x : Ax \leq b\}$ and each cut added to $Ax \leq b,\ Cx \leq d$ is satisfied by all x in $\mathcal{S} \cap \{x : Cx \leq d\}$; this is a variant of the *branch-and-bound* method. In the *branch-and-cut* method, used by Hong [1972], Miliotis [1976], Padberg and Rinaldi [1987, 1991], and others, cuts are restricted to those satisfied by all x in $\mathcal{S}$ and added to $Ax \leq b$; this system, acquiring more and more inequalities as more and more subproblems are being processed, may be used to initialize the cutting-plane method on any as yet unprocessed subproblem. Our computer code uses the branch-and-cut method.

3 Experimental results

Computer codes for the TSP have become increasingly more sophisticated over the years. A conspicuous sign of these improvements is the increasing size of the nontrivial instances that have been solved: a 120-city problem by Grötschel [1980], a 318-city problem by Crowder and Padberg [1980], a 532-city problem by Padberg and Rinaldi [1987], a 666-city problem by Grötschel and Holland [1991], a 1,002-city problem and a 2,392-city problem by Padberg and Rinaldi [1991].

In the table below, we report the results of running our computer code on these instances, as well as on five others. With the exception of the 13,509-city instance, our code was run on a single processor of a Digital AlphaServer 4100 (400 MHz). The 13,509-city TSP was run on a network of 48 workstations, including Digital Alphas, Intel Pentium IIs and Pentium Pros, and Sun UltraSparcs. The

Name	Cities	Tree of subproblems	Running time
gr120	120	1 node	3.3 seconds
lin318	318	1 node	24.6 seconds
pr1002	1,002	1 node	94.7 seconds
gr666	666	1 node	260.0 seconds
att532	532	3 nodes	294.3 seconds
pr2392	2,392	1 node	342.2 seconds
ts225	225	1 node	438.9 seconds
pcb3038	3,038	193 nodes	1.5 days
fnl4461	4,461	159 nodes	1.7 days
pla7397	7,397	129 nodes	49.5 days
usa13509	13,509	9,539 nodes	∼10 years

reported time for this instance is an estimate of the cumulative CPU time spent on the individual machines.

The problems reported in the table come from the set TSPLIB of test instances collected by Reinelt [1991]. We sorted them by their solution time, rather than by their size, to emphasize that the difficulty of an instance depends on factors other than just its number of cities. In particular, ts225 is a contrived nasty instance that was first solved only in 1994—three years after it first appeared in TSPLIB. We will present results for the full set of 110 TSPLIB problems in a comprehensive report of our TSP work that we are preparing.

Our computer code (written in the C programming language) is available for research purposes. It can be obtained over the internet at the page:

`http://www.caam.rice.edu/~keck/concorde.html`

REFERENCES

[1968] M. Bellmore and G.L. Nemhauser, "The traveling salesman problem: a survey", *Operations Research* 16 (1968) 538–558.

[1976] K.S. Booth and G.S. Lueker, "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms", *Journal of Computer Systems and Science* 13 (1976) 335–379.

[1998] R. Carr, "Separation and lifting of TSP inequalities", manuscript, 1998.

[1997] B.V. Cherkassky and A.V. Goldberg, "On implementing the push-relabel method for the maximum flow problem", *Algorithmica* 19 (1997) 390–410.

[1995] T. Christof and G. Reinelt, "Parallel cutting plane generation for the TSP", in: *Parallel Programming and Applications* (P. Fritzson, L. Finmo, eds.), IOS Press, 1995, pp. 163–169.

[1995] V.Chvátal, "Edmonds polytopes and a hierarchy of combinatorial problems", *Discrete Mathematics* 4 (1973) 305–337.

[1993] J.-M. Clochard and D. Naddef, "Using path inequalities in a branch and cut code for the symmetric traveling salesman problem", in: *Third IPCO Conference* (G. Rinaldi and L. Wolsey, eds), 1993, pp. 291–311.

[1985] G. Cornuéjols, J. Fonlupt, and D. Naddef, "The traveling salesman problem on a graph and some related integer polyhedra", *Mathematical Programming* 33 (1985) 1–27.

[1980] H. Crowder and M.W. Padberg, "Solving large-scale symmetric travelling salesman problems to optimality", *Management Science* 26 (1980) 495–509.

[1954] G.B. Dantzig, R. Fulkerson, and S.M. Johnson, "Solution of a large-scale traveling salesman problem", *Operations Research* 2 (1954) 393–410.

[1962] L.R. Ford, Jr. and D.R. Fulkerson, *Flows in Networks*, Princeton University Press, 1962.

[1958] R.E. Gomory, "Outline of an algorithm for integer solutions to linear programs", *Bulletin of the American Mathematical Society* 64 (1958) 275–278.

[1980] M. Grötschel, "On the symmetric travelling salesman problem: solution of a 120-city problem", *Mathematical Programming Study* 12 (1980) 61–77.

[1991] M. Grötschel and O. Holland, "Solution of large-scale symmetric travelling salesman problems", *Mathematical Programming* 51 (1991) 141–202.

[1975] M. Grötschel and M.W. Padberg, *On the Symmetric Travelling Salesman Problem*, Report No.7536-OR, Institut für Ökonometrie und Operations Research, Universität Bonn, 1975.

[1979a] M. Grötschel and M.W. Padberg, "On the symmetric travelling salesman problem I: Inequalities", *Mathematical Programming* 16 (1979) 265–280.

[1979b] M. Grötschel and M.W. Padberg, "On the symmetric travelling salesman problem II: lifting theorems and facets", *Mathematical Programming* 16 (1979) 281–302.

[1986] M. Grötschel and W.R. Pulleyblank, "Clique tree inequalities and the symmetric travelling salesman problem", *Mathematics of Operations Research* 11 (1986) 1–33.

[1972] S. Hong, *A Linear Programming Approach for the Traveling Salesman Problem*, Ph.D. Thesis, The Johns Hopkins University, 1972.

[1942] R.J. Jessen, "Statistical investigation of a sample survey for obtaining farm facts", *Research Bulletin* #304, Iowa State College of Agriculture, 1942.

[1995] M. Jünger, G. Reinelt, and G. Rinaldi, "The traveling salesman problem", in: *Handbook on Operations Research and Management Sciences: Networks* (M. Ball, T. Magnanti, C.L. Monma, and G. Nemhauser, eds.), North-Holland, 1995, pp. 225–330.

[1972] R.M. Karp, "Reducibility among combinatorial problems", in: *Complexity of Computer Computations* (R.E. Miller and J.W. Thatcher, eds.), Plenum Press, 1972, pp. 85–103.

[1955] H.W. Kuhn, "On certain convex polyhedra", Abstract 799t, *Bulletin of the American Mathematical Society* 61 (1955) 557–558.

[1985] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys, eds., *The Traveling Salesman Problem*, Wiley, Chichester, 1985.

[1940] P.C. Mahalanobis, "A sample survey of the acreage under jute in Bengal", *Sankhyu* 4 (1940) 511–530.

[1966] G.T. Martin, "Solving the traveling salesman problem by integer linear programming", *Operations Research* 14 (Supplement 1), Abstract WA7.10.

[1975] J.F. Maurras, "Some results on the convex hull of Hamiltonian cycles of symmetric complete graphs", in: *Combinatorial Programming: Methods and Applications* (B. Roy, ed.), Reidel, Dordrecht, 1975, pp. 179–190.

[1976] P. Miliotis, "Integer programming approaches to the travelling salesman problem", *Mathematical Programming* 10 (1976) 367–378.

[1992] D. Naddef and G. Rinaldi, "The graphical relaxation: A new framework for the symmetric traveling salesman polytope", *Mathematical Programming* 58 (1992) 53–88.

[1998] D. Naddef and G. Rinaldi, "The symmetric traveling salesman polytope: New facets from the graphical relaxation", in preparation.

[1980] M.W. Padberg and S. Hong, "On the symmetric travelling salesman problem: a computational study", *Mathematical Programming Study* 12 (1980) 78–107.

[1982] M.W. Padberg and M.R. Rao, "Odd minimum cut-sets and b-matchings", *Mathematics of Operations Research* 7 (1982) 67–80.

[1987] M. Padberg and G. Rinaldi, "Optimization of a 532-city symmetric traveling salesman problem by branch and cut", *Operations Research Letters* 6 (1987) 1–7.

[1990] M. Padberg and G. Rinaldi, "An efficient algorithm for the minimum capacity cut problem", *Mathematical Programming* 47 (1990) 19–36.

[1991] M. Padberg and G. Rinaldi, "A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems", *SIAM Review* 33 (1991) 60–100.

[1991] G. Reinelt, "TSPLIB - A traveling salesman library", *ORSA Journal on Computing* 3 (1991) 376–384.

[1994] G. Reinelt, *The Traveling Salesman: Computational Solutions for TSP Applications*, Springer-Verlag, Berlin, 1994.

[1949] J.B. Robinson, "On the Hamiltonian game (a traveling-salesman problem)", *RAND Research Memorandum* RM-303, 1949.

David Applegate
Rice University
Computational and Applied Math
Houston, TX 77005-1892
david@caam.rice.edu

Robert Bixby
Rice University
Computational and Applied Math
Houston, TX 77005-1892
bixby@caam.rice.edu

Vašek Chvátal
Rutgers University
Department of Computer Science
New Brunswick, NJ 08903
chvatal@cs.rutgers.edu

William Cook
Rice University
Computational and Applied Math
Houston, TX 77005-1892
bico@caam.rice.edu

Semidefinite Programming and Combinatorial Optimization

Michel X. Goemans[1]

Abstract. We describe a few applications of semidefinite programming in combinatorial optimization.

1991 Mathematics Subject Classification: 90C25, 90C10, 90C27, 05C50, 05C60, 68R10.
Keywords and Phrases: Convex optimization, combinatorial optimization, semidefinite programming, eigenvalue bounds.

Semidefinite programming is a special case of convex programming where the feasible region is an affine subspace of the cone of positive semidefinite matrices. There has been much interest in this area lately, partly because of applications in combinatorial optimization and in control theory and also because of the development of efficient interior-point algorithms.

The use of semidefinite programming in combinatorial optimization is not new though. Eigenvalue bounds have been proposed for combinatorial optimization problems since the late 60's, see for example the comprehensive survey by Mohar and Poljak [20]. These eigenvalue bounds can often be recast as semidefinite programs [1]. This reformulation is useful since it allows to exploit properties of convex programming such as duality and polynomial-time solvability, and it avoids the pitfalls of eigenvalue optimization such as non-differentiability. An explicit use of semidefinite programming in combinatorial optimization appeared in the seminal work of Lovász [16] on the so-called theta function, and this lead Grötschel, Lovász and Schrijver [9, 11] to develop the only known (and non-combinatorial) polynomial-time algorithm to solve the maximum stable set problem for perfect graphs.

In this paper, we describe a few applications of semidefinite programming in combinatorial optimization. Because of space limitations, we restrict our attention to the Lovász theta function, the maximum cut problem [8], and the automatic generation of valid inequalities à la Lovász-Schrijver [17, 18]. This survey is much inspired by another (longer) survey written by the author [7]. However, new results on the power and limitations of the Lovász-Schrijver procedure are presented as well as a study of the maximum cut relaxation for graphs arising from association schemes.

[1]Supported in part by NSF contract 9623859-CCR.

1 PRELIMINARIES

In this section, we collect several basic results about positive semidefinite matrices and semidefinite programming.

Let M_n denote the cone of $n \times n$ matrices (over the reals), and let S_n denote the subcone of symmetric $n \times n$ matrices. A matrix $A \in S_n$ is said to be *positive semidefinite* if its associated quadratic form $x^T Ax$ is nonnegative for all $x \in R^n$. The positive semidefiniteness of a matrix A will be denoted by $A \succeq 0$; similarly, we write $A \succeq B$ for $A - B \succeq 0$. The cone of positive semidefinite matrices will be denoted by PSD_n. The following statements are equivalent for a symmetric matrix A: (i) A is positive semidefinite, (ii) all eigenvalues of A are nonnegative, and (iii) there exists a matrix B such that $A = B^T B$. (iii) gives a representation of $A = [a_{ij}]$ as a *Gram matrix*: there exist vectors v_i such that $a_{ij} = v_i^T v_j$ for all i, j. Given a symmetric positive semidefinite matrix A, a matrix B satisfying (iii) can be obtained in $O(n^3)$ time by a Cholesky decomposition.

Given $A, B \in M_n$, the (Frobenius) inner product $A \bullet B$ is defined by $A \bullet B = Tr(A^T B) = \sum_i \sum_j A_{ij} B_{ij}$. The quadratic form $x^T Ax$ can thus also be written as $A \bullet (xx^T)$. Since the extreme rays of PSD_n are of the form xx^T, we derive that $A \bullet B \geq 0$ whenever $A, B \succeq 0$. We can also similarly derive Fejer's theorem which says that PSD_n is self-polar, i.e. $PSD_n^* = \{A \in S_n : A \bullet B \geq 0 \text{ for all } B \succeq 0\} = PSD_n$.

Semidefinite programs are linear programs over the cone of positive semidefinite matrices. They can be expressed in many equivalent forms, e.g.

$$\begin{array}{rcll} SDP & = & \inf & C \bullet Y \\ \text{subject to:} & & & A_i \bullet Y = b_i \qquad i = 1, \cdots, m \\ & & & Y \succeq 0. \end{array} \tag{1}$$

In general a linear program over a pointed closed convex cone K is formulated as $z = \inf\{c^T x : Ax = b, x \in K\}$, and its dual (see [22]) is $w = \sup\{b^T y : A^T y + s = c, s \in K^*\}$ where $K^* = \{a : a^T b \geq 0 \text{ for all } b \in K\}$. Weak duality always holds: $c^T x - y^T b = (A^T y + s)^T x - y^T Ax = s^T x$ for any primal feasible x and dual feasible y. If we assume that A has full row rank, $\{x \in \text{int}K\} \neq \emptyset$, and $\{(y, s) : A^T y + s = c, s \in \text{int } K^*\} \neq \emptyset$, then $z = w$ and both the primal and dual problems attain their optimum value. In the case of semidefinite programs, the dual to (1) is $\sup\{\sum_{i=1}^n b_i y_i : \sum_i y_i A_i \preceq C\}$.

Semidefinite programs can be solved (more precisely, approximated) in polynomial-time within any specified accuracy either by the ellipsoid algorithm [9, 11] or more efficiently through interior-point algorithms. For the latter, we refer the reader to [22, 1, 24]. The above algorithms produce a strictly feasible solution (or slightly infeasible for some versions of the ellipsoid algorithm) and, in fact, the problem of deciding whether a semidefinite program is feasible (exactly) is still open. However, we should point out that since $\begin{pmatrix} 1 & x \\ x & a \end{pmatrix} \succeq 0$ iff $|x| \leq \sqrt{a}$, a special case of semidefinite programming feasibility is the square-root sum problem: given $a_1, \cdots, a_n$ and k, decide whether $\sum_{i=1}^n \sqrt{a_i} \leq k$. The complexity of this problem is still open.

2 LOVÁSZ'S THETA FUNCTION

Given a graph $G = (V, E)$, a stable (or independent) set is a subset S of vertices such that no two vertices of S are adjacent. The maximum cardinality of a stable set is the stability number (or independence number) of G and is denoted by $\alpha(G)$. In a seminal paper [16], Lovász proposed an upper bound on $\alpha(G)$ known as the theta function $\vartheta(G)$. The theta function can be expressed in many equivalent ways, as an eigenvalue bound, as a semidefinite program, or in terms of orthogonal representations. These formulations will be summarized in this section. We refer the reader to the original paper [16], to Chapter 9 in Grötschel et al. [11], or to the survey by Knuth [15] for additional details.

As an eigenvalue bound, $\vartheta(G)$ can be derived as follows. Consider $P = \{A \in S_n : a_{ij} = 1 \text{ if } (i,j) \notin E \text{ (or } i = j)\}$. If there exists a stable set of size k, the corresponding principal submatrix of any $A \in P$ will be J_k, the all ones matrix of size k. By a classical result on interlacing of eigenvalues for symmetric matrices (see [13]), we derive that $\lambda_{max}(A) \geq \lambda_{max}(J_k) = k$ for any $A \in P$, where $\lambda_{max}(\cdot)$ denotes the largest eigenvalue. As a result, $\min_{A \in P} \lambda_{max}(A)$ is an upper bound on $\alpha(G)$, and this is one of the equivalent formulations of Lovász's theta function.

This naturally leads to a semidefinite program. Indeed, the largest eigenvalue of a matrix can easily be formulated as a semidefinite program: $\lambda_{max}(A) = \min\{t : tI - A \succeq 0\}$. In order to express $\vartheta(G)$ as a semidefinite program, we observe that $A \in P$ is equivalent to $A - J$ being generated by E_{ij} for $(i,j) \in E$, where all entries of E_{ij} are zero except for (i,j) and (j,i). Thus, we can write

$$\begin{aligned} \vartheta(G) \quad = \quad \min \quad & t \\ \text{subject to:} \quad & tI + \sum_{(i,j) \in E} x_{ij} E_{ij} \succeq J. \end{aligned}$$

By strong duality, we can also write:

$$\begin{aligned} \vartheta(G) \quad = \quad \max \quad & J \bullet Y & & (2) \\ \text{subject to:} \quad & y_{ij} = 0 & (i,j) \in E & (3) \\ & I \bullet Y = 1 & (i.e.\ \ Tr(Y) = 1) & (4) \\ & Y \succeq 0. & & (5) \end{aligned}$$

Lovász's first definition of $\vartheta(G)$ was in terms of orthonormal representations. An *orthonormal representation* of G is a system $v_1, \cdots, v_n$ of unit vectors in R^n such that v_i and v_j are orthogonal (i.e. $v_i^T v_j = 0$) whenever i and j are not adjacent. The value of the orthonormal representation is $z = \min_{c:||c||=1} \max_{i \in V} \frac{1}{(c^T u_i)^2}$. This is easily seen to be an upper bound on $\alpha(G)$ (since $||c||^2 \geq \sum_{i \in S} (c^T u_i)^2 \geq |S|/z$ for any stable set S). Taking the minimum value over all orthonormal representations of G, one derives another expression for $\vartheta(G)$. This result can be restated in a slightly different form. If x denotes the incidence vector of a stable set then we have that

$$\sum_i (c^T v_i)^2 x_i \leq 1. \tag{6}$$

In other words, the *orthonormal representation constraints* (6) are valid inequalities for $STAB(G)$, the convex hull of incidence vectors of stable sets of G. Grötschel et al. [10] show that if we let $TH(G) = \{x : x \text{ satisfies (6) and } x \geq 0\}$, then $\vartheta(G) = \max\{\sum_i x_i : x \in TH(G)\}$. Yet more formulations of ϑ are known.

2.1 Perfect Graphs

A graph G is called *perfect* if, for every induced subgraph G', its chromatic number is equal to the size of the largest clique in G'. Even though perfect graphs have been the focus of intense study, there are still important questions which are still open. The strong perfect graph conjecture of Berge claims that a graph is perfect if and only if it does not contain an odd cycle of length at least five or its complement. It is not even known if the recognition problem of deciding whether a graph is perfect is in P or NP-complete. However, the theta function gives some important characterizations (but not a "good" or NP∩co-NP characterization) of perfect graphs.

THEOREM 1 (GRÖTSCHEL ET AL. [10]) *The following are equivalent:*

- *G is perfect,*
- *$TH(G) = \{x \geq 0 : \sum_{i \in C} x_i \leq 1$ for all cliques $C\}$*
- *$TH(G)$ is polyhedral.*

Moreover, even though recognizing perfect graphs is still open, one can find the largest stable set in a perfect graph in polynomial time by computing the theta function using semidefinite programming (Grötschel et al. [9, 11]); similarly one can solve the weighted problem, or find the chromatic number or the largest clique. Observe that if we apply this algorithm to a graph which is not necessarily perfect, we would either find the largest stable set or have a proof that the graph is not perfect.

Although $\vartheta(G) = \alpha(G)$ for perfect graphs, $\vartheta(G)$ can provide a fairly poor upper bound on $\alpha(G)$ for general graphs. Feige [6] has shown the existence of graphs for which $\vartheta(G)/\alpha(G) \geq \Omega(n^{1-\epsilon})$ for any $\epsilon > 0$. See [7] for further details and additional references on the quality of $\vartheta(G)$.

3 The Maximum Cut Problem

Given a graph $G = (V, E)$, the cut $\delta(S)$ induced by vertex set S consists of the set of edges with exactly one endpoint in S. In the NP-hard maximum cut problem (MAX CUT), we would like to find a cut of maximum total weight in a weighted undirected graph. The weight of $\delta(S)$ is $w(\delta(S)) = \sum_{e \in \delta(S)} w_e$. In this section, we describe an approach of the author and Williamson [8] based on semidefinite programming.

The maximum cut problem can be formulated as an integer quadratic program. If we let $y_i = 1$ if $i \in S$ and $y_i = -1$ otherwise, the value of the cut

$\delta(S)$ can be expressed as $\sum_{(i,j)\in E} w_{ij}\frac{1}{2}(1-y_iy_j)$. Suppose we consider the matrix $Y=[y_iy_j]$. This is a positive semidefinite rank one matrix with all diagonal elements equal to 1. Relaxing the rank one condition, we derive a semidefinite program giving an upper bound SDP on OPT:

$$\begin{array}{rcll} SDP & = & \max & \frac{1}{2}\sum_{(i,j)\in E} w_{ij}(1-y_{ij}) \\ \text{subject to:} & & y_{ii}=1 & i\in V \\ & & Y=[y_{ij}]\succeq 0. & \end{array} \tag{7}$$

It is convenient to write the objective function in terms of the (weighted) *Laplacien* matrix $L(G)=[l_{ij}]$ of G: $l_{ij}=-w_{ij}$ for all $i\neq j$ and $l_{ii}=\sum_j w_{ij}$. For any matrix Y, we have $L(G)\bullet Y=\sum_{(i,j)\in E} w_{ij}(y_{ii}+y_{jj}-2y_{ij})$ (in particular, if $Y=yy^T$ then we obtain the classical equality $y^TL(G)y=\sum_{(i,j)\in E} w_{ij}(y_i-y_j)^2$). As a result, the objective function can also be expressed as $\frac{1}{4}L(G)\bullet Y$.

The dual of this semidefinite program is $SDP=\frac{1}{4}\min\{\sum_j d_j : diag(d)\succeq L(G)\}$. This can also be rewritten as

$$SDP=\frac{1}{4}n\min_{u:\sum_i u_i=0}\lambda_{max}(L+diag(u)). \tag{8}$$

This eigenvalue bound was proposed and analyzed by Delorme and Poljak [4, 3]. In their study, they conjectured that the worst-case ratio OPT/SDP is $32/(25+5\sqrt{5})\sim 0.88445$ for nonnegative weights and achieved by the 5-cycle. By exploiting (7), Goemans and Williamson [8] derived a randomized algorithm that produces a cut whose expected value is at least $0.87856\,SDP$, implying that $OPT/SDP\geq 0.87856$ for nonnegative weights. We describe their *random hyperplane technique* and their elementary analysis below.

Consider any feasible solution Y to (7). Since Y admits a Gram representation, there exist unit vectors $v_i\in R^d$ (for some $d\leq n$) for $i\in V$ such that $y_{ij}=v_i^Tv_j$. Let r be a vector uniformly generated from the unit sphere in R^d, and consider the cut induced by the hyperplane $\{x : r^Tx=0\}$ normal to r, i.e. the cut $\delta(S)$ where $S=\{i\in V : r^Tv_i\geq 0\}$. By elementary arguments, the probability that v_i and v_j are separated is precisely θ/π, where $\theta=\arccos(v_i^Tv_j)$ is the angle between v_i and v_j. Thus, the expected weight of the cut is exactly given by:

$$E[w(\delta(S))]=\sum_{(i,j)\in E} w_{ij}\frac{\arccos(v_i^Tv_j)}{\pi}. \tag{9}$$

Comparing this expression term by term to the objective function of (7) and using the fact that $\arccos(x)/\pi\geq\alpha\frac{1}{2}(1-x)$ where $\alpha=0.87856\cdots$, we derive that $E[w(\delta(S))]\geq\alpha\frac{1}{4}L(G)\bullet Y$. Hence if we apply the random hyperplane technique to a feasible solution Y of value $\geq(1-\epsilon)SDP$ (which can be obtained in polynomial time), we obtain a random cut of expected value greater or equal to $\alpha(1-\epsilon)SDP\geq 0.87856\,SDP\geq 0.87856\,OPT$. Mahajan and Ramesh [19] have

shown that this technique can be derandomized, therefore giving a deterministic 0.87856-approximation algorithm for MAX CUT.

The worst-case value for OPT/SDP is thus somewhere between 0.87856 and 0.88446, and even though this gap is small, it would be very interesting to prove Delorme and Poljak's conjecture that the worst-case is given by the 5-cycle. This would however require a new technique. Indeed, Karloff [14] has shown that the analysis of the random hyperplane technique is tight, namely there exists a family of graphs for which the expected weight $E[w(\delta(S)]$ of the cut produced is arbitrarily close to $\alpha\, SDP$.

No better approximation algorithm is currently known for MAX CUT. On the negative side though, Håstad [12] has shown that it is NP-hard to approximate MAX CUT within $16/17 + \epsilon = 0.94117\cdots$ for any $\epsilon > 0$. Furthermore, Håstad shows that if we replace the objective function by $\frac{1}{2}\sum_{(i,j)\in E_1} w_{ij}(1 - y_i y_j) + \frac{1}{2}\sum_{(i,j)\in E_2} w_{ij}(1 + y_i y_j)$, then the resulting problem is NP-hard to approximate within $11/12 + \epsilon = 0.91666\cdots$, while the random hyperplane technique still gives the same guarantee of $\alpha \sim 0.87856$.

The analysis of the random hyperplane technique can be generalized following an idea of Nesterov [21] for more general Boolean quadratic programs. First observe that (9) can be rewritten as $E[w(\delta(S))] = \frac{1}{2\pi} L(G) \bullet \arcsin(Y)$, where $\arcsin(Y) = [\arcsin(y_{ij})]$. Suppose now that we restrict our attention to weight functions for which $L(G) \in K$ for a certain cone K. Then a bound of α would follow if we can show that $L(G) \bullet (\frac{2}{\pi}\arcsin(Y)) \geq L(G) \bullet (\alpha Y)$ or $L(G) \bullet \left(\frac{2}{\pi}\arcsin(Y) - \alpha Y\right) \geq 0$. This corresponds to showing that $\left(\frac{2}{\pi}\arcsin(Y) - \alpha Y\right) \in K^*$, where K^* is the polar cone to K. For several interesting cones K (e.g. the cone of positive semidefinite matrices), this analysis can be performed.

We now describe a situation in which the semidefinite programming relaxation simplifies considerably. This is similar to the well-known LP bound in coding introduced by Delsarte [5] which corresponds to the theta function for graphs arising from association schemes. The results briefly sketched below were obtained jointly with F. Rendl.

Consider graphs whose adjacency matrix can be written as $\sum_{i\in M} A_i$ where $M \subseteq \{1, \cdots, l\}$ and $A_0, A_1, \cdots, A_l$ are $n \times n$ $0-1$ symmetric matrices forming an association scheme (see [2]):

1. $A_0 = I$,
2. $\sum_{i=0}^{l} A_i = J$,
3. there exist p_{ij}^k $(0 \leq i, j, k \leq l)$ such that $A_i A_j = A_j A_i = \sum_{k=0}^{n} p_{ij}^k A_k$.

When $l = 2$, the graph with incidence matrix A_1 (or A_2) is known as a *strongly regular* graph.

We list below properties of association schemes, for details see for example [2]. Since the A_i's commute, they can be diagonalized simultaneously and thus they share a set of eigenvectors. Furthermore, the (Bose-Mesner) algebra $\mathcal{A}$ generated by the A_i's has a unique basis of minimal idempotents (i.e. $E^2 = E$) $E_0, \cdots, E_l$. These matrices E_i's are positive semidefinite (since their eigenvalues are all 0 or 1

by idempotence), and have constant diagonal equal to μ_i/n where μ_i is the rank of E_i.

For association schemes, we can show that the optimum correcting vector in (8) is $u = 0$, giving $SDP = \frac{n}{4}\lambda_{max}(L(G))$, and that the optimum primal solution Y is equal to nE_p/μ_p where p is the index corresponding to the eigenspace of the largest eigenvalue of $L(G)$. To see this optimality, one simply needs to realize that $Z = \lambda_{max}(L(G))I - L(G)$ can be expressed as $\sum_{i \neq p} c_i E_i$ and, as a result, satisfies complementary slackness with nE_p/μ_p: $ZE_p = 0$. Furthermore, if we were to add valid inequalities of the form $C_i \bullet Y \leq b_i$ with $C_i \in \mathcal{A}$ to the primal semidefinite program then the primal and dual SDPs can be seen to reduce to a dual pair of linear programs:

$$\begin{array}{llll} \frac{1}{4}\max & \sum_j (L(G) \bullet E_j)x_j & = & \frac{1}{4}\min \quad ns + \sum_i b_i z_i \\ \text{s.t.} & \sum_j \mu_j x_j = n & & \text{s.t.} \quad \mu_j s + \sum_i (C_i \bullet E_j) z_i \geq L \bullet E_j \quad \forall j \\ & \sum_j (C_i \bullet E_j) x_j = b_i \quad \forall i & & z_i \geq 0 \qquad \forall i \\ & x_j \geq 0 \qquad \forall j & & \end{array}$$

The primal semidefinite solution is then $\sum_j x_j E_j$ and the dual constraints imply that $sI + \sum_i z_i C_i \succeq L(G)$. As an illustration, the triangle inequalities can be aggregated in order to be of the required form, and thus the semidefinite program with triangle inequalities can be solved as a linear program for association schemes.

4 Deriving Valid Inequalities

Lovász and Schrijver [17, 18] have proposed a technique for automatically generating stronger and stronger formulations for integer programs. We briefly describe their approach here and discuss its power and its limitations.

Let $P = \{x \in R^n : Ax \geq b, 0 \leq x \leq 1\}$, and let $P_0 = conv(P \cap \{0,1\}^n)$ denote the convex hull of $0-1$ solutions. Suppose we multiply a valid inequality $\sum_i c_i x_i - d \geq 0$ for P by either $1 - x_j \geq 0$ or by $x_j \geq 0$. We obtain a quadratic inequality that we can linearize by replacing $x_i x_j$ by a new variable y_{ij}. Since we are interested only in 0-1 solutions, we can impose that $x_i^2 = x_i$ for all i. Replacing x_i by y_{ii}, we therefore obtain a linear ("matrix") inequality on the entries of Y. Let $M(P)$ denote the set of all symmetric matrices satisfying all the matrix inequalities that can be derived in this way, and let $N(P) = \{x : Y \in M(P), x = Diag(Y)\}$, where $Diag(Y)$ denotes the diagonal of Y; thus $N(P)$ is a projection of $M(P)$. By construction, we have that $P_0 \subseteq N(P) \subseteq P$. They also consider a much stronger operator involving semidefinite constraints. Observe that, for any 0-1 solution x, the matrix Y defined above as xx^T must satisfy $Y - Diag(Y)Diag(Y)^T = 0$. This is again an (intractable) quadratic inequality but it can be relaxed to $Y - Diag(Y)Diag(Y)^T \succeq 0$. Viewing $Y - Diag(Y)Diag(Y)^T$ as a Schur complement, this is equivalent to

$$\begin{bmatrix} 1 & Diag(Y)^T \\ Diag(Y) & Y \end{bmatrix} \succeq 0. \tag{10}$$

As a result, defining $M_+(P)$ as $\{Y \in M(P)$ satisfying (10)$\}$ and $N_+(P) = \{x : Y \in M_+(P), x = Diag(Y)\}$, we have that $N_0(P) \subseteq N_+(P) \subseteq N(P) \subseteq P$ and optimizing a linear objective function over $N_+(P)$ can be done via semidefinite programming.

Lovász and Schrijver study the operator $N^k(\cdot)$ (resp. $N_+^k(\cdot)$) obtained by repeating $N(\cdot)$ (resp. $N_+(\cdot)$) k times, and show that for any $P \subseteq R^n$ we have $N_+^n(P) = N^n(P) = N_0$. Lovász and Schrijver show that the equivalence between (weak) optimization and (weak) separation [9, 11] implies that one can optimize (up to arbitrary precision) in polynomial time over N_+^k for any fixed value of k. They introduce the N-index (resp. N_+-index) of a valid inequality for P_0 starting from P as the least k such that this inequality is valid for $N^k(P)$ (resp. $N_+^k(P)$).

The N_+-index of an inequality can be much smaller than its N-index. The following theorem gives an upper bound on the N_+-index. The case $k = 1$ appears in [18], while the general case is unpublished by the author. Given a set $Q \subset R^n$, let $Q[I] = \{x \in Q : x_i = 1, i \in I\}$.

THEOREM 2 *Let $a^T x \leq a_0$ be a valid inequality for P with $a \geq 0$. Let $S = \{i : a_i > 0\}$. Assume that $a^T x \leq a_0$ is valid for $P[J]$ whenever (i) $J \subseteq S$, $|J| = k$ and whenever (ii) $J \subseteq S$, $|J| \leq k-1$ and $\sum_{j \in J} a_j \geq a_0$. Then $a^T x \leq a_0$ is valid for $N_+^k(P)$.*

The condition $a \geq 0$ can be satisfied through complementation. This theorem essentially says that if one can derive validity of an inequality by fixing any set of k variables to 1, then we can derive it by k repeated applications of N_+; condition (ii) simply takes care of those sets of k variables that do not satisfy the inequality.

As an illustration, consider the stable set polytope where we can take as initial relaxation the fractional stable set polytope

$$FRAC(G) = \{x : x_i + x_j \leq 1 \text{ if } (i,j) \in E, x_i \geq 0 \text{ for all } i \in V\}.$$

Lovász and Schrijver [18] show that the N-index of a clique constraint on k vertices ($\sum_{i \in S} x_i \leq 1$) is $k-2$ while its N_+-index is just 1, as can be seen from Theorem 2. Odd hole, odd antihole, odd wheel, and orthonormal representation constraints also have N_+-index equal to 1, implying the polynomial time solvability of the maximum stable set problem in any graph for which these inequalities are sufficient (including perfect graphs, t-perfect graphs, etc.).

However, there are also situations where the N_+ operator is not very strong. Consider the matching polytope (the convex hull of incidence vectors of matchings, which can also be viewed as the stable set polytope of the line graph) and its Edmonds constraints: $\sum_{i \in S} x_i \leq (|S|-1)/2$ for $|S|$ odd. Stephen and Tunçel [23] show that their N_+-index (starting from the relaxation with only the degree constraints) is exactly $(|S|-1)/2$, and thus $\Theta(\sqrt{n})$ iterations of N_+ are needed to get the matching polytope where n is its dimension. Although n iterations are always sufficient for N or N_+, here is a situation in which not significantly fewer iterations are sufficient. Let

$$P = \left\{ x \in R^n : \sum_{i \in S} x_i \leq \frac{n}{2} \text{ for all } S : |S| = \frac{n}{2} + 1 \right\}.$$

Thus

$$P_0 = \left\{ x \in R^n : 0 \le x_i \le 1 \text{ for } i = 1, \cdots, n, \text{ and } \sum_{i=1}^{n} x_i \le \frac{n}{2} \right\}.$$

Let z^k and z_+^k denote $\max\{\frac{1}{n}\sum_{i=1}^n x_i\}$ over $x \in N^k(P)$ and $N_+^k(P)$, respectively. Goemans and Tunçel (unpublished) have obtained recurrences for z^k and z_+^k and derived several properties; their most important results are summarized below.

THEOREM 3 *1. For $k \le \frac{n}{2}$, $z^k \ge z_+^k > \frac{n/2-r}{n/2+1-r}$. In particular $z^{n/2-1} > 0.5$.*

2. For $k \le \frac{n}{2} - \sqrt{n} + \frac{3}{2}$, we have $z^k = z_+^k$.

Together with Theorem 2, (i) implies that the N_+-index of $\sum_{i=1}^n x_i \le n/2$ is exactly $n/2$, while one can show that its N-index is $n-2$. Furthermore, (ii) says that semidefinite constraints do not help for $n/2 - o(n)$ iterations.

REFERENCES

[1] F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J. Opt.*, 5:13–51, 1995.

[2] A.E. Brouwer and W.H. Haemers. Association schemes. In *Handbook of Combinatorics*, R.L. Graham, M. Grötschel and L. Lovász, eds, Elsevier, 747–771, 1995.

[3] C. Delorme and S. Poljak. Combinatorial properties and the complexity of a max-cut approximation. *Europ. J. Comb.*, 14:313–333, 1993.

[4] C. Delorme and S. Poljak. Laplacian eigenvalues and the maximum cut problem. *Math. Prog.*, 62:557–574, 1993.

[5] P. Delsarte. An algebraic approach to the association schemes of coding theory. *Philips Research Reports, Supplement*, 10, 1973.

[6] U. Feige. Randomized graph products, chromatic numbers, and the Lovász θ-function. In *Proc. of the 27th ACM Symp. on Theory Comput.*, pages 635–640, 1995.

[7] M.X. Goemans. Semidefinite programming in combinatorial optimization. *Math. Prog.*, 79:143–161, 1997.

[8] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42:1115–1145, 1995.

[9] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1:169–197, 1981.

[10] M. Grötschel, L. Lovász, and A. Schrijver. Relaxations of vertex packing. *JCT B*, pages 330–343, 1986.

[11] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin, 1988.

[12] J. Håstad. Some optimal inapproximability results. In *Proc. of the 29th ACM Symp. on Theory Comput.*, 1997.

[13] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[14] H. Karloff. How good is the Goemans-Williamson MAX CUT algorithm. In *Proc. of the 28th ACM Symp. on Theory Comput.*, pages 427–434, 1996.

[15] D. Knuth. The sandwich theorem. *Elec. J. Comb.*, 1, 1994.

[16] L. Lovász. On the Shannon capacity of a graph. *IEEE Trans. Inform. Th.*, IT-25:1–7, 1979.

[17] L. Lovász and A. Schrijver. Matrix cones, projection representations, and stable set polyhedra. In *Polyhedral Combinatorics*, volume 1 of *DIMACS series in Disc. Math. and Theor. Comp. Sci.*, pages 1–17. AMS, 1989.

[18] L. Lovász and A. Schrijver. Cones of matrices and setfunctions, and 0-1 optimization. *SIAM J. Opt.*, 1:166–190, 1991.

[19] S. Mahajan and H. Ramesh. Derandomizing semidefinite programming based approximation algorithms. In *Proc. 36th Symp. on Found. of Comp. Sci.*, pages 162–169, 1995.

[20] B. Mohar and S. Poljak. Eigenvalue methods in combinatorial optimization. In R. Brualdi, S. Friedland, and V. Klee, editors, *Combinatorial and Graph-Theoretic Problems in Linear Algebra*, volume 50 of *The IMA Volumes in Mathematics and its Applications*, pages 107–151. Springer-Verlag, 1993.

[21] Y. Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. CORE Discussion Paper 9719, Louvain-La-Neuve, Belgium, 1997.

[22] Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Methods in Convex Programming*. SIAM, Philadelphia, PA, 1994.

[23] T. Stephen and L. Tunçel. On a representation of the matching polytope via semidefinite liftings. Unpublished, 1997.

[24] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Rev.*, pages 49–95, 1996.

Michel X. Goemans
M.I.T. and University of Louvain
Mailing address:
CORE, 34 Voie du Roman Pays
B-1348 Louvain-La-Neuve
Belgium
goemans@core.ucl.ac.be

Active Set and Interior Methods for Nonlinear Optimization

Richard H. Byrd and Jorge Nocedal

Abstract. We discuss several fundamental questions concerning the problem of minimizing a nonlinear function subject to a set of inequality constraints. We begin by asking: What makes the problem intrinsically difficult to solve, and which characterizations of the solution make its solution more tractable? This leads to a discussion of two important methods of solution: active set and interior points. We make a critical assessment of the two approaches, and describe the main issues that must be resolved to make them effective in the solution of very large problems.

1991 Mathematics Subject Classification: 65K05 90C30
Keywords and Phrases: nonlinear optimization, large-scale optimization, nonlinear programming

The most important open problem in nonlinear optimization is the solution of large constrained problems of the form

$$\begin{array}{rl} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \\ & g(x) \leq 0, \end{array} \tag{1}$$

where the functions $f : R^n \to R,\ h : R^n \to R^m$ and $g : R^n \to R^t$ are assumed to be smooth.

Assuming that certain regularity assumptions hold, the solution of (1) is characterized by the Karush-Kuhn-Tucker conditions [4]. They state that any solution x^* must satisfy the system

$$\nabla f(x^*) + A_h(x^*)\lambda_h^* + A_g(x^*)\lambda_g^* = 0 \tag{2}$$
$$h(x^*) = 0 \tag{3}$$
$$g(x^*) \leq 0 \tag{4}$$
$$g(x^*)^T\lambda_g^* = 0 \tag{5}$$
$$\lambda_g^* \geq 0, \tag{6}$$

for some Lagrange multiplier vectors λ_h^* and λ_g^*. Here A_h and A_g denote the matrices whose columns are the gradients of the functions h and g. The first equation can be written as $\nabla_x L(x^*, \lambda^*) = 0$, where L is the Lagrangian function

$$L(x, \lambda) = f(x) + \lambda_h^T h(x) + \lambda_g^T g(x). \tag{7}$$

This mathematical characterization is, however, not suitable for computation because finding a pair (x^*, λ^*) that satisfies the Karush-Kuhn-Tucker system (2)-(6) is a very hard problem.

Indeed we could attempt to guess the *optimal active set*, i.e. the set of inequality constraints that will be satisfied as equalities at the solution x^*. Based on this guess, we could then replace (4) by a set of equalities, remove (5) and (6), and define all Lagrange multipliers corresponding to inactive inequality constraints to be zero. This transforms (2)-(6) into a system of nonlinear equations, which is much more tractable. Unfortunately, the set of all possible active sets grows exponentially with the number t of inequality constraints. Moreover, not all pairs (x, λ) satisfying the Karush-Kuhn-Tucker conditions are solutions of (1); some of them could be, for example, maximizers. Therefore this type of approach can only be practical if we make intelligent guesses of the active set. We will return to this question below.

The fact that it is impractical to solve the Karush-Kuhn-Tucker system directly has given rise to a variety of constrained optimization methods which make use of two fundamental ideas:

transformation and approximation.

In the rest of the paper we describe how these ideas are used in some of the most powerful methods for nonlinear optimization.

1 Exact Penalty Functions

A very appealing idea is to replace (1) by a single unconstrained optimization problem. At first glance this may seem to be impossible since the general nonlinear optimization problem (1) must be much more complex than the minimization of any unconstrained function.

Nevertheless, several "exact penalty functions" have been discovered [4], and can be used in practice to solve nonlinear programming problems. The best example is the ℓ_1 penalty function

$$\psi(x; \rho) = f(x) + \rho \sum_{i=1}^{m} |h_i(x)| + \rho \sum_{i=1}^{t} g_i^+(x), \tag{8}$$

where $a^+ = \max\{0, a\}$. Here ρ is a positive penalty parameter whose choice is problem dependent. One can show that if the value of ρ is large enough, then local solutions of the nonlinear program (1) are normally local minimizers of (8).

The beauty and simplicity of this approach is undeniable. But it has two drawbacks. First of all, the function ϕ function is not differentiable, and thus minimizing it is far more difficult than minimizing a smooth function. One could use the tools of non-differentiable optimization, but an approach that may be much more effective is to make linear-quadratic approximations of ϕ, and use them to generate a series of estimates of the solution [4]. Interestingly enough, this leads to a method that is closely related to the active set method described in the next section.

The second drawback may be potentially fatal: the approach appears to be very sensitive to the choice of the penalty parameter ρ. Small values of ρ may lead to unbounded solutions, and excessively large values will slow the iteration because the nonlinear constraints will be followed closely. It is interesting that even though this exact penalty approach [4] was proposed more than 15 years ago, it has not yet been firmly established whether the difficulty in choosing the penalty parameter is serious enough to prevent it from becoming a powerful technique for large-scale optimization.

There is another open question concerning this, and most other methods for constrained optimization. It concerns the use of a merit function to determine whether a step is acceptable. We could regard a step p to be acceptable only if it gives a reduction in ψ. Some analysis, as well as numerical experience indicates that this strategy may be overly conservative and that it may be preferable to allow controlled increases in the merit function. How to do this is still an active area of research; an interesting recent proposal is described in [5].

2 Active Set Methods

Let us now consider a different approach, which is based on the strategy of making a series of intelligent guesses of the optimal active set, mentioned in the introduction.

Suppose that x is an estimate of the solution of (1) and that we wish to compute a displacement p leading to a better estimate $x^+ = x + p$. We can do this by making a linear-quadratic approximation – but this time of the original problem (1) — and solving the following subproblem in the variable p,

$$\begin{aligned} \text{minimize} \quad & \nabla_x L(x,\lambda)^T p + \frac{1}{2} p^T \nabla^2_{xx} L(x,\lambda) p \\ \text{subject to} \quad & h(x) + A_h(x)^T p = 0 \\ & g(x) + A_g(x)^T p \leq 0. \end{aligned} \tag{9}$$

This subproblem is much more tractable than (1). In fact, if $\nabla^2_{xx} L(x,\lambda)$ is positive definite, then (9) is not much more difficult to solve than a linear program. For this reason it is common to either modify $\nabla^2_{xx} L(x,\lambda)$, so that it is always positive definite in the null space of constraints, or to replace it — directly or indirectly — by a positive definite approximation. (A recently developed algorithm [5]) deviates from this standard practice by formulating indefinite quadratic programming subproblems, but it is too early to determine if it will supersede the current approaches.)

The step p is considered to be acceptable only if it leads to a reduction in a *merit function.* An example of such a merit function is (8), but many other choices that combine constraint satisfaction and objective function decrease are possible [4]. This method is called *Sequential Quadratic Programming* and is currently regarded as the most powerful active set method.

There is a good mathematical justification [9, 8] for generating steps by means of the quadratic subproblem (9). One can show that the step is a direction of descent for a variety of merit functions. Moreover, the model (9) has the precise

balance between constraint satisfaction and decrease in the objective function. Unlike approaches, such as reduced gradient methods, that attempt to satisfy the original constraints of the problem at each step (which can be computationally very demanding) the quadratic programming model (9) applies successive linearizations to the constraints – which is the idea behind Newton's method for solving equations. Thus we can expect that the iterates generated by this active set approach will decrease a measure of feasibility at a quadratic rate.

There are really two different ideas in the method we have just described. The first is to use the subproblem (9) to provide us with an informed guess of the optimal active set: our guess is the active set identified in the solution of the quadratic subproblem. The second idea is to use the right level of approximation to the objective function and constraints, as discussed above. In the interior methods described next, we no longer attempt to guess the optimal active set, but retain the idea of making linear-quadratic approximations.

3 Interior Point Methods

Let us use slack variables s to transform (1) into the following equivalent problem in the variables x and s,

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subjet to} && h(x) = 0 \\ &&& g(x) + s = 0 \\ &&& s \geq 0. \end{aligned}$$

Even though the only inequalities are now simple non-negativity constraints, a little reflection shows that this problems is just as complex as (1). Let us now *soften* the inequalities by introducing a barrier term in the objective function to obtain the new problem

$$\begin{aligned} &\text{minimize} && \phi(x;\mu) = f(x) - \mu \sum_{i=1}^{t} \ln s_i \\ &\text{subject to} && h(x) = 0 \\ &&& g(x) + s = 0, \end{aligned} \tag{10}$$

where μ is a positive parameter. Note that we have removed the bound $s \geq 0$ because we will assume that the initial value of s is positive, and the barrier term prevents us from generating negative values of s – or for that matter, values that are close to zero.

Of course, (10) is not equivalent to (1) and we have introduced a parameterization of the problem that is controlled by the barrier parameter μ. Note that (10) contains only equality constraints, and is much simpler to solve than an inequality constrained problem. Once the barrier problem (10) is approximately solved, we decrease μ, and repeat the process. This will lead to a sequence of iterates x_μ that will normally converge to a solution of (1) as $\mu \to 0$.

The set of estimates x_μ obtained by this approach is interior to the region $s > 0$, but is not necessarily feasible with respect to the inequalities $g(x) \leq 0$. Thus the term "interior point method" must be interpreted in a broad sense. The ability to generate infeasible iterates turns out to be highly advantageous in practice because finding a feasible point for a nonlinear system is computationally expensive, and it is more efficient to perform the minimization while searching for a feasible point.

Barrier methods for nonlinear programming have been known for a long time [3]. But they fell out of favor in the 1970s, and have been resurrected only recently, in a variation that we now call interior point methods. There are three recent developments that have made barrier methods more effective in solving large problems. We will discuss each of these separately.

3.1 Primal-Dual Steps

Let us consider the problem of finding an approximate solution of the barrier problem (10) for a fixed value of the parameter μ. The Karush-Kuhn-Tucker conditions take the form

$$\begin{aligned} \nabla f(x) + A_h(x)\lambda_h + A_g(x)\lambda_g &= 0 \\ -\mu S^{-1}e + \lambda_g &= 0 \\ h(x) &= 0 \\ g(x) + s &= 0, \end{aligned} \tag{11}$$

where $e = (1, ..., 1)^T$ and $S = \mathrm{diag}(s_1, ..., s_t)$. This is a nonlinear system of equations in x, λ_h and λ_g. We can ignore (for the moment) the fact that s and λ_g must be positive, and simply apply Newton's method to (11) to compute a displacement p in x and new values of the multipliers. We obtain the iteration

$$\begin{bmatrix} \nabla^2_{xx} L & 0 & A_h(x) & A_g(x) \\ 0 & \Sigma & 0 & I \\ A_h^T(x) & 0 & 0 & 0 \\ A_g^T(x) & I & 0 & 0 \end{bmatrix} \begin{bmatrix} p_x \\ p_s \\ \lambda_h^+ \\ \lambda_g^+ \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ \mu S^{-1}e \\ -h(x) \\ -g(x) - s \end{bmatrix}, \tag{12}$$

where $\Sigma = \mu S^{-2}$. This approach is very similar to the barrier techniques used in the 1980s (cf. [10]) and is called a *primal* barrier method.

An important observation is that (11) is not well suited for Newton's method because the second equation is rational. But if we multiply this equation by S we obtain the equivalent system

$$\begin{aligned} \nabla f(x) + A_h(x)\lambda_h + A_g(x)\lambda_g &= 0 \\ S\lambda_g - \mu e &= 0 \\ h(x) &= 0 \\ g(x) + s &= 0. \end{aligned} \tag{13}$$

This nonlinear transformation is very beneficial because the rational equation has now been transformed into a quadratic – and Newton's method is an excellent technique for solving quadratic equations.

Applying Newton's method to (13) gives the iteration (12) but now Σ is defined as

$$\Sigma = \Lambda S^{-1}, \tag{14}$$

where Λ is a diagonal matrix containing the entries of λ_g. This *primal-dual* iteration is at the heart of most interior point methods. After the step is computed, one can backtrack along it to make sure that s and the λ_g remain positive.

