ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

SECTIONS 15–20

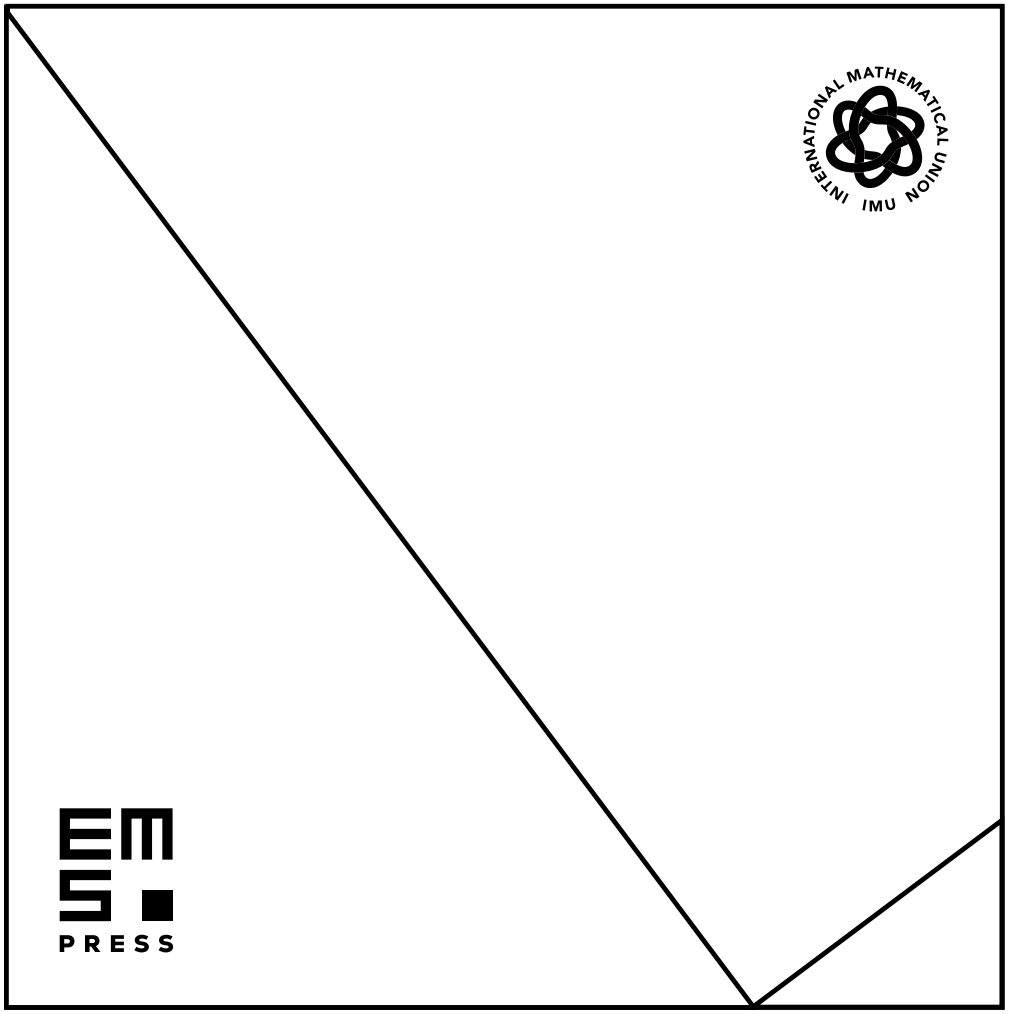
EDITED BY D. BELIAEV AND S. SMIRNOV


EMS
PRESS

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

SECTIONS 15–20

EDITED BY D. BELIAEV AND S. SMIRNOV



Editors

Dmitry Beliaev
Mathematical Institute
University of Oxford
Andrew Wiles Building
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG, UK

Email: belyaev@maths.ox.ac.uk

Stanislav Smirnov
Section de mathématiques
Université de Genève
rue du Conseil-Général 7–9
1205 Genève, Switzerland

Email: stanislav.smirnov@unige.ch

2020 Mathematics Subject Classification: 00B25

ISBN 978-3-98547-058-7, eISBN 978-3-98547-558-2, DOI 10.4171/ICM2022

Volume 1. Prize Lectures
ISBN 978-3-98547-059-4, eISBN 978-3-98547-559-9, DOI 10.4171/ICM2022-1
Volume 2. Plenary Lectures
ISBN 978-3-98547-060-0, eISBN 978-3-98547-560-5, DOI 10.4171/ICM2022-2
Volume 3. Sections 1–4
ISBN 978-3-98547-061-7, eISBN 978-3-98547-561-2, DOI 10.4171/ICM2022-3
Volume 4. Sections 5–8
ISBN 978-3-98547-062-4, eISBN 978-3-98547-562-9, DOI 10.4171/ICM2022-4
Volume 5. Sections 9–11
ISBN 978-3-98547-063-1, eISBN 978-3-98547-563-6, DOI 10.4171/ICM2022-5
Volume 6. Sections 12–14
ISBN 978-3-98547-064-8, eISBN 978-3-98547-564-3, DOI 10.4171/ICM2022-6
⟶ Volume 7. Sections 15–20
ISBN 978-3-98547-065-5, eISBN 978-3-98547-565-0, DOI 10.4171/ICM2022-7

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Published by EMS Press, an imprint of the

European Mathematical Society – EMS – Publishing House GmbH
Institut für Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin, Germany

https://ems.press

Typesetting using the authors' LaTeX sources: VTeX, Vilnius, Lithuania
Printed in Germany
♾ Printed on acid free paper

CONTENTS

VOLUME 1

PRIZE LECTURES

POPULAR SCIENTIFIC EXPOSITIONS

SUMMARIES OF PRIZE WINNERS' WORK

VOLUME 2

SPECIAL PLENARY LECTURES

PLENARY LECTURES

VOLUME 3

1. LOGIC

2. ALGEBRA

3. NUMBER THEORY – SPECIAL LECTURE

3. NUMBER THEORY

4. ALGEBRAIC AND COMPLEX GEOMETRY – SPECIAL LECTURE

4. ALGEBRAIC AND COMPLEX GEOMETRY

VOLUME 4

5. GEOMETRY – SPECIAL LECTURES

5. GEOMETRY

6. TOPOLOGY

7. LIE THEORY AND GENERALIZATIONS

8. ANALYSIS – SPECIAL LECTURE

8. ANALYSIS

VOLUME 5

9. DYNAMICS

10. PARTIAL DIFFERENTIAL EQUATIONS

11. MATHEMATICAL PHYSICS – SPECIAL LECTURE

11. MATHEMATICAL PHYSICS

VOLUME 6

12. PROBABILITY – SPECIAL LECTURE

12. PROBABILITY

13. COMBINATORICS – SPECIAL LECTURE

13. COMBINATORICS

VOLUME 7

18. STOCHASTIC AND DIFFERENTIAL MODELLING

19. MATHEMATICAL EDUCATION AND POPULARIZATION OF MATHEMATICS

20. HISTORY OF MATHEMATICS

15. NUMERICAL ANALYSIS AND SCIENTIFIC COMPUTING

MATHEMATICAL ANALYSIS AND NUMERICAL METHODS FOR INVERSE SCATTERING PROBLEMS

GANG BAO (包 刚)

ABSTRACT

Inverse scattering problems arise in diverse application areas, such as geophysical prospecting, submarine detection, near-field and nanooptical imaging, and medical imaging. For a given wave incident on a medium enclosed by a bounded domain, the scattering (direct) problem is to determine the scattered field or the energy distribution for the known scatterer. An inverse scattering problem is to determine the scatterer from the boundary measurements of the fields. Although significant recent progress has been made in solving the inverse problems, many challenging mathematical and computational issues remain unresolved. In particular, the severe ill-posedness has thus far limited the scope of inverse problem methods in practical applications. This paper is concerned with mathematical analysis and numerical methods for solving inverse scattering problems of broad interest. Based on multifrequency data, effective computational and mathematical approaches are presented for overcoming the ill-posedness of the inverse problems. A brief overview of these approaches and results is provided. Particular attention is paid to inverse medium, inverse obstacle, and inverse source scattering problems. Related topics and open problems are also discussed.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 65N21; Secondary 35R30, 35Q60, 78A46

KEYWORDS

Inverse scattering problems, stability analysis, multiple frequency, stable reconstruction methods

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5034–5055
DOI 10.4171/ICM2022/5

1. INTRODUCTION

Research on scattering and inverse scattering plays a critical role in the advancement of exploration science, especially in medical imaging, stealth technology, oil and gas exploration, nondestructive testing, materials characterization, optical microscopy, and nanooptical imaging. Scattering involves studying the interaction of a medium, often inhomogeneous, with incident waves or particles, while inverse scattering deals with determining the medium, such as location, geometry, or material properties, by the wave field measured externally.

Over the last few decades, the ever-growing practical applications and scientific developments have driven the need for more sophisticated mathematical models and numerical algorithms to describe the scattering of complicated structures, to accurately compute scattered fields and thus to predict the performance of a given structure, as well as to carry out the optimal design of new structures. The rapid growth of computational capability and the development of fast algorithms have also made inverse scattering a viable option for solving many identification problems. Mathematically, inverse scattering has been an emerging and core field of modern mathematical physics. Significant progress has been made in the mathematical studies of uniqueness and stability, as well as the development of numerical methods for solving inverse scattering problems [66,74,78,88,94,102,103]. However, there are outstanding mathematical and computational challenges that remain to be resolved, especially the nonlinearity, ill-posedness, model uncertainty, and large-scale computation. In addition, in the area of nanotechnology and biology, optical measurement techniques are commonly used. Since the size of the measured structure is extremely small, how to overcome the diffraction limit to obtain superresolution imaging presents another key challenge.

This paper is not intended to cover all of the broad topics in inverse scattering theory for wave propagation. It is designed to be an introduction to the work of our research group to overcome the above challenges for solving the inverse scattering problems. Throughout, we are mainly concerned with multifrequency data for the following reasons. First, due to lack of stability, the inverse scattering problems are severely ill-posed at a fixed frequency, that is, small variations in the measured data can lead to large errors in the reconstructions. On the other hand, the problems become well-posed with Lipschitz-type stability estimates when all frequency data, corresponding to the time domain case, is available. Second, the nonlinearity of the inverse scattering problems at high wavenumber leads to many local minima for the associated optimization method. By properly designing a numerical method, such a highly nonlinear problem may be reduced to a set of linear problems at given frequencies. Physically, the approach based on multifrequency data is consistent with the Heisenberg uncertainty principle. According to the principle, one-half of the wavelength is the diffraction limit for resolving the sharpness of details that may be observed by optical microscopy [57,67,79]. The diffraction limit provides a limit on the accuracy of the reconstruction for a given wavelength. To improve the resolution, it is desirable to use an incident field with a shorter wavelength or a higher frequency to illuminate the scatterer.

The goal of this paper is two-fold. Concerning mathematical analysis, our recent stability results for the inverse scattering problems are discussed. Regarding numerical methods, we present the stable recursive linearization method (RLM) for solving quantitatively the inverse scattering problems with increased resolution.

The underlying physical model is usually a wave propagation system decided by means of measuring data. In this work, our primary focus is on acoustic and electromagnetic wave propagation governed by the Helmholtz equation and Maxwell's equations, respectively. Many of the approaches and methods may be extended to study inverse scattering problems in other wave propagation models, especially elastic waves. The inverse scattering problems for wave propagation can be broadly divided into three classes: the inverse medium problem (IMP), the inverse obstacle problem (IOP), and the inverse source problem (ISP), depending on the nature of reconstructions. To emphasize the significance of the spectral information for solving the inverse scattering problems, particular attention is paid to the frequency domain models or the time-harmonic cases. To further limit the scope, the numerical methods discussed here are nonlinear optimization-based iterative methods for solving inverse scattering problems. We refer the reader to [55, 62, 65, 76, 80, 87] and references therein for noniterative, particularly direct imaging methods for solving inverse scattering problems.

The outline of this paper is as follows. In Section 2, the IMP is introduced. Stability results for the multiple frequency models are presented. Section 3 is devoted to the ISP. Stability for the multifrequency ISP of Maxwell's equations is discussed. The recent development of stochastic inverse source problems is provided. The IOP is addressed in Section 4. Of particular interest is the inverse diffraction problem. The paper is concluded with some general remarks and discussions on related problems. Some significant open problems are also presented in Section 5.

2. INVERSE MEDIUM PROBLEM

In this section, we consider the IMP, which is to reconstruct the inhomogeneous medium from boundary measurements of the scattered field surrounding the medium. The main difficulties are the ill-posedness, especially lack of stability, and the nonlinearity. In the static case (zero frequency), the problem is related to the celebrated Calderón problem [56], which is known to be severely unstable, in general [2, 101]. In fact, such severe ill-posedness carries over to the inverse medium problems for acoustic and electromagnetic waves at a fixed frequency [4, 81]. Our remedy to overcome the difficulties is to consider multifrequency boundary data. For the Maxwell equations model, we present a stable reconstruction method based on recursive linearization. The stability of the model IMP is also investigated.

2.1. Model problem

Consider the time harmonic Maxwell equation in three dimensions, namely

$$\nabla \times (\nabla \times E^t) - \kappa^2 (1+q) E^t = 0 \quad \text{in } \mathbb{R}^3, \tag{2.1}$$

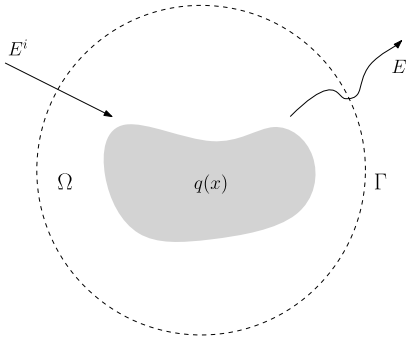


FIGURE 1
The inverse medium problem geometry. A plane wave E^i is incident on the scatterer q with a compact support contained in Ω.

where E^t is the total electric field, $\kappa > 0$ is the wavenumber or frequency, and q is a real function known as the scatterer representing the inhomogeneous medium. The scatterer is assumed to have a compact support contained in a bounded domain $\Omega \subset \mathbb{R}^3$ with boundary Γ, and satisfies $-1 < q \leq q_{\max} < \infty$ where $q_{\max}$ is a positive constant. The problem geometry is shown in Figure 1.

The scatterer is illuminated by a plane wave

$$E^i(x) = \vec{p} e^{\mathrm{i}\kappa x \cdot \vec{n}},$$

where $\vec{n} \in \mathbb{S}^2$ is the propagating direction and $\vec{p} \in \mathbb{S}^2$ is the polarization vector satisfying $\vec{p} \cdot \vec{n} = 0$. Evidently, the incident wave satisfies the homogeneous Maxwell equation

$$\nabla \times (\nabla \times E^i) - \kappa^2 E^i = 0 \quad \text{in } \mathbb{R}^3. \tag{2.2}$$

Since the total field E^t consists of the incident field E^i and the scattered field E, it follows from (2.1)–(2.2) that the scattered field satisfies

$$\nabla \times (\nabla \times E) - \kappa^2 (1 + q) E = \kappa^2 q E^i \quad \text{in } \mathbb{R}^3. \tag{2.3}$$

In addition, the scattered field is required to satisfy the Silver–Müller radiation condition

$$\lim_{r \to \infty} \big((\nabla \times E) \times x - \mathrm{i}\kappa r E \big) = 0, \quad r = |x|.$$

Denote by ν the unit outward normal to Γ. Computationally, it is convenient to reduce the problem to a bounded domain by imposing a suitable (artificial) boundary condition on Γ. For simplicity, we employ the first-order absorbing boundary condition

$$\nu \times (\nabla \times E) + \mathrm{i}\kappa \nu \times (\nu \times E) = 0 \quad \text{on } \Gamma. \tag{2.4}$$

Given the incident field E^i, the direct problem is to determine the scattered field E for the known scatterer q. This work is devoted to the solution of the IMP, i.e., determining the scatterer q from the tangential trace of the electric field, $\nu \times E|_\Gamma$ at multiple frequencies.

Although this is a classical problem in inverse scattering theory, progress has been difficult to make on reconstruction methods, due to the nonlinearity and ill-posedness associated with the inverse scattering problem. We refer to [5, 58, 64, 71, 85, 86, 93, 109] for related results on the IMP.

To overcome the difficulties, an RLM was proposed in [59–61] for solving the IMP of the two-dimensional Helmholtz equation. Based on the Riccati equations for the scattering matrices, the method requires full aperture data and needs to solve a sensitivity matrix equation at each iteration. Due to the high computational cost, it is numerically difficult to extend the method to three-dimensional problems. Recently, new and more efficient RLMs have been developed for solving the two-dimensional Helmholtz equation and the three-dimensional Maxwell equations for both full and limited aperture data by directly using the differential equation formulations [13, 20, 21, 23, 24, 28, 29, 31, 32, 37, 51]. In the case of a fixed frequency, a novel RLM has also been developed by making use of the evanescent waves [22, 25]. Direct imaging techniques have been explored to replace the weak scattering for generating the initial guesses in [19, 33]. More recently, the RLM has been extended to solve the inverse medium scattering problem in elasticity [44].

Next, we present an RLM that solves the IMP of Maxwell's equations in three dimensions, which first appeared in our work [20, 25]. The algorithm requires multifrequency scattering data, and the recursive linearization is obtained by a continuation method on the wavenumber. The algorithm first solves a linear equation under the Born approximation at the lowest wavenumber. Updates are made by using the data at higher wavenumbers sequentially. Following the idea of the Kaczmarz method, we use partial data and solve an underdetermined minimal norm solution at each step. For each iteration, one forward and one adjoint state of the Maxwell equations are solved, which may be implemented by using the symmetric second-order edge elements.

2.2. Born approximation

Rewrite (2.3) as

$$\nabla \times (\nabla \times E) - \kappa^2 E = \kappa^2 q(E^i + E), \tag{2.5}$$

where the incident wave is taken as a plane wave $E^i = \vec{p}_1 e^{\mathrm{i}\kappa x\cdot\vec{n}_1}$. Consider a test function $F = \vec{p}_2 e^{\mathrm{i}\kappa x\cdot\vec{n}_2}$, where $\vec{p}_2, \vec{n}_2 \in \mathbb{S}^2$ satisfy $\vec{p}_2 \cdot \vec{n}_2 = 0$. Clearly, the plane wave F satisfies (2.2).

Multiplying equation (2.5) by F and integrating over Ω on both sides, we have, by integrating by parts and noting (2.2) for F, that

$$\int_\Gamma \big[E \times (\nabla \times F) - F \times (\nabla \times E)\big] \cdot \nu \mathrm{d}s = \kappa^2 \int_\Omega qF \cdot (E^{\mathrm{i}} + E)\mathrm{d}x.$$

A simple calculation yields

$$\int_\Omega q(x)(\vec{p}_1 \cdot \vec{p}_2) e^{\mathrm{i}\kappa x\cdot(\vec{n}_1+\vec{n}_2)}\mathrm{d}x = \frac{\mathrm{i}}{\kappa}\int_\Gamma (\nu \times E) \cdot \big((\vec{n}_2 + \nu) \times \vec{p}_2\big) e^{\mathrm{i}\kappa x\cdot\vec{n}_2}\mathrm{d}s - \int_\Omega q(x)(\vec{p}_2 \cdot E) e^{\mathrm{i}\kappa x\cdot\vec{n}_2}\mathrm{d}x.$$

For the weak scattering, either the wavenumber κ is small, or the domain of support Ω is small, or $\|q\|_{L^\infty(\Omega)}$ is small, we may drop the second (nonlinear) term on the right-hand side of the above equation to obtain the linear integral equation

$$\int_\Omega q(x)e^{\mathrm{i}\kappa x\cdot(\vec{n}_1+\vec{n}_2)}\mathrm{d}x = \frac{\mathrm{i}}{(\vec{p}_1\cdot\vec{p}_2)\kappa}\int_\Gamma (\nu\times E)\cdot\big((\vec{n}_2+\nu)\times\vec{p}_2\big)e^{\mathrm{i}\kappa x\cdot\vec{n}_2}\mathrm{d}s,$$

which is the Born approximation.

Since the scatterer q has a compact support, we use the notation

$$\hat{q}(\xi) = \int_\Omega q(x)e^{\mathrm{i}\kappa x\cdot(\vec{n}_1+\vec{n}_2)}\mathrm{d}x,$$

where $\hat{q}(\xi)$ is the Fourier transform of $q(x)$ with $\xi = \kappa(\vec{n}_1+\vec{n}_2)$. Choose

$$\vec{n}_i = (\sin\theta_i\cos\phi_i, \sin\theta_i\sin\phi_i, \cos\theta_i), \quad i = 1,2,$$

where θ_i, ϕ_i are the latitudinal and longitudinal angles, respectively. It is clear to note that the domain $[0,\pi]\times[0,2\pi]$ of $(\theta_i,\phi_i), i=1,2$, corresponds to the ball $B_{2\kappa} = \{\xi\in\mathbb{R}^3 : |\xi|\le 2\kappa\}$. Thus, the Fourier modes of $\hat{q}$ in the ball $B_{2\kappa}$ can be determined. The scattering data with higher wavenumber κ must be used in order to recover more modes of the scatterer q.

2.3. Recursive linearization

As discussed in the previous subsection, when the wavenumber κ is small, the Born approximation allows the reconstruction of those Fourier modes less than or equal to 2κ for the function $q(x)$. We now describe a procedure that recursively determines q_κ, an approximation of $q(x)$ at $\kappa=\kappa_j$ for $j=1,2,\ldots$, with the increasing wavenumber.

Suppose now that the scatterer $q_{\tilde{\kappa}}$ has been recovered at some $\tilde{\kappa}$, and that κ is slightly larger than $\tilde{\kappa}$. We wish to determine q_κ or to determine equivalently the perturbation

$$\delta q = q_\kappa - q_{\tilde{\kappa}}.$$

Let E and $\tilde{E}$ be solutions of the scattering problem (2.3)–(2.4) corresponding to q_κ and $q_{\tilde{\kappa}}$, respectively. Taking the difference of the scattering problem (2.3)–(2.4) corresponding to q_κ and $q_{\tilde{\kappa}}$, omitting the second-order smallness in δq and in $\delta E = E-\tilde{E}$, we obtain

$$\begin{cases} \nabla\times(\nabla\times\delta E) - \kappa^2(1+q_{\tilde{\kappa}})\delta E = \kappa^2\delta q(E^{\mathrm{i}}+\tilde{E}) & \text{in } \Omega,\\ \nu\times(\nabla\times\delta E) + \mathrm{i}\kappa\nu\times(\nu\times\delta E) = 0 & \text{on } \Gamma. \end{cases} \tag{2.6}$$

For the scatterer q_κ and the incident wave E^{i}, we define the scattering map

$$M(q_\kappa, E^{\mathrm{i}}) = \nu\times E|_\Gamma,$$

where E is the solution of (2.3)–(2.4) with the scatterer q_κ. For simplicity, we denote $M(q_\kappa, E^{\mathrm{i}})$ by $M(q_\kappa)$ since the scattering map $M(q_\kappa, E^{\mathrm{i}})$ is linear with respect to E^{i}.

Next, we examine the boundary data $\nu\times E(x;\theta_1,\phi_1;\kappa)$. Here, the variable x is the observation point which has two degrees of freedom on the artificial boundary Γ, θ_1 and ϕ_1 are latitudinal and longitudinal angles of the incident wave E^{i}, respectively. At each frequency, we have four degrees of freedom, and thus data redundancy, which may be addressed by fixing one of the incident angles, say θ_1.

Let $(\phi_1)_j = 2\pi(j-1)/m$, $j = 1, \ldots, m$, and define the residual operator

$$R_j(q_{\tilde{\kappa}}) = \nu \times E\big(x; \theta_1, (\phi_1)_j; \kappa\big)\big|_\Gamma - \nu \times \tilde{E}\big(x; \theta_1, (\phi_1)_j; \kappa\big)\big|_\Gamma,$$

where $\tilde{E}(x; \theta_1, (\phi_1)_j; \kappa)$ is the solution of (2.3)–(2.4) with the incident longitudinal angle $(\phi_1)_j$ and the scatterer $q_{\tilde{\kappa}}$. For each j, the linearized problem (2.6) can be written as the operator equation

$$DM_j(q_{\tilde{\kappa}})\delta q_j = R_j(q_{\tilde{\kappa}}), \tag{2.7}$$

where $DM_j(q_{\tilde{\kappa}})$ is the Fréchet derivative of the scattering map $M_j(q_\kappa)$ corresponding to the incident angle $(\phi_1)_j$. Applying the Landweber–Kaczmarz iteration [90] to (2.7) yields

$$\delta q_j = \beta_\kappa DM_j^*(q_{\tilde{\kappa}}) R_j(q_{\tilde{\kappa}}),$$

where $\beta_\kappa > 0$ is a relaxation parameter and $DM_j^*(q_{\tilde{\kappa}})$ is the adjoint operator of $DM_j(q_{\tilde{\kappa}})$.

An adjoint state method is adopted to compute the correction δq_j efficiently [25]. For each incident wave with the longitudinal angle $(\phi_1)_j$, it is necessary to solve one direct and one adjoint problem for Maxwell's equations. Since the adjoint problem takes a similar variational form to the direct problem, we need to compute essentially two direct problems at each step. Once δq_j is determined, $q_{\tilde{\kappa}}$ is updated by $q_{\tilde{\kappa}} + \delta q_j$. After the mth sweep is completed, we get the reconstructed scatterer q_κ at the wavenumber κ. Assume that the scattering data is for $\kappa \in [\kappa_{\min}, \kappa_{\max}]$ and let $\kappa_{\min} = \kappa_0 < \kappa_1 < \cdots < \kappa_n = \kappa_{\max}$. The algorithm of the RLM can be illustrated in Table 1.

Start with the Born approximation q_{k_0}.

Do the outer loop on the wavenumber k_i, $i = 1, 2, \ldots, n$.

Let $q_{k_i}^0 = q_{k_{i-1}}$.

Do the inner loop on the incident direction ϕ_j, $j = 1, 2, \ldots, m$,

$\delta q_j = \beta_k DM_j^*(q_{k_i}^{j-1}) R_j(q_{k_i}^{j-1})$,

$q_{k_i}^j = q_{k_i}^{j-1} + \delta q_j$.

End

End

TABLE 1
The algorithm beyond Born approximation

2.4. Numerical experiments

We present an example to illustrate the performance of the method. Let $\tilde{q}(x_1, x_2, x_3) = 2x_1^2 e^{-(x_1^2+x_2^2+x_3^2)}$ and reconstruct the scatterer defined by

$$q(x_1, x_2, x_3) = \tilde{q}(3x_1, 3.5x_2, 3x_3).$$

Figure 2 shows the surface plot of the true scatterer at slices $x_1 = 0.3$, $x_2 = 0$, and $x_3 = 0$, respectively. Six equally-spaced wavenumbers are used in the construction, starting from the lowest wavenumber $\kappa_{\min} = 0.5\pi$ and ending at the highest wavenumber $\kappa_{\max} = 2.5\pi$. The incident fields are taken at 20 randomly chosen directions, which accounts for 20 Landweber iterations at each wavenumber. The relaxation parameter is 0.01 and the noise level of the data is 5%. Figure 3 shows the reconstructed scatterer at the Born approximation with the wavenumber $\kappa = 0.5\pi$. Figures 4–5 illustrate the reconstructed scatterers at the different wavenumbers. It can be observed from the numerical results that the Born approximation generates a poor reconstruction, but the result can be improved as the wavenumber increases.

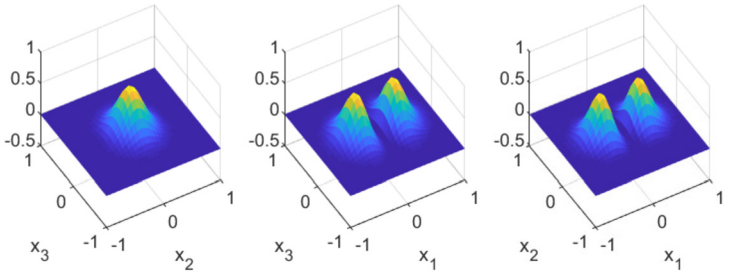


FIGURE 2
The true scatterer: (left) the slice $x_1 = 0.3$; (middle) the slice $x_2 = 0$; (right) the slice $x_3 = 0$.

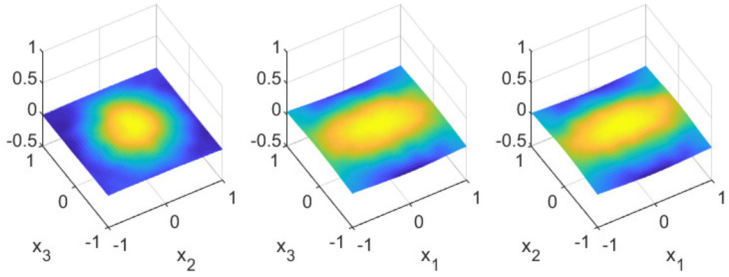


FIGURE 3
The Born approximation at the wavenumber $\kappa = 0.5\pi$: (left) the slice $x_1 = 0.3$; (middle) the slice $x_2 = 0$; (right) the slice $x_3 = 0$.

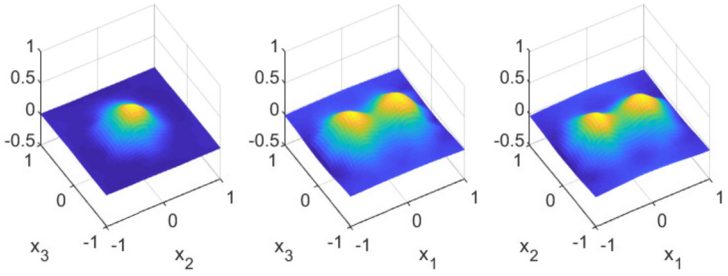


FIGURE 4

The reconstructed scatterer at the wavenumber $\kappa = 1.3\pi$: (left) the slice $x_1 = 0.3$; (middle) the slice $x_2 = 0$; (right) the slice $x_3 = 0$.

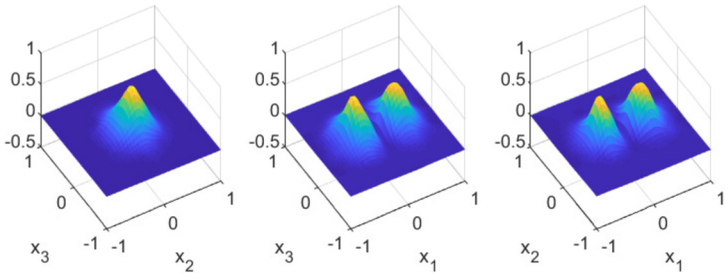


FIGURE 5

The reconstructed scatterer at the wavenumber $\kappa = 2.5\pi$: (left) the slice $x_1 = 0.3$; (middle) the slice $x_2 = 0$; (right) the slice $x_3 = 0$.

2.5. Stability analysis

It is well known that when the data is given for all frequencies and under certain geometrical assumptions, the IMP is well-posed with Lipschitz type stability estimates [45, 46, 96, 108]. However, in practice, the boundary measurements are often taken only at a finite number of frequencies. Our numerical method based on recursive linearization takes advantage of the regularity of the problem at high frequencies without being undermined by local minima. Numerical tests have shown that the method is very stable with data driven accuracy. Some preliminary convergence results of the RLM for solving the IMP with multifrequency are available in [27, 41].

Next, we present stability estimates for the multifrequency IMP in one-dimension. Stability in several dimensions is still open due to the difficulties of strong nonlinearity for high frequencies and trapped rays of the frequency-dependent scattering relation.

Consider the one-dimensional Helmholtz equation

$$\phi''(x,\kappa) + \kappa^2\big(1 + q(x)\big)\phi(x,\kappa) = 0, \quad x \in \mathbb{R},$$

where the scatterer q is assumed to be supported in $(0, 1)$. Denote by ψ_+ and ψ_- the scattering waves corresponding to the left and right excitation $e^{\pm i\kappa x}$ which satisfy

$$\phi_\pm(x,k) = \psi_\pm + e^{\pm i\kappa x}.$$

Assume $q(x) \in C_0^{m+1}([0,1])$ and define the reflection coefficients by

$$\psi_+(x,\kappa) = \mu_+(\kappa)e^{-\mathrm{i}\kappa x}, \quad \psi_-(x,\kappa) = \mu_-(\kappa)e^{\mathrm{i}\kappa x}$$

and their associated measurements

$$d_\pm(\kappa) = \frac{1-\mu_\pm(\kappa)}{1+\mu_\pm(\kappa)}.$$

Given the measurement data $d_\pm(\kappa), \kappa \in (0,\kappa_0)$, the IMP is to reconstruct the scatterer $q(x)$. In the following, we present two stability results obtained recently in [42].

Theorem 2.1. *Assume that $q, \tilde{q}$ are two scatterer functions. Let $d_\pm, \tilde{d}_\pm$ be their boundary measurements in $(0,\kappa_0)$. Then there exist a positive constant C and a function η such that the following estimate holds:*

$$\|q-\tilde{q}\|_{L^\infty(\mathbb{R})} \le C\,\|d_\pm - \tilde{d}_\pm\|^{\eta(\kappa_0)}_{L^\infty(0,\kappa_0)}.$$

Remark 2.2. We refer to [42] for the complete statement. The proof is based on a combination of the trace formula, Hitrik's pole-free strip for the Schrödinger operator, the meromorphic extension, and the Two Constant Theorem. The Hölder exponent $\eta \in (0,1)$ in the estimate is an explicit increasing function of κ_0. It tends to zero when κ_0 tends to zero which shows as expected that the ill-posedness of the inversion increases when the band of frequency shrinks. We conclude from the stability estimate that the reconstruction of the scatterer function is accurate when the band of frequency is large enough and deteriorates when this later shrinks toward zero. These theoretical results confirm the numerical observations and the physical expectations for the increasing stability phenomena by taking multifrequency data.

By taking into account the uncertainty principle, it is reasonable to consider the observable part of the scatterer. In the one-dimensional setting, the observable part of the scatterer q over the frequency band $(0,\kappa_0)$ may be well-defined by using the truncated trace formula [42].

The next theorem gives the stability estimate on the observable part of the scatterer, which shows that the reconstruction of the observable part of the scatterer is stable for κ_0 sufficiently large.

Theorem 2.3. *Assume that $q, \tilde{q}$ are two scatterer functions and $q_{\kappa_0}, \tilde{q}_{\kappa_0}$ are their corresponding observable parts. Let $d_\pm, \tilde{d}_\pm$ be the boundary measurements in $(0,\kappa_0)$. There exist two constants ρ_Q and κ_Q such that the following estimate holds for all $\kappa_0 \ge \kappa_Q$:*

$$\|q_{\kappa_0} - \tilde{q}_{\kappa_0}\|_{L^\infty(\mathbb{R})} \le \rho_Q \left\|d(k) - \tilde{d}(k)\right\|_{L^1(0,\kappa_0)}.$$

3. INVERSE SOURCE PROBLEM

In this section, we consider the ISP that determines the unknown current density function from boundary measurements of the radiated fields at multiple wavenumbers. The ISP has many significant applications in biomedical engineering and antenna synthesis [7, 88].

In medical applications, it is often desirable to use the measurement of the radiated electromagnetic field on the surface of the human brain to infer abnormalities inside the brain [68].

3.1. Model problem

Consider the time-harmonic Maxwell equation in a homogeneous medium

$$\nabla \times (\nabla \times E) - \kappa^2 E = \mathrm{i}\kappa J \quad \text{in } \mathbb{R}^3, \tag{3.1}$$

where $\kappa > 0$ is the wavenumber, E is the electric field, J is the electric current density which is assumed to have a compact support Ω. The Silver–Müller radiation condition is required to ensure the well-posedness of the direct problem

$$\lim_{r\to\infty} \big((\nabla \times E) \times x - \mathrm{i}\kappa r E\big) = 0, \quad r = |x|. \tag{3.2}$$

Given $J \in L^2(\Omega)^3$, it is known that the scattering problem (3.1)–(3.2) has a unique solution

$$E(x,\kappa) = \int_\Omega G(x,y;\kappa) \cdot J(y)\mathrm{d}y,$$

where $G(x,y;\kappa)$ is Green's tensor for the Maxwell system (3.1). Explicitly, we have

$$G(x,y;\kappa) = \mathrm{i}\kappa g(x,y;\kappa) I_3 + \frac{\mathrm{i}}{\kappa} \nabla_x \nabla_x^\top g(x,y;\kappa),$$

where g is the fundamental solution of the three-dimensional Helmholtz equation and I_3 is the 3×3 identity matrix.

Let $B_R = \{x \in \mathbb{R}^3 : |x| < R\}$, where R is a positive constant such that $\Omega \subset\subset B_R$. Denote by Γ_R the boundary of B_R. In the domain $\mathbb{R}^3 \setminus B_R$, the solution of (3.1) has a series expansion in the spherical coordinates which may be used to derive the capacity operator T. In addition, it can be verified that the solution of (3.1) satisfies the transparent boundary condition

$$(\nabla \times E) \times \nu = \mathrm{i}\kappa T(E \times \nu) \quad \text{on } \Gamma_R,$$

where ν is the unit outward normal to Γ_R.

Define the boundary measurement in terms of the tangential trace of the electric field

$$\big\|E(\cdot,\kappa) \times \nu\big\|_{\Gamma_R}^2 = \int_{\Gamma_R} \big(\big|T\big(E(x,\kappa) \times \nu\big)\big|^2 + \big|E(x,\kappa) \times \nu\big|^2\big)\mathrm{d}\gamma(x).$$

Let J be the electric current density with the compact support Ω. The ISP of electromagnetic waves is to determine J from the tangential trace of the electric field $E(x,\kappa) \times \nu$ for $x \in \Gamma_R$.

The ISP for the fixed frequency case has been studied extensively. It is now well known that the problem is ill-posed with nonuniqueness and instability [1, 50, 69, 77, 82]. Due to the existence of infinitely many nonradiating fields, a source with extended support cannot be uniquely determined from surface measurements at a fixed frequency. Therefore, additional constraints need to be imposed in order to obtain a unique solution to the inverse problem. A usual choice is to find the source with a minimum energy norm. However, the difference between the minimum energy solution and the original source function could be significant. Another difficulty of the ISP at fixed frequency is the inherited instability due

to exponential decay of the singular eigenvalues of the forward operator [34, 35, 72]. For the special cases of reconstruction for point sources, we refer to [6, 9, 38, 39, 107] for studies of the unique identifiability and stability of the problem.

The use of the multiple frequency data for the ISP provides an approach to circumvent the difficulties of nonuniqueness and instability presented at a fixed frequency. For the ISP of the Helmholtz equation, uniqueness and stability were established in [34] by multiple frequency measurements. The results indicate that the multifrequency ISP is not only uniquely solvable but also is Lipschitz stable when the highest wavenumber exceeds a certain real number.

In the rest of the section, we present our recent results on uniqueness and stability for the ISP of Maxwell's equations [30], and discuss the recent development on the inverse random source problems, where the current density is a random function.

3.2. Uniqueness and stability

Denote by $\mathbb{X}(B_R)$ the closure of the following set in the $L^2(B_R)^3$ norm:

$$\left\{E \in H(\mathrm{curl}, B_R) : \int_{B_R} \big((\nabla \times E)\cdot(\nabla \times \psi) - \kappa^2 E \cdot \psi\big)\mathrm{d}x = 0,\ \forall \psi \in C_0^\infty(B_R)^3\right\}.$$

We have the following orthogonal decomposition of $L^2(B_R)^3$ [1]:

$$L^2(B_R)^3 = \mathbb{X}(B_R) \oplus \mathbb{Y}(B_R),$$

where $\mathbb{Y}(B_R)$ is an infinite-dimensional subspace of $L^2(B_R)^3$ and the electric current densities in the subspace $\mathbb{Y}(B_R)$ are called nonradiating sources. It corresponds to finding a minimum norm solution when computing the component of the source in $\mathbb{X}(B_R)$.

The following two results characterize clearly the uniqueness and nonuniqueness of the ISP. The proofs can be found in [30].

Theorem 3.1. *Suppose $J \in \mathbb{Y}(B_R)$. Then J does not produce any tangential trace of electric fields on Γ_R and thus cannot be identified.*

Theorem 3.2. *Suppose $J \in \mathbb{X}(B_R)$, then J can be uniquely determined by the data $E \times \nu$ on Γ_R.*

Define a functional space

$$\mathbb{J}_M(B_R) = \big\{J \in \mathbb{X}(B_R) \cap H^m(B_R)^3 : \|J\|_{H^m(B_R)^3} \le M\big\},$$

where $m \ge d$ is an integer and $M > 1$ is a constant. The following theorem concerns the stability for the multifrequency ISP (3.1).

Theorem 3.3. *Let E be the solution of the source problem (3.1)–(3.2) corresponding to $J \in \mathbb{J}_M(B_R)$. Then*

$$\|J\|^2_{L^2(B_R)^3} \lesssim \varepsilon^2 + M^2\left(\frac{K^{\frac{2}{3}}|\ln\varepsilon|^{\frac{1}{4}}}{(R+1)(6m-15)^3}\right)^{5-2m},$$

where

$$\varepsilon = \left(\int_0^K \kappa^2 \left\| E(\cdot, \kappa) \times \nu \right\|_{\Gamma_R}^2 d\kappa \right)^{1/2}.$$

The stability result shows that as the highest frequency increases, the stability continues to improve and approaches the Lipschitz type.

3.3. Inverse random source problems

Stochastic inverse problems refer to inverse problems that involve uncertainties due to the unpredictability of the model and incomplete knowledge of the system and measurements. Compared to deterministic counterparts, stochastic inverse problems have substantially more difficulties from randomness and uncertainties. New models and methodologies must be developed for solving stochastic inverse problems.

When the random source is modeled as the white noise, the stochastic ISP is considered for the Helmholtz equation [12, 15, 68, 98]. The goal is to reconstruct the statistical properties of the random source, such as the mean and variance, from boundary measurements of the radiated random wave field. Since the white noise has independent increments, Itô's calculus can be utilized to derive explicit formulas between the statistics of the wave field and the random source. Recently, the model of the microlocally isotropic Gaussian field is developed to handle stochastic processes with correlated increments [95, 97]. The stochastic inverse problem is to determine the microcorrelation strength in the principal symbol from some statistics of the random wave fields. More recently, a new model of the inverse random source problem has been proposed for the stochastic Helmholtz and Maxwell equations [99, 100], where the source is assumed to be driven by a fractional Gaussian field. The new model covers various stochastic processes and allows to deal with rougher sources.

4. INVERSE DIFFRACTION GRATING PROBLEM

For an IOP, the scattering object is a homogeneous obstacle with a given boundary condition. The inverse problem is to determine the obstacle from knowledge of the scattered field away from the obstacle. In this section, we consider the scattering of a time-harmonic electromagnetic plane wave by a (infinite) periodic structure (Figure 6), also known as a grating in diffractive optics, which may be regarded as a special class of the obstacle problem. The scattering problem in this setting is often referred to as the diffraction problem in the literature.

Due to important applications, especially in the design and fabrication of optical elements such as corrective lenses, antireflective interfaces, beam splitters, and sensors, the diffraction problems in periodic structures have been studied extensively. We refer to [17, 26] and references therein for the mathematical studies of the existence and uniqueness questions of the model problems. Numerical methods can be found in [14, 54, 63, 104] for either an integral equation approach or a variational approach. A comprehensive review can be

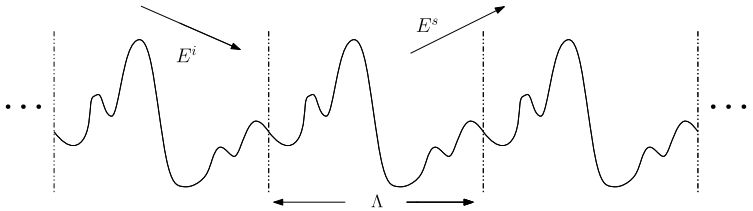


FIGURE 6
The inverse diffraction problem geometry. A plane wave E^i is incident on the surface with period Λ.

found in [16,105] on diffractive optics technology and its mathematical modeling, as well as computational methods.

This section is concerned with the inverse diffraction problem, which is to determine the periodic structure from a reflected field measured at a constant distance away from the structure corresponding to a given incident field. The inverse problem arises naturally in the study of optimal design problems in diffractive optics. The goal is to design a grating structure that gives rise to some specified far-field patterns [11,70].

The mathematical questions on uniqueness and stability for the inverse diffraction problem of both the two-dimensional Helmholtz equation and the three-dimensional Maxwell equations have been studied extensively in [3,10,18,47,49,52,84,92,110]. However, all of the above mentioned results are under fairly restrictive assumptions, or local in nature. A complete answer to the uniqueness question has been given in [47,48] for the determination of a three-dimensional polyhedral periodic diffraction structure by the scattered electromagnetic fields measured above the structure. The result indicates that the uniqueness by any given incident field fails for seven simple classes of regular polyhedral structures. Moreover, if a regular periodic polyhedral structure is not uniquely identifiable by a given incident field, then it belongs to a nonempty class of the seven classes whose elements generate the same total field as the original structure when impinged upon by the same incident field. Problems on global uniqueness or stability for the inverse diffraction problem are still open.

A number of numerical methods have been developed to solve these inverse problems [8,53,73,83,89]. Using a single-layer potential representation, we have presented in [29] an efficient RLM for solving the nonlinear inverse diffraction grating problem in a one-dimensional perfectly reflecting structure. The algorithm requires multifrequency data and the iterative steps are obtained by recursive linearization with respect to the wavenumber: at each step a nonlinear Landweber iteration is applied, with the starting point given by the output from the previous step at a lower wavenumber. Thus, at each stage an approximation to the grating surface filtered at a higher frequency is created. Starting from a reasonable initial guess, the RLM is shown to converge for a larger class of surfaces than the usual Newton's method using the same initial guess.

An extension of the numerical method has been done in [28] for solving the inverse diffraction problem with phaseless data. By using multifrequency data, our algorithm is based on the RLM marching with respect to the wavenumber. With the starting point given

by the output from the previous step at a lower wavenumber, a new approximation to the grating surface filtered at a higher frequency is updated by a Landweber iteration. The numerical results show that the continuation method cannot determine the location of the grating structure, but it can effectively reconstruct the grating shape from the phaseless data.

Another important extension of the method is to solve the inverse diffraction problem by a random periodic structure. Existing studies mostly assume that the periodic structure is deterministic and only the noise level of the measured data is considered for the inverse problem. In practice, however, there is a level of uncertainty of the scattering surface, e.g., the grating structure may have manufacturing defects or it may suffer other possible damages from regular usage. Therefore, in addition to the noise level of measurements, the random surface itself also influences the measured scattered fields. Surface roughness measurements are of great significance for the functional performance evaluation of machined parts and design of microoptical elements. Little is known in mathematics or computation about solving inverse problems of determining random surfaces. One challenge lies in the fact that the scattered fields depend nonlinearly on the surface, which makes the random surface reconstruction problem extremely difficult. Another challenge is to understand to what extend the reconstruction could be made. In other words, what statistical quantities of the profile could be recovered from the measured data? We have recently proposed an efficient numerical method in [36] to reconstruct the random periodic structure from multifrequency scattered fields measured at a constant height above the structure. We demonstrate that three critical statistical properties, namely the expectation, root mean square, and correlation length of the random structure may be reconstructed. Our method is based on a novel combination of the Monte Carlo technique for sampling the probability space, an RLM with respect to the wavenumber, and the Karhunen–Loève expansion of the random structure.

5. DISCUSSIONS AND FUTURE DIRECTIONS

This work is devoted to mathematical analysis and numerical methods for solving inverse scattering problems. On mathematical analysis, we have focused mainly on the stability analysis of the inverse problems. Numerically, we have discussed the recursive linearization approach. These results confirm that the spectral information is vital in stable solution of inverse scattering problems and whenever possible multiple frequency data should be taken and employed for reconstructions. There are tremendous research opportunities for mathematical analysis and numerical methods of inverse scattering problems to meet the continuous growing needs in science and engineering to explore the complex world, from the universe to the new materials, and to the cell. As the computing powers continue to increase and new fast algorithms are developed, inverse scattering problems will continue to contribute to the advancements of the relevant science and engineering.

In the following, we point out some future research directions in line of the research discussed in this work.

For inverse medium scattering problems, we present the stability estimates for the one-dimensional model. In the extreme case when all frequency data is attainable, the esti-

mates have also been obtained in [45, 46]. However, no stability estimate is available for the IMP in several dimensions. By taking into account the uncertainty principle, we conjecture that the reconstruction of the observable part of the scatterer is Lipschitz stable.

For the inverse diffraction problem, global uniqueness remains open. In the polyhedral structure cases, the problem was solved in [47, 48] by using group symmetry properties of the structures and unique continuation properties. For the obstacle scattering problem including the diffraction problem, another interesting problem is to derive the stability estimate with explicit dependence on the wavenumber. The estimate will be particularly useful for convergence analysis of the numerical methods.

In computation, the multifrequency data-based RLM is shown to be stable and effective for solving inverse scattering problems. However, only limited progress has been made on convergence analysis [27, 40, 41, 106]. It is expected that complete analysis should be done by combining the stability estimates and the uncertainty principle. Another difficulty is the incomplete data, including phaseless, limited aperture, or incomplete model. It is interesting to investigate how to employ computational inverse scattering problems to break the diffraction limit. In other words, how to balance the accuracy and resolution. Initial efforts were made on combining the RLM with near-field imaging techniques [31, 32].

Another interesting direction is to study the stochastic inverse scattering problems. As discussed in Section 3.3, initial efforts have been made for solving inverse random source problems. However, little progress is made for inverse medium problems and inverse obstacle problems, where the scatterer and obstacle are respective random functions. It is of interest to consider the more challenging inverse random medium scattering problem. The medium is no longer deterministic and its randomness and uncertainty have to be modeled as well. Since the scattered field depends on the medium or the obstacle nonlinearly, as opposed to the linear dependence on the source, the scattering and inverse scattering problems become much more challenging. In particular, new mathematical and computational frameworks are in demand for solving these problems.

Finally, although the scope of this work is limited to inverse scattering problems in acoustic and electromagnetic waves, we believe many of the methods and techniques discussed here could apply to inverse scattering problems in other wave models. Other emerging topics which beyond the scope of this work but could change the future landscape of solving inverse scattering problems include deep learning type methods [43, 91] and optimal transport methods [75].

ACKNOWLEDGMENTS

The author would like to thank Peijun Li, Jun Lai, Shuai Lu, Faouzi Triki, Xiang Xu, Xiaokai Yuan, and Lei Zhang for useful comments and suggestions during the preparation of the manuscript.

FUNDING

The research was supported in part by an Innovative Group grant of the National Natural Science Foundation of China (No. 11621101).

REFERENCES

[1] R. Albanese and P. Monk, The inverse source problem for Maxwell's equations. *Inverse Probl.* **22** (2006), 1023–1035.

[2] G. Alessandrini, Stable determination of conductivity by boundary measurements. *Appl. Anal.* **27** (1988), 1–3.

[3] H. Ammari, Uniqueness theorems for an inverse problem in a doubly periodic structure. *Inverse Probl.* **11** (1995), 823–833.

[4] H. Ammari, H. Bahouri, D. Dos Santos Ferreira, and I. Gallagher, Stability estimates for an inverse scattering problem at high frequencies. *J. Math. Anal. Appl.* **400** (2013), 525–540.

[5] H. Ammari and G. Bao, Analysis of the scattering map of a linearized inverse medium problem for electromagnetic waves. *Inverse Probl.* **17** (2001), 219–234.

[6] H. Ammari, G. Bao, and J. Fleming, An inverse source problem for Maxwell's equations in magnetoencephalography. *SIAM J. Appl. Math.* **62** (2002), 1369–1382.

[7] T. Angel, A. Kirsch, and R. Kleinmann, Antenna control and generalized characteristic modes. *Proc. IEEE* **79** (1991), 1559–1568.

[8] T. Arens and A. Kirsch, The factorization method in inverse scattering from periodic structures. *Inverse Probl.* **19**(2003), 1195–1211.

[9] A. Badia and T. Nara, An inverse source problem for Helmholtz's equation from the Cauchy data with a single wave number. *Inverse Probl.* **27** (2011), 105001.

[10] G. Bao, A unique theorem for an inverse problem in periodic diffractive optics. *Inverse Probl.* **10** (1994), 335–340.

[11] G. Bao and E. Bonnetier, Optimal design of periodic diffractive structures. *Appl. Math. Optim.* **43** (2001), 103–116.

[12] G. Bao, C. Chen, and P. Li, Inverse random source scattering problems in several dimensions. *SIAM/ASA J. Uncertain. Quantificat.* **4** (2016), 1263–1287.

[13] G. Bao, Y. Chen, and F. Ma, Regularity and stability for the scattering map of a linearized inverse medium problem. *J. Math. Anal. Appl.* **247** (2000), 255–271.

[14] G. Bao, Z. Chen, and H. Wu, Adaptive finite-element method for diffraction grating. *J. Opt. Soc. Amer. A* **22** (2005), 1106–1114.

[15] G. Bao, S.-N. Chow, P. Li, and H. Zhou, An inverse random source problem for the Helmholtz equation. *Math. Comp.* **83** (2014), 215–233.

[16] G. Bao, L. Cowsar, and W. Masters, eds., *Mathematical modeling in optical science*. Frontiers Appl. Math., 22, SIAM, 2001.

[17] G. Bao, D. Dobson, and J. Cox, Mathematical studies in rigorous grating theory. *J. Opt. Soc. Amer. A* **12** (1995), 1029–1042.

[18] G. Bao and A. Friedman, Inverse problems for scattering by periodic structure. *Arch. Ration. Mech. Anal.* **132** (1995), 49–72.

[19] G. Bao, S. Hou, and P. Li, Inverse scattering by a continuation method with initial guesses from a direct imaging algorithm. *J. Comput. Phys.* **227** (2007), 755–762.

[20] G. Bao and P. Li, Inverse medium scattering for three-dimensional time harmonic Maxwell's equations. *Inverse Probl.* **20** (2004), L1–L7.

[21] G. Bao and P. Li, Inverse medium scattering problems for electromagnetic waves. *SIAM J. Appl. Math.* **65** (2005), 2049–2066.

[22] G. Bao and P. Li, Inverse medium scattering for the Helmholtz equation at fixed frequency. *Inverse Probl.* **21** (2005), 1621–1641.

[23] G. Bao and P. Li, Inverse medium scattering problems in near-field optics. *J. Comput. Math.* **25** (2007), 252–265.

[24] G. Bao and P. Li, Numerical solution of inverse scattering for near-field optics. *Optim. Lett.* **32** (2007), 1465–1467.

[25] G. Bao and P. Li, Numerical solution of an inverse medium scattering problem for Maxwell's equations at fixed frequency. *J. Comput. Phys.* **228** (2009), 4638–4648.

[26] G. Bao and P. Li, *Maxwell's equations in periodic structures*. Ser. Appl. Math. Sci. Vol. 208, Springer, 2022.

[27] G. Bao, P. Li, J. Lin, and F. Triki, Inverse scattering problems with multi-frequencies. *Inverse Probl.* **31** (2015), 093001.

[28] G. Bao, P. Li, and J. Lv, Numerical solution of an inverse diffraction grating problem from phaseless data. *J. Opt. Soc. Amer. A* **30** (2013), 293–299.

[29] G. Bao, P. Li, and H. Wu, A computational inverse diffraction grating problem. *J. Opt. Soc. Amer. A* **29** (2012), 394–399.

[30] G. Bao, P. Li, and Y. Zhao, Stability for the inverse source problems in elastic and electromagnetic waves. *J. Math. Pures Appl.* **134** (2020), 122–178.

[31] G. Bao and J. Lin, Imaging of local surface displacement on an infinite ground plane: the multiple frequency case. *SIAM J. Appl. Math.* **71** (2011), 1733–1752.

[32] G. Bao and J. Lin, Imaging of reflective surfaces by near-field optics. *Optim. Lett.* **37** (2012), 5027–5029.

[33] G. Bao, J. Lin, and S. Mefire, Numerical reconstruction of electromagnetic inclusions in three dimensions. *SIAM J. Imaging Sci.* **7** (2014), 558–577.

[34] G. Bao, J. Lin, and F. Triki, A multi-frequency inverse source problem. *J. Differential Equations* **249** (2010), 3443–3465.

[35] G. Bao, J. Lin, and F. Triki, Numerical solution of the inverse source problem for the Helmholtz equation with multiple frequency data. *Contemp. Math.* **548** (2011), 45–60.

[36] G. Bao, Y. Lin, and X. Xu, Inverse scattering by a random periodic structure. *SIAM J. Numer. Anal.* **58** (2020), 2934–2952.

[37] G. Bao and J. Liu, Numerical solution of inverse problems with multi-experimental limited aperture data. *SIAM J. Sci. Comput.* **25** (2003), 1102–1117.

[38] G. Bao, Y. Liu, and F. Triki, Recovering simultaneously a potential and a point source from Cauchy data. *Minimax Theory Appl.* **6** (2021), 227–238.

[39] G. Bao, Y. Liu, and F. Triki, Recovering point sources for the inhomogeneous Helmholtz equation. *Inverse Probl.* **37** (2021), 095005.

[40] G. Bao, S. Lu, W. Rundell, and B. Xu, A recursive algorithm for multi-frequency acoustic inverse source problems. *SIAM J. Numer. Anal.* **53** (2015), 1608–1628.

[41] G. Bao and F. Triki, Error estimates for the recursive linearization for solving inverse medium problems. *J. Comput. Math.* **28** (2010), 725–744.

[42] G. Bao and F. Triki, Stability estimates for the 1D multifrequency inverse medium problem. *J. Differential Equations* **269** (2020), 7106–7128.

[43] G. Bao, X. Ye, Y. Zang, and H. Zhou, Numerical solution of inverse problems by weak adversarial networks. *Inverse Probl.* **36** (2020), 115003.

[44] G. Bao, T. Yin, and F. Zeng, Multifrequency iterative methods for the inverse medium scattering problems in elasticity. *SIAM J. Sci. Comput.* **41** (2019), B721–B745.

[45] G. Bao and K. Yun, On the stability of an inverse problem for the wave equation. *Inverse Probl.* **25** (2009), 045003.

[46] G. Bao and H. Zhang, Sensitive analysis of an inverse problem for the wave equation with caustics. *J. Amer. Math. Soc.* **27** (2014), 953–981.

[47] G. Bao, H. Zhang, and J. Zou, Unique determination of periodic polyhedral structures by scattered electromagnetic fields. *Trans. Amer. Math. Soc.* **363** (2011), 4527–4551.

[48] G. Bao, H. Zhang, and J. Zou, Unique determination of periodic polyhedral structures by scattered electromagnetic fields. Part II, The resonance case. *Trans. Amer. Math. Soc.* **366** (2014), 1333–1361.

[49] G. Bao and Z. Zhou, An inverse problem for scattering by a doubly periodic structure. *Trans. Amer. Math. Soc.* **350** (1998), 4089–4103.

[50] N. Bleistein and J. Cohen, Nonuniqueness in the inverse source problem in acoustics and electromagnetics. *J. Math. Phys.* **18** (1977), 194–201.

[51] C. Borges, A. Gillman, and L. Greengard, High resolution inverse scattering in two dimensions using recursive linearization. *SIAM J. Imaging Sci.* **10** (2017), 641–664.

[52] G. Bruckner, J. Cheng, and M. Yamamoto, An inverse problem in diffractive optics: conditional stability. *Inverse Probl.* **18** (2002), 415–433.

[53] G. Bruckner and J. Elschner, A two-step algorithm for the reconstruction of perfectly reflecting periodic profiles. *Inverse Probl.* **19** (2003), 315–329.

[54] O. Bruno and F. Reitich, Numerical solution of diffraction problems: a method of variation of boundaries. *J. Opt. Soc. Amer. A* **10** (1993), 1168–1175.

[55] F. Cakoni and D. Colton, *Qualitative methods in inverse scattering theory: an introduction*. Springer, Berlin, 2005.

[56] A. Calderón, On an inverse boundary value problem. In *Seminar on numerical analysis and its applications to continuum physics* (Rio de Janeiro, 1980), pp. 65–73, Soc. Brasil. Mat., 1980.

[57] S. Carney and J. Schotland, Inverse scattering for near-field microscopy. *App. Phys. Lett.* **77** (2000), 2798–2800.

[58] S. Carney and J. Schotland, Near-field tomography. *MSRI Ser. Math. Appl.* **47** (2003), 131–166.

[59] Y. Chen, Inverse scattering via Heisenberg uncertainty principle. *Inverse Probl.* **13** (1997), 253–282.

[60] Y. Chen, Inverse scattering via skin effect. *Inverse Probl.* **13** (1997), 649–667.

[61] Y. Chen and V. Rokhlin, On the Riccati equations for the scattering matrices in two dimensions. *Inverse Probl.* **13** (1997), 1–13.

[62] Z. Chen and G. Huang, A direct imaging method for electromagnetic scattering data without phase information. *SIAM J. Imaging Sci.* **9** (2016), 1273–1297.

[63] Z. Chen and H. Wu, An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures. *SIAM J. Numer. Anal.* **41** (2003), 799–826.

[64] W. Chew and Y. Wang, Reconstruction of two-dimensional permittivity distribution using the distorted Born iteration method. *IEEE Trans. Med. Imag.* **9** (1990), 218–225.

[65] D. Colton and A. Kirsch, A simple method for solving inverse scattering problems in the resonance region. *Inverse Probl.* **12** (1996), 383–393.

[66] D. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory*. Springer, Berlin, 1998.

[67] D. Courjon, *Near-field microscopy and near-field optics*. Imperial College Press, London, 2003.

[68] A. Devaney, The inverse problem for random sources. *J. Math. Phys.* **20** (1979), 1687–1691.

[69] A. Devaney, E. Marengo, and M. Li, The inverse source problem in nonhomogeneous background media. *SIAM J. Appl. Math.* **67** (2007), 1353–1378.

[70] D. Dobson, Optimal design of periodic antireflective structures for the Helmholtz equation. *European J. Appl. Math.* **4** (1993), 321–340.

[71] O. Dorn, H. Bertete-Aguirre, J. Berrymann, and G. Papanicolaou, A nonlinear inversion method for 3D electromagnetic imaging using adjoint fields. *Inverse Probl.* **15** (1999), 1523–1558.

[72] M. Eller and N. Valdivia, Acoustic source identification using multiple frequency information. *Inverse Probl.* **25** (2009), 115005.

[73] J. Elschner, G. Hsiao, and A. Rathsfeld, Grating profile reconstruction based on finite elements and optimization techniques. *SIAM J. Appl. Math.* **64** (2003), 525–545.

[74] H. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*. Kluwer, Dordrecht, 1996.

[75] B. Engquist, K. Ren, and Y. Yang, The quadratic Wasserstein metric for inverse data matching. *Inverse Probl.* **36** (2020), 055001.

[76] K. Erhard and R. Potthast, A numerical study of the probe method. *SIAM J. Sci. Comput.* **28** (2006), 1597–1612.

[77] A. Fokas, Y. Kurylev, and V. Marinakis, The unique determination of neuronal currents in the brain via magnetoencephalogrphy. *Inverse Probl.* **20** (2004), 1067–1082.

[78] I. M. Gelfand and B. M. Levitan, On the determination of a differential equation from its spectral functions. *Amer. Math. Soc. Transl. Ser. 2* **1** (1955), 253–304.

[79] C. Girard and A. Dereux, Near-field optics theories. *Rep. Progr. Phys.* **59** (1996), 657–699.

[80] H. Haddar and P. Monk, The linear sampling method for solving the electromagnetic inverse medium problem. *Inverse Probl.* **18** (2002), 891–906.

[81] P. Hähner and T. Hohage, New stability estimates for the inverse acoustic inhomogeneous a medium problem and applications. *SIAM J. Math. Anal.* **62** (2001), 670–685.

[82] S. He and V. Romanov, Identification of dipole equations. *Wave Motion* **28** (1998), 25–44.

[83] F. Hettlich, Iterative regularization schemes in inverse scattering by periodic structures. *Inverse Probl.* **18** (2002), 701–714.

[84] F. Hettlich and A. Kirsch, Schiffer's theorem in inverse scattering theory for periodic structures. *Inverse Probl.* **13** (1997), 351–361.

[85] T. Hohage, On the numerical solution solution of a three-dimensional inverse medium scattering problem. *Inverse Probl.* **17** (2001), 1743–1763.

[86] T. Hohage, Fast numerical solution of the electromagnetic medium scattering problem and applications to the inverse problem. *J. Comput. Phys.* **214** (2006), 224–238.

[87] S. Hou, K. Sølna, and H. Zhao, A direct imaging algorithm using far-field data. *Inverse Probl.* **23** (2007), 1533–1546.

[88] V. Isakov, *Inverse source problems*. Math. Surveys Monogr. 34, American Mathematical Society, Providence, RI, 1989.

[89] K. Ito and F. Reitich, A high-order perturbation approach to profile reconstruction: I. Perfectly conducting gratings. *Inverse Probl.* **15** (1999), 1067–1085.

[90] B. Kaltenbacher, A. Neubauer, and O. Scherzer, Convergence of projected iterative regularization methods for nonlinear problems with smooth solutions. *Inverse Probl.* **22** (2006), 1105–1119.

[91] Y. Khoo and L. Ying, SwitchNet: a neural network model for forward and inverse scattering problems. *SIAM J. Sci. Comput.* **41** (2019), A3182–A3201.

[92] A. Kirsch, Uniqueness theorems in inverse scattering theory for periodic structures. *Inverse Probl.* **10** (1994), 145–152.

[93] M. Klibanov and V. Romanov, Two reconstruction procedures for a 3D phaseless inverse scattering problem for the generalized Helmholtz equation. *Inverse Probl.* **32** (2016), 015005.

[94] P. D. Lax and R. S. Phillips, *Scattering theory*. Academic Press, New York, 1967.

[95] M. Lassas, L. Päivärinta, and E. Saksman, Inverse scattering problem for a two dimensional random potential. *Comm. Math. Phys.* **279** (2008), 669–703.

[96] R.M. Lewis and W. Symes, On the relation between the velocity coefficient and boundary value for solutions of the one-dimensional wave equation. *Inverse Probl.* **7** (1991), 597631.

[97] J. Li, T. Helin, and P. Li, Inverse random source problems for time-harmonic acoustic and elastic waves. *Comm. Partial Differential Equations* **45** (2020), 1335–1380.

[98] P. Li, An inverse random source scattering problem in inhomogeneous media. *Inverse Probl.* **27** (2011), 035004.

[99] P. Li and X. Wang, Inverse random source scattering for the Helmholtz equation with attenuation. *SIAM J. Appl. Math.* **81** (2021), 485–506.

[100] P. Li and X. Wang, An inverse random source problem for Maxwell's equations. *Multiscale Model. Simul.* **19** (2021), 25–45.

[101] S. Nagayasu, G. Uhlmann, and J.-N. Wang, Increasing stability in an inverse problem for the acoustic equation. *Inverse Probl.* **29** (2013), 025012.

[102] F. Natterer, *The mathematics of computerized tomography*. Teubner, Stuttgart, 1986.

[103] J.-C. Nédélec, *Acoustic and electromagnetic equations: integral representations for harmonic problems*. Springer, New York, 2000.

[104] J.-C. Nédélec and F. Starling, Integral equation methods in a quasi-periodic diffraction problem for the time harmonic Maxwell's equations. *SIAM J. Math. Anal.* **22** (1991), 1679–1701.

[105] R. Petit, ed., *Electromagnetic theory of gratings*. Topics in Curr. Phys. 22, Springer, 1980.

[106] M. Sini and N. Thanh, Inverse acoustic obstacle scattering problems using multi-frequency measurements. *Inverse Probl. Imaging* **6** (2012), 749–773.

[107] R. Schmidt, Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propagation* **34** (1986), 276–280.

[108] P. Stefanov and G. Uhlmann, Stability estimates for the hyperbolic Dirichlet to Neumann map in anisotropic media. *J. Funct. Anal.* **154** (1998), 330–358.

[109] M. Vögeler, Reconstruction of the three-dimensional refractive index in electromagnetic scattering by using a propagation–backpropagation method. *Inverse Probl.* **19** (2003), 739–753.

[110] J. Yang and B. Zhang, Uniqueness results in the inverse scattering problem for periodic structures. *Math. Methods Appl. Sci.* **35** (2012), 828–838.

GANG BAO (包 刚)

School of Mathematical Sciences, Zhejiang University, Hangzhou, 310027, China,
baog@zju.edu.cn

TOWARDS ADAPTIVE SIMULATIONS OF DISPERSIVE TSUNAMI PROPAGATION FROM AN ASTEROID IMPACT

MARSHA J. BERGER AND RANDALL J. LEVEQUE

ABSTRACT

The long-term goal of this work is the development of high-fidelity simulation tools for dispersive tsunami propagation. A dispersive model is especially important for short wavelength phenomena such as an asteroid impact into the ocean, and is also important in modeling other events where the simpler shallow water equations are insufficient. Adaptive simulations are crucial to bridge the scales from deep ocean to inundation, but have difficulties with the implicit system of equations that results from dispersive models. We propose a fractional step scheme that advances the solution on separate patches with different spatial resolutions and time steps. We show a simulation with 7 levels of adaptive meshes and onshore inundation resulting from a simulated asteroid impact off the coast of Washington. Finally, we discuss a number of open research questions that need to be resolved for high quality simulations.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 65N50; Secondary 76B15, 76-04, 86-04

KEYWORDS

Adaptive mesh refinement, implicit methods, dispersion, Boussinesq, tsunami, asteroid ocean impact

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5056–5071
DOI 10.4171/ICM2022/73

1. INTRODUCTION

Many steps are required in modeling a tsunami arising from an asteroid impact in the ocean. The impact itself forms a crater that drives the eventual tsunami creation. Modeling this requires a complex three-dimensional multiphysics hydrocode, since there are many physical processes and time scales. Once the tsunami has formed, it propagates hundreds or thousands of kilometers across the ocean. When the shoreline is reached, the ultimate goal is modeling the inundation risk to coastal populations and important infrastructure at a much smaller spatial scale (typically 10 meters or less).

This work addresses the last two steps, the long-distance propagation and coastal inundation. The goal is a high-fidelity model that can accurately determine the inundation risk for particular sites using available bathymetric data sets. Since large-scale ocean simulations are so compute-intensive, for many tsunami modeling problems the two-dimensional depth-averaged shallow water equations (SWEs) are used for the propagation step. These equations assume the wavelength is long relative to the depth of the ocean, as is typical for tsunamis generated by large earthquakes. However, these nondispersive equations are often insufficient for short-wavelength asteroid-generated tsunamis, giving inaccurate results for both tsunami travel time and maximum shoreline run-in, as noted in our own work [5] and in several other studies, e.g., [14, 25, 27]. This is also the case for landslide-generated tsunamis and other short wavelength phenomena; see, e.g., [9]. Shorter waves experience significant dispersion (waves with different periods propagate with different speeds), while the hyperbolic SWEs are nondispersive. Dispersive depth-averaged equations can be obtained by retaining more terms when reducing from the three-dimensional Euler equations to two space dimensions, giving some form of "Boussinesq equations." The additional terms involve higher-order derivatives (typically third order), and several different models have been proposed. When solved numerically, these equations generally require implicit methods in order to remain stable with physically reasonable time steps. By contrast, the hyperbolic SWEs involve only first-order derivatives and explicit methods are commonly used.

Our numerical model is based on the GeoClaw softwave (part of the open source Clawpack software project [1]), which has been heavily used and well validated for modeling earthquake-generated tsunamis using the SWE [4,11,16]. The numerical methods used are high-resolution, shock-capturing finite volume methods based on Riemann solvers, a standard approach for nonlinear hyperbolic problems [15]. In the case of GeoClaw, additional features are included to make the methods "well balanced" so that the steady state of an ocean at rest is preserved. Moreover, the shoreline is represented as an interface between wet and dry cells, and robust Riemann solvers allow the determination of the fluxes at these interfaces. The wet/dry status of a cell can change dynamically as the tsunami advances onshore or retreats. This software implements adaptive mesh refinement (AMR), critical for solving problems with vastly different spatial scales from transocean propagation to community-level inundation modeling. However, the patch-based AMR algorithms are based on the use of explicit solvers, and the extension of the GeoClaw software to also work with implicit solvers for Boussinesq equations has been a major part of this project. The basic approach

used can also be used more generally with the AMR version of Clawpack, which has many potential applications to other wave propagation problems when dispersive or dissipative (e.g., second-order derivative) terms are included and implicit solvers are needed.

2. OVERVIEW OF APPROACH

We are building on the work of [12], in which the Boussinesq equations described in the next section were solved using an extension of GeoClaw, but only for the case of a single grid resolution, without the AMR capability. These equations have the form of the two-dimensional SWEs with the addition of "source terms" involving third-order derivatives. Equations of this type can often be solved by fractional step or splitting methods: first advancing the solution by solving the hyperbolic shallow water equations, and then advancing the solution using terms associated with the higher-order derivatives (or in the opposite order, as we have found to be advantageous). For the SWE part, we can use the standard GeoClaw solver. The third derivative terms require an implicit step as described below. We are currently using a sparse linear system solver called Pardiso [6]; several other groups have found that multigrid works nicely as well.

The difficulty in the solution algorithm comes from the combination of adaptive mesh refinement and implicit solvers. Adaptive mesh refinement is critical in bridging the scales of oceanic tsunamic propagation, which typically needs resolution on the order of kilometers, and inundation modeling, where a resolution on the order of 10 meters or less is required. The patch-based mesh refinement in GeoClaw refines in time as well as space, in order to satisfy the time step stability constraint of explicit methods. If a patch is refined in space by a factor of 4, then we also typically refine in time by the same factor (as required by the CFL condition for explicit methods) and so 4 time steps are taken for the fine patch to "catch up" to the coarse patch in time. The fine patch thus needs to interpolate ghost cells that fill out the stencil at each intermediate time step. Figure 1 indicates this schematically for a refinement by 2.

In our approach, the solution of the implicit equations is stored as additional elements in the solution vector, increasing the number of equations in two horizontal dimensions from 3 to 5. We also reverse the typical splitting order and perform the implicit solving first. At the initial and final times during this time step, the ghost cells are still interpolated in space from the coarse grid, but do not need interpolation in time. The ghost cell values at intermediate times on the fine grid are interpolated from the coarse grid values at times n and $n + 1$.

There are other issues to consider when combining Boussinesq and SWEs in a single solver. The Boussinesq equations give a better model of dispersive waves over some regime, but lack any wave breaking mechanism. As large waves approach shore, this can lead to very large magnitude solitary waves that should break. The nonlinear SWEs perform better at this point; a shock wave develops that is a better representation of the turbulent bore formed by a breaking wave. Very large waves, such as those that might be formed by an asteroid impact, can undergo shoaling far out on the continental shelf and dissipate some of their

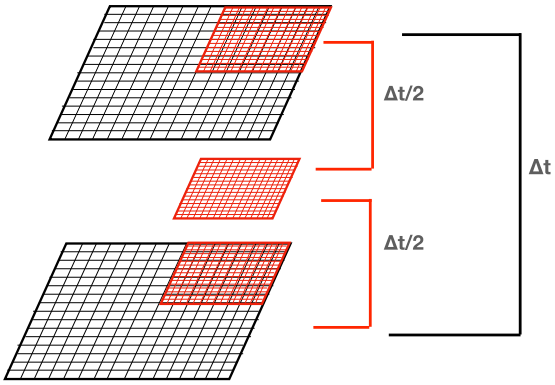


FIGURE 1

Figure shows a coarse grid with coarse cells outlined in the base grid, and a fine patch refined by a factor of 2, also with cells outlined. The fine grid time step is half the coarse grid step. Before the fine grid takes a step, ghost cells are needed to complete the stencil.

energy in this manner (the van Dorn effect [13]). (Shoaling refers to the modification of wave heights when the wave enters shallower water, when the wave steepens and becomes of higher frequency). So it is important to transition from using the Boussinesq equations in deep water to SWEs closer to shore, based on some breaking criterion. Also the onshore flooding is well modeled by SWEs, which is fortunate since the wetting-and-drying algorithms of GeoClaw can then be used. In our initial work we have simply suppressed the higher-order derivative terms (switching to SWEs) wherever the initial water depth was 10 m or less. A better wave breaking model that allows dynamic switching has not yet been implemented but will ultimately be incorporated.

Another issue to consider is the initial conditions for the simulation. For the asteroid impact problem, even the Boussinesq equations are not adequate to model the original generation or evolution of a deep crater in the ocean. We must start with the results of a three-dimensional multiphysics hydrocode simulation and produce suitable initial conditions for the depth-averaged equations. We discuss this further in Section 4.

3. EQUATIONS AND NUMERICAL METHODS

For simplicity we present the equations and algorithm primarily in one space dimension and time, since this is sufficient to illustrate the main ideas.

3.1. The shallow water and Boussinesq equations

The shallow water equations can be written as

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x &= -ghB_x, \end{aligned} \tag{3.1}$$

where $h(x,t)$ = water depth, $B(x)$ = topography ($B<0$ offshore), $\eta(x,t)=B(x)+h(x,t)$, with $\eta=0$ being the sea level. Thus $h_0(x)=-B(x)$ is the depth of water at rest. The depth-averaged horizontal velocity is $u(x,t)$, and so hu is the momentum, and finally, $g=9.81$ is the gravitational constant. These equations are a long wavelength approximation to the Euler equations, in the limit of small ocean depth relative to the wavelength L of the disturbance. For earthquake generated tsunamis, a typical ocean depth might be 4 km, and a subduction zone can have a wavelength of 100 km or more, giving a small ratio.

Equations (3.1) have the form of a hyperbolic system of conservation laws with a source term in the momentum equation that is nonzero only on varying topography. The GeoClaw implementation incorporates the topography term into the Riemann solvers in order to obtain a well-balanced method [16], which amounts to solving equations (3.1) in the nonconservative form

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + (hu^2)_x + ghh_x &= 0. \end{aligned} \tag{3.2}$$

Peregrine [22] derived a Boussinesq-type extension on a flat bottom in the form

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + (hu^2)_x + ghh_x - \frac{1}{3}h_0^2(hu)_{txx} &= 0. \end{aligned} \tag{3.3}$$

These equations have some drawbacks, however, and do not match the dispersion relation of the Euler equations as well as other models developed more recently. (For a historical review of Boussinesq-type models, see [7].)

Madsen and Sorenson [18] and Shaffer and Madsen [26] optimized the equations by adding a term with a parameter B_1 that could be chosen to match the water wave dispersion relation more closely. On general topography, they obtained

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + (hu^2)_x + ghh_x &= \left(B_1+\frac{1}{2}\right)h_0^2(hu)_{txx} + \frac{1}{6}h_0^3(hu/h_0)_{txx} - B_1h_0^2g(h_0\eta_x)_{xx}, \end{aligned} \tag{3.4}$$

where $h_0(x)$ is the initial water depth, and matching the dispersion relation leads to an optimal $B_1=1/15$.

Equations (3.4) appear to have the form of the SWEs (3.2) together with source terms on the right-hand side. However, the standard fractional step approach cannot be used for equations in this form because the source term involves t-derivatives.

These equations can be rewritten as

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + (hu^2)_x + ghh_x - D_{11}((hu)_t) &= gB_1h_0^2(h_0\eta_x)_{xx}, \end{aligned} \tag{3.5}$$

where the differential operator D_{11} is defined by

$$D_{11}(w) = (B_1+1/2)h_0^2w_{xx} - \frac{1}{6}h_0^3(w/h_0)_{xx}.$$

Now subtracting $D_{11}((hu^2)_x + gh\eta_x)$ from both sides of the momentum equation from (3.5) gives

$$\begin{aligned} h_t + (hu)_x &= 0, \\ [I - D_{11}]\big[(hu)_t + (hu^2)_x + gh\eta_x\big] &= -D_{11}\big[(hu^2)_x + gh\eta_x\big] + gh_0^2 B_1 (h_0\eta_x)_{xx}. \end{aligned} \tag{3.6}$$

By inverting $(I - D_{11})$, we get

$$\begin{aligned} h_t + (hu)_x &= 0, \\ (hu)_t + (hu^2)_x + gh\eta_x &= \psi, \end{aligned} \tag{3.7}$$

where ψ is computed by solving an elliptic system:

$$[I - D_{11}]\psi = -D_{11}\big[(hu^2)_x + gh\eta_x\big] + gh_0^2 B_1 (h_0\eta_x)_{xx}. \tag{3.8}$$

System (3.7) now looks like SWEs plus a source term ψ that involves only spatial derivatives.

3.2. Two-dimensional versions

For completeness, we include the two-dimensional version of these equations to show that they have a similar structure. In two dimensions, let $\vec{u} = (u, v)$ be the two horizontal (depth-averaged) velocities. Then the shallow water equations take the form

$$\begin{aligned} h_t + \nabla \cdot (h\vec{u}) &= 0, \\ (h\vec{u})_t + \vec{z} + gh\nabla\eta &= 0, \end{aligned} \tag{3.9}$$

where

$$\vec{z} = \vec{u}\,\nabla \cdot (h\vec{u}) + (h\vec{u} \cdot \nabla)\vec{u} = \begin{bmatrix} (hu^2)_x + (huv)_y \\ (huv)_x + (hv^2)_y \end{bmatrix}. \tag{3.10}$$

The Boussinesq equations of [18, 26], as used in [12], take the form

$$\begin{aligned} h_t + \nabla \cdot (h\vec{u}) &= 0, \\ (h\vec{u})_t + \vec{z} + gh\nabla\eta - D(h\vec{u})_t - gB_1 h_0^2 \nabla\big(\nabla \cdot (h_0\nabla\eta)\big) &= 0, \end{aligned} \tag{3.11}$$

where now the 2×2 matrix D consists of four linear differential operators,

$$D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, \tag{3.12}$$

with

$$\begin{aligned} D_{11}(w) &= (B_1 + 1/2)h_0^2 w_{xx} - \frac{1}{6}h_0^3 (w/h_0)_{xx}, \\ D_{12}(w) = D_{21}(w) &= (B_1 + 1/2)h_0^2 w_{xy} - \frac{1}{6}h_0^3 (w/h_0)_{xy}, \\ D_{22}(w) &= (B_1 + 1/2)h_0^2 w_{yy} - \frac{1}{6}h_0^3 (w/h_0)_{yy}. \end{aligned} \tag{3.13}$$

As in 1D, in order to apply a fractional step method, we subtract $D(\vec{z} + gh\nabla\eta)$ from both sides of the momentum equation of (3.11) so that it becomes

$$[I - D]\big[(h\vec{u})_t + \vec{z} + gh\nabla\eta\big] = -D(\vec{z} + gh\nabla\eta) + gB_1 h_0^2 \nabla\big(\nabla \cdot (h_0\nabla\eta)\big). \tag{3.14}$$

Inverting $[I - D]$ allows rewriting (3.11) as the SWEs with a source term,

$$\begin{aligned} h_t + \nabla \cdot (h\vec{u}) &= 0, \\ (h\vec{u})_t + \vec{z} + gh\nabla\eta &= \vec{\psi}, \end{aligned} \tag{3.15}$$

where $\vec{\psi}$ involves only spatial derivatives and is determined by solving the elliptic equation

$$[I - D]\vec{\psi} = -D(\vec{z} + gh\nabla\eta) + gB_1h_0^2\nabla\big(\nabla \cdot (h_0\nabla\eta)\big). \tag{3.16}$$

Discretizing this elliptic equation leads to a nonsymmetric sparse matrix with much wider bandwidth than the tridiagonal matrix that arises in one dimension due to the cross-derivative terms. Other than the significant increase in computing time required, the fractional step and adaptive mesh refinement algorithms described below carry over directly to the two-dimensional situation.

3.3. Numerics

We return to the one-dimensional equations in order to describe the numerical algorithm and reformulation for patch-based adaptive refinement in space and time.

We solve the one-dimensional Boussinesq equations (3.7), in which ψ is determined as the solution to (3.8), by using a fractional step method with the following steps:

(1) Solve the elliptic equation (3.8) for the source term ψ. After this step the ψ values are saved on each patch to use as boundary conditions for finer patches.

(2) Update the momentum by solving $(hu)_t = \psi$ over the time step (e.g., with forward Euler or two-stage Runge–Kutta method). The depth h does not change in this step since there is no source term in the h_t equation.

(3) Take a step with the homogeneous SWE, using the results of step 2 as initial data. This step uses the regular GeoClaw software and Riemann solvers.

We solve the implicit system first and then take the shallow water step because this facilitates interpolating in time for values required on the edge of grid patches. In order to explain this in more detail, we introduce some notation for a simple case in one space dimension.

First suppose we only have a single grid at one resolution, with no AMR. We denote the numerical solution at some time t_N by $(H, HU)^N$, the cell-averaged approximations to depth and momentum on the grid. We also use Ψ^N for the source term determined by solving the discrete elliptic system defined by $(H, HU)^N$ on this grid. We also assume at the start of the time step that we have boundary conditions for (H, HU) and also for the Boussinesq correction Ψ, provided in the form of "ghost cell" values in a layer of cells surrounding the grid (or two layers in the case of (H, HU) since the high-resolution explicit methods used have a stencil of width 5 because of slope limiters). On a single grid we assume that it is sufficient to use the Dirichlet condition $\Psi = 0$ in all ghost cells surrounding the grid, i.e., that there is no Boussinesq correction in these cells. This is reasonable for a large domain

where the waves of interest are confined to the interior of the domain. We also use zero-order extrapolation boundary condition for (H, HU), which give a reasonable nonreflecting boundary condition for the SWEs step [16].

A single time step of the fractional step algorithm on this grid then takes the following form in order to advance $(H, HU)^N$ to $(H, HU)^{N+1}$ at time $t_{N+1} = t_N + \Delta t$:

(1) Solve the elliptic system for Ψ^N. The right-hand side depends on $(H, HU)^N$.

(2) Advance the solution using the source terms (Boussinesq corrections): $H^* = H^N$, $(HU)^* = (HU)^N + \Delta t \Psi^N$ (using forward Euler, for example).

(3) Take a time step of length Δt with the SWE solver, with initial data $(H, HU)^*$, to obtain $(H, HU)^{N+1}$. We denote this by $(H, HU)^{N+1} = SW((H, HU)^*, \Delta t)$.

In the software it is convenient to store the source term at each time as another component of the solution vector, so we also use $Q^N = (H, HU, \Psi)^N$ to denote this full solution at time t_N.

Now suppose we have two grid levels with refinement by a factor of 2 in time. We denote the coarse grid values at some time t_N as above. We assume that the fine grid is at time t_N, but that on the fine grid we must take two time steps of $\Delta t/2$ to reach time t_{N+1}. We denote the fine grid values at time t_N using lower case, $(h, hu)^N$ and $q^N = (h, hu, \psi)^N$. We also need boundary conditions in the ghost cells of the fine grid patch. If a patch edge is coincident with a domain boundary, then we use the Dirichlet BC $\psi = 0$ and extrapolation BCs for (h, hu), as described above. For ghost cells that are interior to the coarse grid, we let $\mathcal{I}_f(Q)$ represent a spatial interpolation operator that interpolates from coarse grid values to the ghost cells of a fine grid patch at time t_N. This operator is applied to all three components of Q^N, i.e., to the source term, as well as the depth and momentum, in order to obtain the necessary boundary conditions for q^N.

Then one time step on the coarse grid, coupled with two time steps on the fine grid, is accomplished by the following steps:

(1) Coarse grid step:

  (a) Take time step Δt on the coarse grid as described above for the single grid algorithm, but denote the result by $(\tilde{H}, \widetilde{HU})^{N+1}$ since these provisional values will later be updated.

  (b) Using the Dirichlet BCs $\Psi = 0$ on the domain boundary, solve for a provisional $\widetilde{\Psi}^{N+1}$. This will be needed for interpolation in time when determining boundary conditions for ψ on the fine grid, using $\mathcal{I}_f(Q^N)$ and $\mathcal{I}_f(\tilde{Q}^{N+1})$.

(2) Fine grid steps:

  (a) Given $(h, hu)^N$ and boundary conditions $\mathcal{I}_f(Q^N)$, solve the elliptic system for ψ^N.

(b) Update using the source terms, $(h, hu)^* = (h, hu)^N + (0, \frac{\Delta t}{2}\psi^N)$.

(c) Take a shallow water step, $(h, hu)^{N+1/2} = SW((h, hu)^*, \Delta t/2)$. Note that we use $t_{N+1/2} = t_N + \Delta t/2$ to denote the intermediate time.

(d) Obtain BCs at this intermediate time as $\frac{1}{2}(\mathcal{I}_f(Q^N) + \mathcal{I}_f(\tilde{Q}^{N+1}))$.

(e) Solve the elliptic system for $\psi^{N+1/2}$.

(f) Update using the source terms, $(h, hu)^* = (h, hu)^{N+1/2} + (0, \frac{\Delta t}{2}\psi^{N+1/2})$.

(g) Take a shallow water step: $(h, hu)^{N+1} = SW((h, hu)^*, \Delta t/2)$.

(3) Update coarse grid:

(a) Define $(H, HU)^{N+1}$ by the provisional values by $(\tilde{H}, \widetilde{HU})^{N+1}$ where there is no fine grid covering a grid cell, but replacing $(\tilde{H}, \widetilde{HU})^{N+1}$ by the average of $(h, hu)^{N+1}$ over fine grid cells that cover any coarse grid cell.

The final step is applied because the fine grid values $(h, hu)^{N+1}$ are more accurate than the provisional coarse grid values.

We then proceed to the next coarse grid time step. Note that at the start of this step, the updated $(H, HU)^{N+1}$ will be used to solve for Ψ^{N+1}. The provisional $\widetilde{\Psi}^{N+1}$ is discarded. Hence two elliptic solves are required on the coarse level each time step, rather than only one as in the single grid algorithm.

If the refinement factor is larger than 2, then the same approach outlined above works, but there will be additional time steps on level 2. For each time step the ghost cell BCs will be determined by linear interpolation in time between $\mathcal{I}_f(Q^N)$ and $\mathcal{I}_f(\tilde{Q}^{N+1})$.

If there are more than two levels, then this same idea is applied recursively: After each time step on level 2, any level 3 grids will be advanced by the necessary number of time steps to reach the advanced time on level 2. In this case there will also be two elliptic solves for every time step on level 2, once for the provisional values after advancing level 2, and once at the start of the next level 2 time step after (h, hu) on level 2 has been updated by averaging the more accurate level 3 values.

4. COMPUTATIONAL RESULTS

In this section we show an end-to-end simulation of a hypothetical asteroid impact off the coast of Washington, from initial conditions to shoreline inundation. We present our initialization procedure in some detail. The section ends with a discussion of results.

4.1. Initialization procedure

The computational results presented in this section use initial conditions of a static crater, illustrated in Figure 2(a). This is a standard test problem in the literature, from [27]. The crater is 1 km deep, with a diameter of 3 km, in an ocean of depth 4 km. Depth-averaged equations are unsuitable for modeling the generation of the crater and the initial flow, since

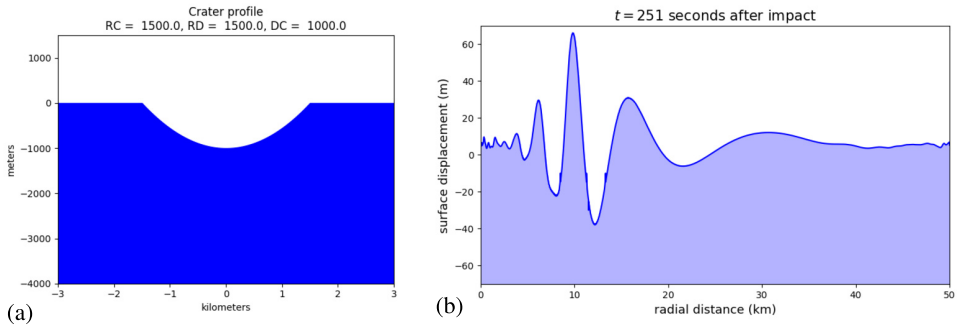


FIGURE 2

(a) Initial conditions of a static crater of depth 1 km and diameter 3 km, which were the initial conditions for the hydrocode simulation. (b) The radially symmetric results of the hydrocode simulation at 251 seconds, used to start the GeoClaw Boussinesq simulation.

the ensuing large vertical velocity components are not modeled. The initial conditions for our depth-averaged simulations are taken from a three-dimensional hydrocode simulation of the first 251 seconds after impact. The hydrocode ALE3D [21] was run by a collaborator [24] in a radially symmetric manner, and the surface displacement at time $t = 251$ seconds was recorded, as shown in Figure 2(b). It proved too noisy to depth-average the horizontal velocity from the hydrocode. Instead, we set the velocity based on the surface displacement and assuming that the wave was a purely outgoing wave satisfying the SWEs, for which the velocity then depends only on the ocean depth and surface displacement. We could then place the "initial" crater anywhere in the ocean.

This procedure for initialization of the velocity can be done because the wave speed for the SWEs is independent of wave number, and the eigenvectors of the linearized Jacobian matrix give the relation between surface elevation and fluid velocity for unidirectional waves. However, in the Boussinesq equations the wave speed depends on wave number and using the initialization based on the SWEs results in a small wave propagating inward as well. Better initialization procedures will be investigated in future research.

4.2. Adaptive simulation

As an illustration, we show a simulation of a hypothetical asteroid impact off the coast of Washington. We place the initial crater approximately 150 km west of Grays Harbor, to study the vulnerable area around Westport, WA, shown in Figure 3. This well-studied area is in close proximity to the Cascadia Subduction Zone, which can generate Mw 9 earthquakes. The Ocosta elementary school in Westport was recently rebuilt to incorporate the first tsunami vertical evacuation structure in the US, due to the low topography of this region, with design work based in part on GeoClaw modeling [10]. Detailed bathymetry and topography data is available in this region at a resolution of 1/3 arcsecond [20], which is roughly 10 m in latitude and 7 m in longitude at this location. For the ocean we use the etopo1 topography DEM [2] at a resolution of 1 arcminute.

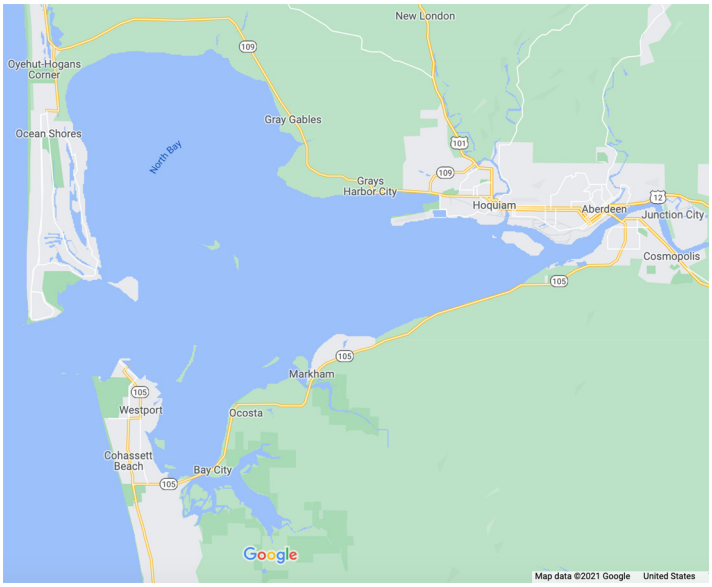


FIGURE 3
Figure shows a Google Maps screenshot of Grays Harbor, on the Washington coast. The community of Westport is on the southern peninsula. This is the focus of the inundation modeling presented in Figure 5.

A 7 level simulation was used, starting with a coarsest level over the ocean with $\Delta y = 10$ arcminutes, and refining by factors 5, 3, 2, 2, 5, 6 at successive levels, with an overall refinement by 1800 for the level-7 grids with $\Delta y = 1/3$ arcsecond. On each grid $\Delta x = 1.5\Delta y$ so that the finest-level computational grids are at a resolution of roughly 10 m in both x and y.

Figures 4 and 5 show snapshots of the simulation at the indicated times. The dispersion is clearly evident, with a much more oscillatory solution than would be obtained with the shallow water equations. The figures show how the fine grid patches move to follow the expanding wave. The refinement is guided to focus at later times only on waves approaching Grays Harbor. There are some reflections at the grid boundaries, but they are much smaller in magnitude than the waves we are tracking. The 6th level refined patch appears approximately 20 minutes after the impact, to track the waves as they approach Grays Harbor. Note how the bathymetry is refined along with the solution when the finer patches appear. The close-up plots near Grays Harbor in Figure 5 shows "soliton fission," a nonlinear dispersive wave phenomenon seen near the coast that can be captured with Boussinesq solvers [3, 19]. In this simulation, we switch from Boussinesq to SWEs at a depth of 10 meters (based on the undisturbed water depth). Figure 5 also shows the waves sweeping over the Westport and Ocean Shores peninsulas. In this calculation, the finest level 7 grid was placed only over Westport, while the level 6 grid covers both peninsulas.

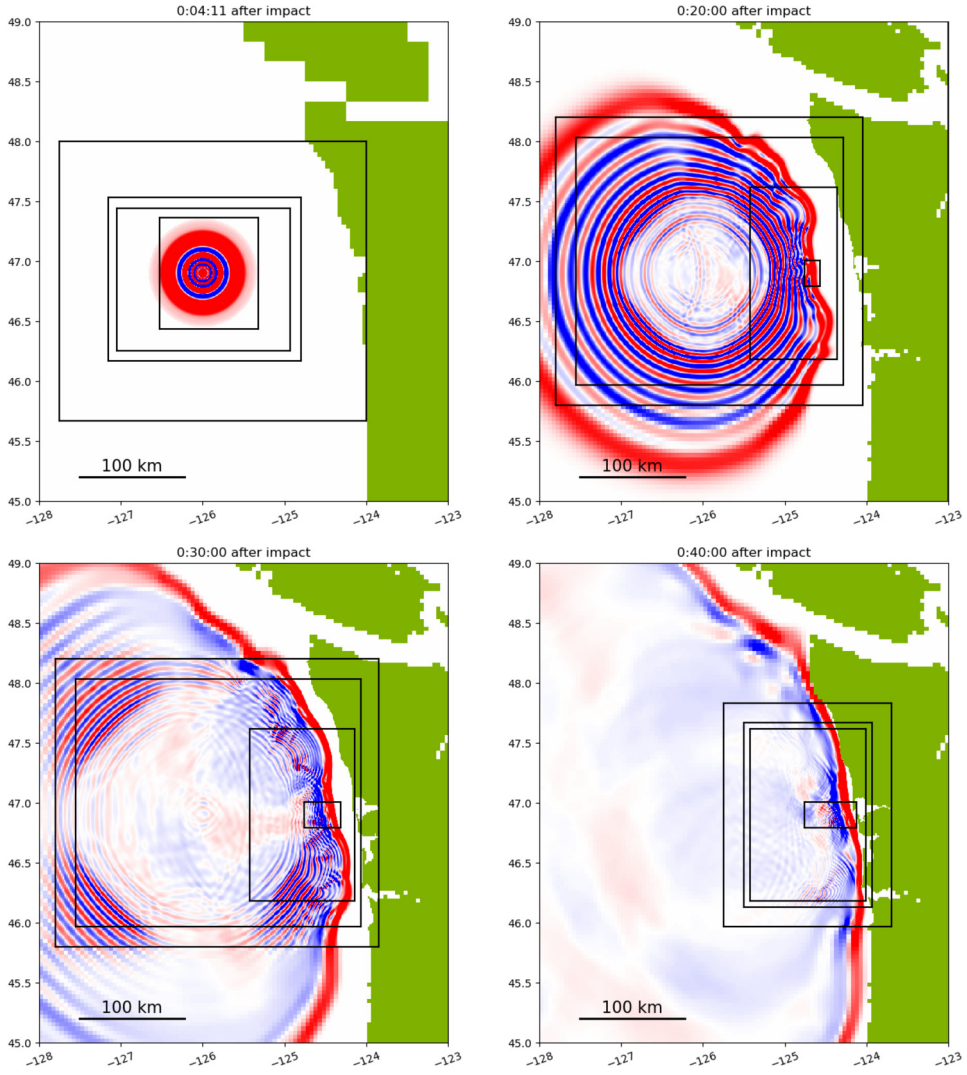


FIGURE 4

Figure shows initial conditions and mesh configuration and three later times during the adaptive solution. The black rectangles show boundaries of refined patches with finer resolution. The colors show elevation above (red) or below (blue) sea level, and saturate at ± 3 m. The waves are larger amplitude but this color range is used to also show the smaller waves in the oscillatory wave train. Note that these are not resolved on coarser levels, and that AMR is guided to focus on the waves approach Grays Harbor, WA.

4.3. Discussion of results

Since the wavelengths of asteroid-generated tsunamis are shorter than those of earthquake-generated tsunamis, dispersive effects may be very important. Dispersion gives rise to a highly oscillatory set of waves that propagate at different speeds, possibly affect-

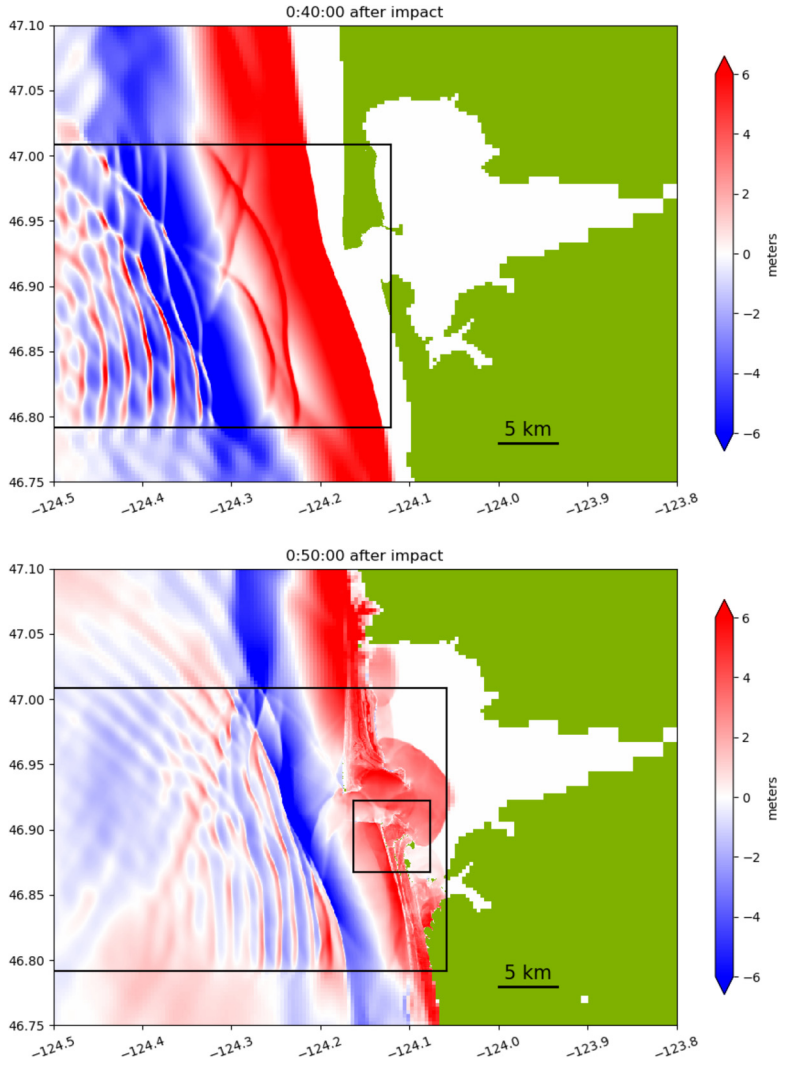


FIGURE 5

The tsunami approaching Grays Harbor and overtopping the Ocean Shores and Westport peninsulas. In this figure, the colors saturate at ± 6 m elevation relative to sea level.

ing the arrival time of the first wave and leading to significant waves over a longer time period. These waves can undergo substantial amplification during the shoaling process on the continental shelf and break up into large solitary waves.

On the other hand, wave breaking can rapidly dissipate the energy in short wavelength waves. Moreover, a train of waves approaching shore results in nonlinear interactions in the swash zone, where waves run up the beach. In the swash zone, a large approaching wave may be largely negated by the rundown of the previous wave. Thus it is not clear a priori whether waves modeled with the Boussinesq equations will result in substantially different

onshore inundation than would be observed is only using the SWE, for which computations are much less expensive. This question still remains to be answered.

5. OPEN PROBLEMS AND FUTURE RESEARCH

We have demonstrated that it is possible to develop a high-fidelity modeling capability that includes propagation and inundation by combining the Boussinesq equations, shallow water equations, and patch-based adaptive mesh refinement in space and time. We are embedding this in the GeoClaw software framework, which has previously been well validated for earthquake-generated tsunamis. This will allow the efficient simulation of dispersive waves generated from asteroid impacts as they propagate across the ocean, combined with high-resolution simulation of the resulting inundation on the coast. This capability could be very important in hazard assessment for an incipient impact.

Unfortunately, the software is not yet robust enough for general use. While it often works well, stability issues sometimes arise at the edges of finer grid patches when refinement ratios greater than 2 are used from one level to the next. Larger refinement ratios are generally desirable, so that an overall refinement factor of several thousand between the coarsest and finest meshes can be achieved with 6 or 7 levels of mesh refinement. Other Boussinesq codes we are aware of use factor of 2 refinement [8, 23], so this may be an inherent instability. Moreover, numerical instabilities have also been observed by other researchers [17] even on a uniform grid when there are sharp changes in bathymetry rather than in the grid resolution. We are investigating this issue theoretically, and may find that a different discretization or even a different fomulation of the Boussinesq equations is required to obtain a sufficiently robust code.

We are also continuing to investigate the shoaling phenomena and the best way to incorporate wave breaking and the transition from Boussinesq to shallow water equations. This can have a significant impact on the resulting onshore run-up and inundation.

Not discussed here but currently under investigation is the possibility of solving the implicit system of equations on multiple levels in a coupled manner, whenever the levels have been advanced to the same point in time. It is an open question whether a coupled system is more accurate and/or stable, and possibly less computationally expensive, in the context of patch-based adaptive mesh refinement that includes refinement in time.

Finally, as mentioned in Section 4.1, additional research is needed on ways to initialize the depth-averaged model. Better initialization procedures will allow a more seamless transition from three-dimentional hydrocode simulations of asteroid impacts in the ocean to our model of tsunami propagation and inundation.

ACKNOWLEDGMENTS

We thank Darrel Robertson, a member of the ATAP team, for providing the hydrocode simulation results for our initial conditions.

FUNDING

This work was partially supported by the NASA Asteroid Threat Assessment Project (ATAP) through the Planetary Defense Coordination Office and BAERI contract AO9667.

REFERENCES

[1] Clawpack Development Team, Clawpack software, 2021 http://www.clawpack.org.

[2] C. Amante and B. W. Eakins, ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis. NOAA technical memorandum NESDIS NGDC-24, 2009, http://www.ngdc.noaa.gov/mgg/global/global.html.

[3] T. Baba, N. Takahashi, Y. Kaneda, K. Ando, D. Matsuoka, and T. Kato, Parallel implementation of dispersive tsunami wave modeling with a nesting algorithm for the 2011 Tohoku tsunami. *Pure Appl. Geophys.* **172** (2015), no. 12, 3455–3472.

[4] M. J. Berger, D. L. George, R. J. LeVeque, and K. T. Mandli, The GeoClaw software for depth-averaged flows with adaptive refinement. *Adv. Water Resour.* **34** (2011), no. 9, 1195–1206.

[5] M. J. Berger and R. J. LeVeque, Modeling issues in asteroid-generated tsunamis. NASA Technical Memorandum NASA/CR-2018-219786, ARC-E-DAA-TN53167, 2018, http://hdl.handle.net/2060/20180006617.

[6] M. Bollhöfer, O. Schenk, R. Janalik, S. Hamm, and K. Gullapalli, State-of-the-art sparse direct solvers. In *Parallel algorithms in computational science and engineering*, edited by A. Grama and A. H. Sameh, pp. 3–33, Springer, Cham, 2020.

[7] M. Brocchini, A reasoned overview on Boussinesq-type models: the interplay between physics, mathematics and numerics. *Philos. Trans. R. Soc. A* **469** (2013), no. 20130496.

[8] D. Calhoun and C. Burstedde, Forestclaw: A parallel algorithm for patch-based adaptive mesh refinement on a forest of quadtrees. 2017, arXiv:1703.03116.

[9] S. Glimsdal, G. K. Pedersen, C. B. Harbitz, and F. Løvholt, Dispersion of tsunamis: does it really matter? *Nat. Hazards Earth Syst. Sci.* **13** (2013), 1507–1526.

[10] F. I. Gonzaléz, R. J. LeVeque, and L. M. Adams, Tsunami hazard assessment of the Ocosta School Site in Westport, WA, 2013, http://hdl.handle.net/1773/24054.

[11] F. González, R. J. LeVeque, J. Varkovitzky, P. Chamberlain, B. Hirai, and D. L. George, GeoClaw results for the NTHMP tsunami benchmark problems, 2011, http://depts.washington.edu/clawpack/links/nthmp-benchmarks.

[12] J. Kim, G. K. Pedersen, F. Løvholt, and R. J. LeVeque, A Boussinesq type extension of the GeoClaw model – a study of wave breaking phenomena applying dispersive long wave models. *Coastal Eng.* **122** (2017), 75–86.

[13] D. G. Korycansky and P. J. Lynett, Offshore breaking of impact tsunami: the Van Dorn effect revisited. *Geophys. Res. Lett.* **32** (2005), no 10.

[14] D. G. Korycansky and P. J. Lynett, Run-up from impact tsunami. *Geophys. J. Int.* **170** (2007), 1076–1088.

[15] R. J. LeVeque, *Finite volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[16] R. J. LeVeque, D. L. George, and M. J. Berger, Tsunami modelling with adaptively refined finite volume methods. *Acta Numer.* **20** (2011), 211–289.

[17] F. Løvholt and G. Pedersen, Instabilities of Boussinesq models in non-uniform depth. *Internat. J. Numer. Methods Fluids* **61** (2009), 606–637.

[18] P. A. Madsen and O. R. Sørensen, A new form of the Boussinesq equations with improved linear dispersion characteristics. Part 2. A slowly-varying bathymetry. *Coastal Eng.* **18** (1992), no. 3–4, 183–204.

[19] M. Matsuyama, M. Ikeno, T. Sakakiyama, and T. Takeda, A study of tsunami wave fission in an undistorted experiment. *Pure Appl. Geophys.* **164** (2007), no. 2, 617–631.

[20] NOAA National Geophysical Data Center, Geophysical Data Center, Astoria, Oregon 1/3 arc-second MHW coastal digital elevation model, 2019, https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.dem:5490.

[21] A. L. Nichols and D. M. D (eds.), ALE3D user's manual: an arbitrary Lagrange/Eulerian 2D and 3D code system. Lawrence Livermore National Laboratory LLNL-SM-726137, 2017.

[22] D. H. Peregrine, Long waves on a beach. *J. Fluid Mech.* **27** (1967), no. 4, 815–827.

[23] S. Popinet, A quadtree-adaptive multigrid solver for the Serre–Green–Naghdi equations. *J. Comp. Physiol.* **302** (2015), 336–358.

[24] D. Robertson, Personal communication, 2019.

[25] D. K. Robertson and G. R. Gisler, Near and far-field hazards of asteroid impacts in oceans. *Acta Astronaut.* **156** (2019), 262–277.

[26] H. A. Schäffer and P. A. Madsen, Further enhancements of Boussinesq-type equations. *Coastal Eng.* **26** (1995), no. 1–2, 1–14.

[27] S. N. Ward and E. Asphaug, Asteroid impact tsunami: a probabilistic hazard assessment. *Icarus* **145** (2000), 64–78.

MARSHA J. BERGER

Courant Institute, New York University, 251 Mercer St., New York, NY 10012, USA, and Flatiron Institute, 162 5th Ave., New York, NY 10010, USA, berger@cims.nyu.edu

RANDALL J. LEVEQUE

Department of Applied Math, University of Washington, Seattle, WA 98195, USA, and HyperNumerics LLC, Seattle, WA, USA, rjl@uw.edu

STRUCTURE-PRESERVING MODEL ORDER REDUCTION OF HAMILTONIAN SYSTEMS

JAN S. HESTHAVEN, CECILIA PAGLIANTINI, AND NICOLÒ RIPAMONTI

ABSTRACT

We discuss the recent developments of projection-based model order reduction (MOR) techniques targeting Hamiltonian problems. Hamilton's principle completely characterizes many high-dimensional models in mathematical physics, resulting in rich geometric structures, with examples in fluid dynamics, quantum mechanics, optical systems, and epidemiological models. MOR reduces the computational burden associated with the approximation of complex systems by introducing low-dimensional surrogate models, enabling efficient multiquery numerical simulations. However, standard reduction approaches do not guarantee the conservation of the delicate dynamics of Hamiltonian problems, resulting in reduced models plagued by instability or accuracy loss over time. By approaching the reduction process from the geometric perspective of symplectic manifolds, the resulting reduced models inherit stability and conservation properties of the high-dimensional formulations. We first introduce the general principles of symplectic geometry, including symplectic vector spaces, Darboux' theorem, and Hamiltonian vector fields. These notions are then used as a starting point to develop different structure-preserving reduced basis (RB) algorithms, including SVD-based approaches, and greedy techniques. We conclude the review by addressing the reduction of problems that are not linearly reducible or in a noncanonical Hamiltonian form.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 65M99; Secondary 70H33, 65P10

KEYWORDS

Reduced order models, structure preservation, Hamiltonian problems

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5072–5097
DOI 10.4171/ICM2022/100

1. INTRODUCTION

The discretization of partial differential equations (PDEs) by classical methods like finite element, spectral method, or finite volume leads to dynamical models with very large state-space dimensions, typically of the order of millions of degrees of freedom to obtain an accurate solution. MOR [48] is an effective method for reducing the complexity of such models while capturing the essential features of the system state. Starting from the Truncated Balanced Realization, introduced by Moore [36] in 1981, several other reduction techniques have been developed and flourished during the last 40 years, including the Hankel-norm reduction [20], the proper orthogonal decomposition (POD) [50], and the Padé-via-Lanczos (PVL) algorithm [14]. More recently, there has been a focus on the physical interpretability of the reduced models. Failure to preserve structures, invariants, and intrinsic properties of the approximate model, besides raising questions about the validity of the reduced models, has been associated with instabilities and exponential error growth, independently of the theoretical accuracy of the reduced solution space. Stable reduced models have been recovered by enforcing constraints on the reduced dynamics obtained using standard reduction tools. Equality and inequality constraints have been considered to control the amplitude of the POD modes [16], the fluid temperature in a combustor [31], and the aerodynamic coefficients [54]. Other methods directly incorporate the quantity of interest into the reduced system, producing *inf–sup* stable [5], flux-preserving [11], and skew-symmetric [3] conservative reduced dynamics. Even though great effort has been spent developing time integrators that preserve the symplectic flow underlying Hamiltonian systems, interest in geometric model order reduction initiated more recently, with efforts to preserve the Lagrangian structures [33].

The remainder of the paper is organized as follows. In Section 2, we present the structure characterizing the dynamics of Hamiltonian systems and the concept of symplectic transformations. In Section 3, we show that linear symplectic maps can be used to guarantee that the reduced models inherit the geometric formulation from the full dynamics. Different strategies to generate such maps are investigated in Section 4, with thoughts on optimality results and computational complexities. A novel approach deviating from the linearity of the projection map is briefly discussed in Section 5. Finally, we discuss applications of structure-preserving reduction techniques to two more general classes of problems in Section 6, and some concluding remarks are offered in Section 7.

2. SYMPLECTIC GEOMETRY AND HAMILTONIAN SYSTEMS

Let us first establish some definitions and properties concerning symplectic vector spaces.

Definition 2.1. Let $\mathcal{M}$ be a finite-dimensional real vector space and $\Omega : \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$ a bilinear map. Ω is called antisymmetric if

$$\Omega(u, v) = -\Omega(v, u), \quad \forall u, v \in \mathcal{M}.$$

It is nondegenerate if

$$\Omega(u, v) = 0, \quad \forall u \in \mathcal{M} \Longrightarrow v = 0.$$

Definition 2.2. Let $\mathcal{M}$ be a finite-dimensional vector space with Ω an antisymmetric bilinear form on $\mathcal{M}$. The pair $(\mathcal{M}, \Omega)$ is a symplectic linear vector space if Ω is nondegenerate. Moreover, $\mathcal{M}$ has to be $2n$-dimensional.

Since we are interested in structure-preserving transformations, preserving the structure means to preserve the antisymmetric bilinear form, as stated in the following definition.

Definition 2.3. Let $(\mathcal{M}_1, \Omega_1)$ and $(\mathcal{M}_2, \Omega_2)$ be two symplectic vector spaces with $\dim(\mathcal{M}_1) \geq \dim(\mathcal{M}_2)$. The differentiable map $\phi : \mathcal{M}_1 \mapsto \mathcal{M}_2$ is called a symplectic transformation (symplectomorphism) if

$$\phi^* \Omega_2 = \Omega_1,$$

where $\phi^* \Omega_2$ is the pull-back of Ω_2 with ϕ.

One of the essential properties of Euclidean spaces is that all the Euclidean spaces of equal dimensions are isomorphic. For the symplectic vector spaces, a similar result holds, since two $2n$-dimensional symplectic vector spaces are symplectomorphic to one another. They therefore are fully characterized by their dimensions (as a consequence of the following theorem).

Theorem 2.1 (Linear Darboux' theorem [13]). *For any symplectic vector space* $(\mathcal{M}, \Omega)$*, there exists a basis* $\{e_i, f_i\}_{i=1}^n$ *of* $\mathcal{M}$ *such that*

$$\Omega(e_i, e_j) = 0 = \Omega(f_i, f_j), \quad \Omega(e_i, f_j) = \delta_{ij}, \quad \forall i, j = 1, \ldots, n. \tag{2.1}$$

The basis is called Darboux' chart or canonical basis.

The proof of Theorem 2.1 is based on a procedure similar to the Gram–Schmidt process to generate the symplectic basis, known as symplectic Gram–Schmidt [4].

The canonical basis allows representing the symplectic form as

$$\Omega(u, v) = \zeta^\top \mathbb{J}_{2n} \eta, \tag{2.2}$$

where $\zeta, \eta \in \mathbb{R}^{2n}$ are the expansion coefficients of $u, v \in \mathcal{M}$ with respect to the basis $\{e_i, f_i\}_{i=1}^n$ and

$$\mathbb{J}_{2n} = \begin{bmatrix} 0_n & \mathbb{I}_n \\ -\mathbb{I}_n & 0_n \end{bmatrix} \in \mathbb{R}^{2n \times 2n}, \tag{2.3}$$

is known as the Poisson tensor, with $0_n \in \mathbb{R}^{n \times n}$ and $\mathbb{I}_n \in \mathbb{R}^{n \times n}$ denoting the zero and identity matrices, respectively. As a direct result, the matrix representation of the symplectic form Ω in the canonical basis is $\mathbb{J}_{2n}$. More generally, using a noncanonical basis, the form reduces to $\Omega(u, v) = \zeta^\top J_{2n} \eta$, with J_{2n} being an invertible constant skew-symmetric matrix.

While symplectic vector spaces are helpful for the analysis of dynamical problems in Euclidean spaces and to define geometric reduced-order models, the constraint to the

Euclidean setting is not generally adequate. In particular, the abstraction of the phase spaces of classical mechanics over arbitrary manifolds requires the definition of more general symplectic manifolds. We refer the reader to [34] for a more comprehensive description of the topic. In this work, we limit ourselves to introducing a significant result regarding the evolution of the state of Hamiltonian systems.

Definition 2.4. Let $(\mathcal{M}, \Omega)$ be a symplectic manifold and $H : \mathcal{M} \mapsto \mathbb{R}$ a 1-form. We refer to the unique vector field $\mathcal{X}_H$, which satisfies

$$i(\mathcal{X}_H)\Omega = \mathbf{d}H,$$

as the *Hamiltonian vector field* related to H, where $i(\mathcal{X}_H)$ denotes the contraction operator and $\mathbf{d}$ is the exterior derivative. The function H is called the *Hamiltonian* of the vector field $\mathcal{X}_H$.

Suppose $\mathcal{M}$ is also compact, then $\mathcal{X}_H$ is complete [22] and can be integrated, i.e., there exists an integral curve of $\mathcal{X}_H$, parametrized by the real variable t, that is, the solution of

$$\dot{y}(t) = \mathcal{X}_H\big(y(t)\big). \tag{2.4}$$

Equation (2.4) is referred to as Hamilton's equation of evolution or Hamiltonian system. Darboux' theorem, as a generalization of Theorem 2.1, states that two symplectic manifolds are only locally symplectomorphic. Using this result, the Hamiltonian vector field $\mathcal{X}_H$ admits the local representation

$$\mathcal{X}_H = \sum_{i=1}^{n} \frac{\partial H}{\partial f_i}\frac{\partial}{\partial e_i} - \frac{\partial H}{\partial e_i}\frac{\partial}{\partial f_i}, \tag{2.5}$$

with $\{e_i, f_i\}_{i=1}^n$ is a local basis, leading to the following representation of (2.4), expressed directly in terms of H.

Proposition 2.1. *Let $(\mathcal{M}, \Omega)$ be a $2n$-dimensional symplectic vector space and let $\{q_i, p_i\}_{i=1}^n$ be a canonical system of coordinates. Hamilton's equation is defined by*

$$\begin{cases} \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \\ \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \end{cases} \tag{2.6}$$

for $i = 1, \ldots, n$, which is a first order system in the (q_i, p_i)-space, or generalized phase-space.

Thus, if the state vector $y = (q_1, \ldots, q_n, p_1, \ldots, p_n)$ is introduced, (2.6) takes the form

$$\dot{y}(t) = \mathbb{J}_{2n}\nabla_y H\big(y(t)\big), \tag{2.7}$$

where $\nabla_y H$ is the naive gradient of H. The flow of Hamilton's equation has some interesting properties.

Proposition 2.2. *Let ϕ_t be the flow of a Hamiltonian vector field $\mathcal{X}_H$. Then $\phi_t : \mathcal{M} \mapsto \mathcal{M}$ is a symplectic transformation.*

We rely on a geometric perspective of linear vector spaces to highlight the importance of Proposition 2.2. Given two coefficient vectors u and v in $\mathbb{R}^{2n}$, the symplectic form (2.2) can be interpreted as the sum of the oriented areas of the orthogonal projection of the parallelogram defined by the two vectors on the (q_i, p_i) planes. Definition 2.3, in case of $2n$-dimensional symplectic vector space $(\mathcal{M}, \Omega)$ with canonical coordinates, is equivalent to stating that a map $\phi : \mathbb{R}^{2n} \mapsto \mathbb{R}^{2n}$ is a symplectic transformation if and only if its Jacobian ϕ' satisfies everywhere

$$(\phi')^\top J_{2n} \phi' = J_{2n}. \tag{2.8}$$

Property (2.8) can be used to show that a symplectic transformation preserves the bilinear form Ω in the sense that [34]

$$\Omega\big(\phi(u), \phi(v)\big) = \Omega(u, v). \tag{2.9}$$

Hence, the symplectic map ϕ represents a volume-preserving transformation. However, being symplectic is a more restrictive condition than being volume-preserving, as shown in the Nonsqueezing Theorem [24].

We conclude this section by noting that if the Hamiltonian function does not depend explicitly on time, its value is conserved along the solution trajectory.

Proposition 2.3. *For Hamiltonian systems (2.7), the Hamiltonian function is a first integral.*

3. SYMPLECTIC GALERKIN PROJECTION

The motivation of MOR is to reduce the computational complexity of dynamical systems in numerical simulations. In the context of structure-preserving projection-based reduction, two key ingredients are required to define a reduced model. First, we need a low-dimensional symplectic vector space that accurately represents the solution manifold of the original problem. Then, we have to define a projection operator to map the symplectic flow of the Hamiltonian system onto the reduced space, while preserving its delicate properties.

Let us assume there exists a canonical basis $\{e_i, f_i\}_{i=1}^n$ such that Hamilton's equation can be written in canonical form

$$\begin{cases} \dot{y}(t) = \mathbb{J}_{2n} \nabla_y H(y(t)), \\ y(0) = y_0, \end{cases} \tag{3.1}$$

and the related symplectic vector space is denoted by $(\mathcal{M}, \Omega)$. Symplectic projection-based model order reduction adheres to the key idea of more general projection-based techniques [30] to approximate y in a low-dimensional symplectic subspace $(\mathcal{A}, \Omega)$ of dimension $2k$. In particular, we aim at $k \ll n$ to have a clear reduction, and therefore, significant gains in terms of computational efficiency. Let $\{\tilde{e}_i, \tilde{f}_i\}_{i=1}^k$ be a reduced basis for the approximate symplectic subspace and construct the linear map $\phi : \mathcal{A} \mapsto \mathcal{M}$ given by

$$y \approx \phi(z) = Az, \tag{3.2}$$

where

$$A = [\tilde{e}_1, \dots, \tilde{e}_k, \tilde{f}_1, \dots, \tilde{f}_k] \in \mathbb{R}^{2n \times 2k}.$$

Then A belongs to the set of symplectic matrices of dimension $2n \times 2k$, also known as the symplectic Stiefel manifold, defined by

$$\mathrm{Sp}(2k, \mathbb{R}^{2n}) := \left\{ L \in \mathbb{R}^{2n \times 2k} : L^\top J_{2n} L = J_{2k} \right\}.$$

Differential maps are often used to transfer structures from well-defined spaces to unknown manifolds. In this context, using the symplecticity of A, it is possible to show [1] that Definition 2.3 holds, with the right inverse of ϕ represented by A, and that there exists a symplectic form on $\mathcal{A}$ given by

$$\tilde{\Omega} = \phi^* \Omega = A^\top \mathbb{J}_{2n} A = \mathbb{J}_{2k}. \tag{3.3}$$

As a result, $(\mathcal{A}, \tilde{\Omega})$ is a symplectic vector space. In the following, for the sake of notation, we use $\mathcal{A}$ to indicate the reduced symplectic manifold paired with its bilinear form.

Given a symplectic matrix $A \in \mathbb{R}^{2n \times 2k}$, its symplectic inverse is defined as

$$A^+ = \mathbb{J}_{2k}^\top A^\top \mathbb{J}_{2n}. \tag{3.4}$$

Even though different from the pseudoinverse matrix $(A^\top A)^{-1} A^\top$, the symplectic inverse A^+ plays a similar role and, in the following proposition, we outline its main properties [44].

Proposition 3.1. *Suppose $A \in \mathbb{R}^{2n \times 2k}$ is a symplectic matrix and A^+ is its symplectic inverse. Then*

- $A^+ A = \mathbb{I}_{2n}$,
- $(((A^+)^\top)^+)^\top = A$,
- $(A^+)^\top \in \mathrm{Sp}(2k, \mathbb{R}^{2n})$,
- *If A is orthogonal then $A^+ = A^\top$.*

Using (3.3), the definition of A^+ and the symplectic Gram–Schmidt process, it is possible to construct a projection operator $\mathcal{P}_{\mathcal{A}} = A \mathbb{J}_{2k}^\top A^\top \mathbb{J}_{2n} = A A^+$, that, differently from the POD orthogonal projection [46], can be used to approximate (3.1) with Hamiltonian system of reduced-dimension $2k$, characterized by the Hamiltonian function

$$H_{\mathrm{RB}}(z) = H(Az). \tag{3.5}$$

In particular, in the framework of Galerkin projection, using (3.2) in (3.1) yields

$$A\dot{z} = \mathbb{J}_{2n} \nabla_y H(Az) + r, \tag{3.6}$$

with r being the residual term. Utilizing the chain rule and the second property of A^+ in Proposition 3.1, the gradient of the Hamiltonian in (3.6) can be recast as

$$\nabla_y H(Az) = (A^+)^\top \nabla_z H_{\mathrm{RB}}(z).$$

By assuming that the projection residual is orthogonal with respect to the symplectic bilinear form to the space spanned by A, we recover

$$\begin{cases} \dot{z}(t) = \mathbb{J}_{2k} \nabla_z H_{\mathrm{RB}}(z(t)), \\ z(0) = A^+ y_0. \end{cases} \tag{3.7}$$

System (3.7) is known as a symplectic Galerkin projection of (3.1) onto $\mathcal{A}$. The preprocessing stage consisting of the collection of all the computations required to assemble the basis A is known as the *offline* stage. The numerical solution of the low-dimensional problem (3.7) represents the *online* stage, and provides a fast approximation to the solution of the high-fidelity model (3.1) by means of (3.2). Even though the offline stage is possibly computationally expensive, this splitting is beneficial in a multiquery context, when multiple instances of (3.7) have to be solved, e.g., for parametric PDEs.

Traditional projection-based reduction techniques do not guarantee stability, even if the high-dimensional problem admits a stable solution [37], often resulting in a blowup of system energy. On the contrary, by preserving the geometric structure of the problem, several stability results hold for the reduced Hamiltonian equation (3.7). In [1, PROPOSITION 15, PAGE A2625], the authors show that the error in the Hamiltonian $|H(y(t)) - H_{\mathrm{RB}}(z(t))|$ is constant for all t. We detail two relevant results in the following, suggesting that structure and energy preservation are key for stability.

Theorem 3.1 (Boundedness result [44]). *Consider the Hamiltonian system* (3.1)*, with Hamiltonian* $H \in C^\infty(\mathcal{M})$ *and initial condition* $y_0 \in \mathbb{R}^{2n}$ *such that* $y_0 \in \mathrm{range}(A)$*, with* $A \in \mathbb{R}^{2n\times 2k}$ *symplectic basis. Let* (3.7) *be the reduced Hamiltonian system obtained as the symplectic Galerkin projection induced by* A *of* (3.1)*. If there exists a bounded neighborhood* $\mathcal{U}_{y_0}$ *in* $\mathbb{R}^{2n}$ *such that* $H(y_0) < H(\tilde{y})$*, or* $H(y_0) > H(\tilde{y})$*, for all* $\tilde{y}$ *on the boundary of* $\mathcal{U}_{y_0}$*, then both the original system and the reduced system constructed by the symplectic projection are uniformly bounded for all* t*.*

Theorem 3.2 (Lyapunov stability [1, 44]). *Consider the Hamiltonian system* (3.1) *with Hamiltonian* $H \in C^2(\mathcal{M})$ *and the reduced Hamiltonian system* (3.7)*. Suppose that* y^* *is a strict local minimum of* H*. Let* S *be an open ball around* y^* *such that* $\nabla^2 H(y) > 0$ *and* $H(z) < c$*, for all* $z \in S$ *and some* $c \in \mathbb{R}$*, and* $H(\bar{y}) = c$ *for some* $\bar{y} \in \partial S$*, where* ∂S *is the boundary of* S*. If there exists an open neighborhood* S *of* y^* *such that* $S \cap \mathrm{range}(A) \neq \emptyset$*, then the reduced system* (3.7) *has a stable equilibrium point in* $S \cap \mathrm{range}(A)$*.*

For the time-discretization of (3.7), the use of a symplectic integrator [26] is crucial for preserving the symplectic structure at the discrete level. In particular, the discrete flow obtained using a symplectic integrator satisfies a discrete version of Proposition 2.2.

In the next section, we introduce different strategies to construct symplectic bases as results of optimization problems.

4. PROPER SYMPLECTIC DECOMPOSITION

Let us consider the solution vectors $y_i = y(t_i) \in \mathbb{R}^{2n}$ (the so-called solution snapshots) obtained, for different time instances $t_i \in [t_0, t_{\mathrm{end}}]$, $\forall i = 1, \dots, N$, by time discretization of (3.1) using a symplectic integrator. Define the snapshot matrix

$$M_y := [y_1 \dots y_N], \tag{4.1}$$

as the matrix collecting the solution snapshots as columns. In the following, we consider different algorithms stemming from the historical *method of snapshots* [50], as the base of the proper orthogonal decomposition (POD). To preserve the geometric structure of the original model, we focus on a similar optimization problem, the proper symplectic decomposition (PSD), which represents a data-driven basis generation procedure to extract a symplectic basis from M_y. It is based on the minimization of the projection error of M_y on $\mathcal{A}$ and it results in the following optimization problem for the definition of the symplectic basis $A \in \mathbb{R}^{2n\times 2k}$:

$$\begin{aligned} &\underset{A\in\mathbb{R}^{2n\times 2k}}{\text{minimize}} && \left\| M_y - AA^{+}M_y \right\|_F, \\ &\text{subject to} && A \in \mathrm{Sp}(2k, \mathbb{R}^{2n}), \end{aligned} \tag{4.2}$$

with $\mathcal{A} = \mathrm{range}(A)$ and $\|\cdot\|_F$ being the Frobenius norm. Problem (4.2) is similar to the POD minimization, but with the feasibility set of rectangular orthogonal matrices, also known as the Stiefel manifold

$$\mathrm{St}(2k, \mathbb{R}^{2n}) := \left\{ L \in \mathbb{R}^{2n\times 2k} : L^{\top}L = \mathbb{I}_{2n} \right\},$$

replaced by the symplectic Stiefel manifold. Recently there has been a great interest in optimization on symplectic manifolds, and a vast literature is available on the minimization of the least-squares distance from optimal symplectic Stiefel manifolds. This problem has relevant implications in different physical applications, such as the study of optical systems [18] and the optimal control of quantum symplectic gates [52]. Unfortunately, with respect to POD minimization, problem (4.2) is significantly more challenging for different reasons. The nonconvexity of the feasibility set and the unboundedness of the solution norm precludes standard optimization techniques. Moreover, most of the attention is focused on the case $n = k$, which is not compatible with the reduction goal of MOR.

Despite the interest in the topic, an efficient optimal solution algorithm has yet to be found for the PSD. Suboptimal solutions have been attained by focusing on the subset of the ortho-symplectic matrices, i.e.,

$$\mathbb{S}(2k, 2n) := \mathrm{St}(2k, \mathbb{R}^{2n}) \cap \mathrm{Sp}(2k, \mathbb{R}^{2n}). \tag{4.3}$$

In [44], while enforcing the additional orthogonality constraint in (4.2), the optimization problem is further simplified by assuming a specific structure for A. An efficient greedy method, not requiring any additional block structures to A, but only its orthogonality and simplecticity, has been introduced in [1]. More recently, in [7], the orthogonality requirement has been removed, and different solution methods to the PSD problem are explored. In the following, we briefly review the above-mentioned approaches.

4.1. SVD-based methods for orthonormal symplectic basis generation

In [44], several algorithms have been proposed to directly construct ortho-symplectic bases. Exploiting the SVD decomposition of rearranged snapshots matrices, the idea is to

search for optimal matrices in subsets of $\mathrm{Sp}(2k, \mathbb{R}^{2n})$. Consider the more restrictive feasibility set

$$\mathbb{S}_1(2k, 2n) := \mathrm{Sp}(2k, \mathbb{R}^{2n}) \cap \left\{ \begin{bmatrix} \Phi & 0 \\ 0 & \Phi \end{bmatrix} \mid \Phi \in \mathbb{R}^{n \times k} \right\}.$$

Then $A^\top \mathbb{J}_{2n} A = \mathbb{J}_{2k}$ holds if and only if $\Phi^\top \Phi = \mathbb{I}_n$, i.e., $\Phi \in \mathrm{St}(k, \mathbb{R}^n)$. Moreover, we have that $A^+ = \mathrm{diag}(\Phi^\top, \Phi^\top)$. The cost function in (4.2) becomes

$$\left\| M_y - A A^+ M_y \right\|_F = \left\| M_1 - \Phi \Phi^\top M_1 \right\|_F, \tag{4.4}$$

with $M_1 = [p_1 \ \ldots \ p_N \ q_1 \ \ldots \ q_N] \in \mathbb{R}^{n \times 2N}$, where p_i and q_i are the generalized phase-space components of y_i. Thus, as a result of the Eckart–Young–Mirsky theorem, (4.4) admits a solution in terms of the singular-value decomposition of the data matrix M_1. This algorithm, formally known as Cotangent Lift, owes its name to the interpretation of the solution A to (4.4) in $\mathbb{S}_1(2k, 2n)$ as the cotangent lift of linear mappings, represented by Φ and $\Phi^\top$, between vector spaces of dimensions n and k. Moreover, this approach constitutes the natural outlet in the field of Hamiltonian systems of the preliminary work of Lall et al. [33] on structure-preserving reduction of Lagrangian systems. However, there is no guarantee that the Cotangent Lift basis is close to the optimal of the original PSD functional.

A different strategy, known as Complex SVD decomposition, relies on the definition of the complex snapshot matrix $M_2 = [p_1 + i q_1 \ \ldots \ p_N + i q_N] \in \mathbb{C}^{n \times N}$, with i being the imaginary unit. Let $U = \Phi + i\Psi \in \mathbb{C}^{n \times N}$, with $\Phi, \Psi \in \mathbb{R}^{n \times k}$, be the unitary matrix solution to the following accessory problem:

$$\begin{aligned} \underset{U \in \mathbb{R}^{n \times 2k}}{\text{minimize}} \quad & \left\| M_2 - U U^* M_2 \right\|_F, \\ \text{subject to} \quad & U \in \mathrm{St}(2k, \mathbb{R}^{2n}). \end{aligned} \tag{4.5}$$

As for the Cotangent Lift algorithm, the solution to (4.5) is known to be the set of the $2k$ left-singular vectors of M_2 corresponding to its largest singular values. In terms of the real and imaginary parts of U, the orthogonality constraint implies

$$\Phi^\top \Phi + \Psi^\top \Psi = \mathbb{I}_n, \quad \Phi^\top \Psi = \Psi^\top \Phi. \tag{4.6}$$

Consider the ortho-symplectic matrix, introduced in [44], and given by

$$A = \begin{bmatrix} E & \mathbb{J}_{2n}^\top E \end{bmatrix} \in \mathbb{R}^{2n \times 2k}, \quad E^\top E = \mathbb{I}_k, \quad E^\top \mathbb{J}_{2n} E = 0_k, \quad \text{with } E = \begin{bmatrix} \Phi \\ \Psi \end{bmatrix}. \tag{4.7}$$

Using (4.6), it can be shown that such an A is the optimal solution of the PSD problem in

$$\mathbb{S}_2(2k, 2n) := \mathrm{Sp}(2k, \mathbb{R}^{2n}) \cap \left\{ \begin{bmatrix} \Phi & -\Psi \\ \Psi & \Phi \end{bmatrix} \mid \Phi, \Psi \in \mathbb{R}^{n \times k} \right\},$$

that minimizes the projection error of $M_r := [M_y \ \ \mathbb{J}_{2n} M_y]$, also known as the rotated snapshot matrix, with M_y given in (4.1). In [7], extending the result obtained in [41] for square matrices, it has been shown that (4.7) is a complete characterization of symplectic matrices with orthogonal columns, meaning that all the ortho-symplectic matrices admit a representation of the form (4.7), for a given E, and hence $\mathbb{S}_2(2k, 2n) \equiv \mathbb{S}(2k, 2n)$. In the same work,

Haasdonk et al. showed that an ortho-symplectic matrix that minimizes the projection error of M_r is also a minimizer of the projection error of the original snapshot matrix M_y, and vice versa. This is been achieved by using an equivalence argument based on the POD applied to the matrix M_r. Thus, combining these two results, the Complex SVD algorithm provides a minimizer of the PSD problem for ortho-symplectic matrices.

4.2. SVD-based methods for nonorthonormal symplectic basis generation

In the previous section, we showed that the basis provided by the Complex SVD method is not only near-optimal in $\mathbb{S}_2$, but is optimal for the cost functionals in the space of ortho-symplectic matrices. The orthogonality of the resulting basis is beneficial [32], among others, for reducing the condition number associated with the fully discrete formulation of (3.7). A suboptimal solution to the PSD problem not requiring the orthogonality of the feasibility set is proposed in [44], as an improvement of the SVD-based generators of ortho-symplectic bases using the Gappy POD [19], under the name of nonlinear programming approach (NLP). Let $A^* \in \mathbb{S}_2(2r, 2n)$ be a basis of dimension $2r$ generated using the Complex SVD method. The idea of the NLP is to construct a target basis $A \in \mathrm{Sp}(2k, \mathbb{R}^{2n})$, with $k < r \ll n$, via the linear mapping

$$A = A^* C, \tag{4.8}$$

with $C \in \mathbb{R}^{2r \times 2k}$. Using (4.8) in (4.2) results in a PSD optimization problem for the coefficient matrix C, of significantly smaller dimension ($4kr$ parameters) as compared to the original PSD problem ($4kn$ parameters) with A unknown. However, no optimality results are available for the NLP method.

A different direction has been pursued in [7], based on the connection between traditional SVD and Schur forms and the matrix decompositions, related to symplectic matrices, as proposed in the following theorem.

Theorem 4.1 (SVD-like decomposition [53, THEOREM 1, P. 6]). *If $B \in \mathbb{R}^{2n \times n_s}$, then there exist $S \in \mathrm{Sp}(2n, \mathbb{R}^{2n})$, $Q \in \mathrm{St}(n_s, \mathbb{R}^{n_s})$, and $D \in \mathbb{R}^{2n \times n_s}$ of the form*

$$D = \begin{array}{c} \begin{matrix} b & q & b & n-2b-q \end{matrix} \\ \begin{bmatrix} \Sigma & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \Sigma & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} b \\ q \\ m-b-q \\ b \\ q \\ m-b-q \end{matrix}, \tag{4.9}$$

with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_b)$, $\sigma_i > 0 \ \forall i = 1, \ldots, b$, such that

$$B = SDQ. \tag{4.10}$$

Moreover, $\mathrm{rank}(B) = 2b + q$ *and σ_i are known as symplectical singular values.*

Let us apply the SVD-like decomposition to the snapshot matrix M_y (4.1), where n_s represents the number of snapshots, and define its weighted symplectic singular values as

$$w_i = \begin{cases} \sigma_i \sqrt{\|S_i\|_2^2 + \|S_{n+i}\|_2^2}, & 1 \le i \le b, \\ \|S_i\|_2, & b+1 \le i \le b+q, \end{cases}$$

with $S_i \in \mathbb{R}^{2n}$ being the ith column of S and $\|\cdot\|_2$ the Euclidean norm. The physical interpretation of the classical POD approach characterizes the POD reduced basis as the set of a given cardinality that captures most of the energy of the system. The energy retained in the reduced approximation is quantified as the sum of the squared singular values corresponding to the left singular vectors of the snapshot matrix representing the columns of the basis. A similar guiding principle is used in [7], where the energy of the system, i.e., the Frobenius norm of the snapshot matrix, is connected to the weighted symplectic singular values as

$$\|M_y\|_F^2 = \sum_{i=1}^{b+q} w_i^2. \tag{4.11}$$

Let $\mathcal{I}_{\mathrm{PSD}}$ be the set of indices corresponding to the k largest energy contributors in (4.11),

$$\mathcal{I}_{\mathrm{PSD}} = \{i_j\}_{j=1}^k = \underset{\mathcal{I}\subset\{1,\dots,b+q\}}{\operatorname{argmax}} \left(\sum_{i\in\mathcal{I}} w_i^2 \right). \tag{4.12}$$

Then, the PSD SVD-like decomposition defines a symplectic reduced basis $A \in \mathrm{Sp}(2k, \mathbb{R}^{2n})$ by selecting the pairs of columns from the symplectic matrix S corresponding to the indices set $\mathcal{I}_{PSD}$

$$A = [s_{i_1} \quad \dots \quad s_{i_k} \quad s_{n+i_1} \quad \dots \quad s_{n+i_k}]. \tag{4.13}$$

Similarly to the POD, the reconstruction error of the snapshot matrix depends on the magnitude of the discarded weighted symplectic singular values as

$$\left\| M_y - AA^+ M_y \right\|_F^2 = \sum_{i\in\{1,\dots,b+q\}\setminus\mathcal{I}_{\mathrm{PSD}}} w_i^2. \tag{4.14}$$

Even though there are no proofs that the PSD SVD-like algorithm reaches the global optimum in the sense of (4.2), some analysis and numerical investigations suggest that it provides superior results as compared to orthonormal techniques [7].

4.3. Greedy approach to symplectic basis generation

The reduced basis methodology is motivated and applied within the context of real-time and multiqueries simulations of parametrized PDEs. In the framework of Hamiltonian systems, we consider the following parametric form of (3.1):

$$\begin{cases} \dot{y}(t,\mu) = \mathbb{J}_{2n} \nabla_y H(y(t,\mu);\mu), \\ y(0,\mu) = y_0(\mu), \end{cases} \tag{4.15}$$

with $\mu \in \mathcal{P} \subset \mathbb{R}^d$ being a d-dimensional parameter space. Let $\mathcal{Z}^{\mathcal{P}}$ be the set of solutions to (4.15) defined as

$$\mathcal{Z}^{\mathcal{P}} = \left\{ y(t,\mu) : t \in [t_0, t_{\mathrm{end}}], \mu \in \mathcal{P} \right\} \subset \mathbb{R}^{2n}.$$

For the sake of simplicity, in the previous sections we have only considered the nonparametric case. The extension of SVD-based methods for basis generations to (4.15) is straightforward on paper, but it is often computationally problematic in practice as the number of snapshots increases. Similar to other SVD-based algorithms, the methods described in the previous sections require the computation of the solution to (4.15) corresponding to a properly chosen discrete set of parameters $S^\mu = \{\mu_j\}_{j=1}^p \subset \mathcal{P}$ and time instances $S^t = \{t_i\}_{i=1}^N$, defined *a priori*, and constituting the sampling set $S^{\mu,t} := S^\mu \times S^t$. Random or structured strategies exist to define the set S^μ, such as the Monte Carlo sampling, Latin hypercube sampling, and sparse grids [12], while S^t is a subset of the time-discretization, usually dictated by the integrator of choice. The set of snapshots corresponding to the sampling set $S^{\mu,t}$ must provide a "good" approximation of the solution manifold and should not miss relevant parts of the time-parameter domain. Once the sampling set $S^{\mu,t}$ has been fixed, the matrix M_y, M_1, or M_2, depending on the method of choice, is assembled, and its singular value decomposition is computed. Even though a certain amount of computational complexity is tolerated in the *offline* stage to obtain a significant speed-up in the *online* stage, the evaluation of the high-fidelity solution for a large sampling set and the SVD of the corresponding snapshot matrix are often impractical or not even feasible. Hence, an efficient approach is an incremental procedure. The reduced basis, in which the column space represents the approximating manifold, is improved iteratively by adding basis vectors as columns. The candidate basis vector is chosen as the maximizer of a much cheaper optimization problem. This summarizes the philosophy of the greedy strategy applied to RB methods [6,9], which requires two main ingredients: the definition of an error indicator and a process to add a candidate column vector to the basis.

Let U_k be an orthonormal reduced basis produced after k steps of the algorithm. In its idealized form, introduced in [51], the greedy algorithm uses the projection error

$$(t^*, \mu^*) := \operatorname*{argmax}_{(t_i,\mu_j)\in S^{\mu,t}} \left\| u(t_i,\mu_j) - U_k U_k^\top u(t_i,\mu_j) \right\|_2, \tag{4.16}$$

to identify the snapshot $u^* := u(t^*, \mu^*)$ that is worst approximated by the column space of U_k over the entire sampling set $S^{\mu,t}$. Let u_{k+1} be the vector obtained by orthonormalizing u^* with respect to U_k. Then the basis U_k is updated as $U_{k+1} = [U_k \;\; u_{k+1}]$. To avoid the accumulation of rounding errors, it is preferable to utilize backward stable orthogonalization processes, such as the modified Gram–Schmidt orthogonalization. The algorithm terminates when the basis reaches the desired dimension, or the error (4.16) is below a certain tolerance. In this sense, the basis U_{k+1} is *hierarchical* because its column space contains the column space of its previous iterations. This process is referred to as *strong greedy* method. Even though introduced as a heuristic procedure, interesting results regarding algebraic and exponential convergence have been formulated in [6,9], requiring the orthogonality of the basis in the corresponding proofs. However, in this form, the scheme cannot be efficiently implemented: the error indicator (4.16) is expensive to calculate because it requires all the snapshots of the training set $S^{\mu,t}$ to be accessible, relieving the computation only of the cost required for the SVD.

An adjustment of the *strong greedy* algorithm, known as *weak greedy* algorithm, assembles the snapshot matrix corresponding to $S^{\mu,t}$ iteratively while expanding the approximating basis. The idea is to replace (4.16) with a surrogate indicator $\eta : S^{\mu,t} \mapsto \mathbb{R}$ that does not demand the computation of the high-fidelity solution for the entire time-parameter domain.

In the case of elliptic PDEs, an *a-posteriori* residual-based error indicator requiring a polynomial computational cost in the approximation space dimension has been introduced in [49]. The substantial computational savings allow the choice of a more refined, and therefore representative, sampling set $S^{\mu,t}$. One might also use a goal-oriented indicator as the driving selection in the greedy process to obtain similar computational benefits. In this direction, in the framework of structure-preserving model order reduction, [1] suggests the Hamiltonian as a proxy error indicator. Suppose $A_{2k} = [E_k \ \mathbb{J}_{2n}^\top E_k]$, with $E_k = [e_1 \ \dots \ e_k]$, is a given ortho-symplectic basis and consider

$$(t^*, \mu^*) := \underset{(t_i,\mu_j)\in S^{\mu,t}}{\operatorname{argmax}} \left| H\big(y(t_i,\mu_j)\big) - H\big(A_{2k}A_{2k}^+ y(t_i,\mu_j)\big)\right|. \tag{4.17}$$

By [1, PROPOSITION 15], the error in the Hamiltonian depends only on the initial condition and the symplectic reduced basis. Hence, the indicator (4.17) does not require integrating in time the full system (4.15) over the entire set S^μ, but only over a small fraction of the parameter set, making the procedure fast. Hence, the parameter space can be explored first,

$$\mu^* := \underset{\mu_j\in S^\mu}{\operatorname{argmax}} \left| H\big(y_0(\mu_j)\big) - H\big(A_{2k}A_{2k}^+ y_0(\mu_j)\big)\right|, \tag{4.18}$$

to identify the value of the parameter that maximizes the error in the Hamiltonian as a function of the initial condition. This step may fail if $y_0(\mu_j) \in \operatorname{range}(A_{2k})$, $\forall j = 1, \dots, p$. Then (4.15) is temporally integrated to collect the snapshot matrix

$$M_g = \big[y(t_1,\mu^*) \quad \dots \quad y(t_N,\mu^*)\big].$$

Finally, the candidate basis vector $y^* = y(\mu^*, t^*)$ is selected as the snapshot that maximizes the projection error

$$t^* := \underset{t_i\in S^t}{\operatorname{argmax}} \left\| y(t_i,\mu^*) - A_{2k}A_{2k}^+ y(t_i,\mu^*)\right\|_2. \tag{4.19}$$

Standard orthogonalization techniques, such as QR methods, fail to preserve the symplectic structure [10]. In [1], the SR method [47], based on the symplectic Gram–Schmidt, is employed to compute the additional basis vector e_{k+1} that conforms to the geometric structure of the problem. To conclude the $(k+1)$th iteration of the algorithm, the basis A_{2k} is expanded in

$$A_{2(k+1)} = \big[E_k \quad e_{k+1} \quad \mathbb{J}_{2n}^\top E_k \quad \mathbb{J}_{2n}^\top e_{k+1}\big].$$

We stress that, with this method, known as symplectic greedy RB, two vectors, e_{k+1} and $\mathbb{J}_{2n}^\top e_{k+1}$, are added to the symplectic basis at each iteration, because of the structure of ortho-symplectic matrices. A different strategy, known as PSD-Greedy algorithm and partially based on the PSD SVD-like decomposition, has been introduced in [8], with the feature of not using orthogonal techniques to compress the matrix M_g. In [1], following the results

given in [9], the exponential convergence of the symplectic strong greedy method has been proved.

Theorem 4.2 ([9, THEOREM 20, P. A2632]). *Let $\mathcal{Z}^{\mathcal{P}}$ be a compact subset of $\mathbb{R}^{2n}$. Assume that the Kolmogorov m-width of $\mathcal{Z}^{\mathcal{P}}$ defined as*

$$d_m(\mathcal{Z}^{\mathcal{P}}) = \inf_{\substack{\mathcal{Z}_* \subset \mathbb{R}^{2n} \\ \dim(\mathcal{Z}_*)=m}} \sup_{v \in \mathcal{Z}^{\mathcal{P}}} \min_{w \in \mathcal{Z}_*} \|v - w\|_2,$$

decays exponentially fast, namely $d_m(\mathcal{Z}^{\mathcal{P}}) \leq c \exp(-\alpha m)$ with $\alpha > \log 3$. Then there exists $\beta > 0$ such that the symplectic basis A_{2k} generated by the symplectic strong greedy algorithm provides exponential approximation properties,

$$\left\| s - A_{2k} A_{2k}^{+} s \right\|_2 \leq C \exp(-\beta k), \tag{4.20}$$

for all $s \in \mathcal{Z}^{\mathcal{P}}$ and some $C > 0$.

Theorem 4.2 holds only when the projection error is used as the error indicator instead of the error in the Hamiltonian. However, it has been observed for different symplectic parametric problems [1] that the symplectic method using the loss in the Hamiltonian converges with the same rate of (4.20). The orthogonality of the basis is used to prove the convergence of the greedy procedure. In the case of a nonorthonormal symplectic basis, supplementary assumptions are required to ensure the convergence of the algorithm.

5. DYNAMICAL LOW-RANK REDUCED BASIS METHODS FOR HAMILTONIAN SYSTEMS

The Kolmogorov m-width of a compact set describes how well this can be approximated by a linear subspace of a fixed dimension m. A problem (4.15) is informally defined *reducible* if d_m decays sharply with m, implying the existence of a low-dimensional representation of $\mathcal{Z}^{\mathcal{P}}$. A slow decay limits the accuracy of any efficient projection-based reduction on linear subspaces, including all the methods discussed so far. For Hamiltonian problems, often characterized by the absence of physical dissipation due to the conservation of the Hamiltonian, we may have $d_m(\mathcal{Z}^{\mathcal{P}}) = \mathcal{O}(m^{-\frac{1}{2}})$ in case of discontinuous initial condition [23] for wave-like problems. Several techniques, either based on nonlinear transformations of the solution manifold to a reducible framework [39] or presented as online adaptive methods to target solution manifolds at fixed time [42], have been introduced to overcome the limitations of the linear approximating spaces. In different ways, they all abandon the framework of symplectic vector spaces. Therefore, none of them guarantees conservation of the symplectic structure in the reduction process. Musharbash et al. [38] proposed a dynamically orthogonal (DO) discretization of stochastic wave PDEs with a symplectic structure. In the following, we outline the structure-preserving dynamic RB method for parametric Hamiltonian systems, proposed by Pagliantini [40] in the spirit of the geometric reduction introduced in [15]. In contrast with traditional methods that provide a global basis, which is fixed in time, the gist of a dynamic approach is to evolve a local-in-time basis to provide an accurate

approximation of the solution to the parametric problem (4.15). The idea is to exploit the local low-rank nature of Hamiltonian dynamics in the parameter space. From a geometric perspective, the approximate solution evolves according to naturally constrained dynamics, rather than weakly enforcing the required properties, such as orthogonality or symplecticity of the RB representation, via Lagrange multipliers. This result is achieved by viewing the flow of the reduced model as prescribed by a vector field that is everywhere tangent to the desired manifold.

Suppose we are interested in solving (4.15) for a set of p vector-valued parameters $\eta_h = \{\mu_i\}_{i=1}^p$, sampled from $\mathcal{P}$. Then the Hamiltonian system, evaluated at η_h, can be recast as a set of ODEs in the matrix unknown $\mathcal{R} \in \mathbb{R}^{2n\times p}$,

$$\begin{cases} \dot{\mathcal{R}}(t) = \mathcal{X}_H(\mathcal{R}(t), \eta_h) = \mathbb{J}_{2n}\nabla_{\mathcal{R}} H(\mathcal{R}(t), \eta_h), \\ \mathcal{R}(t_0) = \mathcal{R}_0(\mu_h), \end{cases} \tag{5.1}$$

where H is a vector-valued Hamiltonian function, the jth column of $\mathcal{R}(t)$ is such that $\mathcal{R}_j(t) = y(t, \mu_j)$, and $(\nabla_{\mathcal{R}} H)_{i,j} := \partial H_j / \partial \mathcal{R}_{i,j}$. We consider an approximation of the solution to (5.1) of the form

$$\mathcal{R}(t) \approx R(t) = A(t)Z(t), \tag{5.2}$$

where $A(t) \in \mathbb{S}(2k, 2n)$, and $Z(t) \in \mathbb{R}^{2k\times p}$ is such that its jth column $Z_j(t)$ collects coefficients, with respect to the basis $A(t)$, of the approximation of $y(t, \mu_j)$. Despite being cast in the same framework of an RB approach, a stark difference between (5.2) and (3.2) lies in the time-dependency of the basis in (5.2).

Consider the manifold of $2n \times p$ matrices having at most rank $2k$, and defined as

$$\mathcal{Z}_{2n}^{\mathcal{P}} := \left\{R \in \mathbb{R}^{2n\times p} : R = AZ \text{ with } A \in \mathbb{S}(2k, 2n), Z \in \mathbb{Z}\right\}, \tag{5.3}$$

with the technical requirement

$$\mathbb{Z} := \left\{Z \in \mathbb{R}^{2k\times p} : \operatorname{rank}(ZZ^\top + \mathbb{J}_{2k}^\top ZZ^\top \mathbb{J}_{2k}) = 2k\right\}. \tag{5.4}$$

This represents a full-rank condition on Z to ensure uniqueness of the representation (5.2) for a fixed basis. The tangent vector at $R(t) = A(t)Z(t) \in \mathcal{Z}_{2n}^{\mathcal{P}}$ is given by $X = X_A Z + AX_Z$, where X_A and X_Z correspond to the tangent directions for the time-dependent matrices A and Z, respectively. Applying the orthogonality and symplecticity condition on $A(t)$, for all times t, results in

$$X_A^\top A + A^\top X_A = 0 \quad \text{and} \quad X_A^\top \mathbb{J}_{2n} A + A^\top \mathbb{J}_{2n} X_A = 0, \tag{5.5}$$

respectively. Using (5.5) and an additional gauge constraint to uniquely parametrize the tangent vectors X by the displacements X_A and X_Z, the tangent space of $\mathcal{Z}_{2n}^{\mathcal{P}}$ at $R = AZ$ can be characterized as

$$T_R\mathcal{M}_{2n}^{\mathcal{P}} = \left\{X \in \mathbb{R}^{2n\times p} : X = X_A Z + AX_Z, \right.$$
$$\left. \text{with } X_Z \in \mathbb{R}^{2k\times p}, X_A \in \mathbb{R}^{2n\times 2k}, X_A^\top A = 0, X_A \mathbb{J}_{2k} = \mathbb{J}_{2n} X_A\right\}.$$

The reduced flow describing the evolution of the approximation $R(t)$ is derived in [40] by projecting the full velocity field $\mathcal{X}_H$ in (5.1) onto the tangent space $T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}$ of $\mathcal{Z}_{2n}^{\mathcal{P}}$ at $R(t)$, i.e.,

$$\begin{cases} \dot{R}(t) = \Pi_{T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}} \mathcal{X}_H(R(t), \eta_h), \\ R(t_0) = U_0 Z_0. \end{cases} \tag{5.6}$$

To preserve the geometric structure of the problem, the projection operator $\Pi_{T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}}$ is a symplectomorphism (see Definition 2.3) for each realization of the parameter $\mu_j \in \eta_h$, in the sense given in the following proposition.