Note that, in contrast to standard practice, we have not used any duality arguments in deriving the primal-dual step computation. Indeed the term "primal-dual" is not very descriptive of the key idea, which consists of applying a nonlinear transformation that changes the optimality conditions (11) into the equivalent system (13). Even though these two systems have the same solutions, Newton's method will produce different iterates, and the primal-dual step is known to be superior [13].

An interesting question is whether the nonlinear transformation we used is the best possible.

3.2 Coping With Ill-Conditioning

The barrier function $\phi(x;\mu)$ defined in (10) is inherently ill conditioned. A simple computation shows that the Hessian of ϕ has condition number of order $O(1/\mu)$. This is reflected in the primal-dual iteration (12) where the matrix $\Sigma = \Lambda S^{-1}$ becomes unbounded as $\mu \to 0$. Nevertheless, solving (12) by a direct method, as is done in most linear programming codes, does not lead to significant roundoff errors, even when μ is very small [11, 12].

The key observation in this roundoff error analysis can be better explained if we consider Newton-like methods for solving the unconstrained problem $\min f(x)$. Here the step p is computed by solving a system of the form

$$Ap = -\nabla f(x),$$

where A is either the Hessian matrix $\nabla^2 f(x)$ or some other related matrix. It is easy to see that the quality of the search direction is very sensitive to the accuracy with which $\nabla f(x)$ is calculated, but is not particularly sensitive to changes in A. The ill-conditioning of the barrier function can cause errors in the factorization of the iteration matrix, but very significant errors can be tolerated before the quality of the iteration is degraded — and simple safeguards ensure that high accuracy is obtained in most cases [12].

All of this assumes that a direct method is used to solve (12). But in many practical applications, the problem is so large that direct methods are impractical due to the great amount of fill that occurs in the factorization. In other applications, the Hessians of f, g or h are not be available, and only products of these Hessians times vectors can be computed. In these cases it is attractive to use the linear conjugate gradient (CG) method to solve the Newton equations (12). This system is indefinite, but by eliminating variables, one obtains a positive definite reduced system to which the projected conjugate gradient method can be applied [2, 1].

When using the conjugate gradient method to solve the Newton equations, ill-conditioning is a grave concern. The unfavorable distribution of eigenvalues of the matrix in (12) may require a large number of CG iterations, and may even prevent us from achieving sufficient accuracy in the step computation. Fortunately, since the barrier function is separable and the portion that gives rise to the ill-conditioning is known explicitly, we can apply preconditioning techniques. To describe them let us recall that the step given by (12) has been decomposed in terms of its x and s-components, $p = (p_x, p_s)$. Then the change of variables

$$\tilde{p}_s = \mu S^{-2} p_s,$$

transforms the primal-dual matrix $\Sigma = \Lambda S^{-1}$ into $\Sigma = \mu^{-1}\Lambda S$. The second equation in (11) implies that ΛS converges to μI, showing that the new matrix Σ will not only be bounded, but will converge to the identity matrix. The CG iteration can now be effectively applied to the transformed system [1]. One should note, however, that this preconditioning comes at a price, and increases the cost of the CG iteration [1].

In summary, we have learned how to cope with ill-conditioning in barrier methods for nonlinear optimization. These observations also indicate that developing quasi-Newton variants of the interior methods just described may not pose significant difficulties provided that we approximate only the Hessian of the Lagrangian (7) of the original problem (1), as opposed to the Hessian of the Lagrangian of the barrier problem (10) which contains structural ill-conditioning.

3.3 Predictor-Corrector Strategy

The third key contribution of interior point methods has been the idea of using probing schemes to determine how fast to reduce the barrier parameter, and at the same time to determine (indirectly) how accurately to solve the barrier problem [7]. We cannot describe these *predictor-corrector* techniques here, and refer the reader to [13] for an excellent treatment of this subject.

We will only outline the key ideas of this approach which, at present, has only been implemented in the context convex optimization. Its most interesting feature is that it goes beyond the principle of Newton's method which computes a step based on an approximation of the problem at the current point. Instead, one first probes the problem by attempting to solve (12) with $\mu = 0$, which amounts to trying to solve the original nonlinear program (1). By gathering information in this probing iteration (the predictor), we can make a decision on how much to decrease the barrier parameter. At the same time, and at minimal cost, we can compute a primal-dual type step that corrects the predictor step and generates an iterate that is closer to the solution of the current barrier problem.

4 Final Remarks

Let us contrast the active set and interior approaches described in the previous sections by comparing the way in which they generate steps.

In the active set method we compute an *exact* solution of the subproblem (9). This is a full-fledged inequality constrained problem which can be costly to solve when the number of variables and constraints is large – particularly if the Hessian of the model is not positive definite. This is the main disadvantage of active set methods.

The great virtue of the active set approach is that it gives us, at every iteration, a guess of the optimal active set. As the iterates approach the solution, the active set of the subproblem (9) does not change, or undertakes minimal changes. This allows great savings in the solution of the subproblem because a warm start can be used: the solution of a new subproblem (9) can start from the active set identified at the previous iteration, and one can also re-use certain matrix factorizations [6].

Let us now consider interior point methods. The primal-dual iteration (12) is only a local method, and must be modified to be capable of dealing with non-convex problems. The interior methods described in [1] and [14] compute the step by solving a quadratic subproblem obtained by making a linear-quadratic approximation of the barrier problem. This approximation is such that, asymptotically, the iteration reduces to the primal-dual iteration (12). In both of these approaches there is an explicit bound on the step p_s in the slack variables. It takes the form

$$p_s \geq 0.995s,$$

and is known as a "fraction to the boundary rule".

This subproblem appears to be very similar to (9) since it also contains inequality constraints, but the presence of the barrier terms in the objective softens these constraints. Whereas in the active set approach the solution of the subproblem will normally lie on the boundary of the feasible region, in the interior approach this will not be the case, and solving the subproblem is simpler. This is one of the great advantages of interior methods.

A drawback of interior methods is that they normally do not provide a clear indication of the optimal active set until the solution is computed to high accuracy. This is undesirable in some applications, and future interior point codes may need to switch to an active set iteration, if necessary. Another weakness of interior methods is that they cannot efficiently re-use information from a previous subproblem. Roughly speaking, the solution of every subproblem requires the same amount of work. Finally, it is not yet known if interior point methods will prove to be as robust as active set methods for solving difficult non-convex problems.

These observations are based on the limited numerical experience that has been accumulated for both approaches when solving large problems. Once we have gained a better understanding of their practical behavior, and after new variants have been proposed, we will undoubtly discover that other unforeseen issues will tilt the balance towards one approach or the other.

REFERENCES

[1] R.H. Byrd, M.E. Hribar, and J. Nocedal. *An Interior Point Algorithm for Large Scale Nonlinear Programming*, Technical Report OTC 97/05, Optimization Technology Center, Northwestern University (1997).

[2] T.F. Coleman and A. Verma. *A Preconditioned Conjugate Gradient Approach to Linear Equality Constrained Minimization*, Technical Report, Computer Science Dept., Cornell University, Ithaca, New York.

[3] A.V. Fiacco and G.P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley & Sons, 1968.

[4] R. Fletcher. *Practical Methods of Optimization, Second Edition*, John Wiley & Sons, New York, 1990.

[5] R. Fletcher and S. Leyffer. *Nonlinear Programming Without a Penalty Function*, Technical Report NA/171, Department of Mathematics, University of Dundee, Dundee, Scotland.

[6] P.E. Gill, W. Murray, and M.A. Saunders. *An SQP algorithm for large-scale optimization*, Technical Report, Mathematics Department, University of California at San Diego, San Diego, California.

[7] S. Mehrotra. *On the Implementation of a Primal-Dual Interior Point Method*, SIAM Journal on Optimization, 2, pp. 575-601, 1992.

[8] M.J.D. Powell. *The Convergence of Variable Metric Methods for Nonlinearly Constrained Optimization Calculations*, in "Nonlinear Programming 3", (O. Mangasarian, R. Meyer and S. Robinson, eds), pp. 27-64, Academic Press, New York.

[9] S.M. Robinson. *Perturbed Kuhn-Tucker Points and Rates of Convergence for a Class of Nonlinear Programming Algorithms*, Math. Programming, 7, pp. 1-16, 1974.

[10] M.H. Wright. *Interior Point Methods for Constrained Optimization*, Acta Numerica 1992 (A. Iserles ed.), pp. 341-407, Cambridge University Press.

[11] M.H. Wright. *Ill-Conditioning and Computational Error in Interior Methods for Nonlinear Programming*, Technical Report 97-4-04, Computing Sciences Research Center, Bell Laboratories, Murray Hill, New Jersey.

[12] S.J. Wright. *Finite-Precision Effects on the Local Convergence of Interior-Point Algorithms for Nonlinear Programming*, Preprint ANL/MCS P705-0198, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois.

[13] S.J. Wright. *Primal-Dual Interior Point Methods*, SIAM, 1997.

[14] H. Yamashita. *A Globally Convergent Primal-Dual Interior Point Method for Constrained Optimization*, Technical Report, Mathematical Systems Institute Inc, Tokyo, Japan.

Richard H. Byrd
Computer Science Dept.
University of Colorado,
Boulder CO 80309
USA
richard@cs.colorado.edu

Jorge Nocedal
ECE Department
Northwestern University
Evanston IL 60208
USA
nocedal@ece.nwu.edu

Column Generation and the Airline Crew Pairing Problem

Ranga Anbil, John J. Forrest and William R. Pulleyblank

Abstract. The cost of flight crews is the second largest operating cost of an airline. Minimizing it is a fundamental problem in airline planning and operations, and one which has leant itself to mathematical optimization. We discuss several recent advances in the methods used to solve these problems. After describing the general approach taken, we discuss a new method which can be used to obtain approximate solutions to linear programs, dramatically improving the solution time of these problems. This is the so-called volume algorithm. We also describe several other ideas used to make it routinely possible to get very good solutions to these large mixed integer programs.

1991 Mathematics Subject Classification: 90B35, 90C09, 90C10
Keywords and Phrases: linear programming, integer programming, volume algorithm, crew pairing, column generation

1 Airline Operations and the Crew Pairing Problem

The Airline Crew Pairing problem has become well known as a prototypical applied problem which lends itself to a column generation approach. A *pairing*, or *tour of duty*, is a sequence of flights to be flown by a crew, starting and ending at the crew member's home base. Typically the length of a pairing will range from a single day to four or five days. The pairing problem is to compute a set of pairings covering all the scheduled flights of an airline, which has minimum, or near minimum, cost. This problem normally forms the third major stage in a four part planning process:

1. Schedule Creation. Flights are planned, based on market demands and competitive analysis. For example, we may schedule a flight from New York's La Guardia Airport to Chicago's O'Hare Airport every non-holiday weekday departing at 6:15 P.M.

2. Fleet Scheduling. Airlines normally have several different types (fleets) of aircraft. A fleet is chosen for each flight scheduled in the first stage, subject to the constraints that the type of aircraft chosen for each flight be of suitable capacity and flying characteristics and subject to "Kirchoff's law"

for aircraft – each aircraft that departs from an airport must have landed there first. Subject to these constraints, the operating cost of the aircraft should be minimized.

3. CREW PAIRING. A set of pairings is created which assigns crews to aircraft flights. This will normally be solved separately for each fleet, as typically a crew member only has current authorization for one type of fleet.

4. ROSTERING. Individual crew members are assigned to each pairing, often based on seniority, sometimes by an auction process.

The difficulty of the crew pairing problem comes from several factors:

1. The rules which govern the feasibility of a pairing are very complex, often combining FAA regulations with sets of rules coming from union negotiations.

   For example, the rules governing a feasible assignment of duties of a major US carrier, include the following rules: the maximum trip length is 4 duty days; the maximum duty flying time is 8 hours; the minimum time allowed between connecting flights is 40 minutes; at most one aircraft changes is permitted in a duty.

   There are also more complicated rules. One example requires that in any 24 hour window of a pairing, a crew member flying up to 8 hours is entitled to an intervening rest period of 540 minutes before the next duty. However, if necessary, this can be shortened to 510 minutes, but if this is done, then the following intervening rest must be at least 630 minutes.

2. The calculation of the cost of a pairing is complex, and usually incorporates several nonlinear components. Most of the cost is for the crew pay. This is based on the amount of actual flight time, but also has minima that are applied based on the time of duty each day and the number of days in the pairing. There are also costs associated with hotel costs and per diem expenses.

3. The number of feasible pairings is very large, so it is not practical to generate the complete set of feasible pairings for a problem.

   The table below illustrates the number of feasible pairings, with maximum length three or four days, for several fleets.

Fleet	Max Days	#Flights	#Bases	#Duties	#Pairings (millions)
AAS80	3	1152	12	690,000	48,400
AA757	3	251	15	7,000	1
AA727	3	375	11	31,000	36
AAF10	4	307	3	55,000	63,200
UA737	4	773	7	568,000	100,000,000
USDC9	4	478	4	562,000	105,000,000

Note that although the number of feasible pairings is very large, the number of feasible one day duties is of a much more manageable size.

2 Overall Approach and Formulation

2.1 Basic formulation

The crew pairing optimization problem is normally solved in several stages. First, a daily problem is solved, which handles all the flights which are scheduled every day. Then a weekly problem is solved which reuses as much of the daily solution as possible, and also handles weekends. Then a monthly or dated problem is solved which keeps as much as possible of the weekly solutions and creates a complete monthly schedule, dealing with holidays and week to week transitions. Each stage of this problem can be formulated as an integer program.

Let P be the set of all feasible pairings. For each $j \in P$, we define a $0,1$ variable x_j with the interpretation $x_j = 1$ if the pairing j is used and $x_j = 0$ if it is not used. Let c_j denote the cost of pairing j, which can be computed based on the cost structure as described above.

Let F be the set of all flights that must be covered in the period of time under consideration. For each $i \in F$, let P^i denote the set of all pairings which cover the flight i. Then the problem becomes:

$$\begin{array}{lll} \text{Minimize} & \sum_{j\in P} c_j x_j & (1)\\ \text{subject to} & x_j \in \{0,1\} \quad \text{for all } j \in P, & (2)\\ & \sum_{j\in P^i} x_j = 1 \quad \text{for all } i \in F. & (3) \end{array}$$

Let $A = (a_{ij} : i \in F,\, j \in P)$ denote the $0,1$ flight-pairing incidence matrix, where $a_{ij} = 1$ if pairing j includes flight i and 0 otherwise. Then this can be written as the set partitioning problem:

$$\text{Minimize } cx \text{ subject to } 0 \le x \le 1, \text{ integer, and } Ax = 1. \tag{4}$$

A is a matrix which typically has several hundred rows (one for each flight to be covered) and many millions of columns, corresponding to the set of feasible pairings. The size of A makes exhaustive column enumeration impractical.

Early works, discussed in Arabeyre et al.[3], Bornemann[11], and Marsten and Shepardson[23], restrict column generation *a priori* to a small, manageable subset of preferred pairings and solve the resulting set partitioning problem by implicit enumeration or branch-and-bound via linear programming or Lagrangian relaxation. Other works, reported in Rubin[26], Baker et al.[4], Ball and Roberts[5], Etschmaier and Mathaisel[15], Gershkoff[18], Anbil et al.[1] and Graves et al.[19] employ iterative column generation schemes centered around local improvement procedures but do not guarantee global convergence.

Anbil, Johnson and Tanga[2], present a method to handle more than 5 million columns and report excellent cost savings over a local improvement scheme, but the search is nonetheless limited to this subset. Recently, Housos and Elmroth

[21] solve the problem a day at a time over a week horizon and report excellent results, however, it is not clear if their iterative scheme has global convergence.

Minoux[24] is perhaps the first to present a globally converging column generation method for the linear relaxation of the crew pairing problem. This is described in the next solution. Crainic and Rousseau[13] also report a similar but less formal procedure.

Desaulniers et al.[14] and Barnhart et al.[9] discuss convergence rate issues and column generation integrated into branch-and-bound for solving the overall mixed integer program. Barnhart, Hatay and Johnson[8] discuss a scheme for interactively expanding the duty network to include deadhead flights as they are needed. Ryan and Falkner[28], Hoffman and Padberg[20], Bixby et al.[10], Chu et al.[12] and Wedelin[30] discuss specialized optimization algorithms for the underlying mixed integer programs. Vance, Barnhart, Johnson and Nemhauser[29] present a different formulation for the crew pairing problem but with similar algorithmic issues.

Some additional constraints may also be added to this formulation. For example, *Crew base constraints* associate each pairing with a crew base from which it can be flown. Then the number of pairings associated with each crew base is given both upper and lower bounds. *Quality constraints* may apply to an entire problem, and restrict the number of three and four day pairings which can be used in an optimal solution. Often these are soft constraints, in that they can be violated, but an artificial penalty cost is added to the solution if this happens.

2.2 Solution Approach

Our solution approach combines rapid solutions of linear programming relaxations with column generation and specialized branching:

1. Generate an Initial Solution. If a warm start solution is available, use it. (A warm start solution is a solution obtained from an earlier, similar run.) If not, use heuristics to create a feasible solution. Add additional columns to the problem, as described in the following section.

2. Solve the linear program using the Sprint method, described below.

3. Generate additional columns to improve the quality of the linear program solution. We focus on flights which are being covered by expensive pairings, but also randomize. Also, we use the column generation method described in the next section.

4. Repeat Steps 2 and 3 until the linear program solution only improves negligibly.

5. Collapse flights based on best follow-on using the linear program solution. This is described in Section 4.

6. Reduce the matrix, generate additional columns to improve the linear programming solution using the reduced matrix, and return to Step 5.

7. When the number of flights in Step 5 is less than 100, solve for the best integer programming solution.

The selection of flights to be used for deadheading creates an additional problem. Airlines generally prefer to use their own flights, but will permit the use of

competitor flights, if it substantially improves the objective function. This then has the potential to substantially increase the problem size, even to an extent that it cannot be solved. For that reason, careful selection of a suitable set of deadheads is crucial.

2.3 THE SPRINT METHOD

The linear programs solved in the above procedure have a small number of rows but a very large number of columns. As the problems get large, the solution times using standard linear programming methods grow too rapidly to be practical. John Forrest[16] introduced the so-called *Sprint* method. The example below, based on a problem from American Airlines, had 837 flights and generated a total of 5.5 million feasible pairings. The problem was solved with IBM's OSL Primal, Dual and Sprint, for successively larger subsets of the 5.5 million columns. In each case, the pairings represented by the first 110 columns gave a good feasible solution. All times are minutes of solution time on a S/390 (in 1989).

Number of Pairings	Primal Iterations	Time	Dual Iterations	Time	Sprint Iterations	Time
5,000	479	.05	203	.05	933	.10
10,000	1416	.35	295	.10	1072	.14
50,000	7239	11.1	1861	3.99	2870	.86
100,000	17273	57.7	4804	21.5	6981	2.68
200,000	33950	226.8	8008	74.3	36356	12.7
250,000	44215	397.8	8727	100.7	57181	21.5
500,000			11105	258.6	106002	54.5

The Sprint method proceeds as follows: It begins by solving the linear program obtained by keeping a small subset of the columns and then uses the optimal dual variables to price out all columns of the original problem. If the solution is not optimal to the original problem, then there are columns having negative reduced cost. A new problem is formed by keeping the columns in the optimal basis and then adding a subset of the best unused columns, based on the current reduced cost. This is done by first bucketing the columns based on the reduced cost, adding the columns in the best buckets, plus a random selection of columns from other buckets. Note that even though the number of Sprint iterations grows rapidly, they are performed very quickly, enabling our low solution times.

3 OPTIMAL COLUMN GENERATION

In the solution approach described above, we go through a number of waves of column generation. Minoux[24] introduced an approach, based on the optimal dual variables to the linear program described above, which would generate new columns that could reduce the cost of the current solution, if any existed.

The linear programming relaxation of the integer program above is

$$\text{Minimize } cx \text{ subject to } 0 \leq x, \text{ and } Ax = 1. \tag{5}$$

We have dropped the upper bounds on the variables, as they are implied by the non negativity constraints, combined with the equations $Ax = 1$. The dual problem will have a variable y_i corresponding to each flight $i \in F$.

Minoux observed that the total number of feasible one-day duties is manageable. (See the table above.) He suggested building a tree of duties to represent pairings as paths in this tree. A path includes both duty arcs and ground arcs that link consecutive duties, and the cost of a path or pairing is viewed as a linear sum of duty and ground arc costs. Initially, some paths are selected and a linear programming solver gives the dual variables corresponding to each flight. The duty arc costs are then reduced by their flight duals and a shortest path algorithm applied to the duty tree gives the most negative reduced cost pairing to enter the problem. It is then resolved and the updated duals drive the next shortest path generation. This process will eventually reach a point when the shortest path yields a non-negative reduced cost. The columns present in the current linear program contain the optimum solution.

Current column generation methods reported in Barnhart et al.[7] Desaulniers et al.[14] Lavoie et al.[22], Wise[31] and Barnhart et al.[9] are mostly further extensions to Minoux's seminal work that address one or more of three main limitations: the path cost is not necessarily linear in the arc costs; all paths may not be legal; there are extra constraints such as crew base constraints that may limit the number of pairings that may be flown from a specific base.

The most common approach is dynamic programming on a shortest-path formulation with side constraints to ensure accurate costing, legal paths and the proper accounting of the extra constraints.

We use a variant of the shortest path column generation scheme. Like Minoux[24], we use a duty tree with duty arcs and ground arcs, but carry a vector of arc costs to enable the different ways of computing the cost. Using the set of dual variables from a previously solved linear program, we reduce the arc costs component-wise. We then perform a depth first traversal of the duty tree, ensuring we are feasible as we proceed, and keeping a component-wise tally of the reduced arc costs. We now add this tally to the cost of the shortest path from our location to the end of the tree, again component-wise, and use the maximum of the component sums as a lower bound on the true shortest path cost. If this lower bound exceeds a threshold, then we backtrack. We ensure that every pairing we generate will be feasible as we proceed, by ensuring that all constraints are satisfied as we go along. This is in contrast with some earlier methods which generated a superset of the feasible columns and then were filtered afterwards. Our method also avoids the excessive memory requirements inherent in dynamic programming.

Each time we generate columns, we set a threshold and generate all columns whose current costs will be less than this value. This value is changed during the running of the method.

4 Branching Methods

Normally, a branch and bound approach is used to repeatedly add constraints to the linear programming relaxation until an integral optimum is found. A key to the success of this approach is the method used to generate these constraints. We use a method called *branch on follow-ons*, introduced by Ryan and Falkner [27]. A *follow-on* is the second of a pair of consecutive flights flown in a pairing. We examine the solution obtained to a linear relaxation, and look for follow-ons which occur in a set of pairings j for which $\sum x_j$ is large, that is, close to 1. (Note that the values x_j will, in general, be fractional.) The sum is over all pairings j in the optimum linear programming solution for which x_j is positive. When we find such a follow on, we fix it, that is, we force it to be part of the solution by creating an artificial flight which combines the two consecutive flights, and reducing the problem.

Note that the column generation phase will still consider the original set of flights, so it is possible to recover from a bad choice.

5 The Volume Algorithm

Barahona and Anbil[6] developed an extension to the subgradient algorithm which will produce approximate optimal feasible primal and dual solutions to a linear program, much more quickly than solving it exactly. These methods yield significant improvements to the running time of our crew pairing approach.

Since the early 1970's, subgradient algorithms have been used to rapidly produce good lower bounds for linear programs, see for example Nemhauser and Wolsey[25]. The method produces a sequence of feasible solutions to the dual problem, which will converge to the optimum. Barahona and Anbil[6] present a new method which also produces a feasible solution to the primal problem. It is based on the following result:

Theorem 1 *Consider the linear program Maximize z, subject to $z + a_i\pi \leq b_i$, for $i = 1, ..., m$, where π is a vector with $n-1$ components. Let $(\hat{z}, \hat{\pi})$ be an optimal solution and suppose that the constraints $1, ..., m', m' \leq m$ are active. Let $\bar{z} < \hat{z}$ and assume that $z + a_i\pi \leq b_i$, for $i = 1, ..., m'; z \geq \bar{z}$ defines a bounded polyhedron.*

For $1 \leq i \leq m'$, let γ_i be the volume between the face defined by $z + a_i\pi \leq b_i$ and the hyperplane defined by $z = \bar{z}$. Then an optimal dual solution is given by

$$\lambda_i = \frac{\gamma_i}{\sum_{j=1}^{m'} \gamma_j}.$$

They also show how this theorem can be used to produce a feasible dual solution which approximates the optimum.

We apply this method to the linear programs we encounter when solving the crew pairing problem. In addition to the performance improvements, the dual solutions produced tend to be very good for the column generation phase. This is because they are highly nonbasic, that is, a large number of variables are nonzero. Dual solutions obtained using the barrier or interior method for linear

programming share this property. However, those used by the simplex algorithm tend to be very sparse, and do not work as well.

We also use the volume algorithm as a crash procedure for the simplex method when the Sprint method performs poorly. These situations arise after we generate columns for a period of time using only duals from the Volume algorithm. The table below summarizes runs on several problems obtained from US Airways and Southwest Airlines. All times are CPU seconds on an RS6000/590.

Using the Volume algorithm dual solution to crash the dual simplex is, on average, nine times faster than the dual simplex algorithm and two times faster than the interior method. We did not consider the primal simplex algorithm, since it performs very poorly for these problems.

No. of Rows	No. of Columns	No. of Elements	Primal-Dual Interior	Dual Simplex	Volume plus Dual Simplex
2504	53226	553148	1341	3747	320
2991	46450	502338	1984	6928	923
4810	95933	1009283	4165	25917	2576

6 Conclusions

The methods we have described here are used successfully by several airlines to solve their monthly crew planning problems. In fact, it is part of a larger system which combines management of a database of previously created good solutions, used for warm starts, with interactive tools permitting the scheduler to manually change the solutions produced.

Presently attention is switching to the *schedule repair problem*. This is the operational problem which occurs when a schedule is disrupted, for example, due to weather or mechanical problems, and it is desired to get back on schedule as efficiently as possible. This problem is complicated by the fact that we must produce feasible solutions in minutes, rather than hours, and by the problems of aircraft availability as well as the possibility of canceling flights.

References

[1] R. Anbil, E. Gelman, B. Patty and R. Tanga, Recent Advances in Crew Pairing Optimization at American Airlines, *Interfaces* 21 (1991), 62-74.

[2] R. Anbil, E.L. Johnson and R. Tanga, A Global Approach to Crew Pairing Optimization. *IBM Systems Journal* 31 (1991), 71-78.

[3] J.P. Arabeyre, J. Fearnley and W. Teather, The Airline Crew Scheduling Problem: A Survey, *Transportation Sci.* 3 (1969), 140-163.

[4] E.K. Baker, L.D. Bodin and M. Fisher, The Development of a Heuristic Set Covering Based System for Aircrew Scheduling, *Transportation Policy Decision Making* 3 (1985), 95-110.

[5] M. Ball and A. Roberts, A Graph Partitioning Approach to Airline Crew Scheduling. *Transportation Sci.* 19 (1985), 95-110.

[6] F. Barahona and R. Anbil, The Volume Algorithm: producing primal solutions with a subgradient method, Research Report RC 21103 (94395), IBM T.J. Watson Research Center, Yorktown Heights, NY, October 1997.

[7] C. Barnhart, E.L. Johnson, R. Anbil and L. Hatay, A Column Generation Technique for the Long-Haul Crew Assignment Problem, Computational Optimization Center Working Paper COC-91-01, Georgia Tech. (1991).

[8] C. Barnhart, L.Hatay and E.L.Johnson, Deadhead Selection for the Long-Haul Crew Pairing Problem, *Operations Research* 43 (1995), 491-499.

[9] C. Barnhart, E.L. Johnson, G.L. Nemhauser, M.W.P.Savelsbergh and P.H.Vance, Branch-and-Price: Column Generation for Solving Huge Integer Programs, *Operations Research* 46, No.3 (1998), to appear.

[10] R. E. Bixby, J.W. Gregory, I.J. Lustig, R.E. Marsten and D.F. Shanno, Very Large-Scale Linear Programming: A Case Study in Combining Interior Point and Simplex Methods, *Operations Research* 40 (1992), 885-897.

[11] D.R. Bornemann, The Evolution of Airline Crew Pairing Optimization, *22nd AGIFORS Symposium Proceedings*, (1982).

[12] H.D. Chu, E. Gelman and E.L. Johnson. Solving Large Scale Crew Scheduling Problems, *European Journal of Operations Research* 97 (1997) 260-268.

[13] T. Crainic and J. Rousseau, The Column Generation Principle and the Airline Crew Scheduling Problem, *INFOR* 25 (1987) 136-151.

[14] G. Desaulniers, J. Desrosiers, Y. Dumas, S. Marc, B. Rioux, M.M. Solomon and F. Soumis, Crew Pairing at Air France. *European Journal of Operations Research* 97 (1997), 245-259.

[15] M.M. Etschmaier and D.F.X. Mathaisel, Airline Scheduling: An Overview, *Transportation Sci.* 19 (1985), 127-138.

[16] J.J. Forrest, Mathematical programming with a library of optimization subroutines, presented at the ORSA/TIMS Joint National Meeting, New York, October 1989.

[17] R. Gerbracht, A New Algorithm for Very Large Crew Pairing Problems. *18th AGIFORS Symposium Proceedings* (1978).

[18] I. Gerschkoff, Optimizing Flight Crew Schedules. *Interfaces* 19 (1989), 29-43.

[19] G.W. Graves, R.D. McBride, I. Gershkoff, D.A. Anderson and D.Mahidhara, Flight Crew Scheduling, *Mgt. Sci.* 39 (1993), 736-745.

[20] L.H. Hoffman and M. Padberg, Solving Airline Crew Scheduling Problems by Banch-and-Cut, *Management Science* 39 (1993), 657-682.

[21] E. Housos and T. Elmroth, Automatic Optimization of Subproblems in Scheduling Airline Crews, *Interfaces* 27 (1997), 68-77.

[22] S. Lavoie, M. Minoux and E.Odier, A New Approach for Crew Pairing Problems by Column Generation with Application to Air Transportation. *European Journal of Operations Research* 35 (1988), 45-58.

[23] R.E. Marsten and F. Shepardson, Exact Solution of Crew Problems using the Set Partitioning Mode: Recent Successful Applications, *Networks* 11 (1981), 65-177.

[24] M. Minoux, Column Generation Techniques in Combinatorial Optimization: A New Application to Crew Pairing Problems, *24th AGIFORS Symposium* (1984), 15-29.

[25] G.L. Nemhauser and L.A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, New York, (1988).

[26] J. Rubin, A Technique for the Solution of Massive Set Covering Problems, with Applications to Airline Crew Scheduling, *Transportation Sci.* 7 (1973), 34-48.

[27] D.M. Ryan and J.C. Falkner, A bus crew scheduling system using a set partitioning mode, *Annals of Operations Research* 4 (1987), 39-56.

[28] D.M. Ryan and J.C. Falkner, On the Integer Properties of Scheduling Set Partitioning Models. *European Journal of Operations Research* 35 (1988), 442-456.

[29] P.H. Vance, C. Barnhart, E.L. Johnson and G.L. Nemhauser, Airline Crew Scheduling: A New Formulation and Decomposition Algorithm. *Operations Research*, Vol. 45. No. 2 (1997), 188-200.

[30] D. Wedelin, An Algorithm for Large Scale 0-1 Integer Programming with Application to Airline Crew Scheduling *Annals of Operations Research*, Vol. 57 (1995), 283-301.

[31] T.H. Wise, *Column Generation and Polyhedral Combinatorics for Airline Crew Scheduling*, Ph.D. Dissertation, Cornell U., Ithaca, N.Y. (1995).

Ranga Anbil
Mathematical Sciences
IBM Research
P.O Box 218
Yorktown Heights, NY 10598
anbil@watson.ibm.com

John J. Forrest
Mathematical Sciences
IBM Research
P.O Box 218
Yorktown Heights, NY 10598
forrest@watson.ibm.com

William R. Pulleyblank
Mathematical Sciences
IBM Research
P.O Box 218
Yorktown Heights, NY 10598
wrp@watson.ibm.com

Routing and Timetabling by Topological Search

Alexander Schrijver

Abstract. We discuss how decomposing the search space into homotopy classes can help in finding solutions to combinatorial optimization problems. Searching any homotopy class then amounts to finding a group function ψ on the arcs of a directed graph such that ψ is cohomologous to a given function ϕ and such that ψ has values in a prescribed range.

We describe applications to two specific classes of NP-complete problems: routing wires on a chip (where the main tool is solving the cohomology problem in a free group, and a main result the polynomial-time solvability of the wire-routing problem for any fixed number of modules), and finding a periodic timetable (applied to the Dutch railway timetable, where liftings of the period group C_{60} to the integers give the classes to be searched).

The methods also imply a characterization of the existence of an isotopy of a compact surface S that brings a given set of disjoint closed curve on S to a given undirected graph embedded on S.

1991 Mathematics Subject Classification: 05C85, 05C90, 90B06, 90B10, 90B35
Keywords and Phrases: homotopy, disjoint paths, routing, timetabling, closed curves, compact surface

1. Introduction

A basic technique in combinatorial optimization, and more generally, in integer programming, is to extend ('relax') the feasible solution set $X \subseteq \mathcal{Z}^k$ to conv.hull$(X) \subseteq \mathcal{R}^k$, and to use the solution of the relaxed problem as a guideline in an approximative method or in a branch and bound process. This is based on the hope that a fractional solution is close to the integer solution, and on the idea that the relaxed problem can be solved fast with linear programming techniques.

Mathematically, the idea can be described as embedding the group $\mathcal{Z}^k$ into the group $\mathcal{R}^k$, where $\mathcal{Z}^k$ is a 'hard' group, while $\mathcal{R}^k$ is a 'tractable' group (as long as the feasible region is convex).

In this survey we describe a different technique of reducing problems on 'hard' groups to problems on 'tractable' groups. Instead of *embedding* the hard group into a tractable group, we *lift* the hard group to a tractable group. We give two examples where this technique can be applied successfully, although it is not as

generally applicable as the embedding technique described above. We will not venture upon describing the method in its full generality, but hope that the reader will see that the frameworks we describe have a common underlying structure.

The type of problems where the technique applies can be described as follows. Let $D = (V, A)$ be a directed graph, and let G be a group. Call two functions $\phi, \psi : A \to G$ *cohomologous* (denoted by $\phi \sim \psi$) if there exists a function $p : V \to G$ such that for each arc $a = (u, v)$ one has

$$\psi(a) = p(u)^{-1}\phi(a)p(v). \tag{1}$$

Consider the following *cohomology feasibility problem:*

$$\begin{aligned} &\textit{given: } \phi : A \to G \text{ and } \Psi : A \to 2^G, \\ &\textit{find: } \psi \sim \phi \text{ such that } \psi(a) \in \Psi(a) \text{ for each } a \in A. \end{aligned} \tag{2}$$

This problem is in general hard to solve, even if $G = C_3$. Then, if $\phi(a) := 1$ and $\Psi(a) := G \setminus \{1\}$ for $a \in A$ (assuming that G is the multiplicative group with three elements), the cohomology feasibility problem has a solution if and only if the directed graph D is 3-vertex-colourable. As the latter problem is NP-complete, the cohomology feasibility problem is NP-complete.

On the other hand, there are groups where the cohomology feasibility problem is solvable in polynomial time, provided that the sets $\Psi(a)$ each are convex in a certain sense. For instance, if $G = \mathcal{R}^k$ and each $\Psi(a)$ is convex, the cohomology feasibility problem can be solved in polynomial time by linear programming methods (assuming that the $\Psi(a)$ are appropriately described).

Another tractable group is the group $\mathcal{Z}$ of integers, where each $\Psi(a)$ is a convex subset of $\mathcal{Z}$. The cohomology feasibility problem then can be solved with a variant of the Bellman-Ford method for finding shortest paths.

As an extension of this, we have shown in [6] that if G is a free group, and each $\Psi(a)$ is *hereditary* (closed under taking contiguous subwords), then again the cohomology feasibility problem is solvable in polynomial time. This holds more generally for free partially commutative groups, if the subsets $\Psi(a)$ are convex in a certain sense — see Section 3 ([7]).

We give two applications in which the cohomology feasibility problem with a hard group shows up ($\mathcal{Z}^k$, C_{60}), and show how a lifting to a tractable group (the free group, $\mathcal{Z}$) can help in solving the problem. (In fact, $\mathcal{Z}^k$ is a special case of a free partially commutative group; however, the subsets in the application are not of the prescribed type.)

2. Disjoint paths in directed planar graphs

The first application is that of routing the wires on a very large-scale integrated (VLSI) circuit (a chip). If we restrict ourselves to one layer, the following k *disjoint paths problem* emerges:

> *given:* a planar directed graph $D = (V, E)$, and distinct vertices $s_1, t_1, \ldots, s_k, t_k$ of D;
> *find:* pairwise disjoint directed paths $P_1, \ldots, P_k$, where P_i runs from s_i to t_i $(i = 1, \ldots, k)$.

For general directed graphs, this problem is NP-complete even when fixing $k = 2$ (Fortune, Hopcroft, and Wyllie [1]). This is in contrast with the undirected case (for those believing NP≠P), where Robertson and Seymour [4] showed that, for any fixed k, the k disjoint paths problem is solvable in polynomial time for any undirected graph (not necessarily planar).

Also, for directed planar graphs, the k disjoint paths problem is NP-complete if we do not fix k (Lynch [2]). However, in [6] it is shown that for fixed k and for directed planar graphs, it is solvable in polynomial time. We sketch the method.

For each $i = 1, \ldots, k$, choose a simple curve C_i in $\mathcal{R}^2$ connecting s_i and t_i, in such a way that the C_i are pairwise disjoint. Now the following k *disjoint homotopic paths problem* is solvable in polynomial time:

> *given:* pairwise disjoint simple curves $C_1, \ldots, C_k$, where C_i connects vertices s_i and t_i of D $(i = 1, \ldots, k)$;
> *find:* pairwise disjoint directed paths $P_1, \ldots, P_k$ in D, such that P_i is homotopic to C_i in the space $\mathcal{R}^2 \setminus \{s_1, t_1, \ldots, s_k, t_k\}$.

The curves C_i can be described equivalently by a *flow* $\phi : A \to \mathrm{FG}_k$, where FG_k denotes the free group with k generators $g_1, \ldots, g_k$. Thus at any vertex $v \notin \{s_1, t_1, \ldots, s_k, t_k\}$, the *flow conservation law* should hold; that is, if $a_1, \ldots, a_n$ are the arcs of D incident with v in clockwise order, then the product

$$\phi(a_1)^{\mathrm{sign}(a_1,v)} \cdot \ldots \cdot \phi(a_n)^{\mathrm{sign}(a_n,v)}$$

equals 1, where $\mathrm{sign}(a, v) = +1$ if a enters v and $\mathrm{sign}(a, v) = -1$ if a leaves v. If $v = s_i$, the product should be a conjugate of g_i^{-1}, and if $v = t_i$, a conjugate of g_i.

Let us call a flow $\phi : A \to \mathrm{FG}_k$ *feasible* if for each arc a: $\phi(a) \in \{1, g_1, \ldots, g_k\}$, and for each vertex v and for each two faces f and g incident with v, if $a_1, \ldots, a_m$ are the arcs incident with v met when going from f to g around v in clockwise order, then

$$\phi(a_1)^{\mathrm{sign}(a_1,v)} \cdot \ldots \cdot \phi(a_m)^{\mathrm{sign}(a_m,v)}$$

belongs to $\{1, g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}\}$.

So feasible flows correspond to solutions to the original k disjoint paths problem. Now given a flow $\phi : A \to \mathrm{FG}_k$, we can find in polynomial time a feasible flow $\psi : A \to \mathrm{FG}_k$ homotopic to ϕ, if it exists. Here ϕ and ψ are called *homotopic* if there exists a function $p : \mathcal{F} \to \mathrm{FG}_k$ such that for each arc a of D, if f and f' denote the faces incident with a at its left-hand and right-hand side, respectively, then $\psi(a) = p(f)^{-1}\phi(a)p(f')$. (Here $\mathcal{F}$ denotes the collection of faces of D.)

This follows from the polynomial-time solvability of the cohomology feasibility problem for free groups, with each $\Psi(a)$ hereditary. Indeed, by passing from the graph D to its (planar) dual graph D^*, the problem of finding a feasible flow homotopic to a given flow, is transformed to the cohomology feasibility problem.

The polynomial-time solvability of the k disjoint homotopic paths problem implies that for fixed k, the k disjoint paths problem in directed planar graphs is polynomial-time solvable: it can be shown that one can enumerate in polynomial time (for fixed k) flows $\phi_1, \ldots, \phi_N : A \to \mathrm{FG}_k$ with the property that each feasible

flow is homotopic to at least one of $\phi_1, \ldots, \phi_N$. (The exponent of the polynomial depends on k.) This is the proof method for:

THEOREM 1. *For each fixed k, the k disjoint paths problem in directed planar graphs can be solved in polynomial time.*

Note that the k disjoint paths problem asks for *any* flow; that is, one not restricted by its homotopy class. In other words, we ask for a feasible flow $\phi : A \to \mathcal{Z}^k$. (So the generators may commute; this corresponds to the possibility that curves may be shifted over each other.) Not fixing k, this is an NP-complete problem. By lifting $\mathcal{Z}^k$ to FG_k, we restrict the solution set, and obtain a polynomial-time solvable problem (also polynomial-time for nonfixed k). As the number of liftings can be bounded by a polynomial for fixed k, we can solve the original problem for fixed k in polynomial time. (In fact, generally there are infinitely many liftings, but only a restricted number of them potentially gives a feasible solution.)

3. Free partially commutative groups

The algorithm for solving the cohomology feasibility problem for free partially commutative groups (with convex sets $\Psi(a)$) implies a necessary and sufficient condition for the existence of a solution ψ, which we describe now.

There is an obvious necessary condition for the existence of such a function ψ. Let us denote a path P in D as a word $a_1 \cdots a_t$ over the alphabet $\{a, a^{-1} | a \in A\}$. In this way we indicate that P traverses the arcs $a_1, \ldots, a_t$ in this order, where $a_i = a^{-1}$ means that arc a is traversed in backward direction. A $v - w$ *path* is a path starting in v and ending in w.

Define $\phi(a^{-1}) := \phi(a)^{-1}$ and $\Psi(a^{-1}) := \Psi(a)^{-1}$. For any path $P = a_1 \cdots a_t$ define $\phi(P) := \phi(a_1) \cdots \phi(a_t) \in G$ and $\Psi(P) := \Psi(a_1) \cdots \Psi(a_t) \subseteq G$.

A necessary condition for the existence of ψ in the cohomology feasibility problem (2) is:

> for each $v \in V$ and each $v - v$ path P there exists an $x \in G$ such that $x^{-1}\phi(P)x \in \Psi(P)$. (3)

Indeed, we can take $x = p(v)$ where p is as in (1).

In some cases this condition is sufficient as well, for instance, if G is the infinite group with one generator g and each $\Psi(a)$ is convex (that is, if $g^i, g^j \in \Psi(a)$ then also $g^k \in \Psi(a)$ whenever k is inbetween i and j).

However, this condition generally is not sufficient. A stronger necessary condition is:

> for each $v \in V$ and each two $v - v$ paths P_1, P_2 there exists an $x \in G$ such that $x^{-1}\phi(P_1)x \in \Psi(P_1)$ and $x^{-1}\phi(P_2)x \in \Psi(P_2)$, (4)

since again we can take $x = p(v)$.

Now for *free partially commutative groups*, condition (4) is also sufficient, for certain subsets $\Psi(a)$. A free partially commutative group is constructed as follows. Let $g_1, \ldots, g_k$ be generators, and let E be a collection of pairs $\{i, j\}$ with

$i, j \in \{1, \ldots, k\}$ and $i \neq j$. Then the group $G = G_{k,E}$ is the group generated by $g_1, \ldots, g_k$ with relations $g_i g_j = g_j g_i$ for each $\{i, j\} \in E$. So if $E = \emptyset$ then $G_{k,E}$ is the free group generated by $g_1, \ldots, g_k$, while if E consists of all pairs from $\{1, \ldots, k\}$ then $G_{k,E}$ is isomorphic to $\mathcal{Z}^k$.

There is the following direct reduction rule for words over the 'symbols' $g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}$: if symbol α commutes with each symbol occurring in word y, then $x\alpha y\alpha^{-1}z = xyz$. It can be shown that repeating this reduction as long as possible starting with a word w, one reaches the empty word 1 if w equals 1 in the group. So the word problem can be solved easily (cf. Wrathall [10]).

Applying this reduction to a general word w, one obtains a shortest possible word w' (shortest among all words w'' that are equal to w in the group). The length of w' is denoted by $|w|$. This defines a 'norm' on $G_{k,E}$, satisfying $|1| = 0, |u^{-1}| = |u|$ and $|uw| \leq |u| + |w|$. So we can define a distance function dist on G by:

$$\text{dist}(x, y) := |x^{-1}y|$$

for $x, y \in G$. For $x, y \in G$ let $[x, y]$ be the set of all $z \in G$ satisfying $\text{dist}(x, z) + \text{dist}(z, y) = \text{dist}(x, y)$. Call a subset H of G *convex* if $1 \in H$, $[x, y] \subseteq H$ for all $x, y \in H$, $[x, y] \subseteq H^{-1}$ for all $x, y \in H^{-1}$.

Note that if G is the free group then $H \subseteq G$ is convex if and only if $H \neq \emptyset$ and H is hereditary.

In [7] the following theorem is proved.

THEOREM 2. *Let G be a free partially commutative group and let each $\Psi(a)$ be convex. Then the cohomology feasibility problem* (2) *has a solution ψ if and only if condition* (4) *is satisfied.*

The proof is based on a polynomial-time algorithm giving either the function ψ or a pair of paths P_1, P_2 violating (4). Therefore we also have:

THEOREM 3. *The cohomology feasibility problem* (2) *is solvable in polynomial time if G is a free partially commutative group and each $\Psi(a)$ is convex.*

We assume here that membership of $\Psi(a)$ of a given word can be checked in polynomial time.

4. DISJOINT CLOSED CURVES IN GRAPHS ON A COMPACT SURFACE

We describe a consequence of Theorem 2. Let S be a compact surface. A *closed curve* on S is a continuous function $C : S^1 \to S$, where S^1 is the unit circle in $\mathcal{C}$. Two closed curves C and C' are called *freely homotopic*, in notation $C \sim C'$, if there exists a continuous function $\Phi : S^1 \times [0, 1] \to S$ such that $\Phi(z, 0) = C(z)$ and $\Phi(z, 1) = C'(z)$ for each $z \in S^1$.

For any pair of closed curves C, D on S, let $\text{cr}(C, D)$ denote the number of crossings of C and D, counting multiplicities. Moreover, $\text{mincr}(C, D)$ denotes the minimum of $\text{cr}(C', D')$ where C' and D' range over closed curves freely homotopic

to C and D, respectively. That is,

$$\text{mincr}(C,D) := \min\{\text{cr}(C',D') | C' \sim C, D' \sim D\}.$$

Let $G = (V,E)$ be an undirected graph embedded on S. (We identify G with its embedding on S.) For any closed curve D on S, $\text{cr}(G,D)$ denotes the number of intersections of G and D (counting multiplicities):

$$\text{cr}(G,D) := |\{z \in S^1 | D(z) \in G\}|.$$

The following was shown in Schrijver [5] (motivated by Robertson and Seymour [3]) — it can also be derived (with surface duality) from Theorem 2.

THEOREM 4. *Let $G = (V,E)$ be an undirected graph embedded on a compact surface S and let $C_1, \ldots, C_k$ be pairwise disjoint simple closed curves on S, each non-nullhomotopic. Then there exist pairwise vertex-disjoint simple circuits $C_1', \ldots, C_k'$ in G such that $C_i' \sim C_i$ $(i = 1, \ldots, k)$, if and only if for each closed curve D on S:*

$$\text{cr}(G,D) \geq \sum_{i=1}^{k} \text{mincr}(C_i, D),$$

with strict inequality if D is doubly odd.

Here we call a closed curve D on S *doubly odd* (with respect to G and $C_1, \ldots, C_k$) if D is the concatenation $D_1 \cdot D_2$ of two closed curves D_1 and D_2 such that $D_1(1) = D_2(1) \notin G$ and such that

$$\text{cr}(G, D_j) \not\equiv \sum_{i=1}^{k} \text{cr}(C_i, D_j) \pmod 2,$$

for $j = 1, 2$.

The essence of the theorem is sufficiency of the condition.

The theorem can be extended to directed circuits in directed graphs embedded on a compact orientable surface, although the condition becomes more difficult to describe. (For the torus, see Seymour [9].) In any case, the method yields a polynomial-time algorithm finding the directed circuits.

5. PERIODIC TIMETABLING

The cohomology feasibility problem also shows up in the problem of making the timetable for Nederlandse Spoorwegen (Dutch Railways), a project currently performed for NS by CWI (with Adri Steenbeek). The Dutch railway system belongs to the busiest in the world, with several short distance trajectories, while many connections are offered, with short transfer time.

Task is to provide algorithmic means to decide if a given set of conditions on the timetable can be satisfied. In particular, the hourly pattern of the timetable is considered. The basis of the NS-timetable is a periodic cycle of 60 minutes.

How can this problem be modeled? First of all, each departure time to be determined is represented by a variable v_t. Here t is a train leg that should go every hour once. So v_t represents a variable in the cyclic group C_{60}. Similarly, the arrival time of leg t is represented by a variable a_t in C_{60}.

In the problem considered, a fixed running time is assumed for each leg. This implies that if train leg t has a running time of, say, 11 minutes, then $a_t - v_t = 11$. The waiting period of a train at a station is prescribed by an interval. E.g., if t and t' are two consecutive train legs of one hourly train, and if it is required that the train stops at the intermediate station for a period of at least 2 and at most 5 minutes, then one poses the condition that $v_{t'} - a_t \in [2, 5]$ (as interval of C_{60}).

This gives relations between train legs of one hourly train. To make connections, one has to consider train legs of two different trains. So if one wants to make a connection from leg t, arriving in Utrecht say, of one train, to a leg t' departing from Utrecht of another train, so that the transfer time is at least 3 and at most 7 minutes, then one gets the condition that $v_{t'} - a_t \in [3, 7]$.

Finally, there is the condition that for safety each two trains on the same trajectory should have a timetable distance of at least 3 minutes. That is, if train leg t of one train and train leg t' of another train run on the same railway section, then one should pose the condition $v_{t'} - v_t \in [3, 57]$.

By representing each variable by a vertex, the problem can be modeled as follows. Let $D = (V, A)$ be a directed graph, and for each $a \in A$, let $\Psi(a)$ be an interval on C_{60}. Find a function $p : V \to C_{60}$ such that $p(w) - p(u) \in \Psi(a)$ for each arc $a = (u, w)$ of D.

This is a special case of the cohomology feasibility problem. Note that (as C_{60} is abelian) one may equivalently find a 'length' function $l : A \to C_{60}$ such that $l(a) \in \Psi(a)$ for each $a \in A$ and such that each undirected circuit in D has length 0. (For arcs a in the circuit traversed backward one takes $-l(a)$ for its length.)

It is not difficult to formulate this problem as an integer linear programming problem. Indeed, if for any arc $a = (u, w)$, $\Psi(a)$ is equal to the interval $[l_a, u_a]$, we can put:

$$l_a \leq x_w - x_u + 60y_a \leq u_a, \tag{5}$$

where y_a is required to be an integer. Thus we get a system of $|A|$ linear inequalities with $|V|$ real variables x_v and $|A|$ integer variables y_a. In fact, if there is a solution, there is also one with the x_v being integer as well (as the x variables make a network matrix).

Now in solving (5), one may choose a spanning tree T in D, and assume that $y_a = 0$ for each arc a in T (cf. Serafini and Ukovich [8]). Alternatively, one may consider the problem as follows.

A *circulation* is a function $f : A \to \mathcal{R}$ such that the 'flow conservation law':

$$\sum_{a \in \delta^-(v)} f(a) = \sum_{a \in \delta^+(v)} f(a)$$

holds for each vertex v of D. Here $\delta^-(v)$ and $\delta^+(v)$ denote the sets of arcs entering v and leaving v, respectively.

Let L be the lattice of all integer-valued circulations. Now one can describe the problem as one of finding a linear function $\Phi : L \to \mathcal{Z}$ such that there exist z_a (for $a \in A$) with the properties that $l_a \leq z_a \leq u_a$ for each arc A and $z^T f = 60\Phi(f)$ for each $f \in L$.

The existence of such z_a can be checked in polynomial time, given the values of Φ on a basis of L. Indeed, for $a \in T$ let $y_a = 0$, and for $a \notin T$ let $y_a = \phi(f)$, where f is the incidence vector of the circuit in $T \cup \{a\}$ (so f is a circulation: $f(a') = 1$ on forward arcs a' in the circuit, $f(a') = -1$ on backward arcs a' in the circuit, and $f(a') = 0$ on each other arc a'). Then for this fixed y we can test (5) in polynomial time (with the Bellman-Ford method), which is equivalent to finding z as required.

Hence, in searching a feasible timetable one can branch on choices of Φ. Each Φ corresponds to a homotopy class of solutions of the timetable problem.

Again, this amounts to a lifting, now from C_{60} to $\mathcal{Z}$. Indeed, we consider for each arc $a \notin T$ a translation by $60y_a$ of the feasible interval, considered as interval on $\mathcal{Z}$, and try to solve the problem over $\mathcal{Z}$.

We also note that, given Φ, if there exist z_a, one can optimize the z_a under any linear (or convex piecewise linear) objective function (for instance, passenger waiting time).

Typically, the problems coming from NS have about 3000 variables with about 10,000 constraints. In a straightforward way they can be reduced to about 200 variables with about 600 constraints. The above observations turn out to require a too heavy framework in order to solve the problem fast in practice (although they are of help in optimizing a given solution).

The package CADANS that CWI is developing for NS for solving the problem above, is based on a fast constraint propagation technique and fast branching heuristics designed by Adri Steenbeek. It gives, in a running time of the order of 1-10 minutes either a solution (i.e., a feasible timetable), or an inclusionwise minimal set of constraints that is infeasible. If CADANS gives the latter answer, the user should drop, or relax, at least one of the constraints in the minimal set in order to make the constraints feasible. Thus CADANS can be used interactively to support the planner. Alternatively, it can uncover bottlenecks in the infrastructure, and indicate where extra infrastructure (viaducts, flyovers, four-tracks) should be built in order to make a given set of conditions feasible.

REFERENCES

[1] S. Fortune, J. Hopcroft, J. Wyllie, The directed subgraph homeomorphism problem, *Theoretical Computer Science* 10 (1980) 111–121.

[2] J.F. Lynch, The equivalence of theorem proving and the interconnection problem, *(ACM) SIGDA Newsletter* 5 (1975) 3:31–36.

[3] N. Robertson, P.D. Seymour, Graph minors. VII. Disjoint paths on a surface, *Journal of Combinatorial Theory, Series B* 45 (1988) 212–254.

[4] N. Robertson, P.D. Seymour, Graph minors. XIII. The disjoint paths problem, *Journal of Combinatorial Theory, Series B* 63 (1995) 65–110.

[5] A. Schrijver, Disjoint circuits of prescribed homotopies in a graph on a compact surface, *Journal of Combinatorial Theory, Series B* 51 (1991) 127–159.

[6] A. Schrijver, Finding k disjoint paths in a directed planar graph, *SIAM Journal on Computing* 23 (1994) 780–788.

[7] A. Schrijver, *Free partially commutative groups, cohomology, and paths and circuits in directed graphs on surfaces*, preprint, 1994.

[8] P. Serafini, W. Ukovich, A mathematical model for periodic scheduling problems, *SIAM Journal on Discrete Mathematics* 2 (1989) 550–581.

[9] P.D. Seymour, Directed circuits on a torus, *Combinatorica* 11 (1991) 261–273.

[10] C. Wrathall, The word problem for free partially commutative groups, *Journal of Symbolic Computation* 6 (1988) 99–104.

CWI, Kruislaan 413,
1098 SJ Amsterdam,
The Netherlands
and
Department of Mathematics,
University of Amsterdam,
Plantage Muidergracht 24,
1018 TV Amsterdam,
The Netherlands.

Open Dynamical Systems and their Control

Jan C. Willems

Abstract. A mathematical framework for studying open dynamical systems is sketched. Special attention is given in the exposition to linear time-invariant differential systems. The main concepts that are introduced are the behavior, manifest and latent variables, controllability, and observability. The paper ends with a discussion of control, which is viewed as system interconnection.

1991 Mathematics Subject Classification: 93B05, 93B07, 93B36, 93B51, 93C05, 93C15
Keywords and Phrases: Dynamical systems, open systems, behaviors, controllability, observability, control, stabilization.

1 Introduction

The purpose of this presentation is to explain some of the main features of the theory of open dynamical systems. The adjective *'open'* refers to systems that interact with their environment. This interaction may take the form of exchange of a physical quantity as mass or energy, or it may simply consist of exchange of information. Closed dynamical systems have been studied very extensively in mathematics. Typically these lead to models of the general form $\frac{d}{dt}x = f(x)$. The evolution of such systems is completely determined by the dynamical laws (expressed by the vector-field of f) and the initial state $x(0)$. In open dynamical systems, however, the evolution of the system variables is determined by the dynamical laws, the initial conditions, and, in addition, by the influence of the environment. This may for instance take the form of an external input function that drives the system. Examples of application areas where this interaction with the environment is essential are signal processing and control. Whereas in signal processing it is reasonable to view the input function as a given (or stochastically described) time-function, this is not the case in application areas as control, since in this case the input function is usually generated by a mechanism which selects the input on the basis of the evolution of output variables in the system itself. This feature leads to *'feedback'* which forms the central concept of control, ever since the subject came into existence.

2 Dynamical systems

A first goal is to put forward a notion that serves to describe open dynamical systems mathematically. A framework that has shown to be quite effective, both in terms of generality and applicability, is called the *'behavioral approach'*. One of its main features is that it does not start from an input/output structure or map, nor from a state space model. Instead, any family of trajectories parameterized by time is viewed as a dynamical system. The theory underlying this approach has been treated in [16, 17, 12]. Here we can only describe a few of the bare essentials.

A *dynamical system* Σ is triple $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ with $\mathbb{T} \subset \mathbb{R}$ the time-set, $\mathbb{W}$ the signal space, and $\mathfrak{B} \subset \mathbb{W}^{\mathbb{T}}$ the *behavior*. The intuition behind this definition is that $\mathbb{T}$ is the set of relevant time-instances; $\mathbb{W}$ is the set in which the signals, whose dynamic relation Σ models, take on their values; the behavior $\mathfrak{B}$ specifies which signals $w : \mathbb{T} \to \mathbb{W}$ obey the laws of the system. The time-set $\mathbb{T}$ equals for example $\mathbb{R}$ or $\mathbb{R}_+$ in continuous-time, and $\mathbb{Z}$ or $\mathbb{Z}_+$ in discrete-time systems. Important properties of dynamical systems are linearity and time-invariance; Σ is said to be *linear* if $\mathbb{W}$ is a vector space and $\mathfrak{B}$ is a linear subspace of $\mathbb{W}^{\mathbb{T}}$, and *time-invariant* (assuming $\mathbb{T} = \mathbb{R}$ or $\mathbb{Z}$) if $\sigma^t \mathfrak{B} = \mathfrak{B}$ for all $t \in \mathbb{T}$, where σ^t denotes the t-shift (defined by $(\sigma^t f)(t') := f(t' + t)$). There is much interest in generalization from a time-set that is a subset of $\mathbb{R}$ to domains with more independent variables (e.g., time and space). These 'dynamical' systems have $\mathbb{T} \subset \mathbb{R}^n$, and are referred to as *n-D* systems.

3 Differential systems

The 'ideology' of the behavioral approach is based on the belief that in a model of a dynamical (physical) phenomenon, it is the behavior $\mathfrak{B}$, i.e., a set of trajectories $w : \mathbb{T} \to \mathbb{W}$, that is the central object of study. But, this set of trajectories must be specified somehow, and it is here that differential (and difference) equations enter the scene. Of course, there are important examples where the behavior is specified in other ways (for example, in Kepler's laws for planetary motion), but differential equations are certainly the most prevalent specification of behaviors encountered in applications. For $\mathbb{T} = \mathbb{R}$, $\mathfrak{B}$ then consists of the solutions of a system of differential equations as $f(w, \frac{d}{dt}w, \ldots, \frac{d^N}{dt^N}w) = 0$. We call these *differential* systems. Of particular interest (at least in control, signal processing, and circuit theory, etc.) are systems with a signal space that is a finite-dimensional vector space and behavior described by linear constant-coefficient differential equations. The fact that non-trivial new things can be said about such systems, which from a mathematical point of view may appear very simple, is due to the many meaningful new concepts originating from the interaction of systems with their environment.

A *linear time-invariant differential system* is a dynamical system $\Sigma = (\mathbb{R}, \mathbb{W}, \mathfrak{B})$, with $\mathbb{W}$ a finite-dimensional (real) vector space, whose behavior consists of the solutions of

$$R(\frac{d}{dt})w = 0, \tag{1}$$

with $R \in \mathbb{R}^{\bullet \times \bullet}[\xi]$ a real polynomial matrix. Of course, the number of columns

of R equals the dimension of $\mathbb{W}$. The number of rows of R, which represents the number of equations, is arbitrary. In fact, when the row dimension of R is less than its column dimension, $R(\frac{d}{dt})w = 0$ is an under-determined system of differential equations which is typical for models in which the influence of the environment is taken into account. The definition of a solution of $R(\frac{d}{dt})w = 0$ is an issue. There is much to be said for considering solutions in $\mathcal{L}^{\mathrm{loc}}(\mathbb{R}, \mathbb{W})$ and interpreting $R(\frac{d}{dt})w$ as a distribution. This allows steps, ramps, etc., which are often used in engineering applications. Nevertheless, for ease of exposition, we define the behavior to be

$$\{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{W}) \mid R(\frac{d}{dt})w = 0\}. \tag{2}$$

We denote this behavior as $\ker(R(\frac{d}{dt}))$, the set of linear time-invariant differential systems by $\mathfrak{L}^\bullet$, and those with $\dim(\mathbb{W}) = \mathtt{w}$ by $\mathfrak{L}^{\mathtt{w}}$. Whence $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \mathfrak{B}) \in \mathfrak{L}^{\mathtt{w}}$ means that there exists a $R \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi]$ such that $\mathfrak{B} = \ker(R(\frac{d}{dt}))$. We call $R(\frac{d}{dt})w = 0$ a *kernel representation* of Σ. Note that we may as well write $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$, instead of $\Sigma \in \mathfrak{L}^{\mathtt{w}}$, since the time-axis ($\mathbb{R}$) and the signal space ($\mathbb{R}^{\mathtt{w}}$) are evident from this notation.

Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$. Define the *consequences* of $\mathfrak{B}$ to be the set $\mathcal{N}_{\mathfrak{B}} := \{n \in \mathbb{R}^{\mathtt{w}}[\xi] \mid n^T(\frac{d}{dt})\mathfrak{B} = 0\}$. It is easy to see that $\mathcal{N}_{\mathfrak{B}}$ is an $\mathbb{R}[\xi]$-submodule of $\mathbb{R}^{\mathtt{w}}[\xi]$, that for $\mathfrak{B} = \ker(R(\frac{d}{dt}))$, $\mathcal{N}_{\mathfrak{B}}$ equals the submodule spanned by the transposes of the rows of R, and that there is a one-to-one relation between $\mathfrak{L}^{\mathtt{w}}$ and the $\mathbb{R}[\xi]$-submodules of $\mathbb{R}^{\mathtt{w}}[\xi]$. This property, however, depends on the fact that we used $\mathfrak{C}^\infty$-solutions. The same one-to-one correspondence holds with distributional solutions, but not with $\mathfrak{C}^\infty$- (or distributional) solutions with compact support. A problem that remains unsolved is to give a crisp characterization for subspaces of $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$ to be elements of $\mathfrak{L}^{\mathtt{w}}$. In the discrete-time case, the analogous systems can be nicely specified: $\mathfrak{B}$ must be a linear, shift-invariant subspace of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}}$, and closed in the topology of point-wise convergence [16, 17].

The one-to-one relationship between certain classes of dynamical systems and certain submodules has been studied in other situations as well [9, 11, 10]. For example, it holds for constant-coefficient *PDE*'s. Let $R \in \mathbb{R}^{\bullet \times \mathtt{w}}[\xi_1, \xi_2, \ldots, \xi_n]$ be a polynomial matrix in n variables. It induces the *PDE*

$$R(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n})w = 0 \tag{3}$$

in the functions $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mapsto (w_1(x), w_2(x), \ldots, w_{\mathtt{w}}(x)) \in \mathbb{R}^{\mathtt{w}}$. Define the behavior of this *PDE* as

$$\{w \in \mathfrak{C}^\infty(\mathbb{R}^n, \mathbb{R}^{\mathtt{w}}) \mid R(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n})w = 0\}. \tag{4}$$

It turns out that, as in the case with one independent variable, there is again a one-to-one relation between these behaviors and the $\mathbb{R}[\xi_1, \xi_2, \ldots, \xi_n]$-submodules spanned by the rows of R [11]. Analogous, but technically more involved, results have been obtained for time-varying linear systems with hyper-functions as solutions and the ring of time-varying differential operators having coefficients in $\mathbb{R}(t)$ without poles on the real axis [10].

4 Latent variables and elimination

Mathematical models of complex systems are usually obtained by viewing the system (often in a hierarchical fashion) as an interconnection of subsystems, modules (standard components), for which a model can be found in a database. This principle of *tearing* and *zooming*, combined with *modularity*, lies at the basis of what is called *object-oriented* modelling, a very effective computer assisted way of model building used in many engineering domains. An important aspect of these object-oriented modelling procedures is that they lead to a model that relates the variables whose dynamic relation one wants to model (we call these *manifest* variables) to auxiliary variables (we call these *latent* variables) that have been introduced in the modelling process, for example as variables that specify the interconnection constraints. For differential systems this leads to equations as

$$f_1(w, \frac{d}{dt}w, \dots, \frac{d^N}{dt^N}w, \ell, \frac{d}{dt}\ell, \dots, \frac{d^N}{dt^N}\ell) = f_2(w, \frac{d}{dt}w, \dots, \frac{d^N}{dt^N}w, \ell, \frac{d}{dt}\ell, \dots, \frac{d^N}{dt^N}\ell),$$

relating the (vector of) manifest variables w to the (vector of) latent variables ℓ. In the linear time-invariant case this becomes

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell, \tag{5}$$

with R and M polynomial. Define the *manifest* behavior of (5) as

$$\{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathbf{w}}) \mid \exists \ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\bullet}) \text{ such that } R(\frac{d}{dt})w = M(\frac{d}{dt})\ell\}. \tag{6}$$

We call (5) *latent variable* representation of (6). The question occurs whether (6) is in $\mathfrak{L}^{\mathbf{w}}$. This is the case indeed.

THEOREM 1 : *For any real polynomial matrices* (R, M) *with* $\operatorname{rowdim}(R) = \operatorname{rowdim}(M)$, *there exists a real polynomial matrix* R' *such that the manifest behavior of* $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$ *has the kernel representation* $R'(\frac{d}{dt})w = 0$.

The above theorem is called the *elimination theorem*. Its relevance in object-oriented modelling is as follows. A model obtained this way usually involves very many variables and equations, among them many algebraic ones. The elimination theorem tells that the latent variables may be eliminated and that the number of equations can be reduced to no more than the number of manifest variables. Of course, the order of the differential equation goes up in the elimination process.

The theoretical basis that underlies the elimination theorem is the *fundamental principle*. It gives necessary and sufficient conditions for solvability for $x \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\bullet})$ in the equation $F(\frac{d}{dt})x = y$ with $F \in \mathbb{R}^{\bullet \times \bullet}[\xi]$ and $y \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\bullet})$ given. Define the *annihilators* of F as $\mathcal{K}_F := \{n \in \mathbb{R}^{\operatorname{rowdim}(F)} \mid n^T F = 0\}$. The fundamental principle states that $F(\frac{d}{dt})x = y$ is solvable if and only if $n^T(\frac{d}{dt})y = 0$ for all $n \in \mathcal{K}_F$. This immediately yields the elimination theorem. For the case at hand, it is rather easy to prove the fundamental principle, but there are interesting generalizations where it is a deep mathematical result. For example, for the constant-coefficient *PDE*'s, and for the time-varying linear systems discussed in section 3. Thus the elimination theorem also holds for these classes of systems. The elimination problem has also been studied For nonlinear systems [4].

5 Controllability

An important property in the analysis and synthesis of open dynamical systems is controllability. Controllability refers to be ability of transferring a system from one mode of operation to another. By viewing the first mode of operation as undesired and the second one as desirable, the relevance to control and other areas of applications becomes clear. The concept of controllability has been introduced around 1960 in the context of state space systems. It is one of the notions that is endogenous to control theory. The classical definition runs as follows. The system described by the controlled vector-field $\frac{d}{dt}x = f(x,u)$ is said to be controllable if $\forall a, b, , \exists u$ and $T \geq 0$ such that the solution to $\frac{d}{dt}x = f(x,u)$ and $x(0) = a$ yields $x(T) = b$. One of the elementary results of system theory [1] states that the finite-dimensional linear system $\frac{d}{dt}x = Ax + Bu$ is controllable if and only if the matrix $[B\ AB\ A^2B \cdots A^{\dim(x)-1}B]$ has full row rank. Various generalizations of this result to time-varying, to nonlinear (involving Lie brackets) [7, 8, 2, 15], and to infinite-dimensional systems exist [3].

A disadvantage of the notion of controllability as formulated above is that it refers to a particular representation of a system, notably a state space representation. Thus a system may be uncontrollable either for the intrinsic reason that the control has insufficient influence on the system variables, or because the state has been chosen in an inefficient way. It is clearly not desirable to confuse these reasons. In the context of behavioral systems, a definition of controllability has been put forward that involves the system variables directly.

Let $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ be a dynamical system with $\mathbb{T} = \mathbb{R}$ or $\mathbb{Z}$, and assume that is time-invariant. Σ is said to be *controllable* if for all $w_1, w_2 \in \mathfrak{B}$ there exists $T \in \mathbb{T}$, $T \geq 0$ and $w \in \mathfrak{B}$ such that $w(t) = w_1(t)$ for $t < 0$ and $w(t) = w_2(t-T)$ for $t \geq T$. Thus controllability refers to the ability to switch from any one trajectory in the behavior to any other one, allowing some time-delay.

Two questions that occur are the following: What conditions on the parameters of a system representation imply controllability? Do controllable systems admit a particular representation in which controllability becomes apparent? For linear time-invariant differential systems, these questions are answered in the following theorem.

Theorem 2 : *Let $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathbb{w}}, \mathfrak{B}) \in \mathfrak{L}^{\mathbb{w}}$. The following are equivalent:*

1. *The system Σ is controllable;*
2. *The polynomial matrix R in a kernel representation $R(\frac{d}{dt})w = 0$ of $\mathfrak{B}$ satisfies $\operatorname{rank}(R(\lambda)) = \operatorname{rank}(R)$ for all $\lambda \in \mathbb{C}$;*
3. *The behavior $\mathfrak{B}$ is the image of a linear constant-coefficient differential operator, that is, there exists a polynomial matrix $M \in \mathbb{R}^{\mathbb{w}\times\bullet}[\xi]$ such that $\mathfrak{B} = M(\frac{d}{dt})\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\operatorname{coldim}(M)})$;*
4. *The compact support trajectories of $\mathfrak{B}$ are dense (in the $\mathfrak{C}^\infty$-topology) in $\mathfrak{B}$;*
5. *The $\mathbb{R}[\xi]$-module $\mathbb{R}^{\mathbb{w}}[\xi]/\mathcal{N}\mathfrak{B}$ is torsion-free.*

There exist various algorithms for verifying controllability of a system $\Sigma \in \mathfrak{L}^{\bullet}$ starting from the coefficients of the polynomial matrix R in a kernel (or a latent variable) representation of Σ, but we will not enter into these algorithmic aspects.

A point of the above theorem that is worth emphasizing is that controllable systems admit a representation as the manifest behavior of the latent variable system of the special form

$$w = M(\frac{d}{dt})\ell. \tag{7}$$

We call this an *image* representation. It follows from the elimination theorem that every system in image representation can be brought in kernel representation. But not every system in kernel representation can be brought in image representation: it is precisely the controllable ones for which this is possible.

The controllability issue has been pursued for many other classes of systems. In particular (more difficult to prove) generalizations have been derived for differential-delay [14, 6], for nonlinear, for *n-D* systems [13, 9], and, as we will discuss soon, for *PDE*'s. Systems in an image representation have received much attention recently for nonlinear differential-algebraic systems, where they are referred to as *flat* systems [5]. Flatness implies controllability, but the exact relation remains to be discovered.

We now explain the generalization to constant-coefficient *PDE*'s. Consider the system defined by (3,4). This system is said to be *controllable* if for all w_1, w_2 in the behavior (4) and for all open subsets O_1, O_2 of $\mathbb{R}^n$ with disjoint closure, there exists w in (4) such that $w|_{O_1} = w_1|_{O_1}$ and $w|_{O_2} = w_2|_{O_2}$. The following result has been obtained in [11].

THEOREM 3 : *The following statements are equivalent:*

1. *(3) defines a controllable system;*

2. *(4) admits an image representation, i.e., there exists a polynomial matrix $M \in \mathbb{R}^{\mathbb{w}\times\bullet}[\xi_1, \xi_2, \ldots, \xi_n]$ such that (4) equals*

$$M(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n})\mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathrm{coldim}(M)});$$

3. *The trajectories of compact support are dense in (4).*

It is a simple consequence of this theorem that a scalar *PDE* in one function (i.e., with $\mathrm{rowdim}(R) = \mathrm{coldim}(R) = 1$) with $R \neq 0$ cannot be controllable. It can be shown, on the other hand, that Maxwell's equations (in which case $\mathrm{rowdim}(R) = 8$ and $\mathrm{coldim}(R) = 10$) are controllable. Note that an image representation corresponds to what in mathematical physics is the existence of a *potential function*. An interesting aspect of the above theorem therefore is the fact that it identifies the existence of a potential function with the system theoretic property of controllability and concatenability of behaviors.

6 Observability

The notion of observability was introduced hand in hand with controllability. In the context of the input/state/output system $\frac{d}{dt}x = f(x,u), y = h(x,u)$, it refers to the possibility of deducing, using the laws of the system, the state from observation of the input and the output. The definition that is used in the behavioral context is more general in that the variables that are observed and the variables that need to be deduced are kept general.

Let $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ be a dynamical system, and assume that $\mathbb{W}$ is a product space: $\mathbb{W} = \mathbb{W}_1 \times \mathbb{W}_2$. Then w_1 is said to be *observable* from w_2 in Σ if $(w_1, w_2') \in \mathfrak{B}$ and $(w_1, w_2'') \in \mathfrak{B}$ imply $w_2' = w_2''$. Observability thus refers to the possibility of deducing the trajectory w_1 from observation of w_2 and from the laws of the system ($\mathfrak{B}$ is assumed to be known).

The theory of observability runs parallel to that of controllability. We mention only the result that for linear time-invariant systems, w_1 is observable from w_2 if and only if there exists a set of consequences of the system behavior of the following form that puts observability into evidence: $w_1 = R_2'(\frac{d}{dt})w_2$.

7 Control

In order to illustrate the idea of the nature of control that we would like to transmit in this presentation, consider the system configuration depicted in figure 1. In the top part of the figure, there are two systems, shown as proverbial black-boxes with terminals. It is through their terminals that systems interact with their environment. The black-box imposes relations on the variables that 'live' on its terminals. These relations are formalized by the behavior of the system in the black-box. The system to the left in figure 1 is called the *plant*, the one to the right the *controller*. The terminals of the plant consist of *to-be-controlled variables* w, and *control variables* c. The controller has only terminals with the control variables c. In the bottom part of the figure, the control terminals of the plant and of the controller are connected. Before interconnection, the variables w and c of the plant have to satisfy the laws imposed by the plant behavior. But, after interconnection, the variables c also have to satisfy the laws imposed by the controller. Thus, after interconnection, the restrictions imposed on the variables c by the controller will be transmitted to the variables w. Choosing the black-box to the right so that the variables w have a desirable behavior in the interconnected black-box is, in our view, the basic problem of control. This point of view is discussed with examples in [18].

In the remainder of this paper we describe one simple controller design problem in this setting. Let the variables w be partitioned into two sets: $w = (d, z)$ with the d's exogenous disturbances, and the z's endogenous to-be-controlled variables. Assume that the plant is a linear time-invariant differential system with behavior $\mathcal{P} \in \mathfrak{L}^{\mathtt{d}+\mathtt{z}+\mathtt{c}}$, called the *plant behavior*. Assume further that the exogenous disturbances d are *free* in $\mathcal{P}$, that is, that for all $d \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{d}})$ there exist (z, c) such that $(d, z, c) \in \mathcal{P}$. Now consider the controller, also assumed to be a linear time-invariant differential system, with behavior $\mathcal{C} \in \mathfrak{L}^{\mathtt{c}}$, called the *controller*

behavior. With the controller put into place, the behavior of the to-be-controlled variables becomes

$$\mathcal{K} = \{(d, z) \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathbf{d}+\mathbf{z}}) \mid \exists c \in \mathcal{C} \text{ such that } (d, z, c) \in \mathcal{P}\}. \tag{8}$$

By the elimination theorem, $\mathcal{K} \in \mathfrak{L}^{\mathbf{d}+\mathbf{z}}$. We call $\mathcal{K}$ the *controlled behavior*.

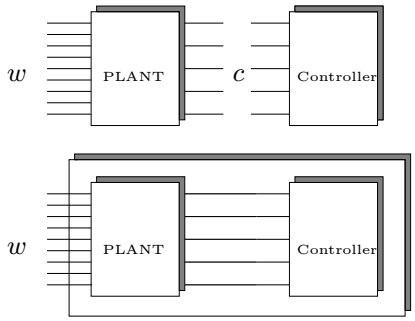


Figure 1: Controller interconnection

The controller $\mathcal{C}$ usually has to satisfy certain practical implementability constraints, perhaps as a signal processor that transforms sensor outputs into actuator inputs, or using physical energy-based constraints, etc. Here, we assume that the controller can be any linear time-invariant differential system that leave the exogenous disturbances free. This is the case if and only if its behavior $\mathcal{C}$ has the property that for all $d \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathbf{d}})$, there exists (z, c) such that both $(d, z, c) \in \mathcal{P}$ and $c \in \mathcal{C}$. We call this set of controllers *admissible controllers*, and denote it by $\mathfrak{C}$.

The control problem that now emerges is that of choosing, for a given plant $\mathcal{P}$, an admissible controller $\mathcal{C} \in \mathfrak{C}$ such that the controlled behavior $\mathcal{K}$ meets certain specifications. We consider *pole placement*, which, as we shall see, implies *stabilization*. We now explain what this means. Consider the controlled system $\mathcal{K} \in \mathfrak{L}^{\mathbf{d}+\mathbf{z}}$. When the controller which generates $\mathcal{K}$ is admissible, d must be free in $\mathcal{K}$. This implies that $\mathcal{K}$ has a kernel representation $P(\frac{d}{dt})z = Q(\frac{d}{dt})d$ with P a polynomial matrix of full row rank. Define the *characteristic polynomial* $\pi_\mathcal{K}$ of $\mathcal{K}$ as follows. If P is not square (and hence wide) $\pi_\mathcal{K} := 0$. Otherwise, $\pi_\mathcal{K} = \det(P)$ where it is assumed that P is chosen such that $\det(P)$ is monic. We call the roots of $\pi_\mathcal{K}$ the *poles* of $\mathcal{K}$. If $\pi_\mathcal{K} \neq 0$, then the behavior $\mathcal{K}_0 = \{(d, z) \in \mathcal{K} \mid d = 0\}$ is finite-dimensional, and the exponents of its exponential responses are the roots of $\pi_\mathcal{K}$.

Note that controllability can be defined when there are more variables in the model than just those that need to be concatenated. Similarly, observability can also be defined when there are more variables in the model that just the observed and the to-be-deduced ones. The definitions are evident. We now state necessary and sufficient conditions for pole assignability.

Theorem 4 : *Let the plant behavior $\mathcal{P} \in \mathfrak{L}^{\mathtt{d}+\mathtt{z}+\mathtt{c}}$ be given. Then there exists, for any monic polynomial $r \in \mathbb{R}[\xi]$, an admissible controller $\mathcal{C} \in \mathfrak{C}$ such that the resulting controlled system $\mathcal{K} \in \mathfrak{L}^{\mathtt{d}+\mathtt{z}}$ has $\pi_{\mathcal{K}} = r$ if the exogenous to-be-controlled variables z are (i) controllable in $\mathcal{P}_0 := \{(d, z, c) \in \mathcal{P} \mid d = 0\}$, and (ii) observable from c in $\mathcal{P}$.*

The controlled behavior $\mathcal{K}$ is said to be *stable* if $(d, z) \in \mathcal{K}$ and $d = 0$ implies that $w(t) \to 0$ as $t \to \infty$. Obviously $\mathcal{K}$ is stable if and only if $\pi_{\mathcal{K}}$ is a Hurwitz polynomial. The above theorem gives controllability and observability conditions that are sufficient for stabilizability. Pole placement and stabilization are very coarse controller design specifications. But also other, more refined, design specifications, for example H_∞-control and robust stability, can be treated in this setting.

These results generalize the classical state space pole placement results in a number of ways. However, we regard the main contribution of the above theorem to be the underlying idea of control. We view *interconnection* as the principle of control. It supersedes the special case of trajectory selection and optimization (often called *open-loop (optimal) control*, and the (very important) special case of feedback control (often called *intelligent control*), in which a signal processor uses the plant sensor outputs in order to select the plant actuator inputs. The latter area is the classical view of control and will undoubtedly gain in importance for technological applications as logical devices and on-line computation becomes cheaper, more reliable, and more powerful. However, by considering interconnection as the basic principle of control, the scope of the subject and its relevance to the design of physical systems can be enhanced in meaningful directions, by making the (optimal) design of subsystems, i.e., integrated system design, as the aim and the domain of the subject.

References

[1] R.W. Brockett, *Finite Dimensional Linear Systems*, Wiley, 1970.

[2] R.W. Brockett, System theory on group manifolds and coset spaces, *SIAM Journal on Control*, volume 10, pages 265-284, 1972.

[3] R.F. Curtain and H.J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, 1995.

[4] S. Diop, Elimination in control theory, *Mathematics of Control, Signals, and Systems,* volume 4, pages 17-32, 1991.

[5] M. Fliess and S.T. Glad, An algebraic approach to linear and nonlinear control, pages 223-267 of *Essays on Control: Perspectives in the Theory and Its Applications*, edited by H.L. Trentelman and J.C. Willems, Birkhäuser, 1993.

[6] H. Glüsing-Lüerssen, A behavioral approach to delay-differential systems, *SIAM Journal on Control and Optimization*, volume 35, pages 480-499, 1997.

[7] A. Isidori, *Nonlinear Control Systems*, Springer-Verlag, 1989.

[8] A. Nijmeijer and A.J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, 1990.

[9] U. Oberst, Multidimensional constant linear systems, *Acta Applicandae Mathematicae*, volume 20, pages 1-175, 1990.

[10] S. Fröhler and U. Oberst, Continuous time-varying linear systems, manuscript, 1998.

[11] H.K. Pillai and S. Shankar, A behavioural approach to control of distributed systems, *SIAM Journal on Control and Optimization*, to appear.

[12] J.W. Polderman and J.C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach,* Springer-Verlag, 1998.

[13] P. Rocha and J.C. Willems, Controllability of 2-D systems, *IEEE Transactions on Automatic Control*, volume 36, pages 413-423, 1991.

[14] P. Rocha and J.C. Willems Behavioral controllability of delay-differential Systems, *SIAM Journal on Control and Optimization*, volume 35, pages 254-264, 1997.

[15] H.J. Sussmann, Lie brackets and local controllability, *SIAM Journal on Control and Optimization*, volume 21, pages 686-713, 1983.

[16] J.C. Willems, Models for dynamics, *Dynamics Reported,* volume 2, pages 171-269, 1989.

[17] J.C. Willems, Paradigms and puzzles in the theory of dynamical systems, *IEEE Transactions on Automatic Control,* volume 36, pages 259-294, 1991.

[18] J.C.Willems, On interconnections, control, and feedback, *IEEE Transactions on Automatic Control*, volume 42, pages 326-339, 1997.

Jan C. Willems
University of Groningen
9700 AV Groningen, NL
email: Willems@math.rug.nl

Free Material Optimization

Michal Kočvara and Jochem Zowe

Abstract. Free material design deals with the question of finding the stiffest structure with respect to one or more given loads which can be made when both the distribution of material and the material itself can be freely varied. We consider here the general multiple-load situation. After a series of transformation steps we reach a problem formulation for which we can prove existence of a solution; a suitable discretization leads to a semidefinite programming problem for which modern polynomial time algorithms of interior-point type are available. Two numerical examples demonstrates the efficiency of our approach.

1991 Mathematics Subject Classification: 73K40; 90C90; 90C25
Keywords and Phrases: Structural optimization; Material optimization; Topology optimization; Semidefinite programming

1 Problem Formulation

In this section we introduce the problem of free material optimization. Only basic description of the problem is given; for more details the reader is refered to [2, 5, 1]. We study the optimization of the design of an elastic continuum structure that is loaded by multiple independent forces. The *material properties at each point* are the design variables. We start from the infinite-dimensional problem setting, show the existence of a solution after a reformulation of the problem and, after discretization, reach a finite-dimensional formulation expressed as a *semidefinite program*, and as such accessible to modern numerical interior-point methods.

First we sketch the single-load model in the two-dimensional space. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain (the elastic body) with Lipschitz boundary Γ. The standard notation $[H^1(\Omega)]^2$ and $[H^1_0(\Omega)]^2$ for Sobolev spaces of functions $v : \Omega \to \mathbb{R}^2$ is used. By $u(x) = (u_1(x), u_2(x))$ with $u \in [H^1(\Omega)]^2$ we denote the *displacement vector* at point x of the body under load. Further, let

$$e_{ij}(u(x)) = \frac{1}{2}\left(\frac{\partial u_i(x)}{\partial x_j} + \frac{\partial u_j(x)}{\partial x_i}\right) \qquad \text{for } i, j = 1, 2$$

denote the *(small-)strain tensor*, and $\sigma_{ij}(x), i, j = 1, 2$, the *stress tensor*. To simplify the notation we will often skip the space variable x in u, e, etc.

Our system is governed by linear Hooke's law, i.e., the stress is a linear function of the strain

$$\sigma_{ij}(x) = E_{ijkl}(x)e_{kl}(u(x)) \quad \text{(in tensor notation)}, \tag{1}$$

where $E(x)$ is the (plain-stress) *elasticity tensor* of order 4; this tensor characterizes the elastic behaviour of material at point x. The strain and stress tensors are symmetric and also E is symmetric in the following sense:

$$E_{ijkl} = E_{jikl} = E_{ijlk} = E_{klij} \qquad \text{for } i,j,k,l = 1,2.$$

These symmetries allow us to and interpret the 2-tensors e and σ as vectors

$$e = (e_{11}, e_{22}, \sqrt{2}e_{12})^T \in \mathbb{R}^3, \qquad \sigma = (\sigma_{11}, \sigma_{22}, \sqrt{2}\sigma_{12})^T \in \mathbb{R}^3.$$

Correspondingly, the 4-tensor E can be written as a symmetric 3×3 matrix

$$E = \begin{pmatrix} E_{1111} & E_{1122} & \sqrt{2}E_{1112} \\ & E_{2222} & \sqrt{2}E_{2212} \\ \text{sym.} & & 2E_{1212} \end{pmatrix}. \tag{2}$$

In this notation, equation (1) reads as $\sigma(x) = E(x)e(u(x))$. Henceforth, E will be understood as a matrix and we will use double indices for its elements. To allow switches from material to no-material, we work with $E \in [L^\infty(\Omega)]^{3\times 3}$.

We consider a partitioning of the boundary Γ into two parts: $\Gamma = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$, where Γ_1 and Γ_2 are open in Γ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. Further we put

$$\mathcal{H} = \{u \in [H^1(\Omega)]^2 \,|\, u_i = 0 \text{ on } \Gamma_1 \text{ for } i = 1 \text{ or } 2 \text{ or any combination}\},$$

i.e., $[H_0^1(\Omega)]^2 \subset \mathcal{H} \subset [H^1(\Omega)]^2$. To exclude rigid-body movements, we assume throughout that

$$\{v \in \mathcal{H} \,|\, v_i = a_i + bx_i,\ a_i \in \mathbb{R},\ i = 1,2,\ b \in \mathbb{R} \text{ arbitrary}\} = \emptyset.$$

For the elasticity tensor E and a given external load $f \in [L_2(\Gamma_2)]^{dim}$ the *potential energy* of an elastic body as a function of the displacement $u \in \mathcal{H}$ is given by

$$-\frac{1}{2}\int_\Omega \langle Ee(u), e(u)\rangle \, dx + F(u) \quad \text{with} \quad F(u) := \int_{\Gamma_2} f \cdot u \, dx. \tag{3}$$

The system is in equilibrium for u^* which maximizes (3), i.e., u^* which solves

$$\sup_{u\in\mathcal{H}} \left\{ -\frac{1}{2}\int_\Omega \langle Ee(u), e(u)\rangle \, dx + F(u) \right\}. \tag{4}$$

Under our assumptions, the supremum in (4) is equal to $\frac{1}{2}F(u^*)$; this value is known as *compliance*. Now the role of the designer is to choose the material function E such that the "sup" in (4) becomes as small as possible, that is, the

body responds with minimal displacements in the direction of the load f. We assume $E(x)$ to be a symmetric and positive semidefinite matrix for almost all $x \in \Omega$ (recall $E \in L^\infty(\Omega)$), what we write as

$$E(x) = E(x)^T \succeq 0 \qquad \text{a.e. in } \Omega. \tag{5}$$

To introduce a resource (cost) constraint for E, we use the (invariant) *trace* of E

$$tr(E(x)) := \sum_{i=1}^{3} E_{ii}(x) \tag{6}$$

and require with some given positive α that

$$\int_\Omega tr(E(x))\, dx \le \alpha. \tag{7}$$

Further, to exclude singularities at isolated points (e.g., at boundary points of Γ_2) we demand that, with some fixed $0 < r^+ \in L^\infty(\Omega)$,

$$tr(E(x)) \le r^+(x) \qquad \text{a.e. on } \Omega. \tag{8}$$

The feasible design functions are collected in a set

$$\mathcal{E} := \left\{ E \in [L^\infty]^{3\times 3}(\Omega) \mid \begin{array}{l} E \text{ is of form (2) and} \\ \text{satisfies (5), (7) and (8)} \end{array} \right\}. \tag{9}$$

With this definition, the single-load problem becomes

$$\inf_{E\in\mathcal{E}} \sup_{u\in\mathcal{H}} \left\{ -\frac{1}{2} \int_\Omega \langle Ee(u), e(u)\rangle\, dx + F(u) \right\}. \tag{10}$$

Let us now assume that the structure must withstand a whole collection of independent loads $f^1, \ldots, f^L$ from $L^2(\Gamma_2)$, acting at different times; further, the design should be the "best possible" one in this framework. This leads to the following multiple-load design (MLD) problem, in which we seek the design function E which yields the smallest possible worst-case compliance

$$\inf_{E\in\mathcal{E}} \sup_{\ell=1,\ldots,L} \sup_{u\in\mathcal{H}} \left\{ -\frac{1}{2} \int_\Omega \langle Ee(u), e(u)\rangle\, dx + F^\ell(u) \right\}; \tag{11}$$

here

$$F^\ell(u) := \int_{\Gamma_2} f^\ell \cdot u\, dx \qquad \text{for } \ell = 1, \ldots, L. \tag{12}$$

2 Existence of a solution

We first eliminate the discrete character of the "$\sup_{\ell=1,\ldots,L}$" in (11). With a *weight vector* λ for the loads, which runs over the unit simplex

$$\Lambda := \left\{ \lambda \in \mathbb{R}^L \mid \sum_{\ell=1}^{L} \lambda_\ell = 1,\ \lambda_\ell \ge 0 \text{ for } \ell = 1, \ldots, L \right\},$$

we get from a standard LP-argument as reformulation of (11):

$$\inf_{E\in\mathcal{E}} \sup_{\lambda\in\Lambda} \sup_{(u^1,\dots,u^L)\in\mathcal{H}\times\cdots\times\mathcal{H}} \sum_{\ell=1}^{L} \left\{ -\frac{1}{2}\int_\Omega \lambda_\ell \langle Ee(u^\ell), e(u^\ell)\rangle\, dx + \lambda_\ell F^\ell(u^\ell) \right\}. \quad (13)$$

The objective function in (13) is linear (thus convex) in the inf-variable E; it is, however, not concave in the sup-argument $(\lambda; u^1, \dots, u^L)$. We will show that a simple change of variable yields a convex-concave version of the problem.

First note that the inf-sup value in (13) remains the same when restricting λ to the half-open set

$$\Lambda^0 := \{\lambda \in \Lambda \mid \lambda_\ell > 0 \text{ for } \ell = 1, \dots, L\}$$

and then pass from the variable $(\lambda; u^1, \dots, u^L)$ to

$$(\lambda; v^1 := \lambda_1 u^1, \dots, v^L := \lambda_L u^L).$$

This converts (13) to

$$\inf_{E\in\mathcal{E}} \sup_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \sum_{\ell=1}^{L} \left\{ -\frac{1}{2}\int_\Omega \lambda_\ell^{-1} \langle Ee(v^\ell), e(v^\ell)\rangle\, dx + F^\ell(v^\ell) \right\}, \quad (14)$$

where we put $\boldsymbol{v} := (v^1, \dots, v^L)$ and

$$\mathcal{V} := \left\{(\boldsymbol{v};\lambda) \mid \boldsymbol{v} \in [\mathcal{H}]^L, \lambda \in \Lambda^0\right\}.$$

The objective function in (14)

$$\mathcal{F}(E; (\boldsymbol{v};\lambda)) := \sum_{\ell=1}^{L} \left\{ -\frac{1}{2}\int_\Omega \lambda_\ell^{-1} \langle Ee(v^\ell), e(v^\ell)\rangle\, dx + F^\ell(v^\ell) \right\} \quad (15)$$

is now concave in $(\boldsymbol{v};\lambda) = (v^1, \dots, v^L; \lambda) \in \mathcal{V}$ and a result due to Moreau ([4]) yileds the following existence theorem.

THEOREM 1 *There exists $E^* \in \mathcal{E}$ such that*

$$\sup_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \mathcal{F}(E^*; (\boldsymbol{v};\lambda)) = \min_{E\in\mathcal{E}} \sup_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \mathcal{F}(E; (\boldsymbol{v};\lambda)).$$

Further

$$\inf_{E\in\mathcal{E}} \sup_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \mathcal{F}(E; (\boldsymbol{v};\lambda)) = \sup_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \inf_{E\in\mathcal{E}} \mathcal{F}(E; (\boldsymbol{v};\lambda)).$$

3 DISCRETIZATION AND SEMIDEFINITE REFORMULATION

Using the well-known identity for the trace of the product of a $d \times d$ matrix A and the rank-one matrix aa^T with $a \in \mathbb{R}^d$:

$$tr(A \cdot aa^T) = \langle Aa, a\rangle \quad (16)$$

we can rewrite the objective function (15) in (14) as

$$\mathcal{F}(E;(\boldsymbol{v};\lambda)) = -\frac{1}{2}\int_\Omega tr\left(E\cdot\sum_{\ell=1}^{L}\lambda_\ell^{-1}e(v^\ell)e(v^\ell)^T\right)dx + \sum_{\ell=1}^{L}F^\ell(v^\ell).$$

Due to Theorem 1, we may switch the order of "inf" and "sup" in (14); further, in order to simplify, let us multiply (14) by -2 to get

$$\inf_{(\boldsymbol{v};\lambda)\in U}\quad \sup_{E\in\mathcal{E}}\left\{\int_\Omega tr\left(E\cdot\sum_{\ell=1}^{L}\lambda_\ell^{-1}e(v^\ell)e(v^\ell)^T\right)dx - 2\sum_{\ell=1}^{L}F^\ell(v^\ell)\right\}. \tag{17}$$

For convenience, we will use the same symbols for the "discrete" objects (vectors) as for the "continuum" ones (functions). Assume that Ω is partitioned into M polygonal elements Ω_m of volume ω_m and let N be the number of nodes (vertices of the elements). We approximate E by a function that is constant on each element Ω_m, i.e., E becomes a vector $(E_1,\ldots,E_M)$ of 3×3 matrices E_m— the values of E on the elements. The feasible set $\mathcal{E}$ is replaced by its discrete counterpart

$$\mathcal{E} := \left\{E\in\mathbb{R}^{3\times 3M}\ \middle|\ \begin{array}{l} E_m = E_m^T \succeq 0 \text{ and } tr(E_m)\le r_m^+ \text{ for } m=1,\ldots,M \\ \sum_{m=1}^{M} tr(E_m)\omega_m \le \alpha \end{array}\right\}.$$

To avoid merely technical details we neglect in the following the constraint

$$tr(E_m)\le r_m^+ \quad \text{for } m=1,\ldots,M.$$

Further assume that the displacement vector u^ℓ corresponding to the load-case ℓ is approximated by a continuous function that is bi-linear (linear in each coordinate) on every element. Such a function can be written as

$$u^\ell(x) = \sum_{n=1}^{N} u_n^\ell \vartheta_n(x)$$

where u_n^ℓ is the value of u^ℓ at n^{th} node and ϑ_n is the basis function associated with this node (for details, see [3]). Recall that, at each node, the displacement has 2 components, hence $u\in\mathbb{R}^D$, $D\le 2N$ (D could be less than $2N$ because of boundary conditions which enforce the displacements of certain nodes to lie in given subspaces of $\mathbb{R}^2$).

For basis functions $\vartheta_n, n=1,\ldots,N$, we define matrices

$$B_n(x) = \begin{pmatrix} \frac{\partial\vartheta_n}{\partial x_1} & 0 \\ 0 & \frac{\partial\vartheta_n}{\partial x_2} \\ \frac{1}{2}\frac{\partial\vartheta_n}{\partial x_2} & \frac{1}{2}\frac{\partial\vartheta_n}{\partial x_1} \end{pmatrix}.$$

For an element Ω_m, let $\mathcal{D}_m$ be an index set of nodes belonging to this element. The value of the approximate strain tensor e on element Ω_m is then (we add the space variable x as a subscript to indicate that $e_x(u^\ell)$ is a function of x)

$$e_x(u^\ell) = \sum_{n \in \mathcal{D}_m} B_n(x) u_n^\ell \qquad \text{on } \Omega_m.$$

Finally, the linear functional $F^\ell(u^\ell)$ reduces to $(f^\ell)^T u^\ell$ with some $f^\ell \in \mathbb{R}^D$.

As discrete version of (17) we thus obtain, after a simple manipulation,

$$\inf_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \sup_{E\in\mathcal{E}} \left\{ \sum_{m=1}^{M} tr\left(E_m \cdot \sum_{\ell=1}^{L} \lambda_\ell^{-1} \int_{\Omega_m} e_x(v^\ell) e_x(v^\ell)^T \, dx \right) - 2 \sum_{\ell=1}^{L} F^\ell v^\ell \right\}. \quad (18)$$

Note that for each element Ω_m the $d \times d$ matrices $\displaystyle\int_{\Omega_m} e_x(v^\ell) e_x(v^\ell)^T \, dx$ can be computed explicitly using the Gaussian integration rule; namely, there exist points $x_{ms} \in \Omega_m$ and weights γ_{ms}^2 for $s = 1, \ldots, S$ such that

$$\int_{\Omega_m} e_x(v^\ell) e_x(v^\ell)^T \, dx = \omega_m \sum_{s=1}^{S} \gamma_{ms}^2 e_{x_{ms}}(v^\ell) e_{x_{ms}}(v^\ell)^T. \quad (19)$$

For instance, for linear $B_n(.)$ (i.e. bilinear ϑ_n) one takes $S = 4$. Hence (18) becomes

$$\inf_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \sup_{E\in\mathcal{E}} \left\{ \sum_{m=1}^{M} \omega_m tr(E_m A_m(\boldsymbol{v},\lambda)) - 2 \sum_{\ell=1}^{L} F^\ell v^\ell \right\} \quad (20)$$

where

$$A_m := A_m(\boldsymbol{v};\lambda) := \sum_{\ell=1}^{L} \lambda_\ell^{-1} \sum_{s=1}^{S} \gamma_{ms}^2 e_{x_{ms}}(v^\ell) e_{x_{ms}}(v^\ell)^T. \quad (21)$$

We now make one further step and introduce a dummy variable ρ_m for $tr(E_m)$ and $m = 1, \ldots, M$. Then the constraint $E \in \mathcal{E}$ in (20) splits into a *global* part (the global material distribution)

$$\rho \in \mathbb{R}_+^M, \quad \sum_{m=1}^{M} \rho_m \omega_m \le \alpha$$

and a *local* one (the local material properties)

$$E_m = E_m^T \succeq 0, \quad tr(E_m) = \rho_m, \qquad \text{for } m = 1, \ldots, M.$$

The "sup" over the local part can be now put under the sum:

$$\inf_{(\boldsymbol{v};\lambda)\in\mathcal{V}} \sup_{\substack{\rho\in\mathbb{R}_+^M \\ \sum \rho_m\omega_m \le \alpha}} \left\{ \sum_{m=1}^{M} \omega_m \sup_{\substack{E_m = E_m^T \succeq 0 \\ tr(E_m) = \rho_m}} tr(E_m \cdot A_m(\boldsymbol{v},\lambda)) - 2 \sum_{\ell=1}^{L} F^\ell v^\ell \right\}. \quad (22)$$

Now we will analytically perform the inner "sup", thus finally reaching a semidefinite programming formulation of the multiple-load problem.