Proposition 5.1 ([40, PROPOSITION 4.3, P. 420]). *Let* $S := ZZ^\top + \mathbb{J}_{2k} ZZ^\top \mathbb{J}_{2k} \in \mathbb{R}^{2k\times 2k}$. *Then, the map*

$$\begin{aligned} \Pi_{T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}} : \mathbb{R}^{2n\times p} &\to T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}, \\ w &\mapsto (I_{2n} - AA^\top)(wZ^\top + \mathbb{J}_{2n} w Z^\top \mathbb{J}_{2k}^\top) S^{-1} Z + AA^\top w, \end{aligned} \tag{5.7}$$

is a symplectic projection, in the sense that

$$\sum_{j=1}^{p} \Omega^{\mu_j}(w - \Pi_{T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}} w, y) = 0, \quad \forall y \in \Pi_{T_{R(t)}\mathcal{Z}_{2n}^{\mathcal{P}}},$$

where Ω^{μ_j} *is the symplectic form associated with the parameter* μ_j.

The optimality of the reduced dynamics, in the Frobenius norm, follows from (5.6), where the flow of R is prescribed by the best low-rank approximation of the Hamiltonian velocity field vector $\mathcal{X}_H$ into the tangent space of the reduced manifold $\mathcal{Z}_{2n}^{\mathcal{P}}$. Using (5.7) and (5.6), it is straightforward to derive the evolution equations for $A(t)$ and $Z(t)$:

$$\begin{cases} \dot{Z}_j(t) = \mathbb{J}_{2n} \nabla_{Z_j} H(AZ_j, \mu_j), \\ \dot{A}(t) = (I_{2n} - AA^\top)(\mathbb{J}_{2n} YZ - YZ\mathbb{J}_{2n}^\top) S^{-1}, \\ A(t_0)Z(t_0) = A_0 Z_0, \end{cases} \tag{5.8}$$

with $Y := [\mathbb{J}_{2n}\nabla H(UZ_1, \mu_1) \ \ldots \ \mathbb{J}_{2n}\nabla H(UZ_p, \mu_p)]$.

The coefficients Z evolve according to a system of p independent Hamiltonian equations, each in $2n$ unknowns, corresponding to the symplectic Galerkin projection onto range(A) for each parameter instance in η_h, similarly to the global symplectic RB method (3.7). In (5.8), however, the basis A evolves in time according to a matrix equation in $2n \times 2k$ unknowns, affecting the projection. A crucial property of the structure of $A(t)$ is given in the following proposition.

Proposition 5.2 ([40, PROPOSITION 4.5, P. 423]). *If* $A_0 \in \mathbb{S}(2k, 2n)$ *then* $A(t) \in \mathbb{R}^{2n\times 2k}$ *solution of* (5.8) *satisfies* $A(t) \in \mathbb{S}(2k, 2n)$ *for all* $t > t_0$.

Standard numerical integrators, applied to (5.8), do not preserve, at the time-discrete level, the property in Proposition 5.2 and the ortho-symplectic structure is compromised after a single time step. In [40], two different intrinsic integrators have been investigated to preserve the ortho-symplecticity of the basis, based on Lie groups and tangent techniques.

Both methods require the introduction of a local chart defined on the tangent space $T_{A(t)}\mathbb{S}$ of the manifold $\mathbb{S}(2k, 2n)$ at $A(t)$, with

$$T_{A(t)}\mathbb{S} := \left\{V \in \mathbb{R}^{2n\times 2k} : A^\top V \in \mathfrak{g}(2k)\right\}$$

and $\mathfrak{g}(2k)$ being the vector space of skew-symmetric and Hamiltonian $2k \times 2k$ real square matrices. In terms of differential manifolds, $\mathfrak{g}(2k)$ represents, together with the Lie bracket $[\cdot,\cdot] : \mathfrak{g}(2k) \times \mathfrak{g}(2k) \mapsto \mathfrak{g}(2k)$ defined as the matrix commutator $[M, L] := ML - LM$, with $M, L \in \mathfrak{g}(2k)$, the Lie algebra corresponding to the Lie group $\mathbb{S}(2k, 2k)$. The idea is to recast the basis equation in (5.8) in an evolution equation in the corresponding Lie algebra. The linearity of Lie algebras allows to compute, via explicit Runge–Kutta methods, numerical solutions that remain on the Lie algebra. Finally, the Cayley transform $\mathrm{cay} : \mathfrak{g}(2k) \mapsto \mathbb{S}(2k, 2k)$ is exploited to generate local coordinate charts and retraction/inverse retraction maps, used to recover the solution in the manifold of rectangular ortho-symplectic matrices. In [29], the structure-preserving dynamical RB-method has been paired with a rank-adaptive procedure, based on a residual error estimator, to dynamically update also the dimension of the basis.

6. EXTENSIONS TO MORE GENERAL HAMILTONIAN PROBLEMS

6.1. Dissipative Hamiltonian systems

Many areas of engineering require a more general framework than the one offered by classical Hamiltonian systems, requiring the inclusion of energy-dissipating elements. While the principle of energy conservation is still used to describe the state dynamics, dissipative perturbations must be modeled and introduced in the Hamiltonian formulation (3.1). Dissipative Hamiltonian systems, with so-called Rayleigh type dissipation, are considered a special case of forced Hamiltonian systems, with the state $y = (q, p) \in \mathbb{R}^{2n}$, with $q, p \in \mathbb{R}^n$, following the time evolution given by

$$\begin{cases} \dot{y}(t) = \mathbb{J}_{2n}\nabla H(y(t)) + \mathcal{X}_F(y(t)), \\ y(0) = y_0, \end{cases} \tag{6.1}$$

where $\mathcal{X}_F \in \mathbb{R}^{2n}$ is a velocity field, introducing dissipation, of the form

$$\mathcal{X}_F := \begin{bmatrix} 0_n \\ f_H(y(t)) \end{bmatrix}. \tag{6.2}$$

We require $\mathcal{X}_F$ to satisfy $(\nabla_y H)^\top \mathcal{X}_F \leq 0$, $\forall y \in \mathbb{R}^{2n}$, to represent a dissipative term, and therefore

$$(\nabla_p H)^\top f_H \leq 0. \tag{6.3}$$

In terms of Rayleigh dissipation theory, there exists a symmetric positive semidefinite matrix $R(q) \in \mathbb{R}^{n\times n}$ such that $f_H = -R(q)\dot{q}(p, q)$ and (6.3) reads

$$(\nabla_p H)^\top f_H = \dot{q}^\top f_H = -\dot{q}^\top R(q)\dot{q} \leq 0.$$

Several strategies have been proposed to generate stable reduced approximations of (6.1), based on Krylov subspaces or POD [27, 45]. In [25], without requiring the symplecticity of the reduced basis, the gradient of the Hamiltonian vector field is approximated using a projection matrix W, i.e., $\nabla_y H(Uz) \approx W \nabla_z H_{\mathrm{RB}}(z)$, which results in a noncanonical symplectic reduced form. The stability of the reduced model is then achieved by preserving the passivity of the original formulation. A drawback of such an approach is that, while viable for nondissipative formulations, it does not guarantee the same energy distribution of (6.1) between dissipative and null energy contributors. In the following, we show that the techniques based on symplectic geometry introduced in the previous sections can still be used in the dissipative framework described in (6.1) with limited modifications to obtain consistent and structured reduced models. Let us consider an ortho-symplectic basis $A \in \mathbb{S}(2k, 2n)$ and the reduced basis representation $y \approx Az$, with $z = (r, s) \in \mathbb{R}^{2k}$ being the reduced coefficients of the representation and $r, s \in \mathbb{R}^k$ being the generalized phase coordinates of the reduced model. The basis A can be represented as

$$A = \begin{bmatrix} A_{qr} & A_{qs} \\ A_{pr} & A_{ps} \end{bmatrix}, \tag{6.4}$$

with $A_{qr}, A_{qs}, A_{pr}, A_{ps} \in \mathbb{R}^{n \times k}$ being the blocks, the indices of which are chosen to represent the interactions between the generalized phase coordinates of the two models, such that $q = A_{qr} r + A_{qs} s$ and $p = A_{pr} r + A_{ps} s$. Following [43], the symplectic Galerkin projection of (6.1) reads

$$\dot{z} = A^+ \big(\mathcal{X}_H(Az) + \mathcal{X}_F(Az) \big) = \mathbb{J}_{2k} \nabla_z H_{\mathrm{RB}}(z) + A^+ \mathcal{X}_F(Az) = \mathcal{X}_{H_{\mathrm{RB}}} + A^+ \mathcal{X}_F, \tag{6.5}$$

with

$$A^+ \mathcal{X}_F = \begin{bmatrix} A_{ps}^\top & -A_{qs}^\top \\ -A_{pr}^\top & A_{qr}^\top \end{bmatrix} \begin{bmatrix} 0_n \\ f_H \end{bmatrix} = \begin{bmatrix} -A_{qs}^\top f_H \\ A_{qr}^\top f_H \end{bmatrix}. \tag{6.6}$$

We note that, in (6.5), the reduced dynamics is described as the sum of a Hamiltonian vector field and a term that, for a general choice of the symplectic basis A and hence of $A_{qs}^\top$, does not represent a dissipative term in the form of a vertical velocity field. The Cotangent Lift method, described in Section 4.1, enforces by construction the structure of vertical velocity field because $A_{qs} = 0$. It can be shown [43] that dissipativity is also preserved since the rate of energy variation of the reduced system is non-positive, i.e.,

$$\nabla_s H_{\mathrm{RB}}(Az)(A_{qr}^\top f_H) = \dot{r}^\top (A_{qr}^\top f_H) = -(A_{qr}\dot{r})^\top R(A_{qr} s)(A_{qr}\dot{r}) \leq 0. \tag{6.7}$$

However, time discretization of the reduced dissipative model is not trivial. Even though the dissipative Hamiltonian structure is preserved by the reduction process, standard numerical integrators do not preserve the same structure at the fully discrete level.

A completely different approach is proposed in [2], where (6.1) is paired with a canonical heat bath, absorbing the energy leakage and expanding the system to the canonical Hamiltonian structure. Consider a dissipative system characterized by the quadratic Hamiltonian $H(y) = \frac{1}{2} y^\top K^\top K y$. Following [17], such a system admits a time dispersive and

dissipative (TDD) formulation

$$\begin{cases} \dot{y} = \mathbb{J}_{2n} K^\top f(t), \\ y(0) = y_0, \end{cases} \tag{6.8}$$

with $f(t)$ being the solution to the integral equation

$$f(t) + \int_0^t \chi(t-s) f(s) ds = Ky, \tag{6.9}$$

also known as a *generalized material relation*. The square time-dependent matrix $\chi \in \mathbb{R}^{2n\times 2n}$ is the *generalized susceptibility* of the system, and it is bounded with respect to the Frobenius norm. Physically, it encodes the accumulation of the dissipation effect in time, starting from the initial condition. When $\chi = 0_{2n}$, (6.8) is equivalent to (3.1). Under physically natural assumptions on χ (see **[17, THEOREM 1.1, P. 975]** for more details), system (6.8) admits a quadratic Hamiltonian extension (QHE) to a canonical Hamiltonian system. This extension is obtained by defining an isometric injection $I : \mathbb{R}^{2n} \mapsto \mathbb{R}^{2n} \times \mathcal{H}^{2n}$, where $\mathcal{H}^{2n}$ is a suitable Hilbert space, and reads

$$\begin{cases} \dot{y} = \mathbb{J}_{2n} K^\top f(t), \\ \partial_t \phi = \theta(t,x), \\ \partial_t \theta = \partial_x^2 \phi(t,x) + \sqrt{2}\delta_0(x) \cdot \sqrt{\chi} f(t), \end{cases} \tag{6.10}$$

where ϕ and θ are vector-valued functions in $\mathcal{H}^{2n}$, δ_0 is the Dirac-delta function, and f solves

$$f(t) + \sqrt{2} \cdot \sqrt{\chi} \phi(t,0) = Ky(t).$$

It can be shown that system (6.10) has the form of a conserved Hamiltonian system with the extended Hamiltonian

$$H_{ex}(y,\phi,\theta) = \frac{1}{2}\left(\|Ky - \phi(t,0)\|_2^2 + \|\theta(t)\|_{\mathcal{H}^{2n}}^2 + \|\partial_x \phi(t)\|_{\mathcal{H}^{2n}}^2 \right),$$

and can be reduced, while preserving its geometric structure, using any of the standard symplectic techniques. We refer the reader to **[2]** for a formal derivation of the reduced model obtained by projecting (6.10) on a symplectic subspace and for its efficient time integration. The method extends trivially to more general Hamiltonian functions, as long as the dissipation is linear in (6.9).

6.2. Noncanonical Hamiltonian systems

The canonical Hamiltonian problem (3.1) has been defined under the assumption that a canonical system of coordinates for the symplectic solution manifold is given, and the Hamiltonian vector can be represented as (2.5). However, many Hamiltonian systems, such as the KdV and Burgers equations, are naturally formulated in terms of a noncanonical basis, resulting in the following description of their dynamics:

$$\begin{cases} \dot{y}(t) = J_{2n} \nabla_y H(y(t)), \\ y(0) = y_0, \end{cases} \tag{6.11}$$

with $J_{2n} \in \mathbb{R}^{2n\times 2n}$ being invertible and skew-symmetric. A reduction strategy, involving the noncanonical formulation (6.11) and based on POD, has been proposed in [21]. Consider the RB ansatz $y \approx Uz$, with $U \in \mathbb{R}^{2n\times k}$ as an orthonormal basis obtained by applying the POD algorithm to the matrix of snapshots collected by solving the full model. The Galerkin projection of (6.11) reads

$$\dot{z} = U^\top J_{2n} \nabla_y H(Uz), \tag{6.12}$$

with the time derivative of the Hamiltonian function, evaluated at the reduced state, given by

$$\dot{H}(Uz) = \dot{z}^\top \big(\nabla_z H(Uz)\big) = \big(\nabla_y H(Uz)\big)^\top J_{2n}^\top U U^\top \nabla_y H(Uz). \tag{6.13}$$

As expected, the Hamiltonian structure is lost in (6.12) and the energy of the system, represented by the Hamiltonian, is no longer preserved in time because $J_{2n} U U^\top$ is not skew-symmetric. Both issues are solved in [21] by considering a matrix W, with the same properties of $\mathbb{J}_{2n}$, such that the relation

$$U^\top J_{2n} = W U^\top \tag{6.14}$$

is satisfied. We stress that a condition similar to (6.14) naturally holds in the canonical Hamiltonian setting for a symplectic basis and has been used to derive Hamiltonian reduced models using the symplectic Galerkin projection. A candidate W is identified in [21] by solving the normal equation related to (6.14), i.e., $W = U^\top J_{2n} U$. For invertible skew-symmetric operators J_{2n} that might depend on the state variables y, Miyatake has introduced in [35] a hyperreduction technique that preserves the skew-symmetric structure of the J_{2n} operator.

Formulation (6.11) is further generalized with the characterization of the phase-space as a Poisson manifold, defined as a $2n_P$-dimensional differentiable manifold $\mathcal{M}_P$ equipped with a Poisson bracket $\{\cdot,\cdot\} : C^\infty(\mathcal{M}_p) \times C^\infty(\mathcal{M}_p) \mapsto C^\infty(\mathcal{M}_p)$ satisfying the conditions of bilinearity, skew-symmetry, the Jacobi identity, and the Leibniz' rule. Since derivations on $C^\infty(\mathcal{M}_P)$ are represented by smooth vector fields, for each Hamiltonian function $H \in C^\infty(\mathcal{M}_P)$, there exists a vector $\mathcal{X}_H$ that determines the following dynamics:

$$\begin{cases} \dot{y}(t) = \mathcal{X}_H(y) = J_{2n_P}(y) \nabla_y H(y(t)), \\ y(0) = y_0, \end{cases} \tag{6.15}$$

with the Poisson tensor J_{2n_p} being skew-symmetric, state-dependent, and generally not invertible. The flow of the Hamiltonian vector field $\mathcal{X}_H(y)$, which is a Poisson map and therefore preserves the Poisson bracket structure via its pullback, also preserves the rank $2n$ of the Poisson tensor $J_{2n_P}(y)$. Moreover, $r = 2n_P - 2n$ represents the number of independent nonconstant functions on $\mathcal{M}_P$ that $\{\cdot,\cdot\}$ commutes with all the other functions in $C^\infty(\mathcal{M}_P)$. These functions are known as Casimirs of the Poisson bracket and their gradients belong to the kernel of $J_{2n_P}(y)$, making them independent of the dynamics of (6.15) and only representing geometric constraints on configurations of the generalized phase-state space.

An interesting relation between symplectic and Poisson manifolds is offered by the Lie–Weinstein splitting theorem, stating that locally, in the neighborhood $\mathcal{U}_{y^*}$ of any point $y^* \in \mathcal{M}_P$, a Poisson manifold can be split into a $2n$-dimensional symplectic manifold $\mathcal{M}$ and

an r-dimensional Poisson manifold M. Following on this result, Darboux' theorem guarantees the existence of local coordinates $(q_1, \dots, q_n, p_1, \dots, p_n, c_1, \dots, c_r)$, where $\{q_i, p_i\}_{i=1}^2$ corresponds to canonical symplectic coordinates and $\{c_i\}_{i=1}^r$ are the Casimirs, such that the Poisson tensor $J_{2n_p}(y)$ is recast, via Darboux' map, in the canonical form $J_{2n_p}^C$, i.e.,

$$J_{2n_p}^C = \begin{matrix} \\ 2n \\ r \end{matrix} \begin{matrix} \begin{matrix} 2n & r \end{matrix} \\ \begin{bmatrix} \mathbb{J}_{2n} & 0 \\ 0 & 0 \end{bmatrix} \end{matrix},$$

with $\mathbb{J}_{2n} \in \mathbb{R}^{2n \times 2n}$ being the canonical Poisson tensor defined in (2.3).

In [28], a quasistructure-preserving algorithm for problems of the form (6.15) has been proposed, leveraging the Lie–Weinstein splitting, an approximation of the Darboux' map and traditional symplectic RB techniques. Let

$$\begin{cases} y^{j+1} = y^j + \Delta t J_{2n_p}(\tilde{y}^j) \nabla_y H(\tilde{y}^j), \\ y^0 = y_0, \end{cases} \tag{6.16}$$

be the fully-discrete formulation of (6.15), where j is the integration index, and $\tilde{y}^j$ represents intermediate state/states dictated by the temporal integrator of choice. Given $\mathcal{M}_{P,j}$, an open subset of $\mathcal{M}_P$ comprising the discrete states y^j, $\tilde{y}^j$, and y^{j+1}, the authors of [28] introduce an approximation $\varphi_{j+\frac{1}{2}} : \mathcal{M}_{P,j} \mapsto \mathcal{M}_s \times \mathcal{N}_j$ of the Darboux' map at $\tilde{y}^j$, with $\mathcal{M}_s$ being a $2N$-dimensional canonical symplectic manifold and $\mathcal{N}_j$ approximating the null space of the Poisson structure. The proposed approximation exploits a Cholesky-like decomposition (see [28, PROPOSITION 2.11, P. 1708]) of the noncanonical rank-deficient $J_{2n_p}(\tilde{y}^j)$ and exactly preserves the dimension of $\mathcal{N}_j$, hence the number of independent Casimirs. By introducing the natural transition map $T_j := \varphi_{j+\frac{1}{2}} \cdot \varphi_{j-\frac{1}{2}}^{-1}$ between the neighboring and overlapping subsets $\mathcal{M}_{j-1}$ and $\mathcal{M}_j$, problem (6.16) is locally recast in the canonical form

$$\begin{cases} \bar{y}^{j+1} = T_j \bar{y}^j + \Delta t J_{2n_p}^C \nabla_{\bar{y}} H^j(\bar{\tilde{y}}^j), \\ \bar{y}^0 = y_0, \end{cases} \tag{6.17}$$

where $\bar{y}^{j+1} := \varphi_{j+\frac{1}{2}} y^{j+1}$, $\bar{y}^j := \varphi_{j+\frac{1}{2}} y^j$, $\bar{\tilde{y}}^{j+1} := \varphi_{j+\frac{1}{2}} \tilde{y}^j$, and $H^j(\bar{\tilde{y}}^j) := H(\varphi_{j+\frac{1}{2}}^{-1}(\bar{\tilde{y}}^j))$. Even though the flow of (6.17) is not a *global* $J_{2n_p}^C$-Poisson map because the splitting is not exact, the approximation is *locally* structure-preserving for each neighborhood $\mathcal{M}_{P,j}$. By exploiting a similar splitting principle, the canonical Poisson manifold $\mathcal{M}_s \times \mathcal{N}_j$ is projected on a reduced Poisson manifold $\mathcal{A} \times \mathcal{N}_j$, with the reduction acting only on the symplectic component of the splitting and $\dim(\mathcal{A}) = 2k \ll 2n$. The corresponding reduced model is obtained via Galerkin projection of (6.17) using an orthogonal J_{2k}^C-symplectic basis of dimension $2k$, generated via a greedy iterative process inspired by the symplectic greedy method described in Section 4.3. Different theoretical estimates and numerical investigations show the proposed technique's accuracy, robustness, and conservation properties, up to errors in the Poisson tensor approximation.

7. CONCLUSION

We provided an overview of model reduction methods for Hamiltonian problems. The symplectic Galerkin projection has been discussed as a tool to generate a reduced Hamiltonian approximation of the original dynamics. PSD algorithms used to compute low-order projection on symplectic spaces have been introduced and compared. Such strategies have been classified in ortho-symplectic and symplectic procedures, depending on the structure of the computed RB. A greedy alternative for the generation of ortho-symplectic basis, characterized by an exponentially fast convergence, has been illustrated as an efficient iterative approach to overcome the computational cost associated with SVD-based techniques that require a fine sampling of the solution manifold of the high-dimensional problem. The potential local low-rank nature of Hamiltonian dynamics has been addressed by a symplectic dynamical RB method. The innovative idea of the dynamical approach consists in evolving the approximating symplectic reduced space in time along a trajectory locally constrained on the tangent space of the high-dimensional dynamics. For problems where the Hamiltonian dynamics is coupled with a dissipative term, structure-preserving reduced models can be constructed with the symplectic reduction process by resorting to an extended nondissipative Hamiltonian reformulation of the system. Finally, we have described RB strategies to reduce problems having a noncanonical Hamiltonian structure that either enforce properties typical of a symplectic basis or use canonical symplectic reductions as an intermediate step to preserve the structure of the original model.

REFERENCES

[1] B. M. Afkham and J. S. Hesthaven, Structure preserving model reduction of parametric Hamiltonian systems. *SIAM J. Sci. Comput.* **39** (2017), no. 6, A2616–A2644.

[2] B. M. Afkham and J. S. Hesthaven, Structure-preserving model-reduction of dissipative Hamiltonian systems. *J. Sci. Comput.* **81** (2019), no. 1, 3–21.

[3] B. M. Afkham, N. Ripamonti, Q. Wang, and J. S. Hesthaven, Conservative model order reduction for fluid flow. In *Quantification of Uncertainty: Improving Efficiency and Technology*, 67–99, Lect. Notes Comput. Sci. Eng. 137, Springer, Cham, 2020.

[4] S. Ahmed, On theoretical and numerical aspects of symplectic Gram–Schmidt-like algorithms. *Numer. Algorithms* **39** (2005), no. 4, 437–462.

[5] F. Ballarin, A. Manzoni, A. Quarteroni, and G. Rozza, Supremizer stabilization of POD–Galerkin approximation of parametrized steady incompressible Navier–Stokes equations. *Internat. J. Numer. Methods Engrg.* **102** (2015), no. 5, 1136–1161.

[6] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.* **43** (2011), no. 3, 1457–1472.

[7] P. Buchfink, A. Bhatt, and B. Haasdonk, Symplectic model order reduction with non-orthonormal bases. *Math. Comput. Appl.* **24** (2019), no. 2, 43.

[8] P. Buchfink, B. Haasdonk, and S. Rave, PSD-greedy basis generation for structure-preserving model order reduction of Hamiltonian systems. In *Proceedings of ALGORITMY*, pp. 151–160, Spektrum STU, 2020.

[9] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici, A priori convergence of the Greedy algorithm for the parametrized reduced basis method. *ESAIM Math. Model. Numer. Anal.* **46** (2012), no. 3, 595–603.

[10] A. Bunse-Gerstner, Matrix factorizations for symplectic QR-like methods. *Linear Algebra Appl.* **83** (1986), 49–77.

[11] K. Carlberg, Y. Choi, and S. Sargsyan, Conservative model reduction for finite-volume models. *J. Comput. Phys.* **371** (2018), 280–314.

[12] W. G. Cochran, *Sampling techniques*. John Wiley & Sons, 2007.

[13] G. Darboux, Sur le probleme de Pfaff. *Bull. Sci. Math. Astron.* **6** (1882), no. 1, 14–36.

[14] P. Feldmann and R. W. Freund, Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **14** (1995), no. 5, 639–649.

[15] F. Feppon and P. F. Lermusiaux, A geometric approach to dynamical model order reduction. *SIAM J. Matrix Anal. Appl.* **39** (2018), no. 1, 510–538.

[16] L. Fick, Y. Maday, A. T. Patera, and T. Taddei, A stabilized POD model for turbulent flows over a range of Reynolds numbers: Optimal parameter sampling and constrained projection. *J. Comput. Phys.* **371** (2018), 214–243.

[17] A. Figotin and J. H. Schenker, Hamiltonian structure for dispersive and dissipative dynamical systems. *J. Stat. Phys.* **128** (2007), no. 4, 969–1056.

[18] S. Fiori, A Riemannian steepest descent approach over the inhomogeneous symplectic group: Application to the averaging of linear optical systems. *Appl. Math. Comput.* **283** (2016), 251–264.

[19] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas, Non-linear model reduction for uncertainty quantification in large-scale inverse problems. *Internat. J. Numer. Methods Engrg.* **81** (2010), no. 12, 1581–1608.

[20] K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ error bounds. *Internat. J. Control* **39** (1984), no. 6, 1115–1193.

[21] Y. Gong, Q. Wang, and Z. Wang, Structure-preserving Galerkin POD reduced-order modeling of Hamiltonian systems. *Comput. Methods Appl. Mech. Engrg.* **315** (2017), 780–798.

[22] W. B. Gordon, On the completeness of Hamiltonian vector fields. *Proc. Amer. Math. Soc.* (1970), 329–331.

[23] C. Greif and K. Urban, Decay of the Kolmogorov N-width for wave problems. *Appl. Math. Lett.* **96** (2019), 216–222.

[24] M. Gromov, Pseudo holomorphic curves in symplectic manifolds. *Invent. Math.* **82** (1985), no. 2, 307–347.

[25] S. Gugercin, C. Beattie, and S. Chaturantabut, Structure-preserving model reduction for nonlinear port-Hamiltonian systems. *SIAM J. Sci. Comput.* **38** (2016), no. 5, B837–B865.

[26] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration*. Springer Ser. Comput. Math. 31, 2006.

[27] C. Hartmann, V. M. Vulcanov, and C. Schütte, Balanced truncation of linear second-order systems: a Hamiltonian approach. *Multiscale Model. Simul.* **8** (2010), no. 4, 1348–1367.

[28] J. S. Hesthaven and C. Pagliantini, Structure-preserving reduced basis methods for Poisson systems. *Math. Comp.* **90** (2021), no. 330, 1701–1740.

[29] J. S. Hesthaven, C. Pagliantini, and N. Ripamonti, Rank-adaptive structure-preserving reduced basis methods for Hamiltonian systems. 2020, arXiv:2007.13153.

[30] J. S. Hesthaven, G. Rozza, and B. Stamm, *Certified reduced basis methods for parametrized partial differential equations*. Springer, 2016. DOI 10.1007/978-3-319-22470-1.

[31] C. Huang, K. Duraisamy, and C. L. Merkle, Investigations and improvement of robustness of reduced-order models of reacting flow. *AIAA J.* **57** (2019), no. 12, 5377–5389.

[32] M. Karow, D. Kressner, and F. Tisseur, Structured eigenvalue condition numbers. *SIAM J. Matrix Anal. Appl.* **28** (2006), no. 4, 1052–1068.

[33] S. Lall, P. Krysl, and J. E. Marsden, Structure-preserving model reduction for mechanical systems. *Phys. D* **184** (2003), no. 1–4, 304–318.

[34] J. E. Marsden and T. S. Ratiu, *Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems*. Texts Appl. Math. 17, Springer, 2013.

[35] Y. Miyatake, Structure-preserving model reduction for dynamical systems with a first integral. *Jpn. J. Ind. Appl. Math.* **36** (2019), no. 3, 1021–1037.

[36] B. Moore, Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Trans. Automat. Control* **26** (1981), no. 1, 17–32.

[37] R. Muruhan and L. R. Petzold, A new look at proper orthogonal decomposition. *SIAM J. Numer. Anal.* **41** (2003), no. 5, 1893–1925.

[38] E. Musharbash, F. Nobile, and E. Vidličková, Symplectic dynamical low rank approximation of wave equations with random parameters. *BIT* **60** (2020), no. 4, 1153–1201.

[39] M. Ohlberger and S. Rave, Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing. *C. R. Math.* **351** (2013), no. 23–24, 901–906.

[40] C. Pagliantini, Dynamical reduced basis methods for Hamiltonian systems. *Numer. Math.* (2021), 1–40.

[41] C. Paige and C. Van Loan, A Schur decomposition for Hamiltonian matrices. *Linear Algebra Appl.* **41** (1981), 11–32.

[42] B. Peherstorfer and K. Willcox, Online adaptive model reduction for nonlinear systems via low-rank updates. *SIAM J. Sci. Comput.* **37** (2015), no. 4, A2123–A2150.

[43] L. Peng and K. Mohseni, Geometric model reduction of forced and dissipative Hamiltonian systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7465–7470, IEEE, 2016. DOI 10.1109/CDC.2016.7799422.

[44] L. Peng and K. Mohseni, Symplectic model reduction of Hamiltonian systems. *SIAM J. Sci. Comput.* **38** (2016), no. 1, A1–A27.

[45] R. V. Polyuga and A. Van der Schaft, Structure preserving model reduction of port-Hamiltonian systems by moment matching at infinity. *Automatica* **46** (2010), no. 4, 665–672.

[46] C. W. Rowley, T. Colonius, and R. M. Murray, Model reduction for compressible flows using POD and Galerkin projection. *Phys. D* **189** (2004), 115–129.

[47] A. Salam and E. Al-Aidarous, Equivalence between modified symplectic Gram–Schmidt and Householder SR algorithms. *BIT* **54** (2014), no. 1, 283–302.

[48] W. H. A. Schilders, H. A. Van der Vorst, and J. Rommes, *Model order reduction: theory, research aspects and applications* 13, Springer, Berlin, 2008.

[49] S. Sen, Reduced-basis approximation and a posteriori error estimation for many-parameter heat conduction problems. *Numer. Heat Transf., Part B, Fundam.* **54** (2008), no. 5, 369–389.

[50] L. Sirovich, Turbulence and the dynamics of coherent structures, Parts I, II and III. *Quart. Appl. Math.* (1987), 561–590.

[51] K. Veroy, C. Prud'homme, D. Rovas, and A. T. Patera, A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *16th AIAA computational fluid dynamics conference*, p. 3847, AIAA, 2003. DOI 10.2514/6.2003-3847.

[52] R. Wu, R. Chakrabarti, and H. Rabitz, Optimal control theory for continuous-variable quantum gates. *Phys. Rev. A* **77** (2008), no. 5, 052303.

[53] H. Xu, An SVD-like matrix decomposition and its applications. *Linear Algebra Appl.* **368** (2003), 1–24.

[54] R. Zimmerman, A. Vendl, and S. Görtz, Reduced-order modeling of steady flows subject to aerodynamic constraints. *AIAA J.* **52** (2014), no. 2, 255–266.

JAN S. HESTHAVEN

Institute of Mathematics, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, jan.hesthaven@epfl.ch

CECILIA PAGLIANTINI
Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands, c.pagliantini@tue.nl

NICOLÒ RIPAMONTI
Institute of Mathematics, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, nicolo.ripamonti@epfl.ch

NUMERICAL STABILITY OF ALGORITHMS AT EXTREME SCALE AND LOW PRECISIONS

NICHOLAS J. HIGHAM

ABSTRACT

The largest dense linear systems that are being solved today are of order $n = 10^7$. Single-precision arithmetic, which has a unit roundoff $u \approx 10^{-8}$, is widely used in scientific computing, and half-precision arithmetic, with $u \approx 10^{-4}$, is increasingly being exploited as it becomes more readily available in hardware. Standard rounding error bounds for numerical linear algebra algorithms are proportional to $p(n)u$, with p growing at least linearly with n. Therefore we are at the stage where these rounding error bounds are not able to guarantee any accuracy or stability in the computed results for some extreme-scale or low-accuracy computations. We explain how rounding error bounds with much smaller constants can be obtained. Blocked algorithms, which break the data into blocks of size b, lead to a reduction in the error constants by a factor b or more. Two architectural features also reduce the error constants: extended precision registers and fused multiply–add operations, either at the scalar level or in mixed precision block form. We also discuss a new probabilistic approach to rounding error analysis that provides error constants that are the square roots of those of the worst-case bounds. Combining these different considerations provides new understanding of the numerical stability of extreme scale and low precision computations in numerical linear algebra.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 65G50; Secondary 65F05

KEYWORDS

floating-point arithmetic, backward error analysis, numerical stability, probabilistic rounding error analysis, blocked algorithm, fused multiply-add, mixed precision computation

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5098–5117
DOI 10.4171/ICM2022/74

1. INTRODUCTION

We are approaching the exascale computing era, in which the world's fastest computers will be able to perform 10^{18} double-precision floating-point operations (flops) per second. With the increased speed comes an increase in the size of problems that can be solved in reasonable time, provided that sufficient memory is available.

A problem of particular interest is solving a dense system of linear equations $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $b \in \mathbb{R}^n$. Table 1 shows the sizes of large systems that have been solved at different points in time. The data is taken from the TOP500[1] list, which ranks the world's fastest computers by their speed (measured in flops per second) in solving a random linear system $Ax = b$ by LU factorization with partial pivoting. Generally, a benchmark run needs to be done with the largest n possible in order to obtain the best performance, so the tabulated values give an indication of the largest systems solved at each point in time. Table 1 suggests that the size of the largest linear systems being solved is growing by roughly a factor 10 each decade.

The standard componentwise backward error result for a solution $\hat{x}$ computed by LU factorization in floating-point arithmetic is as follows [22, THM. 9.4]. We write $|A| = (|a_{ij}|)$ and inequalities between matrices hold componentwise. We need the constant

$$\gamma_n = \frac{nu}{1-nu},$$

where u is the unit roundoff, which is $u = 2^{-t}$ for a base-2 floating-point arithmetic with t bits in the significand.

Theorem 1.1. *Let $A \in \mathbb{R}^{n\times n}$ and suppose that LU factorization produces computed LU factors $\hat{L}$, $\hat{U}$, and a computed solution $\hat{x}$ to $Ax = b$. Then there is a matrix ΔA such that*

$$(A + \Delta A)\hat{x} = b, \quad |\Delta A| \le \gamma_{3n}|\hat{L}||\hat{U}|. \tag{1.1}$$

Ideally, we would like the backward error matrix ΔA in (1.1) to satisfy $\|\Delta A\|_\infty \approx u\|A\|_\infty$. It can be shown that

$$\||\hat{L}||\hat{U}|\|_\infty \le p(n)\rho_n\|A\|_\infty \tag{1.2}$$

for a quadratic polynomial p [22, LEMMA 9.6], where the growth factor $\rho_n \ge 1$ measures the growth of elements during the factorization process. If, however, we make the very favorable assumption that $\||\hat{L}||\hat{U}|\|_\infty \approx \|A\|_\infty$ then we obtain

$$\frac{\|\Delta A\|_\infty}{\|A\|_\infty} \lesssim \gamma_{3n} = 3nu + O(u^2). \tag{1.3}$$

For the largest n in Table 1, in IEEE double-precision arithmetic (see Table 2), we have $nu \approx (2.1 \times 10^7)(1.11 \times 10^{-16}) \approx 2.3 \times 10^{-9}$, so even with these favorable assumptions our bound indicates the potential for a significant loss of numerical stability. If we work in IEEE single precision then $nu \approx (2.1 \times 10^7)(5.96 \times 10^{-8}) \approx 1.25$, and our backward error bound is of order 1, suggesting the possibility of a complete loss of stability.

1 http://www.top500.org.

Machine	Date	n
Fugaku	June 2021	2.1×10^7
Jaguar	June 2010	6.3×10^6
ASCI RED	June 2000	3.6×10^5
CM-5/1024	June 1993	5.2×10^4

TABLE 1
Size of large linear systems solved. The data is from the TOP500.

Precision	Name	(sig, exp)	u	$x_{\min}$	$x_{\max}$
Half	bfloat16	(8, 8)	3.91×10^{-3}	1.18×10^{-38}	3.39×10^{38}
Half	fp16	(11, 5)	4.88×10^{-4}	6.10×10^{-5}	6.55×10^{4}
Single	fp32	(24, 8)	5.96×10^{-8}	1.18×10^{-38}	3.40×10^{38}
Double	fp64	(53, 11)	1.11×10^{-16}	2.22×10^{-308}	1.80×10^{308}
Double extended (Intel)		(64, 16)	5.32×10^{-20}	3.36×10^{-4932}	1.19×10^{4932}
Quadruple	fp128	(113, 15)	9.63×10^{-35}	3.36×10^{-4932}	1.19×10^{4932}

TABLE 2
Parameters for floating-point arithmetics: number of bits in significand (including implicit most significant bit) and exponent (sig, exp), unit roundoff u, smallest normalized positive number $x_{\min}$, and largest finite number $x_{\max}$. The last three columns are given to three significant figures. The arithmetics whose names begin "fp" are from the IEEE standard [26].

Modern hardware increasingly supports half-precision arithmetic, which is attractive because of its speed, lower energy usage, and reduced storage and data movement costs. The two currently available half-precision formats are bfloat16 [27] and IEEE half precision; see Table 2. The optimistic bound (1.3) provides useful information only if $3nu < 1$, but $3nu > 1$ for problems of order $n \geq 684$ in IEEE half precision and $n \geq 86$ in bfloat16. Yet machine learning codes routinely use half precision in inner products and matrix–vector products with $n \gg 682$ with apparent success [19, 38]. Moreover, the machine topping the HPL-AI mixed-precision benchmark [15] in the June 2021 TOP500 list solved a linear system of order 1.6×10^7 using IEEE half-precision arithmetic for most of the computations, and the result was good enough to pass the benchmark's test that the residual is of order the unit roundoff for double precision.

How can this apparent mismatch between theory and practice be explained, and what are the implications for the future as the size of the largest problems continues to increase and the use of low precision arithmetic becomes more common? Have we reached the point where our techniques for analyzing rounding errors, honed over 70 years of digital computation, are unable to predict the accuracy of numerical linear algebra computations that are now routine? I will show that we can, in fact, understand to a considerable extent the behavior of extreme-scale and low accuracy computations. To do so, we need to take account of a number

of algorithmic design techniques and architectural features of processors that help reduce error growth, and we need to exploit a new probabilistic approach to rounding error analysis.

The main purpose of backward error analysis results such as Theorem 1.1 is to show the form of the backward error bound and to reveal the circumstances (if any) in which the backward error could be large. As Wilkinson [37], Parlett [31], and the present author [22, SECT. 3.2] have noted, the constants in a backward error bound are the least important part of it. Wilkinson recommends that if sharp error estimates are required they should be computed a posteriori [36, SECT. 12], [37]. This is indeed good advice, but it nevertheless remains valid to ask what a priori bounds can tell us—about the limits of what can be computed and about whether a successful computation can be guaranteed for a mission-critical application or one that takes up substantial computational resources. Furthermore, this question is also relevant for future benchmarking: will the HPL benchmark [14, 32] (used in the TOP500) or the HPL-AI mixed precision benchmark need modifying in the future because their criteria for successful completion can no longer be satisfied?

We assume that the floating-point arithmetic in use satisfies the standard model [22, SECT. 2.2]

$$\mathrm{fl}(x \mathrm{\ op\ } y) = (x \mathrm{\ op\ } y)(1+\delta), \quad |\delta| \le u, \quad \mathrm{op} = +, -, *, /. \tag{1.4}$$

This model is certainly satisfied by IEEE arithmetic, which defines $\mathrm{fl}(x \mathrm{\ op\ } y)$ to be the rounded exact value. In general, we denote by $\mathrm{fl}(expr)$ the value of the expression *expr* when it is evaluated in floating-point arithmetic.

We begin, in Section 2, by showing how the use of blocked inner products and blocked matrix factorizations reduces constants in rounding error bounds by a factor approximately equal to the block size. In Section 3 we explain how extended precision registers on Intel x86 processors and fused multiply–add operations and their mixed precision block generalizations yield reductions in the error constants. In Section 4 we explain how probabilistic rounding error analysis gives rounding error bounds with constants that are the square roots of the constants in the worst-case bounds. Some other relevant considerations are discussed in Section 5. We offer our conclusions in Section 6.

2. BLOCKED ALGORITHMS

Blocked algorithms,[2] which are primarily designed to give better performance on modern computers with hierarchical memories, also lead to improved rounding error bounds, as we now explain.

2 A blocked algorithm organizes a computation so that it works on separate chunks of data. It is also commonly called a "block algorithm", but the use of "block" is best reserved for properties, factorizations, and algorithms in which scalars are generalized into blocks. For example, a block tridiagonal matrix is not, in general, tridiagonal, and a block LU factorization is different from an LU factorization because it has a block upper triangular U.

2.1. Blocked inner products

Let $x, y \in \mathbb{R}^n$ and consider the inner product $s = x^T y$. If we evaluate s in the natural way as

$$s = x_1 y_1, \quad s \leftarrow s + x_k y_k, \quad k = 2 : n, \tag{2.1}$$

then the computed result $\hat{s}$ satisfies [22, SECT. 3.1]

$$|s - \hat{s}| \le \gamma_n |x|^T |y|. \tag{2.2}$$

In fact, this bound holds no matter what order the terms are summed in. Another way to compute the inner product is by summing two half-length inner products, where we assume $n = 2b$ for simplicity:

$$\begin{aligned} s_1 &= x(1 : b)^T y(1 : b), \\ s_2 &= x(b + 1 : n)^T y(b + 1 : n), \\ s &= s_1 + s_2. \end{aligned}$$

For this formulation the error bound is

$$|s - \hat{s}| \le \gamma_{n/2+1} |x|^T |y|,$$

so the error constant has been reduced by a factor 2. We can generalize this idea. Assuming that[3] $n = kb$, we can compute

$$\begin{aligned} s_i &= x\big((i-1)b + 1 : ib\big)^T y\big((i-1)b + 1 : ib\big), \quad i = 1 : k, \\ s &= s_1 + s_2 + \cdots + s_k, \end{aligned} \tag{2.3}$$

and the error bound is [22, SECT. 3.1]

$$|s - \hat{s}| \le \gamma_{b+n/b-1} |x|^T |y|. \tag{2.4}$$

As long as $b \ll n$, the error constant has been reduced by about a factor b. The reason for the reduction is that whereas for the standard evaluation (2.1) elements of x and y take part in up to $n - 1$ additions, for (2.3) they take part in at most $b + n/b - 2$ additions. The value of b that minimizes the bound (2.4) is $b = \sqrt{n}$, so if we take for b the nearest integer to $\sqrt{n}$ we will have

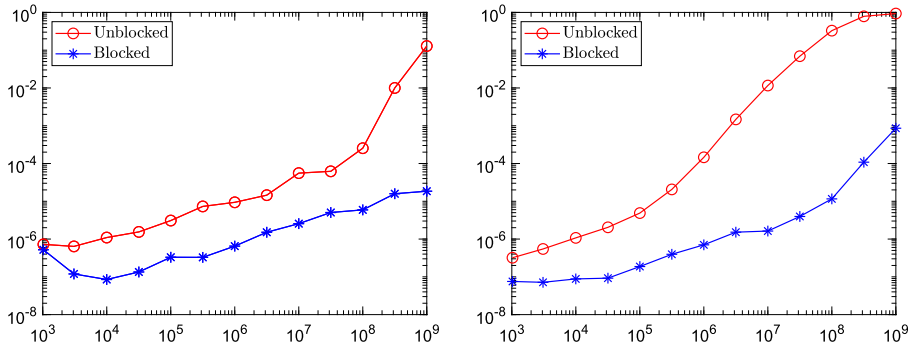
$$|s - \hat{s}| \lesssim \gamma_{2\sqrt{n}} |x|^T |y|.$$

By splitting the inner product into pieces, computing the partial inner products, and summing the results, we have reduced the error constant from n to $2\sqrt{n}$, which is a substantial reduction for large n.

Blocking of inner products is common in practice, though a fixed block size rather than one depending on n is normally taken [8]. It may be done in a low-level kernel for performance considerations and so may be invisible to a user.

---

3 This is not a practical restriction, as for general n we can compute the inner product of the last n mod b elements separately or pad the vectors with zeros so that their dimension is a multiple of b.



(a) Elements of x and y generated as $-0.25 + \text{randn}$, where randn is drawn from the normal $(0, 1)$ distribution.

(b) Elements of x and y generated from the uniform distribution on $(0, 1)$.

FIGURE 1
Relative errors in inner products computed in single precision with no blocking and with block size 256.

To illustrate the benefits of blocking we show in Figure 1 the relative errors in inner products computed with and without blocking for two types of random vector. The dimension n ranges from 10^3 to 10^9, the block size is 256, and the relative errors are averaged over 10 pairs of vectors x and y for each n. The reason for shifting the normally distributed random vectors is to make the mean nonzero, as for a zero mean the errors tends to be much smaller [24]. The more rapid growth of the errors for the unblocked computation that begins around $n = 10^7$ for both distributions is due to stagnation (described in Section 4.3).

The blocking approach just described can be improved by using a combination of two different methods. We will illustrate the idea for summation, but it trivially generalizes to inner products.

Assume that we have at our disposal two summation algorithms: a fast one, referred to as the FastSum algorithm, and an accurate one, referred to as the AccurateSum algorithm. Algorithm 2.1 uses these two algorithms to compute $\sum_{i=1}^{n} z_i$ by an algorithm of Blanchard, Higham, and Mary [6]. The algorithm is called FABsum, which stands for "fast and accurate blocked summation." To compute the inner product $x^T y$, we can take $z_i = x_i y_i$. We assume that n is a multiple of b.

Algorithm 2.1 (FABsum) This algorithm takes as input n summands z_i, a block size b that divides n, and two summation algorithms FastSum and AccurateSum. It returns the sum $s = \sum_{i=1}^{n} z_i$.

1: for $i = 1 : n/b$
2: Compute $s_i = \sum_{j=(i-1)b+1}^{ib} z_j$ with FastSum.
3: end
4: Compute $s = \sum_{i=1}^{n/b} s_i$ with AccurateSum.

Note that for $b = 1$, FABsum reduces to AccurateSum, and for $b = n$ it reduces to FastSum. The motivation for FABsum is that if b is chosen large enough, most of the work is done by FastSum but the use of AccurateSum can lead to improved accuracy.

Assume that for a sum $s = \sum_{i=1}^n z_i$ the computed $\hat{s}$ from FastSum satisfies

$$\hat{s} = \sum_{i=1}^n z_i(1 + \mu_i^f), \quad |\mu_i^f| \le \varepsilon_f(n), \tag{2.5}$$

and the computed $\hat{s}$ from AccurateSum satisfies

$$\hat{s} = \sum_{i=1}^n z_i(1 + \mu_i^a), \quad |\mu_i^a| \le \varepsilon_a(n), \tag{2.6}$$

where $\varepsilon_f(n)$ and $\varepsilon_a(n)$ are $O(u)$ and depend on n and u. With these assumptions on the backward errors of the underlying summation algorithms, we have the following backward error result [6, THM. 3.1].

Theorem 2.2. *Let $s = \sum_{i=1}^n z_i$ be computed by Algorithm 2.1. The computed $\hat{s}$ satisfies*

$$\hat{s} = \sum_{i=1}^n z_i(1 + \mu_i), \quad |\mu_i| \le \varepsilon(n,b) = \varepsilon_f(b) + \varepsilon_a(n/b) + \varepsilon_f(b)\varepsilon_a(n/b).$$

To see the gains in accuracy Algorithm 2.1 can bring, consider the following two choices. For FastSum take recursive summation, which is the usual algorithm that computes $s = z_1 + z_2$, $s \leftarrow s + z_k$, $k = 3 : n$. Then $\varepsilon_f(b) = (b-1)u + O(u^2)$. If AccurateSum is recursive summation at twice the working precision then $\varepsilon_a(n/b) = u + O(u^2)$, and so $\varepsilon(n,b) = bu + O(u^2)$ is independent of n to first order. If AccurateSum is the method known as compensated summation [22, SECT. 4.3], which works entirely in the working precision and for which $\varepsilon_a(n/b) = 2u + O(u^2)$, then $\varepsilon(n,b) = (b+1)u + O(u^2)$, which again does not grow with n to first order. Analysis of the second-order terms in [6, SECT. 3.1.2] shows that they are not significant unless n is extremely large.

Denote by $C(n,b)$ the cost in flops of Algorithm 2.1. If $C_f(n)$ and $C_a(n)$ are the costs for summing n terms by FastSum and AccurateSum, respectively, then

$$C(n,b) = \frac{n}{b}C_f(b) + C_a\left(\frac{n}{b}\right).$$

In particular, if the costs C_f and C_a are linear functions of the number of summands, as is usually the case, $C(n,b)$ simplifies to

$$C(n,b) = C_f(n) + \frac{1}{b}C_a(n) + O\left(\frac{n}{b}\right).$$

Therefore the cost of Algorithm 2.1 can be made close to that of FastSum by taking the block size b sufficiently large. The parameter b can be tuned to achieve the highest possible performance on a given target architecture, while keeping it independent of n to avoid error growth.

We have seen that by using a blocked implementation of summation or an inner product it is possible to reduce the error constant $nu + O(u^2)$ by a constant factor, or even to reduce it to $(b+1)u + O(u^2)$ by using FABsum, while at the same time increasing the performance. The increased performance and reduced error bound go hand in hand.

2.2. Blocked matrix multiplication

The standard error bound for a matrix product $C = AB$, where $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times t}$, is

$$|C - \hat{C}| \le \gamma_n |A||B|,$$

which is an immediate consequence of (2.2), and this bound holds for any order of evaluation. Consider Algorithm 2.3, which is a blocked implementation of matrix multiplication that amounts to computing each element of the product by the blocked inner product (2.3).

Algorithm 2.3 (Blocked matrix multiplication) Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times t}$ be partitioned into $b \times b$ blocks A_{ij} and B_{ij}, where $p = m/b_1$, $q = n/b$, and $r = t/b_2$ are assumed to be integers. This algorithm computes $C = AB$.

```
1: for i = 1 : p
2:     for j = 1 : r
3:         C_ij = 0
4:         for k = 1 : q
5:             X = A_ik B_kj
6:             C_ij = C_ij + X
7:         end
8:     end
9: end
```

We have written lines 5 and 6 as shown in order to make clear that $A_{ik}B_{kj}$ is computed and then added to C_{ij}. The expression "$C_{ij} = C_{ij} + A_{ik}B_{kj}$" would be ambiguous because for an individual element of C_{ij} it has the form

$$c_{i_1,j_1} = c_{i_1,j_1} + \sum_{\ell=1}^{b} a_{i_1,\ell} b_{\ell,j_1}, \tag{2.7}$$

and the order in which the b additions are done is not specified. If the additions are done from left to right then the algorithm is numerically equivalent to standard matrix multiplication. However, in Algorithm 2.3 the addition involving c_{i_1,j_1} is done last.

A rounding error result for Algorithm 2.3 follows readily from that for a blocked inner product: the computed $\hat{C}$ satisfies

$$|C - \hat{C}| \le \gamma_{b+n/b-1} |A||B|. \tag{2.8}$$

Again, if $b \ll n$ then the error constant has been reduced by about a factor b. In a highly optimized matrix multiplication algorithm, there may be multiple levels of blocking [18], which give a further reduction in the error bound.

We note that the FABsum algorithm (Algorithm 2.1) and its rounding error analysis trivially extend to matrix multiplication [6, SECT. 4].

2.3. Blocked matrix factorizations

The LAPACK library [3] pioneered the use of blocked algorithms that compute a matrix factorization a block at a time, where each block is square or rectangular with b columns, with the block size typically $b = 128$ or $b = 256$. These algorithms typically contain operations of the form $A_{ij} = A_{ij} - X_i Y_j$, and these are implemented as calls to a level 3 BLAS gemm (general matrix multiply) routine [13], which computes $C \leftarrow \alpha AB + \beta C$ for arbitrary matrices A, B, and C of conformable dimensions. In view of the error bound (2.8) for Algorithm 2.3, the blocked algorithm will have a constant in a (componentwise) backward error bound that is about b times smaller than for the unblocked algorithm provided that the gemm computes αAB before adding the result to βC. And, of course, for block-level computations that are inner product-based, the blockings of the previous subsections can be applied with a smaller block size, giving a further reduction in error bound.

Most references do not take advantage of blocking when stating error bounds. The LAPACK manual [3] states error bounds of the form $p(n)u$ for $n \times n$ matrices, where $p(n)$ is independent of the block size. Standard texts such as those of Demmel [12], Golub and Van Loan [17], and Higham [22] give error analysis only for unblocked algorithms, so do not derive the b-dependent constants for the blocked algorithms (though [22, SECT. 13.2] derives the constants for blocked LU factorization). We suggest three reasons why error analyses for blocked algorithms are usually not provided. First, as explained in Section 1, there has long been a feeling, going back to Wilkinson, that the most important part of a bound is not the constants but the form of the bound and that optimizing constants is not worthwhile. Second, the precise constants depend on which blocked algorithm variant of a factorization is chosen (there are usually several) and precisely how it is implemented. Third, the error analysis for a blocked algorithm tends to be more complicated than for the unblocked algorithm, which can obscure the main ideas of the analysis.

The important point to note is that with a suitable implementation the constant in a backward error bound for a blocked factorization with block size b will be reduced by a factor of order b or more.

3. ARCHITECTURAL FEATURES

A number of features of modern processors contribute to reducing the error in numerical computations.

3.1. Extended precision registers

Intel x86 processors support an 80-bit extended precision format with a 64-bit significand (see Table 2), which is compatible with that specified in the IEEE standard [26], [11, SECT. 4.2.2], [29, SECT. 3.4.3]. When a compiler uses this format with 80-bit registers to accumulate sums and inner products, it is effectively working with a unit roundoff of 2^{-64} rather than 2^{-53} for double precision, giving error bounds smaller by a factor up to $2^{11} = 2048$. We note, however, that extra precision registers can lead to strange rounding effects, in particular because of double rounding [22, SECT. 2.3, PROBS. 27.1, 27.3], [29].

3.2. Fused multiply–add

Another architectural feature that provides benefits to accuracy is a fused multiply–add (FMA) operation, which computes $x + yz$ with just one rounding error instead of two. Without an FMA,

$$\mathrm{fl}(x + yz) = \big(x + yz(1 + \delta_1)\big)(1 + \delta_2), \quad |\delta_1| \le u,\ |\delta_2| \le u,$$

whereas with an FMA,

$$\mathrm{fl}(x + yz) = (x + yz)(1 + \delta), \quad |\delta| \le u,$$

which means that the result is computed with a relative error bounded by u. The motivation for an FMA is speed, as it is implemented in such a way that it takes the same time as a single multiplication or addition. With the use of an FMA standard error bounds for inner product-based computations are reduced by a factor 2. It should be noted, though, that an FMA can lead to unexpected results when applied to certain expressions **[22, SECT. 2.6]**.

3.3. Mixed precision block fused multiply–add

A mixed precision block FMA takes as input matrices $A \in \mathbb{R}^{b_1 \times b}$, $B \in \mathbb{R}^{b \times b_2}$, and $C \in \mathbb{R}^{b_1 \times b_2}$, where A and B are provided in a given precision u_{low} and C is either in precision u_{low} or in a higher precision u_{high}, and computes

$$\underbrace{D}_{u_{\mathrm{low}} \text{ or } u_{\mathrm{high}}} = \underbrace{C}_{u_{\mathrm{low}} \text{ or } u_{\mathrm{high}}} + \underbrace{A}_{u_{\mathrm{low}}}\ \underbrace{B}_{u_{\mathrm{low}}}, \tag{3.1}$$

returning D in precision u_{low} or u_{high}. We will assume that the internal computations are at precision u_{high}. The output matrix D can be used as the input C to a subsequent FMA, so by chaining FMAs together in this way, larger matrix products can be computed **[5, ALG. 3.1]**. Table 3 gives the precisions and matrix dimensions for some block FMAs available in hardware. These block FMAs are designed to give one result per cycle and so can give significant performance benefits. For example, on the NVIDIA V100 GPU, whose tensor cores implement block FMAs, half-precision arithmetic on the tensor cores runs 8 times faster than single precision arithmetic, which in turn runs at twice the speed of double-precision arithmetic.

When C and D in (3.1) are taken at the higher precision, u_{high}, mixed precision block FMAs give an increase in accuracy compared with computations carried out at the lower precision, u_{low}. Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times t}$ be given in precision u_{high} and partitioned into $b_1 \times b$ blocks A_{ij} and $b \times b_2$ blocks B_{ij}, respectively, where $p = m/b_1$, $q = n/b$, and $r = t/b_2$ are assumed to be integers. When the product $C = AB$ is computed by a sequence of chained block FMAs using **[5, ALG. 3.1]** (which has the same general form as Algorithm 2.3), it can be shown **[5, THM. 3.2]** that the computed $\hat{C}$ satisfies

$$|C - \hat{C}| \le f(n, b, u_{\mathrm{low}}, u_{\mathrm{high}})|A||B|, \tag{3.2}$$

where the first-order part of f is given in Table 4. We see that for $n < 2u_{\mathrm{low}}/u_{\mathrm{high}} =: n_*$ the block FMA constant is independent of n. It is always smaller than the constant for standard multiplication at precision u_{low} and of similar magnitude to the constant for standard

Year of release	Device	Matrix dimensions	u_{low}	u_{high}
2016	Google TPU v2	$128 \times 128 \times 128$	bfloat16	fp32
2017	Google TPU v3	$128 \times 128 \times 128$	bfloat16	fp32
2017	NVIDIA V100	$4 \times 4 \times 4$	fp16	fp32
2018	NVIDIA T4	$4 \times 4 \times 4$	fp16	fp32
2019	ARMv8.6-A	$2 \times 4 \times 2$	bfloat16	fp32
2020	NVIDIA A100	$8 \times 8 \times 4$	bfloat16	fp32
		$8 \times 8 \times 4$	fp16	fp32
		$8 \times 4 \times 4$	TensorFloat-32	fp32
		$2 \times 4 \times 2$	fp64	fp64

TABLE 3

Processing units or architectures equipped with mixed-precision fused multiply–add accelerators. Matrix dimensions are expressed as $b_1 \times b \times b_2$, where b_1 is the number of rows in A, b is the number of columns in A and rows in B, and b_2 is the number of columns in B. The input and output precisions u_{low} and u_{high} are defined in (3.1). Sources [4, 9, 30].

Evaluation method	Bound
Standard in precision u_{low}	$(n+2)u_{\text{low}}$
Block FMA, u_{high} internally, output in u_{high}	$2u_{\text{low}} + nu_{\text{high}}$
Standard in precision u_{high}	nu_{high}

TABLE 4

First order part of constant term $f(n, b, u_{\text{low}}, u_{\text{high}})$ in error bound (3.2) for matrix multiplication with and without the use of a mixed precision block FMA.

multiplication at precision u_{high} for $n > n_*$. When u_{low} corresponds to fp16 or bfloat16 and u_{high} to fp32, we have $n_* = 16{,}384$ and $n_* = 131{,}072$, respectively. Hence while a mixed precision block FMA takes inputs at precision u_{low}, for large n it produces results as good as if the computation were done at precision u_{high}.

Note that (3.2) assumes that, in the notation of Algorithm 2.3, lines 5 and 6 are evaluated as $C_{ij} = C_{ij} + A_{ik}B_{kj}$, in left to right order; if the evaluation uses lines 5 and 6 as stated then the u_{high} term in Table 4 for the block FMA is further reduced. However, NVIDIA tensor cores in the Volta, Turing, and Ampere microarchitectures with $b = 4$ do not use a fixed order when evaluating each individual element $c_{ij} + a_{i1}b_{1j} + a_{i2}b_{2j} + a_{i3}b_{3j} + a_{i4}b_{4j}$ in (3.1), but rather evaluate the expression starting with the largest magnitude term [16].

4. PROBABILISTIC ROUNDING ERROR ANALYSIS

We have now seen two main reasons why standard rounding error bounds may be pessimistic: first, they do not account for block algorithms, and second, architectural features of the computer may provide increased accuracy for certain types of operations. We now discuss a third reason, which has to do with the very nature of rounding error bounds.

In the model (1.4), the relative error δ in $\mathrm{fl}(x \text{ op } y)$ is typically strictly less than u in magnitude, and, of course, it is zero if x op y happens to be a floating-point number. Rounding error analyses apply (1.4) repeatedly. Typically, a product of $1+\delta_i$ terms appears, which can be handled by the next lemma **[22, LEMMA 3.1]**.

Lemma 4.1. *If $|\delta_i| \le u$ and $\rho_i = \pm 1$ for $i = 1:n$, and $nu < 1$, then*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1+\theta_n, \quad |\theta_n| \le \gamma_n. \tag{4.1}$$

This lemma, combined with some useful identities satisfied by the γ_k and θ_k **[22, LEMMA 3.3]**, provides a convenient way to carry out rounding error analyses. However, the proof of the lemma involves multiple uses of the triangle inequality and so the bound $|\theta_n| \le \gamma_n$ can be expected to be potentially weak.

For a given algorithm and a given set of data, we would like to be able to say that there exists a set of rounding errors δ_i that, if they occur, produce an error of roughly the same size as the rounding error bound. This is usually not the case, but it can be true if in every invocation of (1.4) δ has the same sign. For basic kernels, it may be possible to show that the error bound is approximately attainable for a special choice of the data, as is the case for recursive summation **[22, PROB. 4.2]**, **[35, P. 19]**, but such examples do not indicate the quality of the bound in typical cases.

In an early paper on rounding error analysis, Wilkinson derives rounding error bounds for Gaussian elimination, Givens QR factorization, and Householder QR factorization and then states that **[34, P. 318]**

> *"The bounds we have obtained are in all cases strict upper bounds. In general, the statistical distribution of the rounding errors will reduce considerably the function of n occurring in the relative errors. We might expect in each case that this function should be replaced by something which is no bigger than its square root and is usually appreciably smaller."*

He makes similar statements in **[35]**. For many years, primarily because of Wilkinson's comments, it has been regarded as a rule of thumb that a worst-case rounding error bound $f(n)u$ is more realistic if it is replaced by $\sqrt{f(n)}u$. No proof has been given to make this rule of thumb rigorous, but one can argue as follows:

- linearize the error into a sum $e = \sum_{i=1}^{p} t_i\delta_i$, where the δ_i are rounding errors and the t_i depend on the data;
- assume that the δ_i are independent random variables of mean zero;
- apply the central limit theorem to deduce that the probability distribution of $e/(\sum_{i=1}^{n} t_i^2)^{1/2}$ tends towards a normal distribution of mean zero and standard deviation $\sigma \le u$;

- conclude that for sufficiently large n, the probability that $|e|$ will not exceed $u(\sum_{i=1}^n t_i^2)^{1/2}$ times a small constant is very high.

Compared with the worst-case constant $\sum_{i=1}^n |t_i|$, the quantity $(\sum_{i=1}^n t_i^2)^{1/2}$ can be smaller by a factor up to $\sqrt{n}$. This argument, however, has a number of weaknesses. First, it is essentially forward error-based, whereas we prefer to work with backward errors if possible. Second, the argument is based on the first-order part of the error, so says nothing about higher-order terms. Third, it is not clear how large n must be for the application of the central limit theorem to be valid.

Despite the weaknesses of a central limit theorem argument, a probabilistic approach seems to be necessary to obtain substantially better bounds than the worst-case ones. Indeed, as Stewart [33] has noted,

> *"To be realistic, we must prune away the unlikely. What is left is necessarily a probabilistic statement."*

We will discuss probabilistic rounding error analysis in the next two subsections.

4.1. Error analysis for nonrandom data

Higham and Mary [23] introduced a new probabilistic rounding error analysis, making use of a concentration inequality. This work was extended by Higham and Mary for random data [24], and by Ipsen and Zhou [28], and Connolly, Higham, and Mary [10], all of whom use martingales. We will present the most general results for nonrandom data, which are those from [10].

We need the following probabilistic version of Lemma 4.1 [10, LEMMA 4.6], which includes the constant

$$\widetilde{\gamma}_n(\lambda) = \exp\left(\frac{\lambda\sqrt{n}u + nu^2}{1-u}\right) - 1 = \lambda\sqrt{n}u + O(u^2). \tag{4.2}$$

We use $\mathbb{E}$ to denote the expectation of a random variable.

Theorem 4.2. *Let $\delta_1, \delta_2, \dots, \delta_n$ be random variables of mean zero with $|\delta_k| \le u$ for all k such that $\mathbb{E}(\delta_k \mid \delta_1, \dots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$ for $k = 2:n$. Then for $\rho_i = \pm 1$, $i = 1:n$ and any constant $\lambda > 0$,*

$$\prod_{i=1}^n (1+\delta_i)^{\rho_i} = 1 + \theta_n, \quad |\theta_n| \le \widetilde{\gamma}_n(\lambda) \tag{4.3}$$

holds with probability at least $1 - 2\exp(-\lambda^2/2)$.

The key difference between (4.1) and (4.3) is that, to first order, the bound in (4.3) is proportional to $\sqrt{n}u$ rather than nu.

Next, we need the following model of rounding errors.

Model 4.3 (Probabilistic model of rounding errors). *Let the computation of interest generate rounding errors $\delta_1, \delta_2, \dots$ in that order. The δ_k are random variables of mean zero such that $\mathbb{E}(\delta_k \mid \delta_1, \dots, \delta_{k-1}) = \mathbb{E}(\delta_k) = 0$.*

The model says that the rounding errors δ_i are mean independent and of mean zero, but they do not need to be from the same distribution. Mean independence is a weaker condition than independence: if the rounding errors are independent then they can be shown to be mean independent, but the converse implication does not hold. Under the model, Theorem 4.2 holds and allows us to bound rounding error terms that appear in analyses of inner product-based computations. This leads to the following three results [10, THMS. 4.8–4.10], in which

$$Q(\lambda, n) = 1 - 2n \exp(-\lambda^2/2).$$

Theorem 4.4 (Inner products). *Let $s = x^T y$, where $x, y \in \mathbb{R}^n$, be evaluated in floating-point arithmetic. Under Model 4.3, no matter what the order of evaluation, the computed $\hat{s}$ satisfies*

$$\hat{s} = (x + \Delta x)^T y = x^T (y + \Delta y), \quad |\Delta x| \le \widetilde{\gamma}_n(\lambda)|x|, \quad |\Delta y| \le \widetilde{\gamma}_n(\lambda)|y| \tag{4.4}$$

with probability at least $Q(\lambda, n)$.

Theorem 4.5 (Matrix products). *Let $C = AB$ with $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Under Model 4.3, the jth column of the computed $\hat{C}$ satisfies*

$$\hat{c}_j = (A + \Delta A_j) b_j, \quad |\Delta A_j| \le \widetilde{\gamma}_n(\lambda)|A|, \quad j = 1 : n, \tag{4.5}$$

with probability at least $Q(\lambda, mn)$, and hence

$$|C - \hat{C}| \le \widetilde{\gamma}_n(\lambda)|A||B| \tag{4.6}$$

with probability at least $Q(\lambda, mnp)$.

Theorem 4.6 (Linear system). *Let $A \in \mathbb{R}^{n\times n}$ and suppose that LU factorization and substitution produce computed factors $\hat{L}$ and $\hat{U}$ and a computed solution $\hat{x}$ to $Ax = b$. Then, under Model 4.3,*

$$(A + \Delta A)\hat{x} = b, \quad |\Delta A| \le \left(3\widetilde{\gamma}_n(\lambda) + \widetilde{\gamma}_n(\lambda)^2\right)|\hat{L}||\hat{U}| \tag{4.7}$$

holds with probability at least $Q(\lambda, n^3/3 + 3n^2/2 + 7n/6)$.

Matrix multiplication and LU factorization both have triply nested loops, which can be ordered in $3! = 6$ ways. Theorems 4.5 and 4.6 both hold no matter which ordering of the loops is taken.

Let us now focus our attention on Theorem 4.6. For $n = 10^8$, the function $Q(\lambda, n^3/3 + 3n^2/2 + 7n/6)$ approaches 1 rapidly as λ increases and is approximately $1 - 10^{-3}$ for $\lambda = 11$ and $1 - 10^{-8}$ for $\lambda = 12$. Moreover, as shown in [23, SECT. 3.5], $Q(\lambda, f(n))$ remains independent of n as long as λ increases proportionally to $\log n$. Experiments show that the probability $Q(\lambda, f(n))$ is actually very pessimistic and in practice the bounds usually hold with $\lambda = 1$.

Theorems 4.2–4.6 provide a rigorous proof for inner product-based computations of the rule of thumb stated by Wilkinson, under the assumptions of Model 4.3.

Probabilistic error analysis can also be applied to blocked algorithms, with the blocking and the probabilistic approach combining to reduce the error constant. For example, the error constant $(b + n/b - 1)u + O(u^2)$ in (2.4) for a blocked inner product translates to $(\sqrt{b} + \sqrt{n/b})u + O(u^2)$ in a probabilistic bound.

4.2. Error analysis for random data

Numerical experiments show that the bounds in Theorems 4.4–4.6 reflect the actual rate of growth of the error with n for some problems [23,24], but the bounds can, nevertheless, be pessimistic. Higham and Mary [24] investigate the case where the data is random. They use the following model for the data, which is denoted by d_j, $j = 1 : n$.

Model 4.7 (Probabilistic model of the data). *The d_j, $j = 1 : n$, are independent random variables sampled from a given distribution of mean μ_x and satisfy $|d_j| \le \xi_d$, $j = 1 : n$, where ξ_d is a constant.*

A modified version of Model 4.3 is needed.

Model 4.8 (Modified probabilistic model of rounding errors). *Let the computation of interest generate rounding errors $\delta_1, \delta_2, \ldots$ in that order. The δ_i are random variables of mean zero and, for all k, the δ_k are mean independent of the previous rounding errors and of the data, in the sense that*

$$\mathbb{E}(\delta_k \mid \delta_1, \ldots, \delta_{k-1}, d_1, \ldots, d_n) = \mathbb{E}(\delta_k) = 0. \tag{4.8}$$

Under these models, Higham and Mary [24] obtain error bounds for an inner product, a matrix–vector product, and a matrix product. We state the result for matrix products [24, THM. 3.4].

Theorem 4.9. *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ satisfy Model 4.7 with means μ_A, μ_B and bounds ξ_A, ξ_B, and let $C = AB$. Under Model 4.8, the computed $\hat{C}$ satisfies*

$$\max_{i,j} \left|(C - \hat{C})_{ij}\right| \le \left(\lambda|\mu_A\mu_B|n^{3/2} + (\lambda^2 + 1)\xi_A\xi_B n\right)u + O(u^2) \tag{4.9}$$

with probability at least $P(\lambda) = 1 - 2mnp\exp(-\lambda^2/2)$.

The rate of growth of the bound (4.9) is $n^{3/2}$ except when μ_A or μ_B is small or zero, in which case it is just n. Thus the error bound depends on the means of the data. Furthermore, it is shown in [24, THM. 3.3] that for an inner product $x^T y$ in which either x or y has zero mean the backward error is bounded by $c_1 u + O(u^2)$ instead of $c_2\sqrt{n}u + O(u^2)$ as in Theorem 4.4, where c_1 and c_2 are constants.

We note that extending this analysis with random data to the solution of linear systems by LU factorization is an open problem, as noted in [24, SECT. 5].

4.3. Limitations

It is important to realize that the assumptions of the probabilistic rounding error analysis may not hold: the rounding errors may be dependent or may have nonzero mean, and in these cases the error may grow as nu rather than $\sqrt{n}u$. Consider a sum $\sum_{i=1}^{n} x_i$

computed by recursive summation, where the x_i are positive and decrease with i. For a large enough i, the summand x_i may be so small that it does not change the current partial sum in floating-point arithmetic. From this point on, no summand changes the sum so the rounding errors are all negative and Model 4.3 does not hold, and in this circumstance the worst-case linear growth can be achieved, as can be shown by numerical examples [10, 23]. This problem is called stagnation. A cure for stagnation is to randomize the rounding using stochastic rounding [10], which ensures that the sum can increase. Indeed with stochastic rounding, Model 4.3 is always satisfied and so, by Theorem 4.4, the error in the sum grows as $\sqrt{n}u$ instead of nu with high probability.

5. OTHER CONSIDERATIONS

5.1. Sharpness of error bounds

The bound (1.1) of Theorem 1.1 is not the best we can obtain. In the proof of the bound in [22], it is first shown that $A + \Delta A_1 = \hat{L}\hat{U}$, where

$$|\Delta A_1| \le \begin{bmatrix} \gamma_1 & \gamma_1 & \cdots & \cdots & \gamma_1 \\ \gamma_1 & \gamma_2 & \cdots & \cdots & \gamma_2 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \gamma_{n-1} & \gamma_{n-1} \\ \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} & \gamma_n \end{bmatrix} \circ |\hat{L}||\hat{U}| \equiv H \circ |\hat{U}|, \tag{5.1}$$

where $\circ$ is the Hadamard product, $A \circ B = (a_{ij}b_{ij})$. The bound (1.1) corresponds to replacing every element of H by γ_n. Analogous replacements are made in the part of the analysis dealing with the solution of the triangular systems by substitution. Clearly, then, not all n^2 inequalities in (1.1) are sharp. The same is true of (4.7), as its proof is analogous to that of (1.1). However, (5.1) still contains a term γ_n and so this sharper bound does not bring any significant benefits.

5.2. Growth factor at low precisions

Wilkinson [34] showed that with partial pivoting the growth factor ρ_n for LU factorization is bounded by 2^{n-1}, and he noted that ρ_n is nevertheless usually small in practice. Many years of experience have confirmed that ρ_n is indeed usually less than 50 (say) in practice. The growth factor directly affects the backward error bounds, through the size of the elements of $\hat{U}$ and (1.2). When we are working in half precision, with unit roundoff $u \approx 5 \times 10^{-4}$ for fp16 or $u \approx 4 \times 10^{-3}$ for bfloat16, element growth can have a much bigger relative effect on the quality of a solution than for single precision or double precision. Matrices that give large growth factors for partial pivoting are known, and a class of random matrices of arbitrary condition number that typically have $\rho_n \ge n/(4 \log n)$ was recently identified by Higham, Higham, and Pranesh [21]. For the latter class of matrices, growth alone can cause a complete loss of numerical stability for $n \ge 10^5$ in fp16—and, to complicate matters, it can also cause overflow in fp16 [25].

5.3. Iterative refinement

If we solve $Ax = b$ in half-precision arithmetic then, of course, we cannot expect a backward error smaller than the unit roundoff u_h for half precision. However, we can obtain a numerically stable solution at higher precision by using the computed half-precision solution as a first approximation that we improve by iterative refinement at the higher precision, using the half-precision LU factors. This procedure is guaranteed to work only for condition numbers $\kappa(A) = \|A\|\|A^{-1}\|$ up to u_h^{-1}. GMRES-based iterative refinement solves the update equation by GMRES preconditioned by the LU factors and can tolerate much more ill-conditioned A. See [2, 7, 20] for details of GMRES-based iterative refinement. Although this mixed precision algorithm uses higher precision to raise the quality of the initial solution, the conditions for success rest on the rounding error bounds for the factorization, and so the considerations of this paper contribute to our understanding of the algorithm.

6. CONCLUSIONS

We have seen that several factors combine to make errors in inner-product based computations much smaller than worst-case rounding error bounds suggest. Block algorithms can reduce error bounds by a factor of the block size b, and if blocking is used at multiple levels then the reduction factors can accumulate. Extended precision registers and (block) FMAs can give automatic accuracy boosts. With a block size $b = 256$ and the 80-bit registers on Intel x86-64 processors a reduction in an error bound by a constant factor $256 \times 2048 = 5.2 \times 10^5$ is possible for large problems.

The rate of growth of the error can be much smaller than the worst-case bounds because of statistical effects. If the rounding errors are mean independent and of mean zero then, as explained in Section 4.1, for inner products, matrix–vector products, matrix products, and the solution of linear systems by LU factorization, the constant $\gamma_n = nu + O(u^2)$ in a worst-case componentwise backward error bound can be replaced by $\widetilde{\gamma}_n = \sqrt{n}u + O(u^2)$ to obtain a bound that holds with high probability. Even these bounds can be pessimistic because, as explained in Section 4.7, when the data is random with zero mean, the error bound reduces further—to a constant independent of n for an inner product.

Together, these aspects go a considerable way to explaining why linear systems, and other linear algebra problems, are able to be successfully solved with ever growing dimensions and with the use of low precision arithmetics (perhaps within a mixed precision algorithm [1]).

It is pleasing to note that blocked algorithms and (block) FMAs, which were introduced to boost performance, also yield smaller rounding error bounds. It will be important to analyze future developments in algorithms and computer architectures to understand their effects on rounding error analysis.

ACKNOWLEDGMENTS

I thank Jack Dongarra, Massimilano Fasi, Sven Hammarling, Theo Mary, and Mantas Mikaitis for their comments on a draft of this paper.

FUNDING

This work was supported by Engineering and Physical Sciences Research Council grant EP/P020720/1, the Royal Society, and the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U. S. Department of Energy Office of Science and the National Nuclear Security Administration.

REFERENCES

[1] A. Abdelfattah, H. Anzt, E. G. Boman, E. Carson, T. Cojean, J. Dongarra, A. Fox, M. Gates, N. J. Higham, X. S. Li, J. Loe, P. Luszczek, S. Pranesh, S. Rajamanickam, T. Ribizel, B. F. Smith, K. Swirydowicz, S. Thomas, S. Tomov, Y. M. Tsai, and U. M. Yang, A survey of numerical linear algebra methods utilizing mixed-precision arithmetic. *Int. J. High Perform. Comput. Appl.* **35** (2021), no. 4, 344–369.

[2] P. Amestoy, A. Buttari, N. J. Higham, J.-Y. L'Excellent, T. Mary, and B. Vieublé, Five-precision GMRES-based iterative refinement. MIMS EPrint 2021.5, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2021, http://eprints.maths.manchester.ac.uk/id/eprint/2807.

[3] E. Anderson, Z. Bai, C. H. Bischof, S. Blackford, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, and D. C. Sorensen, *LAPACK users' guide. Third edn.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[4] Arm architecture reference manual. Armv8, for Armv8-A architecture profile. ARM DDI 0487F.b (ID040120), ARM Limited, Cambridge, UK, 2020.

[5] P. Blanchard, N. J. Higham, F. Lopez, T. Mary, and S. Pranesh, Mixed precision block fused multiply–add: error analysis and application to GPU tensor cores. *SIAM J. Sci. Comput.* **42** (2020), no. 3, C124–C141.

[6] P. Blanchard, N. J. Higham, and T. Mary, A class of fast and accurate summation algorithms. *SIAM J. Sci. Comput.* **42** (2020), no. 3, A1541–A1557.

[7] E. Carson and N. J. Higham, Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM J. Sci. Comput.* **40** (2018), no. 2, A817–A847.

[8] A. M. Castaldo, R. C. Whaley, and A. T. Chronopoulos, Reducing floating point error in dot product using the superblock family of algorithms. *SIAM J. Sci. Comput.* **31** (2008), 1156–1174.

[9] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, NVIDIA A100 tensor core GPU: performance and innovation. *IEEE Micro* **41** (2021), no. 2, 29–35.

[10] M. P. Connolly, N. J. Higham, and T. Mary, Stochastic rounding and its probabilistic backward error analysis. *SIAM J. Sci. Comput.* **43** (2021), no. 1, A566–A585.

[11] Intel Corporation, Intel 64 and IA-32 architectures software developer's manual. Volume 1: basic architecture, 2021, https://software.intel.com/content/www/us/en/develop/download/intel-64-and-ia-32-architectures-software-developers-manual-volume-1-basic-architecture.html.

[12] J. W. Demmel, *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[13] J. J. Dongarra, J. Du Croz, S. Hammarling, and I. Duff, A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Software* **16** (1990), no. 1, 1–17.

[14] J. J. Dongarra, P. Luszczek, and A. Petitet, The LINPACK benchmark: past, present and future. *Concurr. Comput. Pract. Exp.* **15** (2003), 803–820.

[15] J. J. Dongarra, P. Luszczek, and Y. M. Tsai, HPL-AI mixed-precision benchmark, https://icl.bitbucket.io/hpl-ai/.

[16] M. Fasi, N. J. Higham, M. Mikaitis, and S. Pranesh, Numerical behavior of NVIDIA tensor cores. *PeerJ Comput. Sci.* **7** (2021), e330(1–19).

[17] G. H. Golub and C. F. Van Loan, *Matrix computations. Fourth edn*. Johns Hopkins University Press, Baltimore, MD, USA, 2013.

[18] K. Goto and R. A. van de Geijn, Anatomy of high-performance matrix multiplication. *ACM Trans. Math. Software* **34** (2008), no. 3, 1–25.

[19] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, Deep learning with limited numerical precision. In *Proceedings of the 32nd international conference on machine learning*, pp. 1737–1746 J. Mach. Learn. Res. Workshop Conf. Proc. 37, 2015.

[20] A. Haidar, H. Bayraktar, S. Tomov, J. Dongarra, and N. J. Higham, Mixed-precision iterative refinement using tensor cores on GPUs to accelerate solution of linear systems. *Proc. R. Soc. Lond. A* **476** (2020), no. 2243, 20200110.

[21] D. J. Higham, N. J. Higham, and S. Pranesh, Random matrices generating large growth in LU factorization with pivoting. *SIAM J. Matrix Anal. Appl.* **42** (2021), no. 1, 185–201.

[22] N. J. Higham, *Accuracy and stability of numerical algorithms. Second edn.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

[23] N. J. Higham and T. Mary, A new approach to probabilistic rounding error analysis. *SIAM J. Sci. Comput.* **41** (2019), no. 5, A2815–A2835.

[24] N. J. Higham and T. Mary, Sharper probabilistic backward error analysis for basic linear algebra kernels with random data. *SIAM J. Sci. Comput.* **42** (2020), no. 5, A3427–A3446.

[25] N. J. Higham, S. Pranesh, and M. Zounon, Squeezing a matrix into half precision, with an application to solving linear systems. *SIAM J. Sci. Comput.* **41** (2019), no. 4, A2536–A2551.

[26] IEEE standard for floating-point arithmetic, IEEE Std 754-2019 (revision of IEEE 754-2008). The Institute of Electrical and Electronics Engineers, New York, USA, 2019.

[27] Intel Corporation, BFLOAT16—Hardware Numerics Definition, 2018, https://software.intel.com/en-us/download/bfloat16-hardware-numerics-definition, white paper. Document number 338302-001US.

[28] I. C. F. Ipsen and H. Zhou, Probabilistic error analysis for inner products. *SIAM J. Matrix Anal. Appl.* **41** (2020), no. 4, 1726–1741.

[29] J.-M. Muller, N. Brunie, F. de Dinechin, C.-P. Jeannerod, M. Joldes, V. Lefèvre, G. Melquiond, N. Revol, and S. Torres, *Handbook of floating-point arithmetic. Second edn*. Birkhäuser, Boston, MA, USA, 2018.

[30] T. Norrie, N. Patil, D. H. Yoon, G. Kurian, S. Li, J. Laudon, C. Young, N. Jouppi, and D. Patterson, The design process for Google's training chips: TPUv2 and TPUv3. *IEEE Micro* **41** (2021), no. 2, 56–63.

[31] B. N. Parlett, The contribution of J. H. Wilkinson to numerical analysis. In *A history of scientific computing*, edited by S. G. Nash, pp. 17–30, Addison-Wesley, Reading, MA, USA, 1990.

[32] A. Petitet, R. C. Whaley, J. Dongarra, and A. Cleary, HPL: a portable implementation of the High-Performance Linpack Benchmark for distributed-memory computers, Version 2.3, 2018, https://www.netlib.org/benchmark/hpl/.

[33] G. W. Stewart, Stochastic perturbation theory. *SIAM Rev.* **32** (1990), no. 4, 579–610.

[34] J. H. Wilkinson, Error analysis of direct methods of matrix inversion. *J. ACM* **8** (1961), 281–330.

[35] J. H. Wilkinson, *Rounding errors in algebraic processes*. Notes Appl. Sci. 32, Her Majesty's Stationery Office, London, 1963.

[36] J. H. Wilkinson, Modern error analysis. *SIAM Rev.* **13** (1971), no. 4, 548–568.

[37] J. H. Wilkinson, Numerical linear algebra on digital computers. *IMA Bull.* **10** (1974), no. 2, 354–356.

[38] P. Zamirai, J. Zhang, C. R. Aberger, and C. De Sa, Revisiting bfloat16 training. 2021, arXiv:2010.06192.

NICHOLAS J. HIGHAM

Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK, nick.higham@manchester.ac.uk

THE MATHEMATICS OF ARTIFICIAL INTELLIGENCE

GITTA KUTYNIOK

ABSTRACT

We currently witness the spectacular success of artificial intelligence in both science and public life. However, the development of a rigorous mathematical foundation is still at an early stage. In this survey article, which is based on an invited lecture at the International Congress of Mathematicians 2022, we will in particular focus on the current "workhorse" of artificial intelligence, namely deep neural networks. We will present the main theoretical directions along with several exemplary results and discuss key open problems.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68T07; Secondary 41A25, 42C15, 35C20, 65D18

KEYWORDS

Applied harmonic analysis, approximation theory, deep learning, inverse problems, partial different equations

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5118–5139
DOI 10.4171/ICM2022/141

1. INTRODUCTION

Artificial intelligence is currently leading to one breakthrough after another, both in public with, for instance, autonomous driving and speech recognition, and in the sciences in areas such as medical diagnostics or molecular dynamics. In addition, the research on artificial intelligence and, in particular, on its theoretical foundations is progressing at an unprecedented rate. One can envision that the corresponding methodologies will in the future drastically change the way we live in numerous respects.

1.1. The rise of artificial intelligence

Artificial intelligence is, however, not a new phenomenon. In fact, already in 1943, McCulloch and Pitts started to develop algorithmic approaches to learning by mimicking the functionality of the human brain, through artificial neurons which are connected and arranged in several layers to form artificial neural networks. Already at that time, they had a vision for the implementation of artificial intelligence. However, the community did not fully recognize the potential of neural networks. Therefore, this first wave of artificial intelligence was not successful and vanished. Around 1980, machine learning became popular again, and several highlights can be reported from that period.

The real breakthrough and with it a new wave of artificial intelligence came around 2010 with the extensive application of *deep* neural networks. Today, this model might be considered the "workhorse" of artificial intelligence, and in this article we will focus predominantly on this approach. The structure of deep neural networks is precisely the structure McCulloch and Pitts introduced, namely numerous consecutive layers of artificial neurons. Today two main obstacles from previous years have also been eliminated; due to the drastic improvement of computing power, the training of neural networks with hundreds of layers in the sense of *deep* neural networks is feasible, and we are living in the age of data, hence vast amounts of training data are easily available.

1.2. Impact on mathematics

The rise of artificial intelligence also had a significant impact on various fields of mathematics. Maybe the first area which embraced these novel methods was the area of inverse problems, in particular, imaging science, where such approaches have been used to solve highly ill-posed problems such as denoising, inpainting, superresolution, or (limited-angle) computed tomography, to name a few. One might note that, due to the lack of a precise mathematical model of what an image is, this area is particularly suitable for learning methods. Thus, after a few years, a change of paradigm could be observed, and novel solvers are typically at least to some extent based on methods from artificial intelligence. We will discuss further details in Section 4.1.

The area of partial differential equations was much slower to embrace these new techniques, the reason being that it was not per se evident what the advantage of methods from artificial intelligence for this field would be. Indeed, there seems to be no need to utilize learning-type methods, since a partial differential equation is a rigorous mathematical

model. But, lately, the observation that deep neural networks are able to beat the curse of dimensionality in high-dimensional settings led to a change of paradigm in this area as well. Research at the intersection of numerical analysis of partial differential equations and artificial intelligence therefore accelerated since about 2017. We will delve further into this topic in Section 4.2.

1.3. Problems of artificial intelligence

However, as promising as all these developments seem to be, a word of caution is required. Besides the fact that the practical limitations of methods such as deep neural networks have not been explored at all and at present neural networks are still considered a "jack-of-all-trades," it is even more worrisome that a comprehensive theoretical foundation is completely lacking. This was very prominently stated during the major conference on artificial intelligence and machine learning, which is NIPS (today called NeurIPS) in 2017, when Ali Rahimi from Google received the Test of Time Award and during his plenary talk stated that "Machine learning has become a form of alchemy." This raised a heated discussion to which extent a theoretical foundation does exist and is necessary at all. From a mathematical viewpoint, it is crystal clear that a fundamental mathematical understanding of artificial intelligence is inevitably necessary, and one has to admit that its development is currently in a preliminary state at best.

This lack of mathematical foundations, for instance, in the case of deep neural networks, results in a time-consuming search for a suitable network architecture, a highly delicate trial-and-error-based (training) process, and missing error bounds for the performance of the trained neural network. One needs to stress that, in addition, such approaches also sometimes unexpectedly fail dramatically when a small perturbation of the input data causes a drastic change of the output leading to radically different—and often wrong—decisions. Such adversarial examples are a well-known problem, which becomes severe in sensitive applications such as when minor alterations of traffic signs, e.g., the placement of stickers, cause autonomous vehicles to suddenly reach an entirely wrong decision. It is evident that such robustness problems can only be tackled by a profound mathematical approach.

1.4. A need for mathematics

These considerations show that there is a tremendous need for mathematics in the area of artificial intelligence. And, in fact, one can currently witness that numerous mathematicians move to this field, bringing in their own expertise. Indeed, as we will discuss in Section 2.4, basically all areas of mathematics are required to tackle the various difficult, but exciting challenges in the area of artificial intelligence.

One can identify two different research directions at the intersection of mathematics and artificial intelligence:

- *Mathematical Foundations for Artificial Intelligence*. This direction aims for deriving a deep mathematical understanding. Based on this, it strives to over-

come current obstacles such as the lack of robustness or places the entire training process on a solid theoretical foundation.

- *Artificial Intelligence for Mathematical Problems*. This direction focuses on mathematical problem settings such as inverse problems and partial differential equations with the goal of employing methodologies from artificial intelligence to develop superior solvers.

1.5. Outline

Both research directions will be discussed in this survey paper, showcasing some novel results and pointing out key future challenges for mathematics. We start with an introduction into the mathematical setting, stating the main definitions and notations (see Section 2). Next, in Section 3, we delve into the first main direction, namely mathematical foundations for artificial intelligence, and discuss the research threads of expressivity, optimization, generalization, and explainability. Section 4 is then devoted to the second main direction, which is artificial intelligence for mathematical problems, and we highlight some exemplary results. Finally, Section 5 states the seven main mathematical problems and concludes this article.

2. THE MATHEMATICAL SETTING OF ARTIFICIAL INTELLIGENCE

We now get into more details on the precise definition of a deep neural network, which is after all a purely mathematical object. We will also touch upon the typical application setting and training process, as well as on the current key mathematical directions.

2.1. Definition of deep neural networks

The core building blocks are, as said, artificial neurons. For their definition, let us recall the structure and functionality of a neuron in the human brain. The basic elements of such a neuron are dendrites, through which signals are transmitted to its soma while being scaled/amplified due to the structural properties of the respective dendrites. In the soma of the neuron, those incoming signals are accumulated, and a decision is reached whether to fire to other neurons or not, and also with which strength.

This forms the basis for a mathematical definition of an artificial neuron.

Definition 2.1. An *artificial neuron* with *weights* $w_1, \dots, w_n \in \mathbb{R}$, *bias* $b \in \mathbb{R}$, and *activation function* $\rho : \mathbb{R} \to \mathbb{R}$ is defined as the function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x_1, \dots, x_n) = \rho\left(\sum_{i=1}^{n} x_i w_i - b\right) = \rho\big(\langle x, w\rangle - b\big),$$

where $w = (w_1, \dots, w_n)$ and $x = (x_1, \dots, x_n)$.

By now, there exists a zoo of activation functions with the most well-known ones being as follows:

(1) Heaviside function

$$\rho(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0; \end{cases}$$

(2) Sigmoid function $\rho(x) = \frac{1}{1+e^{-x}}$;

(3) Rectifiable Linear Unit (ReLU) $\rho(x) = \max\{0, x\}$.

We remark that of these examples, by far the most extensively used activation function is the ReLU due to its simple piecewise-linear structure, which is advantageous in the training process and still allows superior performance.

Similar to the structure of a human brain, these artificial neurons are now being concatenated and arranged in layers, leading to an (artificial feed-forward) neural network. Due to the particular structure of artificial neurons, such a neural network consists of compositions of affine linear maps and activation functions. Traditionally, a deep neural network is then defined as the resulting function. From a mathematical standpoint, this bears the difficulty that different arrangements lead to the same function. Therefore, sometimes a distinction is made between the architecture of a neural network and the corresponding realization function (see, e.g., [6]). For this article, we will, however, avoid such technical delicacies and present the most standard definition.

Definition 2.2. Let $d \in \mathbb{N}$ be the dimension of the input layer, L the number of layers, $N_0 := d$, N_ℓ, $\ell = 1, \ldots, L$, the dimensions of the hidden and last layer, $\rho : \mathbb{R} \to \mathbb{R}$ a (non-linear) activation function, and, for $\ell = 1, \ldots, L$, let T_ℓ be the affine functions

$$T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, \quad T_\ell x = W^{(\ell)} x + b^{(\ell)},$$

with $W^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ being the weight matrices and $b^{(\ell)} \in \mathbb{R}^{N_\ell}$ the bias vectors of the ℓth layer. Then $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$, given by

$$\Phi(x) = T_L \rho\big(T_{L-1} \rho\big(\cdots \rho\big(T_1(x)\big)\big)\big), \quad x \in \mathbb{R}^d,$$

is called a *(deep) neural network* of *depth* L.

Let us already mention at this point that the weights and biases are the free parameters which will be learned during the training process. An illustration of the multilayered structure of a deep neural network can be found in Figure 1.

2.2. Application of a deep neural network

Aiming to identify the main mathematical research threads, we first have to understand how a deep neural network is used for a given application setting.

Step 1 (Train-test split of the dataset). We assume that we are given samples $(x^{(i)}, y^{(i)})_{i=1}^{\tilde{m}}$ of inputs and outputs. The task of the deep neural network is then to identify the relation between those. For instance, in a classification problem, each output $y^{(i)}$ is considered to be the label of the respective class to which the input $x^{(i)}$ belongs. One can also take the viewpoint that $(x^{(i)}, y^{(i)})_{i=1}^{\tilde{m}}$ arise as samples from a function such as

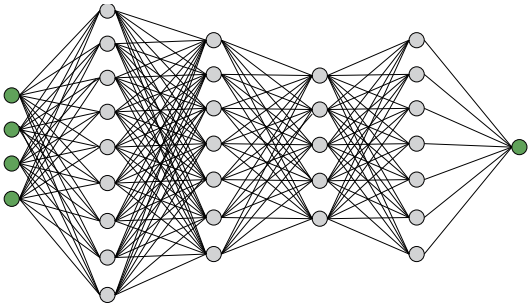


FIGURE 1
Deep neural network $\Phi : \mathbb{R}^4 \to \mathbb{R}$ with depth 5.

$g : \mathcal{M} \to \{1, 2, \dots, K\}$, where $\mathcal{M}$ might be a lower-dimensional manifold of $\mathbb{R}^d$, in the sense of $y^{(i)} = g(x^{(i)})$ for all $i = 1, \dots, \tilde{m}$.

The set $(x^{(i)}, y^{(i)})_{i=1}^{\tilde{m}}$ is then split into a training data set $(x^{(i)}, y^{(i)})_{i=1}^{m}$ and a test data set $(x^{(i)}, y^{(i)})_{i=m+1}^{\tilde{m}}$. The training data set is—as the name indicates—used for training, whereas the test data set will later on be solely exploited for testing the performance of the trained network. We emphasize that the neural network is not exposed to the test data set during the entire training process.

Step 2 (Choice of architecture). For preparation of the learning algorithm, the architecture of the neural network needs to be decided upon, which means the number of layers L, the number of neurons in each layer $(N_\ell)_{\ell=1}^{L}$, and the activation function ρ have to be selected. It is known that a fully connected neural network is often difficult to train, hence, in addition, one typically preselects certain entries of the weight matrices $(W^{(\ell)})_{\ell=1}^{L}$ to already be set to zero at this point.

For later purposes, we define the selected class of deep neural networks by $\mathcal{NN}_\theta$ with θ encoding this chosen architecture.

Step 3 (Training). The next step is the actual training process, which consists of learning the affine functions $(T_\ell)_{\ell=1}^{L} = (W^{(\ell)} \cdot + b^{(\ell)})_{\ell=1}^{L}$. This is accomplished by minimizing the *empirical risk*

$$\hat{\mathcal{R}}(\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}) := \frac{1}{m} \sum_{i=1}^{m} \big(\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}(x^{(i)}) - y^{(i)}\big)^2. \tag{2.1}$$

A more general form of the optimization problem is

$$\min_{(W^{(\ell)},b^{(\ell)})_\ell} \sum_{i=1}^{m} \mathcal{L}(\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}(x_i), y^{(i)}) + \lambda \mathcal{P}\big((W^{(\ell)}, b^{(\ell)})_\ell\big), \tag{2.2}$$

where $\mathcal{L}$ is a loss function to determine a measure of closeness between the network evaluated in the training samples and the (known) values $y^{(i)}$, with $\mathcal{P}$ being a penalty/regularization term to impose additional constraints on the weight matrices and bias vectors.

One common algorithmic approach is gradient descent. Since, however, m is typically very large, this is computationally not feasible. This problem is circumvented by randomly selecting only a few gradients in each iteration, assuming that they constitute a reasonable average, which is coined *stochastic gradient descent*.

Solving the optimization problem then yields a network $\Phi_{(W^{(\ell)},b^{(\ell)})_\ell} : \mathbb{R}^d \to \mathbb{R}^{N_L}$, where

$$\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}(x) = T_L \rho\big(T_{L-1}\rho\big(\cdots \rho\big(T_1(x)\big)\big)\big).$$

Step 4 (Testing). Finally, the performance (often also called generalization ability) of the trained neural network is tested using the test data set $(x^{(i)}, y^{(i)})_{i=m+1}^{\tilde{m}}$ by analyzing whether

$$\Phi_{(W^{(\ell)},b^{(\ell)})_\ell}(x^{(i)}) \approx y^{(i)}, \quad \text{for all } i = m+1, \dots, \tilde{m}.$$

2.3. Relation to a statistical learning problem

From the procedure above, we can already identify the selection of architecture, the optimization problem, and the generalization ability as the key research directions for mathematical foundations of deep neural networks. Considering the entire learning process of a deep neural network as a statistical learning problem reveals those three research directions as indeed the natural ones for analyzing the overall error.

For this, let us assume that there exists a function $g : \mathbb{R}^d \to \mathbb{R}$ such that the training data $(x^{(i)}, y^{(i)})_{i=1}^m$ is of the form $(x^{(i)}, g(x^{(i)}))_{i=1}^m$ and $x^{(i)} \in [0,1]^d$ for all $i = 1, \dots, m$. A typical continuum viewpoint to measure success of the training is to consider the *risk* of a function $f : \mathbb{R}^d \to \mathbb{R}$ given by

$$\mathcal{R}(f) := \int_{[0,1]^d} \big(f(x) - g(x)\big)^2 \, dx, \tag{2.3}$$

where we used the L^2-norm to measure the distance between f and g. The error between the trained deep neural network $\Phi^0 (:= \Phi_{(W^{(\ell)},b^{(\ell)})_\ell}) \in \mathcal{NN}_\theta$ and the optimal function g can then be estimated by

$$\mathcal{R}(\Phi^0) \le \underbrace{\Big[\hat{\mathcal{R}}(\Phi^0) - \inf_{\Phi \in \mathcal{NN}_\theta} \hat{\mathcal{R}}(\Phi)\Big]}_{\text{Optimization error}} + \underbrace{2 \sup_{\Phi \in \mathcal{NN}_\theta} \big|\mathcal{R}(\Phi) - \hat{\mathcal{R}}(\Phi)\big|}_{\text{Generalization error}} + \underbrace{\inf_{\Phi \in \mathcal{NN}_\theta} \mathcal{R}(\Phi)}_{\text{Approximation error}}. \tag{2.4}$$

These considerations lead to the main research threads described in the following subsection.

2.4. Main research threads

We can identify two conceptually different research threads, the first being focused on developing mathematical foundations of artificial intelligence and the second aiming to use methodologies from artificial intelligence to solve mathematical problems. It is intriguing to see how both have already led to some extent to a paradigm shift in some mathematical research areas, most prominently the area of numerical analysis.

2.4.1. Mathematical foundations for artificial intelligence

Following up on the discussion in Section 2.3, we can identify three research directions which are related to the three types of errors which one needs to control in order to estimate the overall error of the entire training process:

- *Expressivity*. This direction aims to derive a general understanding whether and to which extent aspects of a neural network architecture affect the best case performance of deep neural networks. More precisely, the goal is to analyze the approximation error $\inf_{\Phi \in \mathcal{NN}_\theta} \mathcal{R}(\Phi)$ from (2.4), which estimates the approximation accuracy when approximating g by the hypothesis class $\mathcal{NN}_\theta$ of deep neural networks of a particular architecture. Typical methods for approaching this problem are from applied harmonic analysis and approximation theory.

- *Learning/Optimization*. The main goal of this direction is the analysis of the training algorithm such as stochastic gradient descent, in particular, asking why it usually converges to suitable local minima even though the problem itself is highly nonconvex. This requires the analysis of the optimization error, which is $\hat{\mathcal{R}}(\Phi^0) - \inf_{\Phi \in \mathcal{NN}_\theta} \hat{\mathcal{R}}(\Phi)$ (cf. (2.4)) and which measures the accuracy with which the learnt neural network Φ^0 minimizes the empirical risk (2.1), (2.2). Key methodologies for attacking such problems come from the areas of algebraic/differential geometry, optimal control, and optimization.

- *Generalization*. This direction aims to derive an understanding of the out-of-sample error, namely, $\sup_{\Phi \in \mathcal{NN}_\theta} |\mathcal{R}(\Phi) - \hat{\mathcal{R}}(\Phi)|$ from (2.4), which measures the distance of the empirical risk (2.1), (2.2) and the actual risk (2.3). Predominantly, learning theory, probability theory, and statistics provide the required methods for this research thread.

A very exciting and highly relevant new research direction has recently emerged, coined explainability. At present, it is from the standpoint of mathematical foundations still a wide open field.

- *Explainability.* This direction considers deep neural networks, which are already trained, but no knowledge about the training is available; a situation one encounters numerous times in practice. The goal is then to derive a deep understanding of how a given trained deep neural network reaches decisions in the sense of which features of the input data are crucial for a decision. The range of required approaches is quite broad, including areas such as information theory or uncertainty quantification.

2.4.2. Artificial intelligence for mathematical problems

Methods of artificial intelligence have also turned out to be extremely effective for mathematical problem settings. In fact, the area of inverse problems, in particular, in imaging sciences, has already undergone a profound paradigm shift. And the area of numerical

analysis of partial differential equations seems to soon follow the same path, at least in the very high dimensional regime.

Let us briefly characterize those two research threads similar to the previous subsection on mathematical foundations of artificial intelligence.

- *Inverse Problems.* Research in this direction aims to improve classical model-based approaches to solve inverse problems by exploiting methods of artificial intelligence. In order to not neglect domain knowledge such as the physics of the problem, current approaches aim to take the best out of both worlds in the sense of optimally combining model- and data-driven approaches. This research direction requires a variety of techniques, foremost from areas such as imaging science, inverse problems, and microlocal analysis, to name a few.

- *Partial Differential Equations.* Similar to the area of inverse problems, here the goal is to improve classical solvers of partial differential equations by using ideas from artificial intelligence. A particular focus is on high-dimensional problems in the sense of aiming to beat the curse of dimensionality. This direction obviously requires methods from areas such as numerical mathematics and partial differential equations.

3. MATHEMATICAL FOUNDATIONS FOR ARTIFICIAL INTELLIGENCE

This section shall serve as an introduction into the main research threads aiming to develop a mathematical foundation for artificial intelligence. We will introduce the problem settings, showcase some exemplary results, and discuss open problems.

3.1. Expressivity

Expressivity is maybe the richest area at present in terms of mathematical results. The general question can be phrased as follows: Given a function class/space $\mathcal{C}$ and a class of deep neural networks $\mathcal{NN}_\theta$, how does the approximation accuracy when approximating elements of $\mathcal{C}$ by networks $\Phi \in \mathcal{NN}_\theta$ relate to the complexity of such Φ? Making this precise thus requires the introduction of a complexity measure for deep neural networks. In the sequel, we will choose the canonical one, which is the complexity in terms of memory requirements. Notice though that certainly various other complexity measures exist. Further, recall that the $\|\cdot\|_0$-"norm" counts the number of nonzero components.

Definition 3.1. Retaining the same notation for deep neural networks as in Definition 2.2, the *complexity* $C(\Phi)$ of a deep neural network Φ is defined by

$$C(\Phi) := \sum_{\ell=1}^{L} \big(\|W^{(\ell)}\|_0 + \|b^{(\ell)}\|_0 \big).$$

The most well-known—and maybe even the first—result on expressivity is the universal approximation theorem [8, 13]. It states that each continuous function on a compact domain can be approximated up to an arbitrary accuracy by a shallow neural network.

Theorem 3.2. *Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, $\rho : \mathbb{R} \to \mathbb{R}$ continuous and not a polynomial. Then, for each $\varepsilon > 0$, there exist $N \in \mathbb{N}$ and $a_k, b_k \in \mathbb{R}$, $w_k \in \mathbb{R}^d$, $1 \leq k \leq N$, such that*

$$\left\| f - \sum_{k=1}^{N} a_k \rho\big(\langle w_k, \cdot \rangle - b_k\big) \right\|_\infty \leq \varepsilon.$$

While this is certainly an interesting result, it is not satisfactory in several regards: It does not give bounds on the complexity of the approximating neural network and also does not explain why depth is so important. A particularly intriguing example for a result, which considers complexity and also targets a more sophisticated function space, was derived in [31].

Theorem 3.3. *For all $f \in C^s([0,1]^d)$ and $\rho(x) = \max\{0, x\}$, i.e., the ReLU, there exist neural networks $(\Phi_n)_{n\in\mathbb{N}}$ with the number of layers of Φ_n being approximately of the order of $\log(n)$ such that*

$$\|f - \Phi_n\|_\infty \lesssim C(\Phi_n)^{-\frac{s}{d}} \to 0 \quad \text{as } n \to \infty.$$

This result provides a beautiful connection between approximation accuracy and complexity of the approximating neural network, and also to some extent takes the depth of the network into account. However, to derive a result on optimal approximations, we first require a lower bound. The so-called VC-dimension (Vapnik–Chervonenkis-dimension) (see also (3.2)) was for a long time the main method for achieving such lower bounds. We will recall here a newer result from [7] in terms of the optimal exponent $\gamma^*(\mathcal{C})$ from information theory to measure the complexity of $\mathcal{C} \subset L^2(\mathbb{R}^d)$. Notice that we will only state the essence of this result without all technicalities.

Theorem 3.4. *Let $d \in \mathbb{N}$, $\rho : \mathbb{R} \to \mathbb{R}$, and let $\mathcal{C} \subset L^2(\mathbb{R}^d)$. Further, let*

$$\boldsymbol{Learn} : (0,1) \times \mathcal{C} \to \mathcal{NN}_\theta$$

satisfy that, for each $f \in \mathcal{C}$ and $0 < \varepsilon < 1$,

$$\sup_{f\in\mathcal{C}} \big\| f - \boldsymbol{Learn}(\varepsilon, f) \big\|_2 \leq \varepsilon.$$

Then, for all $\gamma < \gamma^*(\mathcal{C})$,

$$\varepsilon^\gamma \sup_{f\in\mathcal{C}} C\big(\boldsymbol{Learn}(\varepsilon, f)\big) \to \infty, \quad \text{as } \varepsilon \to 0.$$

This conceptual lower bound, which is independent of any learning algorithm, now allows deriving results on approximations with neural networks, which have optimally small complexity in the sense of being memory-optimal. We will next provide an example of such a result, which at the same time answers another question as well. The universal approximation theorem already indicates that deep neural networks seem to have a universality property in the sense of performing at least as good as polynomial approximation. One can now ask whether neural networks also perform as well as other existing approximation schemes such as wavelets, or the more sophisticated system of shearlets [16].

For this, let us briefly recall this system and its approximation properties. Shearlets are based on parabolic scaling, i.e.,

$$A_{2^j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \quad j \in \mathbb{Z}$$

and $\tilde{A}_{2^j} = \mathrm{diag}(2^{j/2}, 2^j)$, as well as changing the orientation via shearing defined by

$$S_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad k \in \mathbb{Z}.$$

(Cone-adapted) discrete shearlet systems can then be defined as follows, cf. [17]. A faithful implementation of the shearlet transform as a 2D and 3D (parallelized) fast shearlet transform can be found at www.ShearLab.org.

Definition 3.5. The *(cone-adapted) discrete shearlet system* $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of

$$\begin{aligned} &\{\phi(\cdot - m) : m \in \mathbb{Z}^2\}, \\ &\{2^{3j/4}\psi(S_k A_{2^j} \cdot - m) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}, \\ &\{2^{3j/4}\tilde{\psi}(S_k^T \tilde{A}_{2^j} \cdot - m) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}. \end{aligned}$$

Since multivariate problems are typically governed by anisotropic features such as edges in images or shock fronts in the solution of transport-dominated equations, the following suitable model class of functions was introduced in [9].

Definition 3.6. The set of *cartoon-like functions* $\mathcal{E}^2(\mathbb{R}^2)$ is defined by

$$\mathcal{E}^2(\mathbb{R}^2) = \{f \in L^2(\mathbb{R}^2) : f = f_0 + f_1 \cdot \chi_B\},$$

where $\emptyset \neq B \subset [0,1]^2$ is simply connected with a C^2-curve with bounded curvature as its boundary, and $f_i \in C^2(\mathbb{R}^2)$ with $\mathrm{supp}\, f_i \subseteq [0,1]^2$ and $\|f_i\|_{C^2} \leq 1$, $i = 0, 1$.

While wavelets are deficient in optimally approximating cartoon-like functions due to their isotropic structure, shearlets provide an optimal (sparse) approximation rate up to a log-factor. The following statement is taken from [17], where also the precise hypotheses can be found. Notice that the justification for optimality is a benchmark result from [9].

Theorem 3.7. *Let $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ be compactly supported, and let $\hat{\psi}$, $\hat{\tilde{\psi}}$ satisfy certain decay conditions. Then $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ provides an* optimally sparse approximation *of $f \in \mathcal{E}^2(\mathbb{R}^2)$, i.e.,*

$$\sigma_N(f) \lesssim N^{-1}(\log N)^{\frac{3}{2}} \quad \text{as } N \to \infty,$$

where $\sigma_N(f)$ denotes the L_2-error of best N-term approximation of f.

One can now use Theorem 3.4 to show that indeed deep neural networks are as good approximators as shearlets and, in fact, as all affine systems. Even more, the construction in the proof of suitable neural networks, which mimics best N-term approximations, also leads

to memory-optimal neural networks. The resulting statement from [7] in addition proves that the bound in Theorem 3.4 is sharp.

Theorem 3.8. *Let ρ be a suitably chosen activation function, and let $\varepsilon > 0$. Then, for all $f \in \mathcal{E}^2(\mathbb{R}^2)$ and $N \in \mathbb{N}$, there exist a neural network Φ with complexity $O(N)$ and activation function ρ with*

$$\|f - \Phi\|_2 \lesssim N^{-1+\varepsilon} \to 0 \quad \text{as } N \to \infty.$$

Summarizing, one can conclude that deep neural networks achieve optimal approximation properties of all affine systems combined.

Let us finally mention that lately a very different viewpoint of expressivity was introduced in [21] according to so-called trajectory lengths. The standpoint taken in this work is to measure expressivity in terms of changes of the expected length of a (nonconstant) curve in the input space as it propagates through layers of a neural network.

3.2. Optimization

This area aims to analyze optimization algorithms, which solve the (learning) problem in (2.1), or, more generally, (2.2). A common approach is gradient descent, since the gradient of the loss function (or optimized functional) with respect to the weight matrices and biases, i.e., the parameters of the network, can be computed exactly. This is done via backpropagation [27], which is in a certain sense merely an efficient application of the chain rule. However, since the number of training samples is typically in the millions, it is computationally infeasible to compute the gradient on each sample. Therefore, in each iteration only one or several (a batch) randomly selected gradients are computed, leading to the algorithm of *stochastic gradient descent* [25].

In convex settings, guarantees for convergence of stochastic gradient descent do exist. However, in the neural network setting, the optimization problem is nonconvex, which makes it—even when using a nonrandom version of gradient descent—very hard to analyze. Including randomness adds another level of difficulty as is depicted in Figure 2, where the two algorithms reach different (local) minima.

This area is by far less explored than expressivity. Most current results focus on shallow neural networks, and for a survey, we refer to [6].

3.3. Generalization

This research direction is perhaps the least explored and maybe also the most difficult one, sometimes called the "holy grail" of understanding deep neural networks. It targets the out-of-sample error

$$\sup_{\Phi \in \mathcal{NN}_\theta} \left| \mathcal{R}(\Phi) - \hat{\mathcal{R}}(\Phi) \right| \tag{3.1}$$

as described in Section 2.4.1.

One of the mysteries of deep neural networks is the observation that highly overparameterized deep neural networks in the sense of high complexity of the network do *not*

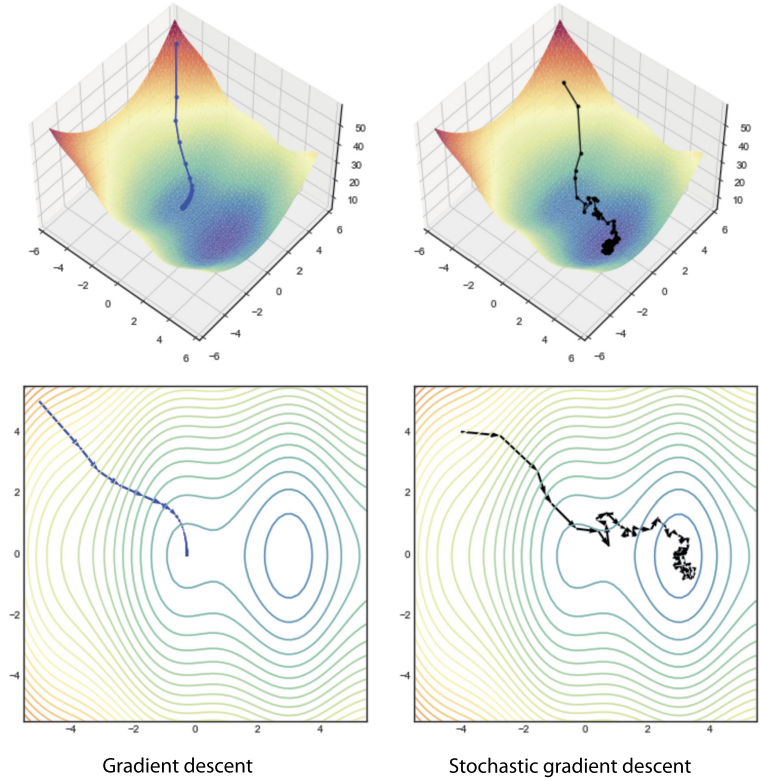


FIGURE 2
Gradient descent versus stochastic gradient descent. Taken from [6]. © Cambridge University Press. Reprinted with permission.

overfit, with overfitting referring to the problem of fitting the training data too tightly and consequently endangering correct classification of new data. An illustration of the phenomenon of overfitting can be found in Figure 3.

Let us now analyze the generalization error in (3.1) in a bit more depth. For a large number m of training samples, the law of large numbers tells us that with high probability $\hat{\mathcal{R}}(\Phi) \approx \mathcal{R}(\Phi)$ for each neural network $\Phi \in \mathcal{NN}_\theta$. Bounding the complexity of the hypothesis class $\mathcal{NN}_\theta$ by the VC-dimension, the generalization error can be bounded with probability $1 - \delta$ by

$$\sqrt{\frac{\mathrm{VCdim}(\mathcal{NN}_\theta) + \log(1/\delta)}{m}}. \tag{3.2}$$

For classes of highly over-parametrized neural networks, i.e., where $\mathrm{VCdim}(\mathcal{NN}_\theta)$ is very large, we need an enormous amount of training data to keep the generalization error under control. It is thus more than surprising that numerical experiments show the phenomenon of a so-called *double descent curve* [5]. More precisely, the test error was found to decrease after

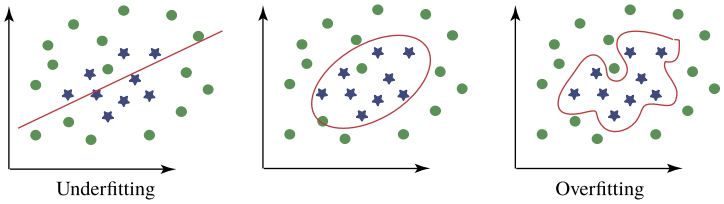


FIGURE 3
Phenomenon of overfitting for the task of classification with two classes.

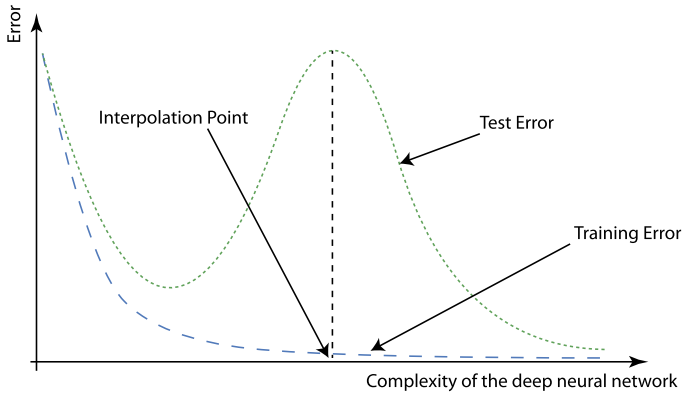


FIGURE 4
Double descent curve.

passing the interpolation point, which follows an increase consistent with statistical learning theory (see Figure 4).

3.4. Explainability

The area of explainability aims to "open the black box" of deep neural networks in the sense as to explain decisions of trained neural networks. These explanations typically consist of providing relevance scores for features of the input data. Most approaches focus on the task of image classification and provide relevance scores for each pixel of the input image. One can roughly categorize the different types of approaches into gradient-based methods [28], propagation of activations in neurons [4], surrogate models [24], and game-theoretic approaches [19].

We would now like to describe in more detail an approach which is based on information theory and also allows an extension to different modalities such as audio data as well as analyzing the relevance of higher-level features; for a survey paper, we refer to [15]. This *rate-distortion explanation (RDE)* framework was introduced in 2019 and later extended by applying RDE to noncanonical input representations.

Let now $\Phi : \mathbb{R}^d \to \mathbb{R}^n$ be a trained neural network, and $x \in \mathbb{R}^d$. The goal of RDE is to provide an explanation for the decision $\Phi(x)$ in terms of a sparse mask $s \in \{0, 1\}^d$

which highlights the crucial input features of x. This mask is determined by the following optimization problem:

$$\min_{s\in\{0,1\}^d} \quad \mathbb{E}_{v\sim\mathcal{V}}\, d\big(\Phi(x), \Phi\big(x\odot s + (1-s)\odot v\big)\big) \quad \text{subject to } \|s\|_0 \le \ell,$$

where $\odot$ denotes the Hadamard product, d is a measure of distortion such as the ℓ_2-distance, $\mathcal{V}$ is a distribution over input perturbations $v \in \mathbb{R}^d$, and $\ell \in \{1,\dots,d\}$ is a given sparsity level for the explanation mask s. The key idea is that a solution s^* is a mask marking few components of the input x which are sufficient to approximately retain the decision $\Phi(x)$. This viewpoint reveals the relation to rate-distortion theory, which normally focusses on lossy compression of data.

Since it is computationally infeasible to compute such a minimizer (see [30]), a relaxed optimization problem providing continuous masks $s \in [0,1]^d$ is used in practice:

$$\min_{s\in[0,1]^d} \quad \mathbb{E}_{v\sim\mathcal{V}}\, d\big(\Phi(x), \Phi\big(x\odot s + (1-s)\odot v\big)\big) + \lambda\|s\|_1,$$

where $\lambda > 0$ determines the sparsity level of the mask. The minimizer now assigns each component x_i of the input—in case of images each pixel—a relevance score $s_i \in [0,1]$. This is typically referred to as *Pixel RDE*.

Extensions of the RDE-framework allow the incorporation of different distributions $\mathcal{V}$ better adapted to data distributions. Another recent improvement was the assignment of relevance scores to higher-level features such as arising from a wavelet decomposition, which ultimately led to the approach *CartoonX*. An example of Pixel RDE versus CartoonX, which also shows the ability of the higher-level explanations of CartoonX to give insights into what the neural network saw when misclassifying an image, is depicted in Figure 5.

4. ARTIFICIAL INTELLIGENCE FOR MATHEMATICAL PROBLEMS

We now turn to the research direction of artificial intelligence for mathematical problems, with the two most prominent problems being inverse problems and partial differential equations. As before, we will introduce the problem settings, showcase some exemplary results, and also discuss open problems.

4.1. Inverse problems

Methods of artificial intelligence, in particular, deep neural networks, have a tremendous impact on the area of inverse problems, as already indicated before. One current major trend is to optimally combine classical solvers with deep learning in the sense of taking the best out of the model- and data-world.

To introduce such results, we start by recalling some basics about solvers of inverse problems. For this, assume that we are given an (ill-posed) inverse problem

$$Kf = g, \tag{4.1}$$

FIGURE 5
Pixel RDE versus CartoonX for analyzing misclassifications of deep neural networks, where the first image is misclassified as "Diaper", the second as "Egyptian Cat", and the third as "Screw".

where $K : X \to Y$ is an operator and X and Y are, for instance, Hilbert spaces. Drawing from the area of imaging science, examples include denoising, deblurring, or inpainting (recovery of missing parts of an image). Most classical solvers are of the form (which includes Tikhonov regularization)

$$f^\alpha := \operatorname*{argmin}_f \Big[\underbrace{\|Kf - g\|^2}_{\text{Data fidelity term}} + \alpha \cdot \underbrace{\mathcal{P}(f)}_{\text{Penalty/Regularization term}} \Big],$$

where $\mathcal{P} : X \to \mathbb{R}$ and $f^\alpha \in X, \alpha > 0$ is an approximate solution of the inverse problem (4.1). One very popular and widely applicable special case is *sparse regularization*, where $\mathcal{P}$ is chosen by

$$\mathcal{P}(f) := \big\| \big(\langle f, \varphi_i \rangle \big)_{i \in I} \big\|_1$$

and $(\varphi_i)_{i \in I}$ is a suitably selected orthonormal basis or a frame for X.

We now turn to deep learning approaches to solve inverse problems, which might be categorized into three classes:

- *Supervised approaches.* An ad hoc approach in this regime is given in [14], which first applies a classical solver followed by a neural network to remove reconstruction artifacts. More sophisticated approaches typically replace parts of the classical solver by a custom-built neural network [26] or a network specifically trained for this task [1].
- *Semisupervised approaches.* These approaches encode the regularization as a neural network with an example being adversarial regularizers [20].
- *Unsupervised approaches.* A representative of this type of approaches is the technique of deep image prior [29]. This method interestingly shows that the structure of a generator network is sufficient to capture necessary statistics of the data prior to any type of learning.

Aiming to illustrate the superiority of approaches from artificial intelligence for inverse problems, we will now focus on the inverse problem of computed tomography (CT) from medical imaging. The forward operator K in this setting is the *Radon transform*, defined by

$$\mathcal{R}f(s,\vartheta) = \int_{-\infty}^{\infty} f\big(s\omega(\vartheta) + t\omega(\vartheta)^{\perp}\big)dt, \quad \text{for } (s,\vartheta) \in \mathbb{R}\times(0,\pi).$$

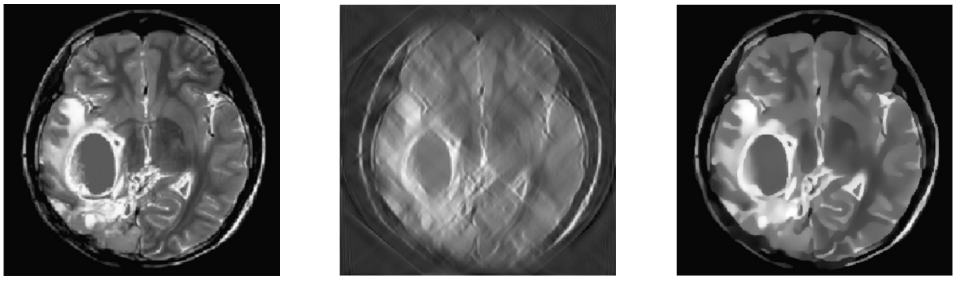
Here $\omega(\vartheta) := (\cos\vartheta, \sin\vartheta)$ is the unitary vector with orientation described by the angle ϑ with respect to the x_1-axis and $\omega(\vartheta)^{\perp} := (-\sin\vartheta, \cos\vartheta)$. Often, only parts of the so-called sinogram $\mathcal{R}f$ can be acquired due to physical constraints as in, for instance, electron tomography. The resulting, more difficult problem is termed *limited-angle CT*. One should notice that this problem is even harder than the problem of low-dose CT, where not an entire block of measurements is missing, but the angular component is "only" undersampled.

The most prominent features in images f are edge structures. This is also due to the fact that the human visual system reacts most strongly to those. These structures in turn can be accurately modeled by microlocal analysis, in particular, by the notion of wavefront sets $WF(f) \subseteq \mathbb{R}^2 \times \mathbb{S}$, which—coarsely speaking—consist of singularities together with their direction. Basing in this sense the application of a deep neural network on microlocal considerations, in particular, also using a deep learning-based wavefront set detector [2] in the regularization term, the reconstruction performance significantly outperforms classical solvers such as sparse regularization with shearlets (see Figure 6, we also refer to [3] for details). Notice that this approach is of a hybrid type and takes the best out of both worlds in the sense of combining model- and artificial intelligence-based approaches.

Finally, the deep learning-based wavefront set extraction itself is yet another evidence of the improvements on the state-of-the-art now possible by artificial intelligence. Figure 7 shows a classical result from [23], whereas [2] uses the shearlet transform as a coarse edge detector, which is subsequently combined with a deep neural network.

4.2. Partial differential equations

The second main range of mathematical problem settings, where methods from artificial intelligence are very successfully applied to, are partial differential equations. Although



Original | Sparse Regularization with Shearlets | Deep Microlocal Reconstruction [3]

FIGURE 6
CT reconstruction from Radon measurements with a missing angle of 40°.

Original | CoShREM [23] | DeNSE [2]

FIGURE 7
Wavefront set detection by a model-based and a hybrid approach.

the benefit of such approaches was not initially clear, both theoretical and numerical results show their superiority in high-dimensional regimes.

The most common approach aims to approximate the solution of a partial differential equation by a deep neural network, which is trained according to this task by incorporating the partial differential equation into the loss function. More precisely, given a partial differential equation $\mathcal{L}(u) = f$, we train a neural network Φ such that

$$\mathcal{L}(\Phi) \approx f.$$

Since 2017, research in this general direction has significantly accelerated. Some of the highlights are the Deep Ritz Method [10] and Physics Informed Neural Networks [22], or a very general approach for high-dimensional parabolic partial differential equations [12].

One should note that most theoretical results in this regime are of an expressivity type and also study the phenomenon whether and to which extent deep neural networks are able to beat the curse of dimensionality. In the sequel, we briefly discuss one such result as an example. In addition, notice that there already exist contributions—though very few—which analyze learning and generalization aspects.

Let $\mathcal{L}(u_y, y) = f_y$ denote a parametric partial differential equation with y being a parameter from a high-dimensional parameter space $\mathcal{Y} \subseteq \mathbb{R}^p$ and u_y the associated solution in a Hilbert space $\mathcal{H}$. After a high-fidelity discretization, let $b_y(u_y^h, v) = f_y(v)$ be the associated variational form with u_y^h now belonging to the associated high-dimensional space U^h, where we set $D := \dim(U^h)$. We, moreover, denote the coefficient vector of u_y^h with respect to a suitable basis of U^h by $\mathbf{u}_y^{\mathrm{h}}$. Of key importance in this area is the *parametric map* given by

$$\mathbb{R}^p \supseteq \mathcal{Y} \ni y \mapsto \mathbf{u}_y^{\mathrm{h}} \in \mathbb{R}^D \quad \text{such that } b_y\big(u_y^h, v\big) = f_y(v) \quad \text{for all } v,$$

which in multiquery situations such as complex design problems needs to be solved several times. If p is very large, the curse of dimensionality could lead to an exponential computational cost.

We now aim to analyze whether the parametric map can be solved by a deep neural network, which would provide a very efficient and flexible method, hopefully also circumventing the curse of dimensionality in an automatic manner. From an expressivity viewpoint, one might ask whether, for each $\varepsilon > 0$, there exists a neural network Φ such that

$$\left\| \Phi(y) - \mathbf{u}_y^{\mathrm{h}} \right\| \leq \varepsilon \quad \text{for all } y \in \mathcal{Y}. \tag{4.2}$$

The ability of this approach to tackle the curse of dimensionality can then be studied by analyzing how the complexity of Φ depends on p and D. A result of this type was proven in [18], the essence of which we now recall.

Theorem 4.1. *There exists a neural network Φ which approximates the parametric map, i.e., which satisfies (4.2), and the dependence of $C(\Phi)$ on p and D can be (polynomially) controlled.*

Analyzing the learning procedure and the generalization ability of the neural network in this setting is currently out of reach. Aiming to still determine whether a trained neural networks does not suffer from the curse of dimensionality as well, in [11] extensive numerical experiments were performed, which indicate that, indeed, the curse of dimensionality is also beaten in practice.

5. CONCLUSION: SEVEN MATHEMATICAL KEY PROBLEMS

Let us conclude with seven mathematical key problems of artificial intelligence as they were stated in [6]. Those constitute the main obstacles in *Mathematical Foundations for Artificial Intelligence* with its subfields being expressivity, optimization, generalization, and explainability, as well as in *Artificial Intelligence for Mathematical Problems,* which focus on the application to inverse problems and partial differential equations:

(1) What is the role of depth?

(2) Which aspects of a neural network architecture affect the performance of deep learning?

(3) Why does stochastic gradient descent converge to good local minima despite the nonconvexity of the problem?

(4) Why do large neural networks not overfit?

(5) Why do neural networks perform well in very high-dimensional environments?

(6) Which features of data are learned by deep architectures?

(7) Are neural networks capable of replacing highly specialized numerical algorithms in natural sciences?

ACKNOWLEDGMENTS

The author would like to thank Hector Andrade Loarca, Adalbert Fono, Stefan Kolek, Yunseok Lee, Philipp Scholl, Mariia Seleznova, and Laura Thesing for their helpful feedback on an early version of this article.

FUNDING

This research was partly supported by the Bavarian High-Tech Agenda, DFG-SFB/TR 109 Grant C09, DFG-SPP 1798 Grant KU 1446/21-2, DFG-SPP 2298 Grant KU 1446/32-1, and NSF-Simons Research Collaboration on the Mathematical and Scientific Foundations of Deep Learning (MoDL) (NSF DMS 2031985).

REFERENCES

[1] J. Adler and O. Öktem, Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Probl.* **33** (2017), 124007.

[2] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen, Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM J. Imaging Sci.* **12** (2019), 1936–1966.

[3] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen, Deep microlocal reconstruction for limited-angle tomography. *Appl. Comput. Harmon. Anal.*, to appear.

[4] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10** (2015), e0130140.

[5] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** (2019), 15849–15854.

[6] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, The modern mathematics of deep learning. In *Mathematical aspects of deep learning*, Cambridge University Press, to appear.

[7] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.* **1** (2019), 8–45.

[8] G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signal* **2** (1989), 303–314.

[9] D. Donoho, Sparse components of images and optimal atomic decompositions. *Constr. Approx.* **17** (2001), 353–382.

[10] W. E and B. Yu, The Deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.* **6** (2018), 1–12.

[11] M. Geist, P. Petersen, M. Raslan, R. Schneider, and G. Kutyniok, Numerical solution of the parametric diffusion equation by deep neural networks. *J. Sci. Comput.* **88** (2021), Article number: 22.

[12] J. Han and A. Jentzen W. E, Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA* **115** (2018), 8505–8510.

[13] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** (1989), 359–366.

[14] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **26** (2017), 4509–4522.

[15] S. Kolek, D. A. Nguyen, R. Levie, J. Bruna, and G. Kutyniok, *A rate-distortion framework forexplaining black-box model decisions*. Springer LNAI, Beyond Explainable AI, to appear.

[16] G. Kutyniok and D. Labate, Introduction to shearlets. In *Shearlets: multiscale analysis for multivariate data*, pp. 1–38, Birkhäuser, Boston, 2012.

[17] G. Kutyniok and W.-Q. Lim, Compactly supported shearlets are optimally sparse. *J. Approx. Theory* **163** (2011), 1564–1589.

[18] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider, A theoretical analysis of DeepNeural networks and parametric PDEs. *Constr. Approx.* **55** (2022), 73–125.

[19] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions. In *NeurIPS*, pp. 4768–4777, 2017.

[20] S. Lunz, O. Öktem, and C.-B. Schönlieb, Adversarial regularizers in inverse problems. In *NIPS*, pp. 8507–8516, 2018.

[21] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, On the expressive power of deep neural networks. In *ICML*, pp. 2847–2854, 2017.

[22] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear pdes. *J. Comput. Phys.* **378** (2019), 686–707.

[23] R. Reisenhofer, J. Kiefer, and E. J. King, Shearlet-based detection of flame fronts. *Exp. Fluids* **57** (2015), 11.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": explaining the predictions of any classifier. In *ACM SIGKDD*, pp. 1135–1144, 2016.

[25] H. Robbins and S. Monro, A stochastic approximation method. *Ann. Math. Statist.* **22** (1952), 400–407.

[26] Y. Romano, M. Elad, and P. Milanfar, The little engine that could: regularization by denoising. *SIAM J. Imaging Sci.* **10** (2017), 1804–1844.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors. *Nature* **323** (1986), 533–536.

[28] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, SmoothGrad: removing noise by adding noise. In: *ICML (2017), workshop on visualization for deep learning*, 2017.

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky, Deep image prior. In *CVPR*, pp. 9446–9454, 2018.

[30] S. Wäldchen, J. Macdonald, S. Hauch, and G. Kutyniok, The computational complexity of understanding network decisions. *J. Artif. Intell. Res.* **70** (2021), 351–387.

[31] D. Yarotsky, Error bounds for approximations with deep ReLU networks. *Neural Netw.* **94** (2017), 103–114.

GITTA KUTYNIOK

Ludwig-Maximilians-Universität München, Department of Mathematics, 80333 München, Germany, kutyniok@math.lmu.de

STOCHASTIC GRADIENT DESCENT: WHERE OPTIMIZATION MEETS MACHINE LEARNING

RACHEL WARD

ABSTRACT

Stochastic gradient descent (SGD) is the de facto optimization algorithm for training neural networks in modern machine learning, thanks to its unique scalability to problem sizes where the data points, the number of data points, and the number of free parameters to optimize are on the scale of billions. On the one hand, many of the mathematical foundations for stochastic gradient descent were developed decades before the advent of modern deep learning, from stochastic approximation to the randomized Kaczmarz algorithm for solving linear systems. On the other hand, the omnipresence of stochastic gradient descent in modern machine learning and the resulting importance of optimizing performance of SGD in practical settings have motivated new algorithmic designs and mathematical breakthroughs. In this note, we recall some history and state-of-the-art convergence theory for SGD which is most useful in modern applications where it is used. We discuss recent breakthroughs in adaptive gradient variants of stochastic gradient descent, which go a long way towards addressing one of the weakest points of SGD: its sensitivity and reliance on hyperparameters, most notably, the choice of step-sizes.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 74P99; Secondary 93E35, 46N30, 46N40

KEYWORDS

Adaptive gradient methods, machine learning, smoothness, stepsize, stochastic approximation

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5140–5153
DOI 10.4171/ICM2022/171

1. INTRODUCTION

In the past several decades, *randomized* algorithms have slowly gained popularity and established legitimacy as scalable extensions of classical deterministic algorithms to large scales. Perhaps the most widely used randomized algorithm today is *stochastic gradient descent*, which has established itself in the past decade as the de facto optimization method for training artificial neural networks.

A common optimization problem in large-scale machine learning involves a training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$ (we suppose that the labels $y_j \in \mathbb{R}$ for simplicity), a parameterized family of prediction functions h, and a least squares minimization problem of the form

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \big(h(x_i; w) - y_i\big)^2. \tag{1.1}$$

In *linear* least squares regression, the prediction function is linear with respect to the weights w; for example, the prediction function is $h(x; w) = \sum_{j=1}^{p} w_j x_j^{p-1}$ in univariate polynomial regression. By contrast, in "neural network" regression, the prediction function h is a parameterized class of highly nonlinear functions inspired by models of how the human brain processes information. The neural network's compositional structure allows for the prediction function $h(x_i; w)$ to be computed at given values of x_i and w by recursively applying successive transformations to the input vector $x_i \in \mathbb{R}^{d_0}$ in layers. For example, a canonical fully-connected layer corresponds to the computation

$$x_i^{(j)} = \sigma(W_j x_i^{(j-1)} + b_j) \in \mathbb{R}^{d_j}, \tag{1.2}$$

where $x_i^{(0)} = x_i$, $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$, the vector $b_j \in \mathbb{R}^{d_j}$ contains the jth layer parameters, and σ is a simple componentwise nonlinear activation function such as the ReLU function $\sigma(x) = \max\{0, x\}$; the total number of parameters $w \in \mathbb{R}^p$ in (1.1) is the sum of the parameters at each of L layers, $w = (W_1, b_1, W_2, b_2, \dots, W_L, b_L)$. "Neural network training" refers to solving the optimization problem (1.1), either exactly or approximately. A particular vector of parameters $w^* \in \mathbb{R}^p$ corresponding to an approximate solution of the optimization problem (1.1) is considered to be a "good" solution if the corresponding neural network function $h(\cdot; w^*)$ has good generalization properties, meaning that when applied to fresh data $\{(\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_m, \tilde{y}_m)\}$ from the same distribution from which the training data was drawn, the distortion $\frac{1}{m} \sum_{i=1}^{m} (h(\tilde{x}_i; w^*) - \tilde{y}_i)^2$ is small. Thus, optimization and generalization must both be taken into account when discussing the performance of a particular algorithm for neural network training. In this note, we will only discuss the optimization component of the stochastic gradient descent algorithm. The generalization of solutions $w^* \in \mathbb{R}^p$ found by SGD tends to be remarkably strong, but this is not as well understood mathematically and represents an important ongoing area of research.

To motivate the stochastic gradient descent algorithm, let us first recall the basic gradient descent procedure for minimizing a differentiable function $F : \mathbb{R}^p \to \mathbb{R}$: starting from an initial point $w_0 \in \mathbb{R}^p$, iterate until convergence

$$w_{j+1} \leftarrow w_j - \eta_j \nabla F(w_j), \tag{1.3}$$

where $\eta_j > 0$ is the step-size prescribed for the jth step. Gradient descent with fixed step-size $\eta_j = \eta$ is guaranteed to converge to a minimizer of F under general conditions, such as if F is smooth (in the sense that F has Lipschitz gradient), convex, and has a finite lower bound. Gradient descent is a first-order iterative algorithm, where first-order means that it only requires computing gradients and not higher-order derivatives. A contributing reason to the feasibility of large-scale neural network training is that neural network optimization is particularly well suited for first-order optimization methods: for neural network prediction functions composed of layers as in (1.2), the gradient of the corresponding objective function $F(w) = \frac{1}{n}\sum_{i=1}^{n}(h(x_i; w) - y_i)^2$ with respect to the parameter vector w can be computed by the chain rule using algorithmic differentiation—a technique referred to as *back propagation* in the machine learning community. However, even a single gradient computation of the form $\nabla F(w) = \frac{1}{n}\sum_{i=1}^{n}\nabla(h(x_i; w) - y_i)^2$ becomes prohibitively expensive as the size of the training set n reaches multiple millions. More generally, when optimizing functions with "finite sum" form, $F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$,[1] a single gradient evaluation $\nabla F(w)$ requires the computation of all n component function gradients $\nabla f_i(w)$. In such settings, it is natural to consider drawing a random subset of component functions and using the gradient of the random batch of components as a computationally efficient surrogate for the full gradient. This is the template for *stochastic gradient descent*, which is described in detail blow.

Algorithm 1 Stochastic Gradient Descent

```
// Return: ŵ, an intended approximation to
// w* ∈ arg min_{w∈ℝ^p} F(w), where F = (1/n) Σ_{i=1}^n f_i
procedure Stochastic Gradient Descent
    Initialize point w^(0). Prescribed step-size schedule {η_t}_{t=1}^∞
    for t := 1 to T − 1 do
        Draw an index i_t uniformly at random from {1, 2, ..., n}
        Iterate w^(t+1) ← w^(t) − η_t ∇f_{i_t}(w^(t)).
    end for
    return ŵ = w^(T)
end procedure
```

It is important that the index i_t is chosen uniformly at random, in which case the random vector $\nabla f_{i_t}(w^{(t)})$ is an unbiased estimate for the full gradient $\nabla F(w^{(t)})$, meaning that $\mathbb{E}_{i_t}\nabla f_{i_t}(w^{(t)}) = \nabla F(w^{(t)})$. In practice, one often implements *minibatch* stochastic gradient descent, which is a compromise between full gradient descent and stochastic gradient descent where a batch of component gradient directions are averaged at each step, to

1 We assume $n \in \mathbb{N}$ is a finite number for simplicity and because it is most relevant for applications, but all results can be extended in theory to continuous parameterizations $F(w) = \int_s f_s(w)d\mu(s)$.

reduce the variance of the stochastic gradient estimate. For a comprehensive overview of stochastic gradient descent methods in large-scale optimization, we refer the reader to the comprehensive article [4].

1.1. Stochastic gradient descent: Background

1.1.1. Stochastic approximation

The idea for stochastic gradient descent appeared almost 70 years ago in a paper by Robbins and Monro [25]. Suppose that $M : \mathbb{R} \to \mathbb{R}$ is a function with a unique root we wish to identify. We do not have access to exact exact evaluations $M(w)$, but rather we can access at a given point w a random variable $N(w)$ such that $\mathbb{E}[N(w)] = M(w)$. Within this framework of *stochastic approximation*, Robbins and Monro proposed the following root-finding algorithm: fix w_0 and a decreasing step-size schedule $\{a_n\}_{n=0}^{\infty}$, then iterate

$$w_{n+1} = w_n - a_n N(w_n). \tag{1.4}$$

Blum [3] subsequently extended the algorithm to the multivariate setting. One recognizes the SGD Algorithm 1 as a special case of the Robbins–Monro root-finding algorithm via the correspondence $M(w) = \nabla F(w)$ and $N(w^{(t)}) = \nabla f_{i_t}(w^{(t)})$. Under certain assumptions akin to strong convexity, smoothness, and bounded noise, Robbins, Monro, and Blum showed that algorithm (1.4) converges with probability 1, provided the step-size schedule $\{a_n\}_{n=0}^{\infty}$ is chosen to decrease at a rate such that

$$\sum_{n=0}^{\infty} a_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} a_n^2 < \infty. \tag{1.5}$$

An important gap between the setting considered by Robbins and Monro and the application of stochastic gradient descent in large scale machine learning is in the model assumptions for the stochastic noise: Robbins–Monro convergence theory and the implied choice of step-sizes (1.5) assume that the stochastic noise on the observations is *uniformly bounded*, $\sup_\theta |N(\theta)| \leq N < \infty$. More realistic in the setting of large-scale machine learning is an *affine-variance* noise model, where the stochastic noise level is proportional to the size of the full gradient at any given point. Specifically, the affine-variance noise model is as follows: for parameters $\sigma_0, \sigma_1 > 0$,

$$\forall w \in \mathbb{R}^p : \mathbb{E}_i \left\| \nabla f_i(w) \right\|_2^2 \leq \sigma_0^2 + \sigma_1^2 \left\| \nabla F(w) \right\|^2. \tag{1.6}$$

We will come back to the discussion about the stochastic noise later on.

Stochastic gradient descent was recognized as a powerful algorithm for training artificial neural networks in 1960, when it was used to train one of the earliest neural networks—the Adaline network (Adaline stands for "adaptive linear unit") [29]. The proposal of Adaline came shortly after Rosenblatt invented the perceptron, widely considered the first artificial neural network. Following Adaline, (stochastic) gradient descent persisted as the de facto algorithm for training artificial neural networks due to its simplicity and ability to extend (using back propagation) to multilayered neural network architectures. The full power of artificial neural networks was not realized until around 2010, when increased computing power

from GPUs and distributed computing allowed the use of larger networks, which became known as "deep learning," and neural networks began winning prizes in image recognition contests, approaching human level performance on various tasks. Just in time for the advent of modern deep learning, researchers began to formalize the nonasymptotic theory for stochastic gradient descent—convergence to global minimizers in the convex setting [2,23] and to stationary points the nonconvex setting [9].

1.1.2. The Kaczmarz method for solving linear systems

Independent of Robbins and Monro's work in stochastic approximation, Stefan Kaczmarz proposed an iterative method (now called the Kaczmarz method) for solving linear equations [15]. Consider an overdetermined system of consistent linear equations, $Aw = b$. Denote the ith row (out of a total of n rows) of A by a_i, and let $w^{(0)}$ be an arbitrary initial approximation to the solution of $Aw = b$. Kaczmarz observed that w^*, the unique solution to the overdetermined consistent system, corresponds to the unique point in the intersection of the hyperplanes $S_i = \{w : \langle a_i, w\rangle = b_i\}$. He proposed an iterative projection algorithm for finding the point w^* whereby one cycles through the hyperplanes in their natural ordering, and projects the current estimate for w^* onto the subsequent subspace, until convergence. That is, starting from an initial guess w^0 and for $t = 1, 2, \ldots$, iterate

$$w^{(t+1)} = w^{(t)} + \frac{b_i - \langle a_i, w^{(t)}\rangle}{\|a_i\|^2} a_i; \quad i = t \bmod n. \tag{1.7}$$

The Kaczmarz method can be viewed as an instance of what is now referred to as the method of successive projections onto convex sets (POCS). In 1933, John von Neumann proved convergence of POCS in the case of two ($n = 2$) hyperplanes [27]; Halperin later extended von Neumann's convergence result to arbitrarily many hyperplanes [11]. Aronszajn [1] later provided an explicit rate of convergence for the case of two hyperplanes—the convergence rate is linear and depends explicitly on the angle between the two hyperplanes. Kayalar and Weinert [17] proved that Aronszajn's rate of convergence is sharp. This sharp analysis has proved difficult to extend beyond the case of two hyperplanes, as it is related to the difficulty of analyzing the product of more than two orthogonal projection operators, see, for example, [5]. The Kaczmarz method was rediscovered in image reconstruction in 1970, where (along with additional positivity constraints) it is called the algebraic reconstruction technique (ART) [10]. ART is used extensively in computed tomography and, in fact, was used in the first medical scanner [13].

Later, in the 1990s, several works, including [8,12], observed that the Kaczmarz algorithm (1.7) tended to converge more quickly and consistently if the algorithm was changed so that the rows are selected in a *random*, rather than cyclic order. In a seminal paper, Strohmer and Vershynin proved in 2007 that if the rows are drawn from a particular weighted random distribution, the Kaczmarz algorithm converges in expectation with a sharp linear convergence rate [26] depending on a condition number of the matrix A. Precisely, the randomized Kaczmarz method proposed by Strohmer and Vershynin is as follows (Algorithm 2):

Algorithm 2 Randomized Kaczmarz method

1: // Return: $\hat{w}$, an intended solution to $Aw = b$
2: **procedure** Randomized Kaczmarz algorithm
3: Initialize point $w^{(0)}$. Denote rows of A by $\{a_i\}_{i=1}^n$.
4: **for** $t := 1$ **to** $T - 1$ **do**
5: Draw a row a_{i_t}, where i_t is chosen from the set $\{1, 2, \ldots, n\}$ at random according to a weighted probability distribution such that $\text{Prob}(i_t = j) \propto \|a_j\|_2^2$.
6: Iterate $w^{(t+1)} \leftarrow w^{(t)} + \frac{b_{i_t} - \langle a_{i_t}, w^{(t)} \rangle}{\|a_{i_t}\|^2} a_{i_t}$.
7: **end for**
8: **return** $\hat{w} = w^{(T)}$
9: **end procedure**

Subsequently, the paper [22] recognized the randomized Kaczmarz method as a special case of stochastic gradient descent Algorithm 1 applied in the setting of linear regression, where the objective function is $F(w) = \|Aw - b\|_2^2 = \frac{1}{n} \sum_{i=1}^n n(\langle a_i, w \rangle - b_i)^2$, and implemented with *importance sampling* so that $\text{Prob}(i_t = j) \propto \|a_j\|_2^2$ and $\eta_{i_t} = \frac{\gamma}{\|a_{i_t}\|_2^2}$ to maintain unbiasedness of the stochastic gradient estimator for the full gradient. Extending Strohmer and Vershynin's analysis beyond the linear regression setting improves on a previous linear convergence rate of Bach and Moulines [2] to show that stochastic gradient descent Algorithm 1 enjoys a linear convergence rate under a general set of conditions including convexity and smoothness. Moreover, the convergence rate can be improved when component functions are allowed to be drawn from an importance sampling weighted distribution, as extended to neural networks in [16, 20].

1.2. Stochastic gradient descent: convergence theory

In this section, we will lay out the convergence theory for stochastic gradient descent precisely. Enforce the following conditions on a loss function of the form $F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w)$:

(a) each f_i is L-smooth: $\forall w, z, \|\nabla f_i(w) - \nabla f_i(z)\|_2 \leq L\|w - z\|_2$;

(b) each f_i is convex;

(c) F is μ-strongly convex.

Under these assumptions, the loss function F has a unique minimizer w^*, and SGD converges to this minimizer as follows [22].

Theorem 1. *Consider constant step-size $\eta \leq \frac{1}{\mu}$. Draw $w^{(0)}$ either as a random initial point or deterministically. Denote $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w^*)\|_2^2$. Under the stated assumptions, the expected error of the SGD Algorithm 1 satisfies*

$$\mathbb{E}\left\|w^{(t)} - w^*\right\|_2^2 \leq \left[1 - 2\eta\mu(1 - \eta L)\right]^t \mathbb{E}\left\|w^{(0)} - w^*\right\|_2^2 + \frac{\eta\sigma^2}{\mu(1 - \eta L)}.$$

The error expression has two terms, highlighting that the algorithm enjoys a linear convergence rate up to a so-called "region of confusion" of radius $\frac{\eta\sigma^2}{\mu(1-\eta L)}$. Optimizing the step-size η to balance the two error terms results in the following sharp convergence rate.

Corollary 1.1. *Enforce the assumptions of Theorem 1. Fix the constant step-size* $\eta = \frac{\mu\epsilon}{2\epsilon\mu L + 2\sigma^2}$. *Denote by* $\epsilon_0 = \mathbb{E}_{w^{(0)}}\|w^{(0)} - w^*\|_2^2$. *After* $T = 2\log(\epsilon_0/\epsilon)(\frac{L}{\mu} + \frac{\sigma^2}{\mu^2\epsilon})$ *iterations, the expected error satisfies*

$$\mathbb{E}\left\|w^{(T)} - w^*\right\|_2^2 \leq \epsilon.$$

While this convergence rate cannot really be improved in the setting where the step-size is fixed, we can improve on this rate slightly by considering a carefully chosen piecewise constant decreasing step-size schedule and applying Corollary (1.1) recursively. We could not find the following result stated explicitly in the literature, so we provide the short proof.

Proposition 1.1. *Enforce the assumptions on smoothness and convexity from Theorem 1. For error function* $h(s) = \frac{\mu s}{2s\mu L + 2\sigma^2}$ *and times*

$$T_J = 2\left(\frac{L}{\mu} + \frac{2^J\sigma^2}{\mu^2\mathbb{E}\|w^{(0)} - w^*\|_2^2}\right), \quad J = 1, 2, \ldots$$

consider the SGD Algorithm 1 with piecewise constant decreasing step-size schedule

$$\eta_t = \eta_1 := h(\epsilon_0 \cdot 2^{-1}), \quad 1 \leq t \leq T_1,$$
$$\eta_t = \eta_J := h(\epsilon_0 \cdot 2^{-J}), \quad 1 + \sum_{j=1}^{J-1} T_j \leq t \leq \sum_{j=1}^{J} T_j,$$

After $T_{(K)} = 2K(\frac{L}{\mu} + \frac{2}{K} \cdot \frac{\sigma^2}{\mu^2\mathbb{E}\|w^{(0)} - w^*\|_2^2}2^K)$ *iterations,*

$$\mathbb{E}\left\|w^{(T_{(K)})} - w^*\right\|_2^2 \leq 2^{-K}\mathbb{E}\left\|w^{(0)} - w^*\right\|_2^2.$$

Comparing the error bounds in Proposition 1.1 and Corollary 1.1, we see that to achieve error $\mathbb{E}\|w^{(T_{(K)})} - w^*\|_2^2 \leq 2^{-K}\mathbb{E}\|w^{(0)} - w^*\|_2^2$, the piecewise constant decreasing step-size schedule requires a number of iterations $T = 2K(\frac{L}{\mu} + \frac{2}{K} \cdot \frac{\sigma^2}{\mu^2\epsilon_0}2^K)$ while the constant step-size schedule requires a larger number of iterations $T' = 2K(\frac{L}{\mu} + \frac{\sigma^2}{\mu^2\epsilon_0}2^K)$. This suggests that to get the best possible convergence rate of SGD when the region of confusion dominates the condition number, piecewise constant decreasing step-size schedules can outperform constant step-size schedules.

Proof of Proposition 1.1. Theorem 1.1 is proved by induction on the bound in Corollary 1.1 with the number of levels K. For the base case $K = 1$, we get the result by applying Corollary 1.1 with $\epsilon_1 = \epsilon_0/2$ and fixed step-size $\eta_1 = h(\epsilon_0/2)$. For the induction, suppose the result holds at $K - 1$, that is, suppose that $\mathbb{E}\|w^{(T_{(K-1)})} - w^*\|_2^2 \leq \epsilon_{K-1}$, where $\epsilon_K := 2^{-K}\mathbb{E}\|w^{(0)} - w^*\|_2^2$. Apply Corollary 1.1 with $\epsilon_0 = \epsilon_{K-1}$ and $\epsilon = \epsilon_0/2$ and with $\eta_K = h(\epsilon_0 \cdot 2^{-K})$ to arrive at the stated bound at K. ∎

We draw the reader's attention to the fact that we have focused on stochastic gradient descent convergence theory under assumptions such as smoothness and convexity which

are not satisfied in the setting of training neural networks. However, increasingly, neural networks are implemented to be highly *overparameterized*, or configured so that the number of parameters p (length of the parameter vector w) is set to be larger than the size of the training data. In this regime, recent works have shown that in a certain "neural tangent kernel regime," the loss function associated to training overparameterized neural networks is locally strongly convex around a random initialization $w^{(0)}$ [6,14]. While it is an active area of research to try and understand the extent to which overparameterized neural networks remain similar to linear systems in regimes where neural networks are most powerful in practice, there is evidence that points to a strong connection. One important piece of evidence is the fact that SGD is typically trained using piecewise constant decreasing step-sizes to optimize convergence speed, just as suggested by Proposition 1.1. Thus, the convergence theory for SGD in the strongly convex setting (and the corresponding step-size schedule which results for optimizing convergence) is surprisingly relevant in the application of training large-scale neural networks.

1.3. Adaptive step-size rules in stochastic gradient descent

Proposition 1.1 suggests that in training neural networks using stochastic gradient descent, piecewise constant decreasing step-sizes should be effective. In practice, neural networks are indeed trained using piecewise constant decreasing step-size schedules; however, the particular choice of step-size schedule in Proposition 1.1 is not so useful in practice as it is a function of several parameters of the optimization problem: the strong convexity parameter $\mu > 0$, the Lipschitz smoothness constant L associated to the loss function, the stochastic noise level $\sigma^2 > 0$, and the error at initialization $\|w^{(0)} - w^*\|_2^2$. In practice, none of these quantities is known to the user in advance. Indeed, this represents a serious disconnect between the theory for SGD and the practical implementation, as the convergence behavior of the basic SGD Algorithm 1 is quite sensitive to the choice of step-size schedule. Fortunately, simple modifications to the basic SGD algorithm have been developed, such as Adagrad [7,21], RMSprop, and Adam [18], which are significantly more robust to the step-size schedule. A convergence theory for these algorithms as adaptive step-size learners in the setting of stochastic gradient descent was initiated independently in [19,28]. We will focus on the results from [28], which focuses on guarantees for the AdaGrad adaptive gradient algorithm.