Fix $m \in \{1, \ldots, M\}$ and consider the inner "sup" in (22):

$$\sup_{\substack{E_m = E_m^T \succeq 0 \\ tr(E_m) = \rho_m}} tr(E_m A_m). \tag{23}$$

We use Lagrange theory to write this as

$$\inf_{\tau \in \mathbb{R}} \{\tau \rho_m + \sup_{E_m = E_m^T \succeq 0} tr(E_m(A_m - \tau I_d))\} \tag{24}$$

with the $d \times d$ identity matrix I_d. The only τ for which the inenr "sup" is finite are those with $A_m - \tau I_d \not\succeq 0$. Hence we get for (24)

$$\sup_{\substack{E_m = E_m^T \succeq 0 \\ tr(E_m) = \rho_m}} tr(E_m A_m) = \rho_m \inf_{\tau I_d - A_m \succeq 0} \tau. \tag{25}$$

With

$$\tau_m := \inf_{\tau I_d - A_m \succeq 0} \tau$$

our discretized problem (22) becomes (note that A_m and thus τ_m depends on $(\boldsymbol{v}; \lambda)$)

$$\inf_{(\boldsymbol{v};\lambda) \in \mathcal{V}} \sup_{\substack{\rho \in \mathbb{R}_+^M \\ \sum \rho_m \omega_m \le \alpha}} \left\{ \sum_{m=1}^{M} \rho_m \omega_m \tau_m - 2 \sum_{\ell=1}^{L} F^\ell v^\ell \right\}.$$

The inner "sup" over ρ is a linear program for each fixed outer variable $(\boldsymbol{v}; \lambda)$. Hence the "sup" is attained at an extreme point of the feasible $\rho-$set and we can continue

$$\inf_{(\boldsymbol{v};\lambda) \in \mathcal{V}} \{ \max_{m=1,\ldots,M} \alpha \tau_m - 2 \sum_{\ell=1}^{L} F^\ell v^\ell \},$$

which in view of (3) is the same as

$$\begin{aligned} &\inf_{\substack{(\boldsymbol{v};\lambda) \in \mathcal{V} \\ \tau \in \mathbb{R}}} \alpha \tau - 2 \sum_{\ell=1}^{L} F^\ell v^\ell \\ \text{s.t.} \quad &\tau I_d - A_m(\boldsymbol{v}; \lambda) \succeq 0 \quad \text{for } m = 1, \ldots, M. \end{aligned} \tag{26}$$

To emphasize the dependence of A_m on $(\boldsymbol{v}; \lambda)$, we have again inserted the variables. With the $(d \times LS)$-matrix

$$Z_m := \left[\gamma_{m1} e_{x_{m1}}(v^1), \ldots, \gamma_{ms} e_{x_{ms}}(v^1), \ldots\ldots, \gamma_{m1} e_{x_{m1}}(v^L), \ldots, \gamma_{ms} e_{x_{ms}}(v^L)\right]$$

and the $(LS \times LS)$-matrix

$$\Lambda(\lambda) := diag(\lambda_1, \ldots, \lambda_1, \ldots\ldots, \lambda_L, \ldots, \lambda_L)$$

the constraints in (26) become

$$\tau I_d - Z_m(\boldsymbol{v})\Lambda(\lambda)^{-1} Z_m(\boldsymbol{v})^T \succeq 0$$

which, using a standard result on Shur complement, is equivalent to

$$\begin{pmatrix} \tau I_d & Z_m(\boldsymbol{v}) \\ Z_m(\boldsymbol{v})^T & \Lambda(\lambda) \end{pmatrix} \succeq 0.$$

We end up with the announced *semidefinite program* for the discretization of (14)

$$\begin{aligned} \inf_{\substack{(\boldsymbol{v};\lambda)\in\mathcal{V} \\ \tau\in\mathbb{R}}} \ & \alpha\tau - 2\sum_{\ell=1}^{L} F^\ell v^\ell \\ \text{s.t.} \ & \\ & \begin{pmatrix} \tau I_d & Z_m(\boldsymbol{v}) \\ Z_m(\boldsymbol{v})^T & \Lambda(\lambda) \end{pmatrix} \succeq 0 \qquad \text{for } m = 1, \ldots, M. \end{aligned} \tag{27}$$

The semidefinite program (27) can be efficiently solved by modern interior-point polynomial time methods. The question of recovering the optimal elasticity matrices $E_1^*, \ldots, E_M^*$ from the solution of (27) is a bit technical; again we refer the reader to [1].

4 EXAMPLES

Results of two numerical examples are presented in this section. The values of the "density" function ρ are depicted by gradations of grey: full black corresponds to high density, white to zero density (no material), etc.

Example 1. We consider a typical example of structural design: The two forces (or force and fixed boundary) are opposite to each other and there is a hole in between because of technological reasons. The geometry of domain Ω and the forces are depicted in Figure 1. The body can be loaded either by the forces on the left or on the right-hand side. Therefore this example has to be considered as MLD (two-load case). Symmetry allows us to compute only one half of the original domain. The resulting values of the "density" function ρ for 37×25 mesh are also presented in Figure 1. Again, the figure is composed from two computational domains to get the full body.

Example 2. In this example we try to model a wrench. The geometry of domain Ω is depicted in Figure 2. The nut (depicted in full black in Figure 2) is considered to present a rigid obstacle for the wrench. Hence the wrench is in unilateral contact with the nut and there are no other boundary conditions. The loads are also shown in Figure 2. Note that the problem is nonlinear because of the unilateral contact conditions and that for positive vertical force we get a different design than for a negative one; hence we have to consider these two forces as two independent loads. The resulting values of the "density" function ρ for 37×22 discretization are shown in Figure 3.

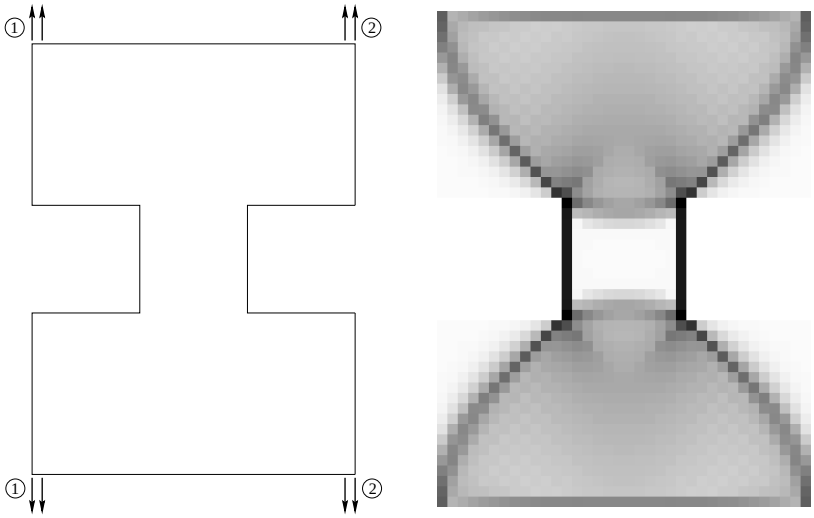


Figure 1: Example 1

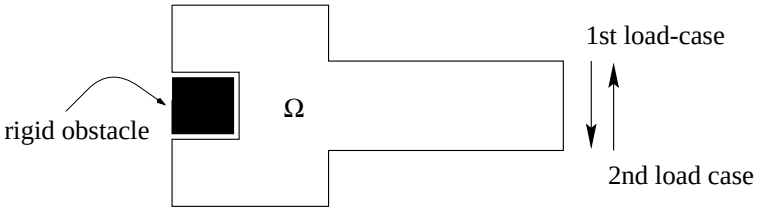


Figure 2: Example 2

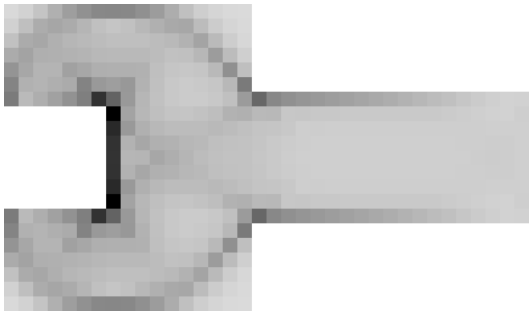


Figure 3: Example 2

ACKNOWLEDGEMENTS

The work on this paper was partly supported by GIF-contract I0455-214.06/95, BMBF-project 03ZO7BAY and grant A1075707 of the Czech Academy of Sciences. The first author is on leave from the Czech Academy of Sciences.

REFERENCES

[1] A. Ben-Tal, M. Kočvara, A. Nemirovski, and J. Zowe. Free material design via semidefinite programming. The multi-load case with contact conditions. Preprint 219, Inst. Appl. Math., Univ. of Erlangen, 1997.

[2] M. P. Bendsøe, J. M. Guades, S. Plaxton, and J. E. Taylor. Optimization of structure and material properties for solids composed of softening material. *Int. J. Solids Struct.*, 33:1179–1813, 1995.

[3] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam, New York, Oxford, 1978.

[4] J.-J. Moreau. Théorèmes "inf-sup". *C. R. Acad. Sc. Paris*, t.258, Groupe 1:2720–2722, 1964.

[5] J. Zowe, M. Kočvara, and M. Bendsøe. Free material optimization via mathematical programming. Preprint 213, Inst. Appl. Math., Univ. of Erlangen, 1997.

Michal Kočvara and Jochem Zowe
Institute of Applied Mathematics
University of Erlangen
Martensstr. 3
91058 Erlangen
Germany

Section 18

Teaching and Popularization of Mathematics

In case of several authors, Invited Speakers are marked with a *.

Mathematics Education: Reform or Renewal?

George E. Andrews

In considering the relationship between traditional methods of instruction and reform proposals, we should first take into account the environment in which instruction takes place.

It is a sad fact that students in the U.S. today are on the average: (1) not well prepared mathematically before entering college, (2) not studying very hard, and (3) distracted by many absorbing extracurricular activities.

The reports of unpreparedness are legion. From the recent TIMSS report and related sources, we find that primary and secondary mathematics education in the U.S. is not doing well. Indeed this problem has prompted the N.C.T.M. to produce (and then revise) a set of national standards for mathematics education at the primary and secondary levels. Even those of us who have criticized this project do recognize that it was undertaken in response to a real need.

Also there is much evidence that generally students do not study enough. The Pace Report suggests that a majority study less than 15 hours a week. A comparable study at my own university (Penn State) confirms this depressing statistic. No published report I know indicates that students are putting in close to the expected 30 hours a week that the old saying "two hours outside of class for each hour in class" suggests.

These factors must also be viewed against an even more disturbing aspect of undergraduate life in the U.S.: alcohol consumption. Graham Spanier (President of Penn State) is not someone with whom I always agree. However, he is making an effort to draw attention to this problem, and he paints a troubling picture:

"A survey conducted by the Harvard School of Public Health in 1993 reported nationally, 44 percent of all college students were binge drinkers, defined as consuming five or more drinks in a sitting for men and four or more drinks in a sitting for women during a two week period.

About half of these binge drinkers, or about one in five students overall, were frequent binge drinkers, drinking heavily three or more times in two weeks.

About two in five students drank without binging.

Only about one in six – 16 percent – were non-drinkers.

There are unmistakeable consequences of such behavioral patterns. Among the Harvard Study respondents, frequent binge drinkers were 25 times more likely than non-binge drinkers to report having had five or more problems such as doing something they regretted, missing a class, forgetting where they were, getting behind in school work,... and so on...

While only a fraction of one percent of the Harvard Study respondents considered themselves to be problem drinkers, 39 percent said they drink to get drunk."

Given these gargantuan problems, what then should we do for the improvement of undergraduate mathematics education? The answers provided by the bulk of the reform movement are:

INCREASED USE OF:

(1) calculators

(2) computer laboratories

(3) group projects

(4) term papers

DECREASED USE OF:

(1) lectures

(2) paper and pencil work and drill

At first glance, these answers seem inappropriate to say the least. We are faced with a large group of poorly prepared students with poor work habits who are prone to difficulties with alcohol, and we proceed to set up a system where "the students will learn by constructing calculus for themselves."

However, the current popularity of reform, indeed its great appeal to many administrators, may have more to do with certain unstated (perhaps unconscious) secondary effects than with its ability to improve mathematics education.

Whatever the virtues of each of the first four items (and we often hear much about their virtues), it is clear that each will serve to MASK THE PROBLEMS alluded to earlier.

If you can't multiply 8 times 7, if you don't know that $1/2 = .5$, if you can't divide 1000 by 10, the calculator is your salvation. If you've learned none of the small bits of information that serve to reinforce and accompany the development of mathematical maturity, have no fear; the calculator will come to your rescue. If some sorehead tells you that you need to know a few of these things so that you won't think 1000 divided by 10 is 10000 because you typed "*" when you meant "/", don't worry; just remember to type carefully.

The same can be said of computer laboratories, with the additional comment that these machines seem to obviate any necessity of gaining facility with the fundamentals of calculus. So what if I don't know the derivative of x^3; DERIVE tells me that it is $3x^2$.

Group projects will, of course, assist the ill-prepared and not-so-hard-working who will be able to do better than they otherwise might because they can be carried along by the stronger members of the group.

Term papers are partly a response to the frequent observation that students not only can't calculate, they also can't write. Again strugglers will have resources available from a variety of sources to help hide their deficiencies.

Since we are awash in enthusiasm for calculators, computer laboratories, group projects and term papers, it would be a serious mistake to ignore their quite obvious potential for abuse. This is important in a world where there is heavy pressure on administrators to cover over problems they can't solve directly. It is especially important when the public will buy the idea that "reform" and "innovation" can be expected to solve these problems eventually.

Does all this mean that all is right with the world of traditional instruction? No, unfortunately! While we did not create the environment I described earlier,

we traditionalists have behaved less than responsibly, or at least have not squarely faced our dilemmas.

Do we as faculty members sometimes shortchange our students because mathematical research is so much more fun than office hours or lecture preparation? You know the answer.

Do we put in front of our freshmen in the U.S. teaching assistants who are uncomfortable with English? Of course! We reply in response that we are hard pressed to avoid this because the pool of mathematically competent, native English speakers in the U.S. is small. Nonetheless, this is a serious problem that can only be mitigated if it is addressed directly.

Do we value research by an order of magnitude over teaching? Anyone who says "no" is not living in the real world. Think about the colleagues in your department who were courted by better universities. Was it their teaching that Princeton was after? Give me a break! What message does this absolutely certain fact of life send to everyone? Don't misunderstand me. The tremendous value we place on our research is positive overall; among other things it generally protects us from the relativism that has undermined standards and pedagogy in so many disciplines. However, although you and I may agree that excellence in research is the core of our profession, we have a duty to address the resulting implications for teaching. And teaching must be done well if we are to flourish.

Let me conclude with the words of Paul Halmos taken from his article "The Calculus Turmoil":

"Yes, there is a disease, but calculus is neither its cause nor its main symptom. We mathematicians can do our small bit to cure it, but not by rewriting calculus books. All that we can do, all that we are professionally able to do, is to insist on raising the quality of primary and secondary education by establishing and maintaining a high quality in college courses, by insisting on and strictly enforcing severe prerequisites, and by encouraging and properly training prospective grade school and high school teachers. That we can do, and I hope we will."

Halmos is clearly calling for renewal rather than reform, and I could not agree more.

George E. Andrews
Department of Mathematics
Pennsylvania State University
410 McAllister Building
University Park, PA 16802-6401
USA
andrews@math.psu.edu

De la Comprehension des Processus d'Apprentissage a la Conception de Processus d'Enseignement

Michèle Artigue

I. Introduction

La recherche en didactique des mathématiques, par ses travaux théoriques et expérimentaux nous donne aujourd'hui les moyens de mieux comprendre les processus d'apprentissage en mathématiques, des premiers apprentissages de l'école élémentaire à ceux en jeu à l'université, ainsi que de mesurer les effets des stratégies d'enseignement usuelles. La question de savoir comment tirer parti de cette connaissance pour améliorer l'enseignement reste cependant largement ouverte. A travers elle, se pose celle de l'utilité des recherches didactiques mais aussi celle des cadres théoriques adéquats pour penser le fonctionnement de systèmes complexes comme le sont les systèmes d'enseignement. C'est à ces questions que nous nous intéressons dans ce texte, en privilégiant au niveau des exemples l'enseignement des mathématiques à l'université ou à la transition lycée/université.

II. Qu'est-ce qu'apprendre des mathématiques ? Comment les apprend-t-on ? La diversité des positions épistémologiques

Toute réflexion sur l'apprentissage ou l'enseignement des mathématiques s'appuie, sur des présupposés épistémologiques, même si ces derniers restent souvent largement implicites. Ces présupposés peuvent être, comme le montrent différents travaux, relativement divers (Steiner, 1988). Ils jouent sans aucun doute un rôle restreint dans le travail quotidien des mathématiciens mais ils façonnent leur vision de cette science, de ce qui fait sa spécificité, de ses rapports avec les autres sciences mais aussi avec les pratiques sociales et donc, de ce fait, leur vision des valeurs à transmettre dans l'enseignement.

Cette diversité des présupposés épistémologiques possibles se retrouve, tout aussi grande, au niveau des théories de l'apprentissage. Les trente dernières années ont été marquées sur ce plan, dans le monde de l'éducation, par une domination nette des approches constructivistes, issues de l'épistémologie génétique piagétienne (Brun, 1996). Dans ces approches, l'apprentissage est conçu comme un processus d'adaptation individuel, basé sur des processus d'assimilation et d'accommodation conduisant à l'élaboration de schèmes. Il y a assimilation lorsque les nouvelles situations rencontrées peuvent être prises en charge par de simples adaptations des schèmes cognitifs déjà construits, accommodation lorsqu'un

déséquilibre cognitif important se fait jour, nécessitant une réorganisation structurelle. La théorie piagétienne ne se réduit bien sûr pas à cela mais ses autres dimensions n'ont pas eu en éducation mathématique une influence aussi durable.

Les approches constructivistes ont permis de porter un nouveau regard sur l'apprentissage, en montrant qu'il n'est pas le fruit d'un simple processus de transmission des savoirs. Elles ont permis de mieux appréhender la complexité des processus cognitifs dont il résulte et de mettre en évidence le rôle joué dans ces processus par les connaissances antérieures du sujet, par ses conceptions initiales. Elles sont cependant aujourd'hui de plus en plus considérées comme insuffisantes pour modéliser, de façon satisfaisante, les processus d'apprentissage en mathématiques car la dimension sociale et culturelle des apprentissages n'y est pas suffisamment bien prise en compte.

Parallèlement à ce qui se produit dans les travaux d'histoire et de philosophie des mathématiques, ce sont ces dernières dimensions que l'on essaie aujourd'hui de mieux prendre en charge dans les cadres théoriques élaborés. Comme le soulignent A. Sierpinska et S. Lerman (1996) dans leur revue de ces questions, ceci conduit à des constructions diverses qui se différencient notamment par la façon dont elles conçoivent les rapports entre l'individuel et le culturel dans l'apprentissage. Ainsi, dans les approches qualifiées de socio-culturelles, l'apprentissage est d'abord vu comme un processus d'enculturation. Il est social avant d'être intériorisé ; les médiations de pairs et d'adultes, comme les médiations instrumentales par des outils culturels y jouent un rôle fondamental. Les approches interactionistes, quant à elles, ne veulent réduire l'apprentissage, ni à un processus d'adaptation individuelle, ni à un processus d'enculturation dans une culture pré-établie. Ce qui y devient central, ce sont les interactions entre individus à l'intérieur d'une culture qui façonne ces interactions en même temps qu'elle est façonnée par elles. C'est à partir de ces interactions que se construisent, via des processus d'interprétation, les significations individuelles ; c'est le type de ces interactions qui conditionne les formes de connaissance accessibles (Cobb, Bauersfeld, 1995).

Comme le soulignent également A. Sierpinska et S. Lerman, la recherche didactique française a suivi depuis les années 70 un chemin analogue mais original. Elle se situe au départ dans le paradigme constructiviste mais la théorie des situations (Brousseau, 1997), qui en est un des piliers, met au centre de l'analyse, non le sujet apprenant mais les relations qu'il entretient avec le savoir mathématique, ses pairs et l'enseignant, au sein de la situation d'enseignement. C'est l'étude de ces relations qui permet de donner sens aux comportements observés et de les interpréter en termes d'apprentissage. L'apprentissage de l'élève y est de plus vu comme résultant d'un équilibre complexe, variable suivant les individus et les contextes, entre une adaptation " mathématique " et une adaptation aux attentes de l'enseignant et donc de l'institution, même si ces dernières restent largement implicites. Cette dimension institutionnelle de l'apprentissage est prise en compte de façon plus centrale dans l'approche anthropologique développée par Y. Chevallard (1990). L'apprentissage, selon lui, va résulter des rapports aux objets mathématiques que l'élève va nouer au sein des différentes institutions auxquelles il appartient. Il ne se constitue pas non plus indépendamment des rapports à d'autres objets : le rapport à l'Ecole, notamment. Le sujet est donc ici présent

avec ses motivations, ses affects, mais ses apprentissages sont contraints par des rapports institutionnels qui se constituent en normes, des normes qui peuvent être, pour un même objet mathématique, sensiblement différentes d'une institution à l'autre.

La diversité des constructions théoriques que nous venons d'évoquer n'exclut pas, fort heureusement, les points de consensus. Sans rentrer plus avant dans l'analyse comparée des différentes positions, nous voudrions en citer quelques uns :

1. L'apprentissage des mathématiques est un processus complexe dans lequel s'imbriquent étroitement l'individuel, le social et le culturel.

2. L'apprentissage des mathématiques n'est pas un processus " continu ". Il nécessite des reconstructions, réorganisations voire parfois de véritables ruptures avec des connaissances et des modes de pensée antérieurs.

3. L'apprentissage des mathématiques ne peut être conçu comme une simple progression vers des niveaux croissants d'abstraction. Il met en jeu de façon tout aussi essentielle la flexibilité du fonctionnement mathématique via notamment l'articulation de points de vue, de registres de représentation, de cadres de fonctionnement mathématique.

4. L'apprentissage des mathématiques est fortement dépendant des instruments matériels et symboliques du travail mathématique. Cette dépendance, qui concerne à la fois ce qui est appris et les modes d'apprentissage, est particulièrement importante à prendre en compte aujourd'hui du fait de l'évolution technologique.

Il s'agit là d'affirmations générales, tout comme le sont les différents cadres théoriques que nous avons évoqués jusqu'ici. Percevoir leur intérêt réel pour l'étude des processus d'apprentissage en mathématiques nécessite, nous semble-t-il, de s'interroger sur des apprentissages précis. C'est ce que nous ferons dans les deux paragraphes suivants, en privilégiant des domaines mathématiques qui posent des problèmes d'enseignement reconnus au niveau universitaire : l'analyse et l'algèbre linéaire. Cette particularisation nous servira également à souligner le rôle essentiel joué dans le travail didactique par l'analyse épistémologique des domaines mathématiques concernés. Vu les contraintes d'espace imposées à ce texte, nous avons choisi de centrer l'exposé sur deux des points évoqués ci-dessus, à savoir les points 2 et 3.

III-Reconstructions et ruptures dans l'apprentissage et l'enseignement mathématique

Si nous avons choisi ce point, c'est que l'enseignement, les recherches le montrent clairement, tend à sous-estimer l'importance de ces reconstructions et ruptures et les difficultés résistantes qu'elles posent à la majorité des élèves et étudiants lorsqu'elles sont laissées à leur seule responsabilité. Nous l'aborderons à partir d'exemples issus du champ de l'analyse élémentaire (" Calculus " dans la culture anglo-saxone). Il montre bien, nous semble-t-il, la nécessité de telles reconstructions et la diversité de leurs types possibles. Les reconstructions vont concerner tout d'abord des objets anciens qui existent pour les élèves avant que ne débute l'enseignement de l'analyse. C'est le cas par exemple pour la notion de tangente qui a été introduite dans un contexte géométrique et qui, pour entrer dans le champ de l'analyse, doit perdre certains de ses attributs et en gagner d'autres,

les représentations mentales associées devant se modifier en conséquence (Castela, 1995).

D'autres reconstructions vont s'avérer nécessaires parce que seules certaines facettes d'un concept seront présentes dans un premier contact mais aussi parce qu'il ne serait pas réaliste.de viser d'emblée les rapports les plus aboutis. H. Poincaré le soulignait déjà au début du siècle, dans une conférence sur les définitions en mathématiques (Poincaré, 1904). Y évoquant les notions de continuité, de dérivabilité et d'intégrabilité des fonctions, il rappelait combien l'intuition avait été trompeuse pour les mathématiciens et il insistait sur le fait que ces problèmes n'avaient pu être surmontés qu'en faisant primer la rigueur logique sur l'intuition. Mais un tel choix lui semblait catastrophique pour un débutant et il écrivait notamment :

" Nous voilà donc obligés de revenir en arrière ; sans doute est-il dur pour un maître d'enseigner ce qui ne le satisfait pas entièrement ; mais la satisfaction du maître n'est pas l'unique objet de l'enseignement ; on doit d'abord se préoccuper de ce qu'est l'esprit de l'élève et de ce qu'on veut qu'il devienne. "

a. Les reconstructions internes au champ de l'analyse : le cas de l'intégrale

Le cas de l'intégrale nous paraît bien illustrer cette situation. En France et dans de nombreux pays, l'intégrale est introduite dans l'enseignement secondaire via la notion de primitive, donc comme processus inverse de la dérivation, puis appliquée à des calculs simples d'aires et de volumes, en se basant sur un rapport intuitif à ces notions. Ce n'est qu'au niveau universitaire qu'est introduite une théorie de l'intégration, via l'intégrale de Riemann puis, à des niveaux plus avancés, la théorie de la mesure et l'intégrale de Lebesgue. Il y a là nécessairement en jeu des reconstructions successives et délicates du rapport à la notion d'intégrale. Les recherches que nous avons menées sur les procédures différentielles et intégrales ont montré les limites évidentes de l'enseignement usuel dans ce domaine (Alibert & al., 1989). Certes les étudiants atteignaient un niveau de performance raisonnable dans la résolution d'exercices mathématiques standard mais, ayant à décider, si telle ou telle situation relevait ou non d'une procédure intégrale, par exemple dans des problèmes de modélisation, ils se trouvaient complètement démunis, ne devant leur salut qu'aux indices linguistiques dont les présentations scolaires de ce genre de problème sont en général truffées (tranches, contributions élémentaires, découpages infinitésimaux...). Pire, un certain nombre, interrogés, n'hésitaient pas à déclarer que, dans ce domaine, le plus sûr était de s'abstenir d'essayer de comprendre et de fonctionner mécaniquement.

La situation que nous allons présenter, élaborée par M. Legrand dans le cadre de cette recherche, a été conçue pour faire face à ce problème, en faisant réellement vivre aux étudiants le besoin de la procédure intégrale. Le problème posé est le suivant : calculer l'intensité de la force d'attraction exercée par un barreau homogène de 6 mètres de longueur, pesant 18kg, sur une masse ponctuelle de 2kg située dans son prolongement, à 3 mètres de son extrémité. On rappelle au départ aux étudiants l'expression de la force d'attraction entre masses ponctuelles.

Exploitée régulièrement depuis plus de dix ans, cette situation a fait la preuve de son efficacité et de sa robustesse. Nous allons essayer d'en faire percevoir les raisons en en démontant les ressorts didactiques. Les étudiants de première année

d'université à qui elle est proposée ne reconnaissent pas d'emblée qu'il s'agit là d'un problème relevant d'une procédure intégrale. Ils n'en sont pas pour autant bloqués, notamment parce qu'ils disposent d'une stratégie pour attaquer le problème, inadaptée dans le cas présent mais souvent utilisée en physique : elle consiste à se ramener au cas de l'attraction entre masses ponctuelles en concentrant la masse de la barre en son centre de gravité. Dans les expérimentations, ce type de solution correspond toujours à un pourcentage important de réponses. Mais, dans un groupe de taille raisonnable, certains manifestent à coup sûr des doutes sur sa validité. Comment la tester ? Un des intérêts de cette situation réside dans le fait qu'un tel test est possible en appliquant la même méthode mais d'une autre façon : si elle est valide, elle doit le rester si l'on partage la barre en deux et si l'on applique le principe du centre de gravité séparément à chacun des morceaux. L'invalidation qui en résulte permet de mettre le doigt sur un facteur clef : la contribution d'un morceau du barreau dépend de sa distance à la masse ponctuelle et, à défaut de valeur précise, de proposer un encadrement de la valeur de la force cherchée. La technique à la base de l'invalidation peut alors être engagée dans un processus de raffinement successif du découpage qui aboutit à la conviction que la force, dont l'existence est physiquement assurée, peut être approchée d'aussi près qu'on le veut. Ce que l'on a mis en jeu n'est autre que le processus fondamental de la procédure intégrale. Il fonctionne ici comme " outil implicite " du travail mathématique, au sens développé par R. Douady (1984).

Dans le scénario didactique élaboré, les étudiants sont ensuite appelés à travailler sur des situations qui, dans des contextes divers, mettent en jeu ce même processus, puis à rechercher et expliciter les analogies existantes entre toutes ces situations pour aboutir aux caractéristiques de la procédure intégrale, en faisant ainsi un " outil explicite ". Ce n'est qu'à la suite de ce travail que tout ceci est mis en forme dans le cadre de la théorie de l'intégrale de Riemann et qu'un travail sur l'intégrale en tant qu'objet est développé. Les évaluations régulièrement faites attestent de l'efficacité du dispositif global.

Nous voudrions insister sur le fait que l'efficacité de la situation décrite ci-dessus n'est pas uniquement liée à ses caractéristiques mathématiques. Le scénario didactique construit pour organiser la rencontre des étudiants avec cette nouvelle facette de l'intégrale est tout aussi crucial. Ce scénario joue de façon essentielle sur le caractère social de l'apprentissage : c'est par les débats au sein du groupe (qui peut dépasser une centaine d'étudiants dans les expérimentations menées) que se régule la situation ; c'est le jeu collectif qui permet de dépasser la stratégie du centre de gravité pour aboutir à la procédure intégrale, dans un temps raisonnable, sans que l'enseignant n'apporte lui-même la solution ; c'est le jeu collectif qui force des régularités dans les déroulements qui seraient beaucoup moins assurées si les étudiants étaient confrontés individuellement à la même situation et en fait sa robustesse didactique. Il en va de même pour l'ensemble du processus d'enseignement construit.

b. Les reconstructions internes au champ de l'analyse : le cas du concept de limite

Le paysage que nous venons de décrire peut paraître idyllique. Il faut toutefois reconnaître que toutes les reconstructions nécessaires au fil de l'apprentissage de

l'analyse ne semblent pas aussi aisément gérables. Les différences sont par exemple sensibles si l'on s'intéresse au concept de limite. Dans beaucoup de pays aujourd'hui, une fois tirées les leçons de la période formaliste des mathématiques modernes, on a renoncé dans l'enseignement secondaire à fonder l'enseignement de l'analyse sur la notion formalisée de limite. Le premier contact avec ce domaine mathématique s'appuie sur des explorations graphiques et numériques aisément accessibles avec les calculatrices actuelles ; il est de l'ordre de l'empirique. On se contente d'une conception dynamique intuitive de la notion de limite, de techniques relevant d'une analyse algébrisée et de quelques théorèmes qui, une fois admis, permettent de gérer des problèmes simples de variation et d'optimisation. La transition vers une analyse formalisée, nécessite des reconstructions coûteuses, à la fois conceptuelles et techniques.

Sur le plan conceptuel, il y a là un saut qualitatif. En effet, ce qui est en jeu, épistémologiquement, à travers la formalisation du concept de limite, c'est avant tout la réponse à des besoins de fondements, de structuration du savoir. Un tel besoin est, les recherches l'attestent, difficile à faire ressentir aux étudiants via des situations analogues à celle décrite ci-dessus pour l'intégrale ; y être sensible nécessite déjà une culture mathématique certaine. C'est pourquoi des chercheurs comme A. Robert préconisent ici des stratégies didactiques spécifiques qui permettent de mieux prendre en compte cette dimension culturelle, en jouant sur des leviers métamathématiques (Robert et Robinet, 1996).

Mais la reconnaissance de ces difficultés de nature conceptuelle ne doit cependant pas conduire à sous-estimer les difficultés techniques de la reconstruction. Dans l'analyse algébrisée des premiers contacts, le travail technique continue à se situer dans la continuité des acquis algébriques. Le passage à une analyse formalisée suppose en particulier une reconstruction des rapports à l'égalité et des modes de raisonnement. L'égalité de deux objets ne résulte plus généralement d'équivalences successives, comme en algèbre, elle résulte d'une proximité à ϵ près, pour tout $\epsilon > 0$. La manipulation des inégalités prend d'ailleurs le pas sur celle des égalités. Parallèlement, aux raisonnements par équivalences successives basés sur la conservation d'égalités, se substituent des raisonnements par conditions suffisantes basés sur la perte contrôlée d'informations dans le traitement d'inégalités. Il y a donc là tout un monde technique nouveau qu'il faut identifier et apprendre à maîtriser. Dans le contexte de la massification de l'enseignement secondaire, une telle reconstruction revient sans aucun doute aujourd'hui à la charge de l'université, pour les filières où elle estime cette évolution de rapport nécessaire. Mais elle doit être alors pensée dans la durée, car elle s'y inscrit nécessairement, vu sa complexité.

Nous nous sommes dans cette partie exprimée en termes de " reconstruction ". Nous voudrions cependant souligner que certains chercheurs s'expriment plus nettement en termes de rupture, en se référant à la notion d'obstacle épistémologique empruntée au philosophe G. Bachelard (1938). C'est le cas par exemple dans divers travaux concernant la notion de limite et l'on pourra sur ce point se référer à la synthèse effectuée par B. Cornu (1991).

V. FLEXIBILITÉ, APPRENTISSAGE ET ENSEIGNEMENT

L'apprentissage mathématique est souvent perçu comme une spirale permettant

d'accéder à des niveaux d'abstraction croissants. C'est cette vision que nous voudrions relativiser ici, en mettant l'accent sur le rôle joué dans l'apprentissage par l'articulation flexible entre cadres, registres de représentation, points de vue et plus généralement entre formes de pensée mathématique (Dreyfus, Eisenberg, 1996). Ceci nous semble d'autant plus important que le développement de ces flexibilités est, pour l'instant, comme le montrent les recherches, mal pris en charge par l'enseignement usuel. Nous le ferons en privilégiant cette fois le domaine de l'algèbre linéaire et en nous appuyant plus particulièrement sur la synthèse des recherches didactiques dans ce domaine que présente l'ouvrage édité par J.L. Dorier (1997). Comme le souligne cet auteur, l'algèbre linéaire trouve sa source dans différents cadres mathématiques qu'elle a permis d'unifier : cadre géométrique, cadre des équations linéaires, en dimension finie et infinie... Le développement d'une articulation flexible entre ces différents cadres, comme entre chacun d'eux et celui de l'algèbre linéaire abstraite qui permet de les réorganiser conceptuellement, apparaît alors comme une composante essentielle de l'apprentissage dans ce domaine. Ce développement s'appuie sur des articulations entre modes de raisonnement, niveaux de langage et de descriptions, points de vue, registres de représentation dont l'apprentissage n'a rien d'évident. Dans l'ouvrage cité, J. Hillel par exemple analyse les différents langages ou niveaux de représentations à l'oeuvre en algèbre linéaire et leur interaction : le langage de la théorie générale, le langage de R^n et le langage géométrique ; A. Sierpinska, A. Defence, T. Khatcherian et L. Saldanha identifient, quant à eux, trois modes de raisonnement : synthétique-géométrique, analytique-arithmétique et analytique structurel. Tout en soulignant le rôle de l'interaction entre ces modes dans le développement de l'algèbre linéaire, ils montrent, par une étude fine de situations de tutorat à l'université, que l'enseignement, tant par les activités qu'il propose, que par les formats d'interaction enseignant-étudiant qu'il utilise favorise peu le développement d'une articulation souple et cohérente de ces trois modes. Dans ce qui suit, nous évoquerons rapidement, vu les contraintes d'espace, des flexibilités qui outillent en quelque sorte les flexibilités précédentes et auxquelles l'enseignement usuel est tout aussi peu sensible.

a. Flexibilité entre registres de représentations

Le travail en algèbre linéaire mobilise divers registres de représentations sémiotiques (graphiques, tableaux, écriture symbolique, langue naturelle...). Comme le souligne R. Duval (1996), les représentations sémiotiques sont absolument nécessaires à l'activité mathématique car ses objets ne sont pas directement accessibles à la perception. Pourtant l'enseignement tend selon lui à les réduire à un rôle d'extériorisation et de communication et à voir dans la capacité à reconnaître, former, traiter ou convertir dans un autre registre, des représentations sémiotiques, un simple sous-produit de la conceptualisation. La recherche de K. Pavlopoulou (1994) sur la coordination des registres de représentation en algèbre linéaire met bien en évidence que les rapports entre appréhension conceptuelle et appréhension sémiotique sont bien plus complexes. Le module d'enseignement expérimental qu'elle a mis en place pour des étudiants redoublants tend de plus à montrer que l'enseignement, lorsqu'il se veut sensible à la dimension sémiotique

du travail mathématique, peut permettre de surmonter des difficultés pourtant apparemment résistantes.

b. Flexibilité entre points de vue

Ce type de flexibilité intervient par exemple en algèbre linéaire dans les rapports entre points de vue cartésien et paramétrique, qui renvoient respectivement, à des caractérisations en termes de systèmes d'équations ou de systèmes de générateurs. Le travail en algèbre linéaire met en effet en jeu régulièrement le passage d'un point de vue à un autre, de façon explicite en dimension finie, de façon plus métaphorique ensuite. La thèse de M. Alves Dias (1998), menée avec à la fois des étudiants français et brésiliens de divers niveaux, met bien en évidence les difficultés résistantes rencontrées par les étudiants à développer une articulation efficace des deux points de vue. En témoignent par exemple les faibles pourcentages de réussite obtenus à un exercice aussi banal que le suivant :

> " On considère dans R^3 les vecteurs suivants : a=(2,3,-1) b=(1,-1,-2) c=(5,0,7) d=(0,0,1). Trouver une représentation cartésienne de l'intersection des sous-espaces vectoriels E et F engendrés respectivement par {a,b} et {c,d} ",

et les nombreux dérapages formels qu'il occasionne (confusion coordonnées/paramètres conduisant à des intersections dans R^2 ou R^4, association brutale d'équations à des vecteurs...), les anticipations et contrôles dans le cadre géométrique ou dans celui des sytèmes linéaires n'étant visiblement ici d'aucun secours pour les étudiants concernés.

Mais ce que montre également cette recherche, à travers l'analyse de manuels représentatifs de l'enseignement dans les deux pays, c'est la très faible sensibilité à ces difficultés que semble manifester l'enseignement. Certes les étudiants disposent, via les techniques de résolution des systèmes linéaires, des moyens de gérer techniquement l'articulation des points de vue, mais ceci ne suffit pas visiblement à leur permettre de lui donner sens, à leur permettre de la gérer et contrôler de façon efficace. La dualité, lorsqu'elle est introduite, devrait leur permettre de repenser cette articulation et de mieux percevoir le rôle qu'y joue l'association vecteur / équation. Mais les deux mondes restent, pour la plupart des étudiants, des mondes trop distants que l'enseignement ne leur donne pas les moyens de connecter de façon efficace.

Cette faible prise en charge institutionnelle de l'articulation ne nous semble pas un cas isolé. Elle semble considérée comme allant de soi, une fois que l'on a " compris " la notion, comme s'il s'agissait d'une pure question d'intendance que l'on pouvait laisser au travail privé de l'étudiant. Les recherches montrent que ce n'est malheureusement pas le cas. La flexibilité n'est pas pour autant hors de portée, si l'on est attentif à son développement. Les travaux déjà cités tendent à le montrer en ce qui concerne l'algèbre linéaire. C'est aussi le cas si nous revenons au champ de l'analyse. De nombreux travaux dans ce domaine montrent que les technologies informatiques, si leur utilisation est soigneusement pensée, peuvent jouer un rôle décisif dans le développement d'articulations flexibles ainsi que dans une rééquilibration des rapports entre registre algébrique et graphique, faisant de ce dernier un instrument réellement efficace de l'activité mathématique (Tall,

1996). Nos propres recherches sur l'enseignement des équations différentielles vont dans le même sens, tout en montrant combien le changement de statut du registre graphique nécessaire à son opérationnalisation s'oppose aux rapports institutionnels dominants et est, de ce fait, difficile à négocier (Artigue, 1992).

VI. Des connaissances à leur exploitation : quelques éléments de réflexion

Nous avons essayé, dans ce qui précède, de montrer sur deux points très précis, des types d'apports que pouvaient fournir les travaux didactiques. Ceci ne permet bien sûr qu'une vision très partielle de la façon dont les questions d'apprentissage sont abordées dans le champ didactique et des résultats auxquels elles ont permis d'arriver. Mais ils nous serviront à revenir dans ce dernier paragraphe sur la question des rapports possibles entre les connaissances acquises et l'action sur le système d'enseignement. Comme nous l'avons souligné dans l'introduction, il ne s'agit pas là d'une question facile.

Les résultats des recherches en didactique nous aident indubitablement à mieux comprendre comment fonctionnent les élèves et étudiants, à identifier les difficultés qui jalonnent l'apprentissage, les raisons de résistances constatées à nos efforts d'enseignants, à analyser les liens que peuvent avoir certaines de ces difficultés avec les stratégies d'enseignement dominantes. Ils nous aident aussi, plus globalement, à comprendre les modes de fonctionnement et les dysfonctionnements des systèmes d'enseignement, à mettre en évidence, à ce niveau aussi, des régularités intéressantes.

La connaissance de ces difficultés, de ces dysfonctionnements ne fournit pas pour autant directement les moyens de les surmonter. Certes les travaux de recherche, ne se bornent pas à effectuer des constats et diagnostics ; dans de nombreux cas, ils ont conduit au développement de produits d'enseignement qui ont été expérimentés et évalués. Mais si l'on considère ces produits, ils ne nous permettent que rarement de penser que, par de minimes adaptations de notre enseignement, nous pourrions obtenir des gains substantiels. En général, au contraire, ils requièrent de la part des enseignants un engagement plus lourd que l'engagement standard et des changements substantiels de pratiques. Car ce qui est à réorganiser, ce n'est pas seulement le contenu de l'enseignement, c'est globalement l'ensemble des formes de travail de l'étudiant, pour lui permettre de rencontrer ces contenus de façon satisfaisante, pour lui permettre d'apprendre. C'est sans doute là le prix à payer pour trouver aux systèmes didactiques dans lesquels nous vivons de meilleurs équilibres de fonctionnement, en particulier dans le contexte actuel de massification de l'enseignement universitaire, mais montre clairement que la réussite de l'action dépend de facteurs et contraintes qui échappent au contrôle de la recherche.

A ceci s'ajoutent sans aucun doute, en particulier au niveau de l'enseignement supérieur, des difficultés spécifiques liées à la complexité des connaissances en jeu. Les apprentissages que nous avons évoqués dans ce texte, qu'il s'agisse de l'analyse ou de l'algèbre linéaire, sont des apprentissages qui s'articulent nécessairement avec de nombreux apprentissages antérieurs, des apprentissages qui ne peuvent être pensés et organisés que dans le long terme. La définition même à ce niveau de processus d'enseignement (en prenant donc en charge non seulement l'organisation

des contenus mathématiques mais aussi leur gestion didactique) et leur évaluation posent des problèmes qui restent aujourd'hui largement ouverts.

Enfin, il faut reconnaître la complexité des systèmes dans lesquels s'inscrit l'apprentissage et l'enseignement des mathématiques. Les connaissances sur le fonctionnement de ces systèmes que nous pouvons inférer des recherches sont bien trop partielles pour permettre d'en contrôler un fonctionnement qui restera nécessairement fortement indéterminé. Ceci marque bien la nécessaire limite d'actions, même fondées sur la recherche, la prudence nécessaire dans les essais de généralisations de dispositifs mis au point dans des conditions expérimentales particulières, l'importance de prévoir des systèmes de régulation de l'action qui permettent de pallier les limites de nos capacités de prédiction. Les idées, mêmes épistémologiquement et cognitivement les plus séduisantes, ne conduisent pas nécessairement à des stratégies pédagogiques viables, dans un enseignement de masse comme celui que nous connaissons aujourd'hui, dans un monde marqué par les incertitudes sociales. C'est ce que nous avons essayé de montrer en analysant l'évolution récente de l'enseignement secondaire de l'analyse en France (Artigue, 1996), c'est sans aucun doute aussi valable pour les enseignements universitaires. Mais, qu'elles qu'en soient les limites, chaque progrès dans la connaissance que nous avons du fonctionnement de cette complexité est précieux, il nous arme pour la comprendre et la piloter, en nous adaptant à des conditions sans cesses changeantes. Ces connaissances méritent d'être capitalisées. Elles le seront nous semble-t-il d'autant plus efficacement que les cadres théoriques qui nous serviront à les organiser ne réduiront pas trop drastiquement la complexité mais prendront en compte l'enseignement, l'apprentissage et leurs rapports, de façon équilibrée, dans leurs composantes non seulement cognitives et épistémologiques mais aussi culturelles et sociales.

RÉFÉRENCES :

Alibert & al., 1989. *Procédures différentielles dans les enseignements de mathématiques et de physique au niveau du premier cycle universitaire*, IREM Paris 7

Artigue M., 1992. Functions from an algebraic and graphic point of view : cognitive difficulties and teaching practices, in E.Dubinski & G.Harel (eds), *The concept of function : some aspects of epistemology and pedagogy,* MAA Notes n°25, 109-132.

Artigue M., 1996. Learning and teaching elementary analysis, *Proceedings of ICMI 8*, Sevilla (à paraître).

Alves Dias M., 1998. *Les problèmes d'articulation entre points de vue "cartésien" et "paramétrique" dans l'enseignement de l'algèbre linéaire*, Thèse, Université Paris 7.

Bachelard G., 1938. *La formation de l'esprit scientifique*, J. Vrin, Paris.

Brousseau G., 1997. *The theory of didactical situations*, Kluwer Academic Publishers, Dordrecht.

Brun J., 1996. Evolution des rapports entre la psychologie du développement cognitif et la didactique des mathématiques, in J.Brun (ed), *Didactique des Mathématiques*, 19-43, Delachaux et Niestlé, Lausanne.

Castela C., 1995. Apprendre avec et contre ses connaissances antérieures, *Recherches en Didactique des Mathématiques*, n°15.1, 7 – 47.
Chevallard Y., 1990. *La transposition didactique* (2ème édition), La Pensée Sauvage, Grenoble.
Cobb P., Bauersfeld H. (eds), 1995. *The emergence of mathematical meaning : interaction in classroom cultures*, Lawrence Erlbaum Associates Publishers, Hillsdale, NJ.
Cornu B., 1991. Limits, in D.Tall (ed), *Advanced Mathematical Thinking*, 153-166, Kluwer Academic Publishers, Dordrecht.
Douady R.,1984. *Dialectique outil-objet et jeux de cadre*, Thèse, Université Paris 7.
Dorier J.L. (ed), 1997. *L'enseignement de l'algèbre linéaire en question*, La Pensée Sauvage, Grenoble.
Dreyfus T., Eisenberg T., 1996. On different facets of mathematical thinking, in Sternberg & Benzeev (eds), *The Nature of Mathematical Thinking*, London : Reidel.
Duval R., 1996. *Semiosis et pensée humaine*, Peter Lang, Berne.
Pavlopoulou K., 1994. *Propédeutique de l'algèbre linéaire : la coordination de registres de représentation sémiotique*, Thèse, Université de Strasbourg I.
Poincaré H., 1904. Les définitions en mathématiques, *L'enseignement des Mathématiques*, n°6, 255 – 283.
Robert A., Robinet J., 1996, La prise en compte du meta en didactique des mathématiques, *Recherches en Didactique des Mathématiques*, n°16.2, 145 – 175.
Sierpinska A., Lerman S., 1996. Epistemologies of mathematics and mathematics education, in A.J.Bishop & al. (eds), *International Handbook of Mathematics Education,* 827-876, Kluwer Academic Publishers, Dordrecht.
Steiner H.G., 1988. Relations between historico-epistemological studies and research in mathematics education, in L.Bazzini & H.Steiner (eds), *Proceedings of the first Italian-German bilateral symposium on didactics of mathematics,* 25-35.
Tall D., 1996. Functions and calculus, in A.J.Bishop & al. (eds), *International Handbook of Mathematics Education,* 289-325, Kluwer Academic Publishers, Dordrecht.

Michèle Artigue
IUFM de Reims et Equipe DIDIREM
Université Paris 7
Case 7018, 2 place jussieu
75251 Paris Cedex 05, France
artigue@gauss.jussieu.fr

Drawing Instruments: Theories and Practices from History to Didactics

Maria G. Bartolini Bussi

Abstract. Linkages and other drawing instruments constitute one of the most effective fields of experience at secondary and university level to approach the theoretical dimension of mathematics. The main thesis of this paper is the following: By exploring, with suitable tasks and under the teacher's guidance, the field of experience of linkages and other drawing instruments, secondary and university students can 1) relive the making of theories in a paradigmatic case of the historical phenomenology of geometry; 2) generate 'new' (for the learners) pieces of mathematical knowledge by taking active part in the production of statements and the construction of proofs in a reference theory 3) assimilate strategies for exploration and representative tools (such as metaphors, gestures, drawings, and arguments) that nurture the creative process of statement production and proof construction. This thesis will be defended by referring to research studies already published or in progress.

1991 Mathematics Subject Classification: MSC 97A15 98A15

Aknowledgements Several individuals and institutions have contributed to the preparation of this paper. Annalisa Martinez, Marcello Pergola and Carla Zanoli have built dozens of beautiful instruments and have used them for years in secondary classrooms : I owe thanks to them all for my introduction to the beautiful world of drawing instruments and for their continuous friendly cooperation. The collection of instruments is now stored in the Laboratorio di Matematica of the Museo Universitario di Storia Naturale e della Strumentazione Scientifica di Modena (http://www.museo.unimo.it/labmat/). In the website there are a lot of photos of the instruments quoted in this paper. My friends and colleagues Ferdinando Arzarello, Paolo Boero and M. Alessandra Mariotti have read and commented on earlier versions of this paper. Jonathan A. Hillman has improved the English of this paper. CNR has funded my research on this topic until 1997. The University of Modena has given financial support in the years 1995-1997. The Department of Pure and Applied Mathematics of the University of Modena has funded my travel to Berlin. The collective project concerning the approach to theoretical thinking and theorems is now a part of the national project 'Ricerche di Matematica e Informatica per la Didattica', coordinated by F. Arzarello.

1. Introduction.

In recent years several efforts have been made at the international level to clarify the objects, the aims, the research questions, the methodologies, the findings and the criteria to evaluate the results of research in didactics of mathematics (or mathematics education, according to the name preferred in some countries). I may quote the volume edited for the 20 years of work at the IDM, Bielefeld University,

and Professor Hans-Georg Steiner' s 65th birthday [BSSW] ; the ICMI Study held in 1994 in Washington DC about 'What is Research in Mathematics Education and what are its Results' [KS]; the Working Group 25 in ICME 8 [Mal]; the International Handbook edited by Bishop [Bi]. Didactics of mathematics as a scientific discipline is fairly young compared to other sciences, yet is deeply rooted in the perennial effort of mathematicians to advance human understanding of mathematics and to transmit mathematics knowledge to future generations. It has become clear that analytical tools are needed from different disciplines (such as epistemology, history, psychology) to obtain results that can increase the knowledge of the teaching and learning processes in the classroom, produce effective innovation in schools and understand why some designed innovation works or does not work, and, at a larger level, influence the development of school systems.