As a precursor to discussing adaptive gradient methods in the context of stochastic gradient descent, let us first understand their behavior in the setting of batch (full) gradient descent (where where the gradients $\nabla F(w)$ are measured exactly).[2] In the batch setting, the AdaGrad algorithm is as follows.

2 We note that in the batch setting, line search methods are efficient black-box plugins for adaptively updating the step-size. However, such methods lose effectiveness in the presence of stochastic noise, and have a tendency to overfit the noisy gradient directions.

Algorithm 3 Gradient Descent with AdaGrad

// Return: $\hat{w}$, an intended approximation to
// $w^* \in \arg\min_{w \in \mathbb{R}^p} F(w)$
procedure Gradient Descent with AdaGrad
    Initialize point $w^{(0)}$, initial step-size parameters $b_0, \eta > 0$. Tolerance $\epsilon > 0$.
    repeat
        $t + 1 \leftarrow t$.
        Update step-size $b_t^2 = b_{t-1}^2 + \|\nabla F(w^{(t-1)})\|^2$

        Iterate $w^{(t)} \leftarrow w^{(t-1)} - \frac{\eta}{b_t} \nabla F(w^{(t-1)})$.
    until $\|\nabla F(w^{(t)})\|^2 \leq \epsilon$
    return $\hat{w} = w^{(T)}$
end procedure

To put our main result in context, let us first review the following classical result (see, for example, [24, (1.2.13)]) on the convergence rate for gradient descent with fixed step-size.

Lemma 1.1. *Suppose that F is L-smooth, and suppose that $F^* = \inf_x F(x) > -\infty$. Fix η and b, consider gradient descent, $w^{(t+1)} = w^{(t)} - \frac{\eta}{b} \nabla F(w^{(t)})$. If $b \geq \eta L$, then*

$$\min_{t=0:T-1} \left\| \nabla F(w^{(t)}) \right\|^2 \leq \varepsilon$$

after at most a number of steps

$$T = \frac{2b(F(w^{(0)}) - F^*)}{\eta \varepsilon}.$$

Alternatively, if $b \leq \frac{\eta L}{2}$, then convergence is not guaranteed at all—gradient descent can oscillate or diverge.

The following result on the convergence of AdaGrad Algorithm 3 from [28] shows that in contrast to fixed step-size gradient descent, AdaGrad always converges, and its convergence rate as a function of the parameters $b_0, \eta > 0$ can be understood in a sharp sense. It suggests that in practice, one should simply initialize AdaGrad with a large step-size $1/b_0$, and the algorithm will adapt on its own by decreasing the step-size to an appropriate limiting value.

Theorem 2 (AdaGrad—convergence). *Consider the AdaGrad Algorithm 3. Suppose that F is L-smooth and suppose that $F^* = \inf_w F(w) > -\infty$. Then*

$$\min_{t=0:T-1} \left\| \nabla F(w^{(t)}) \right\|^2 \leq \varepsilon$$

after ***Case 1:*** $T = 1 + \lceil \frac{2(F(w^{(0)})-F^*)(b_0+2(F(w^{(0)})-F^*)/\eta)}{\eta\varepsilon} \rceil$ *steps if* $\frac{b_0}{\eta} \geq L$, *and*

Case 2: $T = 1 + \lceil \frac{(\eta L)^2-b_0^2}{\varepsilon} + \frac{4((F(w^{(0)})-F^*)/\eta+(\frac{3}{4}+\log\frac{\eta L}{b_0})\eta L)^2}{\varepsilon} \rceil$ *steps if* $\frac{b_0}{\eta} < L$.

In either case, $\max_t \frac{b_t}{\eta} \leq \frac{b_{\max}}{\eta}$ *where* $b_{\max}/\eta = 2L(1+\log(\eta L/b_0)) + \frac{2}{\eta^2}(F(w^{(0)}) - F^*)$.

Comparing the convergence rate of AdaGrad with the convergence rate of gradient descent with fixed step-size, we see that in case $b = b_0 \geq \eta L$, the rates are essentially the same. But in case $b = b_0 < \eta L$, gradient descent can fail to converge as soon as $b \leq \eta L/2$, while AdaGrad converges for any $b_0 > 0$, and is extremely robust to the choice of $b_0 < \eta L$ in the sense that the resulting convergence rate remains close to the optimal rate of gradient descent with fixed step-size $\eta/b = 1/L$, paying only a factor of $\log(\frac{\eta L}{b_0})$ in the constant.

The convergence rate in Theorem 2 represents a worst-case analysis of AdaGrad over the class of L-smooth functions. In practice, the limiting step-size will obtain very quickly, and at a value much *larger* than $1/L$. This is not surprising since the smoothness parameter L represents only the globally worst-case bound on the magnitude of the ratio $\frac{\|\nabla F(w)-\nabla F(z)\|}{\|w-z\|}$ over all $w, z \in \mathbb{R}^p$. In other words, even if one has a priori bound on L, AdaGrad can converge significantly faster than gradient descent with fixed step-size $1/L$, and is thus advantageous to use even with such knowledge.

Now let us turn to the convergence analysis of AdaGrad in the stochastic setting, also from [28]. Recall that in the stochastic setting, instead of observing a full gradient at each iteration, we observe a stochastic gradient $g_t \in \mathbb{R}^p$ which is an unbiased estimator for the true gradient $\nabla F(w^{(t)})$.

We now state Adagrad in the stochastic setting (Algorithm 4):

Algorithm 4 Stochastic Gradient Descent with AdaGrad

1: `//Return` w^*, `an approximation to a stationary point of a smooth function` $F(\cdot)$ `over` $\mathbb{R}^p$.
2: **procedure** AdaGrad in Stochastic Setting
3: Initial point $w^{(0)} \in \mathbb{R}^p$. step-size parameters η, b_0.
4: **for** $t := 1$ **to** $T - 1$ **do**
5: step-size update:

$$\eta_t = \frac{\eta}{\sqrt{b_{t-1}^2 + \|g_t\|^2}} \quad \text{where } b_t^2 = b_{t-1}^2 + \|g_t\|^2$$

6: Iterate $w^{(t+1)} \leftarrow w^{(t)} - \eta_t g_t$.
7: **end for**
8: **return** $\hat{w} = w^{(T)}$
9: **end procedure**

More formally, denote

$$\mathcal{F}_t = \sigma\left\{w^1, g^1, \ldots, w^{(t)}, g^{(t)}, w^{(t+1)}\right\} \tag{1.8}$$

to be the sigma algebra generated by the observations of the algorithm after observing the first t stochastic gradients. We will assume that the stochastic gradients satisfy the following:

Assumptions 2.1 (Unbiased gradients). *For each time t, the stochastic gradient, g_t, is an unbiased estimate of $\nabla F(w^{(t)})$, i.e.,*

$$\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla F(w^{(t)}). \tag{1.9}$$

For the theory in this section, we will assume that the stochastic noise is uniformly bounded, as in the setting of Robbins and Monro.[3]

Assumptions 2.2 (Uniformly bounded gradient and uniformly bounded variance). *We assume $\sup_{w\in\mathbb{R}^p} \|\nabla F(w)\| \leq \gamma$. Moreover, for each time t, the variance satisfies*

$$\mathbb{E}\left[\left\|g_t - \nabla F(w^{(t)})\right\|^2 |\mathcal{F}_{t-1}\right] \leq \sigma^2. \tag{1.11}$$

The AdaGrad step-sizes $\frac{\eta}{b_t}$ in the stochastic setting exhibit quite different behavior than in the deterministic setting. Rather than converging to a fixed value proportional to the Lipschitz smoothness constant as in the batch setting, the step-size decreases to zero in the stochastic setting, roughly at the rate of $\frac{1}{b_t} \approx \frac{1}{\sigma\sqrt{t}}$. This rate is optimal in t in terms of the resulting convergence theorems in the setting of smooth but not necessarily convex F, or convex but not necessarily strongly convex or smooth F. Still, one must be careful with convergence theorems for AdaGrad because the step-size is a random variable and dependent on all previous points visited along the way.

Theorem 3. *Suppose F is L-smooth and $F^* = \inf_w F(w) > -\infty$. Suppose that the random variables $g_t, t \geq 0$, satisfy the above assumptions. Then with probability $1-\delta$,*

$$\min_{t\in[T-1]} \left\|\nabla F(w^{(t)})\right\|^2 \leq \left(\frac{2b_0}{T} + \frac{2\sqrt{2}(\gamma+\sigma)}{\sqrt{T}}\right)\frac{Q}{\delta^{3/2}}$$

where $Q = \frac{F(w^{(0)})-F^*}{\eta} + \frac{4\sigma+\eta L}{2}\log(\frac{20T(\gamma^2+\sigma^2)}{b_0^2} + 10)$.

This result implies that AdaGrad converges starting from any value of b_0 for a given η. To put this result in context, we can compare to Corollary 2.2 of [9], which implies that under similar assumptions, if the Lipschitz constant L and the variance σ are known a priori, and the step-size is

$$\eta_t = \eta = \min\left\{\frac{1}{L}, \frac{1}{\sigma\sqrt{T}}\right\}, \quad t = 0, 1, \ldots, T-1,$$

---

3 These assumptions can be weakened to a single affine-variance assumption: For each time t, the variance only needs to satisfy

$$\mathbb{E}\left[\left\|g_t - \nabla F(w^{(t)})\right\|^2 |\mathcal{F}_{t-1}\right] \leq \sigma_0^2 + \sigma_1^2\|\nabla F(w^{(t)}\|^2. \tag{1.10}$$

then with probability $1-\delta$,

$$\min_{\ell\in[T-1]} \left\|\nabla F(w^{(t)})\right\|^2 \leq \frac{2L(F(w^{(0)})-F^*)}{T\delta} + \frac{(L+2(F(w^{(0)})-F^*))\sigma}{\delta\sqrt{T}}.$$

Thus, essentially, AdaGrad convergence achieves the rate of [9], but without requiring a priori knowledge of L and σ to set the step-sizes. The constant in the $O(1/\sqrt{T})$ rate of AdaGrad scales according to σ^2 (up to a logarithmic factors in σ) while the results with well-tuned step-size scales linearly with σ.

The main technical difficulty in the proof of Theorem 3 is in dealing with the AdaGrad step-sizes which are random variables which depend on the current and all previous stochastic gradients. See [28] for details of the proof.

1.4. Outlook

Stochastic gradient descent is the de facto algorithm used for minimizing functions which arise in deep learning and neural network training. While there are many mysteries surrounding the behavior of stochastic gradient descent in applications, there are also several regimes in which we have a rich and sharp mathematical understanding. Remarkably, the strong linear convergence guarantees for stochastic gradient descent which are guaranteed in the setting of strongly convex finite sums (and the piecewise constant decreasing step-size rules they imply) are empirically verified in practice in training overparameterized neural networks. That is, in many practical settings, seemingly highly nonconvex and highly nonlinear neural network-based regression functions of interest are in reality perturbations of linear regression problems. In this sense, practice caught up with theory. A theoretical understanding of practical methods for making stochastic gradient descent more robust to hyperparameter specifications such as the step-size schedule has begun to emerge in recent years. In this sense, stochastic gradient enhancements developed in practice to meet the needs of large-scale machine learning inspired new theoretical directions in the study of stochastic gradient descent.

FUNDING

This work was partially supported by AFOSR MURI FA9550-19-1-0005, NSF DMS 1952735, NSF HDR-1934932, and NSF 2019844.

REFERENCES

[1] L. Aronszajn, Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** (1950), 337–404.

[2] F. Bach and E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in neural information processing systems. Vol. 24*, 2011.

[3] J. R. Blum, Approximation methods which converge with probability one. *Ann. Math. Stat.* (1954), 382–386.

[4] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning. *SIAM Rev.* **60** (2018), no. 2, 223–311.

[5] F. Deutsch, The rate of convergence for the method of alternating projections, II. *J. Math. Anal. Appl.* **205** (1997), 381–405.

[6] S. S. Du, X. Zhai, B. Poczos, and A. Singh, Gradient descent provably optimizes over-parameterized neural networks. In *International conference on learning representations*, 2018.

[7] J. Duchi, E. Hazan, and Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** (2011), 2121–2159.

[8] H. G. Feichtinger, C. Cenker, M. Mayer, H. Steier, and T. Strohmer, New variants of the POCS method using affine subspaces of finite codimension with applications to irregular sampling. In *Visual Communications and Image Processing'92*, pp. 299–310, Proc. SPIE 1818, International Society for Optics and Photonics, 1992.

[9] S. Ghadimi and G. Lan, Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.* **23** (2013), no. 4, 2341–2368.

[10] R. Gordon, R. Bender, and G. T. Herman, Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and X-ray photography. *J. Theoret. Biol.* **29** (1970), no. 3, 471–481.

[11] I. Halperin, The product of projection operators. *Acta Sci. Math. (Szeged)* **23** (1962), 96–99.

[12] G. T. Herman and L. B. Meyer, Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application). *IEEE Trans. Med. Imag.* **12** (1993), no. 3, 600–609.

[13] G. N. Hounsfield, Computerized transverse axial scanning (tomography): Part 1. description of system. *Br. J. Radiol.* **46** (1973), no. 552, 1016–1022.

[14] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, p. 6, ACM, 2021.

[15] S. Karczmarz, Angenaherte Auflosung von Systemen linearer Gleichungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.* (1937), 355–357.

[16] A. Katharopoulos and F. Fleuret, Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pp. 2525–2534, PMLR, 2018.

[17] S. Kayalar and H. Weinert, Error bounds for the method of alternating projections. *Math. Control Signals Systems* **1** (1988), 43–59.

[18] D. Kingma and J. Ba. Adam, A method for stochastic optimization. 2014, arXiv:1412.6980.

[19] X. Li and F. Orabona, On the convergence of stochastic gradient descent with adaptive stepsizes. 2018, arXiv:1805.08114.

[20] I. Loshchilov and F. Hutter, Online batch selection for faster training of neural networks. 2015, arXiv:1511.06343.

[21] H. B. McMahan and M. Streeter, Adaptive bound optimization for online convex optimization. In *COLT 2010*, p. 244, 2010.

[22] D. Needell, R. Ward, and N. Srebro, Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in neural information processing systems*, 2014.

[23] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** (2009), no. 4, 1574–1609.

[24] Y. Nesterov, *Introductory lectures on convex programming volume I: Basic course*. 1998.

[25] H. Robbins and S. Monro, A stochastic approximation method. *Ann. Math. Stat.* **22** (1951), 400–407.

[26] T. Strohmer and R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.* **15** (2009), no. 2, 262–278.

[27] J. von Neumann, Functional operators. vol. II. The geometry of orthogonal spaces. *Ann. of Math. Stud.* **22** (1950). This is a reprint of mimeographed lecture notes first distributed in 1933.

[28] R. Ward, X. Wu, and L. Bottou, AdaGrad stepsizes: sharp convergence over nonconvex landscapes. In *International conference on machine learning*, pp. 6677–6686, PMLR, 2019.

[29] B. Widrow, An adaptive "Adaline" neuron using chemical "memistors". Technical report No. 1553-2, 1960.

RACHEL WARD

2515 Speedway, Austin, TX 78712, USA, rward@math.utexas.edu

SOLVING INVERSE PROBLEMS WITH DEEP LEARNING

LEXING YING

ABSTRACT

We discuss some recent work on applying deep learning to inverse problems. On the algorithmic side, we propose two new neural network modules, BCR-Net and Switch-Net, motivated by pseudodifferential and Fourier integral operators that commonly appear in the study of inverse problems. On the application side, we propose neural networks for inverse maps in five applications: electric impedance tomography, optical tomography, inverse acoustic scattering, seismic imaging, and traveltime tomography. In each application, the architecture is motivated by perturbation theory and filtered backprojection, and is implemented using the new modules along with convolution layers. When translation and rotation equivariances are available, appropriate reparameterizations in the data and model domains result in convolutional architectures that are both general and effective. These applications demonstrate that our approach provides a seamless way for combining the mathematical structure of the inverse problems with the power of deep neural networks.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 65N21; Secondary 65R32, 74J25, 86A22

KEYWORDS

Inverse problems, neural networks, deep learning

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5154–5175
DOI 10.4171/ICM2022/49

1. INTRODUCTION

In the past decade, deep learning (DL) has become the dominant approach in computer vision, image processing, speech recognition, and many other applications in machine learning and data science [26, 34, 40, 42, 43, 48, 58, 60]. From a technical point of view, this success is a synergy of several key developments: (1) deep neural networks (NNs) as a flexible framework for representing high-dimensional functions and maps, (2) simple algorithms such as backpropagation (BP) and stochastic gradient descent (SGD) for tuning the model parameters, (3) efficient general software packages such as Tensorflow [1] and Pytorch [52], and (4) unprecedented computing power provided by GPUs and TPUs. Despite the successes, however, there remain a number of outstanding challenges: (1) NN architectural design is still an art and lacks basic mathematical principles in many cases; (2) NN training often requires an enormous amount of data, which is infeasible in many applications; and (3) a general mathematical theory of deep learning is still lacking.

Many computational problems in physical sciences face the same challenges as those in data science: high-dimensionality, complex or unspecified models, and high computational costs. Some well-known examples include many-body quantum systems, deterministic and stochastic control, molecular dynamics, uncertainty quantification, inverse problems, etc. There is a clear opportunity to leverage the recent developments of DL in the study of these problems. Indeed, the past few years have witnessed a rise of activities in this direction [2, 5, 7, 8, 12, 16, 18, 19, 30, 31, 35–37, 41, 44, 46, 47, 54, 55, 57, 61].

Among these topics, this paper focuses on inverse problems, i.e., recovering unknown interior parameters from boundary measurements. It is a field of enormous importance, with applications in physics, chemistry, medicine, earth sciences, etc. From a computational perspective, many inverse problems are quite challenging for several well-understood reasons: (1) the inverse map from boundary measurements to interior parameters is high-dimensional and nonlinear; (2) asymptotic methods based on perturbation theory often have low accuracy, while fully optimization-based iterative algorithms are often time-consuming; (3) most solution methods are not designed to adapt to data priors, when they are available.

Contributions. We argue that applying deep learning to the study of inverse problems is a fruitful mathematical research direction. On the one hand, NNs offer a flexible tool for representing the high-dimensional inverse maps. They also learn from the data distribution prior effectively via training. On the other hand, the rich mathematical and physical theories behind inverse problems provide guiding principles for designing compact, yet effective NN architectures. As a result, we avoid the need for enormous amounts of data, which are often not available for inverse problems.

The main contributions of this line of study are two-fold. On the *algorithmic* side, we first identify the mathematical operators commonly used in inverse problems, with two such examples being pseudodifferential operators (PDOs) and Fourier integral operators (FIOs) [59]. By leveraging analytical results from partial differential equation (PDE) theory and numerical linear algebra (NLA), we propose novel NN modules for these key types of operators.

On the *application* side, we apply this approach to five different inverse problems: electric impedance tomography, optical tomography, inverse acoustic scattering, seismic imaging, and traveltime tomography. For each application, using the linearized theory and perturbative expansion as a starting point, we approximate the inverse map with a composition sequence of operators. The NN for the inverse map is then assembled using the corresponding modules, along with existing primitives such as convolution neural networks (CNNs). Finally, the weights of the whole network are trained end-to-end with the available training data.

Organization. The rest of the paper is organized as follows. Section 2 describes new NN modules motivated by PDOs and FIOs. Section 3 details the five inverse problems. Finally, Section 4 concludes with a discussion of future directions.

2. NEW, MATHEMATICALLY-MOTIVATED NN MODULES

If one takes a close look at the successful NN architectures in the literature, it is not hard to see that behind each there is a powerful mathematical structure, tabulated as follows.

NN architecture	Mathematical structure
Fully-connected layer	Dense operator
Convolution layer	Translation-invariant local operator
Recurrent neural network (RNN)	Markov chain
ResNet	ODE/time-stepping/semigroup

For the inverse problem theory, two types of commonly occurring operators are pseudodifferential operators (PDOs) and Fourier integral operators (FIOs). In this section, we propose novel NN modules for efficient and accurate representations of these two types of operators.

2.1. Pseudodifferential operators

A pseudodifferential operator (PDO) K is of the form

$$(Kf)(x) \equiv \int k(x,y) f(y) \mathrm{d}y = \int a(x,\xi) e^{2\pi i x\cdot\xi} \hat{f}(\xi) \mathrm{d}\xi, \tag{2.1}$$

where the *symbol* function $a(x,\xi)$ of the PDO is smooth away from the origin of the frequency variable ξ [59]. PDOs are powerful generalizations of standard differential operators. When applied to a function f, the support of the singularities in the output Kf is contained in the singularity support of the input. Some well-known examples of PDOs include the Green's functions of elliptic operators, fractional Laplacians, etc. When a PDO is translation-equivariant, it becomes a convolution and thus can be represented with a convolution layer,

though this representation is often not effective for highly nonlocal PDOs. More importantly, non-translation-equivariant PDOs cannot be represented using convolution layers.

One of the key properties of PDOs is that, when discretized with local basis functions, the off-diagonal blocks of the matrix form of a PDO are numerically low-rank [28, 29]. This property gives rise to highly effective data-sparse approximations to PDOs, and the one adopted here is based on wavelet analysis. Motivated by this approximation, we propose a novel NN module associated with PDOs by

- representing the data-sparse approximation of PDOs as a (linear) NN,
- enriching its representation power by including intermediate layers and nonlinearities such as the ReLU activation function.

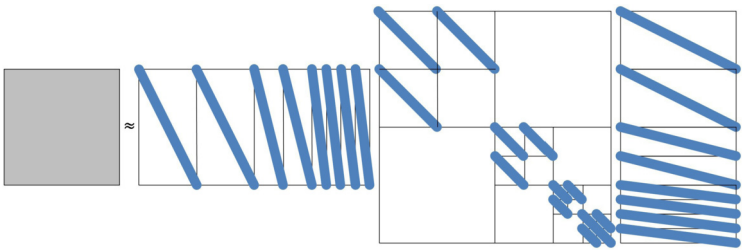


FIGURE 1
The nonstandard wavelet form of a PDO. The first and third matrices on the right-hand side are the inverse and forward transforms for the redundant wavelet/scaling function frame, which can be implemented with fast wavelet transforms in linear complexity. The large middle matrix represents the PDO under this redundant frame, which has a well-defined sparsity pattern with only $O(n)$ nonzero entries.

Wavelet analysis. The data-sparse approximation is based on the nonstandard wavelet form proposed in [9]. Given an $n \times n$ matrix form of a PDO K, the nonstandard form represents the operator in the redundant wavelet/scaling function frame and keeps only $O(n)$ significant coefficients in a well-defined sparsity pattern. Figure 1 illustrates the sparsity pattern, shown in blue.

When applying this PDO to an input vector, the matrix–vector multiplication (MatVec) at each wavelet scale can be written as a three-layer NN with two channels in the middle (see Figure 2 (left) for an illustration). Putting together the networks across all scales gives rise to a linear NN shown in Figure 2 (right).

In order to represent nonlinear operators similar to PDOs, we propose generalizing the architecture in Figure 2 by inserting multiple intermediate layers and including nonlinear activations such as the ReLU function. This results in a new NN module called a *BCR-Net* [17] as shown in Figure 3.

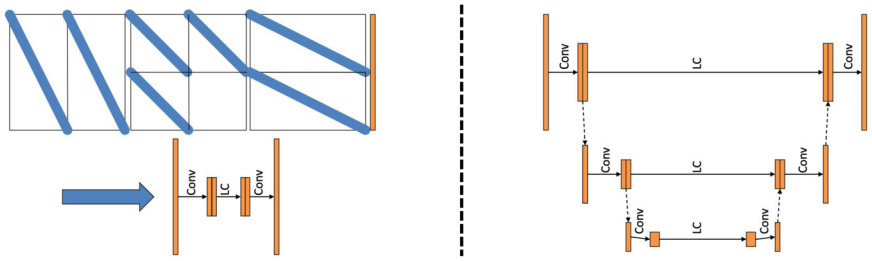


FIGURE 2

A matrix–vector multiplication (MatVec) with an input vector. (Left) The computation at each scale of the wavelet-based data-sparse approximation is a three-layer NN with two channels in the middle. (Right) The NN obtained by merging across all scales. Conv and LC stand for convolution and locally-connected layer, respectively.

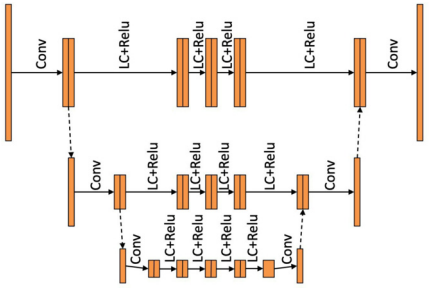


FIGURE 3

The BCR-Net module based on the nonstandard redundant wavelet form for PDOs.

2.2. Fourier integral operators

A Fourier integral operator (FIO) K is of the form

$$(Kf)(x) \equiv \int k(x,y)f(y)\mathrm{d}y = \int a(x,\xi)e^{2\pi i\Phi(x,\xi)}\hat{f}(\xi)\mathrm{d}\xi, \tag{2.2}$$

where the *amplitude* $a(x,\xi)$ of the FIO is smooth away form the origin of ξ and the *phase* $\Phi(x,\xi)$ is homogeneous of degree one in ξ. Viewed as a map from the frequency to the spatial domain (i.e., $\hat{f}$ to Kf), FIOs are generalizations of the Fourier transforms with more general phase and amplitude functions. When applied to a function f, the support of the singularities in the output Kf depends on the input singularities in a well-defined way governed by the Hamiltonian flow of the phase function Φ [59]. Most examples of the FIOs appear in high-frequency wave propagations and scattering theory, and it is for this reason that they are often key to solving wave-based inverse problems.

One key property of FIOs is that, when they are discretized with local basis functions, the resulting matrix representation satisfies the so-called *complementary low-rank property* [45]. More precisely, when the $n \times n$ matrix is partitioned into $\sqrt{n} \times \sqrt{n}$ blocks each of size $\sqrt{n} \times \sqrt{n}$, each block is numerically low-rank. This property allows for an effi-

cient data-sparse approximation, the *butterfly factorization*, to be detailed below. Motivated by the butterfly factorization, we propose a new NN module associated with FIOs by

- representing the butterfly factorization of FIOs as a (linear) NN, and
- enriching its representation power by including intermediate layers and nonlinear activations, such as the ReLU function.

Butterfly factorization. Given an $n \times n$ matrix representation of an FIO K, the simplest form of butterfly factorization partitions the whole matrix into $\sqrt{n} \times \sqrt{n}$ blocks and then computes a low-rank approximation of each block. Figure 4 demonstrates that the low-rank approximations for all blocks can be summarized compactly as the product of three sparse matrices. Notice that the second matrix of the factorization serves essentially as a permutation.

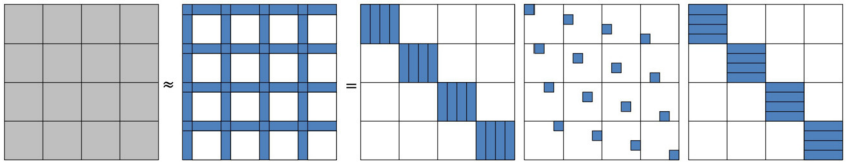


FIGURE 4
Butterfly factorization of an FIO. The middle plot shows the numerical low-rank properties of each $\sqrt{n} \times \sqrt{n}$ block. On the right, the first and third matrices collect the low-rank bases, while the second matrix essentially performs a permutation.

When applying the FIO to an input vector, the MatVec (as shown in Figure 5) can be represented as a three-layer linear NN. Here the first and third matrices become a convolution or locally-connected layer with $\sqrt{n}$ channels, while the second matrix can be implemented with a transpose.

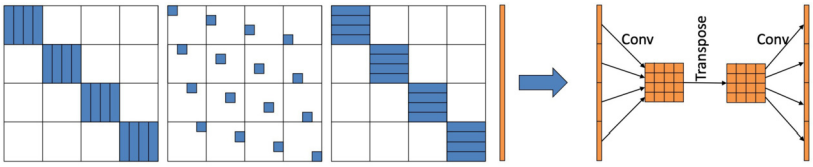


FIGURE 5
A matrix–vector multiplication (MatVec) with an input vector. The computation is represented by a three-layer NN with a transpose operation in the middle. C/LC stands for a convolution or locally-connected layer.

In order to represent nonlinear operators similar to FIOs, we generalize the architecture in Figure 5 by inserting multiple intermediate layers and including nonlinear activations (e.g., ReLU). The resulting new NN module, shown in Figure 6, is called a *Switch-Net* [38].

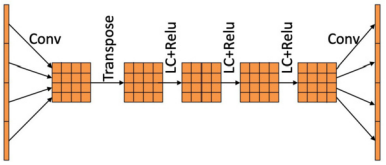


FIGURE 6
The Switch-Net module based on the butterfly factorization for FIOs.

3. INVERSE PROBLEMS

This section describes how to apply deep learning to five inverse problems: electrical impedance tomography, optical tomography, inverse acoustic scattering, seismic imaging, and traveltime tomography. For each problem, we proceed as follows:

- describe the basic setup,
- represent the linearized inverse map as a sequence of operators by following the perturbation theory and filtered backpropagation,
- design the NN architecture by following this sequence and using the new modules in Section 2 as well as CNNs.

Throughout this process, we keep in mind several guiding principles:

- the NN design should adapt to the data collection geometry,
- pre- and post-processing often significantly simplify the NN design, and
- preserving equivariances is the key to efficiency and accuracy.

3.1. Electrical impedance tomography

Consider a rectangular domain Ω (see Figure 7 (left)) with top boundary denoted by Γ. To simplify the presentation, we assume a periodic boundary condition in the horizontal direction. One form of the governing equation for electrical impedance tomography (EIT) is the elliptic equation

$$(Lu)(p) \equiv \big(-\Delta - \eta(p)\big)u(p) = 0, \quad p \in \Omega, \tag{3.1}$$

where we often denote $p = (x, z)$, with x and z being the horizontal and vertical components, respectively. Here $\eta(p)$ is the unknown internal parameter field. In one common form of an EIT experiment, for each boundary point $s \in \Gamma$, we enforce the delta boundary condition $\delta_s(\cdot)$ and then record the normal derivative $\frac{\partial u^s(r)}{\partial n(r)}$ at every point $r \in \Gamma$, where $u^s(\cdot)$ denotes the solution of (3.1) induced by the boundary condition $\delta_s(\cdot)$. The set $d(r, s)$ of boundary measurements is

$$d(r,s) = \frac{\partial u^s(r)}{\partial n(r)} - \frac{\partial u_0^s(r)}{\partial n(r)}, \tag{3.2}$$

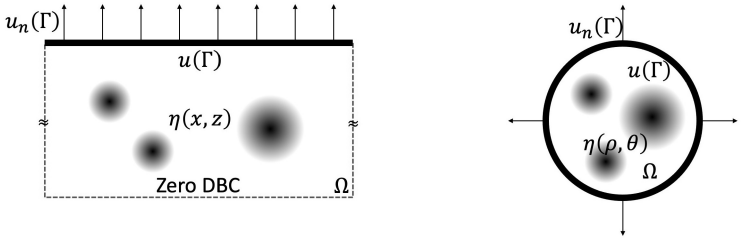


FIGURE 7
Electrical impedance tomography: (left) an experimental setup for a rectangular geometry; (right) an experimental setup in a circular geometry.

where $u_0^s(\cdot)$ stands for the background solution when $\eta(p) \equiv 0$. In technical terms, $d(r,s)$ is the kernel of the *Dirichlet-to-Neumann map* of (3.1). The inverse problem is to recover $\eta(p) \equiv \eta(x,z)$ from $d(r,s)$.

In order to obtain an approximation to the inverse map $d(r,s) \to \eta(x,z)$, we first study how $d(r,s)$ depends on $\eta(x,z)$ in the perturbative regime. Let L_0 be the operator with $\eta(x,z) \equiv 0$ and $G_0 = L_0^{-1}$ be its Green's function. A perturbative analysis in [23] shows that, when η is small, the data $d(r,s)$ can be well approximated with

$$d(r,s) \approx \iint_{(x,z)} \frac{\partial G_0}{\partial n}\big(r,(x,z)\big)\frac{\partial G_0}{\partial n}\big(s,(x,z)\big)\eta(x,z)\mathrm{d}x\mathrm{d}z. \tag{3.3}$$

Due to the translation-equivariance of the background operator L_0, this equation can be simplified when the data d is written in a warped coordinate system (m,h) with $(r,s) \equiv (m+h, m-h)$, namely

$$d(m,h) \approx \iint_{(x,z)} \frac{\partial G_0}{\partial n}\big((m-x)+h,z\big)\frac{\partial G_0}{\partial n}\big((m-x)-h,z\big)\eta(x,z)\mathrm{d}x\mathrm{d}z. \tag{3.4}$$

The key observation is that this is a 1D convolution in m and x with h and z treated as parameters (or as channels in the NN terminology). Furthermore, due to the elliptic nature of the EIT problem, the forward map between η and d is numerically low-rank in h and z. As a result, the number of channels required for this convolution operator is bounded logarithmically in the number of degrees of freedom and the desired accuracy.

The discussion so far shows that, in the small η regime, we can approximate the forward map $K : \eta(x,z) \to d(m,h)$ with a 1D CNN with a small number of channels. The filtered backprojection algorithm suggests that $\eta \approx (K^*K + \varepsilon I)^{-1}K^*d$. This motivates representing the product $(K^*K + \varepsilon I)^{-1}K^*$ as an NN. Regarding K^*, the analysis above for K shows that the adjoint operator K^* can also be approximated with a 1D CNN or BCR-Net with a small number of channels. The operator $(K^*K + \varepsilon I)^{-1}$ is a PDO in the (x,z) domain with global support, which can be approximated with a 2D BCR-Net or even a 2D CNN. Putting them together results in the following NN architecture [23] for the inverse map of the EIT problem:

$$d(m,h) \Rightarrow \text{1D CNN/BCR-Net} \Rightarrow \text{2D CNN/BCR-Net} \Rightarrow \eta(x,z). \tag{3.5}$$

Such an architecture can also be applied directly to the circular geometry (see Figure 7 (right)) if the unknown field η is parameterized in polar coordinates.

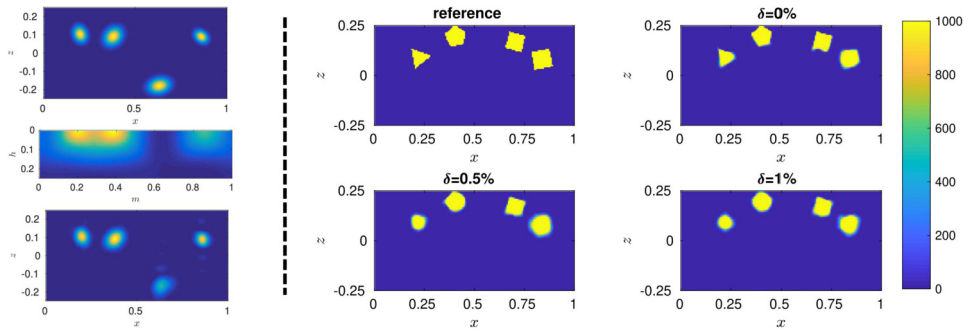


FIGURE 8

Electrical impedance tomography: (left, from top to bottom) the ground truth $\eta(x, z)$, the boundary measurement $d(m, h)$, and the NN reconstruction; (right) the NN reconstructions at different noise levels.

Figure 8 presents a numerical example. The NN has about 70K weights and is trained with about 10K (d, η) training pairs. The left part shows, from top to bottom, the ground truth internal parameters $\eta(x, z)$, the boundary measurements $d(m, h)$ in the warped coordinate system (m, h), and the NN reconstruction obtained by applying the trained NN to $d(m, h)$. The images show that the NN reconstruction is close to the ground truth, though the accuracy gradually deteriorates as the depth z grows due to the nature of the EIT problem. The right part shows how the NN, while trained with noiseless training data, performs under different noise levels in the testing boundary measurement $d(m, h)$. The images demonstrate that the trained NN is robust to measurement noise.

3.2. Optical tomography

Consider a circular domain Ω in 2D (see Figure 9) with boundary $\Gamma = \mathbb{S}^1$. The governing equation for optical tomography (OT) is the radiative transfer equation (RTE)

$$(L\Phi)(p, v) \equiv v \cdot \nabla\Phi(p, v) + \mu_t(x)\Phi(p, v) = \mu(p) \int_{\mathbb{S}^1} \sigma(v \cdot v')\Phi(p, v')\mathrm{d}v', \quad (p, v) \in \Omega \times \mathbb{S}^1, \tag{3.6}$$

where σ is a fixed scattering phase with $\int_{\mathbb{S}^1} \sigma(v \cdot v')\mathrm{d}v = 1$. The transport coefficient $\mu_t(p) = \mu_a + \mu(p)$ measures the total absorption, including the *known* physical absorption constant μ_a and the *unknown* scattering strength quantified by the term $\mu(p)$. In a typical OT experiment, for each boundary point $s \in \mathbb{S}^1$, one specifies at s either an isotropic scattering source or a delta source in the normal direction, and records the outgoing flux (denoted by $f^s(\cdot)$) at each point $r \in \mathbb{S}^1$. The set of boundary measurements is given by

$$d(r, s) = f^s(r) - f_0^s(r), \tag{3.7}$$

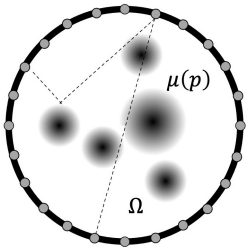


FIGURE 9
Optical tomography. An experimental setup in a circular geometry.

where $f_0^s(\cdot)$ is the outgoing flux when $\mu(x) \equiv 0$. The inverse problem is then to recover $\mu(p)$ from $d(r, s)$.

In order to obtain an approximation for the inverse map $d(r, s) \to \mu(p)$, we study how $d(r, s)$ depends on $\mu(p)$ in the perturbative regime. By using an equivalent integral formulation [21], one can explicitly derive the perturbative relationship between $\mu(p)$ and $d(r, s)$. However, for the purposes of NN design, a simple observation based on the rotation-equivariance of the experimental setup is sufficient. By introducing a warped coordinate system (s, h) with $(r, s) \equiv (h + s, s)$, the data $d(h, s)$ in the new system can be written as

$$d(s, h) \approx \iint_{(\rho,\theta)} k(h, \rho, s - \theta)\mu(\rho, \theta)\mathrm{d}\rho\mathrm{d}\theta, \tag{3.8}$$

which is a 1D convolution in s and θ, with h and ρ treated as parameters (i.e., channels in the NN terminology). Since the RTE (3.6) preserves singularities, especially when the absorption μ_a is weak, this map between η and d is singular in the h and ρ variables. As a result, the number of channels required for the 1D convolution operator can scale with the resolution in ρ and h.

The above discussion shows that, in the small μ regime, we can approximate the forward map $K : \mu(\rho, \theta) \to d(s, h)$ with a 1D CNN with multiple channels. The filtered backprojection algorithm suggests that $\mu \approx (K^*K + \varepsilon I)^{-1}K^*d$. This again motivates the approach of representing the product $(K^*K + \varepsilon I)^{-1}K^*$ as an NN. As the adjoint to K, the operator K^* can also be approximated with a 1D CNN or BCR-Net with multiple channels. The operator $(K^*K + \varepsilon I)^{-1}$ is a PDO in the (ρ, θ) domain with global support, which can be approximated efficiently with a 2D BCR-Net or CNN. Summarizing these discussions results in the following NN architecture [21] for the OT problem:

$$d(s, h) \Rightarrow \text{1D CNN/BCR-Net} \Rightarrow \text{2D CNN/BCR-Net} \Rightarrow \mu(\rho, \theta). \tag{3.9}$$

Figure 10 presents one numerical example. The resulting NN, with about 50K weights, is trained with a dataset of 8K (d, μ) training pairs. The images show the reference (ground truth) parameter μ, along with the NN reconstructions μ^{NN} at different noise levels. The results suggest that the learned NN representation of the inverse map is quite robust to noise, even though the optical tomography problem is (weakly) ill-posed.

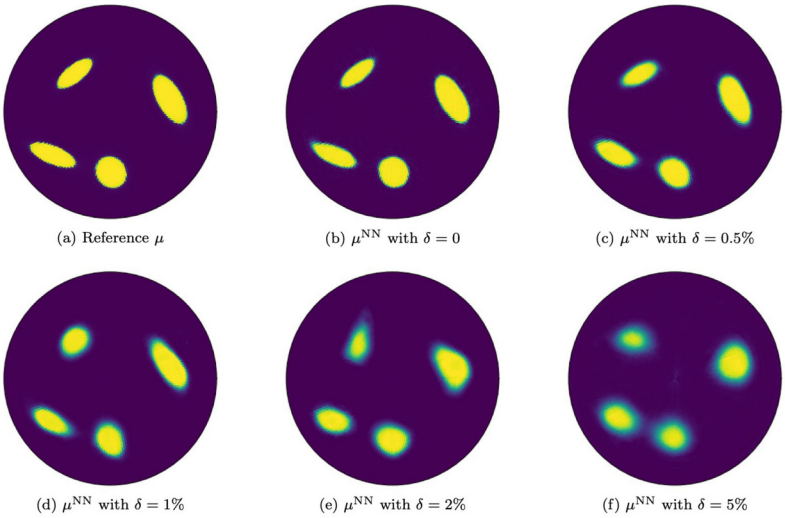


FIGURE 10

Optical tomography. Reference solution μ along with the NN reconstructions under different noise levels in the boundary measurement $d(s,h)$.

3.3. Inverse acoustic scattering

Let us consider the acoustic scattering problem in 2D in the frequency domain. The governing equation is the Helmholtz equation

$$(Lu)(p) = \left(-\Delta - \frac{\omega^2}{c(p)^2}\right)u(p) = 0, \tag{3.10}$$

where ω is a fixed angular frequency and $c(p)$ is the unknown velocity field. Assume that there exists a known constant background velocity c_0 such that $c(p) - c_0$ is compactly supported in a domain Ω (see Figure 11 (left)). In a typical experimental setup, for each incoming direction $s \in \mathbb{S}^1$, the plane wave $e^{i\omega s\cdot p}$ generates an outgoing scattered field $u^s(p)$ such that $u^s(p) + e^{i\omega s\cdot p}$ is a solution of (3.10). At each unit direction $r \in \mathbb{S}^1$, the far field pattern defined as

$$\hat{u}^s(r) \equiv \lim_{\rho\to\infty} u^s(\rho\cdot r)\sqrt{\rho}e^{-i\omega\rho} \tag{3.11}$$

is recorded. The set of boundary measurements is then $d(r,s) = \hat{u}^s(r)$. Instead of trying to recover $c(p)$ directly, it is often convenient to treat a rescaled index-of-refraction field $\eta(p) \equiv \frac{\omega^2}{c(p)^2} - \frac{\omega^2}{c_0^2}$ as the unknown. The inverse problem is then to recover $\eta(p)$ (equivalently to $c(p)$) from $d(r,s)$.

In order to obtain an approximation for the inverse map $d(r,s) \to \eta(x)$, as usual we consider first how $d(r,s)$ depends on $\eta(x)$ in the perturbative regime. A perturbative analysis for planar incoming waves and far field patterns in [20, 38] shows that, when η is small, the data $d(r,s)$ can be approximated up to a smooth amplitude as

$$d(r,s) \approx (K\eta)(r,s) \equiv \int_{p\in\Omega} e^{i\omega(s-r)\cdot p}\eta(p)\mathrm{d}p. \tag{3.12}$$

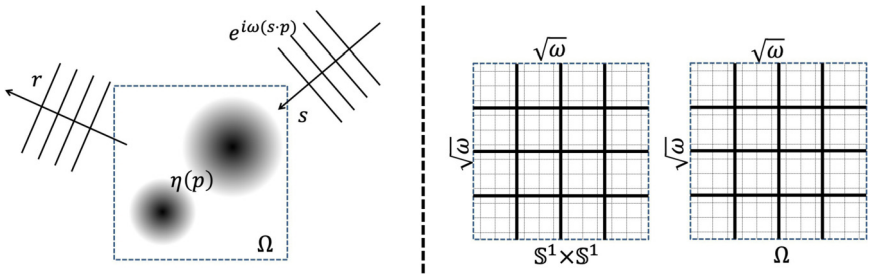


FIGURE 11

Inverse acoustic scattering: (left) an experimental setup in 2D; (right) the complementary low-rank structure of the forward map from $\eta(p)$ to $d(r,s)$.

A rank estimate of the operator kernel [38] shows that K is an FIO from Ω to $\mathbb{S}^1 \times \mathbb{S}^1$ (see Figure 11 (right) for an illustration). As a result, the approximate forward operator from η to d can be represented with a 2D Switch-Net.

The filtered backprojection states that $\eta \approx (K^*K + \varepsilon I)^{-1}K^*d$, thus motivating the approach of representing the product $(K^*K + \varepsilon I)^{-1}K^*$ as an NN. As the adjoint of an FIO is also an FIO [59], the operator K^* can also be approximated with a Switch-Net. The operator $(K^*K + \varepsilon I)^{-1}$ is a PDO in the p variable and can therefore be implemented with a 2D BCR-Net. Concatenating these two modules results in the NN architecture in [38] for the inverse acoustic scattering problem:

$$d(r,s) \Rightarrow \text{2D Switch-Net} \Rightarrow \text{2D BCR-Net} \Rightarrow \eta(p). \tag{3.13}$$

When Ω is a disk, it is natural to parameterize the unknown field η in polar coordinates (ρ,θ). The boundary measurement d is also written in a new coordinate system (h,m) with midpoint $m = \frac{r+s}{2}$ and shift $h = \frac{r-s}{2}$. Under these two new coordinate systems, the rotation-equivariance of the circular geometry implies that the map from $\eta(\rho,\theta)$ to $d(m,h)$ is a 1D convolution in θ and m, with h and ρ treated as the channel dimensions,

$$d(m,h) \approx \int_{\geq 0}\int_0^{2\pi} k(h,\rho,m-\theta)\eta(\rho,\theta)\mathrm{d}\rho\mathrm{d}\theta. \tag{3.14}$$

Following the discussion that leads from (3.8) to (3.9), we can also adopt the following NN architecture [20] for the circular geometry:

$$d(h,m) \Rightarrow \text{1D CNN/BCR-Net} \Rightarrow \text{2D CNN/BCR-Net} \Rightarrow \eta(\rho,\theta). \tag{3.15}$$

Figure 12 gives a numerical example for inverse acoustic scattering. The resulting NN has about 400K weights and is trained with a dataset of 16K (d,η) training pairs. The images show, for two different cases, the reference (ground truth) parameter η, along with the NN reconstructions η^{NN} at different noise levels up to 100%. The results suggest that the learned NN inverse map is highly robust to noise, thanks to the well-posedness of this problem.

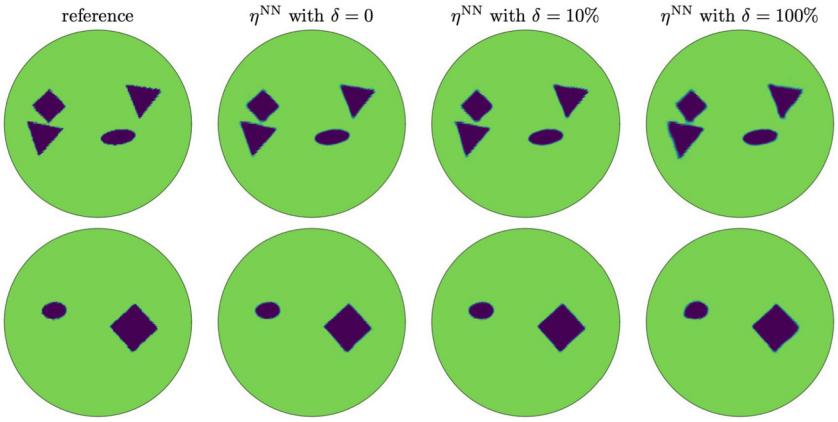


FIGURE 12

Inverse scattering. Each row corresponds to a different test case. For each case, we plot the reference solution, along with the NN reconstructions up to a 100% noise level.

3.4. Seismic imaging

We consider the seismic imaging setting under a simple 2D acoustic model in the frequency domain. The governing equation is again the Helmholtz equation

$$(Lu)(p) = \left(-\Delta - \frac{\omega^2}{c(p)^2}\right)u(p) = f(p), \quad p \in \Omega, \tag{3.16}$$

where ω is a fixed frequency and $c(p)$ is sound speed. We assume that the background velocity $c_0(p)$ is given and the difference between $c(p)$ and $c_0(p)$ is supported in Ω (see Figure 13 (left)). In a typical experimental setup, for each point s on the top surface, one specifies a delta source $f(p) = \delta_s(p)$ and records the wave solution $u^s(\cdot)$ of (3.16) at all points r also on the top surface. The set of boundary measurements is

$$d(r, s) = u^s(r) - u_0^s(r), \tag{3.17}$$

where $u_0^s(\cdot)$ is the solution of some background velocity $c_0(p)$. By again introducing the scaled index-of-refraction field $\eta(p) = \frac{\omega^2}{c(p)^2} - \frac{\omega^2}{c_0(p)^2}$, we obtain the inverse problem of recovering $\eta(p)$ from $d(r, s)$.

As usual, in order to obtain an approximation for the inverse map $d(r, s) \to \eta(p)$, we first study how $d(r, s)$ depends on $\eta(p)$ in the perturbative regime. A perturbative analysis [38] of planar incoming waves and far field patterns shows that, when η is small, the boundary measurement $d(r, s)$ can be well approximated with

$$d(r, s) \approx (K\eta)(r, s) \equiv \int \big(G_0(r, p)G_0(p, s)\big)\eta(p)\mathrm{d}p, \tag{3.18}$$

where $G_0(p)$ is the Green's function of the background operator $L_0 = -\Delta - \omega^2/c_0^2(p)$. A rank estimate of the kernel $G_0(r, p)G_0(p, s)$ in [38] proves that K is an FIO defined between the domain Ω and the product (r, s) space (see Figure 13 (right) for an illustration). As a result, the forward operator from η to d can be approximated with a 2D Switch-Net.

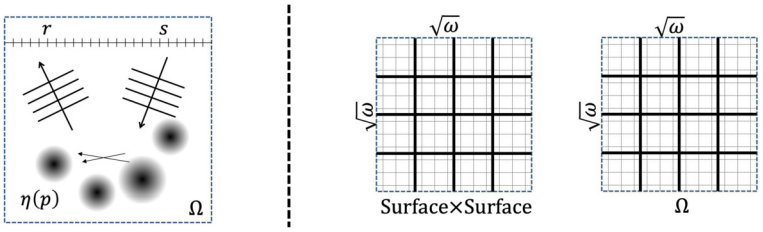


FIGURE 13

Seismic imaging: (left) a simple experimental setup in 2D; (right) the complementary low-rank structure of the forward map from $\eta(p)$ to $d(r,s)$.

The filtered backprojection algorithm again suggests that $\eta \approx (K^*K + \varepsilon I)^{-1}K^*d$, which motivates representing the product $(K^*K + \varepsilon I)^{-1}K^*$ as an NN. As the adjoint of an FIO K, K^* can be approximated with a Switch-Net. Under generic conditions, the operator $(K^*K + \varepsilon I)^{-1}$ is a PDO in the p variable, which can be approximated with a 2D BCR-Net. Putting everything together results in the following NN architecture [38] for the seismic imaging problem:

$$d(r,s) \Rightarrow \text{2D Switch-Net} \Rightarrow \text{2D BCR-Net} \Rightarrow \eta(x,z), \tag{3.19}$$

where x and z are the horizontal and depth coordinates of p, respectively.

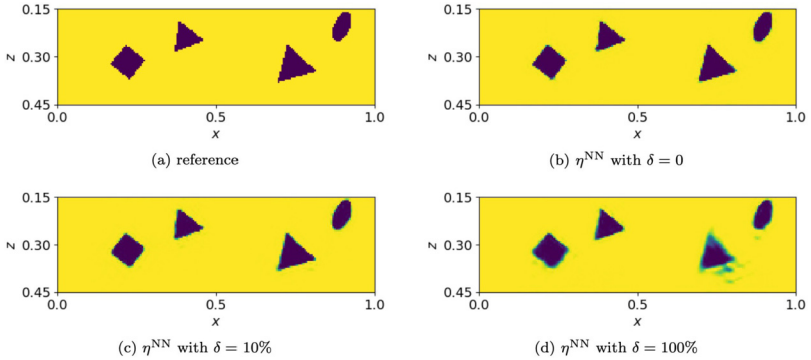
Often in seismic imaging, the background velocity $c_0(p)$ only depends on the depth z (and is independent of the horizontal coordinate x). In this case, we can exploit the translation-equivariance in the horizontal direction and reparameterize the boundary measurement d under the coordinate system (m,h) with $m = \frac{r+s}{2}$ and offset $h = \frac{r-s}{2}$. Under this new coordinate system, the forward map from $\eta(x,z)$ to $d(m,h)$ is a 1D convolution with the offset h and depth z treated as channels,

$$d(m,h) \approx \int_0^Z \int k(h,z,m-x)\eta(x,z)\mathrm{d}x\mathrm{d}z. \tag{3.20}$$

Following the discussion that leads from (3.4) to (3.5), we obtain the following NN architecture [20] for $c_0(p)$ that depends only on depth:

$$d(m,h) \Rightarrow \text{1D CNN/BCR-Net} \Rightarrow \text{2D CNN/BCR-Net} \Rightarrow \eta(x,z). \tag{3.21}$$

Figure 14 shows a numerical example for the seismic inversion problem. The NN has about 1M weights in total and is trained with a dataset of 16K (d,η) pairs. The images show the reference (ground truth) parameter η, along with the NN reconstructions η^{NN} at noise levels up to 100%. The results demonstrate that the learned NN inverse map is quite robust to noise. Notice that the reconstruction quality deteriorates with depth naturally since the boundary measurements are all collected at the top surface.



(a) reference (b) η^{NN} with $\delta = 0$

(c) η^{NN} with $\delta = 10\%$ (d) η^{NN} with $\delta = 100\%$

FIGURE 14

Seismic imaging. The reference solution along with the NN reconstructions with different levels of noise added to the boundary measurements.

3.5. Traveltime tomography

Here we consider a circular domain Ω in 2D with the boundary denoted by Γ (see Figure 15). The governing equation for traveltime tomography (TT) is the eikonal equation

$$|\nabla u(p)| = \frac{1}{c(p)}, \quad p \in \Omega, \tag{3.22}$$

where $c(p)$ is the unknown velocity field. Assuming that $c(p)$ has a background velocity c_0 (taken to be 1 without loss of generality), we introduce the slowness deviation $\eta(p) \equiv \frac{1}{c(p)} - 1$ and write (3.22) as $|\nabla u(p)| = 1 + \eta(p)$. In a typical setup, we specify the zero boundary condition at each boundary point s, solve for the viscosity solution $u^s(x)$ of (3.22), and record $u^s(r)$ at each boundary point r. The set of boundary measurements is then given by

$$d(r, s) = u^s(r) - u_0^s(r), \tag{3.23}$$

where $u_0^s(r) = \|r - s\|$ is the solution for $\eta(x) \equiv 0$. The inverse problem is to recover $\eta(p) \equiv \frac{1}{c(p)} - 1$ from $d(r, s)$.

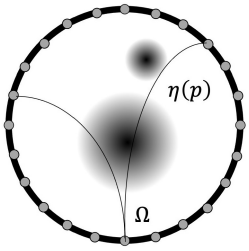


FIGURE 15

Traveltime tomography. Experimental setup in a circular geometry.

To obtain an approximation for the inverse map $d(r,s) \to \eta(p)$, we study how $d(r,s)$ depends on $\eta(p)$ in the perturbative regime. A simple consideration based on the rotation-equivalence of the experimental setup suggests viewing the parameter η in polar coordinates (ρ,θ) and the boundary measurements in the warped coordinates (s,h) with $(r,s) \equiv (h+s,s)$. The data $d(h,s)$ can then be written as [22]

$$d(s,h) \approx (K\eta)(h,s) = \int_{\geq 0} \int_0^{2\pi} k(h,\rho,s-\theta)\eta(\rho,\theta)\mathrm{d}\rho\mathrm{d}\theta, \tag{3.24}$$

which is a 1D convolution in s and θ, with h and ρ treated as parameters (or channels in the NN terminology). Since the viscosity solution of the eikonal equation often has singularities, the number of channels required for the 1D convolution operator can be quite significant.

The discussion above shows that, in the small η regime, we can approximate the forward map $K : \eta(\rho,\theta) \to d(s,h)$ with a 1D CNN or BCR-Net with multiple channels. The filtered backprojection algorithm $\eta \approx (K^*K + \varepsilon I)^{-1}K^*d$ suggests representing the product $(K^*K + \varepsilon I)^{-1}K^*$ as a linear NN and then generalizing to the nonlinear regime. By invoking the same argument used for K, the adjoint operator K^* can also be approximated with a 1D CNN or BCR-Net with a small number of channels. The operator $(K^*K + \varepsilon I)^{-1}$ is a PDO in the (ρ,θ) domain with global support, which can be approximated with a 2D BCR-Net or with a 2D CNN. Summarizing the discussion results in the following NN architecture [22] for traveltime tomography:

$$d(s,h) \Rightarrow \text{1D CNN/BCR-Net} \Rightarrow \text{2D BCR-Net/CNN} \Rightarrow \eta(\rho,\theta). \tag{3.25}$$

Figure 16 gives a numerical example for the traveltime tomography. The NN for the inverse map, with about 640K weights, is trained with a set of 16K (d,η) pairs. The three rows correspond to test examples with negative inclusion $c(p) < 1$, positive inclusion $c(p) > 1$, and mixed inclusion, respectively. For each test example, we plot the reference solution along with the NN reconstructions with different levels of noise added to the boundary measurements. The results show that, even for this ill-posed problem, the NN inverse map is accurate and robust with respect to noise.

4. CONCLUDING REMARKS

In this paper, we discussed our recent work on applying deep learning to inverse problems. On the algorithmic side, we proposed two new NN modules, BCR-Net and Switch-Net. They are motivated by the pseudodifferential and Fourier integral operators, which play key roles in the study of inverse problems. On the application side, we propose NNs that approximate the inverse maps in five settings of interest: electrical impedance tomography, optical tomography, inverse acoustic scattering, seismic imaging, and traveltime tomography. In each application, the architecture is motivated by the perturbation theory and filtered backprojection and is implemented using the new modules along with standard convolution layers. In several cases, we have heavily relied on the special geometry of the domain Ω and the data collection process. When combined with appropriate reparameterizations, this

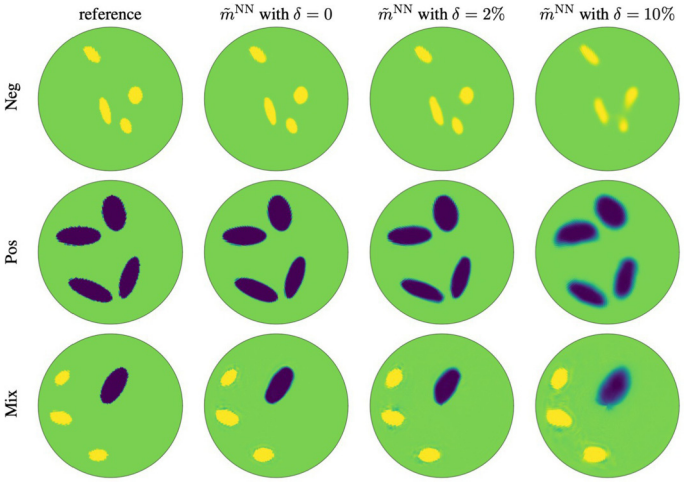


FIGURE 16

Traveltime tomography. The three rows correspond to negative, positive, and mixed inclusions. For each case, the reference solution is shown along with the NN reconstructions with different levels of noise added to the boundary measurements.

often results in NN architectures that are both general and effective. Our approach provides a seamless way that combines the mathematical structure of the inverse problems, the power of deep NNs, and the information in the data prior. Below we list some directions for future research:

- We have considered only the case of complete measurement data. A question of both practical and theoretical importance is how to extend to the case of partial measurement data.
- For wave-based inverse problems, we have focused on a single frequency or a single energy. In many applications, one often has access to boundary measurements at multiple frequencies or energies, or even time-dependent measurements.
- The first part of the proposed NNs is closely related to the migration step in traditional imaging pipelines (such as in seismic imaging). An interesting study would be to compare the intermediate result after the first part of our NN with the migration results to get a more precise understanding of the proposed NNs.

The study of inverse problem using deep learning has grown into a relatively large subject [6, 39, 49, 56, 62]. This paper has solely focused on one approach that is deeply rooted in microlocal analysis (see [3,10] for related work). There are a few other highly active research directions that we have not discussed here, but which may be of interest to the reader:

- We have not discussed the work on (linear) underdetermined inverse problem in imaging [32, 51]. This field is closely connected with sparse recovery problems, such as compressive sensing, matrix completion, phase retrieval, etc. [11].
- In the unrolling or unfolding approach [2, 13, 15, 25, 33, 50, 53, 63] for solving inverse problems, one writes the iterative solution algorithm as a ResNet and then trains the network parameters to minimize the reconstruction error. In many cases, this approach leads to a very high quality reconstruction.
- There is also active work studying stability issues when applying deep learning to inverse problems [4, 14, 24, 27], which is particularly important for applications with ill-posed inverse problems.

ACKNOWLEDGMENTS

The author would like to thank Bin Dong, Zach Izzo, and Ozan Öktem for their help in preparing this manuscript.

FUNDING

This work was partially supported by the National Science Foundation under awards DMS-1818449 and DMS-2011699 and by the U. S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283, 2016.

[2] J. Adler and O. Öktem, Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Probl.* **33** (2017), no. 12, 124007.

[3] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen, Deep microlocal reconstruction for limited-angle tomography. 2021, arXiv:2108.05732.

[4] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci.* **117** (2020), no. 48, 30088–30095.

[5] M. Araya-Polo, J. Jennings, A. Adler, and T. Dahlke, Deep-learning tomography. *Lead. Edge* **37** (2018), no. 1, 58–66.

[6] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, Solving inverse problems using data-driven models. *Acta Numer.* **28** (2019), 1–174.

[7] L. Bar and N. Sochen, Unsupervised deep learning algorithm for PDE-based forward and inverse problems. 2019, arXiv:1904.05417.

[8] J. Berg and K. Nyström, A unified deep artificial neural network approach to partial differential equations in complex geometries. *Neurocomputing* **317** (2018), 28–41.

[9] G. Beylkin, R. Coifman, and V. Rokhlin, Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.* **44** (1991), no. 2, 141–183.

[10] T. A. Bubba, M. Galinier, M. Lassas, M. Prato, L. Ratti, and S. Siltanen, Deep neural networks for inverse problems with pseudodifferential operators: An application to limited-angle tomography. *SIAM J. Imaging Sci.* **14** (2021), no. 2, 470–505.

[11] E. J. Candès, Mathematics of sparsity (and a few other things). In *Proceedings of the International Congress of Mathematicians, Seoul, South Korea* 123, Citeseer, 2014.

[12] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks. *Science* **355** (2017), no. 6325, 602–606.

[13] J. Cheng, H. Wang, Y. Zhu, Q. Liu, Q. Zhang, T. Su, J. Chen, Y. Ge, Z. Hu, X. Liu, et al., Model-based deep medical imaging: the roadmap of generalizing iterative reconstruction model using deep learning. 2019, arXiv:1906.08143.

[14] M. J. Colbrook, V. Antun, and A. C. Hansen, Can stable and accurate neural networks be computed? – On the barriers of deep learning and Smale's 18th problem. 2021, arXiv:2101.08286.

[15] M. V. de Hoop, M. Lassas, and C. A. Wong, Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. 2019, arXiv:1912.11090.

[16] W. E and B. Yu, The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.* **6** (2018), no. 1, 1–12.

[17] Y. Fan, C. O. Bohorquez, and L. Ying, BCR-Net: a neural network based on the nonstandard wavelet form. *J. Comput. Phys.* **384** (2019), 1–15.

[18] Y. Fan, J. Feliu-Fabà, L. Lin, L. Ying, and L. Zepeda-Núñez, A multiscale neural network based on hierarchical nested bases. *Res. Math. Sci.* **6** (2019), no. 2, 21.

[19] Y. Fan, L. Lin, L. Ying, and L. Zepeda-Núnez, A multiscale neural network based on hierarchical matrices. *Multiscale Model. Simul.* **17** (2019), no. 4, 1189–1213.

[20] Y. Fan and L. Ying, Solving inverse wave scattering with deep learning. 2019, arXiv:1911.13202.

[21] Y. Fan and L. Ying, Solving optical tomography with deep learning. 2019, arXiv:1910.04756.

[22] Y. Fan and L. Ying, Solving traveltime tomography with deep learning. 2019, arXiv:1911.11636.

[23] Y. Fan and L. Ying, Solving electrical impedance tomography with deep learning. *J. Comput. Phys.* **404** (2020), 109119.

[24] M. Genzel, J. Macdonald, and M. März, Solving inverse problems with deep neural networks–robustness included? 2020, arXiv:2011.04268.

[25] D. Gilton, G. Ongie, and R. Willett, Neumann networks for linear inverse problems in imaging. *IEEE Trans. Comput. Imaging* **6** (2019), 328–343.

[26] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning* 1, MIT Press, Cambridge, 2016.

[27] N. M. Gottschling, V. Antun, B. Adcock, and A. C. Hansen, The troublesome kernel: why deep learning for inverse problems is typically unstable. 2020, arXiv:2001.01258.

[28] L. Greengard and V. Rokhlin, A fast algorithm for particle simulations. *J. Comput. Phys.* **73** (1987), no. 2, 325–348.

[29] W. Hackbusch, L. Grasedyck, and S. Börm, An introduction to hierarchical matrices. *Math. Bohem.* **127** (2002), no. 229–241.

[30] J. Han and A. Jentzen W. E, Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci.* **115** (2018), no. 34, 8505–8510.

[31] J. Han, L. Zhang, R. Car, and W. E, Deep potential: a general representation of a many-body potential energy surface. *Commun. Comput. Phys.* **23** (2018), no. 3, 629–639.

[32] P. Hand and V. Voroninski, Global guarantees for enforcing deep generative priors by empirical risk. In *Conference on learning theory*, pp. 970–978, PMLR, 2018.

[33] A. Hauptmann, J. Adler, S. Arridge, and O. Öktem, Multi-scale learned iterative reconstruction. *IEEE Trans. Comput. Imaging* **6** (2020), 843–856.

[34] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **29** (2012), no. 6, 82–97.

[35] H. Kabir, Y. Wang, M. Yu, and Q.-J. Zhang, Neural network inverse modeling and applications to microwave filter design. *IEEE Trans. Microwave Theory Tech.* **56** (2008), no. 4, 867–879.

[36] Y. Khoo, J. Lu, and L. Ying, Solving for high-dimensional committor functions using artificial neural networks. *Res. Math. Sci.* **6** (2019), no. 1, 1.

[37] Y. Khoo, J. Lu, and L. Ying, Solving parametric PDE problems with artificial neural networks. *European J. Appl. Math.* **32** (2021), no. 3, 421–435.

[38] Y. Khoo and L. Ying, SwitchNet: a neural network model for forward and inverse scattering problems. *SIAM J. Sci. Comput.* **41** (2019), no. 5, A3182–A3201.

[39] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft, Machine learning in seismology: Turning data into insights. *Seismol. Res. Lett.* **90** (2019), no. 1, 3–14.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25** (2012), 1097–1105.

[41] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider, A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.* (2021), 1–53.

[42] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning. *Nature* **521** (2015), no. 436.

[43] M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, Deep learning of the tissue-regulated splicing code. *Bioinformatics* **30** (2014), no. 12, i121–i129.

[44] Y. Li, J. Lu, and A. Mao, Variational training of neural network approximations of solution maps for physical models. *J. Comput. Phys.* **409** (2020), 109338.

[45] Y. Li, H. Yang, E. R. Martin, K. L. Ho, and L. Ying, Butterfly factorization. *Multiscale Model. Simul.* **13** (2015), no. 2, 714–732.

[46] Z. Long, Y. Lu, X. Ma, and B. Dong PDE-net, Learning PDEs from data. In *International Conference on Machine Learning*, pp. 3208–3216, PMLR, 2018.

[47] A. Lucas, M. Iliadis, R. Molina, and A. K. Katsaggelos, Using deep neural networks for inverse problems in imaging: beyond analytical methods. *IEEE Signal Process. Mag.* **35** (2018), no. 1, 20–36.

[48] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55** (2015), no. 2, 263–274.

[49] M. T. McCann and M. Unser, Biomedical image reconstruction: from the foundations to deep neural networks. 2019, arXiv:1901.03565.

[50] V. Monga, Y. Li, and Y. C. Eldar, Algorithm unrolling: interpretable, efficient deep learning for signal and image processing. *IEEE Signal Process. Mag.* **38** (2021), no. 2, 18–44.

[51] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, Deep learning techniques for inverse problems in imaging. *IEEE J. Sel. Areas Inf. Theory* **1** (2020), no. 1, 39–56.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019), 8026–8037.

[53] P. Putzky and M. Welling, Invert to learn to invert. *Adv. Neural Inf. Process. Syst.* **32** (2019), 446–456.

[54] M. Raissi and G. E. Karniadakis, Hidden physics models: machine learning of nonlinear partial differential equations. *J. Comput. Phys.* **357** (2018), 125–141.

[55] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378** (2019), 686–707.

[56] S. Ravishankar, J. C. Ye, and J. A. Fessler, Image reconstruction: from sparsity to data-adaptive methods and machine learning. *Proc. IEEE* **108** (2019), no. 1, 86–109.

[57] K. Rudd and S. Ferrari, A constrained integration (CINT) approach to solving partial differential equations using artificial neural networks. *Neurocomputing* **155** (2015), 277–285.

[58] J. Schmidhuber, Deep learning in neural networks: an overview. *Neural Netw.* **61** (2015), 85–117.

[59] E. M. Stein and T. S. Murphy, *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals* 3, Princeton University Press, 1993.

[60] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **27** (2014), 3104–3112.

[61] C. Tan, S. Lv, F. Dong, and M. Takei, Image reconstruction based on convolutional neural network for electrical resistance tomography. *IEEE Sens. J.* **19** (2018), no. 1, 196–204.

[62] P. R. Wiecha, A. Arbouet, C. Girard, and O. L. Muskens, Deep learning in nanophotonics: inverse design and beyond. *Photon. Res.* **9** (2021), no. 5, B182–B200.

[63] H.-M. Zhang and B. Dong, A review on deep learning in medical image reconstruction. *J. Oper. Res. Soc. China* (2020), 1–30.

LEXING YING

Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA, lexing@stanford.edu

16. CONTROL THEORY AND OPTIMIZATION

SPECIAL LECTURE

DISCREPANCY THEORY AND RELATED ALGORITHMS

NIKHIL BANSAL

ABSTRACT

We survey some classical results and techniques in discrepancy theory, and the recent developments in making these techniques algorithmic. The previous methods were typically based on non-constructive approaches such as the pigeonhole principle, and counting arguments involving exponentially many objects and volume of convex bodies. The recent algorithmic methods are based on an interesting interplay of methods from discrete Brownian motion, convex geometry, optimization, and matrix analysis, and their study has led to interesting new connections and progress in both discrepancy and algorithm design.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68R; Secondary 05D40, 11K38, 68Q25

KEYWORDS

Discrepancy, combinatorics, algorithms, random walks

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5178–5210
DOI 10.4171/ICM2022/169

1. INTRODUCTION

Combinatorial discrepancy deals with the following type of question. Given a set-system (U, C) with elements $U = [n]$ and a collection $C = \{S_1, \ldots, S_m\}$ of subsets of U, how well can we color the elements red and blue so that each set $S_i \in C$ is colored as evenly as possible. Formally, let us use ± 1 to denote the colors red and blue, so that if $x(j)$ denotes the color of element j, then $|\sum_{j \in S} x(j)|$ is the imbalance for set S. Then the discrepancy of the set system (U, C) is defined as

$$\operatorname{disc}(C) = \min_{x:U \to \{-1,1\}} \max_{i \in [m]} \left| \sum_{j \in S_i} x(j) \right|,$$

that is, the minimum imbalance that must occur in at least one of the sets in C, over all possible bipartitions of U.

More generally, for an $m \times n$ matrix A, the discrepancy of A is defined as

$$\operatorname{disc}(A) = \min_{x \in \{-1,1\}^n} \|Ax\|_\infty. \tag{1.1}$$

This generalizes the definition for set systems by choosing A to be the incidence matrix for the system. Letting $v_1, \ldots, v_n$ denote the columns of A, this is the same as minimizing $\|\sum_j x(j)v_j\|_\infty$ over all ± 1 colorings x, and the problem is also referred to as vector balancing. In some settings, one also considers more general norms besides ℓ_∞, and more general objects v_i than just vectors.

Roughly speaking, discrepancy can be viewed as the study of how to divide a set of objects into two (or more) parts which are as similar as possible, with respect to various criteria. For this reason the problem arises in several applications, often in unexpected ways, and is related to various topics in mathematics and theoretical computer science [22,23,26,47,53]. For example, in computer science it has several applications in areas such as computation geometry, pseudorandomness, approximation algorithms, numerical integration, and differential privacy.

Beating random coloring. For any discrepancy problem, one option is to simply pick a random coloring by setting each $x(j)$ independently and uniformly to ± 1. However, for many problems one can do substantially better, so in a sense discrepancy theory can be viewed as the study of how to improve over the basic probabilistic method [1].

1.1. A brief history

Roughly speaking, there are three classical techniques in discrepancy. One of the earliest techniques was linear algebraic and similar to the well-known iterated-rounding technique [17,19,41]. Though this technique gives surprisingly good bounds for some problems in discrepancy, in general they are quite weak and far from optimal.

A huge breakthrough was made in the 1980s with the partial-coloring method due to Beck [18] and Spencer [63]. A similar approach based on ideas from convex geometry was developed independently by Gluskin [33]. Roughly speaking, this method guarantees the existence of a coloring of a constant fraction of the elements where every set in the set

system incurs a low discrepancy. This method is then repeated $O(\log n)$ times in order to get a full coloring of all the n elements.

The third approach was developed by Banaszczyk [4] in the late 1990s based on sophisticated ideas from convex geometry. His technique produced a full coloring directly, and led to improved bounds for many fundamental discrepancy problems.

Algorithmic aspects. Interestingly, the original proofs of the partial-coloring method and Banaszczyk's method were based on non-constructive approaches such as counting arguments, pigeonhole principle, and volume estimates of convex bodies, and did not give efficient algorithms. It was even conjectured that these results might be inherently non-algorithmic. This was problematic as in many applications of discrepancy one actually needs to be able to find good colorings efficiently.

In recent years, there has been remarkable progress in obtaining algorithmic versions of both the partial-coloring method [6, 28, 36, 45, 59] and Banaszczyk's method [8–10, 24, 35, 42]. There techniques combine ideas from linear algebra, discrete Brownian motion, optimization, and convex geometry in interesting ways, and lead to several new results and insights both in discrepancy and algorithm design. Another remarkable development has been on approximating hereditary discrepancy based on the γ_2-norm from functional analysis and semidefinite programming duality [49, 50].

In this survey, we give a brief overview of both the classical techniques and recent algorithmic results, and sketch the main ideas behind them. We also discuss some recent new directions such as online discrepancy, discrepancy of random instances, matrix discrepancy and mention several conjectures and problems that are still open. Unfortunately, we have to leave out several interesting topics, and in particular the various exciting applications of these results. We also leave out techniques for proving lower bounds for discrepancy problems.

1.2. Some examples

We describe some classical problems to give a flavor of the area, and we will use these throughout as running examples to illustrate the various techniques. Most of these problems have a long and fascinating history, that we will discuss only very briefly here.

(1) *Spencer's problem.* What is the discrepancy of an arbitrary set system with n elements and m sets?

(2) *Beck–Fiala problem.* What is the discrepancy of a set system where each element lies in at most d sets, i.e., the maximum degree is at most d?

(3) *Komlós problem.* What is the discrepancy of a matrix where the columns have ℓ_2-norm at most 1?

(4) *Prefix Komlós.* Given vectors $v_1, \ldots, v_n \in \mathbb{R}^m$ satisfying $\|v_j\|_2 \leq 1$, minimize the maximum discrepancy of prefixes, i.e., minimize $\max_{k\in[n]} \| \sum_{j=1}^{k} x(j)v_j \|_\infty$.

(5) *Tusnady's problem.* Given a set U of n arbitrary points in $[0,1]^d$, what is the discrepancy with respect to rectangles, i.e., sets $R \cap U$ where R ranges over all possible rectangles $\prod_{i=1}^{d}[a_i, b_i] \subset [0,1]^d$?

(6) *Discrepancy of Arithmetic Progressions.* Here $U = \mathbb{Z}_n$ and the sets S are of the form $S_{a,b} = \{a, a+b, a+2b, \ldots\}$. The special case of homogeneous arithmetic progressions (with $a = 0$ for each set) is called the Erdős Discrepancy Problem.

Given a coloring x and a row a, let $\mathrm{disc}(x,a) := |\sum_j x(j)a(j)|$ denote the discrepancy of x for a. For any set S, by standard probabilistic tail bounds

$$\Pr\big[\mathrm{disc}(x,S) \geq c|S|^{1/2}\big] \approx \exp\big(-c^2/2\big), \tag{1.2}$$

and thus a random coloring has discrepancy $\Omega(n^{1/2})$ or worse for all the problems above. We now describe the various improved bounds known for them. We shall give the details in later sections.

1.2.1. Spencer's problem

For an arbitrary set system, (1.2) and a union bound over the m sets implies that a random coloring has discrepancy $O((n \log m)^{1/2})$ with high probability (whp). In an influential work, Spencer [63] and Gluskin [33] showed the following result.

Theorem 1.1. *Any set system with $m \geq n$ sets has discrepancy $O((n \log 2(m/n))^{1/2})$. For $m \leq n$, the discrepancy is $O(m^{1/2})$.*

In particular, for $m = n$ this gives $O(n^{1/2})$ discrepancy, which is also the best possible, answering a question of Erdős. While this $O(\log n)^{1/2}$ factor improvement over the random coloring may seem relative minor, Spencer developed the partial coloring method to prove Theorem 1.1, which has become a key tool and gives huge improvements for many other problems.

1.2.2. Beck–Fiala and Komlós problem

Beck and Fiala [19], in one of the first applications of the iterated rounding technique, showed the following result.

Theorem 1.2. *Any set system with maximum degree d has discrepancy at most $2d-1$.*

A long-standing conjecture is the following.

Conjecture 1.2.1 ([19]). *The discrepancy of any set system with degree d is $O(d^{1/2})$.*

If we allow a mild dependence on n, the partial-coloring method gives a bound of $O(d^{1/2}\log n)$. The best known bound in this direction is $O((d\log n)^{1/2})$ due to Banaszczyk [4], based on a more general result that we shall see later.

Scaling the entries by $d^{-1/2}$, notice that the Beck–Fiala problem is a special case of the Komlós problem. For the Komlós problem, the partial-coloring method gives an $O(\log n)$

bound, and Banaszczyk's method gives the best known bound of $O((\log n)^{1/2})$. The following conjecture generalizes Conjecture 1.2.1.

Conjecture 1.2.2 (Komlós). *Given any $v_1, \ldots, v_n \in \mathbb{R}^m$ satisfying $\|v_j\|_2 \leq 1$ for all j, there is an $x \in \{-1, 1\}^n$ such that $\| \sum_j x(j) v_j \|_\infty = O(1)$.*

1.2.3. Prefix discrepancy

The study of discrepancy problems involving prefixes of a given sequence of vectors also has a long history and several surprising connections to other classical ordering problems. See, e.g., [17] for a fascinating survey and also [5, 20, 27, 47]. We restrict our focus here to the prefix version of the Komlós problem. The best bound known here is $O((\log n)^{1/2})$ due to Banaszczyk [5], where he further extended his method from [4] to handle prefixes.

Given this extension, a natural question is whether the prefix Komlós problem is any harder than the one without prefixes.

Problem 1.3. Is the discrepancy of the prefix Komlós problem $O(1)$?

There is no clear consensus here, and in fact for some discrepancy problems it is known that considering prefixes makes the problem harder [30, 52].

Algorithmic aspect. As we shall see later, there are several algorithmic approaches known by now for the partial-coloring method and for Banaszczyk's method in [4] without prefixes. However, the best algorithmic bound we know for the prefix version is still $O(\log n)$, based on partial coloring approach, and the following question is very interesting.

Problem 1.4. Find an efficient algorithm to obtain an $O((\log n)^{1/2})$ discrepancy coloring for the prefix Komlós problem.

1.2.4. Tusnady's problem

Here one can do exponentially better than random colorings, and these ideas have significant applications in numerical integral and quasi-Monte Carlo methods [23, 47].

The case of $d = 2$ is already instructive to see the relative power of various techniques. Moreover, we still do not know the right answer here. Linear algebraic methods give a bound of $O(\log^4 n)$. Using partial coloring, this can be pushed to about $O(\log^{5/2} n)$ [47]. The current best bound is $O(\log^{3/2} n)$ due to Nikolov [54], based on Banaszczyk's result for prefix discrepancy [5]. On the other hand, the best known lower bound is $\Omega(\log n)$ [49, 61].

For general d, after a long line of work, the current lower and upper bounds are $\Omega_d(\log^{d-1} n)$ [49] and $O_d(\log^{d-1/2} n)$ [54]. Closing this gap is an important open problem.

Conjecture 1.4.1. *The discrepancy of Tusnady's problem in d dimensions is $O(\log^{d-1} n)$.*

1.2.5. Arithmetic progressions

This problem has a long history, including results of Weyl [69] and Roth [58]. Using Fourier analysis, Roth [58] proved a lower bound of $\Omega(n^{1/4})$. Interestingly, Roth believed that

his result might not be best possible and the right exponent might be $1/2$, as suggested by random colorings. Eventually, Matoušek and Spencer [51] gave a matching $O(n^{1/4})$ upper bound using the partial coloring method.

For homogeneous arithmetic progressions, an $O(\log n)$ upper bound follows from a simple explicit construction. A famous question of Erdős was whether the discrepancy is $O(1)$. This was answered negatively in a breakthrough work by Tao [67].

1.3. Hereditary discrepancy and rounding

An important application of discrepancy is in rounding a fractional solution to an integral one without introducing much error, based on the following result of Lovász, Spencer, and Vesztergombi [44].

Theorem 1.5 ([44]). *For any $x \in \mathbb{R}^n$ satisfying $Ax = b$, there is a $\tilde{x} \in \mathbb{Z}^n$ with $\|\tilde{x} - x\|_\infty < 1$, such that $\|A(x - \tilde{x})\|_\infty \leq \operatorname{herdisc}(A)$.*

Here $\operatorname{herdisc}(A)$ is the *hereditary discrepancy* of A, which is a more robust version of discrepancy, and defined as the maximum discrepancy over all column restrictions of A,

$$\operatorname{herdisc} A = \max_{S \subset [n]} \operatorname{disc}(A_{|S}) = \max_{S \subset [n]} \min_{x \in \{-1,1\}^n} \|Ax\|_\infty.$$

For most classes of set systems, any upper bound on discrepancy is also a bound on hereditary discrepancy, as the class itself may be closed under taking subset of columns. For example, this holds for all the problems in Section 1.2, except for the case of arithmetic progressions, which is an example of a particular set system.

Rounding via discrepancy. To see the idea behind Theorem 1.5, suppose that x is $1/2$-integral (i.e., each $x(j)$ has fractional part 0 or $1/2$). Let S be the set of variables with fractional part $1/2$, and let y be ± 1 coloring of S with discrepancy $\operatorname{disc}(A_{|S})$. Then $x' = x + y/2$ is integral and

$$\|Ax' - Ax\|_\infty = \|A(y/2)\|_\infty = \operatorname{disc}(A_{|S})/2 \leq \operatorname{herdisc}(A)/2.$$

That is, the signs of y are used to decide whether to round each $x(j)$ up or down. For arbitrary x, Theorem 1.5 follows by applying this to each bit after the decimal starting from the least significant bit.

The problem of rounding arises naturally for example in designing efficient approximation algorithms for discrete optimization problems. However, note that Theorem 1.5 only shows the existence of a good rounding, and gives no clue on how to actually find one efficiently. We shall see an algorithmic version of Theorem 1.5 in Section 3.1. In general, the recent algorithmic progress on discrepancy has led to several new results in approximation algorithms, a particularly notable result is [60].

2. CLASSICAL TECHNIQUES

We now describe the classical techniques of (i) the linear algebraic method, (ii) partial coloring, and (iii) Banaszczyk's method. Interestingly, the linear algebraic idea will also

play a key role in many of the algorithmic versions of partial coloring and Banaszczyk's method that we shall see later in Sections 3 and 4.

2.1. Linear algebraic method

This technique is simple but it can be surprisingly powerful. It is widely used in combinatorial optimization and is also referred to as iterated rounding [41].

For discrepancy problems it works as follows. Let $B \in \mathbb{R}^{m \times n}$ be the input matrix. The algorithm starts with the all-zero coloring $x_0 = (0, 0, \ldots, 0)$, and updates the coloring over several iterations $t = 1, 2, \ldots, T$, until the final coloring $x_T \in \{-1, 1\}^n$. The intermediate colorings satisfy $x_t \in [-1, 1]^n$, and once some color reaches ± 1 it is never updated again.

It remains to specify how the coloring is updated in each iteration. Call a variable j *alive* at time t, if $x_{t-1}(j) \in (-1, 1)$ and let A_t be the set of alive variables at the beginning of time t. The idea is to pick a suitable subset B_t of rows of B, with $\text{rank}(B_t) < |A_t|$, and consider some nonzero solution v_t satisfying (i) $B_t v_t = 0$ and (ii) $v_t(j) = 0$ for $j \in [n] \setminus A_t$. Such a solution exists as there are $|A_t|$ alive variables, and $\text{rank}(B_t) \leq |A_t|$.

The coloring is updated as $x_t = x_{t-1} + \delta v_t$, where $\delta > 0$ is chosen so that x_t stays in $[-1, 1]^n$ and at least one more color reaches ± 1 compared to x_{t-1}. The ingenuity lies in choosing B_t at each time t.

Let us see how to use this template to prove the Beck–Fiala theorem.

Theorem 1.2. *Any set system with maximum degree d has discrepancy at most $2d - 1$.*

Proof. Let B denote the incidence matrix of the set system. By our assumption, each column of B has at most d ones. Let us apply iterated rounding, where at iteration t we choose B_t to consist of rows S_i with $|A_t \cap S_i| > d$. Call such rows *large*. As each column of B has at most d ones, the number of ones in B restricted to columns in A_t is at most $d|A_t|$, and so the number of large rows is strictly less than $|A_t|$ and thus $\text{rank}(B_t) < |A_t|$.

To bound the final discrepancy, notice that as long as a row is large, its discrepancy stays 0. But once it has at most d alive elements, then no matter how these variables get updated in subsequent iterations, the additional discrepancy must be strictly less than $2d$ (e.g., if all the d alive variables were all -0.999 but get set to 1 eventually). As the final discrepancy of a set system is integral, this gives the bound of $2d - 1$. ∎

For more ingenious applications of this method to discrepancy, we refer to the survey by [17] and references therein.

2.2. Partial coloring method

We now describe the partial coloring lemma of Spencer and give some applications. We then describe the convex geometric proof of this result due to Gluskin [33] based on an exposition of Giannopoulos [32].

Theorem 2.1 (Partial coloring lemma). *Let A be an $m \times n$ matrix with rows $a_1, \ldots, a_m$. For each $i \in [m]$, let $\Delta_i = \lambda_i \|a_i\|_2$ be target discrepancy bound for row i. If the λ_i satisfy*

$$\sum_{i \in [m]} g(\lambda_i) \leq n/5, \tag{2.1}$$

where

$$g(\lambda) = \begin{cases} K \exp(-\lambda^2/9) & \text{if } \lambda > 0.1, \\ K \ln(\lambda^{-1}) & \text{if } \lambda \leq 0.1, \end{cases}$$

and K is some absolute constant. Then there exists $x \in \{-1, 0, 1\}^n$ with $|\{j : |x(j)| = 1\}| \geq n/10$ and $\operatorname{disc}(y, a_i) \leq \Delta_i$ for each $i \in [m]$.

Comparison with union bound. It is instructive to compare this with the standard union bound argument. For a random coloring x, as $\Pr[\operatorname{disc}(a_i, x) \geq \lambda \|a_i\|_2] \approx \exp(-\lambda_i^2/2) \approx g(\lambda_i)$. For the union bound to work, we need to choose λ_i to (roughly) satisfy the condition $\sum_i g(\lambda_i) < 1$. In contrast, Lemma 2.1 allows $\sum_i g(\lambda_i) = \Omega(n)$. This gives substantially more power. For example, suppose A is a 0–1 matrix corresponding to a set system. The union bound argument cannot ensure that $\Delta_i \ll |S_i|^{1/2}$ for even a couple of sets, while Theorem 2.1 allows us to set $\Delta_i < 1$ for $O(n/\log n)$ sets. As $x \in \{-1, 0, 1\}^n$, this in fact gives a partial coloring with exactly zero discrepancy for those sets!

2.2.1. Applications

The partial coloring method is very general and is widely used in discrepancy theory. We show how it directly gives Theorem 1.1 and the $O(d^{1/2} \log n)$ bound for the Beck–Fiala problem.

Proof of Theorem 1.1 for Spencer's problem. Let us assume that $m \geq n$. The case of $m \leq n$ follows by reducing $n = m$ by a standard linear algebraic trick. The coloring is constructed in phases. Let $n_0 = n$ and let n_k be the number of uncolored elements in phase k. In phase k, we apply Theorem 2.1 to the set system restricted to these n_k elements with $\Delta_k = c(n_k \log(2m/n_k))^{1/2}$ for each row, and verify that (2.1) holds for large enough $c = O(1)$. This gives a partial coloring on $\geq n_k/10$ elements, so $n_k \leq (0.9)^k n$ and summing up over the phases, total discrepancy is at most $\Delta_0 + \Delta_1 + \cdots = O((n \log(m/n))^{1/2})$.

***O*($d^{1/2} \log n$) discrepancy for the Beck–Fiala problem.** Again, the coloring is constructed in phases where $n_k \leq n(0.9)^k$ elements are uncolored in phase k. In phase k, let $s_{k,j}$ denote the number of sets with number of uncolored elements in the range $[2^j, 2^{j+1})$. Then $s_{k,j} \leq \min(m, n_k d/2^j)$ as the degree is at most d. A simple computation shows that (2.1) holds for $\Delta_i = cd^{1/2}$ for each i, for large enough $c = O(1)$. The result then follows directly as there are $O(\log n)$ phases and each set incurs $O(d^{1/2})$ discrepancy in each phase.

2.2.2. Proof of the partial coloring lemma

The original proof of Spencer was based on the pigeonhole principle and the entropy method and has several nice expositions, e.g., [1]. We sketch here the convex geometric proof of Gluskin [33].

The simple observation that ties discrepancy to geometry is the following.

Observation 2.1.1. *For a $m \times n$ matrix A, there is a coloring with discrepancy at most Δ_i for row a_i iff the polytope $P = \{x : |a_i x| \leq \Delta_i, i \in [m]\}$, contains some point in $\{-1, 1\}^n$.*

Similarly, Theorem 2.1 is equivalent to showing that the polytope P contains some point in $\{-1, 0, 1\}^n$ with at least $n/10$ nonzero coordinates, if the Δ_i satisfy (2.1).

Gluskin relates this property to the Gaussian volume of P. Call a convex body K symmetric if $x \in K$ implies $-x \in K$. Let $\gamma_n(x) = (2\pi)^{-n/2} \exp(-\|x\|_2^2/2)$ denote the standard n-dimensional Gaussian measure. For ease of exposition, we ignore the constants in Theorem 2.1.

Theorem 2.2 (Gluskin). *There is a small constant $\delta > 0$, such that any symmetric convex body $K \in \mathbb{R}^n$ with $\gamma_n(K) \geq 2^{-\delta n}$ contains $y \in \{-1, 0, 1\}^n$ with at least δn coordinates ± 1.*

The proof is based on a nice counting argument.

Proof. For $x \in \mathbb{R}^n$, let $K_x := K + x$ denote K shifted by x. As K is symmetric, we have that $\gamma_n(K_x) \geq \exp(-\|x\|^2/2)\gamma_n(K)$, as the densities of any two symmetric points y and $-y$ upon shifting by x satisfy,

$$\gamma_n(y + x) + \gamma_n(x - y) \geq 2\big(\gamma_n(x - y)\gamma_n(y + x)\big)^{1/2} = 2\exp\big(-\|x\|^2/2\big)\gamma_n(y).$$

Consider the 2^n copies K_x for all $x \in \{-1, 1\}^n$. As the total Gaussian volume of these copies is at least $2^n \exp(-n/2)\gamma_n(K) = 2^{cn}$, for some $c > 0$, there exists some point z contained in at least 2^{cn} copies. So there must exist some $x, x' \in \{-1, 1\}^n$ differing in $\Omega(n)$ coordinates such that z lies in both K_x and $K_{x'}$. Suppose $z = k_1 + x = k_2 + x'$ for some $k_1, k_2 \in K$. Then $y := (x - x')/2 = (k_2 - k_1)/2 \in K$ as K is symmetric and convex, and, moreover, $y \in \{-1, 0, 1\}^n$ with $\Omega(n)$ coordinates ± 1. ∎

Theorem 2.1 follows by relating the condition (2.1) on λ_i to the volume $\gamma_n(P)$.

Gaussian measure of polytopes. For a vector $a \in \mathbb{R}^n$ and scalar $b > 0$, define the slab $S(a, b) = \{x : |\langle x, a\rangle| \leq b\}$. Then $S(a, \lambda\|a\|_2)$ is symmetric and convex, with measure $\gamma_n(S(a, b)) = \gamma_1([-\lambda, \lambda])$, and $P = \bigcap_{i=1}^m S(a_i, \lambda_i\|a_i\|_2)$ is an intersection of slabs.

By the Sidak–Khatri lemma (see, e.g., [33]), for any symmetric convex body K and slab S, we have that $\gamma_n(K \cap S) \geq \gamma_n(K)\gamma_n(S)$, and hence $\gamma_n(P) \geq \prod_i \gamma_n(S(a_i, \lambda_i\|a_i\|_2)$. As $\gamma_1([-\lambda, \lambda]) \approx 1 - O(\exp(-\lambda^2/2))$ for $\lambda \geq 1$ and $O(\lambda)$ for $\lambda < 1$, we have that $\log(\gamma_1[-\lambda, \lambda]) \approx -g(\lambda)$, and thus condition (2.1) implies that $\gamma_n(P) \geq 2^{-\delta n}$.

2.3. Banaszczyk method

A problem with the partial coloring method is that it requires $O(\log n)$ rounds to obtain a full coloring, which can result in an extra $O(\log n)$ factor loss in the discrepancy bound, as we saw for the Beck–Fiala problem. The following result of Banaszczyk [4] gives a way to find a full coloring directly and can give better results. The form of the result also makes it broadly applicable in other settings.

Theorem 2.3 ([4]). *Given any convex body $K \subseteq \mathbb{R}^m$ of Gaussian measure $\gamma_m(K) \geq 1/2$, and vectors $v_1, \ldots, v_n \in \mathbb{R}^m$ of ℓ_2 norm at most $1/5$, there exists a coloring $x : [n] \to \{-1, 1\}$ such that $\sum_{j=1}^n x(j)v_j \in K$.*

While this statement looks similar to Theorem 2.1, a crucial difference is that the Gaussian measure and convex body K here are in the *output space* and are m-dimensional, while K in Theorem 2.1 is in the *input space* and n-dimensional.

The proof of Theorem 2.3 involves some delicate computation and a non-trivial idea of Ehrhard symmetrization. However, the main idea is very clean that we sketch below.

Proof. The key step is to show that for any convex body K with $\gamma_m(K) \geq 1/2$ and $u \in \mathbb{R}^m$ with $\|u\|_2 \leq 1/5$, there is a convex body $K * u$ contained in $(K - u) \cup (K + u)$ such that $\gamma_m(K * u) \geq 1/2$.

Given this fact, Theorem 2.3 follows by induction on the number of vectors n. It trivially holds for $n = 0$ as $\mathbf{0} \in K$ as $\gamma_m(K) \geq 1/2$. Suppose inductively that it holds for some $n - 1$. Consider the convex body $K' = K * v_n$. As $\gamma_m(K') \geq \gamma_m(K) \geq 1/2$, by induction there exist $x(1), \ldots, x(n-1) \in \{-1, 1\}$ such that $u = x(1)v_1 + \cdots + x(n-1)v_{n-1} \in K'$. But as $K' \subset (K - v_n) \cup (K \cup v_n)$, at least one of $u + v_n$ or $u - v_n$ must lie in K, giving the sign $x(n)$ such that $u + x(n)v_n = \sum_{i=1}^n x(i)v_i \in K$. ∎

Bound for the Komlós problem. Theorem 2.3 directly gives the $O((\log n)^{1/2})$ bound for the Komlós problem. This follows as $\mathbb{E}\|g\|_\infty = O((\log m)^{1/2})$ for a random Gaussian vector g in $\mathbb{R}^m$, and so choosing K to be $2\mathbb{E}\|g\|_\infty$ times the unit ℓ_∞-ball in $\mathbb{R}^m$, by Markov's inequality we have that $\gamma_m(K) \geq 1/2$. As Theorem 2.3 requires $\|v\|_2 \leq 1/5$, we can further scale K by a factor of 5. Finally, we can assume $m \leq n^2$, as $\|v_j\|_2 \leq 1$ for each j implies that at most n^2 rows a_i can have $\|a_i\|_1 \geq 1$.

Banaszczyk's theorem for prefix discrepancy. In a subsequent work, Banaszczyk [5] further extended this result to handle prefixes, using a clever inductive argument.

Theorem 2.4 ([5]). *Given vectors $v_1, \ldots, v_n \in \mathbb{R}^m$ of ℓ_2 norm at most $1/5$ and any convex body $K \subseteq \mathbb{R}^m$ with $\gamma_m(K) \geq 1 - 1/(2n)$, there exists a coloring $x : [n] \to \{-1, 1\}$ such that each for $k = 1, \ldots, n$, the prefix sum satisfies $\sum_{j=1}^k x(j)v_j \in K$.*

Proof. Consider the sequence of symmetric convex bodies K_j defined iteratively as $K_n = K$ and $K_j = (K_{j+1} * v_{j+1}) \cap K$, for $j = n-1, \ldots, 1$. We first show that $\gamma_m(K_j) \geq 1 - (n - j + 1)/2n$ for $j \in [n]$ by backwards induction. Indeed, $\gamma_m(K_n) = \gamma_m(K) \geq 1 - 1/(2n)$ in the base case. If this it holds for some $j \leq n$, then

$$\begin{aligned}\gamma_m(K_{j-1}) &= \gamma_m(K_j * v_j) \cap \gamma_m(K) \geq \gamma_m(K_j * v_j) - \big(1 - \gamma_m(K)\big) \\ &\geq \gamma_m(K_j) - \big(1 - \gamma_m(K)\big) \geq 1 - \frac{n - j + 1}{2n} - \frac{1}{2n} = 1 - \frac{n - j}{2n},\end{aligned}$$

where we use that $\gamma_m(K_j * v_j) \geq \gamma_m(K_j)$ as $\gamma_m(K_j) \geq 1/2$.

So $\gamma_m(K_1) \geq 1/2$ and K_1 is convex, and a simple calculation shows that either v_1 or $-v_1$ lies in K_1. We now apply induction in the forward direction. Suppose there is some

$j \geq 1$ such that there are signs $x(1), \ldots, x(j)$ satisfying (i) $u := \sum_{i=1}^{j} x(i)v_i \in K_j$ and (ii) $\sum_{i=1}^{k} x(i)v_i \in K$ for all $k \leq j$. To continue the induction, we need to show that (i) $u \in K$ and (ii) that there is a sign $x(j+1)$ such that $u + x(j+1)v_{j+1} \in K_{j+1}$. Now, $u \in K$ clearly holds as by (i) we have $u \in K_j \subset K$. Now, for the sake of contradiction suppose that both $u + v_{j+1}$ and $u - v_{j+1} \notin K_{j+1}$. Then $u \notin K_{j+1} + v_{j+1} \cup K_{j+1} - v_{j+1}$ and hence $u \notin K_{j+1} * v_{j+1}$. By definition, as $K_j = K \cap (K_{j+1} * v_{j+1}) \subset K_{j+1} * v_{j+1}$, this contradicts our inductive assumption that $u \in K_j$. ■

Bound for prefix Komlós. Theorem 2.4 directly implies an $O((\log n)^{1/2})$ discrepancy for the prefix version of the Komlós problem. In particular, the condition that $\gamma_m(K) \geq 1 - 1/(2n)$ instead of $\geq 1/2$ in Theorem 2.3 make no difference beyond a constant factor as $\Pr[\|g\|_\infty \geq \mathbb{E}\|g\|_\infty + t] \leq \exp(-t^2/2)$ by concentration for Lipschitz functions of Gaussians, and choosing $t = O((\log n)^{1/2})$.

3. ALGORITHMS FOR PARTIAL COLORING

In the next few sections we describe the progress on making these results algorithmic. We first describe several different algorithmic proofs for partial coloring. In Section 4 we describe the algorithmic approaches for Banaszczyk's method as stated in Theorem 2.3. In Section 5, we describe an algorithm to approximate the hereditary discrepancy of any arbitrary matrix.

The algorithms for partial coloring can be divided into two types: either based on a random walk approach, or a direct optimization based approach.

Random-walk based approaches. Bansal [6] gave the first algorithm for various applications of partial coloring such as the $O(n^{1/2})$ bound for Spencer's problem with $m = O(n)$ sets and the $O(d^{1/2} \log n)$ bound for the Beck–Fiala problem. Subsequently, Lovett and Meka [45] designed an elegant and substantially simpler algorithm that gave an algorithmic version of the full partial coloring lemma as stated in Theorem 2.1.

These algorithms can be viewed as a randomized version of the iterated rounding method, where one starts with the all zero-coloring, and updates the variables gradually using a correlated Brownian motion with small discrete steps. The variables are fixed once they reach ± 1, and correlations between the variables are chosen to ensure that each row has low discrepancy. Bansal's algorithm was based on solving a suitable semidefinite program (SDP) at each time step to generate the covariance matrix for the random walk. Lovett and Meka showed that one can simply do a standard discrete Brownian motion in the subspace orthgonal to tight discrepancy constraints, without the need to solve any SDPs.

Direct methods. Later, Rothvoss [59] further extended the result of Lovett and Meka from polytopes to general symmetric convex bodies and gave an algorithmic version of Theorem 2.2. His algorithm is extremely elegant and simple to describe. A related algorithm was given by Eldan and Singh [28]. Both these algorithms are based on solving a very simple optimization problem.

We now describe these algorithms and sketch the main ideas behind their analysis.

3.1. The SDP-based approach

We start with the SDP-based approach. Even though the latter algorithms are more general and simpler, this approach is very natural and motivates why the Brownian motion is needed. It is also the only approach we know for some problems such as the algorithmic version of Theorem 1.5, that we describe below in Theorem 3.1. More importantly, it is quite flexible and can be extended in various ways by adding new SDP constraints, as we shall see later in Section 4.

A relaxation for discrepancy. Given an input matrix A, a natural approach to find a low discrepancy coloring for it is to first solve some convex programming relaxation and then try to round the solution suitably to ± 1. Let us first consider linear programming relaxations.

Recall that a linear program (LP) consists of variables $x_1, \ldots, x_n \in \mathbb{R}$, and the goal is to optimize some linear objective $c^T x$ subject to linear constraints $a_i^T x \leq b_i$ for $i \in [m]$. LPs can be solved optimally in time polynomial in n, m, and the bit length of the input.

Let a_i denote the ith row of A, then the natural LP relaxation for discrepancy is,

$$\min t \quad \text{s.t.} \quad -t \leq a_i x \leq t, \ \ \forall i \in [m] \ \text{ and } \ -1 \leq x_j \leq 1, \ \ \forall j \in [n].$$

However, this always has the trivial solution $x = \mathbf{0}$ with objective $t = 0$, which is useless.

So let us consider a more general class of optimization problems called semidefinite programs (SDPs). An SDP can be viewed as an LP with variables of the form x_{ij} for $1 \leq i, j \leq n$, arranged as entries of an $n \times n$ matrix X, where we require that X be symmetric and positive semidefinite, denoted by $X \succeq 0$. For matrices A, B, let $\langle A, B \rangle = \mathrm{Tr}(A^T B) = \sum_{ij} A_{ij} B_{ij}$ denote the trace inner product. An SDP is an optimization problem of the form

$$\max \langle C, X \rangle \quad \text{s.t. } \langle A_k, X \rangle \leq b_k, \ 1 \leq k \leq m, \ X \succeq 0,$$

where $C, A_1, \ldots, A_m \in \mathbb{R}^{n \times n}$.

SDPs can be solved to any desired level of accuracy in polynomial time. As $X \succeq 0$ iff it is the Gram matrix of some vectors $w_1, \ldots, w_n \in \mathbb{R}^n$, i.e., $X_{ij} = \langle w_i, w_j \rangle$, SDPs can be viewed as *vector programs* where the variables are the vectors w_i and we can impose any linear constraints on their inner products (but not on the w_i themselves).

SDP relaxation for discrepancy. Let λ be some upper bound on the discrepancy $\mathrm{disc}(A)$, and consider the following SDP:

$$\left\| \sum_j a_{ij} w_j \right\|_2^2 \leq \lambda^2 \quad \text{for } i \in [m], \ \|w_j\|_2^2 = 1, \ i \in [n],$$

Let us call a feasible solution to this SDP a vector-coloring for A, and the smallest λ for which it is feasible as the vector discrepancy, $\mathrm{vecdisc}(A)$. Clearly, $\mathrm{vecdisc}(A) \leq \mathrm{disc}(A)$.

At first glance, this SDP also does not seem useful. For example, for Spencer's problem, the solution $w_i = e_i$, where w_i is the ith standard basis vector, is always feasible with $\lambda = n^{1/2}$, irrespective of the matrix A. However, this SDP becomes quite useful

when $\lambda \ll n^{1/2}$ as it gives nontrivial correlations between the vectors w_i that we can exploit. Below, we describe a very simple algorithm that gives the following algorithmic version of Theorem 1.5.

Theorem 3.1 ([6]). *Given any $A \in \mathbb{R}^{m \times n}$, there is an efficient algorithm that, with high probability, finds a coloring with discrepancy $O((\log m \log n)^{1/2}\, \mathrm{herdisc}(A))$.*

Later, we will also describe a simple $O((n \log\log\log n)^{1/2})$ bound for Spencer's problem with $m = O(n)$ using this SDP. This is perhaps surprising, as apriori the naive solution $w_i = e_i$ does not give any meaningful correlations between the elements and corresponds to random coloring.

3.1.1. Algorithm for Theorem 3.1

Before describing the algorithm, it is instructive to see why a direct approach for rounding the SDP does not work.

Problem with direct rounding. For simplicity, let us suppose that $\lambda = 0$ for some matrix A. Then the vectors $w_1, \ldots, w_n$ produced by the SDP solution are nicely correlated so that $\sum_j a_{ij} w_j = 0$ for each row i.

To convert the w_j into scalars while preserving the correlations, let us pick a random Gaussian vector $g \in \mathbb{R}^n$, with each coordinate $g_k \sim N(0,1)$ independently and project the vectors w_j on g to obtain $y_j = \langle w_j, g\rangle$. Then as the g_k are iid $N(0,1)$, we have that $\langle g, w\rangle = \sum_k g_k w(k) \sim N(0, \|w\|_2^2)$ for any vector $w \in \mathbb{R}^n$, and hence (i) $y_j \sim N(0,1)$ for each j as $\|w_j\|_2^2 = 1$ and (ii) $\sum_j a_{ij} y_j = 0$ for each row i. This seems very close to what we want except that $y_j \sim N(0,1)$ instead of ± 1.

However, the following hardness result of Charikar, Newman, and Nikolov [21] rules out any reasonable way of rounding these y_j to ± 1.

Theorem 3.2 ([21]). *Given a set system on n elements and $m = O(n)$ sets, it is NP-hard to distinguish whether it has discrepancy 0 or $\Omega(\sqrt{n})$.*

In particular, this implies that there must exist set systems with discrepancy $\Omega(\sqrt{n})$ but vector-discrepancy 0 (otherwise solving the SDP would give an efficient way to distinguish between set systems with discrepancy 0 and $\Omega(\sqrt{n})$).

Discrete Brownian motion. So instead of trying to round the y_j's directly to ± 1, the algorithm will gradually obtain a ± 1 coloring by combining solutions of various SDPs over time. We first give a overview of the algorithm.

More precisely, at time 0, we start with the coloring $x_0 = (0, \ldots, 0)$ and modify it over time as follows. Let x_{t-1} denote the fractional coloring at time $t-1$. Then $x_t = x_{t-1} + \Delta x_t$ is obtained by adding a small update vector Δx_t to x_{t-1}. As the perturbations are added, the colors evolve over time, and once a color reaches ± 1 it is frozen and no longer updated. The updates Δx_t are obtained by solving the SDP with $\lambda = \mathrm{herdisc}(A)$, restricted to the alive elements and setting $\Delta x_t(j) = \gamma \langle g, w_j\rangle$, where g is a random gaussian and γ is a small multiplier.

Formal description. Let $\gamma = \max_{i,j} |A_{ij}|/n$ and $\ell = 8 \log n/\gamma^2$. Let x_t and A_t denote the coloring and the set of alive (unfrozen) variables at the end of time t. Let $\lambda = \text{herdisc}(A)$.

(1) Initialize $x_0(j) = 0$ for $j \in [n]$ and $A_0 = \emptyset$.

(2) At each time step $t = 1, 2, \ldots, \ell$, do the following.
Solve the following SDP:

$$\left\| \sum_j a_{ij} w_j^t \right\|_2^2 \leq \lambda^2 \quad \forall i \in [m], \quad \|w_j^t\|_2^2 = 1 \text{ if } j \in A_{t-1}, \text{ else } \|w_j^t\|_2^2 = 0.$$

Pick a random Gaussian $g_t \in \mathbb{R}^n$, and set $x_t(j) = x_{t-1}(j) + \gamma \langle g_t, w_j^t \rangle$.
Set $A_t = \{j : |x_t(j)| < 1\}$.

(3) Set $x_\ell(j) = -1$ if $x_\ell(j) < -1$ and $x_\ell(j) = 1$ otherwise. Output x_ℓ.

Analysis. We now sketch the ideas behind Theorem 3.1. First, notice that as $\lambda = \text{herdisc}(A)$, the SDP above is always feasible no matter which variables are alive.

Let us now see how the colors of the elements and the discrepancies of the rows evolve over time. Fix some element j. Its color $x_t(j)$ starts at 0 at $t = 0$ and evolves as a martingale with updates $\Delta x_t(j) = \gamma \langle w_j^t, g_t \rangle$ until it is frozen. As $\|w_j^t\| = 1$, we have $\Delta x_t(j) \sim N(0, \gamma^2)$ and thus $x_t(j)$ will reach ± 1 in $O(1/\gamma^2)$ steps with constant probability. As there are $\ell = O(\log n/\gamma^2)$ steps, whp all elements will reach ± 1, by the end of the algorithm.

Now fix some row i. Its discrepancy $x_t(a_i) := \sum_j a_{ij} x_t(j)$ is 0 at $t = 0$, and evolves as $\sum_j a_{ij} \Delta x_t(j) = \sum_j \gamma \langle g_t, \sum_j a_{ij} w_j^t \rangle$ at step t. As $\| \sum_j a_{ij} w_j^t \|^2 \leq \lambda^2$, the sequence $x_t(a_i)$ forms a martingale with Gaussian increments with variance at most $\gamma^2 \lambda^2$. As $\ell = O(\log n/\gamma^2)$, by standard martingale concentration and union bound over the m constraints, each row has final discrepancy $O(\ell^{1/2} \cdot \gamma\lambda \cdot (\log m)^{1/2}) = O(\lambda (\log m \log n)^{1/2})$ whp.

Finally, whp truncating $x_\ell(j)$ to ± 1 introduces negligible error for any row. This follows as $\Delta x_t(j) \sim N(0, \gamma^2)$ we have that whp $|x_t(j)| < 1 + \gamma \cdot O((\log n)^{1/2})$ when it freezes. As $\text{herdisc}(A) \geq \max_{ij} |A_{ij}|$ and $\gamma = \max_{ij} |A_{ij}|/n \leq \text{herdisc}(A)/n$, the rounding error is negligible.

3.1.2. Algorithmic version of Spencer's result

The above approach is quite flexible, e.g., the discrepancy bounds λ_i^t for each row i and be chosen adaptively at time t. We describe a simple version of this idea that already gives a $\beta n^{1/2}$ for $\beta = c(\log \log \log n)^{1/2}$ bound for Spencer's problem with $m = n$ sets, and thus beats random coloring.

As in Section 2.2.1, it suffices to obtain a partial coloring with $O(\beta n^{1/2})$ discrepancy. Let us run the algorithmic template above for $\ell = 100/\gamma^2$ steps, using the following

SDP relaxation for partial coloring at each time t:

$$\left\| \sum_{j \in S_i} w_j \right\|_2^2 \leq \lambda_i^2 \quad \text{for } i \in [m], \tag{3.1}$$

$$\sum_{j \in A_{t-1}} \|w_j\|_2^2 \geq |A_{t-1}|/10, \tag{3.2}$$

$$\|w_j\|_2^2 \leq 1 \quad \forall j \in A_{t-1}, \text{ else } \|w_j\|^2 = 0.$$

Notice that the λ_i on the right hand side in (3.1) can be different for each rows. The constraint (3.2) says that at least $|A_{t-1}|/10$ elements must be colored.

The bounds λ_i are set as follows. Initially, $\lambda_i = cn^{1/2}$ for each S_i where c is a large enough constant. If the discrepancy $|x_t(S_i)|$ for S_i exceeds $\beta n^{1/2}$ at any time, we label S_i *dangerous* and set $\lambda_i^2 = n/\log n$ at all future time steps.

The result follows from the following two observations.

Lemma 3.3. *If the SDPs are feasible at all time steps, then whp each set has discrepancy $O(\beta n^{1/2})$, and at least $\Omega(n)$ elements are colored ± 1 at the end of the algorithm.*

Proof. (Sketch) By the choice of the λ_i, once a set becomes dangerous, its discrepancy evolves as a martingale with Gaussian increments with variance at most $\gamma^2 n/\log n$. As there at most $\ell = O(\gamma^{-2})$ time steps, whp each set incurs an additional discrepancy of at most $O(n^{1/2})$.

Next, the variance $\mathbb{E}[\Delta x_t(j)^2]$ increases by at least $\gamma^2/10$ on average for the alive variables at each step t by the constraint (3.2). As $\ell = 100\gamma^{-2}$, a simple Markov argument shows that a constant fraction of the elements will reach ± 1 with at least constant probability. ∎

Lemma 3.4. *With probability $1 - o(1)$, all the SDPs are feasible.*

Proof. (Sketch) As $\lambda_i \leq O(n^{1/2})$ at each time and $\ell = O(\gamma^{-2})$, each set S_i has discrepancy $O(n^{1/2})$ in expectation. So by standard martingale concentration, with probability $1 - o(1)$, the fraction of sets that ever become dangerous $2\exp(-\Omega(\beta^2)) \ll (\log\log n)^{-2}$ for c large enough. Let us condition on this event. We will show that the SDP is feasible at each step using Theorem 2.1. Indeed, as each dangerous set S_i contributes $g(\Delta_i/|S_i|^{1/2}) \leq g(1/\log n) \leq K \log\log n$ to (2.1), the dangerous sets contribute

$$O\big(n/(\log\log n)^2\big) \cdot K \log\log n = o(n)$$

in total. As $\lambda_i = cn^{1/2}$ for the other sets and $m = n$, their total contribution is also at most $n/10$ for c large enough. ∎

3.2. The Lovett–Meka algorithm

Lovett and Mekka [45] substantially simplified the random-walk approach and extended it to give the following algorithmic version of the general partial coloring lemma.

Theorem 3.5. *Given an input matrix $A \in \mathbb{R}^{m\times n}$ and some fractional coloring $x_0 \in [-1,1]^n$ with k alive elements, for $i \in [m]$ let λ_i be such that*

$$\sum_i \exp\left(-\lambda_i^2/16\right) \le k/16. \tag{3.3}$$

Then there is a randomized polynomial time algorithm to find a coloring x with at most $k/2$ alive variables such that $|x(a_i) - x_0(a_i)| \le \lambda_i \|a_i\|_2$ for each row $i \in [m]$.

We remark that the colors produced by Theorem 3.5 lie in $[-1, 1]$, in contrast to $\{-1, 0, 1\}$ in Theorem 2.1, but this does not make any difference. Theorem 3.5 is also slightly stronger than Theorem 2.1 for $\lambda_i \ll 1$.

The key idea of the algorithm is that whenever a discrepancy constraint becomes tight or some variable reaches ± 1, one can simply do a random walk orthogonal to it. We now describe it formally. Without loss of generality, we assume that all variables are initially alive, i.e., $k = n$.

3.2.1. The algorithm

Let A_{t-1} be the set of alive variables at the beginning of time t. The algorithm maintains a linear subspace $V_{t-1} \subset \mathbb{R}^n$. Initially at $t = 1$, the coloring is x_0, $A_0 = [n]$ and $V_0 = \mathbb{R}^n$. The following is repeated for $\ell = O(\gamma^{-2})$ steps.

At time t, the algorithm chooses a random gaussian vector g_t in the subspace V_{t-1} and updates $x_t = x_{t-1} + \gamma g_t$, where γ is a small step size as usual.

(1) If $|x_t(j)| \ge 1$, set $V_t = V_{t-1} \cap e_j^{\perp}$, so that $x(j)$ will not be updated anymore.

(2) If $|x_t(a_i)| \ge \lambda_i \|a_i\|_2$, set $V_t = V_{t-1} \cap a_i^{\perp}$, so that row i incurs no further discrepancy.

Analysis. We assume that γ is small enough so that we can ignore the rounding error in the sketch below. By design, the algorithm ensures that $x_t(j) \in [-1, 1]$ for all j and that $x_t(a_i) \le \lambda_i \|a_i\|_2$ for all i. We now show that, with constant probability, at least half the variables reach ± 1.

For a linear subspace V, let $N(V)$ denote the standard multidimensional Gaussian distribution supported on V. By rotational invariance, a random vector $g \sim N(V)$ can be written as $g = g(1)v_1 + \cdots + g(d)v_d$ for some orthonormal basis $\{v_1, \ldots, v_d\}$ for V and $g(1), \ldots, g(d)$ iid $N(0, 1)$. We note the following fact.

Lemma 3.6. *Let V be a d-dimensional subspace of $\mathbb{R}^n$ and $g \sim N(V)$. Then for all $u \in \mathbb{R}^n$, $\langle g, u\rangle \sim N(0, \sigma^2)$ where $\sigma^2 \le \|u\|^2$. Moreover, for $i = 1, \ldots, n$ let σ_i be such that $\langle g, e_i\rangle \sim N(0, \sigma_i^2)$. Then $\sum_{i=1}^n \sigma_i^2 = d$.*

Proof. Let u' denote the projection of u onto V. Clearly, $\|u'\| \le \|u\|$. As $g \in V$, $\langle g, u\rangle = \langle g, u'\rangle$ and hence $\langle g, u\rangle \sim N(0, \|u'\|^2)$. For the second part, if $v_1, \ldots, v_d$ is an orthogonal basis for V, then $\sigma_i^2 = \sum_{j=1}^d \langle e_i, v_j\rangle^2$. Thus $\sum_{i=1}^n \sigma_i^2 = \sum_{i=1}^n \sum_{j=1}^d \langle e_i, v_j\rangle^2 = \sum_{j=1}^d \sum_{i=1}^n \langle v_j, e_i\rangle^2 = \sum_{j=1}^d \|v_j\|^2 = d$. ∎

Proof of Theorem 3.5. (Sketch) First we claim that in expectation, not many discrepancy constraints become tight in step (2) of the algorithm. This follows as for any time t, by Lemma 3.6 the discrepancy increment for each row a_i is distributed as $N(0, \leq \gamma^2 \|a_i\|^2)$. As $\ell = O(\gamma^{-2})$, by standard tail bounds $\Pr[|x_\ell(a_j) - x_0(a_j)| \geq \lambda_i \|a_i\|_2] = \exp(-\Omega(\lambda_i^2))$. As the λ_i satisfy (3.3), choosing the constants appropriately, the probability that more than $n/8$ discrepancy constraints becomes tight is at most $1/8$.

Let us condition on the above event. The proof now follows from a win–win argument. If more than $n/2$ elements reach ± 1, we are already done. If this does not happen, then at any time during the algorithm the subspace V_t has dimension at least $n - n/2 - n/8 \geq 3n/8$. By Lemma 3.6, as $\sum_j \mathbb{E}[\Delta x_t(j)^2] \geq (3n/8)\gamma^2$ and $\ell = O(\gamma^{-2})$ steps, the *energy* $\sum_j (x_\ell(j)^2 - x_0(j)^2)$ must increase by $\Omega(n)$ in expectation. But as $x_\ell(j)^2 - x_0(j)^2 \in [-1, 1]$ for all j, a simple argument can be used to show that at least $\Omega(n)$ variables reach ± 1 in expectation.

3.3. Direct approaches

The Lovett–Meka algorithm crucially uses the face structure of the polytope and does not seem to generalize to general convex bodies in the sense of Theorem 2.2. In particular, even if $\gamma_n(K) \geq 2^{-\delta n}$, condition (2.1) may not hold as it might require exponentially many facets to obtain any reasonable approximation of a general convex body K.

We now describe an extremely elegant and simple to state the algorithm due to Rothvoss [59], that finds a partial coloring in general convex bodies. We then describe a related algorithm by Eldan and Singh [28].

3.3.1. Rothvoss' algorithm

Let K be a symmetric convex body with $\gamma(K) \geq 2^{-\delta n}$. Take a random Gaussian $g \in \mathbb{R}^n$, and output the point closet to g in the body $K \cap [-1, 1]^n$, i.e., output

$$x^* = \operatorname{argmin}\{\|g - x\|_2 : x \in K \cap [-1, 1]^n\}.$$

That's it! The point x^* can be computed by a convex program, using a membership oracle for K.

Theorem 3.7 ([59]). *Let $\varepsilon > 0$ be a sufficiently small constant and $\delta := (3/2)\varepsilon \log_2(1/\varepsilon)$, and let K be a symmetric convex body with $\gamma_n(K) \geq \exp(-\delta n)$. Then whp, x^* has at least εn many coordinates ± 1.*

Analysis. The proof is also very elegant and uses Gaussian concentration for Lipschitz functions and the Sidak–Khatri lemma in a clever way.

The starting observation is that the distance $d(g, x^*)$ is at least $n^{1/2}/5$ with probability $1 - \exp(-\Omega(n))$. This follows as $x^* \in [-1, 1]^n$ and as $g(j) \sim N(0, 1)$ for each coordinate j, we have $\Pr[|g(j)| \geq 2] \geq 1/25$. On the other hand, $d(g, K) \leq 3(\delta n)^{1/2}$ with probability $1 - \exp(-\Omega(n))$ by Gaussian concentration for Lipschitz functions as $\gamma_n(K) \geq \exp(-\delta n)$.

Now, suppose for the sake of contradiction that fewer than εn coordinates of x^* are ± 1 for some g. Let I be the set of these coordinates. A key observation is that if x^* is an optimum solution to some convex program, and some constraint is not tight at x^*, then x^* remains optimum even when this constraint is removed. So x^* would still be the optimum solution, if replace $K \cap [-1,1]^n$ in the convex program by $K \cap S(I)$, where $S(I) = \{x : |x(j)| \le 1, j \in I\}$ is the intersection of the slabs corresponding to coordinates in I.

By the Sidak–Khatri lemma, $\gamma_n(K \cap S(I)) \ge \gamma_n(K)\gamma_n(S(I)) \ge \exp(-(\epsilon + 2\delta)n)$, and hence by Gaussian concentration the distance $d(g, x^*) = d(g, K \cap S(I)) \le 6(\varepsilon + \delta)n^{1/2}$ with probability $1 - \exp(-\Omega(n))$. So even if after a union bound over the $\approx \exp(\delta \ln(1/\delta)n)$ possible choices for I, one has $d(x^*, g) = O((\varepsilon + \delta)n^{1/2})$ whp. This contradicts the first observation that $d(x^*, g) \ge n^{1/2}/5$ whp.

3.3.2. Eldan–Singh algorithm

This algorithm is as simple to state and only requires linear optimization: Pick a random direction $c \in R^n$ and optimize over $K \cap [-1,1]^n$, i.e., output

$$x^* = \operatorname{argmax}\{c^T x : x \in K \cap [-1,1]^n\}.$$

Eldan and Singh [28] showed a result similar to Theorem 3.7. That is, for any $\epsilon > 0$ small enough, there is a $\delta > 0$ such that if $\gamma_n(K) \ge 2^{-\delta n}$ then whp x^* has at least εn coordinates ± 1 with constant probability.

4. ALGORITHMIC VERSION OF BANASZCZYKS'S RESULT

We now consider the algorithmic approaches for Banaszczyk's method. The first progress was by Bansal, Dadush, and Garg [8], who gave an efficient SDP-based algorithm to find an $O((\log n)^{1/2})$ discrepancy coloring for the Komlós problem. A deterministic algorithm for the problem was subsequently obtained by Levy, Ramadas, and Rothvoss [42].

Later, Bansal, Dadush, Garg, and Lovett [9] gave an algorithm for the general case of Banaszczyk's theorem with arbitrary convex body K. Their algorithm, called the Gram–Schmidt walk, combines linear algebra and random walks. Recently, Harshaw et al. [35] gave an optimal analysis of this walk.

We describe both these approaches below. We mention that finding an efficient algorithm for the prefix version of Banaszczyk's problem in Theorem 2.4 is still open.

Problem 4.1. Find an efficient algorithm for the prefix version of Banaszczyk's theorem. The case of prefix Komlós (Problem 1.4) would already be very interesting.

4.1. The Komlós problem

We describe the following result of Bansal, Dadush, and Garg [8].

Theorem 4.2 ([8]). *Given vectors $v_1, \dots, v_n \in \mathbb{R}^m$ with $\|v_j\|_2 \le 1$ for $j \in [n]$, there is a polynomial time algorithm that finds an $O((\log n)^{1/2})$ discrepancy coloring whp.*

The algorithm is based on SDPs and is similar to that in Section 3.1, but it adds some extra constraints to the SDP so that the resulting solution has some additional desirable properties. To understand these properties, it is instructive to see what can go wrong with the partial coloring approach. To focus on the main ideas we consider the special case of Beck–Fiala problem, with the goal of finding an $O((d \log n)^{1/2})$ discrepancy coloring.

Recall that the $O(d^{1/2} \log n)$ bound using partial coloring was obtained by requiring zero discrepancy for large sets S, say of size $> 10d$. For small sets of size $s \leq 10d$, we can set the bound roughly $O(s^{1/2})$ (in Section 2.2.1 we used the bound $d^{1/2}$, but $O(s^{1/2} \ln(20d/s))$ also works). So as long as a set is large, it incurs zero discrepancy, and once it is small it incurs at most $O(d^{1/2})$ discrepancy in each partial coloring step.

The ideal process. Ideally, one would expect that once a set S becomes small, then whenever a constant fraction of the elements get colored colored globally in a partial coloring step, the size of S should also decrease geometrically. If so, this would actually give an $O(d^{1/2})$ discrepancy. However, the problem is that partial coloring does not give much control on which elements get colored, e.g., sets can incur discrepancy $O(d^{1/2})$ even if only $O(d^{1/2})$ of their elements get colored. This imbalance between the discrepancy and the *progress* a set makes in getting colored is the main barrier to improving the $O(d^{1/2} \log n)$ bound.

A concrete bad example. To see this more explicitly, let us consider the Lovett–Meka algorithm. Suppose the subspace V_{t-1} at time t is spanned by the orthonormal basis b, $e_{d+1}, \ldots, e_n$ where $b = d^{-1/2}(e_1 + \cdots + e_d)$. Then any update $\Delta x_t \in V_{t-1}$ has $\Delta x_t(1) = \cdots = \Delta x_t(d)$, and for the set $S = \{1, \ldots, d\}$, all variables get updated by the same amount, so if it incurs discrepancy $d^{1/2}$, the coloring progress is only $d^{1/2}$. In contrast, if the $\Delta x_t(1), \ldots, \Delta x_t(d)$ were independent, $\Omega(d)$ elements would get colored in expectation while incurring a discrepancy of $d^{1/2}$.

The key idea behind the algorithm of [8] is to ensure that even though the update Δx_t lies in some subspace that we cannot control, the coordinates $\Delta x_t(j)$ behave roughly independently in the sense that

$$\mathbb{E}\left[\left(\sum_j b(j)\Delta x_t(j)\right)^2\right] \leq \eta\left(\sum_j b(j)^2 \mathbb{E}\left[\Delta x_t(j)^2\right]\right) \quad \forall b \in \mathbb{R}^n, \tag{4.1}$$

where $\eta \geq 1$ is some fixed constant. Notice that if the $\Delta x_t(j)$ were independent or even pairwise independent, then (4.1) would be an equality with $\eta = 1$.

The algorithm will add an additional SDP constraint to ensure property (4.1). We describe this below and then give a sketch of the analysis.

4.1.1. Algorithm

Let (U, C) be the input set system. As usual, the algorithm starts with the coloring $x_0 = 0^n$. Let x_{t-1}, A_{t-1} denote the coloring and the set of alive variables at the beginning of t. Call a set $S \in C$ large if $|S \cap A_{t-1}| \geq 10d$.

Repeat the following for $t = 1, 2, \ldots, \ell$ until $A_\ell = \emptyset$.

(1) Solve the following SDP:

$$\left\| \sum_{j \in S} w_j \right\|^2 = 0 \quad \text{for all large } S, \tag{4.2}$$

$$\left\| \sum_j b(j) w_j \right\|^2 \leq 2 \sum_j b(j)^2 \|w_j\|^2 \quad \forall b \in \mathbb{R}^n, \tag{4.3}$$

$$\|w_j\|^2 \leq 1 \quad \text{for } j \in A_{t-1}, \text{ and else } \|w_j\|^2 = 0, \tag{4.4}$$

$$\sum_j \|w_j\|^2 \geq |A_{t-1}|/4. \tag{4.5}$$

(2) Let $\Delta x_t(j) = \gamma \langle g, v_i \rangle$ where g is a random Gaussian vector. Set $x_t = x_{t-1} + \Delta x_t$, and update A_t accordingly.

The infinitely many constraints (4.3) can be written compactly as $X \preceq 2\,\mathrm{diag}(X)$, where X is the Gram matrix with $X_{ij} = \langle w_i, w_j \rangle$.

4.1.2. Analysis

The constraints (4.2) ensures that $\Delta x_t(S) = 0$ for large sets, which are at most $|A_{t-1}|/10$ in number. The constraints (4.3) imply the property (4.1). The feasibility of the SDP follows from the following geometric result.

Theorem 4.3 ([10]). *Let $G \subset \mathbb{R}^n$ be an arbitrary subspace with dimension $\dim(G) = \delta n$. For any $\zeta > 0$ and $\eta > 1$ with $1/\eta + \zeta \leq \delta$, there is a $n \times n$ PSD matrix X satisfying:*

(i) $\langle hh^T, X \rangle = 0$ *for all $h \in G^\perp$, where $G^\perp$ is the subspace orthogonal to G.*

(ii) $X_{ii} \leq 1$ *for all $i \in [n]$.*

(iii) *The trace* $\mathrm{tr}(X) \geq \zeta n$.

(iv) $X \preceq \eta\,\mathrm{diag}(X)$.

In particular, choosing G to be the subspace orthogonal to all large rows and setting $\delta = 0.9$, $\eta = 2$, and $\zeta = 0.1$, Theorem 4.3 implies that the SDP is always feasible.

This algorithm can be viewed as an interesting extension of iterated-rounding, where the update lies in a subspace, and yet has interesting random-like properties.

Let us see why this helps. At any time t, the discrepancy for set S has Gaussian increments with variance $\mathbb{E}[(\sum_{j \in S} \Delta x_t(j))^2]$, which by (4.1) is at most $2 \sum_{j \in S} \mathbb{E}[\Delta x_t(j)^2]$, i.e., twice the variance injected into the elements of S. We will show that

$$\sum_t \left(\sum_{j \in S} \Delta x_t(j)^2 \right) = O(d)$$

whp, and hence the discrepancy of S will be a Gaussian with standard deviation $O(d^{1/2})$. A union bound over the sets then gives the desired $O((d \log n)^{1/2})$ bound.

To this end, let us define $\sum_{j \in S} x_t(j)^2$ as the energy of S at time t. By (4.2), any S incurs discrepancy only after it becomes small, and so from that time onward its energy

can increase by at most $O(d)$. A priori there is no reason why the total increase in energy of S should be related to $\sum_t \sum_{j\in S} \mathbb{E}[\Delta x_t(j)^2]$ (the total variance injected into the elements of S). For example, even for a single variable j if $x_t(j)$ fluctuates a lot over time, $\sum_t \Delta x_t(j)^2$ could be arbitrarily large, while the final energy is ≤ 1. More precisely, the change in energy of S at time t is

$$\sum_{j\in S}\left(x_t(j)^2 - x_{t-1}(j)^2\right) = \underbrace{2\sum_{j\in S} x_{t-1}(j)\Delta x_t(j)}_{I} + \underbrace{\sum_{j\in S}\Delta x_t(j)^2}_{II}.$$

Summing up over t, the left-hand side telescopes and equals the total increase in energy of S. But $\sum_t \Delta x_t(j)^2$ can be much larger than this if the sum of term I over time is very negative. However, constraint (4.1) turns out to be very useful again. In particular, term I is a mean-zero update, and by (4.1) its variance can be bounded as

$$\mathbb{E}\left[\left(\sum_{j\in S} x_{t-1}(j)\Delta x_t(j)\right)^2\right] \leq 2\sum_{j\in S} x_{t-1}(j)^2\Delta x_t(j)^2 \leq 2\sum_{j\in S}\Delta x_t(j)^2.$$

This implies that the contribution of I is quite small compared to $\sum_t \sum_{j\in S} \mathbb{E}[\Delta x_t(j)^2]$. A clean exposition based on supermartingale concentration is in [7].

4.2. The general setting

We now describe the algorithmic version of Theorem 2.3. For simplicity, we will assume that K is symmetric. This is almost without loss of generality, because if K is asymmetric with $\gamma_m(K) \geq 3/4$, then $K \cap -K$ is symmetric and $\gamma_m(K \cap -K) \geq 1/2$.

An immediate issue with making Theorem 2.3 algorithmic is that any explicit description of K to a reasonable accuracy could already require exponential space. A crucial first step was by Dadush, Garg, Nikolov, and Lovett [24] who reformulated Theorem 2.3 without any reference to K. To state this result, recall that a random vector $Y \in \mathbb{R}^m$ is σ-sub-Gaussian if for all test directions $\theta \in \mathbb{R}^m$,

$$\mathbb{E}\left[e^{\langle\theta, Y\rangle}\right] \leq e^{\sigma^2\|\theta\|_2^2/2}.$$

Roughly, this means that $\langle Y, \theta\rangle$ looks like a Gaussian random variable with variance at most σ^2 for every unit vector θ. Simplifying slightly to symmetric K, [24] showed the following.

Theorem 4.4 ([24]). *For any symmetric convex body K, Theorem 2.3 (up to the exact value of c) is equivalent to the following: Let $v_1, \ldots, v_n \in \mathbb{R}^m$ be vectors with $|v_j\|_2 \leq 1$. Then there exists a distribution D on colorings $\{-1, 1\}^n$, such that for x sampled from D, the random vector $\sum_{j=1}^n x(j)v_j$ is σ-sub-Gaussian for some $\sigma = O(1)$.*

Moreover, to get a constructive version of Theorem 2.3 for any K, it suffices to give an algorithm that can efficiently sample a coloring from D.

The idea behind Theorem 4.4 is that as $\gamma_m(K) \geq 1/2$, a random Gaussian $g \in \mathbb{R}^m$ satisfies $\Pr[g \in K] \geq 1/2$, or equivalently, $\Pr[\|g\|_K \leq 1] \geq 1/2$ where $\|\cdot\|_K$ is the norm with K as its unit ball. By standard tail bounds, this gives $\mathbb{E}[\|g\|_K] = O(1)$. The following result of Talagrand [66], together with Markov's inequality, directly gives Theorem 4.4.

Theorem 4.5 ([66]). *Let $K \subset \mathbb{R}^m$ be a symmetric convex body and $Y \in \mathbb{R}^m$ be a σ-sub-Gaussian random vector. Then for the standard Gaussian $g \in \mathbb{R}^m$,*

$$\mathbb{E}\big[\|Y\|_K\big] \le O(\sigma) \cdot \mathbb{E}\big[\|g\|_K\big].$$

Bansal, Dadush, Garg, and Lovett [9] designed an algorithm called the Gram–Schmidt walk (GS-walk), with the following guarantee.

Theorem 4.6 ([9]). *Given vectors $v_1, \dots, v_n \in \mathbb{R}^m$ with $\|v_j\|_2 \le 1$, GS-walk outputs a coloring $x \in \{-1, 1\}^n$ such that $\sum_{j=1}^n x(j)v_j$ is sub-Gaussian with $\sigma \approx 6.32$.*

Harshaw, Sävje, Spielman, and Zhang [35] gave an improved analysis of the algorithm and showed that $\sigma = 1$, which is the best possible.

4.2.1. Gram–Schmidt walk algorithm

Before we describe the algorithm, we give some intuition. Suppose first that the vectors $v_1, \dots, v_n$ are orthogonal. Then, in fact a random coloring suffices. This follows as for any $\theta \in \mathbb{R}^m$, we have $\langle \theta, \sum_j x(j)v_j \rangle = \sum_j x(j)\langle \theta, v_j \rangle$, which for a random ± 1 coloring x is distributed as a sub-Gaussian with variance $\sum_j \langle \theta, v_j \rangle^2$, which is at most $\|\theta\|_2^2$ as the v_j are orthogonal and have at most unit length.

On the other extreme, suppose that $v_1, \dots, v_n$ are all identical and equal to some unit vector v. Then a random coloring is very bad and has variance n (instead of $O(1)$) in the direction $\theta = v$. The right thing here, of course, is to pair up the signs of $x(j)$. The general algorithm will handle these two extreme examples in a unified way, by trying to exploit the linear dependencies as much as possible while also using randomness.

We now describe the algorithm formally.

The Gram–Schmidt walk. Let $v_1, \dots, v_n$ be the input vectors. Let x_{t-1}, A_{t-1} denote the coloring and the set of alive elements at the beginning of time t.

Let $n(t) \in A_{t-1}$ be the largest indexed element alive at time t. This is called the *pivot* at time t and will play a special role. Let W_t be subspace spanned by the vectors in $A_{t-1} \setminus \{n(t)\}$ (i.e., all vectors alive at time t except $n(t)$). Let $v^\perp(t)$ be the orthogonal projection of the pivot $v_{n(t)}$ on $W_t^\perp$.

The algorithms works as follows. Initialize $x_0 = (0, \dots, 0)$ and $A_0 = [n]$.

At $t = 1, \dots, n$, do the following:

(1) Compute the update direction $u_t = (u_t(1), \dots, u_t(n)) \in \mathbb{R}^n$ as follows. Set $u_t(j) = 1$ for the pivot $j = n(t)$ and $u_t(j) = 0$ for $j \notin A_{t-1}$.
The $u_t(j)$ for the remaining $j \in A_{t-1} \setminus \{n(t)\}$ are defined by writing

$$v^\perp(t) = v_{n(t)} + \sum_{j \in A_{t-1} \setminus \{n(t)\}} u_t(j)v_j.$$

(2) Let $\delta_t^- < 0 < \delta_t^+$ be the unique negative and positive solutions for δ, respectively, to $\max_{j \in A_{t-1}} |x_{t-1}(j) + \delta u_t(j)| = 1$. Let

$$\delta_t = \begin{cases} \delta_t^- & \text{with probability } \delta_t^+/(\delta_t^+ - \delta_t^-), \\ \delta_t^+ & \text{with probability } -\delta_t^-/(\delta_t^+ - \delta_t^-). \end{cases}$$

(3) Update x_{t-1} randomly as $x_t = x_{t-1} + \delta_t u_t$. Update A_t accordingly.

Remark. Let us first see what the algorithm does for the two cases mentioned above. If the v_i are orthonormal, then $v^\perp(t) = v_{n(t)}$ as $v_{n(t}$ is orthogonal to W_t, and the algorithm only updates the color of the pivot. Moreover, at each time t $x(n(t))$ is set independently to ± 1, and so the algorithm eventually produces a completely random coloring. On the other hand, in the case where the v_i are identical, at each step t, as long as $n_t \geq 2$, the algorithm will exactly pair up the color of the pivot with the alive vector with the lowest index, resulting in overall discrepancy of at most 1.

Sketch of analysis. At each step, at least one element reaches -1 or 1, so the algorithm terminates in at most n steps.

Fix a vector $\theta \in \mathbb{R}^m$ with respect to which we want to show sub-Gaussianity of the discrepancy vector. Let $Y_t := \sum_{i=1}^n x_t(i) v_i$ and let $\text{disc}_t = \langle \theta, Y_t \rangle$. The goal is to show that

$$\mathbb{E}\big[e^{\text{disc}_n}\big] \leq e^{(\sigma^2/2)\|\theta\|_2^2}, \quad \text{for } \sigma = O(1).$$

Let us denote $\Delta x_t := x_t - x_{t-1} = \delta_t u_t$ and $\Delta \text{ disc}_t := \text{disc}_t - \text{disc}_{t-1}$. A key observation is that as u_t is chosen to satisfy $v^\perp(t) = \sum_{i=1}^n u_t(i) v_i$, we have

$$\Delta \text{ disc}_t = \sum_{i=1}^n \langle \theta, v_i \rangle \Delta x_t(i) = \delta_t \sum_{i=1}^n \langle \theta, v_i \rangle u_t(i) = \delta_t \big\langle \theta, v^\perp(t) \big\rangle \tag{4.6}$$

and hence depends only on the vector $v^\perp(t)$.

Proving sub-Gaussianity. We sketch the main idea. Let us first make a simplifying assumption that at each time t, the element to reach ± 1 is the pivot. So the elements get colored in the order $n, n-1, \ldots, 1$ and the pivot at time t is $n(t) = n - t + 1$. Let $w_1, \ldots, w_n$ be the orthonormal vectors obtained by applying the Gram–Schmidt orthonormalization procedure (GS) on the vectors $v_1, \ldots, v_n$ in that order. That is, $w_1 = v_1/\|v_1\|$ and for $i > 1$, w_i is the projection of v_i orthogonal to $v_1, \ldots, v_{i-1}$, normalized to have unit norm. Then $v^\perp(t) = \langle v_{n(t)}, w_{n(t)} \rangle w_{n(t)}$.

By (4.6), the overall discrepancy along θ is $\text{disc}_n(\theta) = \sum_{t=1}^n \delta_t \langle \theta, v^\perp(t) \rangle$. As δ_t is a mean-zero random variable chosen independently at time t, and $|\delta_t| \leq 2$, we have

$$\mathbb{E}\big[e^{\text{disc}_n(\theta)}\big] = \mathbb{E}\big[e^{\sum_{t=1}^n \delta_t \langle \theta, v^\perp(t) \rangle}\big] \leq e^{O(1) \cdot \sum_{t=1}^n \langle \theta, v^\perp(t) \rangle^2}.$$

But this is at most $e^{O(1) \cdot \|\theta\|_2^2}$, as desired, because

$$\sum_t \big\langle \theta, v^\perp(t) \big\rangle^2 = \sum_t \big\langle \theta, \langle v_{n(t)}, w_{n(t)} \rangle w_{n(t)} \big\rangle^2 \leq \sum_t \langle \theta, w_{n(t)} \rangle^2 \leq \|\theta\|_2^2,$$

as $|\langle v_{n(t)}, w_{n(t)}\rangle| \le \|w_{n(t)}\|_2 \|v_{n(t)}\|_2 \le 1$, and $\sum_i \langle \theta, w_i\rangle^2 \le \|\theta\|_2^2$ as the w_i are orthonormal.

In general the analysis needs some more care as non-pivot elements will also get colored during the process. But, roughly speaking, this only improves the bounds. If some non-pivot element x_k is colored at some time t, then the GS procedure (without v_k) will produce a different set of orthonormal vectors $\{w'_i\}$, but the increase in $\langle \theta, w'_{n(t)}\rangle^2 - \langle \theta, w_{n(t)}\rangle^2$, can be charged against the fact that k will never be a pivot anymore in the the future. We refer to [9] for the formal analysis.

5. APPROXIMATING HEREDITARY DISCREPANCY

In the previous sections we obtained bounds on the discrepancy of various classes of set systems and matrices. One can ask whether given a particular matrix A, can we efficiently determine $\mathrm{disc}(A)$. However, as described earlier in Theorem 3.2, discrepancy is hard to approximate in a very strong sense [21]. Intuitively, this is because discrepancy can be quite brittle, e.g., consider some matrix A with large discrepancy; however, if we duplicate each column of A, the resulting matrix has discrepancy 0.

Even though discrepancy is hard to approximate, in a surprising and remarkable result Matoušek, Nikolov, and Talwar [50] showed that $\mathrm{herdisc}(A)$ can be well approximated. Note that a priori it is not even clear how to certify (even approximately) that $\mathrm{herdisc}(A) \le k$, as it is the maximum over exponentially many quantities that themselves cannot be certified.

Theorem 5.1 ([50]). *There is an $O(\log m)^{3/2}$) approximation algorithm for computing the hereditary discrepancy of any $A \in \mathbb{R}^{m\times n}$.*

This is based on relating the hereditary discrepancy of a matrix to its γ_2-norm.

The γ_2-norm. For a matrix A, let $r(A) = \max_i (\sum_j A_{ij}^2)^{1/2}$ and $C(A) = \max_j (\sum_i A_{ij}^2)^{1/2}$ denote the largest ℓ_2-norm of rows and columns A. The $\gamma_2(A)$-norm of A is defined as

$$\gamma_2(A) = \min\{r(U)c(V) : UV = A\},$$

the smallest product $r(U)c(V)$ over all possible factorizations of A.

The quantity $\gamma_2(A)$ is efficiently computable using an SDP as follows. Consider vectors $w_1, \ldots, w_m$ corresponding to rows of U and $w_{m+1}, \ldots, w_{m+n}$ to columns of V. As $\alpha U, V/\alpha$ is also a valid factorization for any $\alpha > 0$, we can assume that $r(U) = c(V)$. Then, it is easily seen that $\gamma_2(A)$ is the smallest value t for which the following SDP is feasible.

$$\langle w_i, w_{j+m}\rangle = A_{ij} \quad \forall i \in [m], j \in [n] \text{ and } \langle w_i, w_i\rangle \le t \ \forall i \in [m+n]. \tag{5.1}$$

Theorem 5.1 follows from the following two facts.

Lemma 5.2. *For any $A \in \mathbb{R}^{m\times n}$ and factorization $A = UV$ with U, V arbitrary, we have that* $\mathrm{disc}(A) \le O(r(U)c(V)(\log 2m)^{1/2})$. *In particular,* $\mathrm{disc}(A) \le O(\gamma_2(A)(\log 2m)^{1/2})$.

This also implies that $\mathrm{herdisc}(A) \le O(\gamma_2(A)(\log 2m)^{1/2})$ as $\gamma_2(\cdot)$ itself is a hereditary function. Indeed, for any subset of columns S, we have $\gamma_2(A_{|S}) \le \gamma_2(A)$ as $A_{|S} = UV_{|S}$

and $C(V_{|S}) \leq C(V)$. The proof of Lemma 5.2 uses Banaszczyk's theorem in an interesting way.

Proof. Define the body $K = \{y : \|Uy\|_\infty \leq 2r(U)(\log 2m)^{1/2}\}$. Then $\gamma(K) \geq 1/2$ because for a random gaussian $g \sim N(0, I)$, $\Pr_g[\|Ug\|_\infty \geq 2r(U)(\log 2m)^{1/2}] \leq 1/2$.

As the columns of V have length at most $c(V)$ and $\gamma(K) \geq 1/2$, by Theorem 2.3 there exists $x \in \{-1, 1\}^n$ such that $y := Vx \in 5c(V)K$. By definition of K, this gives $\|Uy\|_\infty \leq 10r(U)c(V)(\log 2m)^{1/2}$, and as $Ax = Uy$, the result follows. ∎

Lemma 5.3. *For any $A \in \mathbb{R}^{m\times n}$, we have* $\text{herdisc}(A) \geq \Omega(\gamma_2(A)/\log m)$.

The proof of Lemma 5.3 establishes an interesting connection between the γ_2-norm and the determinant lower bound defined as follows.

$$\text{detlb}(A) = \max_k \max_{S\subset[m],T\subset[n],|S|=|T|=k} \left|\det(A_{S,T})\right|^{1/k},$$

where $A_{S,T}$ is the submatrix of A restricted to row and columns in S and T.

In a classical result, Lovász, Spencer, and Vesztergombi [44] showed that $\text{herdisc}(A) \geq \text{detlb}(A)/2$ for any matrix A. using a geometric view of hereditary discrepancy similar to that in Observation 2.1.1. In the other direction, Matoušek [48] showed that $\text{herdisc}(A) \leq O(\log(mn)(\log n)^{1/2}\,\text{detlb}(A))$. Interestingly, Matoušek's proof used Theorem 3.1 and duality for the SDP considered in Section 3.1. In particular, if the vector discrepancy is large for some subset of columns, there there must exist a sub-matrix with large detlb. This result was improved recently by Jiang and Reis [39] to $\text{herdisc}(A) \leq O((\log m \log n)^{1/2}\,\text{detlb}(A))$, and this bound is the best possible.

To prove Lemma 5.3, [50] show that $\text{detlb}(A) \geq \gamma_2(A)/\log m$ using the duality of the SDP (5.1) together with ideas of Matoušek [48].

The bounds in both Lemmas 5.2 and 5.3 are the best possible. However, the following conjecture seems quite plausible.

Conjecture 5.3.1. *There is an $O(\log mn)$ approximation algorithm for computing the hereditary discrepancy of any matrix A.*

As $\text{detlb}(A)$ and $\text{herdisc}(A)$ are within an $O(\log mn)$ factor, by the results of [44] and [39], one possible way to prove Conjecture 5.3.1 would be to give an $O(1)$ approximation for computing $\text{detlb}(A)$.

6. OTHER RECENT DIRECTIONS

We now discuss some other recent directions. First, we consider an interesting line of work on understanding the discrepancy of random instances. Next, we consider some results in the online setting where the vectors v_j are revealed over time and the sign $x(j)$ must be chosen immediately and irrevocably when v_j is revealed. Finally, we consider some matrix discrepancy problems, where one considers signed sums of matrices, instead of signed sums of vectors.

6.1. Random instances

In this survey, we restrict our attention to the work on the Beck–Fiala problem [2, 14, 29, 37, 57]. There are two natural probabilistic models here. Either each column has a 1 in exactly k positions chosen randomly out of the m choices, or the Bernoulli ensemble where each entry is 1 with probability $p = k/m$. The latter is slightly easier due to the lack of dependencies. For both these settings, an $O(k^{1/2})$ discrepancy can be achieved for the entire range of n and m, under fairly general conditions [14, 57]. These results are also algorithmic.

An interesting recent line of work shows that in fact much smaller discrepancy is possible if $n \gg m$. Franks and Saks [31] showed that $\operatorname{disc}(A) \leq 2$ with high probability for a fairly general class of random matrices A if $n = \Omega(m^3 \log^2 m)$. Independently, Hoberg and Rothvoss [37] showed that $\operatorname{disc}(A) \leq 1$ whp for the Bernoulli ensemble if $n = \Omega(m^2 \log m)$, provided that $mp = \Omega(\log n)$. Both these results use Fourier based techniques and are non-algorithmic.

Let us note that $n = \Omega(m \log m)$ is necessary to achieve $O(1)$ discrepancy, provided that p is not too small. Indeed, if we fix any coloring x, and consider a random instance, the probability that a fixed row has discrepancy $O(1)$ is $O((pn)^{-1/2})$, so the probability that each row has discrepancy $O(1)$ is at most $(pn)^{-\Omega(m)}$. As there are (only) 2^n possible colorings, a first moment argument already requires that $2^n (pn)^{-m} = \Omega(1)$.

So a natural question is whether the discrepancy is actually $O(1)$ for $n = \Omega(m \log m)$. Curiously, the Fourier-based methods seem to require $n = \Omega(m^2)$ even for $p = 1/2$. However, subsequent results show this optimal dependence using the second moment method. Potukuchi [56] showed that $\operatorname{disc}(A) \leq 1$ if $n = \Omega(m \log m)$ for the dense case of $p = 1/2$. The sparse setting with $p \ll 1$ turns out to be more subtle, and was only recently resolved by Altschuler and Weed [2] using a more sophisticated approach based on the conditional second moment method together with Stein's method of exchangeable pairs. They show the following result.

Theorem 6.1 ([2]). *Let $A \in \{0,1\}^{m \times n}$ be a random matrix with each entry independently chosen to be 1 with probability $p := p(n)$. Then there is a constant $c > 0$ such that if $n \geq cm \log m$, then $\operatorname{disc}(A) \leq 1$ whp.*

The results of [2, 56] are also non-algorithmic, and given the use of the probabilistic method it seems unlikely that they can be made algorithmic. However, one may wonder if this can be done under weaker assumptions such as when $n \gg m^{10}$.

Problem 6.2. Is there an efficient algorithm to find a coloring with expected discrepancy $O(1)$ for random instances of the Beck–Fiala problem when $n = m^{\Omega(1)}$.

Smoothed analysis. A substantial generalization of the random setting is the smoothed analysis setting, where the instance is obtained by taking underlying worst-case instance and perturbing it by a small random noise [65]. Recently, [12] studied the prefix-Komlós problem in this setting, where the vectors $v_1, \ldots, v_n$ are chosen adversarially and then v_j is perturbed by an independent random noise vector u_j.

Theorem 6.3 ([12]). *If the covariance* $\mathrm{Cov}(u_j) \succcurlyeq \varepsilon^2 I_m$ *for some* $\epsilon \geq 1/\mathrm{poly}(m, \log n)$. *Then, whp each prefix has discrepancy* $O((\log m + \log\log n)^{1/2})$.

This improves the dependence on n in Theorem 2.4 to doubly logarithmic, even if the noise is quite small, e.g., a vector is changed only with probability $1/\mathrm{poly}(m, \log n)$ in a single random coordinate. The techniques for random instances do not directly work here as these methods crucially use various special properties of random instances.

An interesting question is whether Theorem 6.1 can be extended to the smoothed setting.

Problem 6.4. Does the Beck–Fiala problem have $O(1)$ expected discrepancy in the smoothed setting, for a reasonably small noise rate, when $n = m^{\Omega(1)}$.

6.2. Online setting

In all the results considered thus far, we assumed that the vectors $v_1, \ldots, v_n \in \mathbb{R}^m$ are all given in advance. Another natural model is the online setting, first studied by Spencer [62], where the vector v_t is revealed at time t and a sign $x(t)$ must be chosen irrevocably without the knowledge of the vectors that will arrive in the future. The goal is to keep the discrepancy $\|d_t\|_\infty$ any time t as small as possible, where $d_t = x(1)v_1 + \cdots + x(t)v_t$ is the discrepancy at end of time t.

We restrict our focus here to the online Komlós setting. Notice that setting $x(t)$ randomly to ± 1 also works in the online setting, but this gives $\Omega(n^{1/2})$ dependence on n. Unfortunately, this dependence is unavoidable in general—at each step t an adversary can choose the vector v_t to be orthogonal to the current discrepancy vector d_{t-1} causing $\|d_t\|_2$ (and hence $\|d_t\|_\infty$) to grow as $\Omega(t^{1/2})$ with time. More refined lower bounds are also known [16, 64].

Interestingly, it turns out that the dependence on n can be substantially improved if the vectors v_t are chosen in a less adversarial manner.

Stochastic model. Here the vectors are chosen randomly and independently from some distribution D, that is known to the algorithm [11, 13, 15, 34]. For the Komlós setting, [11] showed the following.

Theorem 6.5 ([11]). *Let* D *be any distribution on unit vectors in* $\mathbb{R}^m$. *There is an online algorithm that given vectors sampled iid from* D, *achieves discrepancy* $O(\log^4 mn)$ *whp.*

These results are based on a greedy deterministic algorithms that choose the sign $x(t)$ based on a suitable potential function.

Let us consider the simpler setting of ℓ_2 discrepancy and where D is the uniform distribution over the unit sphere S^{m-1}. We sketch the proof of an $O(m^{1/2})$ bound (which is the best possible for ℓ_2-discrepancy even offline, e.g., for m orthonormal vectors).

Consider the potential $\Phi_t = \|d_t\|_2^2$. Upon given v_t, the algorithm chooses $x(t)$ to minimize the increase in $\Delta\Phi_t = \Phi_t - \Phi_{t-1}$. This evaluates to

$$\|d_{t-1} + x(t)v_t\|^2 - \|d_{t-1}\|_2^2 = 2x(t)\langle d_t, v_t\rangle + |v_t|_2^2 = 2x(t)\langle d_t, v_t\rangle + 1,$$

and hence setting $x(t) = -\text{sign}(\langle d_t, v_t\rangle)$ gives $\Delta\Phi_t = -2|\langle d_t, v_t\rangle| + 1$.

As v_t is uniform in S^{m-1}, we have that in expectation $\mathbb{E}_D|\langle d_{t-1}, v_t\rangle| \approx m^{-1/2}\|d_{t-1}\|_2$ for any d_{t-1}. This gives that $\mathbb{E}[\Delta\Phi_t] \ll 0$, and hence Φ_t has a strong negative drift whenever $|d_{t-1}| \gg m^{1/2}$. Using standard arguments, this implies that the discrepancy is $O(m^{1/2})$ whp at any given time.

The case of general distributions D is harder as $\mathbb{E}_D|\langle d_{t-1}, v_t\rangle|$ need not be large for every d_{t-1}. For example, if most of the probability mass of D lies in some subspace M, and d_{t-1} is orthogonal to M. However, one can still make this approach work by considering more complicated potential functions, that in addition to penalizing d_{t-1} with large norm, also penalize d_{t-1} if it gets close to certain undesirable regions.