Analytical tools from history and epistemology are necessary to tackle one issue which is perhaps crucial: the nature of mathematics knowledge. One of the distinctive features of mathematics is theoretical organisation. This has created a very specific mathematician's style, with a very impressive form, that alternates definitions and theorems. Yet, when a mathematician reads a theorem and, in particular, its proof, it is not the form that commands most attention, but rather the process by means of which mathematical ideas have been generated or have been illuminated by the proof in a new way. If we look at the 'confessions' of working mathematicians [T], we have an idea of a continuous (not always individual) process: the major discontinuity seems to happen in the final phase of written communication in Journals, where the leading ideas, the intuitions, the associations, the metaphors or the explorations of special cases are hidden by the formidable and conventional mathematician's style. Unfortunately the curriculum revolution of the sixties gave too much importance to the product (i. e. the form) and put in shadow the process (i. e. the construction of reasoning and arguments). But it was realised soon that teaching beginners the formalities of proof might be very difficult (and, perhaps, meaningless). Instead of scrutinising the reason for failure, what happens now is that, in some countries, proving processes are being eliminated from mathematics curriculum, not taking into account that giving up proofs for a sheer acquisition of isolated facts and notions hides the theoretical organisation of mathematics (for a detailed discussion of these issues see [Ha]).

This is the scenario in which a collective project has been set some years ago by a group of Italian researchers [MBBFG], [AMORP1]. The project highlights the permanent value of proof in mathematics and didactics of mathematics and aims to design, implement and analyse effective teaching experiments, that can introduce students to the theoretical dimension of mathematical culture up to the construction of theorems and proofs. As far as the activity of mathematicians is concerned, from a didactic perspective, we are much more interested in the hidden process of conjecture production and proof construction than in the final product: this very process does offer suggestions on the way of organising effective classroom activity. In particular, whenever the process of producing conjectures about something may evolve continuously and smoothly into the process of constructing proofs, the task of producing 'new' theorems is proved to be easier for students. In confirmation of that, we may recall a typical strategy, used by good teachers.

When a difficult and crucial theorem is introduced in the standard lecture format, before giving the proof, the students are presented with examples, counterexamples and reasons for the plausibility of the statement to make them relive the intellectual experience of the prior inventor of the theorem, although they have been deprived of the long process of generating the conjecture by themselves.

The issue of continuity between the production of conjectures and the construction of proofs has been raised from a cognitive perspective in a study carried out in the 8th grade [GBLM], concerning the production of a theorem of geometry about a problem situation in the field of sunshadows. The authors have described the cognitive continuity as a process with the following characteristics. During the production of the conjecture, the student progressively works his/her statement through an intense argumentative activity; during the subsequent statement proving stage, the student links up with this process in a coherent way, organising some of the justifications ('arguments') produced during the construction of the statement according to a logical chain. The construct of cognitive continuity, further developed by Arzarello & al. [AMORP1] to include also the case of advanced learners has proved to be useful to interpret existing teaching experiments and to design new ones.

In recent years several experiments in different fields have been carried out at very different school level, from primary to tertiary education (primary school : [B2], [BBFG]; middle school: [BGM], [GBLM], [BPR1], [BPR2]; secondary school: [B1], [BP], [Mar], [AMORP2] [MB]; tertiary: [AMORP1]). Some characteristics are shared by nearly all the experiments: 1) the selection, on the basis of historic-epistemological analysis, of fields of experience, rich in concrete and semantically pregnant referents (e. g. perspective drawing; sunshadows; Cabri-constructions; gears; linkages and drawing instruments); 2) the design of tasks, which require the students to take part in the whole process of production of conjectures, of construction of proofs and of generation of theoretical organisation; 3) the use of a variety of classroom organisation (e. g. individual problem solving, small group work, classroom discussion orchestrated by the teacher, lectures); 4) the explicit introduction of primary sources from the history of mathematics into the classroom at any school level.

In my own research, I have found that linkages and other drawing instruments might be one of the most effective fields of experience at secondary and university level. In the following I shall give some details on this case, by analysing the activities designed and implemented for approaching mathematical theorems and more generally the theoretical organisation of mathematics.

2. Linkages and Drawing Instruments: An Historical Digression.

In this section, I shall outline the history of linkages and other drawing instruments by using the metaphor of a theatre play. Only planar drawing instruments will be considered; however spatial drawing instruments such as perspectographs have also played a relevant role in specific practices (e. g. painting, architecture) and have given rise to specific theories (such as projective geometry). But this is another story and, maybe, the topic of a different paper (examples in http://www.museo.unimo.it/labmat/)

2. 1. THE PROLOGUE : EUCLID AND THE CLASSICAL AGE. Drawing instruments have been considered in geometry treatises from the time of Euclid, whose first postulates implicitly define the kind of instruments that are allowed for geometrical constructions [He] : '1) Let the following be postulated : to draw a straight line from any point to any point; 2) To produce a finite straight line continuously in a straight line ; 3) To describe a circle with any centre and distance.'

Even if the description is supposed to recall a practical use of instruments, there is no doubt that the intention is theoretical. Actually the instruments are never quoted directly, not even in the large number of constructions that are discussed in the following books. Moreover, the problem is never to find the approximate solution that could be useful for applications: rather a theoretical solution by straight lines and circles is looked for. Other drawing instruments (and curves) were known at the time of Euclid, yet not included in the set of accepted theoretical tools (e. g. the conchoid of Nicomedes, [He]). They were rather used to solve practical problems. For instance, by means of the conchoid it is possible to find two mean proportionals between two straight lines and, hence, to construct a cube which is in any given ratio to a given cube. This allows to find a set of weights in given proportion to calibrate catapults.

2.2. THE FIRST ACT : DESCARTES AND SEVENTEEN CENTURY GEOMETERS. Descartes, like most scientists of his age, was deeply involved in the study of mechanisms for either practical or theoretical purposes. A famous example of the former kind (i. e. the machine to cut hyperbolic lenses) is described in the 'Dioptrique'. The latter issue forms the core of the 'Géométrie', where two methods of representing curves are clearly stated: the representation by a continuous motion and the representation by an equation [Bos]. Descartes deals with the following question: 'Which are the curved lines that can be accepted in geometry? (p. 315)' and gives an answer (or, better, two answers) different from the one of classical geometers : 1) '[...] we can imagine them as described by a continuous motion, or by several motions following each other, the last of which are completely regulated by those which precede. For in this way one can always have an exact knowledge of their measure (Géométrie p. 316)'; 2) '[...] those which admit some precise and exact measure, necessarily have some relation to all points of a straight line, which can be expressed by some equation, the same equation for all points (Géométrie, p. 319)' The goal of Descartes was related to the very foundations of geometry: if a curve (e. g. a conic or a conchoid) is to be accepted as a tool to solve geometrical problems, one must be sure that, under certain conditions, the intersection points of two such curves exist. Hence, pointwise generation is not sufficient and the continuum problem is called into play: by the standards of the seventeenth century mathematicians, it is solved by referring to one of the most primitive intuitions about the continuum, i. e. the movement of an object. Descartes did not confront the question whether the two given criteria - i. e. the mechanical and the algebraic - are equivalent or not. This problem actually requires constructing more advanced algebraic tools and, what is more important, changing the status of drawing instruments from tools for solving geometric problems to objects of a theory. The importance of the generation of curves by movement is proved by the flourish of innumerable treatises of 'organic' geometry (i. e. geometry developed

by instruments), thanks to leading mathematicians, such as Cavalieri , L'Hospital, Newton, or van Schooten. They designed and studied dozens of different drawing instruments for algebraic curves (incidentally, in the same age when the very concept of algebraic curve started to be worked out).

2.3. The Second Act : Kempe and the Nineteenth Century Geometers. In the nineteenth century there was a shift from studying individual drawing instruments to developing a theory of drawing instruments, in the special case of linkages. On the one side, geometers started to study which curves could be drawn by any n-bar linkage; on the other side they asked which linkages could be used to draw any curve. The curve that resisted longest the attack of geometers was the simplest one, i. e. the straight line. After the approximate 3-bar solution offered by Watt in 1784 (that is still used in nearly every beam-engine), only in 1864 Peaucellier presented a 7-bar linkage, that embodies a rigorous solution based on the properties of circular inversion [K2]. The general problem of drawing any algebraic curve of any degree was temporarily solved by Kempe, a few years later (1876), with the paper entitled 'On a General Method of Describing Plane Curves of the nth Degree by Linkwork' [K1]. The structure of Kempe's proof is quite interesting. Starting from the equation F(x,y)=0 of any plane algebraic curve and from a particular point P of the curve, the polynomial is expanded into a linear combination of cosines of suitable angles. For each element of the sum, an elementary linkage is provided. By combining such linkages, a new linkwork is obtained, that has the effect of 'drawing' the given curve in the neighbourhood of P. Rather than an actual linkwork, the theorem gives an algorithm to construct a (virtual) linkwork, that depends on the equation of the curve.

2.4. The Third Act : Modern Revival of Curve Drawing Devices. The study of linkages is reconsidered in today's mathematics from two different, yet related, perspectives. The problem of drawing curves is reread as the problem of forcing a point of a robot to execute a given trajectory [Ba], [HJW]. The study of abstract linkages and their realisation is related to the study of algebraic varieties and of immersed submanifolds of Euclidean space [GN], [KM]. According to Kapovich & Millson, a major role in the revival of this field of research has been played by Thurston, who has given lectures on this topic since the late seventies. The new theory is completely algebraized and, at a first glance, has nothing to share with the problems that have been described in the previous acts. Yet, the very theorem of Kempe, combined with the work of today's mathematicians, has lead to proving general realizability theorems for vector-valued polynomial mappings, real-algebraic sets and compact smooth manifolds by moduli spaces of planar linkages. Kempe's proof has been carefully scrutinised, revealing some weakness related for instance to the presence of some 'degenerate' configurations of linkages appearing during the movement. However, the structure of the proof, based on the recourse to elementary linkages as building blocks, is still the original one. Hence Kempe's theorem might be considered an hinge: on the one side it closes Descartes' implicit problem to relate motion of instruments and equations and on the other side it opens the way to the modern theory of abstract linkages.

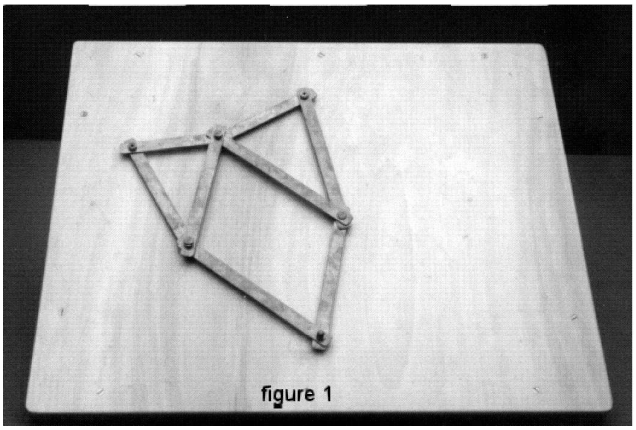
2.5. The History Goes On: The Critical Impact of Drawing Instruments. The historical analysis sketched in the previous 'acts' suggests that in the

domain of geometry the relationship between theoretical and practical issues has always been very rich and complex. On the one side, drawing instruments, intimately connected with the development of algebraic tools, are theoretical products of the continuous modelling effort, that aims at rationalising the perception and the production of shapes. On the other side, drawing instruments are physical objects of the world to be modelled: to understand their functioning means to be able to design instruments which fulfil a desired action. Theories and practices might have been developed for some time independently, but in each age they happen to nurture each other: a double arrow describes the dialectic relationship between them, that is constructed anew repeatedly with shifts of meaning. In the teaching of mathematics such general complex ideas are to be translated into ordered activities for the classroom. If the ideas are interconnected as in a loop, as in this case, an apparently obvious solution is supposed to be to cut the loop somewhere, so that the double arrow becomes a single arrow from theories to practices (i. e. practices are applications of theories) or from practices to theories (i. e. practices are motivations for theories). These are the most common options. It is far beyond the scope of this paper to discuss them in detail. I intend to defend a different option: to put drawing instruments in the centre and to use them as mediators for both theories and practices. This idea is not new: drawing instruments were part of the education of gentlemen in arts such as the military art or the art of navigating since the 17th century [Tu]; they were used in prestigious Institutes of Mathematics (e. g. Goettingen [Mu]) to educate generations of leading mathematicians; drawing instruments are even on show in Scientific Museums for the popularisation of mathematics. In each of these uses, the visibility of theoretical aspects is surely different, because, when concrete referents come into play, the risk is always that the attention is captured by isolated facts and that the argument, if any, is not detached from everyday styles of reasoning [S]. For instance, the very possibility of making 'infinitely many' experiments by dynamic exploration might help, on the one side, the production of conjectures, but, on the other side, might render things self-evident and destroy the need of constructing proofs. If the theoretical aspects of mathematics are central in didactics of mathematics, as we have argued in the introduction, a careful didactic treatment of concrete referents is always needed. Whether an object is considered from a practical or from a theoretical perspective depends on the habits of the students, acquired through a slow process, on the types of exploration tasks and on the issues raised by the teacher in the classroom interaction. This is true for drawing instruments too, for both the material copies and the virtual copies of ancient instruments produced by computer (such as the simulations produced by means of software with graphic interface - such as Cabri - or by means of Java) and for the computer itself considered as the most flexible drawing instrument. In this part of the study, the function of analytical tools from the psychology of mathematics education appears to be relevant.

3. Drawing Instruments in the Classroom.

3. 1. Exploring Linkages. This example concerns the study of one of the pantographs (i. e. the pantograph of Sylvester), which were designed in the 19th century to realise elementary geometric transformations and to give the elemen-

tary blocks of Kempe's theorem. The study was originally carried out with 11th graders [B1],[BP], but we have collected later items of anedoctal evidence that confirm the emergence of similar processes (with the same slowness) when similar tasks are given to undergraduate students, to graduate students or to teachers of mathematics or mathematics educators. Hence, what follows is supposed to apply to both novices and expert explorers. The pantograph of Sylvester is an 8-bar linkage (see the figures).



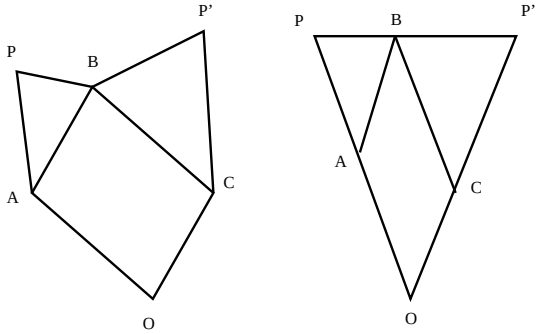


figure 2a figure 2b

figure 2

The students had already been given an introductory lecture concerning the early history of drawing instruments in Euclid's age. They were given a specimen of the pantograph and a set of eight tasks to guide the exploration in small group work. Two tasks are especially relevant for our discussion [B1]: '1) Represent the linkage with a schematic figure and describe it to somebody who has to build a similar one on the basis of your description alone; 2) Are there any geometric properties that are related to all the configurations of the linkage? State a conjecture and try to prove your statement'.

The first task aimed at encouraging students' manipulation of the linkage. It

has to be said that the tradition of abstract and symbolic work has often the effect of inhibiting recourse to manipulation in mathematics lessons. In this case, on the contrary, the students had to measure bars and angles and to try to connect these empirical data with the pieces of geometrical knowledge that were part of their past experience. Actually the small group debated for a long time whether the imaginary addressee had to build an 'equal' linkage (i. e. with the same measure) or a 'similar' one (i. e. capable for working in the same way). In the first case, it would have been enough to write down the length of each bar and to give the instructions for assembling the linkage. When they decided for the second solution, they had to cope with the problem of identifying the structural features of the linkage (i. e. the presence of a parallelogram and of two similar isosceles triangles) from the empirical evidence offered by perception and by measuring. The process of solving the second task resulted in three interlaced phases: (1) producing the conjecture; (2) arguing about the conjecture; (3) constructing a proof.

Producing the conjecture was difficult and slow. The linkage actually realises a rotation as for every configuration, a) OP=OP'; b) POP' = PAB = BCP'. Yet the rotation is approached at as a correspondence between two points that has no transparent relationships with the motion of the linkage. The teacher had a helping attitude, but the whole exploring process was carried out by the students, who at the end agreed with the proposal of one of them, who had 'seen' suddenly the invariant during the exploration. The suggestion was checked experimentally in different configurations and then accepted by the whole group.

Arguing about the conjecture and constructing the proof were actually interlaced processes. The students were helped by the large amount of exploration they had made before. For instance the observation of an intermediate limit case (figure 2b when two sides of the parallelogram and two sides of the triangles are aligned) was considered empirical evidence that the triangles POP', PAB and BCP' are similar. While trying to defend the conjecture by arguments, the students mixed continuously experimental data (obtained by direct manipulation of the mechanism) and statements deduced logically from already accepted statements. Whilst the verbal proof was eventually complete, the process of polishing the entire reasoning in order to give it the form of a logical chain and to write it down was slow and not complete, as the students' text shows:

'Thesis: POP' is constant (see figure 2 for notation).

The angle POP' is constant as the triangles POP' obtained by means of the deformations of the mechanism are always similar, whatever the position of P and P'. In fact OP=OP', because the triangles OCP' and OAP are congruent, as CP'=OA,CO=AP and OCP'=OAP (BCO=OAB and P'CB=BAP). The above triangles are also similar to a third triangle PBP', because, as the triangles BCP' and BAP are similar, it follows that BP' : BP = CP' : CO and the angle P'BP = OCP' as (setting CP'B = CBP' = a and CBA = b) we have PBP' = 360 - (2a+b); OCP' = 360 - (2a+b).

This is true because prolonging the line BC from the side of C the angle supplementary to BCP' is equal to 2a and the angle supplementary to BCO is equal to b as two contiguous angles of a parallelogram are always supplementary'.

Surely this written text in neither complete nor well ordered, according to the mathematician's style: the order of the steps recalls the sequence of production of statements, as observed during the small group work, rather than the logical chain that could have been used by an expert. Nevertheless it was easily transformed later with the teacher's help into the accepted format with reference to elementary euclidean geometry; yet, what is important, the time given to laboriously produce their own proof ensured that the final product in the mathematician's style, where the genesis of the proof was eventually hidden, retained meaning for the students.

3.2. Theoretical Framing of Drawing Instruments and Linkages. The small group study was only one step in a long term teaching experiment. The study was done in the frame of Euclid's elementary geometry. From a cultural perspective, students must be introduced to the different theories which have been invented later, with their own goals and objects, and to the different practices which have been developed, from beam-engines to robotics, otherwise we would have relapsed into the standard linear teaching path from concrete referents to a geometrical study framed by Euclid's geometry, where practices are only the starting point, i. e. motivations for theories.

The students who had realised the study of the pantograph, together with their schoolfellows who had studied other pantographs according to the same tasks, took part in lessons where each small group presented the results of the guided study. The teacher related the different pantographs to each other, generating an embryo of a theory of linkages, where the same proof could be applied, with small adaptations, to different instruments [BP]. The shifts in meaning from considering an individual linkage to developing a theory of linkages was introduced by means of guided reading of some historical sources, like the ones quoted in the theatre play of the section 2; historical sources were assimilated by students, producing explorations and proofs according to the inquiry style of each age.

This is only a prototype of teaching experiments which are made every year with secondary school students (by Marcello Pergola) and with university students (by the author). The difference between secondary school and university students concerns the length of the play: the second act is within the reach of secondary school students, whilst university students can understand the whole play.

3.3. Some Issues to be Deepened. In the above sections, a complex teaching experiment has been outlined. Different classroom organisations have been shown with different roles for the teacher: lectures, small group works, whole class discussions. In small group work phases of joint activity between the teacher and the students were accomplished. In the theoretical framing the teacher acted, by his own words or by quoting historical sources, as a cultural mediator. The study of the teacher's role is a crucial problem of didactics of mathematics, whose discussion is far beyond the scope of this paper: it is related to the possibility of reproducing the teaching experiments in different classrooms. For a partial account about this issue, the interested reader could refer to [MB] for the analysis of the teacher's role in a classroom discussion when the object is the theoretical meaning of geometric construction. Further investigations are planned.

In the theoretical framing episode students coped with a cultural problem, i. e. the construction of a balanced image of mathematics, where theories and

practices are strictly intertwined yet not confused. In the direct and guided manipulation of instruments, students experienced, at an appropriate slow pace, the continuous and smooth transition from physical experience (gestures and manipulation) to the production of their own conjectures and to the construction of a proof. In this process, they used different linguistic tools to express their ideas, from the metaphors taken from everyday language to the fixation of the procedures according to the speech genre of elementary geometry. The study of student processes is a crucial problem of the psychology of mathematics education. Finer grain analyses are an unavoidable part of each of the research studies quoted in the introduction and of ongoing research.

4. Some Implications for Teaching.

The case of linkages and other drawing instruments gives only one among several examples of teaching experiments about the theoretical organisation of mathematics and the approach to theorems. Systematic experiments in this field have been carried out mainly at secondary and university levels, but the activity with drawing instruments has proven to be effective with younger students too, because the difference between a practical and a theoretical use of instruments might be approached (yet is seldom emphasised) also in primary school. For instance, in an experiment carried out in primary school [BBFG], pupils have become aware that they can use a compass in two very different ways: 1) to imitate a round shape (practical use); 2) to construct (if possible) a triangle with sides of given length (theoretical use). In the former case the focus is on a careful use of compass that assures the precision of the drawing. In the latter case the focus is on the definition of the circle: even a free-hand rough sketch could be effective as the compass is meant as a mental instrument.

What implications for curricula could the quoted experiments have? To give an answer, we can contrast our approach to geometry with the traditional one in a very special case: the case of conics. When this topic is considered, it is usually introduced according to some standard steps: 1) A short introduction, concerning the space generation of conics as conic sections, limited to explaining the origin of the name. 2) A metric definitions of conics as loci determined by the focal properties; in this case a particular drawing instrument for obtaining the so-called gardener ellipse is described. 3) The canonical equations; then every problem is considered in this analytic setting. From a cultural perspective, this path conveys a one-sided image of mathematics, i. e. the physical generation of conics (as conic sections or as drawings by instruments) is nothing but a rough introduction to the very important things, that are, on the contrary, metric definitions and equations. What is even more disappointing is the cognitive counterpart: by this approach (even if it is completed by a careful study of quadratic forms, as in the case of university students of mathematics), students do nor learn how to relate their spatial intuitions (on which heuristics might be based) with the plane synthetic or analytic study [BM].

In this paper I have proposed an alternative approach with two different, yet related, arguments. The cultural argument: for centuries curves have been considered as trajectories determined by linkages and other drawing instruments; only later, the mechanical study has been complemented by the algebraic study, arous-

ing theories which retain the links with the spatial referents and which has proven to be relevant for the development of today mathematics The cognitive argument: the very manipulation of drawing instruments provides students with heuristics and representative tools (such as metaphors, gestures, drawings and arguments) that foster the production of conjectures and the construction of related proofs within a reference theory, with a slow and laborious process that recalls the one of professional mathematicians. Reliving the making of theories and producing one's own theorems is a way to appreciate and assimilate the theoretical dimension of mathematics.

References

[AMORP1] Arzarello F., Michelotti C., Olivero F., Robutti O. & Paola D., in press, A Model for Analysing the Transition to Formal Proofs in Geometry, Proc. 22 PME, Stellenbosch (South Africa), 1998.

[AMORP2] Arzarello F., Michelotti C., Olivero F., Robutti O. & Paola D., in press, Dragging in Cabri and Modalities of Transition from Conjectures to Proofs in Geometry, Proc. 22 PME, Stellenbosch (South Africa), 1998.

[B1] Bartolini Bussi M., Geometrical Proofs and Mathematical Machines: An Exploratory Study, Proc. 17 PME , 2 (97–104), Tsukuba (Japan): 1993.

[B2] Bartolini Bussi M., Mathematical Discussion and Perspective Drawing in Primary School, Educ. Stu. in Maths, 31 (1–2), 11–41: 1996.

[Ba] Baker D. R., Some Topological Problems in Robotics, The Math. Intell., 12 (1), 66–76: 1990.

[BBFG] Bartolini Bussi M., Boni M., Ferri F. and Garuti R., in press, Early Approach to Theoretical Thinking: Gears in Primary School, Educ. Stu. in Maths.

[BGM] Boero P., Garuti R. & Mariotti M. A., Some Dynamic Mental Processes Underlying Producing and Proving Conjectures, in Proc. 20 PME, 2 (121–128), Valencia (Spain): 1996.

[Bi] Bishop A. (ed.), International Handbook for Mathematics Education, Kluwer Publ.: 1997.

[BM] Bartolini Bussi M. & Mariotti M. A.., in press, Which is the Shape of an Ellipse? A Cognitive Analysis of an Historical Debate, Proc. 22 PME, Stellenbosch (South Africa), 1998.

[Bos] Bos H. J. M., On the Representation of Curves in Descartes' Géométrie, Arch. Hist. Ex. Sci., 24, 295–338: 1991.

[BP] Bartolini Bussi M. & Pergola M., History in the Mathematics Classroom: Linkages and Kinematic Geometry, in Jahnke H. N., Knoche N. & Otte M. (hrsg.), Geschichte der Mathematik in der Lehre, 39–67, Vandenhoeck & Ruprecht: 1996.

[BPR1] Boero P., Pedemonte B. & Robotti E., Approaching Theoretical Knowledge through Voices and Echoes: a Vygotskian Perspective, Proc. 21 PME, Lahti (Finland), 2 (81–88): 1997.

[BPR2] Boero P., Pedemonte B. & Robotti E., in press, The 'Voices and Echoes Game': and the Interiorisation of Crucial Aspects of the Theoretical Knowledgein a Vygotskian Perspective, Proc. 22 PME, Stellenbosch (South Africa): 1998.

[BSSW] Biehler R. , Scholz R. W., Straesser R. & Winkelmann B., Didactics of

Mathematics as a Scientific Discipline, Kluwer Publishers: 1994.
[GBLM] Garuti R., Boero P., Lemut E., Mariotti M. A.., Challenging the Traditional School Approach to Theorems, Proc. 20 PME, 2 (113–120), Valencia (Spain): 1996.
[GN] Gibson C. G. & Newstead P. E., On the Geometry of the Planar 4–Bar Mechanisms, Acta Appl. Math. 7, 113–135: 1986.
[Ha] Hanna G., More than Formal Proof, For the Learn. of Maths, 9 (1), 20–23: 1989.
[He] Heath T. L. (ed.), Euclid: The Thirteen Books of the Elements, Cambridge Univ. Press: 1908.
[HJW] Hopcroft J., Joseph D. & Whitesides S., Movement Problems for 2-dimensional Linkages, SIAM J. Comp. 13 (3), 610–629: 1984.
[K1] Kempe A. B., On a General Method of Describing Plane Curves of the nth Degree by Linkwork, Proc. London Math. Soc., 7 (102), 213–216: 1876.
[K2] Kempe A. B., How to Draw a Straight Line?, Macmillan & C.: 1877(reprinted by NCTM: 1977).
[KM] Kapovich M. & Millson J. J., preprint, Universality Theorems for Configuration Spaces of Planar Linkages: 1998.
[KS] Kilpatrick J. & Sierpinska A. (eds.), Mathematics Education as a Research Domain: A Search for Identity, Kluwer Publ.: 1998.
[Mal] Malara N. A. (ed.), An International View on Didactics of Mathematics as a Scientific Discipline, University of Modena: 1997.
[Mar] Mariotti M. A., Justifying and Proving: Figural and Conceptual Aspects, Proc. ERCME, Praha: 1997.
[MB] Mariotti M. A. & Bartolini Bussi M., in press, From Drawing to Construction: Teacher's Mediation within the Cabri Environment, Proc. 22 PME, Stellenbosch (South Africa): 1998.
[MBBFG] Mariotti M. A., Bartolini Bussi M., Boero P., Ferri F. & Garuti R., Approaching Geometry Theorems in Contexts: From History and Epistemology to Cognition, Proc. 21 PME ,1 (180–195), Lahti (Finland): 1997.
[Mu] Muehlhausen E., Riemann Surface-Crocheted in Four Colors, The Math. Intell., 15 (3), 49–53: 1993.
[S] Sierpinska A., Mathematics: 'In Context', 'Pure' or 'with Applications'?, For the Lern. of Maths, 15 (1), 2–15: 1995.
[T] Thurston W., On Proof and Progress in Mathematics, Bull. Amer. Math. Soc., 30 (2), 161–177: 1994.
[Tu] Turner A. J., Mathematical Instruments and the Education of Gentlemen, Annals of Science, 30 (1), 51–88: 1973.

Maria G. Bartolini Bussi
Dipartimento di Matematica
Università di Modena
via G. Campi 213/B
I 41100 Modena - Italia
bartolini@unimo.it

Difficulties in the Passage from Secondary to Tertiary Education

Miguel de Guzmán, Bernard R. Hodgson,
Aline Robert and Vinicio Villani

Abstract. For an important part of those students who take mathematics courses at the tertiary level, the transition from secondary to tertiary education presents major difficulties. This is true whether the students are specializing in mathematics or are registered in a program for which mathematics is a service subject. The purpose of this paper is to identify some relevant difficulties related to this passage and to examine possible causes. Such a study can be done from a broad spectrum of perspectives which will be commented briefly: epistemological and cognitive, sociological and cultural, didactical. We also consider actions which could help to improve the situation.

1991 Mathematics Subject Classification: 00
Keywords and Phrases: Mathematics education, tertiary education, secondary-tertiary transition, teaching and learning contexts.

The passage from secondary to tertiary mathematics education is determined by procedures varying considerably from one country to the other, and even within one country, from one institution to another. But whatever the context, this transition often presents major difficulties for an important part of those students who take mathematics courses at the tertiary level. This is true whether the students being considered are specializing in mathematics or are registered in a program for which mathematics is a service subject.

The problem of the transition to the post-secondary level in by no means a new issue in mathematics education. For instance, the very first volume in the Unesco series *New Trends in Mathematics Teaching* includes a report from a conference devoted to this problem (see [18]). This same topic was also discussed in various settings at ICME congresses — see for instance the paper by Cross [5] presented at ICME-4, as well as the report [13] of Action Group 5 at ICME-6. But still today the secondary–tertiary transition can be seen as a major stumbling block in the teaching of mathematics.

This paper, prepared in connection with a round-table discussion at ICM'98, is concerned with various groups of students taking mathematics courses at the university level: students of science (vg, mathematics, physics, chemistry), engineering, economics, preservice secondary school teacher education, etc. After

presenting an overview of the respective points of view of the student and of the teacher on the passage from secondary to tertiary education in mathematics, we shall consider three different types of difficulties then encountered by students: epistemological/cognitive; sociological/cultural; didactical. We conclude the paper with some possible actions, both from an institutional and from a pedagogical perspective, which could help to improve the conditions under which the transition takes place.

1 The point of view of the student

In order to better assess the perception that students may have of the transition from secondary to tertiary mathematics, a questionnaire was recently given to various first-year groups in our respective universities, asking students for their opinion about three possible types of sources for the difficulties they might have encountered with university mathematics: (i) difficulties linked to the way teachers present mathematics at the university level and to the organization of the classroom; (ii) difficulties coming from changes in the mathematical ways of thinking at the higher level; and (iii) difficulties arising from the lack of appropriate tools to learn mathematics. Students were asked to express their degree of agreement with various statements on a five-item Likert scale (from 1/total disagreement/ to 5/total agreement/, 3 being a neutral point). Here are some of the outcomes of this informal survey.

First, it should be stressed that the perception of students can vary considerably according to the type of mathematics they are taking and the program of study to which they belong. This is the case for instance for the results obtained at Université Laval when students were asked for their overall perception of how they went themselves through the secondary-tertiary transition. From a cohort of 250 students, 91 (36%) were in partial or total agreement (items 4 and 5 on the Likert scale) with the statement *"Transition to university mathematics was difficult for me"*, while 127 (51%) expressed disagreement (items 1 and 2 on the Likert scale). However if the results are considered according to the program of study of the students, the picture gets quite diversified. In the following table, we compare three different groups of students from Université Laval, namely[1]

- Group I: *students specializing in mathematics* (first-year and final-year);
- Group II: *preservice secondary school mathematics teachers* (first-year);
- Group III: *engineering students* (first-year).

[1] It should be noted that transition to university education in Québec typically happens as students are age 19, after a two-year intermediate level following secondary school (the so-called "cégep" level). Students entering university are divided into groups, already in their first year, according to their specific domain (mathematics, physics, engineering, etc.).

"Transition to university mathematics was difficult for me."								
Likert scale	Group I		Group II		Group III		Totals	
1	7	(12%)	3	(4%)	35	(30%)	45	(18%)
2	17	(28%)	19	(26%)	46	(39%)	82	(33%)
3	14	(23%)	5	(7%)	11	(9%)	30	(12%)
4	17	(28%)	37	(51%)	15	(13%)	69	(28%)
5	5	(8%)	8	(11%)	9	(8%)	22	(9%)
no reply	0	(0%)	0	(0%)	2	(2%)	2	(1%)
Totals	60	(100%)	72	(100%)	118	(100%)	250	(100%)

One notes that 22 out of 60 mathematics students (37%) agree that the transition was difficult for them (items 4 and 5 on the Likert scale), as opposed to 45 out of 72 (63%) for the preservice teachers and only 24 out of 118 (20%) in the case of engineering students. Analogous differences can be seen when considering other items of the scale.

The three groups often reacted also quite differently to more targeted questions. For instance,

- more than 85% of the students of mathematics and 75% of the preservice secondary school teachers from Université Laval see assessment at the university level as bearing upon more abstract mathematics than previously (the replies from these two cohorts give a mean of 4,2 on a scale of 5), as opposed to only 38% of the engineering students (mean of 3,0);

- similarly, more than 55% of the non-engineers see the mathematics problems they have to solve at the university level as substantially more difficult than at the secondary level (mean of 3,5), which is the case for only 28% of the engineering students (mean of 2,7).

In fact, a clear outcome of the data from Université Laval is that the transition to university mathematics appears much smoother for engineering students than for preservice secondary school teachers or for students of the mathematics program. (The same questionnaire was used with first-year and final-year students of the undergraduate mathematics program; although the answers were not identical, the variations observed appear far less significant than when comparing with future engineers or secondary school teachers.)

The questionnaire was used in France (Université de Versailles and Université de Montpellier) but gave rise to a somewhat different response from the students: much less diversity was observed in the patterns of answers than at Université Laval. Possibly this results from the fact that the French education system attracts a majority of the best students in special classes ("classes préparatoires") leading to the "grandes écoles", so that university-bound students form a rather homogeneous group. It is interesting to note that, from a cohort of 190 university students in the first year of a scientific program, more than 70% are in partial or total agreement with the following statements:

- *I am not used to proofs and abstract developments*;
- *I would prefer to have a textbook, as in secondary school*;

and more than 66% agree with the statements:

- *it is not always clear what is expected of me regarding what is seen in the classroom*;
- *we are not indicated what is essential and what is accessory*;
- *teachers are too abstract, they don't care to present concrete examples*;

and, more surprisingly,

- *there is not enough time spent in the classroom.*

Two groups of students from the Universidad Complutense de Madrid were given the questionnaire, namely some 70 students from the first three years at the Facultad de Matemáticas and 100 first-year and fifth-year students from the Facultad de Ciencias Económicas. The answers collected are quite similar to those from France, the statements receiving a degree of agreement greater than 3,5 on the Likert scale being those which deal with such aspects as: the high level of abstraction, the use of proofs in the mathematical development, the lecturing style (the fast pace, the ignorance of where it is heading), the abstract nature of part of what is being assessed on exams, the need for textbooks.

The questionnaire invited students to write comments on their own. Needless to say, the spectrum of opinions expressed is extremely wide, but it is interesting to consider a few comments made spontaneously by students of Université Laval. Some are quite severe on the university teachers:

- *Many university teachers do not care whether we understand or not what they are teaching us.*
- *A majority of teachers do not understand that we do not understand.*
- *It is hard for them to make us understand what is evident for them.*
- *Passing from secondary school to university mathematics was not as hard as I was told. But what makes it somewhat hard are the changes in the teachers: many of them are not at all suited for teaching. Here, we have teachers who are topnotch mathematicians. But their pedagogical skills will never outmatch those of my high school teachers.*

Other comments have to do with the background of the students or the autonomy expected of them:

- *It seems that I am lacking a lot of prerequisites. It is as if I should know 100% of my high school maths.*
- *In high school, I never learned to do proofs, and now it seems to be taken for granted that we know how to do proofs.*

- *My answers to these questions would vary considerably according to the courses and the instructors. But a general trend is that courses include many many topics which are covered very quickly, so that we need to work a lot on our own outside the classroom.*

But quite a few students did express a positive opinion about their encounter with university mathematics, reflecting the fact that the transition gives no or little problems to a number of students:

- *I appreciate much more university math, because we try to understand where the results we are using, and were using in high school, come from.*

- *Going from high school to university did not raise special problems for me, as the level of difficulty of high school math prepared us well for that.*

2 Perceptions of the university teacher

When seen from the point of view of the university teacher, the transition from secondary to tertiary mathematics is considered to be problematic for a majority of students. Such is the observation we have made from an informal survey of a small number of teachers regularly involved in the teaching of first-year university mathematics. We think that this survey, however limited, still provides us with a good idea of the perceptions of a lot of our university colleagues.

Those involved in the teaching of first-year university mathematics are often rather dissatisfied with the weaknesses they perceive in their students. Many have the feeling that students are not interested in the mathematics itself covered in a course, but only in succeeding at the exams — this might be especially the case in contexts where mathematics is used as a sieve for accessing to other professional fields, for instance for admission to the medicine or law school. University teachers deplore the lack of prerequisite knowledge which makes the beginning at the tertiary level painful and difficult for many of their students; even the contents indicated in the secondary syllabi (where there is such a thing common to secondary school students) cannot be taken as understood and mastered. They also deplore the learning style of students, many of whom have concentrated in the past on the acquisition of computational skills (often, it must be said, so to meet the requirements of university entrance examinations). They lament over the thinking and working habits of their students in mathematics, their lack of organization and of mathematical rigour, as well as their difficulty in acquiring and consolidating knowledge through personal work.

Acquisition of a certain level of autonomy in learning is often seen by university teachers as the main stumbling block in the secondary-tertiary passage. Zucker [27] has expressed as follows the idea that significant individual activity outside the mathematics class becomes an absolute necessity when moving to the higher level: "The fundamental problem is that most of our current high school graduates don't know how to *learn* or even what it means to learn (a fortiori to understand) something. In effect, they graduate high school feeling that learning must come down to them from their teachers. [...] *That the students must also*

learn on their own, outside the classroom, is the main feature that distinguishes college from high school."

3 Types of difficulties in the secondary-tertiary transition

The nature of the difficulties related to the passage from secondary to post-secondary mathematics and the reasons for their occurrences can be seen from a broad spectrum of perspectives.

3.1 Epistemological and cognitive difficulties

As shown by many research works, an important conceptual leap takes place, with respect to the mathematical contents taught and asked practices, when passing from the secondary to the tertiary level. This transition corresponds to a significant shift in the kind of mathematics to be mastered by students: the mathematics is different not only because the topics are different, but more to the point because of an increased depth, both with respect to the technical abilities needed to manipulate the new objects and to the conceptual understanding underlying them. This shift has sometimes been described as corresponding to a move from *elementary* to *advanced mathematical thinking* (see Tall [23]) : secondary school students often succeed in mathematics by relying on their ability to perform algorithms and in spite of a lack of a real understanding of the mathematical concepts with which they are working; they may then experience substantial difficulties, when moving to the tertiary level, in being able to participate by themselves in the process of mathematical thinking, and not merely learn to reproduce mathematical information. In a word, they may have problems in becoming autonomous, mathematically speaking. Moreover, it is no more possible to limit themselves to put isolated theorems in practice, they need to enter into deeper and richer thought processes.

A word of caution is in order here: in a given course, the needs of the students, as perceived by them, are dictated mainly by the exams. In a context where assessment is not congruent with the intended level of the course, being in fact lower, then it would be totally possible for the students, should this be known to them, to succeed in the course without entering into more advanced mathematical thinking. Success in such a context would by no means testify to adequate learning. What we have in mind here is a system where the gap between the level of the course and that of assessment is not too important.

In many countries, the passage to tertiary mathematics coincide with the introduction of new abstract notions such as vector spaces or formalized limits. This is a difficult step because these notions are not in the strict continuity of what students already know (even though vector spaces take their origin in the "spaces of (physical) vectors" $\mathbf{R}^2$ or $\mathbf{R}^3$). We can speak of these notions as "unifying and generalizing concepts", in the following sense (see [7]): such concepts unify and generalize different methods, tools and objects existing previously in a variety of settings; they are formal concepts which unify the various objects from which they have been abstracted. They have not necessarily been created to solve new

problems, but to make the solution of many problems easier or more similar to each other. Moreover, these concepts represent a change of perspective which induces a sophisticated change of level in mental operations.

Other concepts are acquiring a different status, when passing from one level of education to another. For instance, the equality of elementary arithmetic becomes in high school algebra a notion of identity. And in analysis equality incorporates the complex idea of "local infinite proximity", i.e., of an arbitrarily good approximation, such as in the expression $\lim_{x\to a} f(x) = L$. The manipulation of equalities in such a context rests in an essential way on inequalities, which fact contributes to the difficulties linked to the passage from algebraic to analytic thinking (see Artigue [1]).

Students entering tertiary education are facing, in the words of Tall [24, p. 495], "a difficult transition, from a position where concepts have an intuitive basis founded on experience, to one where they are specified by formal definitions and their properties reconstructed through logical deductions". Consequently proofs acquire a new and important status. They have to be complete and established through logical deductions from the formal definitions and properties. Only elementary logic is necessary for overcoming this difficulty, but this is far above what is being asked in secondary mathematics. Moreover, the basic logic one uses in mathematics is different from the ordinary logic of everyday life, just as the mathematical language differs from the natural language.

Even for those students already familiar with proofs, new difficulties may arise. For instance existence proofs are notoriously difficult for most students: on the one hand, it is not easy for them to recognize their need, as this type of situation is rarely raised in secondary mathematics — when given a problem, high school students can (almost) always take it for granted that it has a solution; and also existence questions are difficult to solve because one often has to imagine a certain mathematical object — an analysis-synthesis approach can be useful here, but is not always easy to implement. Sufficiency arguments are generally difficult because there is often a choice to be made. Sometimes, a proof requires not only to apply directly a theorem in a particular case, but also to adapt or even to transform a theorem before recognizing and/or using it. In other occasions, a proof involves a multi-stage process. For instance one often encounters situations in analysis where in order to find a limit, a given expression (vg, a sum or an integral) must be "broken" into two parts to be treated separately by different methods; a strong qualitative intuition is essential for one to succeed in such an approach.

Research shows that when facing a new complex mathematical task or notion for which intuition may not be sufficient to represent the situation, some students react by introducing simplistic procedures, like trying to reduce everything to algorithms, or by developing for themselves simplistic models — such is the case for instance with the notion of limit, as observed by Robert [19]. Other students' errors are more linked to the mathematical domain involved. For example, many difficulties encountered by students in analysis have to do with the structure of $\mathbf{R}$, especially the order relation. In algebra, students do not realize all the consequence of structures in terms of the constraints thus being introduced, because structures

refer to a new idea. For instance, they are surprised (and even bewildered) by the proof — and even more by the need of a proof — of the fact that in a group there exists only one identity element, which is at the same time the only "left" and the only "right" identity element; or by the fact that there is only one group with three elements. This latter example indicates that students are often unaware of the impact of a given definition on the "degree of freedom" of the elements: in a group with three elements, there is simply no room for freedom!

The above comments deal with epistemological and cognitive difficulties which, in a certain way, are "intrinsic" to mathematics, since they concern the change in the type of mathematics to be mastered by students as they move to tertiary education. We would now like to consider some "extrinsic" difficulties.

One such difficulty has to do with the students themselves: there is a substantial heterogeneity in the mathematical background of students entering university education. Some students are fully ready for the transition to the tertiary level, but others are not. And university teachers often do not care to make sure, at the entrance to university, that each one of them masters the basic notions and skills required for an understanding of their course.

Other difficulties concern more directly the university teachers; for instance, the expectations they might have regarding their students: many university teachers develop a distorted image of students and tend to identify their "average" student with an ideal student who has successfully attended a highly scientific track in one of the best secondary schools. But in an actual class this kind of student may be only a very small minority, or even not exist at all. Teachers must also be sensitive to the importance of making explicit to the students what exactly they are doing and learning, and where they are heading. In other words, they must provide students with identifiable goals, not expecting such insights to emerge naturally by themselves.

Another difficulty concerning university teachers is that they expect students to develop from the beginning an active attitude toward "doing mathematics". But students are often not prepared for this kind of work. The situation is vividly documented in an inquiry which involved several classes of Italian first-year university students enrolled in scientific faculties [2]. More than one third of these students (who had learned many proofs in elementary geometry during their secondary school years) share the belief that *if a proof of some theorem in elementary geometry has been produced by a secondary school pupil (say in grade 9 or 10), then even a clever university student is not entitled to check the correctness or the incorrectness of the proposed proof, without the help of books or experts.* They believe that the authority of a professional mathematician is needed, since *only he knows whether the proposed argument is true.*

A last type of cognitive difficulties we would like to consider is linked to an indispensable organization (or reorganization) of knowledge by students. In order to reach the "advanced mathematical thinking" capacities which are expected of them, students must acquire "the ability to distinguish between mathematical knowledge and meta-mathematical knowledge (e.g. of the correctness, relevance, or elegance of a piece of mathematics)" [21, p. 131], they must come to stand back from the computations and to contemplate the relations between concepts.

It is not possible for students, even through extensive personal work, to have met all possible types of problems pertaining to a specific topic. They thus need to develop a global view bringing forth the connections they need to make. But even when students expect that they will have to modify their view of certain mathematical objects and establish links between them, they often encounter great difficulty in doing so because of a lack of organization of their knowledge. A typical instance of such a situation is found in linear algebra. Students may have learned to solve $f(u) = ku$ using determinants, and also to find eigenvalues. But they may well be unable to recognize these concepts in other situations. For instance they may directly meet the equation $f(w) = qw$ arising from a certain context not immediately linked to eigenvalues; but in order to solve this equation with the techniques they now know, they need to recognize that the crux of the problem has to do with eigenvalues. Still more difficult, after having studied straight lines invariant under an affine application, they must learn to link this problem to the eigenvalues of the associated linear application.

3.2 Sociological and cultural difficulties

A second type of difficulties concerns sociological and cultural factors, especially as seen from an institutional point of view. There is a great diversity of such difficulties and local differences can be quite important, which makes it very difficult to present regularities. We limit ourselves here to a few aspects.

More often than not the size of groups at the tertiary level can be very large, especially in the first year, so that a student is often only one in a crowd. For many students, this represents a major change with respect to secondary school, as was clearly shown in the answers to our questionnaire mentioned in Section 1. While some students deal quite easily with the new environment, others find that moving from a "human-size" high school, where most people know each other, to the anonymity of a large university campus is quite a frightful experience. It is only in the rare case that the student will be known as an individual to the teacher. Moreover, groups may be re-formed every semester, so that there is often little or no "sense of community" developed in the classroom. As a consequence, it is very difficult for students to receive help either from the teacher, who frequently has very little time available, or from peer students. And in contexts where students have access to teaching assistants, this systems often prove to be rather unreliable for a variety of reasons (lack of familiarity of the assistant with the content of the course, lack of perspective, problems of communication because of a language barrier, etc.).

Moreover, notwithstanding the size of groups, some students are not comfortable with the climate which may prevail in the classroom. Here is what Tobias writes about science students who do not pursue science study — but this may well apply to mathematics students: "Some students don't decide to reject science per se. They reject the culture of competition that they see as an unavoidable aspect of undergraduate science study. These students don't drop science because they fail in the competition. Often they do very well. Rather for them issues of 'culture' [...] are as important as the actual subject matter of their studies. They

value such qualities as love for one's subject and intrinsic motivation in one's work, and want these qualities to be part of their academic efforts. They see the culture of college science study, in contrast, as emphasizing extrinsic rewards like getting good grades, and objective goals like getting into graduate or medical school." [25, p. 74] In such a competitive atmosphere, the attention of students concentrates on success at the exams, and not on learning.

In many countries, the democratization of teaching has had as a consequence that many weak students are getting access to university. For such students, their relation to knowledge is often not up to what is being expected of them: they meet difficulties in reaching the required level of abstraction and they confine themselves in mere actions and applications of recipes, unaware of the conceptual shift they must accomplish. As a result of this fragility, these students are in a great need of a highly personalized relationship with their instructor, which would allow for the numerous explanations they require. But the current structure of university makes almost impossible such a contact.

A frequent difficulty with students taking mathematics as a service subject is their underestimation of the role of mathematics with respect to their future career (we do not have in mind here the screening role sometimes forced upon mathematics in certain fields). Many students will have chosen a field of specialization in university in which they were not expecting to have to study mathematics. They will often be at a loss to relate their calculus or linear algebra course to their foreseen profession. The instructor needs to address this issue and convince students of the importance of mathematics for their career.

A final cultural difficulty we would like to mention concerns the general conception of the task of teachers at the university level. The lack of pedagogical awareness of some teachers may stem from the fact that they "are expected to conduct research, and thus their motivation and commitment to teaching may not be as strong as that of secondary school teachers, whose sole responsibility is teaching." [9, p. 676] Moreover university teachers, in most cases, have received their professional training as if their only occupation in mathematics is research. Consequently, they have to develop by themselves the pedagogical and communication capabilities they need with their students — and this is an arduous task! The professional reward system in university mathematics is almost universally focused on success in research, and not in pedagogy. The situation is surely much less dramatic in "teaching-oriented" than in "research-oriented" universities, but still this can be seen as a major impediment to the pedagogical dedication of university teachers on a large scale.

3.3 DIDACTICAL DIFFICULTIES

In this Section we ask ourselves to what point the style of teaching and the performance of teachers, at the university level, might be the cause for difficulties experienced by students. It is quite clear that some of these difficulties arise from the way students have been practicing and learning mathematics at the secondary level; for instance, many students arriving at university do not know how to take notes during a lecture, how to read a textbook, how to plan for the study of a

topic, which questions to ask themselves before they get asked by the teacher. The solution is not for university teachers to get closer to the secondary style of teaching in this respect, as students would not get prepared to become autonomous learners. It would take us too far to discuss here how the teaching and learning of mathematics in the secondary school could be changed to improve the situation; moreover, most of the people taking part in this round-table in one way or another are much nearer to university teaching than to teaching at the secondary level.

Among the various circumstances related to university teachers that might cause some problems to their students, we consider the following.

• *Lack of pedagogical and didactical abilities.* "I know the subject and this is sufficient" is here the customary underlying philosophy. In many places it is a rather common belief among university teachers that all one needs in order to teach mathematics at the university level is to deeply know and understand the subject. However, the one who teaches well a subject at any level is not necessarily the one who knows it the most deeply, but the one who achieves that students learn those ideas and methods they should learn. This requires from the teacher many different skills (mathematical as well as didactical ones) that are rarely present in a spontaneous form. It is very important for teachers to be aware of their own possible deficiencies and to try to remedy them.