Oblivious adversary model. Recently, these results were considered in the much more general *oblivious adversary* model. Here, the adversary knows the online algorithm and can pick the vectors accordingly, but it must choose them in advance before the online algorithm begins its execution. Equivalently, it cannot see the internal random choices made by the algorithm.

Notice that the oblivious setting generalizes both the stochastic setting and the worst case offline setting. Moreover, unlike for the stochastic model, here the $\Omega(n^{1/2})$ lower bound holds for any deterministic online algorithm, as d_{t-1} is completely determined by $v_1, \ldots, v_{t-1}$ and the adversary can always pick v_t orthogonal to d_{t-1}. So any nontrivial algorithm in this model must use its internal randomness cleverly.

In a recent breakthrough, Alweiss, Liu, and Sawhney [3] showed the following remarkable result.

Theorem 6.6 ([3]). *For any $\delta > 0$, vectors $v_1, v_2, \ldots, v_n \in \mathbb{R}^m$ with $\|v_t\|_2 \le 1$ for all $t \in [n]$, the algorithm maintains $\|d_t\|_\infty = O(\log(mn/\delta))$ for all $t \in [n]$ with probability $1 - \delta$.*

Choosing $\delta = 1/n^2$ gives that each prefix has discrepancy $O(\log mn)$ whp, almost matching the offline $O((\log mn)^{1/2})$ bound for prefix discrepancy given by Theorem 2.4. Moreover, the algorithm is extremely elegant and simple to describe.

Self-balancing walk algorithm. Let $c = 30 \log mn/\delta$. At each time t,

(1) If $|d_{t-1}|_\infty > c$ or if $|\langle d_{t-1}, v_t\rangle| > c$, declare failure.

(2) Set $x_t = 1$ with probability $1/2 - \langle d_{t-1}, v_t\rangle/2c$ and $x_t = -1$ otherwise.

The algorithm can be viewed as a randomized version of the greedy algorithm that picks the sign randomly if v_t and d_{t-1} are orthogonal, and otherwise uses the correlation between them to create a bias to move d_t closer to the origin.

The proof is a based on a clever stochastic domination argument and induction, and shows that as long as the algorithm does not declare failure, the distribution of d_t is less spread out than $N(0, 2\pi c I)$.

Theorem 6.6 is remarkable in many ways. First, it gives a simple linear time algorithm to obtain $O(\log n)$ discrepancy for the Komlós problem. Second, it also matches the

best algorithmic bound that we currently know for the prefix-Komlós problem in the offline setting. So any improvement of Theorem 6.6 would be extremely interesting.

Problem 6.7. Design an online algorithm for the Komlós problem, or for the prefix Komlós problem, that achieves an $O((\log mn)^{1/2})$ discrepancy.

For recent partial progress in this direction, see [43].

6.3. Matrix discrepancy

So far we only considered problems involving a signed sum of vectors. It is also very interesting to consider signed sums of more general objects such as matrices. An important problem of this type, with application to various fields, is the Kadison–Singer problem [40]. Below is an equivalent formulation in terms of discrepancy due to Weaver [68].

Kadison–Singer problem [68]. Let $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ be rank-1 Hermitian matrices satisfying $\sum_j A_j = I$ and $\|A_j\|_{\mathrm{op}} \le \delta$ for all $j \in [n]$, where $\delta \le 1/2$. Is there a ± 1 coloring x such that $\|\sum_j x(j)A_j\|_{\mathrm{op}} \le 1-\eta$, for some fixed constant $\eta > 0$ independent of n and d?

More generally, one can ask how small can the *discrepancy* $\|\sum_j x(j)A_j\|_{\mathrm{op}}$ be over all possible ± 1 colorings x. For a random coloring, standard matrix concentration results [55] give a bound of $O((\delta \log d)^{1/2})$, which does not give anything useful for the Kadison–Singer problem for large d. In a major breakthrough, Marcus, Spielman, and Srivastava [46] showed a bound of $O(\sqrt{\delta})$, without any dependence on d, using the method of interlacing polynomials. This bound is also the best possible [68]. These techniques are very different and we do not discus them here.

Their result however is non-constructive and obtaining an algorithmic version in an outstanding open question.

Problem 6.8. Is there an algorithmic version for the Kadison–Singer problem, even for the weaker bound of $1-\eta$ instead of $O(\sqrt{\delta})$.

Matrix Spencer problem. Another very interesting question, proposed originally by Raghu Meka, is the following matrix version of the Spencer's problem: given symmetric matrices $A_1, \ldots, A_n \in \mathbb{R}^{n\times n}$ with $\|A_j\|_{\mathrm{op}} \le 1$, find a ± 1 coloring x to minimize $\|\sum_j x(j)A_j\|_{\mathrm{op}}$.

Notice that if the A_j are diagonal, this is equivalent to Spencer's problem for $m = n$. Again, standard matrix concentration bounds imply a $O((n \log n)^{1/2})$ bound for random coloring, and the question is whether better bounds are possible.

Conjecture 6.8.1. *The matrix Spencer problem has discrepancy* $O(n^{1/2})$.

Very recently, Hopkins, Raghavendra, and Shetty [38] proved Conjecture 6.8.1 when the A_j have rank $n^{1/2}$, or, more generally, when $\|A_j\|_F \le n^{1/2}$. This result is based on an interesting new connection between discrepancy and communication complexity, and they also use this to give an alternate new proof of Spencer's result in classical setting. Another related result is due to Dadush, Jiang, and Reis [25].

FUNDING

This work was partially supported by the NWO VICI grant 639.023.812.

REFERENCES

[1] N. Alon and J. H. Spencer, *The probabilistic method*. John Wiley & Sons, 2016.

[2] D. J. Altschuler and J. Niles-Weed, The discrepancy of random rectangular matrices. 2021, arXiv:2101.04036.

[3] R. Alweiss, Y. P. Liu, and M. Sawhney, Discrepancy minimization via a self-balancing walk. In *Proceedings of STOC*, pp. 14–20, ACM, 2021.

[4] W. Banaszczyk, Balancing vectors and Gaussian measures of n-dimensional convex bodies. *Random Structures Algorithms* **12** (1998), no. 4, 351–360.

[5] W. Banaszczyk, On series of signed vectors and their rearrangements. *Random Structures Algorithms* **40** (2012), no. 3, 301–316.

[6] N. Bansal, Constructive Algorithms for Discrepancy Minimization. In *Proceedings of FOCS 2010*, pp. 3–10, IEEE, 2010.

[7] N. Bansal, On a generalization of iterated and randomized rounding. In *Symposium on theory of computing, STOC*, pp. 1125–1135, ACM, 2019.

[8] N. Bansal, D. Dadush, and S. Garg, An algorithm for Komlós conjecture matching Banaszczyk's bound. In *Proceedings of FOCS 2016*, pp. 788–799, IEEE, 2016.

[9] N. Bansal, D. Dadush, S. Garg, and S. Lovett, The Gram–Schmidt walk: a cure for the Banaszczyk blues. *Theory Comput.* **15** (2019), 1–27.

[10] N. Bansal and S. Garg, Algorithmic discrepancy beyond partial coloring. In *Proceedings of STOC 2017*, pp. 914–926, ACM, 2017.

[11] N. Bansal, H. Jiang, R. Meka, S. Singla, and M. Sinha, Online discrepancy minimization for stochastic arrivals. In *Proceedings of SODA*, pp. 2842–2861, ACM-SIAM, 2021.

[12] N. Bansal, H. Jiang, R. Meka, S. Singla, and M. Sinha, Prefix discrepancy, smoothed analysis, and combinatorial vector balancing. 2021, arXiv:2111.07049.

[13] N. Bansal, H. Jiang, S. Singla, and M. Sinha, Online vector balancing and geometric discrepancy. In *Proceedings of STOC*, pp. 1139–1152, ACM, 2020.

[14] N. Bansal and R. Meka, On the discrepancy of random low degree set systems. In *Proceedings of SODA 2019*, pp. 2557–2564, ACM-SIAM, 2019.

[15] N. Bansal and J. H. Spencer, On-line balancing of random inputs. *Random Structures Algorithms* **57** (2020), no. 4, 879–891.

[16] I. Bárány, On a Class of Balancing Games. *J. Combin. Theory Ser. A* **26** (1979), no. 2, 115–126.

[17] I. Bárány, On the power of linear dependencies. In *Building bridges*, pp. 31–45, Springer, 2008.

[18] J. Beck, Balanced two-colorings of finite sets in the square I. *Combinatorica* **1** (1981), no. 4, 327–335.

[19] J. Beck and T. Fiala, "Integer-making" theorems. *Discrete Appl. Math.* **3** (1981), no. 1, 1–8.

[20] J. Beck and V. Sós, Discrepancy theory. In *Handbook of combinatorics 2*, pp. 1405–1446, North-Holland 1995.

[21] M. Charikar, A. Newman, and A. Nikolov, Tight hardness results for minimizing discrepancy. In *Symposium on discrete algorithms, SODA*, pp. 1607–1614, ACM-SIAM, 2011.

[22] B. Chazelle, *The discrepancy method: randomness and complexity*. Cambridge University Press, 2001.

[23] W. Chen, A. Srivastav, G. Travaglini, et al., *A panorama of discrepancy theory*. Lecture Notes in Math. 2107, Springer, 2014.

[24] D. Dadush, S. Garg, S. Lovett, and A. Nikolov, Towards a constructive version of Banaszczyk's vector balancing theorem. In *Proceedings of APPROX/RANDOM*, pp. 28:1–28:12, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2016.

[25] D. Dadush, H. Jiang, and V. Reis, A new framework for matrix discrepancy: partial coloring bounds via mirror descent. 2021, arXiv:2111.03171.

[26] M. Drmota and R. F. Tichy, *Sequences, discrepancies and applications*. Springer, 1997.

[27] F. Eisenbrand, D. Pálvölgyi, and T. Rothvoß, Bin packing via discrepancy of permutations. *ACM Trans. Algorithms* **9** (2013), 24:1–24:15.

[28] R. Eldan and M. Singh, Efficient algorithms for discrepancy minimization in convex sets. *Random Structures Algorithms* **53** (2018), no. 2, 289–307.

[29] E. Ezra and S. Lovett, On the Beck–Fiala conjecture for random set systems. *Random Structures Algorithms* **54** (2019), no. 4, 665–675.

[30] C. Franks, A simplified disproof of Beck's three permutations conjecture and an application to root-mean-squared discrepancy. 2018, arXiv:1811.01102.

[31] C. Franks and M. Saks, On the discrepancy of random matrices with many columns. *Random Structures Algorithms* **57** (2020), no. 1, 64–96.

[32] A. A. Giannopoulos, On some vector balancing problems. *Studia Math.* **122** (1997), no. 3, 225–234.

[33] E. D. Gluskin, Extremal properties of orthogonal parallelepipeds and their applications to the geometry of Banach spaces. *Math. USSR, Sb.* **64** (1989), no. 1, 85.

[34] N. Haghtalab, T. Roughgarden, and A. Shetty, Smoothed analysis with adaptive adversaries. 2021, arXiv:2102.08446.

[35] C. Harshaw, F. Sävje, D. A. Spielman, and P. Zhang, Balancing covariates in randomized experiments using the Gram–Schmidt walk. 2019, arXiv:1911.03071.

[36] N. J. A. Harvey, R. Schwartz, and M. Singh, Discrepancy without partial colorings. In *Approx/random*, pp. 258–273, LIPIcs 28, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2014.

[37] R. Hoberg and T. Rothvoss, A Fourier-analytic approach for the discrepancy of random set systems. In *Proceedings of SODA*, pp. 2547–2556, ACM-SIAM, 2019.

[38] S. B. Hopkins, P. Raghavendra, and A. Shetty, Matrix discrepancy from quantum communication. 2021, arXiv:2110.10099v1.

[39] H. Jiang and V. Reis, A tighter relation between hereditary discrepancy and determinant lower bound. 2021, arXiv:2108.07945.

[40] R. V. Kadison and I. M. Singer, Extensions of pure states. *Amer. J. Math.* **81** (1959), 383–400.

[41] L.-C. Lau, R. Ravi, and M. Singh, *Iterative methods in combinatorial optimization*. Cambridge University Press, 2011.

[42] A. Levy, H. Ramadas, and T. Rothvoss, Deterministic discrepancy minimization via the multiplicative weight update method. In *Proceedings of IPCO 2017*, pp. 380–391, Springer, 2017.

[43] Y. P. Liu, A. Sah, and M. Sawhney, A gaussian fixed point random walk. 2021, arXiv:2104.07009.

[44] L. Lovász, J. Spencer, and K. Vesztergombi, Discrepancy of set-systems and matrices. *European J. Combin.* **7** (1986), no. 2, 151–160.

[45] S. Lovett and R. Meka, Constructive discrepancy minimization by walking on the edges. *SIAM J. Comput.* **44** (2015), no. 5, 1573–1582.

[46] A. W. Marcus, D. A. Spielman, and N. Srivastava, Interlacing families II: mixed characteristic polynomials and the Kadison–Singer problem. *Ann. of Math.* **182** (2015), 327–350.

[47] J. Matoušek, *Geometric discrepancy: an illustrated guide*. Algorithms Combin. 18, Springer, 2009.

[48] J. Matoušek, The determinant bound for discrepancy is almost tight. *Proc. Amer. Math. Soc.* **141** (2013), no. 2, 451–460.

[49] J. Matoušek and A. Nikolov, Combinatorial discrepancy for boxes via the γ_2 norm. In *Symposium on computational geometry, SoCG*, pp. 1–15, LIPIcs 34, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2015.

[50] J. Matoušek, A. Nikolov, and K. Talwar, Factorization norms and hereditary discrepancy. *Int. Math. Res. Not.* **2020** (2018), no. 3, 751–780.

[51] J. Matoušek and J. Spencer, Discrepancy in arithmetic progressions. *J. Amer. Math. Soc.* **9** (1996), no. 1, 195–204.

[52] A. Newman, O. Neiman, and A. Nikolov, Beck's three permutations conjecture: a counterexample and some consequences. In *Proceedings of FOCS 2012*, pp. 253–262, IEEE, 2012.

[53] A. Nikolov, *New computational aspects of discrepancy theory*. PhD thesis, Rutgers University, 2014.

[54] A. Nikolov, Tighter bounds for the discrepancy of boxes and polytopes. *Mathematika* **63** (2017), no. 3, 1091–1113.

[55] R. Oliveira, Sums of random hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.* **15** (2010).

[56] A. Potukuchi, Discrepancy in random hypergraph models. 2018, arXiv:1811.01491.

[57] A. Potukuchi, A spectral bound on hypergraph discrepancy. In *Proceedings of ICALP*, pp. 93:1–93:14, LIPIcs 168, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2020.

[58] K. F. Roth, Remark concerning integer sequences. *Acta Arith.* **9** (1964), 257–260.

[59] T. Rothvoß, Constructive discrepancy minimization for convex sets. In *Proceedings of FOCS 2014*, pp. 140–145, IEEE, 2014.

[60] T. Rothvoss, Better bin packing approximations via discrepancy theory. *SIAM J. Comput.* **45** (2016), no. 3, 930–946.

[61] W. M. Schmidt, Irregularities of distribution vii. *Acta Arith.* **21** (1972), 45–50.

[62] J. Spencer, Balancing games. *J. Combin. Theory Ser. B* **23** (1977), no. 1, 68–74.

[63] J. Spencer, Six standard deviations suffice. *Trans. Amer. Math. Soc.* **289** (1985), no. 2, 679–706.

[64] J. Spencer, Balancing vectors in the max norm. *Combinatorica* **6** (1986), no. 1, 55–65.

[65] D. A. Spielman and S. Teng, Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Commun. ACM* **52** (2009), no. 10, 76–84.

[66] M. Talagrand, *The generic chaining: upper and lower bounds of stochastic processes*. Springer, 2005.

[67] T. Tao, The Erdős discrepancy problem. *Discrete Anal.* (2016).

[68] N. Weaver, The Kadison–Singer problem in discrepancy theory. *Discrete Math.* **278** (2004), 227–239.

[69] H. Weyl, Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.* **77** (1916), 313–352.

NIKHIL BANSAL

Computer Science and Engineering (CSE), University of Michigan, Ann Arbor, MI 48109-2121, USA, bansaln@umich.edu

16. CONTROL THEORY AND OPTIMIZATION

ENLARGEMENTS: A BRIDGE BETWEEN MAXIMAL MONOTONICITY AND CONVEXITY

REGINA S. BURACHIK

ABSTRACT

Perhaps the most important connection between maximally monotone operators and convex functions is the fact that the subdifferential of a convex function is maximally monotone. This connects convex functions with a proper subset of maximally monotone operators (i.e., the cyclically monotone operators). Our focus is to explore maps going in the opposite direction, namely those connecting an arbitrary maximally monotone map with convex functions. In this survey, we present results showing how enlargements of a maximally monotone operator T provide this connection. Namely, we recall how the family of enlargements is in fact in a bijective correspondence with a whole family of convex functions. Moreover, each element in either of these families univocally defines T. We also show that enlargements are not merely theoretical artifacts, but have concrete advantages and applications, since they are, in some sense, better behaved than T itself. Enlargements provide insights into existing tools linked to convex functions. A recent example is the use of enlargements for defining a distance between two point-to-set maps, one of them being maximally monotone. We recall this new distance here, and briefly illustrate its applications in characterizing solutions of variational problems.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 47H04; Secondary 47H05, 49J53, 90C25, 58E35

KEYWORDS

Maximally monotone operators, set valued maps, enlargements, convex analysis, convex functions, inclusion problem, variational inequalities, Bregman distances

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5212–5233
DOI 10.4171/ICM2022/107

1. INTRODUCTION

In the framework of real Banach spaces, convex analysis and the theory of monotone operators have shown a fascinating interplay: the theory of one can be used for developing further the theory of the other, and the advances in one of them generates advances of the other. This cross-fertilization between both theories is a key topic in modern Functional Analysis, and has been captivating mathematicians since the middle of the 20th century. The natural way of going from convex analysis to maximally monotone maps is via the subdifferential of a convex function, which is maximally monotone [29, 35]. More precisely, subdifferentials are the (proper) subset of *cyclically monotone* maps [33].

Given a maximally monotone operator T, we approximate T by another set-valued map, called from now on an *enlargement* of T. Our goal is to show that enlargements provide a fruitful way of going from maximally monotone operators to convex functions. As an example, enlargements are used to prove a formula involving infimal convolution of convex functions [20], or to show new equivalences for the situation in which two convex functions differ by a constant [13], or for characterizing solutions of difference of convex (DC) problems [7]. More examples of this interplay can be found, for instance, in [9, CHAPTER 5], as well as in [31, 37]. We will see in what follows the crucial rôle of convex analysis in establishing outer-semicontinuity of the enlargements.

The concept of enlargement was first hinted in 1996 by Martínez-Legaz and Théra in [28]. Independently, the enlargement was formally defined and studied for the first time in the 1997 paper [10]. The results we quote in this survey span a time period from the late 1990s until today. We also quote crucial results obtained by Svaiter in [38], where the formal definition of the family of enlargements is introduced, and a fundamental link with convexity is established. Another key result mentioned here is the introduction of a family of convex functions associated with T, suggested by Fitzpatrick in 1988 in [23]. A beautiful fact is that these two families, seemingly independent from each other, are actually in a bijective correspondence, as we will see in Theorem 30.

A main motivation for studying maximally monotone maps and their approximations is the inclusion problem, stated as

$$\text{Find } x^* \in X \text{ such that } z \in T(x^*), \tag{1.1}$$

where $T : X \rightrightarrows X^*$ is a maximally monotone operator between a Banach space X and its dual X^*. Model (1.1) is used for solving fundamental problems, such as optimality conditions for (smooth and nonsmooth) optimization problems, fixed point problems, variational inequalities, and solutions of nonlinear equations. If T is point-to-point, the inclusion above becomes an equality.

This survey is organized as follows. In Section 2 we give the theoretical setting and the main definitions and basic results that we will need in later sections. In this section we recall the Fenchel–Young function, and also the Fitzpatrick function. In Section 3 we define the family of enlargements and give prototypical examples. In this section we describe the structure of the family (in terms of smaller and larger elements), and recall some of its continuity properties and the Brøndsted–Rockafellar property. We end this section recalling

a bijective correspondence between the family of enlargements of T and a family of convex functions. In Section 4 we define a family of convex functions associated with the family of enlargements, and illustrate this definition with examples. In this section we also describe the structure of this family, with smallest and largest elements, in analogy with the situation for the family of enlargements. In Section 5 we recall a new distance between point-to-set maps induced by the family of convex functions associated with T. We illustrate this new concept with applications to variational problems, and we include some open questions on these distances. Section 6 contains a few final words.

To facilitate reading, most technical proofs are kept to a minimum, and the focus is set on the main ideas and key points of the results. Interested readers can consult the references given regarding each result. Some results combine several existing facts, and their proofs are given to illustrate the type of analysis used in this topic.

2. PRELIMINARIES

Throughout this paper, X is a real *reflexive* Banach space with topological dual X^* and duality pairing between them denoted by $\langle \cdot, \cdot \rangle$. The norm in any space is denoted by $\| \cdot \|$. We use w to represent the weak topologies both on X and X^*. When using the weak topology, we will mention it explicitly, otherwise the strong topology is assumed. Let Z be a topological space and consider a subset $A \subset Z$, we denote by $\overline{A}$ its closure with respect to the strong topology, by $\text{int}(A)$ the *interior* of A and by $\text{co}(A)$ the *convex hull* of A.

2.1. Basic facts and tools

Recall the following definitions concerning extended real valued functions.

Definition 1. Let Z be a topological space and consider a function $f : Z \to \mathbb{R} \cup \{+\infty\}$.

(i) The *epigraph* of f is the set

$$\text{epi}(f) := \big\{(z,t) \in Z \times \mathbb{R} : f(z) \leq t\big\}.$$

(ii) The *domain* of f is the set

$$\text{dom } f := \big\{x \in Z : f(x) < +\infty\big\}.$$

(iii) The function f is said to be *proper* if $\text{dom } f \neq \emptyset$.

(iv) The function f is said to be *lower-semicontinuous (lsc)* if $\text{epi}(f)$ is closed.

The following definitions are relevant to convex functions.

Definition 2. Let $f : X \to \mathbb{R} \cup \{+\infty\}$ will be a convex function.

(i) The *subdifferential* of f is the point-to-set map $\partial f : X \rightrightarrows X^*$ defined by

$$\partial f(x) := \big\{x^* \in X^* : f(y) - f(x) \geq \big\langle y - x, x^* \big\rangle, \forall\, y \in X\big\}, \quad \text{if } x \in \text{dom } f, \tag{2.1}$$

and $\partial f(x) := \emptyset$, otherwise.

(ii) Given $\epsilon \geq 0$, the *ϵ-subdifferential of* f is the point-to-set map $\partial_\epsilon f : X \rightrightarrows X^*$ defined by

$$\partial_\epsilon f(x) := \big\{x^* \in X^* : f(y) - f(x) \geq \langle y - x, x^* \rangle - \epsilon, \forall\, y \in X\big\}, \quad \text{if } x \in \operatorname{dom} f, \tag{2.2}$$

and $\partial_\epsilon f(x) := \emptyset$, otherwise. Note that $\partial_0 f = \partial f$. To keep ϵ as a variable, we restate the ϵ-subdifferential of f in a way that does not involve ϵ. Namely, we consider the point-to-set map $\breve{\partial} f : \mathbb{R}_+ \times X \rightrightarrows X^*$ defined by

$$\breve{\partial} f(\varepsilon, x) := \partial_\epsilon f(x), \tag{2.3}$$

and call the ensuing enlargement the *Brøndsted–Rockafellar* enlargement.

(iii) The *Fenchel–Moreau conjugate* of f, denoted as $f^* : X^* \to \mathbb{R} \cup \{+\infty\}$, is defined by

$$f^*(x^*) := \sup\big\{\langle x, x^* \rangle - f(x) : x \in X\big\}. \tag{2.4}$$

(iv) The *Fenchel–Young function* associated to f is the function $f^{\mathrm{FY}} : X \times X^* \to \mathbb{R} \cup \{+\infty\}$ defined by

$$f^{\mathrm{FY}}(x, x^*) := f(x) + f^*(x^*) \quad \text{for all } (x, x^*) \in X \times X^*. \tag{2.5}$$

Remark 1. Note that f^* is always convex and (weakly) lsc. Hence, it is also strongly lsc by convexity. Therefore, f^{FY} is a convex, proper, and $(\|\cdot\| \times w)$-lsc function on $X \times X^*$. A remarkable and well-known fact is that f^{FY} completely characterizes the operator ∂f, in the following sense:

$$\partial f(x) = \big\{x^* \in X^* : f^{\mathrm{FY}}(x, x^*) = \langle x, x^* \rangle\big\}. \tag{2.6}$$

More precisely, the definitions yield

$$\begin{aligned} &f(x) + f^*(v) \geq \langle x, v \rangle, \quad \forall (x, v) \in X \times X^*, \\ &f(x) + f^*(v) = \langle x, v \rangle \iff x \in X,\ v \in \partial f(x). \end{aligned} \tag{2.7}$$

Moreover, f^{FY} completely characterizes the map $\breve{\partial} f$ (see (2.3)). Indeed, the definitions yield

$$\breve{\partial} f(\varepsilon, x) = \big\{v \in X^* : f^{\mathrm{FY}}(x, v) = f(x) + f^*(v) \leq \langle x, v \rangle + \varepsilon\big\}. \tag{2.8}$$

Since we always have $f^{\mathrm{FY}}(x, v) \geq \langle x, v \rangle$, when $\epsilon = 0$, (2.8) collapses into (2.6). System (2.7) constitutes the main inspiration for defining a family of convex functions induced by a maximally monotone operator T.

The map $\breve{\partial} f$ has a fundamental rôle in variational analysis. It is used for (i) developing algorithms for nonsmooth optimization (e.g., the ε-subgradient method, bundle methods, perturbed proximal methods), (ii) characterizing minimizing/stationary sequences, and (iii) characterizing approximate solutions of optimization problems. A crucial theoretical impact is the fact that it was used by Rockafellar to show maximality of ∂f in [35].

All information on a point-to-set map is encapsulated in its graph.

Definition 3. Let Z, Y be topological spaces and $F : Z \rightrightarrows Y$ a point-to-set map. The *graph* of F is the set

$$G(F) := \{(z, y) \in Z \times Y : y \in F(z)\}.$$

Given a point-to-set map F, we define $\overline{F}$ as the point-to-set map by the following equality:

$$G(\overline{F}) := \overline{G(F)},$$

where the closure in the right-hand side is taken with respect to the product topology in $Z \times Y$.

We next recall well known concepts related with point-to-set maps from X to X^*.

Definition 4. Let $T : X \rightrightarrows X^*$ be a point-to-set map.

(i) The *domain* of T is denoted by $D(T)$ and defined by

$$D(T) := \{x \in X : T(x) \neq \emptyset\},$$

and the *range* of T, denoted by $R(T)$, is defined by

$$R(T) := \{v \in X^* : \text{exists } x \in D(T) \text{ such that } v \in T(x)\},$$

(ii) T is said to be *monotone* when

$$\langle y - x, y^* - x^* \rangle \geq 0 \quad \forall (x, x^*), (y, y^*) \in G(T).$$

(iii) A monotone operator T is called *maximally monotone* provided

$$\langle y - x, y^* - x^* \rangle \geq 0 \quad \forall (y, y^*) \in G(T) \text{ implies } (x, x^*) \in G(T).$$

Equivalently, T is maximally monotone when $G(T)$ cannot be properly extended (in the sense of the inclusion), without violating the monotonicity condition given in (ii).

Continuity properties are associated with closedness of the graph. Hence, the topology we use determines the continuous maps. Besides from closedness w.r.t. the strong and weak topologies, we will consider sequential closedness with respect to the strong topology in X and the weak topology in X^*.

Recall the standard notation for strong and weak convergence in a reflexive Banach space: Given a sequence $(z_n) \subset X$, and an element $z \in X$, we denote by $z_n \to z$ the strong convergence of (z_n) to z. Given a sequence $(w_n) \subset X^*$ and an element $w \in X^*$,we denote by $w_n \rightharpoonup w$ the weak convergence of (w_n) to w. In this situation, we say that the sequence (z_n, w_n) *converges (sw)* (for strong-weak) to (z, w).

Definition 5. Let X be a reflexive Banach space and fix $S \subset X \times X^*$. We say that S is *sequentially strong-weak-closed*, denoted as $(sw)_s$-closed, if for every sequence $(x_n, v_n) \subset S$ the following condition holds:

$$\text{If } x_n \to x, \quad v_n \rightharpoonup v, \text{ then } (x, v) \in S.$$

We say that S is *sequentially weakly closed*, if S contains all weak limits of its weakly convergent sequences.

Remark 2. If $S \subset X \times X^*$ is weakly closed, then it is sequentially weakly closed (this is true for every topological space). The converse is in general not true (see, e.g, [2, EXAMPLE 3.33]), unless S is convex.

Fact 6. (i) *If S is weakly closed then S is $(sw)_s$-closed.*

(ii) *Assume that S is convex and S is strongly closed. Then, it is weakly closed (and hence $(sw)_s$-closed).*

Proof. (i) If S is weakly closed then, by the previous remark, S is sequentially weakly closed. Since every strongly convergent sequence is weakly convergent, this implies that S is $(sw)_s$-closed.

(ii) If S is strongly closed and convex, by [9, COROLLARY 3.4.16], S is weakly closed. By (i), it is $(sw)_s$-closed. ■

2.2. The Fitzpatrick function

From Remark 1, we see that the function f^{FY} completely characterizes the operator $T := \partial f$ and its enlargement $\partial^{\check{}} f$. A fundamental step in extending this type of link to an arbitrary maximally monotone operator T was performed by Fitzpatrick in 1988 in [23], who defined the following function, now called the *Fitzpatrick function associated with T*:

$$\begin{aligned} F_T(x,v) &:= \sup\big\{\langle y, v\rangle + \langle x - y, u\rangle \;:\; (y,u) \in G(T)\big\} \\ &= \sup\big\{\langle y - x, v - u\rangle + \langle x, v\rangle \;:\; (y,u) \in G(T)\big\}. \end{aligned}$$

By [23, THEOREMS 3.4 AND 3.8], we have that

$$\begin{aligned} F_T(x,v) &= \langle x, v\rangle \quad \text{if and only if } (x,v) \in G(T), \\ F_T(x,v) &\geq \langle x, v\rangle \quad \text{or every } x \in X, v \in X^*, \end{aligned} \tag{2.9}$$

Note the similarity with system (2.7). In other words, F_T characterizes $G(T)$, in a similar way as f^{FY} characterizes $G(\partial f)$. The Fitzpatrick function remained unnoticed for several years until it was rediscovered in [28]. In [24] Flåm gave an economic interpretation of the Fitzpatrick function, and also mentioned that this function was already used in 1982 by Krylov in [27]. The Fitzpatrick function allows for tractable reformulations of hard problems, including variational representation of (nonlinear) evolutionary PDEs, and the development of variational techniques for the analysis of their structural stability; see, e.g., [25, 32, 40, 41]. In [23, THEOREM 3.10], Fitzpatrick proved that F_T is the smallest function among all those that verify system (2.9). We will further explore this fact in later sections. Namely, we will revisit system (2.9) when defining the family of convex functions associated with the enlargements of T.

3. ENLARGEMENTS OF MAXIMALLY MONOTONE MAPS

We start this section by recalling the definition of enlargement, introduced by Svaiter in [38]. Then, we will focus on some well-known properties of set valued maps: local boundedness, outersemicontinuity, Lipschitz continuity, and the Brøndsted–Rockafellar property.

To explore continuity properties, we will often consider the closure of a point-to-set map $E : \mathbb{R} \times X \rightrightarrows X^*$. By Definition 3, for $F := E$, $\overline{E}$ is the point-to-set map such that $G(\overline{E}) = \overline{G(E)}$, where the closure is taken with respect to the strong topology in all spaces. In what follows, T is a fixed maximally monotone operator. From now on, most of the proofs will be omitted to alleviate the reading. All these proofs can be found in the references given for every result. The proofs which I do provide are meant to illustrate the type of mathematical tools used in the analysis, without making the text too technical.

3.1. The family of enlargements $\mathbb{E}(T)$

The theoretical framework that follows is based on the groundbreaking definition of a family of enlargements of T, introduced by Svaiter in [38].

Definition 7. Let $T : X \rightrightarrows X^*$ be a maximally monotone map. We say that a point-to-set mapping $E : \mathbb{R}_+ \times X \rightrightarrows X^*$ is an *enlargement* of T if the following hold:

(E_1) $T(x) \subset E(\epsilon, x)$ for all $\epsilon \geq 0$, $x \in X$;

(E_2) If $0 \leq \epsilon_1 \leq \epsilon_2$, then $E(\epsilon_1, x) \subset E(\epsilon_2, x)$ for all $x \in X$;

(E_3) The *transportation formula* holds for E. More precisely, let $x_1^* \in E(\epsilon_1, x_1)$, $x_2^* \in E(\epsilon_2, x_2)$, and $\alpha \in [0, 1]$. Define

$$\begin{aligned} \hat{x} &:= \alpha x_1 + (1-\alpha)x_2, \quad \tilde{x}^* := \alpha x_1^* + (1-\alpha)x_2^*, \\ \epsilon &:= \alpha\epsilon_1 + (1-\alpha)\epsilon_2 + \alpha\langle x_1 - \hat{x}, x_1^* - \tilde{x}^*\rangle + (1-\alpha)\langle x_2 - \hat{x}, x_2^* - \tilde{x}^*\rangle \\ &= \alpha\epsilon_1 + (1-\alpha)\epsilon_2 + \alpha(1-\alpha)\langle x_1 - x_2, x_1^* - x_2^*\rangle. \end{aligned}$$

Then $\epsilon \geq 0$ and $\tilde{x}^* \in E(\epsilon, \hat{x})$.

The set of all maps verifying (E_1)–(E_3) is denoted by $\mathbb{E}(T)$. We say that E is *closed* if $G(E)$ is (strongly) closed. The set of all closed enlargements is denoted by $\mathbb{E}_c(T)$.

Remark 3. Condition (E_1) ensures that E is an enlargement of T, while (E_2) indicates that the enlargement is increasing with respect to ϵ. Condition (E_3) allows constructing new elements in $G(E)$ by using convex combinations of known elements in $G(E)$. As we will see below, this condition is essential for establishing the link between enlargements of maximally monotone operators and convex functions. Note that if conditions (E_1)–(E_3) hold for E, then they also hold for $\overline{E}$, hence if $E \in \mathbb{E}(T)$, then $\overline{E} \in \mathbb{E}(T)$. By taking $\epsilon_1 = \epsilon_2 = 0$ and $\alpha_1, \alpha_2 \in (0, 1)$ in (E_3), we deduce that $E(0, \cdot)$ is a monotone map. By (E_1), we also have $E(0, \cdot) \supset T$. By maximality, we must have $E(0, \cdot) = T$.

Example 8. The set $\breve{\partial} f(\epsilon, x)$ is nonempty for every $\epsilon > 0$ if and only if f is lower semicontinuous at x. The map $\breve{\partial} f$ is an enlargement of $T = \partial f$. The fact that it verifies (E_1)–(E_2)

follows directly from the definitions. The proof of condition (E_3) for $\breve{\partial} f$ can be found, e.g., in [14, LEMMA 2.1]. It follows from the definitions that $\breve{\partial} f$ is a closed enlargement of ∂f, namely, $\breve{\partial} f \in \mathbb{E}_c(\partial f)$.

We recall next an example of an enlargement of an arbitrary maximally monotone operator T.

Example 9. Define the point-to-set map $T^e : \mathbb{R}_+ \times X \rightrightarrows X^*$ as follows:

$$T^e(\varepsilon, x) := \begin{cases} \{v \in X^* : \langle u - v, y - x\rangle \geq -\varepsilon, \forall (y,u) \in G(T)\}, & \forall x \in D(T), \\ \emptyset, & \text{if } x \notin D(T). \end{cases}$$

As mentioned in the Introduction, this enlargement of T was explicitly defined for the first time in [10]. The fact that it verifies conditions (E_1)–(E_2) follows directly from the definition. The transportation formula, i.e., condition (E_3) for T^e, is established in [17, 21]. It follows from the definitions that it is a closed enlargement of T, namely, $T^e \in \mathbb{E}_c(T)$. The enlargement T^e has been used for developing (i) inexact prox-like methods for variational inequalities [8, 10, 11, 15, 18], (ii) bundle-type methods for finding zeroes of maximally monotone operators [21, 30], and (iii) a unifying convergence analysis for algorithms for variational inequalities [16]. More recently, T^e has been used for developing inexact versions of the Douglas–Rachford algorithm for finding zeroes of sums of maximally monotone operators [1, 22, 39]. This list is by no means complete, but serves as evidence of the impact this concept has had on the development of inexact methods for variational inequalities and related problems. One of the reasons for this enlargement to have so many applications is the fact that, as we will see in Section 3.2, it has better continuity properties than the original T.

Remark 4. We mentioned above the fact that F_T characterizes T (see (2.9)). Moreover, F_T also characterizes T^e. Indeed, it follows directly from the definitions that

$$F_T(x,v) \leq \langle x, v\rangle + \epsilon \quad \text{if and only if } v \in T^e(\epsilon, x). \tag{3.1}$$

Remark 5. When $T = \partial f$, we always have from the definitions that $\breve{\partial} f(\epsilon, x) \subset (\partial f)^e(\epsilon, x)$. The opposite inclusion can be strict, as observed in [28], see also [9, EXAMPLE 5.2.5(IV)].

We mentioned above that $T^e \in \mathbb{E}_c(T)$, we can say more about its "location" within this family. The following result was established in [38].

Theorem 10. *The family $\mathbb{E}(T)$ has a largest and a smallest element (with respect to the inclusion of their graphs). The largest element is T^e, and the smallest element is*

$$T^s(\epsilon, x) = \bigcap_{E \in \mathbb{E}(T)} E(\epsilon, x).$$

Moreover, T^e is the largest element in $\mathbb{E}_c(T)$, and $\overline{T^s}$ is the smallest element in $\mathbb{E}_c(T)$. In other words, for every $E \in \mathbb{E}_c(T)$, we have

$$G(\overline{T^s}) \subset G(E) \subset G(T^e).$$

3.2. Local boundedness

Consider an iterative method for solving problem (1.1) that generates a sequence $(x^k, v^k) \subset G(T)$. Assume that the sequence $(x^k) \subset X$ is convergent. In this situation, can we say something about the behavior of the sequence (v^k)? The answer to this question requires more knowledge about $G(T)$. In fact, if T is maximally monotone, and (x^k) has its limit in the interior of $D(T)$, then $(v^k) \subset X^*$ is bounded and hence has a weakly convergent subsequence by Bourbaki–Alaoglu's theorem. The fundamental property needed here is the *local boundedness* of T in the interior of its domain.

Definition 11. Let X be a topological space and Y a metric space. A point-to-set map $F : X \rightrightarrows Y$ is said to be *locally bounded at* $x \in D(F)$ if there exists an open neighborhood U of x such that $F(U) := \bigcup_{z \in U} F(z)$ is bounded, and it is said to be *locally bounded* when it is locally bounded at every $x \in D(F)$.

Maps that are monotone are locally bounded at every point of the interior of their domains. When they are also maximal, they are not locally bounded at any point of the boundary of their domains. The latter fact means that we cannot expect enlargements to be locally bounded at any point of the boundary of their domains. Hence, we concentrate on points in the interior of their domains. Since $G(E) \supset \{0\} \times G(T)$ and we use enlargements to approximate T, we need to ensure that the local boundedness property is not lost when replacing T by E. In fact, we will see that $G(E)$ is not "too large," in the sense that the local boundedness property in the interior of the domains is still preserved.

Local boundedness of maximally monotone maps was established by Rockafellar in [34], and later extended to more general cases by Borwein and Fitzpatrick [3]. To make our study specific for enlargements, we will use a refined notion of local boundedness for point to set maps defined on $\mathbb{R}_+ \times X$.

Definition 12. Let $E : \mathbb{R}_+ \times X \rightrightarrows X^*$ be a point-to-set mapping. We say that E is *affine locally bounded* at $x \in X$ when there exists an open neighborhood V of x and positive constants L, M such that

$$\sup_{\substack{y \in V, \\ v \in E(\varepsilon, y)}} \|v\| \leq L\varepsilon + M.$$

Remark 6. By Theorem 10, $G(T^e) \supset G(E)$ for every $E \in \mathbb{E}(T)$. Therefore, all local boundedness properties enjoyed by T^e are inherited by $E \in \mathbb{E}(T)$. This means that it is enough to study the local boundedness property for T^e. This result is [17, COROLLARY 3.10], which we recall next.

Theorem 13 (Affine local boundedness). *If $T : X \rightrightarrows X^*$ is monotone, then T^e is affine locally bounded in* $\operatorname{int} D(T)$. *In other words, for all $x \in \operatorname{int} D(T)$ there exist a neighborhood V of x and positive constants L, M such that*

$$\sup\{\|v\| : v \in T^e(\varepsilon, y), y \in V\} \leq L\varepsilon + M$$

for all $\varepsilon \geq 0$.

Remark 7. The local boundedness property allows us to give an answer to the question posed at the start of Section 3.2. We mentioned there that local boundedness of T implies that, when the limit of (x^k) is in the interior of $D(T)$, the sequence (v^k) has a subsequence which is a weakly convergent. Theorem 13 and Remark 6 show that this fact is still true for any enlargement $E \in \mathbb{E}(T)$.

3.3. Lipschitz continuity

Consider again a method that generates a sequence $(x^k) \subset D(T)$ and assume that $X = X^* = \mathbb{R}^n$. Assume again that your sequence (x^k) converges to some $x \in D(T)$. Given any fixed $v \in Tx$, can you find a sequence $v^k \in Tx^k$ such that (v^k) converges to v? Enlargements of T do verify this property. They actually verify a much stronger property: Lipschitz continuity. On the other hand, the above mentioned property is not true for maximally monotone operators. Indeed, if $f(t) = |t|$ then $T = \partial f$ does not verify it at $t = 0$. Indeed, maximally monotone operators satisfy this property at a point x if and only if Tx is a single point. The latter fact is shown in **[34]** (see also **[9, THEOREM 4.6.3]**).

Definition 14. Let Z and Y be Banach spaces and $F : Z \rightrightarrows Y$ a point-to-set map. Let U be a subset of $D(F)$ such that F is closed-valued on U. The mapping F is said to be *Lipschitz continuous* on U if there exists a *Lipschitz constant* $\kappa > 0$ such that for all $x, x' \in U$ it holds that

$$F(x) \subset F(x') + \kappa \|x - x'\| B(0, 1),$$

where $B(0, 1) := \{y \in Y : \|y\| \leq 1\}$. In other words, for every $x \in U$, $v \in F(x)$, and $x' \in U$, there exists $v' \in F(x')$ such that

$$\|v - v'\| \leq \kappa \|x - x'\|.$$

The fact that enlargements of T are Lipschitz continuous at every point in the interior of their domains was proved in **[17, THEOREM 3.14]**. For more details on these properties, see **[9, CHAPTER 5]**.

3.4. On graphs, closedness, and convexity

Assume now that your method generates a sequence $(x^k, v^k) \subset G(T)$ which has a limit point $(x, v) \in X \times X^*$. When can you ensure that $(x, v) \in G(T)$? In other words, is $G(T)$ closed? Closedness of the graph of a point-to-set map is a type of continuity called *outer-semicontinuity* **[9, DEFINITION 2.5.1(A) AND THEOREM 2.5.4]**. When T is maximally monotone, $G(T)$ is strongly closed and also $(sw)_s$-closed, see **[9, PROPOSITION 4.2.1]**. We will show in this section that enlargements enjoy the same kind of outer-semicontinuity. We will see that convexity has a crucial rôle in establishing this. Let $E : \mathbb{R}_+ \times X \rightrightarrows X^*$ be any point-to-set map. Definition 3 gives

$$G(E) := \{(t, x, v) \in \mathbb{R}_+ \times X \times X^* : v \in E(t, x)\}.$$

To see $G(E)$ as the epigraph of a (possibly not convex) function, rearrange the order of its variables and consider the set:

$$\tilde{G}(E) := \big\{(x, v, t) \in X \times X^* \times \mathbb{R}_+ : v \in E(t, x)\big\}.$$

We see next how this set determines a convex function that characterizes $G(E)$. The next result from [38, LEMMA 3.2] is the key in linking enlargements with convexity.

Lemma 15. *Let $\Phi : X \times X^* \times \mathbb{R} \to X \times X^* \times \mathbb{R}$ be defined as*

$$\Phi(x, v, \epsilon) = \big(x, v, \epsilon + \langle v, x\rangle\big),$$

and let $E : \mathbb{R}_+ \times X \rightrightarrows X^*$ be any point-to-set map. The following statements are equivalent:

(i) *E verifies (E_3);*

(ii) *$\Phi(\tilde{G}(E)) \subset X \times X^* \times \mathbb{R}$ is a convex set.*

Fact 16. *Let $E \in \mathbb{E}(T)$.*

(i) *$\tilde{G}(E)$ is $(sw)_s$-closed if and only if $\Phi(\tilde{G}(E))$ is $(sw)_s$-closed.*

(ii) *$\tilde{G}(E)$ is (strongly) closed if and only if $\Phi(\tilde{G}(E))$ is (strongly) closed.*

Proof. (i) Assume that $\tilde{G}(E)$ is $(sw)_s$-closed and take a sequence $(x_n, v_n, s_n) \in \Phi(\tilde{G}(E))$ such that $x_n \to x$, $v_n \rightharpoonup v$, and $s_n \to s$. By definition of Φ, this means that $(x_n, v_n, s_n - \langle x_n, v_n\rangle) \in \tilde{G}(E)$ and

$$x_n \to x, \quad v_n \rightharpoonup v, \quad \text{and} \quad s_n - \langle x_n, v_n\rangle \to s - \langle x, v\rangle.$$

Since $\tilde{G}(E)$ is $(sw)_s$-closed, we deduce that $(x, v, s - \langle x, v\rangle) \in \tilde{G}(E)$. Equivalently, $(x, v, s) \in \Phi(\tilde{G}(E))$ and therefore $\Phi(\tilde{G}(E))$ is sequentially $(sw)_s$-closed. The converse implication follows identical steps, mutatis mutandis. The proof of (ii) follows from the fact that Φ and Φ^{-1} are continuous with respect to the strong topology. ∎

Convexity is crucial for ensuring outer-semicontinuity of the enlargements of T.

Corollary 17. *Let $E \in \mathbb{E}_c(T)$. Then E is $(sw)_s$-outer-semicontinuous.*

Proof. Since $E \in \mathbb{E}_c(T)$, we know that $G(E)$ (and equivalently, $\tilde{G}(E)$) is (strongly) closed. By Fact 16(ii) and Lemma 15, $\Phi(\tilde{G}(E))$ is (strongly) closed and convex. By Fact 6(ii), it is $(sw)_s$-closed. Finally, Fact 16(i) yields that $\tilde{G}(E)$ $(sw)_s$-closed, showing the outer-semicontinuity. ∎

3.5. Brøndsted and Rockafellar property

Since $\{0\} \times G(T) \subset G(E)$, the transportation formula allows producing elements in $G(E)$ by using elements $G(T)$. Can we use elements in $G(E)$ to approach those in $G(T)$?

For $T = \breve{\partial} f$, Brøndsted and Rockafellar [5] showed that any $(\varepsilon, x, v) \in G(\breve{\partial} f)$ can be approximated by an element $(x', v') \in G(\partial f)$ in the following way:

$$\text{For all } \eta > 0, \text{ there exists } v' \in \partial f(x') \text{ such that } \|x - x'\| \leq \frac{\varepsilon}{\eta} \text{ and } \|v - v'\| \leq \eta.$$

This result is known as *Brøndsted–Rockafellar's lemma*. The remarkable fact is that enlargements enjoy this property, too. While this property holds for $T = \breve{\partial} f$ in any Banach space, more general enlargements require reflexivity of the space. A consequence of the Brøndsted and Rockafellar property is the fact that the domain and range of an enlargement is dense in the domain and range of T, respectively. We sketch the proof of the Brøndsted and Rockafellar property here because it is beautiful and is based on crucial results on maximally monotone operators. For that, we recall that the duality mapping in a Banach space is defined as $J := \partial g$ for $g(x) := \frac{1}{2}\|x\|^2$. Using the definition of the subdifferential, it can be shown [9, PROPOSITION 4.4.4(I)] that the duality mapping verifies

$$J(x) := \{v \in X^* : \langle x, v\rangle = \|x\|^2, \|x\| = \|v\|\}. \tag{3.2}$$

The proof uses a key property of T in reflexive Banach spaces. Namely, the surjectivity of $T + \alpha J$ for $\alpha > 0$. This surjectivity property, which is, moreover, a characterization of maximally monotone operators in reflexive spaces, was established by Rockafellar in [36]. Since $G(T^e) \supset G(E)$ for every $E \in \mathbb{E}_c(T)$, it is enough to show that the Brøndsted–Rockafellar property holds for T^e.

Theorem 18. *Let $(\varepsilon, x_\varepsilon, v_\varepsilon) \in G(T^e)$ be given. For all $\eta > 0$, there exists $(x, v) \in G(T)$ such that*

$$\|v - v_\varepsilon\| \leq \frac{\varepsilon}{\eta} \quad \textit{and} \quad \|x_\varepsilon - x\| \leq \eta.$$

Proof. The claim trivially holds if $\varepsilon = 0$ because in this case by (E_1) we can take $(x, v) = (x_\varepsilon, v_\varepsilon) \in G(T)$. Assume that $\varepsilon > 0$. For any fixed $\beta > 0$, define $T_\beta(\cdot) := \beta T(\cdot) + J(\cdot - x_\varepsilon)$. Since T is maximally monotone, the surjectivity property mentioned above implies that there exist $x \in X$ and $v \in Tx$ such that $\beta v_\varepsilon \in \beta v + J(x - x_\varepsilon)$, which rearranges as $\beta(v_\varepsilon - v) \in J(x - x_\varepsilon)$. Using the fact that $(\varepsilon, x_\varepsilon, v_\varepsilon) \in G(T^e)$ and the definition of J in (3.2), we have

$$-\varepsilon \leq \langle v_\varepsilon - v, x_\varepsilon - x\rangle = -\frac{1}{\beta}\|x - x_\varepsilon\|^2 = -\beta\|v - v_\varepsilon\|^2,$$

and the result follows by taking $\beta := \eta^2/\varepsilon$ and rearranging the expression above. ∎

Since $G(E) \supset \{0\} \times G(T)$, we may wonder whether the range and domain of E might be much larger that those of T. The precise situation is a consequence of Theorem 18, and is stated next. Again, it is enough to establish this result for T^e.

Corollary 19. *The following hold:*

(i) $R(T) \subset R(T^e) \subset \overline{R(T)}$;

(ii) $D(T) \subset D(T^e) \subset \overline{D(T)}$.

Proof. The rightmost inclusions in (i) and (ii) follow from the previous theorem by taking $\eta \to +\infty$ for (i) and $\eta \to 0$ for (ii), respectively. The leftmost inclusions follow from (E_1). ∎

4. A FAMILY OF CONVEX FUNCTIONS ASSOCIATED WITH $\mathbb{E}(T)$

We have seen that the convexity emanating from condition (E_3) ensures that enlargements are outer-semicontinuous. How can this fact be used to associate enlargements with convex functions? All information on E is encapsulated in the set $\tilde{G}(E)$. Hence, we start by using this set to define the epigraph of a function defined in $X \times X^*$. The results in this section either directly use those in [19], or combine these for ease of presentation.

Definition 20. Let $S \subset X \times X^* \times \mathbb{R}$. The *lower envelope* of S is the function $\gamma : X \times X^* \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\gamma(x, v) := \inf\{t \in \mathbb{R} : (x, v, t) \in S\},$$

with the convention that $\inf \emptyset = +\infty$.

Fact 21. *Let $S \subset X \times X^* \times \mathbb{R}$ be a nonempty set and let γ be its lower envelope as in Definition 20. The following properties hold:*

(i) $S \subset \operatorname{epi}(\gamma)$.

(ii) *If S is closed and has epigraphical structure (i.e., if $(x, v, t) \in S$ then $(x, v, s) \in S$ for every $s > t$) then $S = \operatorname{epi}(\gamma)$.*

(iii) *S is closed if and only if γ is lower semicontinuous.*

(iv) *S is closed and convex if and only if γ is convex, proper, and lower semicontinuous.*

Proof. The proofs of (i), (iii), and (iv) follow directly from the definitions. For (ii), use the definition of infimum and the closedness of S to deduce that $(x, v, \gamma(x, v)) \in S$. Now the epigraphical structure yields $\operatorname{epi}(\gamma) \subset S$. ∎

The following simple lemma shows when $\tilde{G}(E) = \operatorname{epi}(\lambda)$ for some function λ.

Lemma 22. *Let $E : \mathbb{R}_+ \times X \rightrightarrows X^*$ and let $\lambda : X \times X^* \to \mathbb{R} \cup \{+\infty\}$. The following statements are equivalent:*

(i) $\tilde{G}(E) = \operatorname{epi}(\lambda)$.

(ii) $\lambda \geq 0$ *and* $E(t, x) = \{v \in X^* : \lambda(x, v) \leq t\}$ *for all* $t \geq 0$, $x \in X$.

Proof. [(i) → (ii)] By definition of E, $\tilde{G}(E) \subset X \times X^* \times \mathbb{R}_+$, so $D(E) \subset \mathbb{R}_+ \times X$. Hence, $\lambda \geq 0$. Now use (i) to write

$$v \in E(b, x) \leftrightarrow (x, v, b) \in \tilde{G}(E) \overset{\text{(i)}}{=} \operatorname{epi}(\lambda) \leftrightarrow \lambda(x, v) \leq b.$$

[(ii) → (i)] We have

$$(x,v,b)\in\tilde{G}(E)\leftrightarrow b\geq 0,\ v\in E(b,x)\overset{\text{(ii)}}{\leftrightarrow}\lambda(x,v)\leq b\leftrightarrow(x,v,b)\in\operatorname{epi}(\lambda).$$ ■

Imposing closedness assumptions, we obtain lsc of λ, as we see in the next result, which is **[19, PROPOSITION 3.1]**.

Theorem 23. *Let $E:\mathbb{R}_+\times X\rightrightarrows X^*$ be a point-to-set map such that $G(E)$ is closed and verifies (E_2). Define $\lambda_E:X\times X^*\to\mathbb{R}\cup\{+\infty\}$ as*

$$\lambda_E(x,v):=\inf\big\{t\geq 0:(x,v,t)\in\tilde{G}(E)\big\}=\inf\big\{t\geq 0\ :\ v\in E(t,x)\big\}.$$

In other words, λ_E is the lower envelope of $\tilde{G}(E)$. The following hold:

(i) $\tilde{G}(E)=\operatorname{epi}(\lambda_E)$.

(ii) *λ_E is strongly lsc.*

(iii) $\lambda_E\geq 0$.

(iv) $E(t,x)=\{v\in X^*:\lambda_E(x,v)\leq t\}$ *for all* $t\geq 0$, $x\in X$.

Furthermore, λ_E is the unique function that verifies (iii) and (iv). In particular, the mapping $E\mapsto\lambda_E$ is one-to-one in $\mathbb{E}_c(T)$.

Proof. Items (i) and (ii) follow from Fact 21(ii)–(iii). Item (iii) follows from the definition, and item (iv) follows from (iii) and (i). For the uniqueness of λ_E, use Lemma 22. The injectivity of the mapping $E\mapsto\lambda_E$ follows from (i). ■

The set $\tilde{G}(E)$ is (in general) not convex, and the same holds for λ_E. For generating a convex function, we use again the map Φ.

Definition 24. Let $E\in\mathbb{E}_c(T)$ and let λ_E be as in Theorem 23. Define $\Lambda_E:X\times X^*\to\mathbb{R}\cup\{+\infty\}$ by the equality

$$\operatorname{epi}(\Lambda_E):=\Phi\big(\operatorname{epi}(\lambda_E)\big).$$

Namely,

$$\Lambda_E(x,v)=\lambda_E(x,v)+\langle x,v\rangle,\tag{4.1}$$

for every $(x,v)\in X\times X^*$.

The following result constitutes the link between enlargements and convex functions.

Theorem 25. *Let $E\in\mathbb{E}_c(T)$ and let Λ_E as in Definition 24. The following hold:*

(i) *Λ_E is convex and (strongly) lsc.*

(ii) *Λ_E verifies*

$$\Lambda_E(x,v)\geq\langle x,v\rangle,\quad \textit{for every } (x,v)\in X\times X^*,$$
$$\Lambda_E(x,v)=\langle x,v\rangle,\quad \textit{if } v\in Tx.$$

Moreover, the mapping $E\mapsto\Lambda_E$ is one-to-one in $\mathbb{E}_c(T)$.

Proof. To prove (i), use Definition 24 to write

$$\operatorname{epi}(\Lambda_E) = \Phi\big(\operatorname{epi}(\lambda_E)\big) = \Phi\big(\tilde{G}(E)\big),$$

where we used Theorem 23(i) in the second equality. The convexity now follows from Lemma 15. To prove that the function Λ_E is lsc, use the fact that $E \in \mathbb{E}_c(T)$ and Fact 16(ii). The inequality in (ii) follows from (4.1) and Theorem 23(iii). As for the equality in (ii), assume that $v \in Tx$. By (E_1), this implies that $v \in E(0,x)$ and, by definition, $\lambda_E(x,v) \leq 0$. Since $\lambda_E \geq 0$, we must have $\lambda_E(x,v) = 0$. The latter, combined with (4.1), yields $\Lambda_E(x,v) = \langle x, v\rangle$. The last assertion holds by (4.1) and the fact that, given E, the function λ_E is uniquely defined. ∎

The following result is **[19, COROLLARY 3.2]** and characterizes E in terms of the convexity of Λ_E.

Corollary 26. *Let E be a point-to-set map with closed graph which verifies (E_1)–(E_2). Let Λ_E be as in Definition 24. The following statements are equivalent:*

(i) $E \in \mathbb{E}_c(T)$.

(ii) Λ_E *is convex.*

(iii) E *verifies* (E_3).

Proof. The equivalence between (i) and (iii) follows directly from the assumptions and the definitions. The equivalence between (ii) and (iii) follows from the definition of Λ_E and Lemma 15. Indeed, the definition of Λ_E and Theorem 23(i) gives

$$\operatorname{epi}(\Lambda_E) = \Phi\big(\operatorname{epi}(\lambda_E)\big) = \Phi\big(\tilde{G}(E)\big).$$

By Lemma 15, E verifies (E_3) if and only if $\Phi(\tilde{G}(E))$ is convex, and by the above expression the latter is equivalent to the convexity of Λ_E. ∎

Remark 8. Let $E = T^e \in \mathbb{E}_c(T)$. In this case, we have that $\Lambda_{T^e} = F_T$, the Fitzpatrick function associated with T. This is a consequence of (2.9), and Remark 4. Indeed, (2.9) and the latter remark imply that conditions (iii) and (iv) in Theorem 23 hold for $\lambda(x,v) := F_T(x,v) - \langle x, v\rangle$. By the uniqueness property stated in the same theorem, we must have $\lambda_{T^e}(x,v) = F_T(x,v) - \langle x, v\rangle$. Since we know that $T^e \in \mathbb{E}_c(T)$, (4.1) now yields $\Lambda_{T^e} = F_T$.

Theorem 25 helps us identify the relevant set of convex functions, which we define next.

Definition 27. Let $\mathcal{H}_0$ be the set of all convex and (strongly) lower semicontinuous functions defined on $X \times X^*$. The *Fitzpatrick family of* T is the set

$$\mathcal{H}(T) := \big\{h \in \mathcal{H}_0 : h(x,v) \geq \langle x, v\rangle \text{ for all } (x,v) \in X \times X^*, \\ \text{and } h(x,v) = \langle x, v\rangle \text{ whenever } v \in Tx\big\}.$$

Remark 9. By taking $S = \mathrm{epi}(h)$ in Fact 6, we see that h is (strongly) lower semicontinuous and convex if and only if it is weakly lower semicontinuous, and the latter is equivalent to h being $(sw)_s$-lower semicontinuous.

Example 28. Using Definition 27, Remark 1 states that $f^{\mathrm{FY}} \in \mathcal{H}(\partial f)$. Similarly, the system (2.9) implies that $F_T \in \mathcal{H}(T)$ for every T maximally monotone operator.

Theorem 25 provides a map from $\mathbb{E}(T)$ to $\mathcal{H}(T)$. Namely, the map $E \mapsto \Lambda_E$, with Λ_E fully characterizing $G(T)$ in the sense of condition (ii) in the statement of the theorem. We will define next the inverse of this map. Define $\pi : X \times X^* \to \mathbb{R}$ as $\pi(x, v) := \langle x, v\rangle$ for every $(x, v) \in X \times X^*$.

Theorem 29. *Let $h : X \times X^* \to \mathbb{R} \cup \{+\infty\}$ be a convex and lsc function. Consider the point-to-set map $L^h : \mathbb{R} \times X \rightrightarrows X^*$ defined by*

$$L^h(t, x) := \big\{v \in X^* : h(x, v) \le t + \langle x, v\rangle\big\},$$

for every $t \in \mathbb{R}$, $x \in X$. The following statements are equivalent:

(i) $h \in \mathcal{H}(T)$.

(ii) $h \ge \pi$, $D(L^h) \subset \mathbb{R}_+ \times X$ *and* $L^h \in \mathbb{E}_c(T)$.

(iii) $\Lambda_{L^h} = h$ *and* $L^h(0, x) \supset Tx$.

Proof. [(i) $\to$ (ii)] The definition on $\mathcal{H}(T)$ directly gives $h \ge \pi$. The latter inequality also yields $D(L^h) \subset \mathbb{R}_+ \times X$. Indeed, $v \in L^h(t, x)$ if and only if $h(x, v) \le t + \pi(x, v)$. Since $h \ge \pi$, this implies that $t \ge 0$, so $D(L^h) \subset \mathbb{R}_+ \times X$. We need to show that $E := L^h$ verifies (E_1)–(E_3). Condition (E_2) follows trivially from the definition of L^h. We next check (E_1). By (i), we know that $h(x, v) = \langle x, v\rangle$ for every $v \in Tx$. Therefore,

$$Tx \subset \big\{u \in X^* : h(x, u) \le \langle x, v\rangle\big\} = L^h(0, x),$$

so (E_1) holds. To verify (E_3) we will show that $\Phi(\tilde{G}(L^h))$ is convex and use Lemma 15. Indeed, by definition of Φ and L^h,

$$\begin{aligned}\Phi\big(\tilde{G}(L^h)\big) &= \big\{\big(x, v, t + \langle x, v\rangle\big) : v \in L^h(t, x)\big\}\\ &= \big\{\big(x, v, t + \langle x, v\rangle\big) : h(x, v) \le t + \langle x, v\rangle\big\} =: \Phi_h.\end{aligned}$$

We claim that $\Phi_h = \mathrm{epi}(h)$. Indeed, it is clear that $\Phi_h \subset \mathrm{epi}(h)$. To show that $\mathrm{epi}(h) \subset \Phi_h$, take any $(x, v, s) \in \mathrm{epi}(h)$. We need to show that we can write $s = t + \langle x, v\rangle$ for some $t \ge 0$. Indeed, by (i), we know that $h(x, v) \ge \langle x, v\rangle$. Hence, $\langle x, v\rangle \le h(x, v) \le s$ so $t = s - \langle x, v\rangle \ge 0$. This shows that the claim is true and $\Phi_h = \mathrm{epi}(h)$. Since h is convex, the above expression gives the convexity of $\Phi(\tilde{G}(L^h))$, and Lemma 15 furnishes (E_3).
[(ii) $\to$ (iii)] The inclusion $L^h(0, x) \supset Tx$ is (E_1), which holds because $L^h \in \mathbb{E}_c(T)$. Let us show that $\Lambda_{L^h} = h$. By (ii), we can apply Theorem 23(iv) to $E := L^h$ and obtain

$$\begin{aligned}&L^h(t, x) = \big\{v \in X^* : h(x, v) - \langle x, v\rangle \le t\big\} = \big\{v \in X^* : \lambda_{L^h}(x, v) \le t\big\},\\ &\quad \forall t \ge 0,\ \forall x \in X,\end{aligned}$$

where used the definition of L^h in the first equality. Use $t := \lambda_{L^h}(x,v)$ in the middle set to obtain $h(x,v) - \langle x,v\rangle \le \lambda_{L^h}(x,v)$. By (ii), we have that $h \ge \pi$ so that $t := h(x,v) - \pi(x,v) \ge 0$ can be used in the right-most set to derive $\lambda_{L^h}(x,v) \le h(x,v) - \langle x,v\rangle$. Therefore, $\lambda_{L^h}(x,v) + \langle x,v\rangle = h(x,v)$. Using this equality and the definition of Λ_{L^h}, we have that

$$\Lambda_{L^h}(x,v) = \lambda_{L^h}(x,v) + \langle x,v\rangle = h(x,v),$$

so (iii) is proved.

[(iii) $\to$ (i)] Since h is lsc, L^h is a point-to-set map with closed graph that verifies (E_2). By Theorem 23(iii), we have that $\lambda_{L^h} \ge 0$. The definitions and (iii) now yield

$$h = \Lambda_{L^h} = \lambda_{L^h} + \pi \ge \pi,$$

which gives the inequality in the definition of $\mathcal{H}(T)$. Now we need to prove that $h(x,v) = \pi(x,v)$ whenever $v \in Tx$. Since $Tx \subset L^h(0,x)$, we can use Theorem 23(iv) for $t = 0$ to deduce that $\lambda_{L^h}(x,v) \le 0$. By Theorem 23(iii) again, we have that $\lambda_{L^h} \ge 0$. Altogether, $\lambda_{L^h}(x,v) = 0$ if $v \in Tx$. Using (iii), we can write

$$h(x,v) = \Lambda_{L^h}(x,v) = \lambda_{L^h}(x,v) + \pi(x,v) = \pi(x,v),$$

when $v \in Tx$. Therefore, $h \in \mathcal{H}(T)$. ∎

Remark 10. For every $h \in \mathcal{H}(T)$, we have that $h = \pi$ if and only if $v \in Tx$. Indeed, the "if" part follows because $h \in \mathcal{H}(T)$. Conversely, let $h = \pi$ and take $E := L^h \in \mathbb{E}_c(T)$. By Theorem 23(iv), $v \in L^h(0,x)$ and by Theorem 29 and Remark 3 $T = L^h(0,\cdot)$ so $v \in Tx$.

To complete this section, we combine the results above to obtain a bijection between $\mathbb{E}(T)$ and $\mathcal{H}(T)$. In the following result, we consider the sets $\mathbb{E}(T)$ and $\mathcal{H}(T)$ as partially ordered. In $\mathbb{E}(T)$ we use the partial order of the inclusion of the graphs: $E_1 \preceq E_2$ if and only if $G(E_1) \subset G(E_2)$. In $\mathcal{H}(T)$ we use the natural partial order of pointwise comparison in $X \times X^*$: $h_1 \le h_2$ if and only if $h_1(x,v) \le h_2(x,v)$ for every $(x,v) \in X \times X^*$.

Theorem 30. *The map $\Theta : \mathbb{E}_c(T) \to \mathcal{H}(T)$ defined as $\Theta(E) := \Lambda_E$ is a bijection, with inverse $\Theta^{-1}(h) = L^h$. Considering the partially ordered spaces $(\mathbb{E}(T), \preceq)$ and $(\mathcal{H}(T), \le)$, the bijection Θ is "order reversing", i.e., if $E_1 \preceq E_2$ then $h_2 := \Theta(E_2) \le h_1 := \Theta(E_1)$. Therefore, $\Theta(T^e) = F_T$ and F_T is the smallest element of the family $\mathcal{H}(T)$.*

Proof. By Theorem 25, we know that $\Theta(E) = \Lambda_E \in \mathcal{H}(T)$ and the function Θ is injective. The function Θ is surjective because by Theorem 29, $L^h \in \mathbb{E}_c(T)$ and hence $\Theta(L^h) = \Lambda_{L^h} = h$ for every $h \in \mathcal{H}(T)$. The fact that Θ is order reversing is proved via the following chain of equivalent facts:

$$E_1 \preceq E_2 \overset{\text{definition}}{\longleftrightarrow} \tilde{G}(E_1) \subset \tilde{G}(E_2) \overset{\text{Theorem 23(i)}}{\longleftrightarrow} \operatorname{epi}(\lambda_{E_1}) \subset \operatorname{epi}(\lambda_{E_1})$$
$$\leftrightarrow \lambda_{E_2} \le \lambda_{E_1} \overset{(4.1)}{\leftrightarrow} \Theta(E_2) = \Lambda_{E_2} \le \Theta(E_1) = \Lambda_{E_1}.$$

Since $G(T^e) \supset G(E)$ for every $E \in \mathbb{E}_c(T)$, we can use the fact that Θ is order reversing and Remark 8 to obtain

$$\Theta(T^e) = \Lambda_{T^e} \overset{\text{Remark 8}}{=} F_T \leq \Theta(E),$$

for every $E \in \mathbb{E}_c(T)$. Since $\Theta(\mathbb{E}_c(T)) = \mathcal{H}(T)$, the claim is proved. ■

5. A DISTANCE INDUCED BY $\mathcal{H}(T)$

In this section we recall how the family $\mathcal{H}(T)$ is used for defining a new distance between set-valued maps. More results can be found in [6, 7, 12]. Let $h \in \mathcal{H}(T)$, and let $S : X \rightrightarrows X^*$ be any point-to-set operator. Following [12, DEFINITION 3.1], for each $(x, y) \in X \times X$, define

$$\mathcal{D}_S^{\flat,h}(x, y) := \begin{cases} \inf_{v \in Sy}(h(x, v) - \langle x, v\rangle) & \text{if } (x, y) \in \operatorname{dom} S \times \operatorname{dom} T, \\ +\infty & \text{otherwise} \end{cases} \tag{5.1a}$$

and

$$\mathcal{D}_S^{\sharp,h}(x, y) := \begin{cases} \sup_{v \in Sy}(h(x, v) - \langle x, v\rangle) & \text{if } (x, y) \in \operatorname{dom} S \times \operatorname{dom} T, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.1b}$$

We call these distances *Generalized Bregman distances* (GBDs). When S is point to point, all three collapse into one, $\mathcal{D}_S^h := \mathcal{D}_S^{\flat,h} = \mathcal{D}_S^{\sharp,h}$. The GBDs specialize to the Bregman distance. Given a proper, convex, and differentiable function $f : X \to \mathbb{R}$, the (classical) associated Bregman distance [4] is defined as

$$\mathcal{D}_f(x, y) := f(x) - f(y) + \big\langle y - x, \nabla f(y)\big\rangle. \tag{5.2}$$

It is easy to check that GBDs reduce to (5.2) when $T = S = \nabla f$ and $h := f^{\mathrm{FY}}$ [12, PROPOSITION 3.5]. We show next that GBDs can be used to characterize approximate solutions of the sum problem:

$$\text{find } x \in X \text{ such that } 0 \in Sx + Tx, \tag{5.3}$$

where $S, T : X \rightrightarrows X^*$ are point-to-set maps, with T being maximally monotone. The proof of the next result, inspired by [12, PROPOSITION 3.7], is [7, PROPOSITION 2.2].

Proposition 31. *Suppose that $T : X \rightrightarrows X^*$ is a maximally monotone operator and $S : X \rightrightarrows X^*$ is a point-to-set operator. Fix any $h \in \mathcal{H}(T)$, $\varepsilon \in \mathbb{R}_+$, and $x \in X$. Consider the following statements:*

(a) $0 \in L^h(\varepsilon, x) + Tx$.

(b) $\mathcal{D}_{-S}^{\flat,h}(x, x) \leq \varepsilon$.

Then (a) $\implies$ (b). Moreover, if dom S *is open and S is locally bounded with weakly closed images, then the two statements are equivalent.*

Next we illustrate how optimality conditions for minimizing a DC (difference of convex) function can be expressed by means of a particular type of GBD. Let $f : X \to$

$\mathbb{R} \cup \{+\infty\}$ and $g : X \to \mathbb{R}$ be proper and convex lsc functions. We consider the problem of finding $x \in X$ that globally minimizes $f - g$. In the proposition below, the equivalence between statements (a) and (b) are well known in finite-dimensional spaces (see, e.g., [26, THEOREM 3.1]). Its extension to reflexive Banach spaces can be found in [7, PROPOSITION 2.3].

Proposition 32. *The following statements are equivalent:*

(a) x *is a global minimum of* $f - g$ *on* X.

(b) *For all* $\varepsilon \geq 0$, $\check{\partial} g(\varepsilon, x) \subseteq \check{\partial} f(\varepsilon, x)$.

(c) *For all* $\varepsilon \geq 0$, $\mathcal{D}_{\check{\partial} g}^{\sharp, f^{\mathrm{FY}}}(x, x) \leq \varepsilon$.

5.1. Some open problems related to GBDs

(a) When $T = S = \partial f$, will the GBD induced by $h = F_T$ have some advantages when compared with the classical Bregman distance (induced by $h = f^{\mathrm{FY}}$)? What do the resulting *generalized projections* look like, when compared with the classical Bregman projections?

(b) Can these distances be used to regularize/penalize proximal-like iterations for variational inequalities?

(c) In view of Proposition 31, can we use the GBDs to develop new algorithms for solving problem (5.3)?

(d) In view of Proposition 32, can we obtain an optimality condition for the inclusion problem $0 \in Tx - Sx$ with T, S maximally monotone?

(e) Can a similar result to Proposition 32 be obtained for enlargements of $T = \partial f$ different from $\check{\partial} f$?

6. FINAL WORDS

Many crucial properties have been left uncovered, and it is my hope that those mentioned here will motivate researchers to explore the yet undiscovered paths that link maximally monotone operators with convex functions.

I conclude with a tribute to Asen Dontchev, who passed away on 16 September 2021. Asen was an outstanding mathematician with crucial contributions to set-valued analysis, and especially to the topic of inclusion problems and their approximations.

REFERENCES

[1] M. M. Alves and M. Geremia, Iteration complexity of an inexact Douglas–Rachford method and of a Douglas–Rachford–Tseng's F-B four-operator splitting method for solving monotone inclusions. *Numer. Algorithms* **82** (2019), 263–295.

[2] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

[3] J. M. Borwein and S. Fitzpatrick, Local boundedness of monotone operators under minimal hypotheses. *Bull. Aust. Math. Soc.* **39** (1989), 439–441.

[4] L. M. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **7** (1967), 200–217.

[5] A. Brøndsted and R. T. Rockafellar, On the subdifferentiability of convex functions. *Proc. Amer. Math. Soc.* **16** (1965), 605–611.

[6] R. S. Burachik, M. N. Dao, and S. B. Lindstrom, The generalized Bregman distance. *SIAM J. Optim.* **31** (2021), 404–424.

[7] R. S. Burachik, M. N. Dao, and S. B. Lindstrom, Generalized Bregman envelopes and proximity operators. *J. Optim. Theory Appl.* **190** (2021), 744–778.

[8] R. S. Burachik and J. Dutta, Inexact proximal point methods for variational inequality problems. *SIAM J. Optim.* **20** (2010), 2653–2678.

[9] R. S. Burachik and A. N. Iusem, *Set-valued mappings and enlargements of monotone operators*. Springer, 2008.

[10] R. S. Burachik, A. N. Iusem, and B. F. Svaiter, Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Var. Anal.* **5** (1997), 159–180.

[11] R. S. Burachik, J. Lopes, and G. Da Silva, An inexact interior point proximal method for the variational inequality problem. *Comput. Appl. Math.* **28** (2009), 15–36.

[12] R. S. Burachik and J. E. Martínez-Legaz, On Bregman-type distances for convex functions and maximally monotone operators. *Set-Valued Var. Anal.* **26** (2018), 369–384.

[13] R. S. Burachik and J. E. Martínez-Legaz, On a sufficient condition for equality of two maximal monotone operators. *Set-Valued Var. Anal.* **18** (2020), 327–335.

[14] R. S. Burachik, J. E. Martínez-Legaz, M. Rezaie, and M. Théra, An additive subfamily of enlargements of a maximally monotone operator. *Set-Valued Var. Anal.* **23** (2015), 643–665.

[15] R. S. Burachik, C. A. Sagastizábal, and S. Sheimberg, An inexact method of partial inverses and a parallel bundle method. *Optim. Methods Softw.* **21** (2006), 385–400.

[16] R. S. Burachik, S. Sheimberg, and B. F. Svaiter, Robustness of the hybrid extragradient proximal-point algorithm. *J. Optim. Theory Appl.* **111** (2001), 117–136.

[17] R. S. Burachik and B. F. Svaiter, ε-enlargements of maximal monotone operators in Banach spaces. *Set-Valued Var. Anal.* **7** (1999), 117–132.

[18] R. S. Burachik and B. F. Svaiter, A relative error tolerance for a family of generalized proximal point methods. *Math. Oper. Res.* **26** (2001), 816–831.

[19] R. S. Burachik and B. F. Svaiter, Maximal monotone operators, convex functions and a special family of enlargements. *Set-Valued Var. Anal.* **10** (2002), 297–316.

[20] R. S. Burachik and B. F. Svaiter, Operating enlargements of monotone operators: new connection with convex functions. *Pac. J. Optim.* **2** (2006), 425–445.

[21] R. S. Burachik, C. A. Sagastizábal, and B. F. Svaiter, Bundle methods for maximal monotone operators. In *Ill-posed variational problems and regularization techniques*, edited by M. Théra and R. Tichatschke, pp. 49–64, Springer, Berlin, Heidelberg, 1999.

[22] J. Eckstein and W. Yao, Relative-error approximate versions of Douglas–Rachford splitting and special cases of the ADMM. *Math. Program.* **170** (2017), 417–444.

[23] S. Fitzpatrick, Representing monotone operators by convex functions. In *Workshop/miniconference on functional analysis and optimization*, pp. 59–65, Proc. Centre Math. Anal. Austral. Nat. Univ. 20, Austral. Nat. Univ., Canberra, 1988.

[24] S. D. Flåm, Monotonicity and market equilibrium. *Set-Valued Var. Anal.* **24** (2016), 403–421.

[25] N. Ghoussoub, A variational theory for monotone vector fields. *J. Fixed Point Theory Appl.* **4** (2008), no. 1, 107–135.

[26] J.-B. Hiriart-Urruty, Conditions for global optimality. In *Handbook of global optimization*, edited by R. Horst and P. M. Pardalos, pp. 1–26, Springer, 1995.

[27] N. Krylov, Properties of monotone mappings. *Lith. Math. J.* **22** (1982), 140–145.

[28] J. E. Martínez-Legaz and M. Théra, ε-subdifferentials in terms of subdifferentials. *Set-Valued Var. Anal.* **4** (1996), 327–332.

[29] J. Moreau, Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France* **93** (1965), 273–299.

[30] L. Nagesseur, A bundle method using two polyhedral approximations of the ϵ-enlargement of a maximal monotone operator. *Comput. Optim. Appl.* **64** (2016), 75–100.

[31] J.-P. Penot, The relevance of convex analysis for the study of monotonicity. *Nonlinear Anal.* **58** (2004), 855–871.