• *Lack of adequate models.* It often happens that the university teacher, especially the young one — who is usually also the one charged with the responsibility of teaching entering students —, looks at the university professor as an example to imitate. But much too often this is rather a counterexample showing how NOT to teach mathematics. Our university culture has stimulated in many places mathematicians to disregard, if not to despise, any preoccupation about teaching.

• *Disregard for the importance of the methodology of the subject.* Study and work in mathematics require a different kind of approach than study of, say, history or chemistry. Perhaps it belongs to the secondary school teacher to introduce the student to the style of work needed in each one of the subjects. But since this is usually not done, this specificity should be contemplated during the initial years at the tertiary level.

• *Lack of innovative teaching methods.* Many teachers tend to confine themselves to "unimaginative teaching methods" [21, p. 129], the style of teaching most frequently practiced at the university level being that of a lecture presentation of polished mathematics ("the teacher talks and the student takes notes"). It is sad that many university teachers have never heard, for instance, of the so-called "Moore method" or possible modifications of it (see [4]), or of many other different ways to actively engage students, individually or in groups, in the "discovery" and the development of the subject. Such approaches can give a much more exact measure of what each student is able to do and moreover motivate them more intensely towards the study of mathematics. Finally, it has to be acknowledged that while the recent developments in computer hardware and software (vg, symbolic and/or graphical software) have led quite a few teachers to rethink, totally or partially, their approach to various topics in mathematics, a majority of teachers have never considered seriously how these new tools could be used so to foster students participation, inside and outside the classroom.

• *Carelessness in the design of the course.* University teachers often pay little attention to the actual knowledge and preparation of their students and do not care about the pace which would be at every concrete moment the most adequate for the majority of the students. They may offer no clear guidelines about the course content or the exact objectives their students have to meet. Students sometimes miss well-defined material support (vg, a textbook or duplicated lecture notes) on which to rely — it seems that any decent support may be better, for the profit of the student, than the best set of notes the student would have to take hastily in class, thereby loosing opportunity for active learning. Teachers often offer little help to students, through frequent examples, exercises or problems, for digesting the subject and acquiring a concrete idea about the really important concepts and problems of the subject.

• *Lack of feedback procedures.* In the typical university classroom, there is not much interaction which might help the teacher to know, while the course is still in progress, to what extent what has been "taught" has really been "learned", and why it is so. Due to a lack of know-how (or perhaps of interest), it is often only at the very end of the course that teachers get a picture of their groups, sometimes to find out, perhaps by means of some rather unrealistic examination, that almost all of what they thought students had learnt is completely absent from their minds.

• *Lack of assessment skills.* A crucial component of the process of teaching and learning mathematics is to resort to an adequate way of assessing students' work, designed for their benefit and stimulus. But such an assessment scheme is far from trivial to implement. It is sad to observe that many university teachers make no effort to familiarize themselves with different ways of evaluating students, while many others perhaps learn such methods just after many years of ad-hoc experimentation. Very seldom are alternate assessment processes used, such as portfolios, oral interviews, discussions, proposals of open-ended questions given in advance to students so they have the opportunity to think about the problems in an autonomous study, etc. The easiest solution to evaluation, and possibly the poorest one, is of course the written examination — eventually, for large groups, one which the machine can take care of.

4 Possible actions in order to help alleviating these difficulties

We would like to conclude this paper by listing possible actions which might help alleviating the difficulties in the passage from secondary to tertiary mathematics education. We do not claim that all these actions can actually be implemented, nor that they should have the desired effect. Still it is our hope that such a list, however limited and succinct, can foster discussion around the issues raised in the paper. Some of these actions concern institutional aspects surrounding the transition, while others deal with pedagogical ones.

• Establish a better dialogue between secondary educators and tertiary educators. Such a dialogue can (and must) take place both inside and outside "formal channels" of communication. An interesting example of such a dialogue on a national level is provided by the interest recently aroused in the USA mathematical community by the undergoing revision of the so-called "NCTM Standards" (see

[10]). A more local example is to be found in [6].

• Provide students with orientation activities. This can begin already in secondary school, for example by setting up activities in order to help students individually to choose the track that seems the most appropriate for them, in the perspective of their future university career. In-coming university students should be welcomed with information helping them to better understand the place of mathematics in their university education. In a given course, an orientation document can allow to make explicit to the students the expectations of the teacher, for example that students should work right from the very first day of classes. An example of such a document distributed to students in a first-year calculus course is given in [27, p. 865].

• Provide students with individualized help. One possible solution to the fact that teachers of large groups are often totally unable to provide individual support to students is to create a "Students Help Center" in mathematics. The fact that this becomes a highly visible institutionalized activity could make it a little easier to find the necessary financial support.

• Disseminate information about "success stories". A number of institutions are renowned for their exceptional pedagogical performance. Such situations should be better documented, so to help others to develop the necessary local "culture" (commitment to graduating the students admitted, support provided to students, accessibility of professors, etc.) Successful programs in undergraduate mathematics in the USA are presented in [17] and [26].

• Change the context of the transition step. For instance, have the secondary school courses in the upper grades be delivered at a "higher" and more abstract level, so to get closer to the university teaching style. Or, in the opposite direction, make the first-year university courses closer to secondary school teaching style, i.e., delivered at a "lower" and more intuitive level. Or even have both secondary and university courses change drastically in style and content, taking into account, for instance, the possibilities offered by new technologies and the emerging needs of society (see [12] and [15]). Another approach is to create "bridge courses" for specific groups of students, between secondary and tertiary education, in order to help them to fill their gaps with regard to content, methodology and skills. Or still to introduce selective entrance examination to university, in order to ensure a more homogeneous audience for mathematics classes. A report on the use of a diagnostic placement testing in helping entering university students to choose an appropriate sequence of calculus courses is given in [14].

• Create a context propitious to faculty development. Universities have the responsibility of providing faculty members with a context fostering their general pedagogical development, and especially their awareness of the difficulties experienced by students. Those interested in changing their teaching practices need support, training, team-work, access to forums where pedagogical issues are discussed, etc. It is important that teachers be encouraged to take lots of small initiatives that work — eventually the process will lead to a larger result.

• Help students use resources. All the information cannot come from the teacher in the classroom. Students must get used to choosing, reading and understanding on their own appropriate mathematical information in various forms:

textbooks, library materials, internet, etc.

• Change the "culture" of the students. Whether they are specializing in mathematics or taking mathematics as a service subject, students must come to appreciate the mode of thought specific to mathematics. They need to learn why pure mathematics is at least as important as applied mathematics, and that many of the connections between science and mathematics involve theoretical concepts better understood from the point of view of mathematics. They must realize that mathematics is above all a question of ideas and insight, and not mere techniques — although technical skills do play an important role. Teachers must be aware of the need for students to develop insight; they should not expect that this will come naturally by itself from the experience of solving quantitative problems.

• Change the "culture" of the teachers. Traditional lecturing is just one style of teaching — and often not the most appropriate one. Using different ways of teaching can help students develop different ways of learning. One can propose students alternative types of work, like small group discussions; knowledge understood by the best students can be shared with the others, not in order to give the solutions to problems, but to illustrate what it means to do mathematics. In aiming at helping students change their perception of mathematics and make the transition to "advanced mathematical thinking", teachers must realize that "the formalizing and systematizing of the mathematics is the final stage of mathematical thinking, not the total activity". [24, p. 508–509]

• Establish a better dialogue between mathematicians and users of mathematics. The case of mathematics taught as a service subject (see [3] and [11]) needs special attention. Mathematics should be taught to, say, engineering students by someone who has an adequate understanding of the role played by mathematics in engineering and who can relate mathematics to the interests of the students. Contacts with specialists of the specific domain is essential. It was remarked by Murakami [16, p. 1680] that "it is difficult to see how people of such profile might easily come out of the present educational establishment in any significant numbers".

• Meta-cognitive actions. Students' success is linked to a great extent to their capacity of developing "meta-level" skills allowing them, for instance, to self-diagnose their difficulties and to overcome them, to ask proper questions to their tutors, to optimize their personal resources, to organize their knowledge, to learn to use it in a better way in various modes and not only at a technical level (see [7], [8], [20], [22]). For this to happen, teachers have to make explicit to the students the emergence of new "rules" in mathematics (vg, new concepts) and in learning (vg, need for organization instead of pure memory). Teachers also have to build problems adapted to the various modes of thinking they want the students to acquire, and not only present them problems dealing with technical aspects or preparing for the examination day.

• "Less is more." Decrease the quantity of content covered (provided the teacher does have some control over the content of a course), and engage students in a deeper and more adequate understanding.

References

[1] M. Artigue, "Teaching and learning elementary analysis." In: C. Alsina, J.M. Alvarez Falcón, B.R. Hodgson, C. Laborde and A. Pérez Jiménez, eds., *Selected Lectures from the Eight International Congress on Mathematical Education.* To appear.

[2] P. Boero, "A proposito di intuizione e di rigore nell'insegnamento–apprendimento della geometria: il problema dell'approccio agli enunciati e alle dimostrazioni." *Supplemento al Notiziario dell'Unione Mat. Ital.* 22 (1995) 95–101.

[3] R.R. Clements, P. Lauginie and E. de Turckheim, eds. *Selected Papers on the Teaching of Mathematics as a Service Subject.* (CISM Courses and Lectures no. 305) Springer-Verlag, 1988.

[4] D.W. Cohen, "A modified Moore method for teaching undergraduate mathematics." *American Mathematical Monthly* 89 (1982) 473–474 and 487–490.

[5] K. Cross, "Sixth form mathematics — changes in the curriculum and its effect on preparation for higher education." In: M. Zweng, T. Green, J. Kilpatrick, H. Pollak and M. Suydam, eds., *Proceedings of the Fourth International Congress on Mathematical Education.* Birkhäuser, 1983, pp. 70–72.

[6] F. Demana, "Improving college readiness through school/university articulation." In: N. Fisher, H. Keynes and P. Wagreich, eds., *Mathematicians and Education Reform: Proceedings of the 1988 Workshop.* (CBMS Issues in Mathematics Education, volume 1) American Mathematical Society, 1990, pp. 131–143.

[7] J.-L. Dorier, "Meta level in the teaching of unifying and generalizing concepts in mathematics." *Educational Studies in Mathematics* 29 (1995) 175–197.

[8] J.-L. Dorier, ed., *L'enseignement de l'algèbre linéaire en question.* La pensée sauvage (Grenoble), 1997.

[9] G. Harel and J. Trgalová, "Higher mathematics education." In: A.J. Bishop, K. Clements, C. Keitel, J. Kilpatrick and C. Laborde, eds., *International Handbook of Mathematics Education.* Kluwer, 1996, pp. 675–700.

[10] R. Howe, "The AMS and mathematics education: the revision of the 'NCTM Standards'." *Notices of the American Mathematical Society* 45 (1998) 243–247.

[11] A.G. Howson, J.-P. Kahane, P. Lauginie and E. de Turckheim, eds., *Mathematics as a Service Subject.* (ICMI Study Series) Cambridge University Press, 1988.

[12] D. Hughes-Hallett, "Changes in the teaching of undergraduate mathematics: the role of technology." In: S.D. Chatterji, ed., *Proceedings of the International Congress of Mathematicians 1994.* Birkhäuser, 1995, pp. 1546-1550.

[13] J. Mack, "Transition secondary–postsecondary." (Part of the report of Action Group 5 on Tertiary Academic Institutions.) In: A. Hirst and K. Hirst, eds., *Proceedings of the Sixth International Congress on Mathematical Education.* János Bolyai Mathematical Society, 1988, pp. 159–162.

[14] A. Manaster, "Diagnostic testing: one link between university and high school mathematics." In: N.D. Fisher, H.B. Keynes and P.D. Wagreich, eds., *Mathematicians and Education Reform 1989–1990.* (CBMS Issues in Mathematics Education, volume 2) American Mathematical Society, 1991, pp. 25–37.

[15] D. Mumford, "Calculus reform — for the millions." *Notices of the American Mathematical Society* 44 (1997) 559–563.
[16] H. Murakami, "Teaching mathematics to students not majoring in mathematics — present situation and future prospects." In: I. Satake, ed., *Proceedings of the International Congress of Mathematicians 1990.* Springer, 1991, pp. 1673-1681.
[17] J. Poland, "A modern fairy tale?" *American Mathematical Monthly* 94 (1987) 291–295.
[18] Report on the Conference "Mathematics at the coming to university: real situation and desirable situation." In: *New Trends in Mathematics Teaching*, volume I (1966). Prepared by the International Commission on Mathematical Instruction (ICMI). Unesco, 1967, pp. 366.
[19] A. Robert, *L'acquisition de la notion de convergence des suites numériques dans l'enseignement supérieur.* Thèse de doctorat d'état, Université de Paris VII, 1982.
[20] A. Robert, "Outils d'analyse des contenus enseignés au lycée et à l'université." *Recherches en didactique des mathématiques* 18(2) (1998), to appear.
[21] A. Robert and R. Schwarzenberger, "Research in teaching and learning mathematics at an advanced level." In: [23], pp. 127–139.
[22] A.H. Schoenfeld, *Mathematical Problem Solving.* Academic Press, 1985.
[23] D. Tall, ed., *Advanced Mathematical Thinking.* Kluwer, 1991.
[24] D. Tall, "The transition to advanced mathematical thinking: functions, limits, infinity and proof." In: D.A. Grouws, ed., *Handbook of Research on Mathematics Teaching and Learning.* Macmillan, 1992, pp. 495–511.
[25] S. Tobias, *They're not Dumb, They're Different: Stalking the Second Tier.* Research Corporation, 1990.
[26] A.C. Tucker, *Models That Work: Case Studies in Effective Undergraduate Mathematics Programs.* (MAA Notes 38) Mathematical Association of America, 1995.
[27] S. Zucker, "Teaching at the university level." *Notices of the American Mathematical Society* 43 (1996) 863–865.

Miguel de Guzmán
Facultad de Matemáticas
Universidad Complutense
de Madrid
28040 Madrid, Spain
mdeguzman@bitmailer.net

Bernard R. Hodgson
Département de mathématiques
et de statistique
Université Laval
Québec, G1K 7P4, Canada
bhodgson@mat.ulaval.ca

Aline Robert
IUFM de Versailles
54 avenue des États-Unis
78000 Versailles, France
alr@ccr.jussieu.fr

Vinicio Villani
Dipartimento di Matematica
Università di Pisa
Via Buonarroti 2
I - 56127 Pisa, Italy
villani@dm.unipi.it

Mathematics Instruction in the Twenty-first Century

D. J. Lewis[1]

1991 Mathematics Subject Classification: Primary 00A05
Keywords and Phrases: mathematics education, science education

For the past fifty years, mathematics and science education in the United States, both collegiate and precollegiate, have been criticized as being inadequate and frequently irrelevant. Much of the criticism has focussed on the instruction provided in the elementary and secondary schools and the criticism gets quite intense each time there is an international study of student achievements. Considering the low standing of US students in these comparative studies, it should not be surprising there is discontent with the quality of instruction provided. In the last fifteen years the criticism has been expanded to include mathematics and science instruction at the collegiate level. Here we shall restrict one attention to post-secondary instruction, although what happens in the elementary and secondary schools has a decided impact on the instruction at the next level and the teachers in these schools are educated by the collegiate faculty.

The instructional problems facing the mathematical faculty are not different from that facing science educators in general. Consider the following paragraphs[2] written by Purnell W. Choppin, M.D, President, Howard Hughes Medical Institute.

> "There are two revolutions going on in science education these days. One concerns how teachers teach and students learn, the other concerns technology. As with all revolutions, they carry with them a certain uproar and sense of unease.
>
> In the first revolution – changing how students learn science and how teachers teach science – debate has been raging among those who advocate a more inquiry-based, problem-solving approach to education and those who believe that content should prevail above all else. Obviously, the ideal lies somewhere in between. Students must learn how to solve problems, but they must also respect that there is a clearly defined body of facts and principles that guide thinking and the pursuit of truth. And let it not be forgotten that what drives most scientists to their work is not the desire to merely ask questions,

[1] This discussion is from a US perspective, since that is what we know. While the situation varies from country to country, the evolution occuring worldwide in educational systems suggests many of the problems facing US mathematical educators will probably become universal.

[2] *Making New Connections in Science Education*, 1997, Howard Hughes Medical Institute.

but to find answers about some part of the world that fascinates them and captures their interest. It is the content, after all, that gives meaning to our investigations.

The second revolution concerns the set of powerful new computer-based tools available in a growing number of education settings. It is not that these technologies necessarily change how we learn, but they can transform the speed, intensity, and environment in which we master our universe. Grappling with how to integrate these innovations into the classroom and assess their impact is an exercise we can expect to engage in for some time.

When these revolutions flow together, and they increasingly do, they force us to re-evaluate so many aspects of education that it can, at times seem overwhelming. We want to know if computer-based instruction can draw out talents in students who have been previously difficult to engage. We want to know if student-centered learning via the Internet will satisfy our desire that content be mastered. And we are concerned about conveying the message that learning is always fun, because sometimes it is very hard work. Science is a rigorous endeavor.

Teachers who use technology tell us that its integration into their classrooms has forced a sometimes painful transition in the way they teach, but one that they now feel was worth the effort. In the Institute's undergraduate and precollege education programs we have witnessed firsthand the fruits of these labors; elegant education software, large networks of teachers working together on line, and research products of students who have been guided in the use of these tools for their own learning.

The challenge for science educators is to find the balance between content and pedagogy to ensure that we reach the desired educational outcomes for our children – scientific literacy and finely tuned analytical skills."

Certainly the US mathematical educators face the two revolutions identified by Dr. Choppin and have been struggling with how to reconcile the two. Indeed much of the debate that has arisen is around this reconciliation, and much of the criticism is whether the students do develop "finally tuned analytic skills." Mathematical instruction in the US faces a more complicated situation than do the sciences.

The bulk of US collegiate mathematics instruction deals with the calculus and linear algebra, and with still lower level courses in the case of many state universities, especially regional ones. Too many students are admitted with low mathematical credentials – a fault of the institutions, but a political reality given that a college education is the path to well paying positions. While typically the number of faculty in a US mathematics department is large compared to other countries, the ratio of students to faculty is extremely large; large even when one includes the very substantial numbers of part-timers and graduate student assistants. Engineering, all the sciences (biological, physical, social and financial) and many professional schools require competency in the calculus, linear algebra, and frequently statistics. Mathematics has become intrinsic to all these disciplines and they demand competency on the part of students. As a consequence 70-90% of first-year students will enroll in a mathematics course. The top 10-20% of these students will have completed their study of the calculus in high school, so the mathematical ability of those enrolled in the calculus is not the best, and their motivation is quite varied. In contrast to the past, today's students do not accept the concept of deferred gratification; they demand to know that mathematics is relevant to their career goals.

The mathematical instructors have diligently striven to provide meaningful instruction to these masses of students, trying to meet contradictory goals arising from the students and the various nonmathematics faculties. A very fundamental issue for the faculty are the demands that calculus be presented as a basic tool for science and engineering and their own desire to show it as the great intellectual achievement that it is. Thus we find some instructors treat the calculus and linear algebra as an algorithmic tool; others will present it as a theory, but usually only with heuristic justifications; others will provide a highly rigorous Satz-Beweis course in analysis; and still others, responding to the revolutions described by Dr. Choppin, will emphasize real world modeling and problem solving. The use of technology varies greatly from no use to extensive use of symbolic packages. There are those who do all the instruction via the computer, allowing students to go at their own pace and attempting to combine Dr. Choppin's two revolutions.

Clearly there is merit in each approach and to present calculus and linear algebra maximally one needs to involve all these approaches. This can seldom be done since time allocated to these subjects is limited by the various curriculums. As might be expected, the various approaches have strong advocates, which has made for stringent debate. Despite the tensions, a very positive outcome is that pedagogy and content of first and second year collegiate mathematics instruction is being seriously examined and discussed. Teaching has become an important issue even in the most sophisticated and research intensive departments. A major problem is that there has been little in depth longitudinal evaluations of any of the approaches and there has been little use of cognitive studies to justify any approach. Without such, the debate is mostly opinion.

From my perspective, the American faculty are held to too great a responsibility for what the student learns. This is oxymoron. The faculty can present material, but it falls to the student to do the learning. From my observation, the more the responsibility is laid on the student, the more the course has been deemed a success, whatever the approach.

There is mounting pressure that US research university instruction should be based on discovery learning guided by mentoring rather than on the transmission of knowledge,[3] and that before graduating the student should have experience in research. The motivation for this approach comes from the repeated observation that "the teacher who puts his hand on your shoulder is the one who has had an impact on your life." Given the large number of students enrolled and the rather low preparation with which they arrive and the pressure on the faculty to do research, this may be wishful thinking on the part of our academic leaders. But given the outlook of young Americans, it probably will be the only way to attract and retain the very best students in mathematics and science. I would expect a research experience that goes beyond the current undergraduate expository thesis,

[3] *Reinventing Undergraduate Education*, Boyer Commission on the Education of Undergraduates in the Research University, Carnegie Foundation for the Advancement of Teaching, 1998. This report is highly critical of the education undergraduates receive at US Research Universities, and it includes suggestions for improvement.

will become the norm for the honor's students. These developments suggest that the discussion is about to expand to all undergraduate instruction.

The US academic community faces many challenges in the 21^{st} Century. As higher education becomes more universal world-wide, it is likely some of the challenges will come to others.

D. J. Lewis
National Science Foundation
Arlington VA 22230
USA
dlewis@nsf.gov

Aspects of the Nature and State of Research in Mathematics Education

Mogens Niss

Abstract. This paper offers an outline and a characterisation of the didactics of mathematics, alias the science of mathematics education, as a scientific and scholarly discipline. It further presents a number of major, rather aggregate findings in the discipline, including *the astonishing complexity of mathematical learning*, *the key role of domain specificity*, *obstacles produced by the process-object duality*, *students' alienation from proof and proving*, and *the marvels and pitfalls of information technology in mathematics education*.

1991 Mathematics Subject Classification: 00A35, 00–02
Keywords and Phrases: the didactics of mathematics, mathematics education research

1 Introduction

During the last three decades or so mathematics education has become established as an academic discipline on the international scene. This discipline is given slightly different names in different quarters, such as *mathematics education research*, *science of mathematics education*, and *the didactics of mathematics*. In the following I shall use the names interchangeably.

What are the issues and research questions of the didactics of mathematics, what are its methodologies, and what sorts of results or findings does it offer? In this paper attempts will be made to characterise this discipline, in particular as regards its nature and state, and to present and discuss some of its major findings. I shall begin by offering a definition of the field.

2 Characterising the field

A definition

Subject *The didactics of mathematics, alias the science of mathematics education, is the scientific and scholarly field of research and development which aims at* identifying, characterising*, and* understanding *phenomena and processes actually or potentially involved in the* teaching *and* learning of mathematics at any educational level.

ENDEAVOUR *As particularly regards 'understanding' of such phenomena and processes, attempts to uncover and clarify* causal relationships *and* mechanisms *are in focus.*

APPROACHES *In pursuing these tasks, the didactics of mathematics addresses* all matters *that are pertinent to the teaching and learning of mathematics, irrespective of which scientific, psychological, ideological, ethical, political, social, societal, or other spheres this may involve. Similarly, the field makes use of considerations, methods, and results from* other fields and disciplines *whenever this is deemed relevant.*

ACTIVITIES *The didactics of mathematics comprises different kinds of activities, ranging from theoretical or empirical* fundamental research*, over* applied research *and* development*, to* systematic, reflective practice*.*

It is important to realise a peculiar but essential aspect of the didactics of mathematics: its *dual nature*. As is the case with any academic field, the didactics of mathematics addresses what we may call *descriptive/explanatory* issues, in which the generic questions are 'what *is* (the case)?' and '*why* is this so?'. Objective, neutral answers are sought to such questions by means of empirical and theoretical data collection and analysis without any intrinsic involvement of values (norms). However, by its nature mathematics education implies the fundamental presence of values and norms. So, in addition to its descriptive/explanatory dimension, the didactics of mathematics also has to contain a *normative* dimension, in which the generic questions are 'what *ought to be* the case?' and '*why* should this be so?'. Both dimensions are essential constituents of the science of mathematics education, but they should not be confused with one another.

In a brief outline of the main areas of investigation the two primary ones are *the teaching of mathematics*, and *the learning of mathematics*. A closely related area of investigation is the *outcomes* (results and consequences) of the teaching and the learning of mathematics, respectively.

We may depict, as in Figure 1, these areas in a 'ground floor'. The investigation of these areas leads to derived needs to investigate certain auxiliary areas related to the primary ones but not in themselves of primary didactic concern, such as aspects of mathematics as a discipline, of cognitive psychology, or of curriculum design. As is the case with any new or established scientific field, the didactics of mathematics reflects on its own nature, issues, methods, and results (e.g., Grouws, 1992; Biehler et al., 1994; Bishop et al., 1996; Sierpinska & Kilpatrick, 1998). Theoretical or empirical studies in which the field as such is made subject of investigation form part of the field itself, although at a meta-level, which we depict as an 'upper floor' plane. We may think of it as being transparent so as to allow for contemplation of the ground floor from above. Finally, let us imagine a vertical plane cutting both floors as a common wall. The two half-spaces thus created may be thought of as representing the descriptive/explanatory and the normative dimensions, respectively. If we imagine the vertical wall to be transparent as well, it is possible to look into each dimension from the perspective of the other.

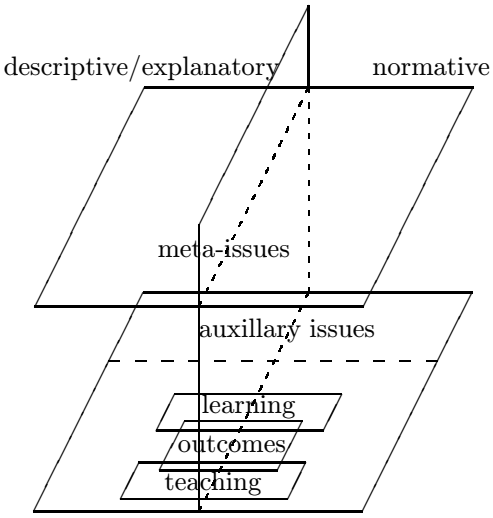


Figure 1: Survey map

Let us sum up the ultimate goals of the didactics of mathematics as follows: We want to be able to specify and characterise *satisfactory learning* of mathematics, including the mathematical competencies we should like to see different categories of individuals possessing. We want to be able to devise and implement *effective mathematics teaching* that can serve to bring about satisfactory learning. We finally want to construct and implement valid and reliable *ways to detect and assess*, without destructive side effects, the results of learning and teaching of mathematics.

For all this to be possible we have to be able to identify and understand the role of mathematics in science and society; what learning of mathematics is, what its conditions are, how it may take place, how it may be hindered, how it can be detected, and how it can be influenced, all with respect to different categories of individuals. We further have to understand what takes place in existing forms mathematics teaching, both as regards the individual student, groups of students, and entire classrooms. We have to invent and investigate new modes of teaching. We have to investigate the relationships between teaching modes and learning processes and outcomes, and the influence of teachers' backgrounds, education, and beliefs on their teaching. We have to examine the properties and effects of established and experimental modes of assessment in mathematics education, with particular regard to the ability to provide valid insight into what students know, understand, and can do.

Traditionally, fields of research within the sciences produce either *empirical findings* of facts', through some form of data collection assisted by theoretical considerations, or they produce *theorems*, i.e. statements derived by means of logical deduction from a collection of 'axioms'. If we go beyond the predominant

paradigms in the sciences and look at the humanities and the social sciences, other aspects have to be added to the ones just considered. In philosophical disciplines, the proposal and analysis of distinctions and concepts, and concept clusters, introduced to specify and represent matters from the real world, serve to create a platform for discussion of these matters in a clear and systematic way. Such disciplines often produce *notions*, *distinctions*, *terms*, amalgamated into *concepts*, or extensive hierarchical networks of concepts connected by formal or material reasoning, called *theories*. Disciplines dealing with human beings, as individuals, as members of different social and cultural groups, and as citizens, or with communities and societies at large, primarily produce *interpretations* and *models*, i.e. hypotheses of individual or social forces and mechanisms that may account for phenomena and structures observed in the domain under consideration. Sometimes sets of interpretations are organised and assembled into systems of interpretation, also called 'theories' which we shall refer to as *interpretative theories*. Finally, there are disciplines within all scientific spheres that produce *designs* (and eventually *constructions*) for which the ultimate test is their functioning in the realm in which they are put into practice. However, as designs and constructions are often required to have certain properties before installation, design disciplines are scientific only to the extent they can provide well-founded reasons to believe that their designs possess certain such properties to a satisfactory degree.

The didactics of mathematics contains instances and provides findings of all the categories of disciplines mentioned, but to strongly varying degrees. There are empirical findings as well as 'theorems' (but, in the honour of truth, these are derived within mathematics itself). There are terms, concepts and theories for analysis of a philosophical nature, and there are models, interpretations and interpretative theories of a pscyhological, sociological or historical nature. Finally there are multitudes of designs and constructions of curricula, teaching approaches, instructional sequences, learning environments, and materials.

Some researchers in mathematics education are hesitant to use the term 'finding', in order to avoid too narrow expectations of what the field has to offer. They prefer to see the didactics of mathematics as providing generic tools for analysing teaching/learning situations. Others emphasise that the field offers illuminating case studies which are not necessarily generalisable beyond the cases themselves, but are nevertheless stimulating for thought and practice. However, as long as we keep in mind that the notion of finding is a broad one, I don't see any severe problems in using this term in the didactics of mathematics.

A major portion of recent research has focused on students' *learning processes and products* as manifested on the individual, small group, and classroom levels, and as conditioned by a variety of factors such as mathematics as a discipline; curricula; teaching; tasks and activities; materials and resources, including text books and information technology; assessment; students' beliefs and attitudes; educational environment, including classroom communication and discourse; social relationships amongst students and between students and teacher(s); teachers' education, backgrounds, and beliefs; and so forth. The typical findings take the shape of models, interpretations, and interpretative theories, but often also of solid empirical facts. We know a lot about the possible mathematical learning

processes of students and about how these may take place within different areas of mathematics and under different circumstances and conditions, as we know a lot about factors that may hinder or simply prevent successful learning.

We have further come to know a great deal about what happens in *actual mathematics teaching* in classrooms at different levels and in different places (Cobb & Bauersfeld, 1995). However, we are still left with hosts of unanswered questions as to how to design, organise, and carry out teaching-learning environments and situations that to a reasonable degree of certainty lead to satisfactory learning outcomes for various categories of students. This is not to say that we don't know anything in this respect, but as yet our knowledge is more punctual and scattered than is the case with our insights into students' mathematical learning. Based on our growing insight into mathematical learning processes and teaching situations, we know more and more about what *is not* effective teaching vis-à-vis various groups of recipients. Moreover, the didactic literature displays numerous examples of experimental teaching designs and practices that are judged highly successful, without this success being easily analysed and documented in scientific terms.

3 Examples of major findings

In this section, we shall consider a few significant findings, of a pretty high level of aggregation, which can serve to illustrate the range and scope of the field. As it is not possible here to provide full documentation of the findings selected, a few recent references, mainly of survey or review type, have to suffice.

The astonishing complexity of mathematical learning *An individual student's mathematical learning often takes place in immensely complex ways, along numerous strongly winding and frequently interrupted paths, across many different sorts of terrain. Some elements are shared by large classes of students, whereas others are peculiar to the individual.*

Students' misconceptions and errors tend to occur in systematic ways in regular and persistent patterns, which can often be explained by the action of an underlying tacit rationality put to operation on a basis which is distorted or insufficient.

The learning processes and products of the student are strongly influenced by a number of crucial factors, including the epistemological characteristics of mathematics and the student's beliefs about them; the social and cultural situations and contexts of learning; primitive, relatively stable implicit intuitions and models that interact, in a tacit way, with new learning tasks; the modes and instruments by which learning is assessed; similarities and discrepancies between different 'linguistic registers'.

This over-arching finding is an agglomeration of several separate findings, each of which results from extensive bodies of research. The roles of epistemological issues and obstacles in the acquisition of mathematical knowledge have been studied, for instance, by Sierpinska and others (for an overview, see Sierpinska & Lerman, 1996). Social, cultural, and contextual factors in mathematical learning

have been investigated from many perspectives, e.g. Bishop, 1988, and Cobb & Bauersfeld, 1995. Pehkonen (e.g. Pehkonen & Törner, 1996), among others, have investigated students' (and teachers') belief's. Fischbein and his collaborators have studied the influence of tacit models on mathematical activity (Fischbein, 1989). The influence of assessment on the learning of mathematics has been subject of several theoretical and empirical studies (e.g. Niss 1993). The same is true with the role of language and communication (see Ellerton & Clarkson, 1996, for an overview).

The studies behind these findings teach us to be cautious when dealing with students' learning of mathematics. Neither processes nor outcomes of mathematical learning are in general logically ordered. For instance, research has shown that many students who are able to correctly solve an equation such as $7x-3 = 13x+15$ are unable to subsequently correctly decide whether $x = 10$ is a solution. The explanation normally given to this phenomenon is that *solving* equations resides in one domain, strongly governed by rules and procedures with no particular attention being paid to the objects involved, whereas examining whether or not a given element solves the equation requires an understanding of what a *solution* means. So, the two facets of the solution of equations, intimitely linked in the mind of the mature knower, need not even both exist in the mind of the novice mathematical learner, let alone be intertwined.

THE KEY ROLE OF DOMAIN SPECIFICITY *For a student engaged in learning mathematics, the specific nature, content and range of a mathematical concept that he or she is acquiring or building up are, to a large part, determined by the set of specific domains in which that concept has been concretely exemplified and embedded for that particular student.*

The finding at issue is closely related to the finding that students' *concept images* are not identical with the *concept definitions* they are exposed to (for overviews, see Vinner, 1991, and Tall, 1992). The concept images are generated by previous notions and experiences as well as by the examples against which the concept definitions have been tested.

The range and depth of the instances of this finding have far-reaching bearings on the teaching and learning of mathematics. Thus, not only are most 'usual' students unable to grasp an abstract concept, given by a definition, in and of itself unless it is elucidated by multiple examples (which is well known), but, more importantly, the scope of the notion that a student forms is often barred by the very examples studied to support that notion. For example, even if students who are learning calculus or analysis are presented with full theoretical definitions, say of $\epsilon - \delta$ type, of function, limit, continuity, derivative, and differentiability, their actual notions and concept images will be shaped, and limited, by the examples, problems, and tasks on which they are actually set to work. If these are drawn exclusively from objects given as standard expressions of familiar, well-behaved objects, the majority of students will gradually tie their notions more and more closely to the specimens actually studied. Thus, the general concept image becomes equipped with properties resulting from an over-generalisation of properties held by the special cases but not implied by the general concept. Remarkably enough, this does not prevent many of the very same students from correctly re-

membering and citing general theoretical definitions. These definitions seem to just be parked in mental compartments detached from the ones activated in the study of the cases. In other words, if average students are to understand the range of a mathematical concept, they have to experience this range by exploring a large variety of manifestations of the concept in various domains.

The danger of forming too restricted images of general concepts seems to be particularly manifest in domains — such as arithmetic, calculus, linear algebra, statistics — that lend themselves to an algorithmic 'calculus', in a general sense. In such domains, algorithmic manipulations — procedures — tend to attract the main part of students' attention so as to create a 'concept filter': Only those instances (and aspects) of a general concept that are relevant in the context of the 'calculus' are preserved in students' minds. In severe cases an over-emphasis in instruction on procedures may even prevent students from developing further understanding of the concepts they experience through manipulations only.

The present finding shows that it is a non-trivial matter of teaching and learning to establish mathematical concepts with students so as to be both sufficiently general and sufficiently concrete. Research further suggests that for this to happen, several different *representations* (e.g. numerical, verbal, symbolic, graphical, diagrammatical) of concepts and phenomena are essential, as are the links and transitions between these representations.

There is a large and important category of mathematical concepts of which the acquisition becomes particularly complex and difficult, namely concepts generated by *encapsulating* specific processes into objects. Well-known examples of this are the concept of function as an *object*, encapsulating the mechanisms that *produce the values* of the function into an entity, and the concept of derivative, encapsulating the processes of differentiating a function pointwise, and of amalgamating the outcomes into a new function. This *process-object duality*, so characterstic of many mathematical concepts, is referred to in the research literature by different terms, such as 'tool-object' (Douady, 1991), 'reification' (Sfard, 1991), 'procept', a hybrid of process and concept, (Tall, 1991, Chapter 15). It constitutes the following finding:

Obstacles produced by the process-object duality *The process-object duality of mathematical concepts that are constituted as objects by encapsulation/reification of specific processes, typically gives rise to serious learning obstacles for students. They often experience considerable problems in leaving the process level and entering the object level.*

For example, many students conceive of an equation as signifying a prompt to perform certain operations, without holding any conception of an equation as such, distinct from the operations to be performed. To them, an equation simply does not constitute a mathematical entity, such as a statement or a predicate.

Undoubtedly, the notions of mathematical proof and proving are some of the most crucial, demanding, complex, and controversial, in all of mathematics education. Deep scientific, philosophical, psychological, and educational issues are involved in these notions. Hence it is no wonder that they have been made subject of discussion and study in didactic research to a substantial extent over the years (for a recent discussion, see Hanna & Jahnke, 1996). Here, we shall confine

ourselves to indicating but one finding pertinent to proof and proving.

STUDENTS' ALIENATION FROM PROOF AND PROVING *There is a wide gap between students' conceptions of mathematical proof and proving and those held by the mathematics community. Typically, students experience great problems in understanding what a proof is supposed to be, and what its purposes and functions are, as they have substantial problems in proving statements themselves. Research further shows that many students who are able to correctly reproduce a (valid) proof, do not see the proof to have, in itself, any bearing on the truth of the proposition being proved.*

The fact that proof and proving represent such great demands and challenges to the learning of mathematics implied that proof and proving have received, in the '80's and '90's, a reduced emphasis in much mathematics teaching. However, there seems to be a growing recognition that there is a need to revitalise them as central components in mathematics education. Also there is growing evidence that it is possible design and stage teaching-learning environments and situations so as to successfully meet parts of the demands and challenges posed by proof and proving.

The last finding to be discussed here, briefly, is to do with the role and impact of information technology on the teaching and learning of mathematics. This is perhaps the single most debated issue in mathematics education during the last two decades, and one which has given rise to large amounts of research (for recent overviews, see Balacheff & Kaput, 1996; and Heid, 1997). The following finding sums up the state-of-the-art:

THE MARVELS AND THE PITFALLS OF INFORMATION TECHNOLOGY IN MATHEMATICS EDUCATION *Information technology has opened avenues to new ways of teaching and learning which may help to greatly expand and deepen students' mathematical experiences, insights, and abilities. However, this does not happen automatically but requires the use of technology to be embedded, with reflection and care, as one element amongst others into the overall design and implementation of teaching-learning environments and situations. The more students can do in and with information technology in mathematics, the greater is the need for their understanding and critical analysis of what they are doing.*

One pitfall of information technology indicated in the research literature is that the technological system itself can form a barrier and an obstacle to learning, either by simply becoming yet another topic in the curriculum, or by distracting students' attention to the system and away from the learning of mathematics. Once again, for this to be avoided it is essential that information technology be assigned a role and place in the entire teaching- learning landscape on the basis of an overall reflective and analytic strategy.

In other words, it is not a simple matter to make information technology assume a role in mathematics education which serves to extend and amplify students' general mathematical capacities rather than replacing their intellects. There is ample research evidence for the claim that when it is no longer our task to train the 'human calculator', some of the traditional drill does becomes obsolete. However, we have yet to see research pointing out exactly what and how much procedural

ability is needed for understanding the processes and products generated by the information technology.

4 Conclusion

In a short paper it is not possible to do justice to then entire field of the didactics of mathematics. Instead of the few findings put forward here, hosts of other findings could equally well have been selected in their place.

Important findings concerning the demands and potentials of *problem solving* and *applications and modelling*; the problems and potentials of *assessment*; the values and efficiency of *collaborative learning* and *innovative teaching approaches and forms of study*, such as project work; the significance of carefully balanced, innovative *multifaceted curricula*, elucidating historical, philosophical, societal, and applicational aspects of mathematics; the impact of *social, cultural and gender factors* on mathematics education; and many others, have not, regrettably, been given their due shares in this presentation. The same is true with the findings contributed by impressive bodies of research on the teaching and learning of specific mathematical *topics*, such as arithmetic, abstract and linear algebra, calculus/analysis, geometry, discrete mathematics, and probability and statistics, and with the findings represented by the instrumental interpretative theories. Also the extensive and elaborate examples of didactical engineering (design and construction) contributed by a number of research and development centres in different countries have been left out of this survey.

Nevertheless, the findings which we have been able to present suffice to teach us two lessons which we might want to call *super-findings*. If we want to teach mathematics to students other than the rather few who can succeed without being taught, or the even fewer who cannot learn mathematics irrespective of how they are taught, two matters have to be kept in mind at all times:

1. We have to be infinitely careful not to jump to conclusions and make false inferences about the processes and outcomes of students' learning of mathematics.

2. If there is something we want our students to know, understand, or be able to do, we have to make it object of explicit and carefully designed teaching. There is no such thing as guaranteed transfer of knowledge, insight and ability from one context or domain to another, it has to be cultivated.

5 Acknowledgements

Key sections of this paper have been greatly inspired by a number of the world's leading researchers in mathematics education. Sincere thanks are due to C. Alsina, M. Artigue, A. Bishop, M. Bartolini Bussi, O. Björkqvist, R. Douady, T. Dreyfus, P. Ernest, J. Fey, P. Galbraith, G. Gjone, J. Godino, G. Hanna, K. Heid, B. Hodgson, C. Laborde, G. Leder, D. Mumford, M. Neubrand, E. Pehkonen, L. Rico, K. Ruthven, A. Schoenfeld, A. Sfard, O. Skovsmose, H. Steinbring, V. Villani, and E. Wittmann, for their advice. Needless to say, the responsibility for the entire paper, especially for any flaws or biases it may contain, is mine alone.

REFERENCES

Balacheff, N. and Kaput, J. (1996). 'Computer-Based Learning Environments in Mathematics'. In Bishop et al., 1996, chapter 13, pages 469–501.

Biehler, R., Scholz, R., Strässer, R., and Winkelmann, B., eds. (1994). *Didactics of Mathematics as a Scientific Discipline*. Dordrecht: Kluwer Academic Publishers.

Bishop, A. (1988). *Mathematical Enculturation. A Cultural Perspective on Mathematics Education*. Dordrecht: Kluwer Academic Publishers.

Bishop, A., Clements, K., Keitel, C., Kilpatrick, J., and Laborde, C., eds. (1996). *International Handbook of Mathematics Education*, volume 1–2. Dordrecht: Kluwer Academic Publishers.

Cobb, P. and Bauersfeld, H. (1995). *The Emergence of Mathematical Meaning: Interaction in Classroom Cultures*. Hillsdale (NJ): Lawrence Erlbaum.

Douady, R. (1991). 'Tool, Object, Setting, Window'. In Bishop, A., Mellin-Olsen, S., and van Dormolen, J., eds., *Mathematical Knowledge: Its Growth Through Teaching*, pages 109–130. Kluwer Academic Publishers, Dordrecht.

Ellerton, N. and Clarkson, P. (1996). 'Language Factors in Mathematics Teaching and Learning'. In Bishop et al., 1996, chapter 26, pages 987–1033.

Fischbein (1989). 'Tacit Models and Mathematics Reasoning'. *For the Learning of Mathematics*, 9(2), 9–14.

Grouws, D., ed. (1992). *Handbook of Research on Mathematics Teaching and Learning*. New York (NY): Macmillan Publishing Company.

Hanna, G. and Jahnke, H. (1996). 'Proof and Proving'. In Bishop et al., 1996, chapter 23, pages 877–908.

Heid, K. (1997). 'The Technological Revolution and the Reform of School Mathematics'. *American Journal of Education*, 106(1), 5–61.

Niss, M., ed. (1993). *Investigations into Assessment in Mathematics Education*. Dordrecht: Kluwer Academic Publishers.

Pehkonen, E. and Törner, G. (1996). 'Mathematical beliefs and different aspects of their meaning'. *Zentralblatt für Didaktik der Mathematik*, 28(4), 101–108.

Sfard, A. (1991). 'On the Dual Nature of Mathematical Conceptions: Reflections on Processes and Objects as Different Sides of the Same Coin'. *Educational Studies in Mathematics*, 22(1), 1–36.

Sierpinska, A. and Kilpatrick, J., eds. (1998). *Mathematics Education as a Research Domain*, volume 1–2. Dordrecht: Kluwer Academic Publishers.

Sierpinska, A. and Lerman, S. (1996). 'Epistemologies of Mathematics and of Mathematics Education'. In Bishop et al., 1996, chapter 22, pages 827–876.

Tall, D., ed. (1991). *Advanced Mathematical Thinking*. Dordrecht: Kluwer Academic Publishers.

Tall, D. (1992). 'The Transition to Advanced Mathematical Thinking: Functions, Limits, Infinity, and Proof'. In Grouws, 1992, pages 495–511.

Vinner, S. (1991). 'The Role of Definitions in Teaching and Learning'. In Tall, 1991, chapter 5, pages 65–81.

Mogens Niss
Roskilde University, IMFUFA,
P.O.Box 260,
DK-4000 Roskilde, Denmark
e-mail: mn@mmf.ruc.dk

Renewal in Collegiate Mathematics Education

To the memory of James R. C. Leitzel

David A. Smith

Abstract. The content and pedagogy of college courses in mathematics and science are not well aligned with the desired outcomes of college education. This is due in part to a professoriate that is largely unaware of pedagogical "best practice." Recent research on neurobiology confirms research on the psychology of learning, and both support best practice in pedagogy. The Calculus Reform Movement has developed courses that focus on student-centered learning and show that new knowledge can be translated into effective learning programs. Computer and calculator technologies offer opportunities to rethink a mathematics curriculum heavily weighted with pre-computer techniques, to create learning environments that accord with best practice, and to shift the primary focus in our courses from manipulation to thinking.

1991 Mathematics Subject Classification: 00A35
Keywords and Phrases: Calculus reform, student-centered learning, cognitive psychology, neurobiology, good practice in pedagogy, Kolb learning cycle, instructional technology

1 Calculus: Reform or Renewal?

"The great obstacle to progress is not ignorance but the illusion of knowledge."[1]

The primary qualification for teaching mathematics in an American university or college is a Ph.D. in mathematics. We take for granted that anyone who has mastered the subject at this level is prepared to teach. If we do what our teachers did, we will be successful — it worked for us. This is not ignorance but a dangerous illusion of knowledge: Good teaching engendered learning in us, so our job is good teaching — learning will follow. If it doesn't, the students must be at fault.

In the mid-1980's there was widespread recognition that something was wrong with this theory, at all levels of mathematics education. Calculus was chosen as the first target for "reform" because it was both the capstone course for secondary education and the entry course for collegiate mathematics. Thus was born the

[1] Daniel Boorstin, former director of the Library of Congress ([2], p. 57).

Calculus Reform Movement, whose history, philosophy, and practice are described in [9], [11], [13].

The first National Science Foundation calculus grants were awarded 10 years ago. Since then we have seen development and implementation of several new approaches to teaching calculus, with widespread acceptance on some campuses, and rejection and backlash on others. Our own approach is to treat calculus as a laboratory science course that emphasizes real-world problems, hands-on activities, discovery learning, writing, teamwork, intelligent use of tools, and high expectations of students.

At the time of development, we had little or no theoretical support for our choice of strategies. In place of theory, we relied on careful empirical work. The following sections develop the theoretical base that we lacked 10 years ago. The results from cognitive psychology were in the literature then but unknown to us and most of the other developers. The results from neurobiology have come to fruition just in this decade, and they confirm the cognitive theories that fit with our empirical observations. Thus, we are replacing the illusion of knowledge with real knowledge about learning and the teaching strategies that engender learning.

In hindsight, "reform" was not a good choice of name. The word has stuck, and most people recognize the course types to which it refers. However, it is an emotionally charged word — in the area of religion, wars have been fought over it. One source of the current controversy is that people with deeply held beliefs feel they are under attack. "Renewal" would be a better descriptor — perhaps we can discuss rationally whether the new aspects are also good, and whether renewal of pedagogical strategies from time to time is itself a good thing to do.

2 Who studies calculus and why

Some 700,000 students enroll in college-level calculus courses in the U. S. in any given year. Of these, 100,000 are in Advanced Placement courses in high schools, 125,000 in two-year colleges, and the rest in four-year colleges or universities [11]. A very small percentage of these students intend to take any mathematics beyond calculus, let alone major in mathematics or do graduate study or become a mathematician. Most of this enrollment is generated either by general education requirements or by prerequisites for subsequent course work. To cite just one example, Duke University has 24 major programs that require one or more semesters of calculus. Even though many students enter with Advanced Placement credits, some 80% of our first-year students take a calculus course. About 2% of each class graduates with a major in mathematics. Thus, most students are not motivated to study calculus except as it serves some other goal — e.g., keeping open options for a major.

American colleges provide liberal, vocational, and/or pre-professional education to students who overwhelmingly see themselves as participants in pre-professional or vocational programs. A small percentage contemplate academic graduate study, but only the tiniest fraction have any concept of liberal education and its potential importance in their lives. Parents usually see things the same way: The objective is for their child to become productive and self-supporting.

Potential employers of graduates at all levels have definite expectations for the skills and abilities of their employees. Collectively, these employers influence support for and accountability from institutions of higher education, public or private. Here is what they want, expressed in seven "skill groups" [1]:

1. The foundation: knowing how to learn
2. Competence: reading, writing, and computation
3. Communication: listening and speaking
4. Adaptability: creative thinking and problem-solving
5. Personal management: self esteem, goal setting and motivation, personal and career development
6. Group effectiveness: interpersonal skills, negotiation, and teamwork
7. Influence: organizational effectiveness and leadership

Students enter college lacking most of these skills, so college must be where they learn them. Indeed, this list defines the goals of higher education in the broad sense: liberal, vocational, and pre-professional. The job of teaching these skills belongs to the entire faculty, including the Mathematics Department — and not just for "computation" and "problem-solving." To get a consistent message from the faculty and to have a good chance of graduating with these skills in place, students must encounter most of them in almost every course.

3 Problems with American collegiate education in mathematics

What was wrong with mathematics education in colleges and universities in the 1980's that led to a perceived need for reform? Many have described the turned-off students and jaded faculty in our classrooms and lecture halls, usually with the intention of blaming someone — teachers at a lower level, society, administrators, or the students themselves. A more constructive description appears in a recent essay [8], a product of discussions among a group of 35 science and mathematics faculty, administrators, foundation officers, and program directors. Their thesis is that there is broad consensus on what constitutes effective science education, but institutional barriers to change have thus far prevented widespread implementation. We quote selected parts of their description of the problem. (The word "science" here is shorthand for "science, mathematics, engineering, and technology.")

> "The traditional approach is to conceive of science education as a process that sifts from the masses of students a select few deemed suitable for the rigors of scientific inquiry. It is a process that resembles what most science faculty remember from their own experiences, beginning with the early identification of gifted students before high school, continuing with the acceleration of those students during grades 9 to 12, fostering in them the disciplined habits of inquiry through their undergraduate majors, and culminating in graduate study and the earning of a Ph.D. Forgotten ... are most students for whom a basic knowledge of science is principally a tool for citizenship, for personal enlightenment,

> for introducing one's own children to science, and for fulfilling employment. Forgotten as well are those students who will become primary and secondary school teachers and, as such, will be responsible for the general quality of the science learning most students bring with them to their undergraduate studies. ...
>
> "Although it is widely recognized that an inquiry-based approach to science increases the quality of learning, introductory-level students are often not given to understand what it means to be a scientist at work. ...
>
> "... science faculty have at times openly acknowledged their tendency to gear instruction to the top 20 percent of the class — to those students whose native ability and persistence enable them to keep pace with the professor's expectations. The fact that others are falling behind and then dropping out is seen not as a failure of pedagogy but as an upholding of standards."

In short, when we use ourselves as models for our students, we get it all wrong. Hardly any entry-level mathematics and science students are like us. In particular, most students in most calculus courses are in their *last* mathematics course. And these students are the next generation's parents, workers, employers, doctors, lawyers, schoolteachers, and legislators. It matters to us how they regard mathematics.

It's not hard to trace how we got out of touch with the needs of our students. Those of us educated in the Sputnik era were in the target population of that "traditional approach" — just at the end of a time when it didn't matter much that the majority of college graduates (an elite subset of the population) didn't know much about science or mathematics. As we became the next generation of faculty, the demographics of college-going broadened significantly, new money flowed to support science, and broad understanding of science became much more important. The reward structure for faculty was significantly altered in the direction of research — away from teaching — just when we were confronted with masses of students whose sociology was quite different from our own.

This oversimplifies a complex story, but our response was to water down expectations of student performance, while continuing to teach in the only way we knew how. We created second-tier courses (e.g., calculus for business and life sciences), we wrote books that students were not expected to read, and we dropped test questions we didn't dare ask. The goal for junior faculty was to become senior faculty so we wouldn't have to deal with freshman courses. Along the way, we produced high-quality research and excellent research-oriented graduate students to follow in our footsteps. But seldom was there any opportunity or incentive to learn anything about learning — in particular, about how our students learn.

4 Messages from cognitive psychology

In 1987, Chickering and Gamson [2], building on an exhaustive review of "50 years of research on the way teachers teach and students learn," enunciated Seven

Principles of Good Practice in Undergraduate Education:

1. Encourages student-faculty contact.
2. Encourages cooperation among students.
3. Encourages active learning.
4. Gives prompt feedback.
5. Emphasizes time on task.
6. Communicates high expectations.
7. Respects diverse talents and ways of learning.

They also published detailed inventories for faculty and administrators ([2], Appendices B and C) to assess the extent to which a school, its departments, and its faculty do or do not follow these principles. One does not need an inventory to see that much of the traditional teaching practice in mathematics is not in accord with these principles. But it doesn't have to be that way. Indeed, [2] is a handbook for implementing these principles.

Research in cognitive psychology has been sending us consistent messages for a half-century, but few mathematicians were listening until the current decade. As Chickering and Gamson summarize,

> "While each practice can stand on its own, when they are all present, their effects multiply. Together, they employ six powerful forces in education:
> - Activity
> - Cooperation
> - Diversity
> - Expectations
> - Interaction
> - Responsibility."

Another result from cognitive research is the Kolb learning cycle ([6], pp. 128-133). The four stages of this cycle are

- Concrete Experience (CE)
- Reflection/Observation (RO)
- Abstract Conceptualization (AC)
- Active Experimentation (AE)

The ideal learner cycles through these stages in each significant learning experience. The AE stage represents testing in new situations the implications of concepts formed at the AC stage. Depending on the results of that testing, the cycle starts over with a new learning experience or with a revision of the current one. The ideal learning environment is designed to lead the learner through these stages and not allow "settling" in a preferred stage. But there are few ideal learners. Most have preferred learning activities and styles, and they are not all alike. This is one reason why learning experiences work better for everyone in a diverse, cooperative, interactive group.

The action-reflection axis (AE-RO) and the concrete-abstract axis (CE-AC) divide the Kolb cycle into four quadrants associated with the four dominant learning styles ([6], pp. 131-132): *Converger* (AC, AE), *Diverger* (CE, RO), *Assimilator* (AC, RO), and *Accommodator* (CE, AE). Most people are not rooted at a single point in the learning style plane, but rather move around in some subset of this plane, depending on the task at hand. However, most mathematicians spend most of their time in the Assimilator quadrant, whereas the students in a calculus class are likely to come from at least three quadrants. If our pedagogical strategies address only the students who are "like us," we are not likely to succeed in reaching all of them.

5 Messages from modern brain research

This is the Decade of the Brain, an exciting period of advances in neurobiology. This work builds on research with animal models and with epileptics after split-brain surgery, but the most exciting advances have come from imaging techniques — CAT, PET, MRI. We can now study functioning human brains for biological insights into the processes of reasoning, memory, and learning in the normal brain.

An important message of brain research for learning is "selection, not instruction" [4]. Evolutionary theory tells us that at birth we have our entire neural system — and it has not changed significantly in the last 10,000 years. Learning takes place by construction of neural networks. External challenges (sensory inputs) select certain neural connections to become active. Inputs enter the brain through old networks — there aren't any others. Each input can trigger memory if it is not new or learning if it is new. The cognitive term for this process is *constructivism*: The learner builds knowledge on what is already known, but only in response to a challenge. In particular, knowledge is not a commodity that can be transferred from knower to learner.

Selection also means that some potential neural pathways are *not* selected, that is, they become dormant through lack of use. The message for collegiate education: If we want to foster such skills as problem solving, creative thinking, and critical thinking, our task is much easier if educational challenges have been developing these skills from infancy. We have a stake in what happens at all levels before college.

Memory is an intricate collection of neural networks. Most experiences initially form relatively weak neural connections in "working memory," necessarily of short duration. The biochemical connections become stronger with use, weaker with disuse. The stabilized networks of long-term memory are accessed mainly by numerous connections to the emotional centers of the brain, but working memory has hardly any connections to the emotional brain. That is, working memory is not related to emotions — just facts — but formation of long-term memory strongly involves emotion [3], [7]. The message: We need to stimulate emotional connections to our subject matter if we expect it to transfer to long-term memory.

Similarly, there are strong connections between the emotional and rational centers in the brain. Indeed, emotional pathways can sometimes direct rational

decision making before the learner is consciously aware of the decision process. It's not hard to see the evolutionary connection here. Since all of these structures are 10,000 years old, they are intimately related to fight-or-flight reactions and other survival strategies [3].

Just as emotion is linked in the brain to learning, memory, and rationality, so are the motor centers of the brain, and by extension, the rest of the body. Body movement facilitates learning — sitting still inhibits learning [5].

We have already linked brain research to constructivism. Now we connect with Kolb's learning cycle. The concrete experience (CE) phase is input to the sensory cortex of the brain: hearing, seeing, touching, body movement. The reflection/observation (RO) phase is internal, mainly right-brain, producing context and relationship, which we need for understanding. Because the right brain is slower than the left, this takes time. The abstract conceptualization (AC) phase is left-brain activity, developing interpretations of our experiences and reflections. These are action plans, explanations to be tested. They place memories and reflections in logical patterns, and they trigger use of language. Finally, the active experimentation (AE) phase calls for external action, for use of the motor brain. Deep learning, based on understanding, is *whole brain* activity. Effective teaching must involve stimulation of all aspects of the learning cycle [12], [14].

6 Technology and learning

In the minds of many, "reform" is strongly associated with introduction of electronic technologies: graphing calculators, symbolic computer systems, the Internet. These technologies have become widely available, increasingly powerful, and increasingly affordable during the same decade as reform efforts. Is this good or bad or neutral for education? The short answer is "yes" — that is, use of technology is good or bad or neutral, depending on who's doing what. There is already an embarrassingly large literature addressing such questions as "Do students learn better with calculators (or Maple, or whatever)?", questions that are just as meaningless as they would have been for earlier technologies, such as blackboards, pencil and paper, slide rules, textbook graphics, or overhead projectors. There are also substantial numbers of thoughtful papers that compare particular classroom technology experiments with traditionally taught classes and measure whatever can be measured. The typical conclusion is that students in the experimental group did as well (or only slightly worse) on traditional skills, and they learned other things as well.

There are also *costs* associated with new technologies, just as there were with older technologies that we now take for granted. We don't know much about cost-effectiveness of new (or old) technologies, because we don't have good ways to measure *effectiveness* of education. Our effectiveness at addressing the goals in Section 2 may not be known until long after the students have left us, and maybe not even then. A more productive line of inquiry is to examine the costs of *not* using technology, in light of the current context of education, of reasonable projections about the world our students will live in, and of what we now know about learning.

Technology is a fact of life for our students — before, during, and after college. Most students entering college now have experience with a graphing calculator, and a growing percentage of students have computer experience as well. Many colleges require computer purchase or at least expect use of technology in a variety of courses. After graduation, it is virtually certain that, whatever the job is, there will be a computer close at hand. And there is no sign that increase in power or decrease in cost will slow down any time in the near future. We know these tools can be used stupidly or intelligently, and intelligent choices often require knowledge of mathematics, so this technological environment is our business. Since most of our curriculum was assembled in a pre-computer age, we need to rethink whether this curriculum still addresses the right issues in the right ways.

But calculus renewal is not primarily about whether we have been teaching the "right stuff." Rather, it is about what students are *learning* and how we can tell. To review, we have seen that the external world (employers) has certain expectations that turn out to be highly consistent with both learning theories and good practice. Neurobiologists have provided the biological basis for accepting sound learning theories and practices, while rejecting unsound ones. What does technology have to do with this?

Looking first at the Kolb cycle, we see that computers and calculators can facilitate the concrete experience (CE) and active experimentation (AE) phases — but *not* the other two phases, which are right brain and left brain activities. Thus, if the activity allows the student to go directly from CE to AE without engaging the brain, it may do more harm than good. Well designed learning activities usually involve the entire cycle. Technology can also support each of the Seven Principles.

7 Technology and curriculum

Developers of new curricula have found most of the traditional content still to be relevant, but not necessarily in the same order or with the same emphases or with the same allotment of time. Here is an example of how technology permits rethinking content and pedagogy in accord with sound theory and good practice.

The *raison d'etre* of calculus is differential equations. Never mind that most calculus students never get there — the interesting problems involve ODE's. Traditionally, understanding ODE's required lots of technique, and that in turn required practically all of Calculus I and II. Now we can pose the problem embodied in a differential equation on Day 1 of a calculus course: The time-rate of change of some important quantity has a certain form — what can we say about the time-evolution of the quantity? We can also draw a picture of the problem: a slope field. The meaning of solution is then clear: We seek a function whose graph fits the slope field. Even the essential content of the existence-uniqueness theorem is intuitively clear — the details can wait for that course in ODE's. By that time, the survivors will have a clear idea of what that course is going to be about and why the details matter.

To be more specific, suppose our question is "What can we say about growth of the human population, past, present, and future?" Students recognize that this

is important, and they start to engage with *ideas*. They can make conjectures about growth rates, such as proportionality to the population, and explore where they lead. They can trace solutions using the same technique as for the slope field: That's Euler's Method. Observing that human population is changing more or less "continuously," they are led naturally to the derivative concept and to what's "natural" about the natural exponential function.

There are many models students might pose for population growth, but we don't have to keep guessing. We have 1000 years of more or less reliable data to which we can fit a model. Using logarithmic graphing, we can find that the historic data are *not* exponential. Rather, the growth rate is proportional to the *square* of the population, so the data fit a hyperbola with a vertical asymptote — which occurs within their lifetime (about 2030). Then they really have to think about what all this means. (See [10], Chapter 7 Lab Reading.)

The details involve substantial mathematics — numerical, symbolic, and graphical. Note the echoes of the Kolb cycle: concrete experience with data plots, reflective observation about what the plots mean, abstraction in the symbolic models and their solutions, and active testing of the symbolic solutions against the reality of the data. Then the cycle starts again with the vertical asymptote: What does it mean? How can we fit it into an abstract scheme? How can we test whether our scheme fits with reality?

8 Renewal in calculus courses

It would be foolish to pretend that reformed calculus courses were designed to implement the messages of cognitive psychology or neurobiology. Few of the developers a decade ago had any knowledge of these subjects. Rather, we had some instinctive ideas about what to try. Some of those ideas were reinforced by our experiences and became the basis of our courses. Some ideas didn't work and were quickly forgotten. This is selection at work — but, in order for it to work, we had to challenge our prior knowledge.

Reformers became committed constructivists, even though few of us knew that word (in the cognitive sense). In varying degrees, we discovered empirically all seven principles of good practice. Our best materials encourage students to complete the learning cycle — often. Our best programs incorporate in some measure all seven of the skill groups identified by employers. And we have learned appropriate ways to use technology to serve learning objectives.

References

[1] A. P. Carnevale, L. J. Gainer, and A. S. Meltzer, *Workplace Basics: The Skills Employers Want*, The American Society for Training and Development and the U. S. Department of Labor, Washington, DC, 1988.

[2] A. W. Chickering and Z. F. Gamson, eds., *Applying the Seven Principles of Good Practice in Undergraduate Education*, New Directions for Teaching and Learning No. 47, Jossey-Bass Publishers, San Francisco, 1991.

[3] A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, G. P. Putnam's Sons, New York, 1994.

[4] M. S. Gazzaniga, *Nature's Mind: The Biological Roots of Thinking, Emotions, Sexuality, Language, and Intelligence*, BasicBooks, New York, 1992.

[5] C. Hannaford, *Smart Moves: Why Learning is Not All in Your Head*, Great Ocean Publishers, Arlington, VA, 1995.

[6] D. A. Kolb, I. M. Rubin, and J. M. McIntyre, eds., *Organizational Psychology: Readings on Human Behavior in Organizations*, 4th ed., Prentice-Hall, Englewood Cliffs, NJ, 1984.

[7] J. LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*, Simon & Schuster, New York, 1996.

[8] Pew Higher Education Roundtable, "A Teachable Moment," *Policy Perspectives* 8 (1) (June, 1998), 1-10, Institute for Research on Higher Education, Philadelphia, PA.

[9] A. W. Roberts, ed., *Calculus: The Dynamics of Change*, MAA Notes No. 39, Mathematical Association of America, Washington, DC, 1996.

[10] D. A. Smith and L. C. Moore, *Calculus: Modeling and Application*, Houghton Mifflin, Boston, 1996.

[11] A. Solow, ed., *Preparing for a New Calculus*, MAA Notes No. 36, Mathematical Association of America, Washington, DC, 1994.

[12] R. Sylwester, *A Celebration of Neurons: An Educator's Guide to the Human Brain*, Association for Supervision and Curriculum Development, Alexandria, VA, 1995.

[13] A. C. Tucker and J. R. C. Leitzel, eds., *Assessing Calculus Reform Efforts*, The Mathematical Association of America, Washington, DC, 1995.

[14] J. E. Zull, "The Brain, The Body, Learning, and Teaching," *National Teaching & Learning Forum* 7 (3) (1998), 1-5.

David A. Smith
Department of Mathematics
Duke University
Box 90320
Durham, NC 27708-0320, USA
das@math.duke.edu

Section 19

History of Mathematics

In case of several authors, Invited Speakers are marked with a *.

History of Mathematics in China: A Factor in World History and a Source for New Questions

Karine Chemla

Keywords and Phrases: Cultural history of mathematics from an international perspective

In the last decades much research has been devoted to the *Jiuzhang suanshu* or *The nine chapters on mathematical procedures* (hereafter abbreviated *The nine chapters*), a book which played a crucial role in the mathematical traditions written in Chinese characters, quite comparable to that of Euclid's *Elements* of geometry in the West. Compiled during the Han dynasty (206 B.C.E. – 221 C.E.), around the beginning of the common era, after the unification of the Chinese empire, the book was to become a "Classic" from which most subsequent Chinese mathematicians drew inspiration. It constitutes the earliest known Chinese source devoted to mathematics to have been handed down by a written tradition. With the discovery, in a grave, of a *Book on mathematical procedures* from the first half of the 2nd century B. C. E., archeologists have recently started to unearth documents that survived in an entirely different way. When they become available, we may expect our understanding of mathematics in early China to be radically changed, especially as regards the background of the composition of *The nine chapters* during the Han dynasty and the modalities of its compilation.

As with all other writings which were granted the status of "Classics" in China, commentaries were composed on *The nine chapters*, some of which were selected to be handed down together with the book. This is how commentaries ascribed to Liu Hui (third century) and Li Chunfeng (seventh century) survived until today.

This paper presents some recent observations on the book itself and its commentaries[1]. It then discusses how the mathematical results obtained in ancient China can be embedded in a world history of mathematics. The examples selected

[1]Since 1984, Professor Guo Shuchun and myself have been collaborating on a critical edition and a French translation of *The nine chapters* and its commentaries within the framework of an agreement between the Academia Sinica (Beijing, China) and the CNRS (France) *18. My ideas on the topic certainly benefited from this joint work, and I am pleased to express my gratitude towards Prof. Guo. Given the limits of this paper, I can unfortunately not do justice to all publications on the subject. The reader is referred to the bibliography in *18. I list below only critical editions of the text published recently *20, 23*, and the references for ideas sketched here. It is my pleasure to thank B. Belhoste, F. Bray, B. Chandler and J. Peiffer for very helpful discussions.

give various reasons why only an international approach to history of mathematics can provide an adequate framework to capture the historical processes which have constituted mathematical lores around the world. Finally, some new questions for the study of mathematical activity raised by research on *The nine chapters* are discussed.

I. Algorithms and their proofs in Early Imperial China

The nine chapters consist of problems and general algorithms with which to solve them. Their terms regularly evoke concrete questions with which the bureaucracy of the Han dynasty was faced, and, more precisely, questions that were the responsability of the "Grand Minister of Agriculture" (*dasinong*), such as remunerating civil-servants, managing granaries or enacting standard grain measures. Moreover, the sixth of *The nine chapters* takes its name from an economic measure actually advocated by a Grand Minister of Agriculture, Sang Hongyang (152-82 B.C.E.), to levy taxes in a fair way, a program for which the Classic provides mathematical procedures. These echoes between the duties of specific sectors of the bureaucracy and some of the mathematical problems tally with the fact that several scholars known in Han times for their ability in mathematics are also recorded as having at some point worked for this very administration. One of them, Geng Shouchang, is one of the two to whom Liu Hui's preface ascribes the composition of *The nine chapters*, whereas the other, Zhang Cang, also dealt with accounting and finance at high levels of the bureaucracy. Hence mathematics seem to have historically developed in Han dynasty China in relation with an administration in charge of economic matters *15. On another hand, some problems of *The nine chapters* were read by later scholars in ancient China to relate to astronomical questions *19. These practitionners hence identified within the book a reflection of an interaction between astronomy and mathematics, long stressed as crucial for the way in which the latter developed in China. Sources also record that both Zhang Cang and Geng Shouchang worked in astronomy.

However, the problems that evidence shows were quoted in the context of astronomical discussions may be perceived as recreational by some readers of today, because of the terms in which they are cast. The historian is thus warned against the assumption that the category of "mathematical problem" remained invariant in time, and is instead invited to describe the practice of problems with respect to which a text was written, before setting out to read it *16. In our case, despite the fact that *The nine chapters* usually present a problem within a particular concrete context, the first readers that we can observe, namely the commentators, read it as exemplifying a set of problems sharing a similar structure and solved by the same algorithm. They felt free to have a problem "circulate" between different contexts, without reformulating it either in other concrete terms or in abstract ones. Such a historical reconstruction guards us from mistaking a problem as merely particular or practical, when Chinese scholars read it as general and meaningful beyond its own context, or mistaking it as merely recreational when it was put to use in concrete situations. This is a crucial point, since it prevents us from jumping to the conclusion that mathematics in China was merely practical, simply because ancient Chinese texts attest to ways of managing the relationships between abstraction and

generality, between pure and practical mathematics, which are different from those we expect.

If a problem was not presented abstractly, that seemingly did not affect the value of generality attached to it. But the commentators expected that the algorithm given for solving it be general, if not abstract. For Liu Hui would criticize an algorithm provided, if it appeared to be less general than it could be and if it made use of inessential particular circumstances in the problem *16. In such cases, and within the framework of the same problem, he would restate a more general algorithm. Presumably, an algorithm's efficiency should extend as widely as possible beyond the scope of the problem for which it was formulated. Generality was thus expected for the operations rather than the situations themselves, and the commentators read the algorithm as determining the domain of problems which a particular one was exemplifying. Moreover, the Classic displays a rational architecture in nine chapters, based on the constitution of the algorithms, and not on the themes of the problems *12. This again highlights the authors' main emphasis on operations. *The nine chapters* thus articulate, within a theoretical framework, problems still bearing the marks of the contexts in which they were put to use or for which specific algorithms were developed. In the authors' opinion, the flavour of practice seemingly did not deprive theory of its glamour.

Mathematical knowledge was cast under the form of algorithms, for arithmetical as well as geometrical matters (computing the area of a circle or the volume of a pyramid). Inspired by Donald Knuth, who suggested reading Old-Babylonian clay-tablets from the point of view of algorithmic theory, Wu Wenjun initiated a new approach to ancient Chinese mathematical sources along similar lines *29. The properties that the algorithms in *The nine chapters* display confirm that they constituted a basis for mathematical effort. For instance, algorithms given for square and cube root extractions of integers and fractions bring into play the place-value decimal numeration system representing numbers on a counting board on which mathematics was practised: the sophistication of resources to which their description testifies – assignment of variables, conditionals, iterations – implies that lists of operations as such were compared, rewritten to be unified *1. This conclusion, drawn by observing how the algorithms are described, fits with what was noted above: an algorithm should be written so as to work for as many situations as possible. Moreover, should the algorithm not have exhausted the integer N when the units of the root are obtained, the Classic prescribed that the result must be given as "side of N", i. e. $\sqrt{N}$ *27, 22, 4.

Again, the algorithm to solve systems of n simultaneous linear equations with n unknowns, amounting to "Gauss elimination method" *22, 21*, puts into play a place-value notation for the equations on the counting board, and its description displays the same properties as listed above *9. It brings in marked numbers ("positive", "negative") and "missing" coefficients, as well as rules for computing with them, to achieve the utmost efficiency *7. First introduced in the flow of the computations, such numbers were then reused to represent any linear equation on the board and have the algorithm cover all possible such systems of linear equations. This way of instituting the general linear equation evokes how quadratic equations appear in *The nine chapters*: the algorithm for square root equation

deprived of a first step, as well as the state of the counting board at this point of the computation, were granted autonomy, the latter yielding the concept of quadratic equation, the former, the algorithm to compute "its" root *7. Both cases attest to the same specific way of defining new objects: algorithms operate on configurations of numbers on the board, and, in both cases, some of their temporary states as well as the part of the algorithm flowing from them received the status of autonomous mathematical objets. Algebraic equations were to develop in China in that way, exclusively as numerical operations depending on n-th root extraction, until the 13th century.

The positive and negative numbers introduced, however, differ in nature from the quadratic irrationals mentioned above: the results could not be such marked numbers, which also betrays that they differ from the modern concepts. They functioned rather as algorithmic marks, exclusively within the context of systems of linear equations, and it was as such that in the 13th century they were exported into a second mathematical domain: used to represent the coefficients of any algebraic equation, they provided the basis for extending the Ruffini-Horner algorithm to obtain "the" root in the most general case *7. The treatment of algebraic equations was thus completed within the framework in which these equations had appeared in *The nine chapters*.

A last group of algorithms, the "rules of false double position", which have disappeared from today's mathematics, betray in yet another way the Classic's interest in algorithms encompassing the widest range of situations possible. A common list of operations is obtained to solve problems of two intrinsincally different types. It takes different meanings when applied to these different cases, but formal identity of the solving procedures served as a basis for a unique algorithm in *The nine chapters*. Again the result of a formal work on operations themselves, such a property epitomizes the development, within this algorithmic framework, of a kind of algebra *3.

In contrast with *The nine chapters* themselves, the commentaries explicitly set out to prove the correctness of the algorithms provided by the Classic. Since they systematically dealt with algorithms, their proofs developed within a context differing from what can be found in Greek texts of Antiquity, where mathematicians addressed establishing the truth of statements. The description of these proofs, besides acquainting us further with the conception of algorithms in ancient China, brings to light what constituted an original practice of proof *3, 14. When proving that the given algorithm for the area of the circle or the volume of the pyramid is correct, Liu Hui brings into play infinitesimal reasonings, using inscribed polygons for the circle, and smaller and smaller similar solids for the pyramid *28, 22, 21. Their detailed structural similarity indicates that these reasonings may have been ruled by patterns or fulfilled constraints *10. Concluding his proof that the algorithm for the circle (i.e. "multiplying half the circumference by half the diameter, one obtains the area of the circle") works, Liu Hui stresses that this algorithm, correct when it involves the actual dimensions of the circle, allows no computation. This singular situation induces him to make explicit a distinction crucial for us to understand how an algorithm was conceived, since he contrasts the algorithm as *prescription for computation* –to produce a value–, from the algorithm as *relation*

of transformation between magnitudes, essential for the proofs *10. As regards the area of the circle, where the two do not run in parallel, this causes a division in the proof. Liu Hui first addresses the latter aspect, before turning to the former and examining how computations can provide approximations. More generally, the problem after which an algorithm is stated offers a context of interpretation of its operations as relations of transformation, which the commentators may bring into play in the proof *16. In the course of proving that an algorithm works, problems occur in another way. Willing to establish that an algorithm actually yields the sought-for unknown, Liu Hui may first himself produce a list of computations performing the same task as follows: he decomposes it into a sequence of auxiliary tasks in which he recognizes known problems and concatenates the algorithms for their solution. The second part of the proof then consists in transforming the algorithm obtained into equivalent ones, until he gets to the algorithm he was originally considering. To this end, Liu Hui applies rules of rewriting to lists of operations, which include: deleting inverse operations such as division and multiplication; reversing the order of operations; merging multiplications and divisions together; inverting algorithms. This kind of formal transformations, operating on an algorithm as such, attests to the development of a form of *algebraic proof again within an algorithmic framework*. The key point here is that Liu Hui relates the validity of this form of proof to the fact that various kinds of numbers were introduced by the Classic (fractions and quadratic irrationals) to provide divisions and root extractions with exact results *17. This makes the opposition between multiplication and division operate with full generality and efficiency in mathematics *12. The interest in pairs of opposed but complementary operations echoes the numerous quotations from the *Yijing (Classic of changes)* in the commentaries. Considering, further, that Liu Hui refers to algorithms – the core of mathematical activity – as embodying change (*bianhua*) within mathematics, we may conjecture that philosophical inquiries into change in ancient China influenced mathematical research or benefited from meditating on mathematics *14, 15.

II. A Factor in World History

Embedding Chinese sources in the world corpus of mathematical writings discloses that their authors shared topics of interest and results with other communities on the planet. This raises various kinds of question. *The nine chapters* share with the earliest extant Indian mathematical writing (6th c.) basic common knowledge, among which is the use of a place-value decimal numeration system. Such evidence allows no conclusion as to where this knowledge originated, a question which the state of the remaining sources may prevent us from ever answering. Instead, it suggests that, from early on, communities practising mathematics in both areas must have established substantial communication.

Later on, Arab scholars became interested in this scientific world through India. This is documented. However, several elements common to Chinese and Arabic sources from the 9th century onwards, and so far not found in the known Indian sources, seem to indicate that there were also direct contacts between the Chinese- and Arabic-speaking intellectual communities. One of these, the topic of a treatise by Qusta ibn-Luqa (9th c.) before spreading westwards, was the set

of rules of false double position. Interacting with their new intellectual context, these rules were proved with Euclidean geometry, which required drawing a diagram representing the relation between what goes in and out of an algorithm *13. Some centuries later, several similarities occur between Chinese and Arabic sources. As-Samawa'l (1172) extracted a 5-th root with a Ruffini-Horner algorithm *25*, as Jia Xian (11th c.) did *6*, and considered polynomials written in a place-value notation for the powers of the indeterminate, which is similar to the notation for polynomials found in sources from Northern China in the 13th cent. *On equations*, by Sharaf-al-Din al Tusi (12th c.), articulates improving approaches to quadratic and cubic equations previously developed in Arabic with finding roots using tabular, numerical algorithms cognate to those traditionally used in China *8* and later to be used by Viète *24. Of course, the earliest evidence available today proves nothing about where a result was obtained. Such a conclusion might be contradicted by finding new manuscripts. However, another type of conclusion can more safely be drawn: some Chinese and Arabic mathematical communities must have been in close enough contact to share a whole group of results *6. These contacts were probably not very intimate, since we have no evidence that Euclidean geometry as widely practised then in the Arab world received mathematicians' attention in China before the arrival of European missionaries at the end of 16th century. Conversely, we so far have found no echo in Arabic sources of the algorithms for solving systems of linear equations which were continuously used in China.

The history of algebraic equations, however, raises many general issues other than the question of "transmission". The sources prior to Tusi's *On equations* in which we recognize such equations and modes of resolution, be they Babylonian, Greek, Chinese, Indian, or Arabic, attest in fact to different concepts and practices, presenting, despite the transformations they underwent, stable features over long periods of time *8. Hence different mathematical traditions elaborated in diverse ways an object that today's readers recognize as the same. The description of these different elaborations, all the more precise when it involves comparing the various treatments to distinguish them, displays the conceptual variety likely to affect what we would conceive of as a unique mathematical object. Considering these sources as a whole, we also see that the approach to equations devised in China can be found in no other corpus of ancient texts. As a result, this gives us a precious piece of historical information which enables us to tackle questions of transmission with greater precision. In another respect, some of these sources display concepts of equation that in turn become ingredients that other sources articulate in their own treatment of equation. For instance, Tusi inherited al-Khwarizmi's theory of quadratic equations (9th century), itself a framework based on blending two different ingredients: Babylonian algorithms solving particular equations by radicals and Diophantos (ca. 2nd c.)' *Arithmetics'* handling of equations as statements of equality involving an unknown. Onto this, Tusi articulated Khayyam (11th c.)'s geometrical theory of cubic equations – an elaboration merging al-Khwarizmi's concept and framework for equations with Greek approaches using conics to problems only later conceived of as equations *24* – and numerical algorithms echoing with Chinese sources. Tusi's *On equations* attests to new

developments concerning equations not only because it systematically provided the numerical algorithms with proofs and conceptually improved the geometrical approach to equations *26*, but also because it bears witness to a synthesis of different concepts and approaches to equations. The mathematical work required to perform this synthesis also needs to be stressed and studied for itself as, more generally, one kind of the processes forming mathematical knowledge *5. It demanded that mathematical bridges be built between different concepts, thus the concepts were melted into a unique one. Such a work, however, may become invisible today to those who inherited concepts depending on this synthesis. Cumulative progress is by no means the only process accounting for the constitution of mathematical knowledge. Non-linear processes took place, the study of which requires that all traditions be taken into account and that the old and yet too widespread global framework of the history of mathematics, drawing a line between the Greeks and the so-called "Renaissance", be revised. The new picture may well bring to light the crucial part played by Arabic-speaking mathematicians of the Middle Ages in bringing together traditions from everywhere, carrying out synthesis of this type and elaborating on them, thereby changing the nature of mathematics *5.

The rules of false double position show the limits of an account in terms of cumulative progress from another perspective. If Western sources in which they could be found after the 9th century enriched them with proofs in the Euclidean manner, they lost the subtelty of *The nine chapters*, since they no longer presented algorithms able to solve problems of two kinds. In fact, these Western sources (Arabic writings and European commercial arithmetics of the Middle Ages) contained only problems of one kind. However, transmission was not smoother in China, where, at the end of 16th century, after a period of mathematical decline, the algorithms were used only for the second kind of problems. When Jesuits brought European mathematical writings to China, Chinese scholars found themselves confronted with cognate algorithms applied to different kinds of problem and coming from two different sources, and could not figure out their relation. They gathered them together again and concluded that Western mathematics was superior *13, 9!

III. New questions

Even though *The nine chapters* contain concepts and results, the international circulation of which shows their potential universality, the Classic adheres in several ways to the local cultural contexts within which it was produced and handed down. We saw above how the emphasis placed on algorithms and on opposed operations might relate to a wider interest in change. In another respect, the status of "Classic" granted to the book refers to a category of writings typical of the history of Chinese literature. It implies that the book was to be treated in a special way and called for peculiar modes of reading from its commentators. In yet another respect, specific literary techniques were used in writing *The nine chapters*: The algorithms for square and cube root extraction were described, sentence by sentence, in parallel with each other, thus bringing to light the ways in which the operations could be considered as analogous to each other. This relates to the fact that, more generally, Chinese texts abound in such parallel sentences which correspond to each other character by character and are read as expressing

a correspondence between their topics *2. Chinese mathematical writings hence demonstrate that they stick to, and benefit from, a given set of common scholarly practices. The historian must take into account the adherence to scholarly practices in order to interpret mathematical texts in a better way. Conversely, because of the characteristics of the subject, mathematical writings could provide a useful observatory to describe widespread scholarly practices *2, 11, 16. This delineates a research program for the study of mathematics as part of a wider cultural context.

These facts cast light on a more general phenomenon: *The nine chapters* attest to a specific concrete work environment for practising mathematics. Ways of dealing with problems and algorithms, of using visual aids, of handling the counting board also demonstrate specific features in the way mathematicians used them in ancient China. On another hand, characteristic interests (in algorithms with an emphasis on generality, or in opposed operations) coalesce with specific features of the mathematical objects (recurring place-value notations *11*, singular concepts like equations as numerical operations, designed as a temporary state in a flow of computation) to form the image of a particular mathematical world. This raises two kinds of problems. How was the work environment designed and used to pursue specific interests? How far can we correlate the questions raised and the means designed to tackle them with the specificities of the notations, concepts and results produced? The case of the counting board yields interesting elements in both respects. A surface handled according to strict and particular rules, this board offers positions where numbers can be placed and transformed in order to carry out computations. For example, multiplication and division both require three positions (top, middle, bottom). The key point is that these positions appear to be subjected to either opposed or the same sequences of events when multiplying or dividing *12. A kind of object, consisting of a position and the sequence of events to which it is subjected in the concrete course of a computation, draws our attention. This kind of object enabled one to display in a specific way the opposition between operations on the board. But it is also involved in working out the similarity between square and cube root extractions, or root extractions and division *2. There, the same names are conferred to positions affected by similar sequences of events: the nature of some basic concepts seems to originate from observing mathematical reality as given shape through computations on the board. The Ruffini-Horner algorithms found in 11th century China make sense in relation to this context too: this way of carrying out root extraction reduces the algorithm to repeating, on a succession of positions, the same sequences of events as found in either a division or a multiplication *6. This again emphasizes the interest in finding out a list of operations the applicability of which is as broad as possible. Moreover such positions enable root extraction to appear as composed of alternately opposed operations, and they eventually formed the place-value notation for algebraic equations and polynomials as they appear in 13th century texts *11. The hypothesis that algorithms were worked out on the board through such objects as positions and the sequence of events on them thus ties together features which we underlined: the interest in generality of the algorithms, in opposed operations, in results such as Ruffini-Horner algorithms, and the recurring of place-value notations over centuries *11. It links the specific practice of math-

ematics to the concepts and results on which mathematicians focused in ancient China. However, this is not peculiar to China: more generally, the products of mathematical activity that eventually become universal are worked out by resorting to specific forms of practice, partly inherited, partly reworked according to the problems addressed or the conditions available. Clearly, describing such regimes of mathematical activity and their relations to the mathematical results produced is an agenda for which ancient China provides a unique contribution.

Elements of bibliography

1. Chemla, K., "Should they read FORTRAN as if it were English ?", *Bulletin of Chinese Studies*, 1, 1987, 301–16.
2. — "Qu'apporte la prise en compte du parallélisme dans l'étude de textes mathématiques chinois ? Du travail de l'historien à l'histoire du travail", *Extrême-Orient, Extrême-Occident*, 11, 1989, 53–80.
3. — "Theoretical Aspects of the Chinese Algorithmic Tradition (first to third century)", *Historia Scientiarum*, 42, 1991, 75-98 +errata in the following issue.
4. — "Des irrationnels en Chine entre le premier et le troisième siècle", *Revue d'histoire des sciences, XLV*, 1992, 135–40.
5. — "De la synthèse comme moment dans l'histoire des mathématiques", *Diogène*, 160, 1992, 97–114.
6. — "Similarities between Chinese and Arabic Mathematical Writings (I) : root extraction", *Arabic Sciences and Philosophy*, 4, 1994, 207–66.
7. — "Different Concepts of Equations in *The Nine Chapters on Mathematical Procedures* and in the Commentary on it by Liu Hui (3rd century)", *Historia Scientiarum*, 4, 1994, 113–37.
8. — "Algebraic Equations East and West until the Middle Ages", in K. Hashimoto, C. Jami, L. Skar (eds), *East Asian Science: Tradition and Beyond*, Kansai Univ. Press, 1995, 83–9.
9. — "Que signifie l'expression 'mathématiques européennes' vue de Chine ?", in C. Goldstein, J. Gray, J. Ritter (éds.), *L'europe mathématique. Histoires, Mythes, Identités. Mathematical Europe. History, Myth, Identity*, Editions de la MSH, 1996, 220–45.
10. — "Relations between procedure and demonstration. Measuring the circle in the *Nine Chapters on Mathematical Procedures* and their commentary by Liu Hui (3rd century)", in H. N. Jahnke, N. Knoche & M. Otte (eds.), *History of Mathematics and Education: Ideas and Experiences*, Vandenhoeck & Ruprecht, 1996, 69–112.
11. — "Positions et changements en mathématiques à partir de textes chinois des dynasties Han à Song-Yuan. Quelques remarques", *Extrême-Orient, Extrême-Occident*, 18, 1996, 115–47.
12. — "Le jeu d'opérations opposées mais complémentaires dans les textes mathématiques chinois anciens. Premières remarques", in Siegmund Probst, K. Chemla, Agnès Erdély, Antonio Moretto, *Ceci n'est pas un festschrift pour Imre Toth*, 29-12-1996.
13. — "Reflections on the world-wide history of the rule of false double position, or: how a loop was closed", *Centaurus*, 39, 1997, 97–120.

14. — "What is at Stake in Mathematical Proofs from Third Century China?", *Science in Context*, 10, 1997, 227–51.
15. — "Croisements entre réflexion sur le changement et pratique des mathématiques en Chine ancienne", in J. Gernet et M. Kalinowski, *En suivant la voie royale*, Presses de l'EFEO, 1997, 191–205.
16. — "Qu'est-ce qu'un problème dans la tradition mathématique de la Chine ancienne ?", *Extrême-Orient, Extrême-Occident*, 19, 1997, 91–126.
17. — "Fractions and irrationals between algorithm and proof in ancient China", *Studies in History of Medicine and Science*, (forthcoming).
18. Chemla, K. and Guo Shuchun, *The nine chapters on mathematical procedures*. A Critical edition, translation and presentation of the classic and the commentaries on it ascribed to Liu Hui (third century) and Li Chunfeng (7th century), Diderot multimedia (to appear).
19. Eberhard, A., *Re-création d'un concept mathématique dans le discours chinois : Les "séries" du Ier au XIXe siècle*, Ph. D., University Paris 7 & TU Berlin, 2 volumes.
20. Guo Shuchun, Jiuzhang suanshu *huijiao*, Liaoning jiaoyu chubanshe, 1990, 550 p.
21. — *Gudai shijie shuxue taidou Liu Hui*, Shandong kexue jishu chubanshe, 1992, 468 p.
22. Li Jimin, *Dongfang shuxue dianji* Jiuzhang suanshu *ji qi Liu Hui zhu yanjiu*, Shaanxi renmin jiaoyu chubanshe, 1990, 492 p.
23. — Jiuzhang suanshu *jiaozheng*, Shaanxi kexue chubanshe, 1993, 590 p.
24. Rashed, Roshdi, "Résolution des équations numériques et algèbre : Saraf-al-Din al Tusi, Viète", *Archive for History of Exact Sciences*, 12, 1974, 244–290.
25. — "L'Extraction de la Racine $n^{ième}$ et l'Invention des Fractions Décimales (XIe–XIIe Siècles)", *Archive for History of Exact Sciences*, 18, 1978, 191–243.
26. — *Sharaf al-Din al-Tusi : Oeuvres mathématiques. Algèbre et géométrie au XII^o siècle*, 2 volumes, Les Belles Lettres, 1986.
27. Volkov, Alexei K., "Ob odnom driévniékitaïskom matiématitchiéskom tiérminié", *Tiézisy konfiériéntsii aspirantov i molodykh naoutchnykh sotroudnikov IV AN SSSR*, Nauka Press, volume 1.1, 1985, 18–22 (in Russian).
28. Wagner, Donald, "An Early Chinese Derivation of the Volume of a Pyramid : Liu Hui, Third Century A.D.", *Historia Mathematica*, 6, 1979, 164–88.
29. Wu Wenjun, "Recent Studies of the History of Chinese Mathematics", *Proceedings of the International Congress of Mathematicians, Berkeley*, California, USA, 1986, 1657–67.

Karine Chemla
3 Square Bolivar
F-75019 Paris
France
chemla@paris7.jussieu.fr

Marx, Mao and Mathematics: The Politics of Infinitesimals

Joseph W. Dauben

Abstract. The "Mathematical Manuscripts" of Karl Marx were first published (in part) in Russian in 1933, along with an analysis by S. A. Yanovskaya. Friedrich Engels was the first to call attention to the existence of these manuscripts in the preface to his Anti-Dühring [1885]. A more definitive edition of the "Manuscripts" was eventually published, under the direction of Yanovskaya, in 1968, and subsequently numerous translations have also appeared. Marx was interested in mathematics primarily because of its relation to his ideas on political economy, but he also saw the idea of variable magnitude as directly related to dialectical processes in nature. He regarded questions about the foundations of the differential calculus as a "touchstone of the application of the method of materialist dialectics to mathematics." Nearly a century later, Chinese mathematicians explicitly linked Marxist ideology and the foundations of mathematics through a new program interpreting calculus in terms of nonstandard analysis. During the Cultural Revolution (1966–1976), mathematics was suspect for being too abstract, aloof from the concerns of the common man and the struggle to meet the basic needs of daily life in a still largely agrarian society. But during the Cultural Revolution, when Chinese mathematicians discovered the mathematical manuscripts of Karl Marx, these seemed to offer fresh grounds for justifying abstract mathematics, especially concern for foundations and critical evaluation of the calculus. At least one study group in the Department of Mathematics at Chekiang Teachers College issued its own account of "The Brilliant Victory of Dialectics - Notes on Studying Marx's 'Mathematical Manuscripts'." Inspired by nonstandard analysis, introduced by Abraham Robinson only a few years previously, some Chinese mathematicians adapted the model Marx had laid down a century earlier in analyzing the calculus, and especially the nature of infinitesimals in mathematics, from a Marxist perspective. But they did so with new technical tools available thanks to Robinson but unknown to Marx when he began to study the calculus in the 1860s. As a result, considerable interest in nonstandard analysis has developed subsequently in China, and almost immediately after the Cultural Revolution was officially over in 1976, the first all-China conference on nonstandard analysis was held in Xinxiang, Henan Province, in 1978.

CHINESE VERSIONS OF THE "MATHEMATICAL MANUSCRIPTS" OF KARL MARX

There were two editorial groups working in the early 1970s on Chinese translations of Marx's "Mathematical Manuscripts," one in Shanghai, the other in Beijing; the Shanghai group was the first to publish trial editions and then excerpts of Marx's "Mathematical Manuscripts." Working initially from a Japanese translation, the "Fu Dan University Scientific Reference Section" (复旦大学理科资料组) completed a first draft which was circulated for discussion in 1971. Two years later, with a copy of the Russian-German edition in hand (which provided transcriptions of the original manuscripts in German), a revised trial edition was printed and in 1974, translations of Marx's essays on derivatives, differentials, and the history of the calculus were published in two successive issues of the Shanghai journal, *Dialectics of Nature* (自然辩证法). A year later, the entire translation appeared as a special edition (专辑) of the *Journal of Fu Dan University* (复旦学报), along with a brief "Remark on the Translation" (译者的话). Meanwhile, in the same year that the Shanghai edition of the manuscripts was printed, a study group at Beijing University published its own translation of three of Marx's essays on the history of the differential calculus, interpreted specifically within a Marxist framework as a "stage in the development of history." When these appeared in the *Acta Mathematica Sinica* in 1975, they were preceded by a half-page of explanatory remarks from the "main editorial committee," wherein it was emphasized that this was a proletarian work, published by the People's Press (人民出版社), and meant to contribute to the socialist revolution and to socialist reconstruction:

> To promote the great campaign criticizing Lin Biao (林彪) and Confucius, the *Mathematical Manuscripts* of [Karl] Marx, who inspired the proletarian revolution, were translated and edited by the Mathematical Manuscripts Study Group of Beijing University, and published by the People's Press (人民出版社). This is a great event on our ideological battlefield.
>
> Lenin pointed out that "with material dialectics to improve essentially the entire political economy, using dialectical materialism to elucidate history, natural science, philosophy, and the policies and strategies of the working class is the most important thing of concern to Marx and Engels, whereby they made their most important and novel contributions, and brilliantly took a giant step in revolutionary intellectual history."

Marx, the preface points out, used dialectical materialism to evaluate the history of the calculus, and was especially critical of what he took to be its idealistic, metaphysical foundations. Chairman Mao himself emphasized repeatedly that dialectics was the key to proper understanding of the sciences. Dialectical materialism was the weapon, literally, that Mao expected Chinese revisionists to use-even revisionist mathematicians-to root out any bourgeois elements and advance mathematics down "Chairman Mao's revolutionary route." Mathematicians thus took their publication of the mathematical manuscripts of Karl Marx as the perfect blueprint showing how their own criticism of mathematics should proceed:

> The great leader, Chairman Mao, has written that "you who study the natural sciences should learn how to use dialectics." By studying Marx's *Mathematical Manuscripts*, our theoretical understanding will reach a higher level,

> and will help us to take hold of the perfect weapons, advancing criticism of revisionism and of bourgeois world outlooks, [thereby] joining the battlefield with Marxism. People who study or teach mathematics should study and use dialectical materialism, which is clarified in the *Mathematical Manuscripts* of [Karl] Marx, to guide their practice and conscientiously improve their world outlooks, pushing the study of mathematics very quickly along Chairman Mao's revolutionary route, making a greater contribution to the socialist revolution and socialist construction.

Within months the Beijing University study group was satisfied that its entire translation was ready for publication, and in July of 1975 issued its definitive edition which included photocopies of several pages from Marx's original manuscripts. Part II reproduced verbatim the sections already issued previously that year. Although the Beijing translation differs in choice of words from time to time from the Shanghai translation, what sets the Beijing edition apart is its inclusion of explanatory terms from the original German version from which the Beijing translation was made. For example, terms like *"Differentiation," "abgeleitete Funktion,"* and *"Grenzwert"* appear, parenthetically, to explain Chinese terminology when new terms/characters are first introduced.

FIRST REACTIONS TO PUBLICATION OF THE MATHEMATICAL MANUSCRIPTS

No sooner had the first two parts of the translation of the manuscripts by the Shanghai group appeared in print than the editors of the *Journal of the Dialectics* of Nature (自然辩证法杂志) began to receive letters from a wide variety of readers. The next number of the journal to appear contained a selection of these letters in a section entitled "Discussion of Problems Concerning Differentials and Limits" (关于微积分和极限问题的讨论). This began with a note from the editors explaining all of the mail the journal had received. Several letters were then published in their entirety, with excerpts from a number of others. The first letter was from a second-year student at Beijing Middle School No. 144, He Fang (何放), who asked "How Should the Concept of Limit be Understood?" (应当怎样认识极限?). The next contribution was from a worker at Factory No. 5703 in Shanghai, Fu Xi-tao (傅锡涛), who was interested in: "Trying to Say Something Concerning my Feelings About Improving Teaching of the Calculus Using Dialectics" (试用辩证法改革微积分教学的一点体会). Fu Xi-tao explained how dialectics could be applied to reform calculus teaching. A third letter came from Zheng Li-xing (郑礼星) of the Fujian Electrical Engineering School (福建机电学校) in Fuzhou, Fujian Province. Zheng took up one side of the debate over whether the differential dx was zero or not, arguing: "The Differential is Comparable to Zero," (微分是相对的零).

Along with their publication of "Selections from Manuscripts Received" (来稿摘登), the editors of *Dialectics of Nature* included excerpts from letters by readers who had studied the translation of the mathematical manuscripts published in the preceding two issues of the journal. The first was taken from a letter by Xu Ting-dong (徐庭栋), who identified himself as a young worker in the Qing-Hai Tractor Factory (西宁青海拖拉机厂). His comments were devoted to "The Differential is a Unity of Zero and Non-Zero" (微分是零和非零的统一),

and drew on similar dialectical criticism of the foundations of the calculus already raised by Zheng Li-xing. But Xu Ting-dong also considered the calculus applied to motion, and was especially interested in discussing acceleration and the derivative. The next letter, attributed to Wu Guang-xia(吳光夏) of Bao Tou Teachers School (包头师范学校) in Inner Mongolia, was also concerned with the zero/non-zero aspect of the differential. Another letter along these same lines came from Chen Ke-jian (陈克艰), a "knowledgeable youth" (知识青年) from Shang Shan Xia Xiang(上山下乡). Again, his analysis was devoted to considering the differential as "zero" and "non-zero," interpreting the calculus as it applied to motion and the paradoxes that arise from trying to consider a moving point as being in any "one" place.

From Harbin Industrial University (哈尔滨工业大学), Shen Tian-ji (沈天骥) wrote to suggest that "The Differential Reflects Quantitative Change from (Two) Different Points of View" (微分反映量变在不同层次的关节点). Here the two different points of view were of Δx versus δx, and the difference between non-zero and zero, as well as the meaning of $\delta y = f(x)\Delta x$. The last letter in this collection of differing points of view prompted by publication of the "Mathematical Manuscripts" was from a young worker at a Shanghai machine packing plant, Chen Li-qin (陈利钦), who insisted that "The Differential Must be Considered as Zero" (微分应当归结为零). Chen's argument was based on his understanding of the limit: $\lim_{\Delta x \to 0} = \dfrac{\Delta y}{\Delta x} = \dfrac{dy}{dy} = f'(x)$.

MATHEMATICIANS BEGIN TO RESPOND

Thus in 1975, two definitive editions of Marx's "Mathematical Manuscripts" appeared in Chinese. The Beijing edition differed only slightly from the Shanghai version, and in some cases they paralleled each other *verbatim* in the Chinese. But with the entire collection of Marx's "Mathematical Manuscripts" now at their disposal, it was not only high school students and factory workers who took an interest, but so too professional mathematicians. For example, writing in the *Journal of Beijing Normal University*, Zhi Zhou (郅洲) of the Philosophy Department explained "How to Understand Derivatives (导函数)— Notes on studying Marx's Mathematical Manuscripts" (怎样理解导函数— 学习马克思≪数学手稿≫笔记). In his introduction, Zhi Zhou explained that whereas the calculus as a scientific subject came into being at the end of the 17th century, it did not develop into a satisfactory theory until the middle of the 19th century. As a result of work done in the 17th and 18th centuries, when metaphysical concepts dominated the natural sciences, the fundamental concept of the calculus, namely the derivative, was also subjected to strong metaphysical influences. In the 1870s, the revolutionary teachings of Marx severely criticized such metaphysical foundations for the derivative, and advocated a correct interpretation on the basis of dialectical materialism. In the first section of his paper, Zhi Zhou examined the history of the derivative, and explained how Newton and Leibniz had introduced the concept as a ratio of differentials. He added that according to Bishop Berkeley (大主教), a representative of "English subjective idealism" (英国的主观唯心主义), the differential dx was literally, in Chinese,

exactly what Berkeley had said it was in English, "the ghost of a departed quantity" (逝去量的鬼魂) [Zhi Zhou 1975, p. 19]. Zhi Zhou also considered both d'Alembert's approach to the calculus in terms of differences Δt and differentials dt, as well as Lagrange's approach expanding functions in terms of their Taylor series, for which the derivative was taken as the coefficient of the linear term of the infinitesimal h, i.e.:

$$f(x+h) = f(x) + hp_1(x) + h^2p_2(x)/2! + h^3p_3(x)/3! + \ldots$$

where $p_1(x)$ was taken to define the derivative, i.e. $f'(x) = p_1(x)$. But whereas Marx stopped with his analysis of the historical development of the calculus at this point, Zhi Zhou went further to consider the contributions made by the French mathematician Cauchy (哥西), specifically the definition of the limit that Cauchy gave in his *Cours d'analyse* of 1821 (分析教程). This was all discussed expressly in terms that Engels had used in 1830 in his *Dialectics of Nature*. Although Zhi Zhou was aware of the fact that Marx never expressly mentioned Cauchy by name, he could not believe that Marx was unaware of the basic ideas used by Cauchy, since Cauchy's point of view was represented in many of the most popular scientific books of his day. Zhi Zhou, who asks why Cauchy did not permit the variable x to actually reach or attain the limit $x = 0$, explained that it was because he feared this would lead to the "monster" (怪物) $0/0$. He notes that later, in the 1850s, the $\epsilon-\delta$ method of proof appeared, and a few decades later, in the 1870s, this was linked to a thorough critique of the real numbers (实数理论). Nevertheless, the first person to "strip away the appearances" and submit the concept of the derivative (导函数) to a thorough metaphysical analysis was, in Zhi Zhou's opinion, none other than Karl Marx. Zhi Zhou devoted the second part of his paper to describing Marx's analysis of the derivative, especially the differential quotient dy/dx in terms of the paradoxical nature of $0/0$, which strictly speaking was undefined, or could represent any value at all. He concluded his essay by returning to the founders of the calculus, to Newton and Leibniz. The characterization was classic Marxism: "Newton's and Leibniz's contributions to the calculus, are great pioneering works in the development of mathematics, but due to the constraints of metaphysical ideology, their works could not avoid being colored by mysticism (神秘主义)." From Newton until the time of Marx, although the calculus underwent considerable development, and despite the fact that the concept of the derivative also had a rich dialectical context, it was still trapped in a web of metaphysical ideology. Owing to the constraints of metaphysics, and even though they raised their voices against "old fashioned orthodox schools of thought," mathematicians could find no alternatives:

> (Our) revolutionary leader Marx, because of his deep grasp of the method of dialectical materialism, thus focused on the idea of the derivative and advanced a series of brilliant dialectical thoughts, even though over the past 200 years mathematicians have been working but have not yet been able to make a great contribution.
>
> Marx's "Mathematical Manuscripts" are one part of a brilliant, monumental mathematical work, and are a precious scientific legacy Marx has left to us. It is not only part of the mathematical writings, but is also a part

> of his philosophy which uses the methods of dialectics as a model for studying mathematics. Teaching and studying the "mathematical manuscripts" is necessary for today's revolution in education, and is needed in the battle to conquer mathematics.
>
> Scientists and especially mathematicians, from the study and research of the "Mathematical Manuscripts," [have] a powerful ideological weapon to transform directly the old mathematical system and to reform the study and teaching of mathematics.

JOURNAL OF FU DAN UNIVERSITY (复旦学报)—1975

Meanwhile, in Shanghai the editors of the *Journal of Fu Dan University* continued to publish new manuscripts submitted in the wake of their publication of the *Mathematical Manuscripts of Karl Marx*. In the second number of 1975, for example, Ou Yang Guang-zhong (欧阳光中) and Zhu Xue-yan (朱学炎) offered a "Discussion on some Ways of Looking at the Calculus of Functions of Several Variables" (谈谈对多元函数微积分的一些看法). Ou Yang was a prominent mathematician who published a considerable amount during the Cultural Revolution; the article begins with strong praise for Marx:

> One hundred years ago the great revolutionary teacher Marx wrote his mathematical manuscripts, and although in the course of these hundred years mathematics has undergone tremendous development, Marx's mathematical manuscripts nevertheless still shine with a brilliant radiance. Marx in his mathematical manuscripts used the special materialist dialectics of Marxism to criticize every shade of idealist metaphysics, tearing the mysterious veil from the deceptive derivatives and differentials, and bringing to light their true essence, thereby setting a brilliant example for us.

In addition to political rhetoric, Ou Yang provided some sophisticated mathematics as well. This was more technical than anything written to this point in connection with the mathematical manuscripts of Karl Marx, for Ou Yang considered vector analysis, potential differences, gradients, the Poisson integral, triple integrals, and a host of related subjects. On a more elementary level, in the next issue of the *Journal of Fu Dan University*, a mathematician by the name of Shu Zuo (舒左) offered a paper meant to serve as "A Starting Point for Calculating with Differentials," (微分演算的出发). Here the name "Shu Zuo" was not only a pseudonym, but also a play on words and characters, for the characters 舒左 ("Shu Zuo") literally mean "Unfold the Left," but with a slightly different change in tone, the characters 数作 (also "Shu Zuo") mean "Do Mathematics":

> In the midst of the movement to study the theory of the dictatorship of the proletariat now surging forward with great momentum, publication of the translation of Marx's *Mathematical Manuscripts* is of great significance. "The proletariat must include in its superstructure (上层建筑) all areas of culture to exercise its dictatorship in every respect over the bourgeoisie." While in practice the domain of natural sciences is opposed to the universal dictatorship of the bourgeoisie, it is necessary to submit

> the development of the domain of the [sciences] to the great revolutionary criticism. Marx's mathematical manuscripts constitute a brilliant model for our great revolutionary criticism of the development of the domain of this subject.

After a study of the derivative, the very popular example $y = x^3$, and the problem of how to interpret dy/dx as $0/0$, Shu Zuo's paper draws to a close with a citation from the well-known letter Engels wrote to Marx on November 21, 1882, in which he discussed the meaning of $x + h$ as a point moving from position x to x_1. The paper ends on a typically Marxist note:

> We certainly must take this sharp weapon of dialectical materialism, to develop the great revolutionary criticism of the domain of our subject, dare to revolt, and know how to revolt! We are full of confidence that the mysterious veil of every shade enshrouding the natural sciences will certainly be torn away completely, and the domination of the natural sciences by idealism and metaphysics of the past systems will be thoroughly smashed, and the red flag of Marxism, Leninism and the thoughts of Mao Zedong will flutter high above the front position of the natural sciences. This is the universal truth that enlightens us by studying Marx's "Mathematical Manuscripts."

Shu Zuo's paper was immediately followed by another concerned with the "Mathematical Manuscripts," this one by Wu Wen-jing (吴文京) who identified himself as a worker at the Birch Woods Rubber Factory in Mu Dan Jiang (牡丹江桦林橡胶厂), a town in the North-East of China. Wu Wen-jing's paper was devoted to "The Differential and Dialectics" (微分和辩证法), and interpreted dialectics in terms of change, translated into an analysis of the mathematics of motion, a favorite Maoist theme among Marxist mathematicians. The paper discusses velocity and acceleration in terms of derivatives. Through a proper application of dialectical materialism, Wu Wen-jing insisted that a critical evaluation of the calculus would reveal its true essence. He also introduced another familiar theme as well, that it was the forces of production in society that spurred development of the natural sciences, in the course of which mathematics changed from a study of constants to variables, from static situations to ones that were dynamic and constantly changing. The last paper to be discussed here that was devoted to Marx's "Mathematical Manuscripts" in the 1975 issue of the *Journal of Fu Dan University* was a contribution by Yan Shao-zong(严绍宗), who presented his thoughts on "Basing the Concept of the Derivative on the Law of Opposites" (对立统一法则支配≪导函数概念≫ 的建立). Here "对立统一," the popular Maoist expression meaning "the unity of opposites," was nothing other than the familiar Hegelian or Marxist doctrine of the dialectical polarities of antithesis/synthesis. Yan asked the usual question, "What is to be understood by dy/dx?" Yan also cited Marx and the problem of interpreting $0/0$, and then took up the ubiquitous analysis of the equation $y = x^3$, in terms of which he discussed derivatives and distinguished between quotients of differences $\Delta y/\Delta x$ and differentials dy/dx.

CHAIRMAN MAO SPEAKS!

In 1975 a special issue of the journal *Practice and Understanding of Mathematics* (数学的实践与认识) opened with two slogans from *The Collected Sayings of Chairman Mao* (毛主席语录):

> From a certain point of view, the most talented and able soldier is one who has had the most practical experience.
>
> Our improvement is improvement on the basis of popularization; our popularization is popularization under the guidance of improvement.

These slogans were meant to reflect the ideology of the journal, as well as the articles in an issue devoted to popularizing mathematics while emphasizing its practical applications. The opening contribution was by the pseudonymous Shu Zuo (舒左), who also contributed a paper that year to the *Journal of Fu Dan University*. This time his article was devoted to a report of a meeting held to study the mathematical manuscripts of Karl Marx, in the spirit of popularization that Chairman Mao himself had admonished everyone to pursue, all of which was reflected directly in the aphorism at the head of the journal. Another attempt to present the basic ideas found in Marx's "Mathematical Manuscripts" to a wider audience was a series of lectures devoted to "Studying Marx's Mathematical Manuscripts" that appeared in the popular journal, *Chinese Science* (中国科学). The first of these was written by Shu Li (舒立) from Beijing University, and was devoted to "Using Marxism to Conquer the Battlefield of Mathematics" (用马克思主义占领数学阵阵地). The allusion to conquering the battlefield was a rhetorical flourish drawing on language Mao himself often used in referring to the struggles China had to face on all fronts. In this case, the point was to advance the battle using dialectical materialism to criticize and revise the foundations of mathematics.

1976—YEAR OF THE DRAGON

On January 8, 1976, Premier Zhou Enlai died. Six months later, on July 28, the industrial and mining city of T'ang-Shan was destroyed by a major earthquake, killing 655,000 people and leaving more than a million people homeless. The third cataclysmic event that year occurred on September 9, when Chairman Mao died. The cover of the journal *Practice and Understanding of Mathematics* (数学的实践与认识) immediately carried a portrait of the Chairman, adorned with the slogan "Eternal Glory to the Mighty Leader and Teacher Chairman Mao Ze-Dong!" (伟大的领袖和导师毛泽东主席永垂不朽!).

The opening paper in this memorial issue commemorating Chairman Mao was a joint work from the study group for Marx's "Mathematical Manuscripts" in the Department of Mathematics at Beijing Normal Teacher's College. The article, "Studying Different World Outlooks from Two Different Mathematical Approaches" (从两种不同的数学推导看不同的世界观), contrasted d'Alembert's approach to the calculus with the foundations advocated by Marx. Admittedly a preliminary study, it was based on a "first reading" of the "Mathematical Manuscripts," but nevertheless reflected a remarkably sophisticated view of the historical differences between d'Alembert's theory of limits and the critical views

of foundations of the calculus held by Marx. The same slogan—"Eternal Glory to the Mighty Leader and Teacher Chairman Mao Ze-Dong!"—also ran across the cover of the third number of the *Journal of Central China Industrial College* (华中工学院学报), atop its third issue for 1976, along with the same portrait of Chairman Mao that appeared virtually everywhere throughout China. Inside, however, a paper said to have been written by Shu Xuan (舒煊) in Wuhan was devoted to "Continuing to Use Marxism to Study Nonstandard Analysis" (坚持用马克思主义考察非标准分析). Although this article does not go into the technicalities of nonstandard analysis with actual applications to mathematics, pure or applied, it does try to develop the value of using nonstandard analysis in a spirit of evaluation and criticism of mathematics compatible with the views of Marx and Engels, both of whom are cited extensively in the article. The main point Shu Xuan makes here is that despite its suspect ideology, nonstandard analysis is nevertheless an important tool in reevaluating calculus along lines inspired by Marx and Engels.

Serious notice of nonstandard analysis

1976, the Year of the Dragon, was also the first in which a serious attempt was made in China to relate the technical details of Abraham Robinson's nonstandard analysis to proper understanding of the calculus. Written under a pseudonym, Shu Ji (舒基), an article appeared in the *Journal of North-West University* (西北大学学报) devoted to: "Discussing the Physical Origins of the Mathematical Structure of *R" (谈谈数学结构*R的现实原型). The major point of this paper was to introduce the nonstandard continuum *R, which included both infinitesimals and transfinite numbers as legitimate real numbers. Shu Ji sought to justify these, as well as nonstandard analysis in general, in terms of Marxist dialectical materialism. Once the theory was on firm ideological ground, the article proceed with deeper technical discussion of nonstandard analysis on its own terms. The article itself, and the views it introduced concerning nonstandard analysis, were prompted, Shu Ji notes, by opinions formed "after studying the dialectics of nature and Marx's mathematical manuscripts."

Shu Ji (舒基) devotes an entire section of his article to arguing that "the infinitely small (large) really are real numbers" (无限小(大)量是实在的数), where "really are real" means that the real numbers are ontologically real, concrete—in physical, material terms. After quoting from Marx's mathematical manuscripts, Engels' *Dialectics of Nature*, and Chairman Mao's "On the Correct Handling of Contradictions Among the People" (关于正确处理人民内部矛盾的问题), Shu Ji claims that Robinson himself recognized that nonstandard analysis was grounded in a concrete, material way in so far as the usefulness of infinitesimals was best seen in applications to real-world problems.

1977

In 1977 the first draft of a course of lectures given at Beijing Normal University were published by Huang Shun-Ji (黄顺基) and Wu Yan-Fu (吴延涪) in the journal *Understanding and Practice of Mathematics*. The opening lecture began

with an introduction to studying the "Mathematical Manuscripts," noting that these constituted a "brilliant document" using dialectical materialism to analyze mathematics, and were a "treasure trove" (宝库) of dialectics. The first lecture follows Marx very closely in offering a critical analysis of the foundations of the calculus through its historical development. The authors point out that studying the "Mathematical Manuscripts" confirms what Engels said at Marx's graveside: that Marx had a special interest in mathematics and made fundamental contributions of his own to the subject. The contributions were primarily in applications to Marx's theory of surplus value, and in applications revealing the special laws of change underlying the evolution of capitalism and patterns of development reflected in modern society. As Huang and Wu emphasized in their introduction:

> The times we are facing today "are times when everything is turned upside down, to which nothing in past history can compare." To strengthen and reinforce the dictatorship of the proletariat, using Marxist-Leninism, the thoughts of Mao Ze-dong have taken command of every position, pioneered study of the manuscripts and research of very important practical significance.

The authors' introductory lecture is divided into four parts, the first devoted to describing the aims Marx had in mind when he wrote the manuscripts. Then comes a section devoted to the major contents and basic ideas of the manuscripts, followed by a third section explaining the process of writing and publishing the manuscripts. The last and most interesting part of this introduction to Marx's mathematical manuscripts considers their practical significance. Here Huang and Wu list a number of major practical results that follow from study of the "Mathematical Manuscripts." Above all, they note that in every branch of science the manuscripts may be used as "a pioneering weapon of revolutionary criticism."

The final article to be discussed here was published by Zhou Guan-xiong (周冠雄) in 1977: "Using the Philosophy of Marxism to Evaluate Nonstandard Analysis" (以马克思主义哲学为指导评价非标准分析). This appeared in the *Journal of Central China Industrial College*, and summarized its main argument as follows:

> The study and discussion of Marx's "Mathematical Manuscripts" are of real and profound value in helping us to understand dialectical materialism, and in studying mathematics using Marxism.... Chairman Mao's directive identifies how we should approach our study of foreign things, how the accounts of Marx, Engels, Lenin and Mao of the infinite and of higher mathematics supply theoretical weapons for evaluating nonstandard analysis. In his "Mathematical Manuscripts," Marx traced the history of the calculus from Newton to Lagrange, acknowledging their contributions and pointing out their idealistic and metaphysical errors.
>
> Marx also analyzed the concepts of derivative, differential, differential operations, etc. Using his own philosophy, Marx outlines a series of very important results, which constitute a glorious model for examining nonstandard analysis....

> The core of nonstandard analysis provides a foundation for higher mathematics [with] infinitesimals. In [his] nonstandard analysis, [Abraham] Robinson (鲁滨逊) shows there is a certain infinitesimal between zero and any positive number using the methods of mathematical logic. The entire theory of nonstandard analysis constructs a mathematical system based on infinitesimals. The system provides another interpretation for the [viability] of the calculus, and another (mathematical) method distinct from the method of limits. We should accept the contributions Robinson has made, but object to the influence of Robinson's formalism, which in a system of natural science has its limitations. We must criticize Robinson's idealism as it appears in his works.

Conclusion

Since the founding of the People's Republic of China in 1949, Chinese scholars have produced a series of studies meant to explain, popularize and establish the methods and philosophy of dialectical materialism in virtually every field of study. In the sciences this has led to criticism, if not condemnation, of Mendelian genetics, of physics in both its Newtonian and Einsteinian interpretations, and in mathematics, of Euclidean geometry and—as has been described in some detail here—of the infinitesimal calculus. But unlike many of their colleagues in the Soviet Union, the Chinese avoided the disastrous consequences of Lysenko's triumph over Mendel by allowing that successful scientists, despite faulty philosophies, nevertheless unconsciously must have used dialectical materialism in guiding their research.

Throughout the Cultural Revolution (1966-1976), Mao Ze-dong promoted Marxism and dialectics to encourage reforms in all fields of endeavor, including the sciences. In mathematics, this encouraged, as it had Marx, an appreciation (with criticism) of the infinitesimal calculus. For Chinese mathematicians, application of Abraham Robinson's newly created nonstandard analysis not only rehabilitated infinitesimals in a technical sense, but (when understood within an appropriate materialist framework), could be used to justify and promote two new fields of study in China—model theory and nonstandard analysis.

[A complete text of this paper, including notes and bibliography, is available upon request from the author].

Joseph W. Dauben
Herbert H. Lehman College and
Ph.D. Program in History,
The Graduate Center
The City University of New York
New York, NY, USA

The Riemann-Roch Theorem and Geometry, 1854-1914

Jeremy J Gray

Abstract. The history of the Riemann-Roch Theorem, from its discovery by Riemann and Roch in the 1850's to its use by Castelnuovo and Enriques in from 1890 to 1914, offers one of the most instructive examples in the history of mathematics of how a result stays alive in mathematics by admitting many interpretations. Various mathematicians over the years took the theorem to be central to their researches in complex function theory, and in the study of algebraic curves and surfaces in a variety of algebraic and geometric styles. In surveying their interpretations and extensions of the theorem, the historian traces the creation of a general theory of complex algebraic curves and surfaces in the period, and uncovers lively agreements and disagreements. This paper provides an overview of the field; the Congress lecture will concentrate on the route from Riemann and Roch via Brill and Noether to Castelnuovo and Enriques. For reasons of space a number of the better-known developments have been omitted. One may consult Dieudonné [1976].

1991 Mathematics Subject Classification: 1, 14, 30
Keywords and Phrases: Riemann-Roch Theorem, algebraic curve, algebraic surface

1 Curves

In the 1850s (see his [1857] and Laugwitz [1996]) Riemann put together a theory of complex functions defined on some 2-dimensional domain, which might be any simply-connected domain, or the whole complex z-sphere (what we call the Riemann sphere) or a finite covering of the z-sphere branched over some points (what we call a Riemann surface). He showed how to define such a function with poles on a patch using his version of the Dirichlet principle. His motivation was the example of algebraic curves, and the outstanding topic of abelian integrals.

He established the existence of complex functions on a surface with no boundary by the Riemann inequality - his contribution to the Riemann-Roch Theorem. His imprecise argument retains its heuristic value. He supposed the surface was $(2p + 1)$-fold connected, which means that it is rendered simply-connected by $2p$ cuts, when it forms a $4p$-sided polygon. He showed that there are p linearly

independent everywhere holomorphic functions defined inside the polygon by considering what would happen if the real parts of their periods all vanished (using the Dirichlet principle again). Later he showed that the differentials of these functions are everywhere-defined holomorphic integrands. Then he specified d points at which the function may have simple poles, again imposing the condition that the functions jump by a constant along the cuts. Now he argued that to create functions with only simple poles and constant jumps one took a sum of p linearly independent functions with no poles plus functions of the form $\frac{1}{z}$ at one of the specified points, and added a constant term. The resulting expression depends linearly on $p + d + 1$ constants. The jumps therefore depend linearly on $p + d + 1$ constants, and there are $2p$ of them to be made to vanish (if the function is single-valued as required). So there will be non-constant meromorphic functions when $p + d + 1 - 2p \geq 2$, i.e. $d > p$. This result, today called the Riemann inequality, says there is a linear space of complex functions of dimension $h^0 \geq d + 1 - p$, and this contains non-constant functions as soon as $d + 1 - p > 1$, or $d > p$.

Roch was a gifted student of Riemann who died of tuberculosis in 1866 aged only 26. He was able to interpret analytically the difference $d + 1 - p$ as the dimension of a certain space of holomorphic integrands, those that vanish at some of the points where the function may have poles. This implies that the difference $h^1 = h^0 - (d+1-p)$ is an analytically meaningful quantity. In Roch's terminology: if a function w has d simple poles, and if q linearly independent integrands can vanish at these poles, then w depends on $d-p+q+1$ arbitrary constants (Roch [1865]).

Riemann showed that any two meromorphic functions on an algebraic curve are algebraically related, whence a theorem establishing that a 'Riemann surface' branched like an algebraic curve over the Riemann sphere can be mapped into the projective plane. This established a close relationship between intrinsic curves and embedded curves, and Riemann showed that any two polynomial equations for the same algebraic curve are birationally related (the variables in one equation for the curve are rational functions of the variables in any other equation for the curve). He also used his inequality to calculate that the dimension of the moduli space of algebraic curves of genus $p > 1$ is $3p - 3$.

The approach of Riemann and Roch was aimed at extending complex function theory. It made abundant use of the Dirichlet principle (which Riemann attempted to prove; his proof was refuted by Prym in 1870). But by directing attention to a theory based on a topological concept of connectivity, Riemann opened the way to elucidate intrinsic properties of curves independent of their embeddings in the plane.

Riemann died in 1866. His eventual successor at Göttingen was Clebsch, who had already pioneered the application of Riemann's ideas to geometry in his [1863]. His initial response to these ideas had been to find them very hard - the topological nature of a Riemann surface was difficult to grasp and the

Dirichlet principle appeared confusing - and Roch's paper struck him as almost incomprehensible. So he initially defined the genus of a curve as the number of linearly independent holomorphic integrands on it. If the curve is non-singular, these, he observed, are integrands of the form $\phi dz/\frac{\partial F}{\partial w}$ where $\phi(z,w)$ is of degree at most $n-3$. The condition on the degree of ϕ arises by considering what happens at infinity, and was dealt with by passing to homogeneous coordinates. If the curve has a k-fold point (it passes k times through a point) then the curve $\phi = 0$ is required to pass $k-1$ times through that singular point (such curves are called adjoint curves). In his book of 1866 with Paul Gordan, Clebsch restricted his attention to curves having only double points and cusps, for which a purely algebraic definition of the number p (called the genus by Clebsch) is possible: $p = \frac{1}{2}(n-1)(n-2) - d - r$ where n is the degree of the defining equation, d is the number of double points and r the number of simple cusps.

The response of Riemann's former student Prym was harsh. He wrote to Casorati (2 December 1866): "They would never have dared publish the foreword in Riemann's lifetime. The attempt to base function theory on algebra can be regarded as completely useless . . . . On the contrary, algebra is an outcome of function theory and not the other way round. (In Neuenschwander, [1978], p. 61). But Clebsch was a charismatic teacher, Klein [1926, p. 297] called him divinely inspired in that respect, and when he too died young, in 1872, he left behind a vigorous group of mathematicians who were to become the custodians of the Riemann-Roch Theorem. They regarded him as having led German mathematicians into the newer geometry and algebra - precisely the subjects, one might note, upon which Gauss did not work.

Prominent among them were Brill and Noether, and their critique of Clebsch-Gordan was that it had not gone far enough in embracing algebra. For them algebra was the source of rigour, and moreover, in Brill's opinion Riemann's work on the Riemann-Roch Theorem was in a form foreign to geometry. This was a sound, critical response, but the price was high: the very definition of genus became entangled with the nature of the singular points a plane curve might have, and the invariance of genus under birational transformations now had to be proved. Clebsch and Gordan had given such a geometric proof by means of a subtle elimination process, which Brill and others wanted to simplify.

The first problem is to define the multiplicity of a singular point on an algebraic curve, the second is to show that by suitable birational transformations any curve can be reduced to one having only what were called ordinary singularities, that is, singular points where all the tangent directions are distinct. Such a point was said to be of multiplicity k if there are k branches at that point. Noether broached the first of these topics in his paper of 1873 with a theorem which gave conditions for a curve that passes through the common points of two curves with equations $F = 0$ and $G = 0$ to have an equation of the form $AF + BG = 0$, where A, B, F, and G are polynomials in the complex variables x and y. It is indicative of the subtleties involved that the English mathematician F.S. Macaulay in the

1890s was among the first to pay scrupulous attention to the cases where the tangent directions are not separated.

The second problem is also difficult. It must be shown that a birational transformation can be found to simplify any given singularity, which is not obvious, and then, since the transformation necessarily introduces new singular points on the transformed curve, it must be shown that these can be made ordinary singularities. The consensus in the literature as to when this was achieved is as late as Walker's (unpublished) Chicago thesis of 1906. There is a significant papers on the topic by Bertini in 1888, and Bliss made it the subject of a Presidential address as late as 1923.

Brill and Noether were the first to call the Riemann-Roch Theorem by that name, in their [1874]. They took from Clebsch the idea that it was to be studied geometrically, that is, in terms of a linear family of adjoint curves. It followed from their definition of the genus (a generalisation of the Clebsch-Gordan definition to a curve having arbitrary ordinary singularities) that the number of free coefficients in the equation for an adjoint curve of degree $n-3$ is $p-1$. The total number of intersection points of the curve and its adjoint apart from the multiple points is $2p-2$, so at most $p-1$ of these are determined by the rest, equivalently, at least $2p-2-(p-1)=p-1$ can be chosen arbitrarily. It follows that q, the dimension of the space of adjoint curves of order $n-3$ that cut out a set of Q points, satisfies the inequality $q \geq Q-p+1$.

Using induction on q and Q, Brill and Noether proved the converse: there is a family of dimension q of adjoint curves of degree $n-3$ that cut out a set of Q points, provided $q \geq Q-p+1$. This is their version of the Riemann inequality. They now came to their version of the Riemann-Roch Theorem, which they stated in terms of what they called special families. By definition a special family satisfies the strict inequality $q > Q-p+1$. The Brill-Noether version of the Riemann-Roch Theorem then says: If an adjoint curve of order $n-3$ is drawn through a special set of Q points in a q-dimensional family of points, for which $q = Q-p+1+r$ (where $0 < r < p-1$), then this curve meets the given curve in $2p-2-Q = R$ further points that themselves belong to a special set of R points in an r-dimensional family, where $r = R-p+1+q$.

Their strong preference for algebra and geometry over function theory was criticised by Klein in his [1892]. Relations between Klein and the followers of Clebsch became strained as he moved in the late 1880s to adopt the mantle of Riemann and became his true successor at Göttingen. His enthusiasm for intuitive geometry clashed with their preference for the certainties of algebra.

For Brill and Noether, the Riemann-Roch Theorem was a theorem about families of plane curves. The first to use higher-dimensional geometry in this context were L. Kraus (who had studied under Klein and Weierstrass and died at the age of 27) and E.B. Christoffel, although credit has usually been given to

Noether. All started from the observation that an algebraic curve of genus $p > 1$ has a p-dimensional space of holomorphic 1-forms. While Christoffel and Noether pursued the analytic implications of taking a basis for these 1-forms, say $\omega_1,\ldots,\omega_p$, Kraus [1880] thought of the p-tuple $(\omega_1(z), \ldots,\omega_p(z)) = (f_1(z)dz, \ldots, f_p(z)dz)$ as giving a map from the curve to a projective space of dimension $p-1 : z \longrightarrow [f_1(z), \ldots, f_p(z)]$. That this map is well-defined (independent of the coordinate system used) and is a map into projective space (the f_i never simultaneously vanish) was fudged by Kraus, (a modern simple proof uses the Riemann-Roch Theorem). That the degree of the map is $2p - 2$ was explicit in Riemann's work. As Kraus saw, the case where the curve is hyper-elliptic also causes problems: here one gets a 2-1 map from the curve to the Riemann sphere. The novelty of Kraus's insight, which Klein appreciated, is the emphasis on higher-dimensional geometry. Whenever there is such a map, questions about the curve, or whole families of curves, are reduced to questions in projective geometry.

First Dedekind and Weber [1882] then Hensel and Landsberg (see Gray [1997]) took up the study of algebraic curves via their associated function fields. Landsberg's [1898a] gave a new proof of the Riemann-Roch theorem, as he put it 'in full generality and without birational transformations'. In the same issue of *Mathematische Annalen* Lansdberg also formulated and proved what he called an analogue of the Riemann-Roch Theorem in the theory of algebraic numbers, and observed that Hilbert had told him that an analogous result held for algebraic number fields. In 1902 Hensel and Landsberg published their joint book on the subject, which was to be the foundation of subsequent work in this direction.

The weak point of this approach is that it does not generalise automatically to algebraic surfaces. Nonetheless in his [1909] (corrected and simplified in his [1910a, b]) Heinrich Jung was able to extended the ideas of Hensel and Landsberg to cover function fields in two variables, and in this way he was able to obtain a Riemann-Roch theorem within the arithmetic tradition (see Gray [1994b]). As he pointed out, his proof was not that different from the Italian one, except that it also applied to divisors that were not integral, which in his view was an improvement.

The markedly algebraic approaches just described were different in spirit from those adopted by Klein and Poincaré, who hewed more closely to the ideas first elaborated by Riemann. The uniformisation theorem, conjectured by Poincaré and Klein in 1881 and eventually proved by

Poincaré and Koebe in 1907, (see Gray [1994a]) opened another route, if one could count the free constants in the Fuchsian functions having at most m poles on a given Riemann surface. The first to try was a former student of Klein's, Ernst Ritter, who made a spirited attempt in his [1894] to connect Klein's work to Poincaré's. Ritter was led to what he called an extended Riemann-Roch Theorem not for functions but for pairs of automorphic forms of particular kinds, and formulated for fractional divisors (a concept first introduced by Klein in his lectures [1892, p.65]). Ritter died age 28, but his ideas were taken up by Robert Fricke

and incorporated into the second volume of Fricke-Klein [1912]. Hermann Weyl, in his famous [1913], proved both the usual Riemann-Roch Theorem and what he called Ritter's extended Riemann-Roch Theorem. The first to give a proof of the Riemann-Roch Theorem using the uniformisation theorem was probably Osgood, who communicated such a proof to his Harvard colleague Coolidge in 1927 and later published it in the second volume of his *Funktiontheorie*, 1929.

2 Surfaces

In the wake of work by Cayley and Clebsch (see Gray [1989]) Noether defined what became known as the arithmetic genus of a surface of degree n in his [1871]. It was a number, p_a, obtained by counting coefficients, which was related to the dimension of the space of adjoint surfaces of order $n-4$ passing $(i-1)$-times through each i-fold curve of F and $(k-2)$-times through each k-fold point. Zeuthen had shown it was a birational invariant, and so one which would survive attempts to resolve the singularities. In his [1875], Noether defined the surface genus, later called the geometric genus, p_g, as the actual number of linearly independent surfaces of degree $n-4$ adjoint to a surface F of degree n. He called the genus of the intersection of F with an adjoint surface the linear genus and showed that surfaces with small values of these genera yielded immediately to classification, as surfaces defined by a polynomial equation of a certain degree with such-and-such double curves and multiple points.

Then in a short paper of 1886 Noether gave the first statement of a Riemann-Roch theorem for algebraic surfaces. Although hopelessly flawed, Noether's mistakes give a good indication of the difficulties inherent in the new subject. Noether took a curve C of genus π on a surface F, and supposed it belongs to an r-dimensional linear system, $|C|$, of curves of the same order, and that C meets a generic curve of this system in a set, G_s, of s points (called the characteristic series on C). This set of points belongs to a linear series on the curve C of dimension $r-1$. If ρ denotes the dimension of the space of adjoints of degree $n-4$, that also pass through C, then Noether's Riemann-Roch theorem asserts that

$$r \geq p_g + s - \pi - \rho + 1,$$

where p_g is the geometric genus of the surface F. For, by the Riemann-Roch theorem on the curve C, there is a linear system $|C'|$ residual to $|C|$ that cuts C in point group consisting of $2\pi - 2 - s$ points. This residual linear system has dimension $d = r - s + \pi - 1$, and arises by cutting C with adjoint surfaces that pass through a fixed G_s. These include a fixed surface through the G_s and a linear system of free surfaces adjoint to F of dimension $p_g - \rho$. Therefore, said Noether,

$$r - s + \pi - 1 \geq p_g - \rho.$$

As Enriques and Castelnuovo showed, Noether's account rests on two crucial, and doubtful, statements. First, the application of the Riemann-Roch theorem

on C assumed that the characteristic series G_s was complete (i.e. of maximal dimension), but this not obvious, and indeed is not always true. One can only say that $d = r - s + \pi - 1 + \delta$, for some $\delta \geq 0$. Second, Noether's claim that $d = p_g - \rho$, is again neither obvious nor always true. One can only say that the dimension d is $p_g - \rho + \eta$, where $\eta \geq 0$. Consequently one only has

$$r = s - \pi + p_g + 1 - \rho - \delta + \eta,$$

which is a disaster, because the correction terms δ and η, each non-negative, enter with opposite signs, and not even an inequality can be disentangled from the correct formula. In the absence of a Noether *Nachlass*, we may never know why Noether offered only this brief, and flawed, sketch.

Another approach to the study of algebraic surfaces was initiated by the Italian mathematician Veronese, who in his [1881] used the method of projection and section to show how curves and surfaces in the plane or in 3-space with singularities could profitably thought of a non-singular objects in a higher-dimensional space; the singularities were the result of the projection of the object into 3-space. Veronese's insight, together with that of Kraus as taken up by Klein, suggested to Corrado Segre that the best approach to surfaces would be to study them birationally, and to look for families of curves sufficiently well behaved to yield an embedding of the surface in some suitable projective space. So Segre advocated a third approach to the Riemann-Roch Theorem, also algebrao-geometrical, but with the emphasis on higher-dimensional projective geometry, which was to prove characteristically Italian. On this approach all birational images of a surface in any projective space were treated equally.

The aim became to find systems of canonical curves on an algebraic surface that yield embeddings in some projective space. Canonical curves K on the surface should be cut out by appropriate adjoint surfaces (as in Noether's approach). The adjoint, $A(C)$, of a curve C should be the sum $C + K$. A suitable generalisation of the Riemann-Roch Theorem should apply to the surface and a curve C or the maximal linear system $|C|$ to which C belongs and evaluate dimensions of linear systems of curves. In particular, if the dimension of the space of canonical curves (or some multiple of them) is large enough, the adjoint surfaces will yield an embedding of the surface in projective space.

However, as Enriques observed in his first major paper, his [1893], Noether's definition of an adjoint surface invokes the degree, so it is projective but not birational. Another definition of the terms 'adjoint' and 'canonical' must be sought. Moreover, linear families of curves on the surface will yield maps to projective space, but if the curves have base points (points common to all the curves), the image of the surface that they provide will have new singularities (the base points will 'blow up' into curves). Similarly components common to all curves of a family may blow down to points. So a way must be found of controlling, and ideally eliminating, these exceptional curves.

In an interesting split in the development of the theory, Italian algebraic geometers offered definitions that had nothing to do with holomorphic integrands, whereas the study of single and double integrals on an algebraic surface (1- and 2-forms) was energetically taken up by the French, notably Picard but also Humbert (see Houzel [1991]). The algebraic and transcendental theories were developed in parallel, with each side reading the other's work, but not merged.

When Castelnuovo and Enriques began their work, little was known about the nature of algebraic surfaces, and there was no method available for the resolution of their possible singularities (one was later developed in Jung [1908]). Much of their work behind the scenes is documented in the recently published letters of Enriques to Castelnuovo (see Bottazzini et al, [1996]). They came to favour a characterisation of surfaces in terms of integers, generalising the arithmetic and geometric genera, and the crucial result that gave them control over these numbers was their formulation of a Riemann-Roch Theorem.

To produce birationally invariant definitions in his [1893], Enriques excluded irregular surfaces (those for which the geometric genus exceeds the arithmetic genus, such as ruled surfaces) and surfaces of genus 0. For regular algebraic surfaces of genus greater than zero he could give quite general conditions that ensured that the characteristic series was complete. Under these conditions he considered a curve C of genus π on the surface that belongs to a linear system of curves of dimension r, such that C meets a generic curve of $|C|$ in s points. If the system $|C|$ is not contained in the canonical system he supposed that through the points common to two curves $|C|$ there passed a space of adjoint curves of dimension $2p + \omega$; he called the non-negative number ω the super-abundance of $|C|$. If the system $|C|$ is contained in the canonical system and the residual system has dimension $i - 1$, the space of adjoint curves has dimension $2p + \omega - i$. He then established the first Riemann-Roch Theorem for algebraic surfaces:

$$r = s - \pi + p_g + 1 + \omega - i,$$

by an ingenious application of the Riemann-Roch theorem on the curve C. His friend Castelnuovo then showed how irregular surfaces could be treated, in his [1896].

Enriques soon became dissatisfied with the arithmetic and geometric genera, because they did not characterise surfaces. In 1896 (see Bottazzini et al [1996, p. 278]) Enriques considered a tetrahedron in $\mathbf{CP}^3$ and observed that there was a surface of degree 6 which had the edges of the tetrahedron as its double curves. Its adjoint surfaces must be of degree 6 - 4 = 2, and must pass through the double curves of the surface. But plainly there is no quadric surface through the 6 edges of a tetrahedron. However, there is a surface of degree $2(n - 4) = 2.(6 - 4) = 4$ which passes twice through the edges of the tetrahedron: the surface composed of the four planes that form the faces of the tetrahedron. So the surface of degree 6 has no adjoint surface and its genus is zero, but it does have what is called a bi-adjoint surface and its bi-genus, the dimension of the space of bi-adjoint

surfaces, is P_2 is 1, not zero. In his [1896] Castelnuovo showed that a surface with arithmetic and geometric genera equal to zero and bi-genus $P_2 = 0$ is indeed a rational surface. This was the first birational characterisation of a surface. It also marks the moment when the so-called plurigenera decisively enter the analysis. The i^{th} plurigenus, P_i, is defined as one more than the dimension of the i^{th} multiple of the canonical system, $|iK|$. In their [1901] Castelnuovo and Enriques used the Riemann-Roch Theorem to obtain lower bounds on the plurigenera (see below).

Enriques' [1896] marks a considerable advance on his [1893] in its level of generality. Irregular surfaces could now be treated, because of recent discoveries by Castelnuovo in his [1896], and a Riemann-Roch Theorem proved about them. The characteristic property of a canonical curve was now that it was a residual curve of any linear system $|C|$ with respect to the adjoint system $A(C)$. Enriques pointed out that this indirect definition had the advantage of being independent of the nature of the fundamental curves of $|C|$ which were therefore subject to no restriction. In his opinion, this made it most appropriate to a birational theory of surfaces.

In their [1901] Castelnuovo and Enriques made notable simplifications to the theory, when they showed that a non-ruled surface can be transformed to one without exceptional curves (curves obtained as the blow-up of points under a birational transformation). This established the existence and uniqueness of what became called minimal models. For surfaces without exceptional curves, Castelnuovo's formula for the plurigenera applied to linear system $|C|$ of genus π and degree n, for which $n < 2\pi - 2$ (the genus of $|C|$ is the genus of a generic member of $|C|$, the degree the number of points in the generic intersection of two members of $|C|$). It asserts that

$$P_i \geq p_a + \frac{1}{2}i(i-1)(p^{(1)} - 1) + 1,$$

for all $i > 1$. The number $p^{(1)}$ is Noether's linear genus of the surface, but with a new birational definition that applies where Noether's does not, as Castelnuovo and Enriques pointed out. The formula indicates that the cases $p^{(1)} > 1$ and $p^{(1)} \leq 1$ will be very different. In fact, the plurigenera can grow in essentially four ways: they might all be 0, they might be 0 or 1, they might grow linearly with i or quadratically with i. This distinction lies at the heart of the subsequent classification of algebraic surfaces.

Castelnuovo and Enriques now gave a preliminary classification of surfaces, characterising rational and ruled surfaces in terms of the values of $p^{(1)}$ and the plurigenera. In 1905 Enriques characterised ruled surfaces in similar terms as the surfaces for which $p_g = 0 = P_4 = P_6$. In his [1906] Enriques characterised his sextic surface birationally as the only surface for which $p_a = 0 = P_3$, $P_2 = 1$. This characterised the class of surfaces nowadays called Enriques surfaces. In his [1906] with Castelnouvo (published as an appendix in the second volume of Picard and Simart's book) and again in his [1907b] Enriques analysed surfaces for which

$p^{(1)} = 1$, and connected the values of $p^{(1)}$ and p_g to growth of the plurigenera.

In his [1914a, b] Enriques turned back to the study and classification of algebraic surfaces. He reported in more detail in his essay with Castelnuovo published in the *Encyklopädie der Mathematischen Wissenschaften* for 1914. The behaviour of the plurigenera led him to argue now that the crucial feature was the value of P_{12}. If $P_{12} = 0$ the surface was ruled; if $P_{12} = 1$ then $p^{(1)} = 1$ and many of the surfaces discovered by Enriques and Severi, Picard, and Bagnera and de Francis belong in this family. If $P_{12} > 1$ and $p^{(1)} > 1$ then the surface has effective canonical and pluri-canonical curves of some positive order (which in turn meant that it could be embedded in some projective space). Or rather, a model of the surface containing no exceptional curves could be mapped birationally onto its image in a projective space of an appropriate dimension. The classification is actually somewhat finer, but it is clear that the broad outlines of the classification are provided by numbers determined, in fair part, by the Riemann-Roch Theorem.

I wish to thank A. Beardon, D. Eisenbud, G.B. Segal, N.I. Shepherd-Barron, and M.H.P. Wilson for commenting helpfully on various versions of this paper.

BIBLIOGRAPHY

Bertini, E. 1888, Sopra alcuni teoremi fondamentali delle curve piane algebriche, *R. Ist. Lomb.* (2) 21, 326-333, 416-424

Bliss, G.A. 1923 The Reduction of singularities of plane curves by birational transformation, *Bull. AMS*, 29, 161-183

Bottazzini, U., Conte, A., Gario, P. 1996 *Riposte Armonie, Lettere di Federigo Enriques a Guido Castelnuovo*, Bollati Boringhieri

Brill, A., Noether, M. 1874 Über die algebraischen Functionen und ihre Anwendung in der Geometrie, *Math. Ann.*, 7, 269-310

Brill, A., and Noether, M. 1893 Die Entwicklung der Theorie der algebraischen Functionen in älterer und neuerer Zeit, *Jahresbericht DMV*, 3, 107-566.

Castelnuovo, G. 1896 Alcuni risultati sui sistemi lineari di curve appartenenti ad una superficie algebrica, *Mem. Soc. It. sci. XL.* (3) 10, 82-102, in *Memorie*, 335-360

Castelnuovo, G. 1897 Alcune proprietà fondamentali sui sistemi lineari di curve tracciati sopra ad una superficie algebrica, *Ann. mat. pura appl.* (2) 25, 235-318, in *Memorie*, 361-442

Castelnuovo, G., Enriques F. 1901 Sopra alcune questioni fondamentali nella teoria delle superficie algebriche, *Ann. di mat.* (3) 6, 165, in Enriques F. *Memorie scelte*, 2, 85-144.

Castelnuovo, G., Enriques F. 1906 Sur quelques résultats nouveaux dans la théorie des surfaces algébriques, in Picard and Simart, [1906], 2, 485-522.

Cayley, A. 1871 On the deficiency of certain surfaces, *Math. Ann.*, 3, 526-529.

Clebsch, R. F. A. 1863 Über die Anwendung der Abelschen Functionen in der Geometrie, *J. für Math.*, 63, 189-243.

Clebsch, R.F.A., Gordan, P. 1866 *Theorie der Abelschen Functionen*, Teubner, Leipzig

Dedekind, R., Weber, H. 1882 Theorie der algebraischen Functionen einer Veränderlichen, *J. für Math.*, 92, 181-290, in R. Dedekind *Ges. Math. Werke*, 1, Chelsea, New York, 1969, 238-350

Dieudonné, J. 1974 *Cours de géométrie algébrique*, Paris, PUF

Enriques, F. 1893a Ricerche di geometria sulle superficie algebriche, Mem. Acc. Torino (2) 44 171-232, in *Memorie scelte*, 1, 31-106.

Enriques, F. 1896 Introduzione alla geometria sopra le superficie algebriche, *Mem. Soc. Ital. delle Scienze*, (3) 10, 1-81 in *Memorie scelte*, 1, 211-312.

Enriques, F. 1905 Sulle superficie algebriche di genere geometrico zero, *Rend. Circ. Mat. Palermo*, 20 1. in *Memorie scelte*, 2, 169-204

Enriques, F. 1907a Sopra le superficie algebriche di bigenere uno, *Mem. Soc. Ital. delle Scienze*, (3) 14, 327-352 in *Memorie scelte*, 2, 241-272.

Enriques, F. 1907b Intorno alle superficie algebriche di genere $p^{(1)} = 1$, *Rend. Acc. Bol.*, 11, 11-15, in *Memorie scelte*, 2, 279-282

Enriques, F. 1914a,b Sulla classificazione delle superficie algebriche e particolarmente sulle superficie di genere lineare $p^{(1)} = 1$. Note I e II, *Rend. Accad. Lincei*, (5) 23, 206-214 and 291-297, in *Memorie scelte*, 3, 173-182 and 183-190

Enriques, F., Castelnuovo, G. 1914 Die algebraischen Flächen vom Gesichtspunkte der birationalen Transformationen aus, in *Encyklopädie der Mathematischen Wissenschaften*, III.2.1 C, 674-768

Fricke, R., Klein, C.F., 1897, 1912 *Vorlesungen über die Theorie der Automorphen Functionen*, 2 vols, Teubner, Leipzig and Berlin

Gray, J.J. 1989 Algebraic and projective geometry in late 19th century, in *Symposium on the history of mathematics in the 19th century*, 361-388, ed. J. McCleary, D.E. Rowe, Academic Press, New York

Gray, J.J. 1994a On the history of the Riemann mapping theorem, *Supp. di Rend. circ. mat. Palermo*, (2) 34, 47-94.

Gray, J.J. 1994b German and Italian Algebraic Geometry, *Supp. di Rend. circ. mat. Palermo*, (2) 36, 151-183

Gray, J.J. 1997 Algebraic geometry between Noether and Noether - A forgotten chapter in the history of algebraic geometry, *Revue d'hist. Math.*, 3.1, 1-48

Hensel, K., Landsberg, G., 1902 *Theorie der algebraischen Functionen einer Variablen*, Teubner, Leipzig

Houzel, C. 1991 Aux origines de la géométrie algébrique: les travaux de Picard sur les surfaces (1884-1905), *Cahiers d'hist. phil. des sciences*, 34, 243-276.

Jung, H. W. E. 1908 Darstellung der Funktionen eines algebraischen Körpers zweier unabhängiger Veränderlicher, *J. für Math.*, 133, 289-314

Jung, H. W. E.. 1909 Der Riemann-Roch Satz für algebraische Funktionen zweier Veränderlicher, *Jahresbericht DMV*, 18, 267-339.

Jung, H. W. E.. 1910a Zur Theorie der algebraischen Flächen, *Jahresbericht DMV*, 19 172-176.

Jung, H. W. E. 1910b Zur Theorie der Kurvenscharen auf einer algebraischen Fläche, *J. für Math.*, 138 77-95.

Klein, F. 1892 *Riemannsche Flächen, Vorlesungen*, Teubner-Archiv, 5

Klein, F. 1926 *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, Teubner, Berlin, Chelsea, New York reprint 1950.
Kraus, L. 1880 Note über aussergewöhnliche Specialgruppen auf algebraischen Curven, *Math. Ann.* 16, 245-259
Landsberg, G., 1898a Algebraische Untersuchungen über den Riemann-Roch Satz, *Math. Ann.* 50, 333-380
Landsberg, G. 1898b Über das Analogon des Riemann-Roch Satzes in der Theorie der algebraischen Zahlen, *Math. Ann.* 50, 577-582
Laugwitz, D. 1996 *Bernhard Riemann*, Birkhäuser, Vita Mathematica, Basel
Macaulay F. S. 1900 The theorem of residuation, being a general treatment of the intersections of plane curves at multiple points, *Proc. LMS* (1), 31, 381-423.
Neuenschwander, E. 1978 Der Nachlass von Casorati (1835-1890) in Pavia, *A. H. E. S.* 19.1, 1-89.
Noether, M. 1871 Sulle curve multiple di superficie algebriche, *Ann di mat*, 2, 5, 163-178.
Noether, M. 1873, 'Ueber einen Satz aus der Theorie der algebraischen Functionen', *Math. Ann.*, 6, 351-359
Noether, M. 1875 Zur Theorie des eindeutigen Entsprechens algebraischer Gebilde, Zweiter Aufsatz, *Math. Ann.*, 8, 495-533.
Noether, M. 1886 Extension du théorème de Riemann-Roch aux surfaces algébriques, *Comptes rendus*, 103 734-7.
Osgood, W. F. 1929 *Lehrbuch der Funktionentheorie*, 2.1, Teubner, Leipzig, Chelsea, New York, 1965
Picard, E., Simart, G. *Théorie des fonctions algébriques de deux variables indépendantes*, 1, 1897, 2, 1906, Chelsea, New York, 1971
Riemann, B. 1857 Theorie der Abel'schen Functionen, *J. für Math.*, 54 in *Gesammelte mathematische Werke* (1990) 120-144.
Ritter, E. 1894 Die multiplicativen Formen auf algebraischem Gebilde beliebigen Gechlechts mit Anwendung auf die Theorie der automorphen Formen, *Math. Ann.* 44, 261-374
Roch, G. 1865 Ueber die Anzahl der willkürlichen Constanten in algebraischen Functionen, *J. für Math.* 64, 372-376.
Salmon, G. 1847 On the degree of a surface reciprocal to a given one, *Cam. Dublin Math. J.*, (2), 2, 65-73
Veronese, G. 1882 Behandlung der projectivischen Verhaältnisse der Räume von verschiedenen Dimensionen, *Math. Ann.* 19, 161-234
Weyl, H. 1913 *Die Idee der Riemannschen Fläche*, Teubner, Leipzig.

Faculty of Mathematics and Computing
The Open University
Walton Hall
Milton Keynes
MK7 6AA
UK

Author Index for Volumes II, III