[32] T. Roche, R. Rossi, and U. Stefanelli, Stability results for doubly nonlinear differential inclusions by variational convergence. *SIAM J. Control Optim.* **52** (2014), no. 2, 1071–1107.

[33] R. T. Rockafellar, Characterization of the subdifferentials of convex functions. *Pacific J. Math.* **17** (1966), 497–510.

[34] R. T. Rockafellar, Local boundedness of nonlinear monotone operators. *Michigan Math. J.* **6** (1969), 397–407.

[35] R. T. Rockafellar, On the maximal monotonicity of subdifferential mappings. *Pacific J. Math.* **33** (1970), 209–216.

[36] R. T. Rockafellar, On the maximality of sums of nonlinear monotone operators. *Trans. Amer. Math. Soc.* **149** (1970), 75–88.

[37] S. Simons, *From Hahn–Banach to monotonicity. 2nd edn.*, Lecture Notes in Math., Springer, NY, 2008.

[38] B. F. Svaiter, A Family of Enlargements of Maximal Monotone Operators. *Set-Valued Var. Anal.* **8** (2000), 311–328.

[39] B. F. Svaiter, A weakly convergent fully inexact Douglas–Rachford method with relative error tolerance. *ESAIM Control Optim. Calc. Var.* **25** (2019), no. 57, 311–328.

[40] A. Visintin, Variational formulation and structural stability of monotone equations. *Calc. Var. Partial Differential Equations* **47** (2013), no. 1–2, 273–317.

[41] A. Visintin, An extension of the Fitzpatrick theory. *Commun. Pure Appl. Anal.* **13** (2014), no. 5, 2039–2058.

REGINA S. BURACHIK

University of South Australia, Adelaide, Australia, regina.burachik@unisa.edu.au

NONLINEAR EIGENVALUE PROBLEMS FOR SEMINORMS AND APPLICATIONS

MARTIN BURGER

Dedicated to the Memory of Joyce McLaughlin and Victor Isakov.

ABSTRACT

The aim of this paper is to discuss recent progress in nonlinear eigenvalue problems for seminorms (absolutely one-homogeneous convex functionals), which find many applications in data science, inverse problems, and image processing. We provide a unified viewpoint on the notion of nonlinear singular vectors and eigenvectors for homogeneous nonlinear operators respectively functionals. We further discuss in particular ground states, i.e., the first eigenvector or eigenfunction. Moreover, we review a recent approach to the analysis of eigenvectors based on duality, which has implications to the possible computation of spectral decompositions, i.e., signal dependent linear expansions in a system of eigenvectors.

Moreover, we discuss some relevant implications such as the refined analysis of variational regularization methods and their bias, as well as the analysis of some iteration methods and time-continuous flows. Finally, we provide more direct applications of the nonlinear eigenvalue problems such as nonlinear spectral clustering on graphs.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 35P30; Secondary 47J10, 68T09, 94A08

KEYWORDS

Nonlinear eigenvalue problems, spectral decomposition, spectral clustering, gradient flows, variational methods

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5234–5255
DOI 10.4171/ICM2022/140

1. INTRODUCTION

Eigenvalue problems are not just a basic technique in linear algebra (cf. [39, 40]), they also find many applications in several branches of sciences, more recently also in data or image analysis. Prominent examples are the computation of states in quantum mechanics, Fourier decompositions—i.e., expansion in Laplacian eigenvalues—of audio signals or images (cf., e.g., [8]), or spectral clustering based on graph Laplacians (cf. [42]). In most of these applications the eigenvector or eigenfunction is of more importance than the exact eigenvalue, e.g., spectral clustering is based on dividing into the sets where the first nontrivial eigenfunction is negative or positive, respectively. Thus, particular focus is pot on the computation of eigenvectors respectively eigenfunctions.

While eigenvalue problems for linear operators are well understood, nonlinear eigenvalue problems, in particular those being nonlinear in the eigenvector or eigenfunction (cf. [1, 38]), are still a lively topic with many different directions of research. In physics, eigenvalue problems for nonlinear Schrödinger equations are a prominent example (cf. [24, 43]), while eigenvalue problems for p-Laplacian operators (and their graph equivalents) received strong recent attention in partial differential equations and data science (cf., e.g., [12, 21, 33–35]).

In this paper we want to focus on a special type of eigenvalue problems for (positively) zero-homogeneous operators related to the subdifferential of absolutely one-homogeneous functionals, more precisely we look for $\lambda > 0$ and $u \in H$, H a Hilbert space, such that

$$\lambda u \in \partial J(u). \tag{1.1}$$

Here $J : H \to \mathbb{R} \cup \{+\infty\}$ is assumed to be convex and absolutely one-homogeneous, thus it is effectively a seminorm on a subspace of H (cf. [15]). The assumption of one-homogeneity is less restrictive than it seems, since many other homogeneous eigenvalue problems can be reformulated equivalently as one-homogeneous problems, as we shall see in the p-Laplacian case below. Such eigenvalue problems can be rephrased in a variational setting, since we look for stationary points of the Rayleigh-quotient

$$R(u) = \frac{J(u)}{\|u\|}. \tag{1.2}$$

Indeed, (iterative) minimization of the Rayleigh quotient is a key technique for the computation of eigenvectors or eigenfunctions (cf. [13, 25, 29, 31, 32]).

Let us mention a related notion of nonlinear singular values (cf. [3]), given by

$$\lambda K^* K u \in \partial J_0(u), \tag{1.3}$$

where $J_0 : X \to \mathbb{R} \cup \{+\infty\}$ is a convex and absolutely one-homogeneous functional on a Banach space X, and $K : X \to Y$ is a bounded linear operator into the Hilbert space Y. This notion generalizes the linear singular value problem

$$K^* K u = \sigma^2 u, \tag{1.4}$$

with the obvious relation $\sigma = \frac{1}{\sqrt{\lambda}}$ to a nonlinear setting, and finds interesting applications in the regularization theory of inverse problems (cf. [3, 4]). We shall see below that indeed there

is a reformulation of the singular value problem (1.3) as a nonlinear eigenvalue problem of the form (1.1).

Besides some basic issues for eigenvalue problems and their direct applications, we will also discuss the issue of *spectral decompositions* (cf. [15, 18, 19, 27, 28, 30]), i.e., the possibility to develop signals in a systematic way into nonlinear eigenvectors, e.g., as a sum

$$f = \sum_{k=1}^{\infty} c_k u_k,$$

for $f \in H$ and u_k being the eigenvector with eigenvalue λ_k. In a general setting we rather look for a decomposition of the form

$$f = \int_0^{\infty} d\phi_\lambda, \tag{1.5}$$

with a measure ϕ on $\mathbb{R}_+$ valued in the Hilbert space H. Such a decomposition will be called spectral decomposition if the polar composition

$$\phi_\lambda = u_\lambda |\phi_\lambda| \tag{1.6}$$

is such that for each λ in the support of $|\phi_\lambda|$ the unit vector $u_\lambda \in H$ is an eigenvector for the eigenvalue λ. Fundamental questions, only partly answered so far, are the existence of nonlinear spectral decompositions as well as a systematic way to compute such decompositions from data. A particular advantage of a spectral decomposition is the possibility to define filtered versions of f,

$$f_\psi = \int_0^{\infty} \psi(\lambda)\, d\phi_\lambda, \tag{1.7}$$

e.g., with ψ being zero on a certain interval to suppress certain scales related to a range of eigenvalues. Such approaches find applications, e.g., in image or geometry processing (cf. [26, 30]). Moreover, the spectral decompositions of two different data f_1 and f_2 can be mixed, which finds interesting applications, e.g., in image fusion (cf. [5]).

The remainder of this paper is organized as follows: In Section 2 we provide some notations and fundamental properties of eigenvalue problems for seminorms, as well as first examples. We also discuss the motivation for a nonlinear spectral decomposition. Section 3 is devoted to the study of ground states, the eigenvectors for the first nontrivial eigenvalue, which are of particular relevance and also the easiest to compute numerically. Section 4 discusses the relation between eigenvalue problems, on the one hand, and variational methods, iterative schemes, and time-continuous flows, on the other. Here we see that eigenvectors and eigenfunctions yield structured examples of exact solutions for those methods. On the other hand, these methods, in particular gradient flows and time-continuous versions, can be used to compute eigenvectors and possibly even spectral decompositions.

2. BASIC PROPERTIES AND FORMULATIONS

In the following we fix some notation, discuss some basic properties of nonlinear eigenvalue problems such as (1.1), and unify the formulations of eigenvalues and singular values.

2.1. Seminorms, duality, and subdifferentials

Throughout the whole paper we assume that $J : H \to \mathbb{R} \cup \{+\infty\}$ is convex and absolutely one-homogeneous, i.e.,

$$J(tu) = |t|J(u), \quad \forall t \in \mathbb{R}. \tag{2.1}$$

This implies that J satisfies a triangle inequality, since

$$J(u_1 + u_2) = 2J\left(\frac{1}{2}u_1 + \frac{1}{2}u_2\right) \leq 2\left(\frac{1}{2}J(u_1) + \frac{1}{2}J(u_2)\right) = J(u_1) + J(u_2).$$

Moreover, the set

$$H_0 = \{u \in H \mid J(u) < \infty\}$$

is a subspace of H on which J is a seminorm, hence our nomenclature as eigenvalue problems for seminorms.

For completeness, let us recall the definition of the subdifferential of a convex functional J,

$$\partial J(u) = \{p \in H^* \mid \langle p, v - u\rangle \leq J(v) - J(u), \forall v \in H\}, \tag{2.2}$$

and the polar function (or convex dual),

$$J^*(p) = \sup_{u \in H} \langle p, u\rangle - J(u). \tag{2.3}$$

Note that for $p \in \partial J(u)$ we have

$$\langle p, u\rangle = J(u)$$

and

$$\langle p, v\rangle \leq J(v)$$

for each $v \in H$, and these properties are actually an equivalent characterization of subgradients under our assumptions (cf. [15]). Since J is a norm on a subspace, we can define a dual norm

$$\|p\|_* = \sup_{u \in H, J(u) \leq 1} \langle p, u\rangle, \tag{2.4}$$

which is interesting for the analysis of subgradients. Indeed, it can be shown that

$$\partial J(u) \subset \partial J(0) = \{p \in H^* \mid \|p\|_* \leq 1\},$$

for each $u \in H$, i.e., subdifferentials are contained in the dual unit ball.

The eigenvalue problem (1.1) can be interpreted in a dual way, by noticing that each eigenvector u is also a multiple of a subgradient p, respectively as $p \in \lambda \partial J^*(p)$. A key observation made in [15] is that these subgradients arising in the eigenvalue problems are of minimal norm.

Proposition 2.1. *Let u be an eigenvector of J satisfying $\lambda u = p \in \partial J(u)$. Then p is a subgradient of minimal norm, i.e.,*

$$\|p\| \leq \|q\|, \quad \forall q \in \partial J(u).$$

In [15] a further geometric characterization of eigenvectors has been derived, which relates to the minimal norm property.

Proposition 2.2. *An element $p \in \partial J(0)$ defines an eigenvector $u = \frac{1}{\lambda} p$ for some $\lambda > 0$ if and only if p satisfies the extremal property,*

$$\langle p, p - q \rangle \geq 0, \quad \forall q \in \partial J(0). \tag{2.5}$$

At least from a theoretical point of view, this yields an option to obtain all eigenvectors of functional as follows: first of all, compute for each $u \in H$ with $\|u\| = 1$ the subgradient of minimal norm, i.e.,

$$p = \arg\min\{\|q\| \mid q \in \partial J(u)\},$$

and subsequently check condition (2.5). In case of satisfaction, u is an eigenvector.

Example 2.3. Consider the simple example $J(u) = \sqrt{\langle u, Au \rangle}$ for a positive semidefinite operator A. In this case

$$\partial J(u) = \frac{1}{J(u)} Au$$

for $u \neq 0$, and it is easy to see that

$$\partial J(0) = \{p = Aw \mid w \in H \langle w, Aw \rangle \leq 1\}.$$

Let u be a linear eigenvector with eigenvalue $\lambda \neq 0$, i.e., $\lambda u = Au$, then (2.5) with $p = \frac{1}{J(u)} Au = \frac{\lambda}{J(u)} u$ becomes

$$\frac{\lambda}{J(u)} \left\langle u, \frac{1}{J(u)} Au - Aw \right\rangle \geq 0.$$

This is satisfied, since it is equivalent to

$$\sqrt{\langle u, Au \rangle} \geq \langle u, Aw \rangle,$$

and this inequality holds due to the Cauchy–Schwarz inequality in the scalar product induced by A.

Example 2.4. Consider a polyhedral functional, i.e.,

$$J(u) = \chi_C^* = \sup_{p \in C} \langle p, u \rangle,$$

with the symmetric polyhedral set

$$C = \operatorname{conv}(\{p_1, \ldots, p_m, -p_1, \ldots, -p_m\}).$$

Then p_j satisfies (2.5) if the plane orthogonal to p_j only intersects C in p_j.

Let us make this more concrete in $\mathbb{R}^2$ in polyhedra with $m = 2$. We start with the example $p_1 = (1, 1)$ and $p_2 = (-1, 1)$, i.e., C is the unit ball in ℓ^∞. The lines orthogonal to $\pm p_j$ only intersect C in p_j, thus all p_j are eigenvectors. As a specific case, we explicitely compute (2.5) for p_1 and $q = (r, s) \in C$,

$$\langle p_1, p_1 - q \rangle = 2 - r - s \geq 0,$$

since $r, s \in [-1, 1]$.

As a second example consider $p_1 = (1, 1)$ and $p_2 = (\varepsilon, 1)$ with $\varepsilon \in (0, 1)$. With the reasoning as above, we see that p_1 is an eigenvector. However, p_2 is not as we see with $q = (1, 1) \in C$,

$$\langle p_2, p_2 - q\rangle = \varepsilon^2 + 1 - \varepsilon - 1 = \varepsilon(\varepsilon - 1) < 0.$$

2.2. Singular values and eigenvalues

In the following we discuss the reformulation of the nonlinear singular value problem (1.3) as a nonlinear eigenvalue problem. Throughout this section we assume that J_0 is a seminorm on a subspace of X extended by $+\infty$, and $K : X \to Y$ is a bounded linear operator. In order to unify the formulation with (1.1), we want to define a functional J on a subspace of Y, respectively on values $v = Ku$. Hence, it is natural to define the space H as the closure of the range of K in Y. If K has a nontrivial nullspace, the definition of $J_0(v)$ as $J(u)$ with $Ku = v$ is not unique, however. We thus first provide a property of eigenvectors that will enable a unique definition.

Lemma 2.5. *Let u be a nonlinear singular vector according to (1.3). Then $J_0(u) \le J_0(w)$ for all w such that $Kw = Ku$.*

Proof. We take a duality product of λK^*Ku with $u - w$ to obtain

$$\lambda\big\langle K^*Ku, w - u\big\rangle = \lambda\langle Ku, Ku - Kw\rangle = 0.$$

On the other hand, from the singular value equation (1.3) we find

$$\lambda\big\langle K^*Ku, w - u\big\rangle = \langle p, w - u\rangle = \langle p, w\rangle - J_0(u) \le J_0(w) - J_0(u).$$

Hence, $J_0(u) \le J_0(w)$. ∎

From this result we see that we need to define J via the minimal value of J_0, more precisely

$$J(v) := \inf_{u, Ku = v} J_0(u). \tag{2.6}$$

It is straightforward to check that $J : H \to \mathbb{R} \cup \{+\infty\}$ is an absolutely one-homogeneous convex functional. Moreover, there is a direct relation between subgradients: we find $p \in \partial J(Ku)$ if and only if $K^*p \in \partial J_0(u)$. Thus, we find the equivalence between

$$\lambda v = \lambda Ku \in \partial J(v) \quad \text{and} \quad \lambda K^*Ku = \lambda K^*v \in \partial J_0(u).$$

2.3. Spectral decomposition

An interesting question is the possible existence of a spectral decomposition in the nonlinear case. Let us recall the well-known spectral decomposition of a positive semi-definite linear operator A on a Hilbert space H: there exists an operator-valued spectral measure E supported on the spectrum of A such that

$$A = \int_0^\infty \lambda \, dE_\lambda.$$

This allows extending functions $f : \mathbb{R}^+ \to \mathbb{R}$ to the operator A as

$$f(A) = \int_0^\infty f(\lambda)\, dE_\lambda.$$

In the case of a compact operator, the spectral measure is concentrated on a countable set and takes the form

$$E = \sum_{k=1}^{\infty} u_k \otimes u_k \delta_{\lambda_k},$$

where u_k is an eigenvector for the eigenvalue λ_k and $\otimes$ denotes the outer product. A positive semidefinite linear operator is the canonical choice in our setting, since it defines a convex absolutely one-homogeneous functional

$$J(u) = \sqrt{\langle u, Au \rangle}.$$

In general, we cannot expect to obtain some kind of spectral decomposition from a convex functional J, respectively its subdifferential ∂J, but we can hope to have a pointwise decomposition, corresponding in the linear case to

$$Au = \int_0^\infty \lambda\, d(E_\lambda u) = \int_0^\infty \lambda\, d\phi_\lambda,$$

with a spectral measure ϕ valued in the Hilbert space H. In particular, this allows for the reconstruction of u from the spectral measure via

$$u = \int_0^\infty d\phi_\lambda,$$

as well as some spectral filtering by integrating some function of λ, e.g., a characteristic function in some region.

In general, there is no unique way to construct a unique spectral decomposition of this kind. For example, for total variation regularization in one dimension (with appropriate definition of the variation on the boundary), it was shown in [3] that the Haar wavelet basis is an orthogonal basis of nonlinear eigenfunctions, hence there exists an atomic spectral decomposition in this basis. However, it also has been shown that there is a continuum of further eigenfunctions, necessarily linearly dependent, hence further spectral decompositions can be obtained by exchanging parts of the Haar wavelet basis. An interesting question is to define a generic spectral decomposition by a natural technique.

3. GROUND STATES

In the following we investigate the first nontrivial eigenvalue and its corresponding eigenvector or eigenfunction, which we call ground state. More precisely, let

$$\mathcal{N}(J) = \{u \in H \mid J(u) = 0\}$$

be the nullspace of J. Due to the properties of J, the nullspace is indeed a linear subspace (cf. [3,15]), and we can define its orthogonal complement $H_0 := \mathcal{N}(J)^\perp$ in H. It can further be shown that for each $u \in H$ and $u_0 \in \mathcal{N}(J)$ the identity

$$J(u + u_0) = J(u) + J(u_0)$$

holds, i.e., there is an analogue of the orthogonal decomposition at the level of J. Finally, let $u \in H$ be an eigenvector for the eigenvalue $\lambda \neq 0$. Then we find for $u_0 \in \mathcal{N}(J)$,

$$\lambda \langle u, u_0 \rangle = \langle p, u_0 \rangle \leq J(u + u_0) - J(u) = 0.$$

Since $-u$ is also an eigenvector, we obtain the opposite inequality and hence the orthogonality of u and u_0. Thus, we get rid of the trivial eigenvalues and the corresponding eigenvectors by restriction to H_0, which leads in particular to the definition of a ground state according to [3].

Definition 3.1. Let $J : H \to \mathbb{R} \cup \{+\infty\}$ be an absolutely one-homogeneous convex functional and H_0 be the orthogonal of its nullspace as above. Then we call $u \in H$ a *ground state* of J if

$$u \in \arg\min_{u \in H_0} \frac{J(u)}{\|u\|}. \tag{3.1}$$

Let us mention that we can rescale u in the above definition and consider equivalently a ground state as a minimizer of J on the unit sphere $\{u \in H_0 \mid \|u\| = 1\}$. The latter is useful for proving the existence of ground states. If J is lower semicontinuous and the sublevel sets of J are precompact, existence follows from a standard argument (cf. [3]). It is apparent for normalized eigenvectors u that $\lambda = J(u)$, thus

$$\lambda_0 := \min_{u \in H_0} \frac{J(u)}{\|u\|} \leq \lambda$$

for each nontrivial eigenvalue. On the other hand, λ_0 is indeed an eigenvalue for each eigenvector u_0 minimizing the Rayleigh-quotient. For this, define $p_0 = \lambda_0 u_0$. Then we have

$$\langle p_0, u_0 \rangle = \lambda_0 \langle u_0, u_0 \rangle = \lambda_0 = J(u_0),$$

and for arbitrary $u \in H \setminus \{0\}$,

$$\langle p_0, u \rangle = \lambda_0 \langle u_0, u \rangle \leq \lambda_0 \|u\| \leq \frac{J(u)}{\|u\|} \|u\| = J(u).$$

Thus $p_0 = \lambda_0 u_0 \in \partial J(u_0)$.

We finally recall the relation to the case of nonlinear singular values. The ground state in this case can be equivalently computed from minimizing

$$u \in \arg\min_{u \in X} \frac{J_0(u)}{\|Ku\|}, \tag{3.2}$$

which is often more accessible.

3.1. p-Laplacian eigenvalues

Ground states of the p-Laplacian are a well-studied problem in partial differential equations, as well as on graphs. In the standard setting, one would look for the first eigenvalue in the problem

$$-\nabla \cdot \left(|\nabla u|^{p-2} \nabla u \right) = \lambda_1 u |u|^{p-2},$$

in a domain Ω with homogeneous Neumann or Dirichlet boundary values. This is, however, related to the eigenvalue of the p-Laplacian energy

$$E_p(u) = \int_\Omega \left|\nabla u(x)\right|^p dx$$

in $L^p(\Omega)$, while our Hilbert space setting corresponds to solving

$$-\nabla \cdot \left(|\nabla u|^{p-2}\nabla u\right) = \lambda_1 u \|u\|_{L^2}^{p-2}.$$

Since $u \mapsto E_p(u)^{1/p}$ is an absolutely one-homogeneous convex functional, the ground state can be computed by minimizing the corresponding Rayleigh quotient

$$R(u) = \frac{E_p(u)^{1/p}}{\|u\|_{L^2}},$$

which corresponds to our setup in this paper. For $p = 2$, all formulations simply yield the standard linear eigenvalue problem for the Laplacian and, indeed, the formulation with the Rayleigh quotient is related to the fact that the first eigenvalue is the best constant in the Poincaré-inequality. On graphs, the corresponding problem for the graph Laplacian is fundamental for spectral clustering techniques (cf. [42]).

Particularly interesting cases are, of course, the limiting ones $p = 1$ and $p = \infty$. For $p = 1$, the ground state is the first eigenfunction of total variation, and, due to area and coarea formula, the L^2-norm and total variation can be related to the volume, respectively perimeter, of level sets (cf. [22]). In this way and similar to the classical Cheeger problems (cf. [37]), it can be shown that, indeed, ground states only take two-function values and the interface between solves an isoperimetric problem. On a graph the ground states of total variation can be related in a similar way to a graph cut, the so-called Cheeger cut (cf. [41]). In one dimension, for a modified version of total variation that takes into account also the variation across the boundary (assuming extension by zero outside), the ground state can be computed as a piecewise constant function with single discontinuity in the midpoint of the interval. For this approach, also scaling of the eigenfunction is possible. For simplicity, consider $\Omega = (0, 1)$ and let u_1 be the ground state. Then indeed, for $s < 1$, the function

$$u_1^s(x) = \begin{cases} \frac{1}{\sqrt{s}} u_1(\frac{x}{s}) & \text{if } x < s, \\ 0 & \text{if } x > s \end{cases}$$

is another eigenfunction for a larger eigenvalue. Moreover, the dilation

$$u_1^{s,t}(x) = \begin{cases} 0 & \text{if } x < t, \\ \frac{1}{\sqrt{s}} u_1(\frac{x-t}{s}) & \text{if } x < s + t, \\ 0 & \text{if } x > s + t \end{cases}$$

is another eigenfunction if $t \leq 1 - s$. Indeed, such results can be generalized to anisotropic total variation in multiple dimension by scaling and dilation along the coordinate axes.

In the case $p = \infty$, the setup in a Hilbert space is not the one usually referred to as ∞-Laplacian, which rather corresponds to the treatment of the ∞-Laplacian energy

$$J(u) = \|\nabla u\|_\infty$$

in $L^\infty(\Omega)$, while we use the Hilbert space $H = L^2(\Omega)$. In this case an interesting problem is to consider an extended version of J, defined as $+\infty$ if u does not have homogeneous boundary values, i.e., we effectively solve the homogeneous Dirichlet problem. Here the ground state can be computed more easily with a different scaling $J(u) = 1$, i.e., we effectively maximize $\|u\|_2$ subject to $|\nabla u(x)| \leq 1$ almost everywhere. For a domain Ω, this is indeed the case if u is the distance function to the boundary, thus computing the ground state yields an alternative way to compute distance functions, respectively solve the eikonal equation (cf. [17] for further details). On a graph, the computation of the ground state with the above normalization yields a way to define and compute a distance function.

3.2. Ground states and sparsity

An interesting class of functionals J in many applications in signal and image processing are ℓ^1-norms or their continuum counterpart, the total variation of a measure. Such functionals are used to enforce sparsity of their minimizers, i.e., a minimal size of the support.

Let us start in the finite-dimensional case $X = \mathbb{R}^n$, $K : \mathbb{R}^n \to \mathbb{R}^m$ (the latter equipped with the Euclidean norm), and

$$J_0(u) = \|u\|_1 = \sum_{i=1}^{n} |u_i|.$$

Here we use the singular value formulation (3.2), i.e., we want to compute

$$u \in \arg\min_{u \in X} \frac{\|u\|_1}{\|Ku\|_2}.$$

Indeed, the sparsity is present also in the ground state. To see this, let e_i be the ith unit vector and $\tilde{e}_i = \operatorname{sign}(u_i) e_i$. Then, for arbitrary u, we have

$$\|Ku\|_2 = \|u\|_1 \left\| \sum_{i=1}^{n} \sigma_i K \tilde{e}_i \right\|_2,$$

with $\sigma_i = \frac{|u_i|}{\|u\|_1}$. By convexity, we find further

$$\|Ku\|_2 \leq \|u\|_1 \sum_{i=1}^{n} \sigma_i \|K\tilde{e}_i\|_2,$$

and equality holds if u has a single nonzero entry. In particular, we find

$$R(u) = \frac{\|u\|_1}{\|Ku\|_2} \geq \frac{1}{\max_i \|Ke_i\|_2},$$

and thus $u = e_j$ with j such that

$$\|Ke_j\|_2 \geq \|Ke_i\|_2, \quad \forall i \in \{1, \dots, n\}$$

is a ground state. Moreover, the proof shows that there are no other ground states, i.e., all of them have maximal sparsity.

In the infinite-dimensional case, we have $X = \mathcal{M}(\Omega)$ for some domain Ω and K mapping to some Hilbert space, typically with the assumption that K is the adjoint of an

operator L mapping from the Hilbert space to the predual space $C_b(\Omega)$ (cf. [9]). Using the latter issues with the dual space of $\mathcal{M}(\Omega)$ can be avoided and the analysis can be carried out in the predual space. The functional J_0 is the total variation norm

$$J_0(u) = \sup_{\varphi \in C_b(\Omega), \|\varphi\|_\infty \leq 1} \int_\Omega \varphi \, du.$$

With analogous reasoning as above, interpreting a general measure after division by its total variation norm as a convex combination of signed concentrated measures, we see that the ground states are of the form $u = \delta_z$ for $z \in \Omega$ such that

$$\|K\delta_z\| \geq \|K\delta_x\|, \quad \forall x \in \Omega.$$

Example 3.2. Let us consider $\Omega \subset \mathbb{R}^d$ and let $k : \mathbb{R}^d \to \mathbb{R}$ be a continuous and integrable kernel. We consider the convolution operator

$$K : \mathcal{M}(\Omega) \to L^2(\mathbb{R}^d), \quad u \mapsto \int_\Omega k(\cdot - y) du(y).$$

We see that $K\delta_z = k(\cdot - z)$ and thus

$$\|K\delta_z\|_{L^2}^2 = \int_{\mathbb{R}^d} k(x - z)^2 \, dx = \int_{\mathbb{R}^d} k(y)^2 \, dy.$$

Hence, the maximum of $\|K\delta_z\|$ is attained for any $z \in \Omega$, which implies that the concentrated measure δ_z is a ground state for any $z \in \Omega$.

Example 3.3. We return to the case of a polyhedral regularization $J = \chi_C^*$, but actually the argument holds for general convex sets C. Since we know that the ground state is an eigenvector and thus a subgradient in $\partial J(0) = C$, it suffices to minimize the Rayleigh quotient over C. Moreover, the extremal property (2.5) can be satisfied only for $p \in \partial C$, hence we further restrict the possible minimization. Let p be the solution of

$$p = \arg\min_{q \in \partial C} \|q\|,$$

i.e., the element of minimal norm in ∂C. Then

$$R(p) = \frac{J(p)}{\|p\|} = \frac{1}{\|p\|} \sup_{q \in C} \langle q, p \rangle = \frac{1}{\|p\|} \langle p, p \rangle = \|p\|.$$

By analogous reasoning, we can show

$$R(q) \geq \|q\| \geq \|p\| = R(p).$$

Thus, p is indeed a ground state of J.

4. VARIATIONAL PROBLEMS, ITERATIONS, AND FLOWS

A first motivation of the definition of nonlinear singular values was to obtain exact solutions of variational problems of the form

$$u \in \arg\min_{u \in X} \frac{1}{2} \|Ku - f\|^2 + \alpha J_0(u),$$

which frequently arise in the regularization of inverse problems, image processing, and data analysis (cf. [4] and the references therein). It was further extended to some iterative methods and time-continuous flows, which in turn can be used as methods to compute eigenvectors or singular vectors. We shall discuss these developments in the following. For the sake of a simple notation, we denote by u_λ an eigenvector with eigenvalue λ, i.e.,

$$\lambda u_\lambda = p_\lambda \in \partial J(u_\lambda). \tag{4.1}$$

Moreover, throughout the whole section we will use data $f \in H_0$, since for arbitrary $f \in H$ we can factor out the component in $\mathcal{N}(J)$. The latter is just technical and beyond our interest of highlighting the main points of the analysis in this paper.

4.1. Variational regularization methods

We start with a discussion of variational regularization methods, which we rephrase as in Section 2 in a Hilbert space setting, i.e.,

$$u \in \arg\min_{u \in H} \frac{1}{2}\|u - f\|^2 + \alpha J(u). \tag{4.2}$$

We consider f being a multiple of the eigenvector u_λ, i.e., $f = cu_\lambda$, $c > 0$, and look for a solution of the form $u = C(\alpha, \lambda)u_\lambda$. The basis for this investigation is the optimality condition

$$u - f + \alpha p = 0, \quad p \in \partial J(u)$$

satisfied by a solution u of (4.2). Making the ansatz $p = p_\lambda = \lambda u_\lambda$, which is in the subdifferential of $C(\alpha, \lambda)u$ due to the zero-homogeneity of ∂J, we arrive at the scalar relation

$$C(\alpha, \lambda) - c + \alpha\lambda = 0,$$

which yields a positive solution if $c > \alpha\lambda$. If $c \leq \alpha\lambda$, we obtain a solution by choosing $C(\alpha, \lambda) = 0$, since $\frac{c}{\alpha\lambda} \in \partial J(0)$. Thus, we find

$$C(\alpha, \lambda) = (c - \alpha\lambda)_+, \tag{4.3}$$

i.e., the solution is a multiple of an eigenvector with the magnitude obtained by a shrinkage formula. We see that obviously the shrinkage is stronger for larger α, but also for larger λ. Hence, there is less change in smaller eigenvalues (low frequencies) than in larger eigenvalues (high frequencies).

The solutions of this kind can be investigated with respect to their robustness with respect to noise (errors in f) and bias (errors due to positive values of α), see [3]. Let us detail some aspects of bias in the following, a particularly interesting property is that the ground state yields the minimal bias (cf. [3]).

Theorem 4.1. *Let $\alpha > 0$ and $u \notin \mathcal{N}(J)$ be a solution of (4.2). Then*

$$\|u - f\| \geq \alpha\lambda_0,$$

where λ_0 is the minimal eigenvalue of J.

Proof. We employ the optimality condition $p = \frac{1}{\alpha}(f - u)$ with $p \in \partial J(u)$ to obtain

$$J(u) = \langle p, u \rangle = \frac{1}{\alpha}\langle f - u, u \rangle \leq \frac{1}{\alpha}\|u\|\|u - f\|.$$

Moreover, due to our assumption on f and since $p \in H_0$ for every subgradient, we also conclude $u \in H_0$. Due to the definition of the minimal eigenvalue via the ground state, we conclude

$$\|u\| \leq \frac{1}{\lambda_0} J(u).$$

Inserting this relation into the above inequality and canceling $J(u)$, which is possible due to $u \notin \mathcal{N}(J)$, yields the assertion. ∎

In order to get rid of bias effects for low frequencies, several two-step approaches have been proposed in literature for examples of J. A structured approach has been derived in [11], which computes a solution v via minimizing

$$\|v - f\| \to \min_v \quad \text{subject to } J(v) - J(u) - \langle p, v - u \rangle = 0,$$

where u is the solution of (4.2) and p the corresponding subgradient arising in the optimality condition. We can elucidate this scheme in the case of $f = cu_\lambda$. If $c > \alpha\lambda$, we have a nontrivial solution $u = C(\alpha, \lambda)u_\lambda$, thus $v = cu_\lambda$ satisfies $J(v) - J(u) - \langle p, v - u \rangle = 0$ and clearly minimizes $\|v - f\|$. This means that low frequencies are exactly reconstructed by this two-step procedure. An alternative approach to reduce bias is iterative regularization, in particular the Bregman iteration, which can be interpreted as an inexact penalization of the above constraint. We will further investigate the behavior of singular vectors in this iteration in the next part.

4.2. Bregman iterations and inverse scale space flows

The Bregman iteration is obtained by subsequently computing

$$u^{k+1} \in \arg\min_u \frac{1}{2}\|u - f\|^2 + \alpha\big(J(u) - J\big(u^k\big) - \big\langle p^k, u - u^k \big\rangle\big),$$

the penalty being the Bregman distance between u and the last iterate u^k, and p^k the subgradient from the optimality condition for u^k (cf. [36]). The optimality condition directly yields an update formula for the subgradients in the form

$$p^{k+1} = p^k + \frac{1}{\alpha}(f - u^{k+1}).$$

Let us mention that for consistency with the variational method the choice $p^0 = 0$ and $u^0 \in \mathcal{N}(J)$ is usually assumed, without loss of generality we can choose $u^0 = 0$. In this case the variational method (4.2) is just the first step of the Bregman iteration. In order to obtain a suitable result, α has to be chosen large in the Bregman iteration, however.

It is instructive to investigate again the case $f = cu_\lambda$ in the Bregman iteration, looking for a solution of the form $u^k = C^k(\alpha, \lambda)u_\lambda$. If α is large, we may expect to have $c < \alpha\lambda$ and hence $u^1 = 0$, which yields $p^1 = \frac{c}{\alpha}u_\lambda$. Indeed, we obtain

$$p^k = \frac{ck}{\alpha}u_\lambda \quad \text{for } k \leq \frac{\alpha\lambda}{c}.$$

The first iteration step with a nonzero solution u^k is given by $k = K(\alpha, \lambda)$ with

$$K(\alpha, \lambda) = \min\left\{k \in \mathbb{N} \mid k > \frac{\alpha\lambda}{c}\right\}.$$

Here we can easily compute $p^k = \lambda u_\lambda$ and thus

$$u^k = cu_\lambda + \alpha\left(\frac{c(k-1)}{\alpha} - \lambda\right)u_\lambda,$$

i.e.,

$$C^k(\alpha, \lambda) = c - \big(\alpha\lambda - c(k-1)\big) = ck - \alpha\lambda.$$

For $k = K(\alpha, \lambda) + 1$, we obtain again u^k, being a nontrivial multiple of u_λ, and the corresponding subgradient is $p^k = \lambda u_\lambda = p^{k-1}$, which implies $u^k = f$. For further iterations, the result clearly does not change anymore. Hence, the number of iterations needed to obtain the exact solution behaves like $\frac{\alpha\lambda}{c}$, and we see again that eigenvectors for smaller eigenvalues (low frequency) are reconstructed faster, while eigenvectors for larger eigenvalues will appear only very late in the iteration.

The computations are a bit more precise in the limit $\alpha \to \infty$, which yields (after appropriate rescaling of step sizes) a time-continuous flow, the so-called inverse scale space method (cf. [20])

$$\partial_t p(t) = f - u(t), \quad p(t) \in \partial J\big(u(t)\big).$$

By analogous reasoning, we can compute the solution for $u(0) = p(0) = 0$ and $f = cu_\lambda$ as

$$u(t) = \begin{cases} 0, & t < \frac{\lambda}{c}, \\ cu_\lambda, & t > \frac{\lambda}{c}. \end{cases}$$

Thus, the reconstruction becomes exact at a time proportional to the eigenvalue.

4.3. Gradient flows

Another iterative scheme obtained from the variational method (4.2) is to start with $u = f$ and solve for

$$u^{k+1} \in \arg\min_u \frac{1}{2}\|u - u^k\|^2 + \alpha J(u).$$

Again, the first step is consistent with (4.2), but the dynamics is very different from the Bregman iteration, in particular for small α, which is the relevant case here. Choosing $f = cu_\lambda$, the optimality condition

$$u^{k+1} - u^k + \alpha p^{k+1} = 0, \quad p^{k+1} \in \partial J(u^{k+1})$$

yields $u^1 = (c - \alpha\lambda)u_\lambda$, and by analogous reasoning

$$u^k = (c - k\alpha\lambda)u_\lambda,$$

as long as $k < \frac{c}{\alpha\lambda}$. For $k > \frac{c}{\alpha\lambda}$, we can indeed verify that $u^k = 0$ solves the problem. Here we see that eigenvectors related to larger eigenvalues (high frequencies) shrink faster to zero, whereas the ground state is the last to disappear.

Again, this can be made more precise for the time-continuous variant, this time obtained as $\alpha \to 0$. The above iteration scheme is also known as minimizing movement scheme (cf. [23]) and the limit for appropriately scaled time is the gradient flow

$$\partial_t u(t) = -p(t), \quad p(t) \in \partial J\big(u(t)\big), \tag{4.4}$$

with initial value $u(0) = f$. Again, the solution of the gradient flow has a simple form if $f = cu_\lambda$, namely

$$u(t) = \begin{cases} (c - \lambda t)u_\lambda, & t < \frac{c}{\lambda}, \\ 0, & t > \frac{c}{\lambda}. \end{cases}$$

This means that solutions shrink to zero linearly in time, and the extinction time $\frac{c}{\lambda}$ again changes with the eigenvalue. Due to the inverse relation with λ, low frequencies get extinct later than high ones.

The behavior on eigenvectors motivates studying the gradient flow also for arbitrary initial values f. First of all, we can generalize the finite time extinction. For initial value $f \in H_0$, it is easy to show that $u(t) \in H_0$ for all $t > 0$, since for $v \in \mathcal{N}(J)$ we have

$$\big\langle u(t), v\big\rangle = \langle f, v\rangle - \int_0^t \big\langle p(s), v\big\rangle\, ds = 0.$$

Now we can use we use the standard dissipation relation

$$\big\|u(t)\big\|^2 + 2\int_0^t J\big(u(s)\big)\, ds \le \|f\|^2,$$

and $J(u(s)) \ge \lambda_0 \|u(s)\|$, resulting in

$$\big\|u(t)\big\|^2 + 2\lambda_0 \int_0^t \big\|u(s)\big\|\, ds \le \|f\|^2.$$

Similar to the proof of the Gronwall inequality, this allows deducing

$$\big\|u(t)\big\| \le \|f\| - \lambda_0 t, \quad \text{for } t < \frac{\|f\|}{\lambda_0}.$$

Thus $u(t) = 0$ for $t = \frac{\|f\|}{\lambda_0}$, and it is easy to show that for $t > \frac{\|f\|}{\lambda}_0$ the unique solution of the gradient flow is given by $u(t) = 0$ and $p(t) = 0$. Thus, the gradient flow exhibits a finite extinction phenomenon, the solution vanishes after finite time. We define the extinction time as

$$t_*(f) = \inf\big\{t > 0 \mid u(t) = 0\big\}. \tag{4.5}$$

Our analysis above yields the following upper bound on the extinction time:

Theorem 4.2. *Let $f \in H_0$ and $u \in C(0, T; H)$ be a solution of the gradient flow (4.4). Then the extinction time defined by (4.5) satisfies*

$$t_*(f) \le \frac{\|f\|}{\lambda_0}, \tag{4.6}$$

where λ_0 is the minimal nontrivial eigenvalue of J.

From the special case of f being a multiple of a ground state, we see that (4.6) is sharp for suitable initial values. In order to gain further understanding, it is instructive to investigate scalar products of the solution u with eigenvectors. This leads to

$$\langle u(t), u_\lambda \rangle = \langle f, u_\lambda \rangle - \int_0^t \langle p(s), u_\lambda \rangle \, ds \geq \langle f, u_\lambda \rangle - \lambda t.$$

Thus we obtain lower bounds on the extinction time of the form

$$t_*(f) \geq \frac{1}{\lambda} \big| \langle f, u_\lambda \rangle \big|$$

and see that the extinction time will be larger the more the initial value is correlated with low frequencies.

The extinction time is not the only relevant quantity, but also the so-called *extinction profile* is of high relevance. The extinction profile v_f is defined as

$$v_f = \lim_{\tau \downarrow 0} \frac{1}{\tau} u\big(t_*(f) - \tau\big),$$

i.e., it is the left-sided derivative of the gradient flow at the extinction time. Surprisingly, it can be shown that v_f is an eigenvector of J, under suitable conditions even that it is the ground state. This was shown first for the total variation flow (cf., e.g., [2]) and later also other zero-homogeneous evolution equations such as the fast diffusion equation (cf. [6,7]). In [15,16] this has been reconsidered in the abstract setting of eigenvectors of seminorms and general results on the extinction profile could be obtained. Let us just motivate formally why it can be expected that the extinction profile is an eigenvector. From the optimality condition in the minimizing movement scheme with τ chosen appropriately, we obtain

$$\frac{1}{\tau} u\big(t_*(f) - \tau\big) = \frac{1}{\tau} \big(u\big(t_*(f) - \tau\big) - u\big(t_*(f)\big) \big) = p\big(t_*(f)\big).$$

Hence, if $p(t_*(f))$ is not vanishing, the limit $\tau \downarrow 0$ yields $v_f = p(t_*(f))$ and, due to the homogeneity of the subdifferential, we also obtain $p(t_*(f)) \in \partial J(v_f)$ in the limit. Thus, v_f, respectively its rescaled version, is an eigenvector of J.

We finally mention that an extension of the results on the extinction profile has been carried out in [14], which analyzes the fine asymptotics for gradient flows of p-homogeneous functionals. In the case $p < 2$, there is still an extinction profile with similar properties, for $p \geq 2$ there is only decay as $t \to \infty$, however. Appropriately rescaled versions of the asymptotics of the solution are again eigenvectors of the underlying functionals.

4.4. Gradient flows and spectral decompositions

Gradient flows are particularly interesting for computing eigenvectors and even spectral decompositions, since the classical theory by Brezis (cf. [10]) implies that indeed the solution selects subgradients of minimal norm, i.e.,

$$\partial_t u(t) = -p^0(t), \quad p^0(t) = \arg\min \big\{ \|p\| \mid p \in \partial J\big(u(t)\big) \big\}.$$

Thus, we obtain a spectral decomposition into eigenvectors $p^0(t)$ with the Lebesgue measure on $\mathbb{R}_+$ if all subgradients of minimal norm are indeed eigenvectors. This means that (2.5)

needs to be satisfied for the subgradients of minimal norm in $\partial J(v)$ for all $v \in H$. Then we have indeed

$$f = \int_0^\infty p(t)\, dt,$$

but this is not yet a spectral decomposition in the above sense, since the integration is not with respect to a measure of the eigenvalue. However, as seen in [15], a change of measure from t to $\lambda(t)$ indeed yields a spectral decomposition. Note, however, that since $p(t)$ and thus $\lambda(t)$ can be piecewise constant, the arising measure in the spectral domain is not absolutely continuous with respect to the Lebesgue measure in typical cases.

In [16] an alternative way to obtain a spectral decomposition in separable Hilbert spaces was derived via extinction profiles. This countable spectral decomposition is obtain by first computing the extinction profile of the gradient flow with starting value f and then projecting f onto the space orthogonal to the first extinction profile. This projection is used again as starting value of the gradient flow and then again f is projected onto the space orthogonal to the new extinction profile. Iterating this procedure the projections converge to zero and the sum of the orthogonal components yields an atomic spectral decomposition.

There are several examples of flows that yield a spectral decomposition, the most prominent one being the one-dimensional total variation (cf. [15]). Other examples are polyhedral regularizations with sufficiently regular convex sets C (cf. [18, 19]) and one-homogeneous functionals vector fields using divergence and rotation (cf. [15]).

5. APPLICATIONS

In order to illustrate the use of nonlinear eigenvalue problems in data science, we discuss two toy examples representing wider classes of applications in this section.

5.1. 1-Laplacian graph clustering

We start with a common technique for data clustering, namely the computation of the first eigenfunction of the 1-Laplacian on graphs. For this, we acquire data on a surface by a laser scanner with random sampling, as illustrated in Figure 1. This resembles the classical two-moons data set frequently used for the evaluation of clustering methods. Based on those data points we build a nearest neighbor graph as illustrated in the right image of Figure 1.

On the arising graph we compute the first nontrivial eigenfunction of the classical graph Laplacian, which is shown in the left part of Figure 2. This serves mainly for comparison with the eigenfunction of the 1-Laplacian on the graph (the ground state of the graph total variation), which is shown in the right part. The ground state can be computed as an extinction profile of the gradient flow with the graph Laplacian eigenfunction as a starting value (cf. [16]). It is apparent that the eigenfunction of the 1-Laplacian has a much sharper transition between positive and negative values, which corresponds closely to the geometric structure in the data. This leads to improved spectral clustering as shown in Figure 3, namely the sub- and superlevel sets at zero (in red, respectively blue). One observes a rather linear

FIGURE 1
Image of traditional Austrian Christmas cookies (Vanillekipferl) and random sampling of points on the surface (left, respectively middle) and neighborhood graph built out of the sample points.

FIGURE 2
First nontrivial eigenfunction of the graph (2-)Laplacian (left) and the graph 1-Laplacian (right).

FIGURE 3
Spectral clustering based on the graph (2-)Laplacian (left) and the graph 1-Laplacian (right).

structure in the clustering with the graph Laplacian, while the clustering with the 1-Laplacian perfectly adapts to the structure in the data set.

5.2. Distance functions from ∞-Laplacians

In the following we illustrate the computation of distance functions by minimizing the ∞-Laplacian energy

$$J(u) = \operatorname{ess\,sup}_x \left|\nabla u(x)\right|. \tag{5.1}$$

We use the graph Laplacian energy on a grid graph built on the map of the United Kingdom with a large stencil. In this case we compute the (nonnegative) ground state over the set of functions on the graph vanishing on a predefined boundary (corresponding to the geographical boundary). We then normalize it such that $J(u) = 1$, which implies that u becomes the

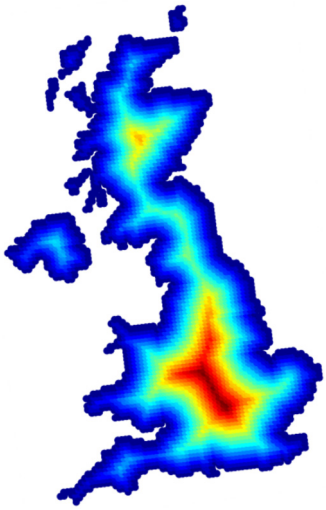


FIGURE 4
Distance function on a grid graph for the geometry of the United Kingdom.

distance function to the boundary. The result is shown in Figure 4 and generalizes results obtained by solving the eikonal equation in the continuum setting.

ACKNOWLEDGMENTS

The author thanks Daniel Tenbrinck (FAU Erlangen-Nürnberg) for measurements and computations in Section 5.1 and Leon Bungert (Bonn University) for computations in Section 5.2.

FUNDING

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 777826 (NoMADS) and by the ERC via Grant EU FP7—ERC ConsolidatorGrant 615216 LifeInverse.

REFERENCES

[1] H. Amann, Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces. *SIAM Rev.* **18** (1976), 620–709.

[2] G. Bellettini, V. Caselles, and M. Novaga, The total variation flow in $\mathbb{R}^n$. *J. Differential Equations* **184** (2002), 475–525.

[3] M. Benning and M. Burger, Ground states and singular vectors of convex variational regularization methods. *Methods Appl. Anal.* **20** (2013), 295–334.

[4] M. Benning and M. Burger, Modern regularization methods for inverse problems. *Acta Numer.* **27** (2018), 1–111.

[5] M. Benning, M. Möller, R. Z. Nossek, M. Burger, D. Cremers, G. Gilboa, and C. Schönlieb, Nonlinear spectral image fusion. In *International conference on scale space and variational methods in computer vision*, pp. 41–53, Springer, 2017.

[6] M. Bonforte and A. Figalli, Total variation flow and sign fast diffusion in one dimension. *J. Differential Equations* **252** (2012), 4455–4480.

[7] M. Bonforte, G. Grillo, and J. Vazquez, Behaviour near extinction for the Fast Diffusion Equation on bounded domains. *J. Math. Pures Appl.* **97** (2012), 1–38.

[8] R. Bracewell, *Fourier analysis and imaging*. Springer, 2004.

[9] K. Bredies and H. Pikkarainen, Inverse problems in spaces of measures. *ESAIM Control Optim. Calc. Var.* **19** (2013), 190–218.

[10] H. Brezis, *Operateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*. Elsevier, 1973.

[11] E. M. Brinkmann, M. Burger, J. Rasch, and C. Sutour, Bias reduction in variational regularization. *J. Math. Imaging Vision* **59** (2017), 534–566.

[12] T. Bühler and M. Hein, Spectral clustering based on the graph p-Laplacian. In *Proceedings of the 26th annual international conference on machine learning*, pp. 81–88, ACM, 2009.

[13] L. Bungert and M. Burger, Gradient flows, nonlinear power methods, and computation of nonlinear Eigenfunctions. In *Handbook of numerical analysis*, to appear.

[14] L. Bungert and M. Burger, Asymptotic profiles of nonlinear homogeneous evolution equations of gradient flow type. *J. Evol. Equ.* **20**, 1061–1092.

[15] L. Bungert, M. Burger, A. Chambolle, and M. Novaga, Nonlinear spectral decompositions by gradient flows of one-homogeneous functionals. *Anal. PDE* **14** (2021), 823–860.

[16] L. Bungert, M. Burger, and D. Tenbrinck, Computing nonlinear eigenfunctions via gradient flow extinction. In *International conference on scale space and variational methods in computer vision*, pp. 485–497, Springer, 2019.

[17] L. Bungert, Y. Korolev, and M. Burger, Structural analysis of an L^∞ variational problem and relations to distance functions. *Pure Appl. Anal.* **2** (2020), 703–738.

[18] M. Burger, L. Eckardt, G. Gilboa, and M. Moeller, Spectral representations of one-homogeneous functionals. In *International conference on scale space and variational methods in computer vision*, pp. 16–27, Springer, 2015.

[19] M. Burger, G. Gilboa, M. Moeller, L. Eckardt, and D. Cremers, Spectral decompositions using one-homogeneous functionals. *SIAM J. Imaging Sci.* **9** (2016), 1374–1408.

[20] M. Burger, G. Gilboa, S. Osher, J. Xu, and W. Yin, Nonlinear inverse scale space methods. *Commun. Math. Sci.* **4** (2006), 179–212.

[21] J. Calder, The game theoretic p-Laplacian and semi-supervised learning with few labels. *Nonlinearity* **32** (2018), 301.

[22] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, An introduction to total variation for image analysis. In *Theoretical foundations and numerical methods for sparse recovery*, pp. 1455–1499, De Gruyter, Berlin, 2010.

[23] E. De Giorgi, New problems on minimizing movements. In *Ennio de Giorgi: selected papers*, pp. 699–713, Springer, 1993.

[24] M. Esteban, M. Lewin, and E. Sere, Variational methods in relativistic quantum mechanics. *Bull. Amer. Math. Soc.* **45** (2008), 535–593.

[25] T. Feld, J. Aujol, G. Gilboa, and N. Papadakis, Rayleigh quotient minimization for absolutely one-homogeneous functionals. *Inverse Probl.* **35** (2019), 06400.

[26] M. Fumero, M. Moeller, and E. Rodola, Nonlinear spectral geometry processing via the TV transform. *ACM Trans. Graph.* **39** (2020), 1–16.

[27] G. Gilboa, A spectral approach to total variation. In *International conference on scale space and variational methods in computer vision*, pp. 36–47, Springer, 2013.

[28] G. Gilboa, A total variation spectral framework for scale and texture analysis. *SIAM J. Imaging Sci.* **7** (2014), 1937–1961.

[29] G. Gilboa, Iterative methods for computing Eigenvectors of nonlinear operators. In *Handbook of mathematical models and algorithms in computer vision and imaging: mathematical imaging and vision*, pp. 1–28, Springer, 2021.

[30] G. Gilboa, M. Moeller, and M. Burger, Nonlinear spectral analysis via one-homogeneous functionals: Overview and future prospects. *J. Math. Imaging Vision* **56** (2016), 300–319.

[31] M. Hein and T. Bühler, An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In *Advances in neural information processing systems 23*, edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, pp. 847–855, Curran Associates, 2010.

[32] R. Hynd and E. Lindgren, Approximation of the least Rayleigh quotient for degree-p homogeneous functionals. *J. Funct. Anal.* **272** (2012), 4873–4918.

[33] J. Jost, R. Mulas, and D. Zhang, p-Laplace operators for oriented hypergraphs. *Vietnam J. Math.* (2021), 1–36.

[34] B. Kawohl and P. Lindqvist, Positive eigenfunctions for the p-Laplace operator revisited. *Analysis* **26** (2006), 545–550.

[35] A. Le, Eigenvalue problems for the p-Laplacian. *Nonlinear Anal.* **64** (2006), 1057–1099.

[36] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, An iterative regularization method for total variation-based image restoration. *Multiscale Model. Simul.* **4** (2005), 460–489.

[37] E. Parini, An introduction to the Cheeger problem. *Surv. Math. Appl.* **6**, 9–21.

[38] P. H. Rabinowitz, Some global results for nonlinear eigenvalue problems. *J. Funct. Anal.* **7** (1971), 487–513.

[39] Y. Saad, *Numerical methods for large eigenvalue problems, revised edn*. Society for Industrial and Applied Mathematics, Philadelphia, 2011.

[40] G. Strang, *Introduction to linear algebra. 5th edn*. Wellesley-Cambridge Press, Wellesley, 1993.

[41] A. Szlam and X. Bresson, Total variation and Cheeger cuts. In *ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 1039–1046, Ominpress, Haifa, 2010.

[42] U. von Luxburg, A tutorial on spectral clustering. *Stat. Comput.* **17** (2007), 395–416.

[43] M. I. Weinstein, Nonlinear Schrödinger equations and sharp interpolation estimates. *Comm. Math. Phys.* **87** (1982), 567–576.

MARTIN BURGER

Department Mathematik, Friedrich-Alexander Universität Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany, martin.burger@fau.de

THE EVALUATION COMPLEXITY OF FINDING HIGH-ORDER MINIMIZERS OF NONCONVEX OPTIMIZATION

CORALIA CARTIS, NICHOLAS I. M. GOULD, AND PHILIPPE L. TOINT

ABSTRACT

We introduce the concept of strong high-order approximate minimizers of nonconvex optimization problems. These apply in both standard smooth and composite nonsmooth settings, and additionally allow convex or inexpensive constraints. An adaptive regularization algorithm is then proposed to find such approximate minimizers. Under suitable Lipschitz continuity assumptions, the evaluation complexity of this algorithm is investigated. The bounds obtained not only provide, to the best of our knowledge, the first known result for (unconstrained or inexpensively-constrained) composite problems for optimality orders exceeding one, but also give the first sharp bounds for high-order strong approximate qth order minimizers of standard (unconstrained and inexpensively constrained) smooth problems, thereby complementing known results for weak minimizers.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 90C60; Secondary 90C26, 90C30, 49M15, 68Q25

KEYWORDS

Nonconvex optimization, composite optimization, regularization methods, complexity bounds, global rates of convergence

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5256–5289
DOI 10.4171/ICM2022/95

1. INTRODUCTION

We consider composite optimization problems of the form

$$\min_{x \in \mathcal{F}} w(x) \stackrel{\text{def}}{=} f(x) + h\big(c(x)\big), \tag{1.1}$$

where f, c are smooth and h possibly nonsmooth but Lipschitz continuous, and where $\mathcal{F}$ is a feasible set associated with inexpensive constraints (which are discussed in the next paragraph). Such problems have attracted considerable attention, due to the their occurrence in important applications such as LASSO methods in computational statistics [26], Tikhonov regularization of underdetermined estimation problems [21], compressed sensing [16], artificial intelligence [22], penalty or projection methods for constrained optimization [8], least Euclidean distance and continuous location problems [17], reduced-precision deep-learning [27], image processing [2], to cite but a few examples. We refer the reader to the thorough review in [23]. In these applications, the function h is typically globally Lipschitz continuous and cheap to compute—common examples include the Euclidean, ℓ_1, or ℓ_∞ norms.

Inexpensive constraints defining the feasible set $\mathcal{F}$ are constraints whose evaluation or enforcement has negligible cost compared to that of evaluating f, c and/or their derivatives. They are of interest here since the evaluation complexity of solving inexpensively constrained problems is dominated solely by the number of evaluations of f, c and their derivatives. Inexpensive constraints include, but are not limited to, convex constraints with cheap projections (such as bounds or the ordered simplex). Such constraints have already been considered elsewhere [3, 12].

Of course, problem (1.1) may be viewed as a general nonsmooth optimization problem, to which a battery of existing methods may be applied (for example, subgradient, proximal gradient, and bundle methods). However, this avenue ignores the problem's special structure, which may be viewed as a drawback. More importantly for our purpose, this approach essentially limits the type of approximate minimizers one can reasonably hope for to first-order points (see [18, CHAPTER 14] for a discussion of second-order optimality conditions and [8, 20] for examples of structure-exploiting first-order complexity analysis). However, our first objective in this paper is to cover *approximate minimizers of arbitrary order* (obviously including first- and second-order ones), in a sense that we describe below. This, as far we know, precludes a view of (1.1) that ignores the structure present in h.

It is also clear that any result we can obtain for problem (1.1) also applies to standard smooth problems (by letting h be the zero function), for which evaluation complexity results are available. Most of these results cover first- and second-order approximate minimizers (see [7, 10, 15, 24, 25] for a few references), but two recent papers [11, 12] propose an analysis covering our stated objective to cover arbitrary-order minimizers for smooth nonconvex functions. However, these two proposals significantly differ, in that they use different definitions of high-order minimizers, by no means a trivial concept. The first paper, focusing on trust-region methods, uses a much stronger definition than the second one which covers adaptive regularization algorithms. Our second objective in the present paper is to strengthen these latter results to *use the stronger definition of optimality for adaptive regu-*

larization algorithms and therefore bridge the gap between the two previous approaches in the more general framework of composite problems.

Contributions and motivation. The main contributions of this paper may be summarized as follows:

(1) We formalize the notion of strong approximate minimizer of arbitrary order for standard (noncomposite) smooth problems and extend it to composite ones, including the case where the composition function is nonsmooth, and additionally allow inexpensive constraints. This notion is stronger than that of "weak" approximate minimizers used in [3, 12, 14].

(2) We provide a conceptual adaptive regularization algorithm whose purpose is to compute such strong approximate minimizers.

(3) We analyze the worst-case complexity of this conceptual algorithm both for composite and standard problems, allowing arbitrary optimality order and any degree of the model used within the algorithm. For composite problems, these bounds are the first ones available for approximate minimizers of order exceeding one. For smooth problems, the bounds are shown to improve on those derived in [11] for trust-region methods, while being less favorable (for orders beyond the second) than those in [12] for approximate minimizers of the weaker sort. These bounds are summarized in Table 1.1 in the case where all ε_j are identical. Each table entry also mentions existing references for the quoted result. Sharpness (in the order of ε) is also reported when known.

We acknowledge upfront that our approach is essentially theoretical, because it depends, in its present incarnation, on computing global minimizers of Taylor series within Euclidean balls, a problem which is known to be a very hard for high orders [1]. Although these calculations do not involve any evaluation of the problem's objective function or of its derivatives (and thus do not affect evaluation complexity bounds), this is a significant hurdle. While realistic algorithms may have to resort to inexact global minimization (we discuss the necessity and impact of such approximations in Section 7), the case of exact ones can be viewed as an idealized, aspirational setting and the complexity results derived therein as "best possible." Thus we ask here the mathematically important questions: what would be achievable in this idealized setting? Or if constrained global minimizers of polynomials were computable because of special problem structure? A second motivation is that high-order models have already proved their usefulness in practice, in particular in the solution of highly nonlinear low-dimensional least-squares problems [19], even if implementing algorithms using them is far from obvious [4]. The identification of approximate minimizers of orders matching the degree of the models is, in our view, an obvious, yet unexplored question. Moreover, the consideration of such approximate minimizers results in new insights in the definition of approximate minimizers and prompts a proposal for a new approximate optimality measure (see Section 2). At variance with standard ones, this proposal has the

	inexpensive constraints	Weak minimizers		Strong minimizers					
		smooth ($h = 0$)		smooth ($h = 0$)		composite			
						h convex		h nonconvex	
$q = 1$	none	$\varepsilon^{-\frac{p+1}{p}}$	sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$	sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$	sharp †	ε^{-2}	[8,20]
	convex	$\varepsilon^{-\frac{p+1}{p}}$	sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$	sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$	sharp	ε^{-2}	
	nonconvex	$\varepsilon^{-\frac{p+1}{p}}$	sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$	sharp [5,12]	ε^{-2}		ε^{-2}	
$q = 2$	none	$\varepsilon^{-\frac{p+1}{p-1}}$	sharp [12]	$\varepsilon^{-\frac{p+1}{p-1}}$	sharp [12]	ε^{-3}		ε^{-3}	
	convex	$\varepsilon^{-\frac{p+1}{p-1}}$	sharp [12]	$\varepsilon^{-\frac{2(p+1)}{p}}$	sharp	ε^{-3}		ε^{-3}	
	nonconvex	$\varepsilon^{-\frac{p+1}{p-1}}$	sharp [12]	$\varepsilon^{-\frac{2(p+1)}{p}}$	sharp	ε^{-3}		ε^{-3}	
$q > 2$	none, or general	$\varepsilon^{-\frac{p+1}{p-q+1}}$	sharp [12]	$\varepsilon^{-\frac{q(p+1)}{p}}$	sharp	$\varepsilon^{-(q+1)}$		$\varepsilon^{-(q+1)}$	

TABLE 1.1
Order bounds (as multiples of powers of the accuracy ε) on the worst-case evaluation complexity of finding weak/strong (ε, δ)-approximate minimizers for composite and smooth problems, as a function of optimality order (q), model degree (p), convexity of the composition function h and presence/absence/convexity of inexpensive constraints. The dagger indicates that this bound for the special case when $h(\cdot) = \|\cdot\|_2$ and $f = 0$ is already known [9].

advantage of being well-defined and consistent across all orders and it is obviously also applicable (and computationally cheap) for orders one and two.

Outline. The paper is organized as follows. Section 2 outlines some useful background and motivation on high-order optimality measures. In Section 3, we describe our problem more formally and introduce the notions of weak and strong high-order approximate minimizers. We describe an adaptive regularization algorithm for problem (1.1) in Section 4, while Section 5 discusses the associated evaluation complexity analysis. Section 6 then shows that several of the obtained complexity bounds are sharp, while Section 7 discusses the necessity of global minimizations and the impact of allowing them to be inexact. Some conclusions and perspectives are finally outlined in Section 8.

2. A DISCUSSION OF qTH-ORDER NECESSARY OPTIMALITY CONDITIONS

Before going any further, it is best to put our second objective (establishing strong complexity bounds for arbitrary qth order using an adaptive regularization method) in perspective by briefly discussing high-order optimality measures. For this purpose, we now digress slightly and first focus on the standard unconstrained (smooth) optimization problem where one tries to minimize an objective function f over $\mathbb{R}^n$. The definition of a jth-order approximate minimizer of a general (sufficiently) smooth function f is a delicate question. It was argued in [11] that expressing the necessary optimality conditions at a given point x in terms of individual derivatives of f at x leads to extremely complicated expressions involving the potential decrease of the function along all possible feasible arcs emanating from x. To avoid this, an alternative based on Taylor expansions was proposed. Such an expansion is given by

$$T_{f,q}(x,d) = \sum_{\ell=0}^{q} \frac{1}{\ell!} \nabla_x^\ell f(x)[d]^\ell \tag{2.1}$$

where $\nabla_x^\ell f(x)[d]^\ell$ denotes the ℓth-order cubically[1] symmetric derivative tensor (of dimension ℓ) of f at x applied to ℓ copies of the vector d. The idea of the *approximate* necessary condition that we use is that, if x is a local minimizer and $q \leq p$ is an integer, there should be a neighborhood of x of radius $\delta_j \in (0,1]$ in which the decrease in (2.1), which we measure by

$$\phi_{f,j}^{\delta_j}(x) \stackrel{\text{def}}{=} f(x) - \min_{d \in \mathbb{R}^n, \|d\| \leq \delta_j} T_{f,j}(x,d), \tag{2.2}$$

must be small. In fact, it can be shown [11, LEMMA 3.4] that

$$\lim_{\delta_j \to 0} \frac{\phi_{f,j}^{\delta_j}(x)}{\delta_j^j} = 0, \tag{2.3}$$

whenever x is a local minimizer of f. Making the ratio in this limit small for small enough δ_j therefore seems reasonable. Let ε_j is a prescribed order-dependent accuracy parameter, and $\varepsilon \stackrel{\text{def}}{=} (\varepsilon_1, \ldots, \varepsilon_q)$. Also let $\delta \stackrel{\text{def}}{=} (\delta_1, \ldots, \delta_q)$ be a vector of associated "optimality radii." Then we will say that x is a *strong* (ε, δ)-approximate qth-order minimizer if, for all $j \in \{1, \ldots, q\}$, there exists a $\delta_j > 0$ such that

$$\phi_{f,j}^{\delta_j}(x) \leq \varepsilon_j \frac{\delta_j^j}{j!}. \tag{2.4}$$

(The factor $j!$ is introduced for notational convenience.) The δ_j are called optimality radii because they are the radii of the *neighborhood of x in which the Taylor series $T_{f,j}(x,d)$ cannot decrease more than ε_j (appropriately scaled)*. Thus δ_j and ε_j are tightly linked (see Lemma 4.4 below) and the limit (2.3) (which applies at true local minimizers) is conceptually achieved when ε_j itself tends to zero. Note that (2.4) reduces to the condition $\|\nabla_x^1 f(x)\| \leq \varepsilon_1$ for $j = 1$, and that, for $j = 2$, $\phi_{f,2}^{\delta_j}(x)$ is obtained by solving a trust-region subproblem,

1 Meaning all its dimensions are the same.

a process whose cost is comparable to that of computing the leftmost eigenvalue of the Hessian, as would be required for the standard second-order measure.

The definition (2.4) should be contrasted with notion of weak minimizers introduced in [12]. Formally, x is a *weak* (ε, δ)-approximate qth-order minimizer if there exists $\delta_q \in \mathbb{R}$ such that

$$\phi_{f,q}^{\delta_q}(x) \leq \varepsilon_q \chi_q(\delta_q) \quad \text{where } \chi_q(\delta) \stackrel{\text{def}}{=} \sum_{\ell=1}^{q} \frac{\delta^\ell}{\ell!}. \tag{2.5}$$

Obviously, (2.5) is less restrictive than (2.4) since it is easy to show that $\chi_q(\delta) \in [\delta, 2\delta)$ and is thus significantly larger than $\delta_q^q/q!$ for small δ_q. Moreover, (2.5) is a single condition, while (2.4) has to hold for all $j \in \{1, \ldots, q\}$. The interest of considering weak approximate minimizers is that they can be computed faster than strong ones. It is shown in [12] that the evaluation complexity bound for finding them is $O(\varepsilon^{-\frac{p+1}{p-q+1}})$, thereby providing a smooth extension to high-order of the complexity bounds known for $q \in \{1, 2\}$. However, the major drawback of using the weak notion is that, at variance with (2.4), it is not coherent with the scaling implied by (2.3).[2] Obtaining this coherence therefore comes at a cost for orders beyond two, as will be clear in our developments below.

If we now consider that inexpensive constraints are present in the problem, it is easy to adapt the notions of weak and strong optimality for this case by (re)defining

$$\phi_{f,j}^{\delta_j}(x) \stackrel{\text{def}}{=} f(x) - \min_{x+d \in \mathcal{F},\, \|d\| \leq \delta_j} T_{f,j}(x, d), \tag{2.6}$$

where $\mathcal{F}$ is the feasible set. We then say that x is a strong inexpensively constrained (ε, δ)-approximate qth-order minimizer if, for all $j \in \{1, \ldots, q\}$, there exists a $\delta_j > 0$ such that (2.4) holds with this new definition.

3. THE COMPOSITE PROBLEM AND ITS PROPERTIES

We now return to the more general composite optimization (1.1), and make our assumptions more specific:

AS.1 The function f from $\mathbb{R}^n$ to $\mathbb{R}$ is p times continuously differentiable and each of its derivatives $\nabla_x^\ell f(x)$ of order $\ell \in \{1, \ldots, p\}$ are Lipschitz continuous in a convex open neighborhood of $\mathcal{F}$, that is, for every $j \in \{1, \ldots, p\}$, there exists a constant $L_{f,j} \geq 1$ such that, for all x, y in that neighborhood,

$$\left\| \nabla_x^j f(x) - \nabla_x^j f(y) \right\| \leq L_{f,j} \|x - y\|, \tag{3.1}$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced operator norm for matrices and tensors.

AS.2 The function c from $\mathbb{R}^n$ to $\mathbb{R}^m$ is p times continuously differentiable and each of its derivatives $\nabla_x^\ell c(x)$ of order $\ell \in \{1, \ldots, p\}$ are Lipschitz continuous in

2 In the worst case, it may lead to the origin being accepted as a second-order approximate minimizer of $-x^2$.

a convex open neighborhood of $\mathcal{F}$, that is, for every $j \in \{1, \ldots, p\}$ there exists a constant $L_{c,j} \geq 1$ such that, for all x, y in that neighborhood,

$$\left\|\nabla_x^j c(x) - \nabla_x^j c(y)\right\| \leq L_{c,j}\|x - y\|, \tag{3.2}$$

AS.3 The function h from $\mathbb{R}^m$ to $\mathbb{R}$ is Lipschitz continuous, subbadditive, and zero at zero, that is, there exists a constant $L_{h,0} \geq 0$ such that, for all $x, y \in \mathbb{R}^m$,

$$\left|h(x) - h(y)\right| \leq L_{h,0}\|x - y\|, \tag{3.3}$$

$$h(x + y) \leq h(x) + h(y) \quad \text{and} \quad h(0) = 0. \tag{3.4}$$

AS.4 There is a constant w_{low} such that $w(x) \geq w_{\text{low}}$ for all $x \in \mathcal{F}$.

Assumption AS.3 allows a fairly general class of composition functions. Examples include the popular $\|\cdot\|_1$, $\|\cdot\|$, and $\|\cdot\|_\infty$ norms, concave functions vanishing at zero and, in the unidimensional case, the ReLu function $\max[0, \cdot]$ and the periodic $|\sin(\cdot)|$. As these examples show, nonconvexity and nondifferentiability are allowed (but not necessary). Note that finite sums of functions satisfying AS.3 also satisfy AS.3. Note also that h being subadditive does not imply that h^α is also subadditive for $\alpha \geq 1$ ($h(c) = c$ is, but $h(c)^2$ is not), or that it is concave [6]. Observe finally that equality always holds in (3.4) when h is odd.[3]

When h is smooth, problem (1.1) can be viewed either as composite or smooth. Does the composite view present any advantage in this case? The answer is that the assumptions needed on h in the composite case are weaker in that Lipschitz continuity is only required for h itself, not for its derivatives of orders 1 to p. If any of these derivatives are costly, unbounded or nonexistent, this can be a significant advantage. However, as we will see below (in Theorems 5.5 and 5.6) this comes at the price of a worse evaluation complexity bound. For example, the case of linear h is simple to assess, since in that case $h(c)$ amounts to a linear combination of the c_i, and there is obviously no costly or unbounded derivative involved: a smooth approach is therefore preferable from a complexity perspective.

Observe also that AS.1 and AS.2 imply, in particular, that

$$\left\|\nabla_x^j f(x)\right\| \leq L_{f,j-1} \quad \text{and} \quad \left\|\nabla_x^j c(x)\right\| \leq L_{c,j-1} \quad \text{for } j \in \{2, \ldots, p\} \tag{3.5}$$

Observe also that AS.3 ensures that, for all $x \in \mathbb{R}^m$,

$$\left|h(x)\right| = \left|h(x) - h(0)\right| \leq L_{h,0}\|x - 0\| = L_{h,0}\|x\|. \tag{3.6}$$

For future reference, we define

$$L_w \stackrel{\text{def}}{=} \max_{j \in \{1,\ldots,p\}} (L_{f,j-1} + L_{h,0}L_{c,j-1}). \tag{3.7}$$

We note that AS.4 makes the problem well-defined in that its objective function is bounded below. We now state a useful lemma on the Taylor expansion's error for a general function r with Lipschitz continuous derivative.

3 Indeed, $h(-x - y) \leq h(-x) + h(-y)$ and thus, since h is odd, $-h(x + y) \leq -h(x) - h(y)$, which, combined with (3.4), gives that $h(x + y) = h(x) + h(y)$.

Lemma 3.1. *Let $r : \mathbb{R}^n \to \mathbb{R}$ be p times continuously differentiable and suppose that $\nabla_x^p r(x)$ is Lipschitz continuous with Lipschitz constant $L_{r,p}$, Let $T_{r,p}(x,s)$ be the pth degree Taylor approximation of $r(x+s)$ about x given by (2.1). Then for all $x, s \in \mathbb{R}^n$,*

$$\left|r(x+s) - T_{r,p}(x,s)\right| \leq \frac{L_{r,p}}{(p+1)!}\|s\|^{p+1}, \tag{3.8}$$

$$\left\|\nabla_x^j r(x+s) - \nabla_s^j T_{r,p}(x,s)\right\| \leq \frac{L_{r,p}}{(p-j+1)!}\|s\|^{p-j+1} \quad (j = 1,\ldots,p). \tag{3.9}$$

Proof. See [12, LEMMA 2.1] with $\beta = 1$. ∎

We now extend the concepts and notation of Section 2 to the case of composite optimization. Abusing notation slightly, we denote, for $j \in \{1,\ldots,p\}$,

$$T_{w,j}(x,s) \stackrel{\text{def}}{=} T_{f,j}(x,s) + h\big(T_{c,j}(x,s)\big) \tag{3.10}$$

($T_{w,j}(x,s)$ it is *not* a Taylor expansion). We also define, for $j \in \{1,\ldots,q\}$,

$$\begin{aligned}\phi_{w,j}^{\delta}(x) &\stackrel{\text{def}}{=} w(x) - \min_{x+d\in\mathcal{F},\|d\|\leq\delta}\big[T_{f,j}(x,s) + h\big(T_{c,j}(x,s)\big)\big] \\ &= w(x) - \min_{x+d\in\mathcal{F},\|d\|\leq\delta} T_{w,j}(x,s)\end{aligned} \tag{3.11}$$

by analogy with (2.6). This definition allows us to *consider (approximate) high-order minimizers of w, despite h being potentially nonsmooth*, because we have left h unchanged in the optimality measure (3.11), rather than using a Taylor expansion of h.

We now state a simple first-order necessary optimality condition for a global minimizer of composite problems of the form (1.1) with convex h.

Lemma 3.2. *Suppose that f and c are continuously differentiable and that AS.3 holds. Suppose in addition that h is convex and that x_* is a global minimizer of w. Then the origin is a global minimizer of $T_{w,1}(x_*,s)$ and $\phi_{w,1}^{\delta}(x_*) = 0$ for all $\delta > 0$.*

Proof. Suppose now that the origin is not a global minimizer of $T_{w,1}(x_*,s)$, but that there exists an $s_1 \neq 0$ with $T_{w,1}(x_*,s_1) < T_{w,1}(x_*,0) = w(x_*)$. By Taylor's theorem, we obtain that, for $\alpha \in [0,1]$,

$$f(x_*+\alpha s_1) = T_{f,1}(x_*,\alpha s_1) + o(\alpha), \quad c(x_*+\alpha s_1) = T_{c,1}(x_*,\alpha s_1) + o(\alpha) \tag{3.12}$$

and, using AS.3 and (3.6),

$$\begin{aligned}h\big(c(x_*+\alpha s_1)\big) &= h\big(T_{c,1}(x_*,\alpha s_1) + o\big(\alpha\|s_1\|\big)\big) \leq h\big(T_{c,1}(x_*,\alpha s_1)\big) + h\big(o(\alpha)\|s_1\|\big) \\ &\leq h\big(T_{c,1}(x_*,\alpha s_1)\big) + o(\alpha)L_{h,0}\|s_1\| = h\big(T_{c,1}(x_*,\alpha s_1)\big) + o(\alpha).\end{aligned} \tag{3.13}$$

Now note that the convexity of h and the linearity of $T_{f,1}(x_*,s)$ and $T_{c,1}(x_*,s)$ imply that $T_{w,1}(x_*,s)$ is convex, and thus that $T_{w,1}(x_*,\alpha s_1) - w(x_*) \leq \alpha[T_{w,1}(x_*,s_1) - w(x_*)]$. Hence, using (3.12) and (3.13), we deduce that

$$\begin{aligned}0 \leq w(x_*+\alpha s_1) - w(x_*) &\leq T_{w,1}(x_*,\alpha s_1) - w(x_*) + o(\alpha) \\ &\leq \alpha\big[T_{w,1}(x_*,s_1) - w(x_*)\big] + o(\alpha),\end{aligned}$$

which is impossible for α sufficiently small, since $T_{w,1}(x_*, s_1) - w(x_*) < 0$ by construction of s_1. As a consequence, the origin must be a global minimizer of the convex $T_{w,1}(x_*, s)$ and therefore $\phi^{\delta}_{w,1}(x_*) = 0$ for all $\delta > 0$. ∎

Unfortunately, this result does not extend to $\phi^{\delta}_{w,q}(x)$ when $q = 2$, as is shown by the following example. Consider the univariate $w(x) = -\frac{2}{5}x + |x - x^2 + 2x^3|$, where h is the (convex) absolute value function satisfying AS.3. Then $x_* = 0$ is a global minimizer of w (plotted as unbroken line in Figure 3.1) and yet

$$T_{w,2}(x_*, s) = T_{f,2}(x_*, s) + \left|T_{c,2}(x_*, s)\right| = -\frac{2}{5}s + \left|s - s^2\right|$$

(plotted as dashed line in the figure) admits a global minimum for $s = 1$ whose value $(-\frac{2}{5})$ is smaller that $w(x_*) = 0$. Thus $\phi^{1}_{w,2}(x_*) > 0$ despite x_* being a global minimizer. But it is clear in the figure that $\phi^{\delta}_{w,2}(x_*) = 0$ for δ smaller than $\frac{1}{2}$.

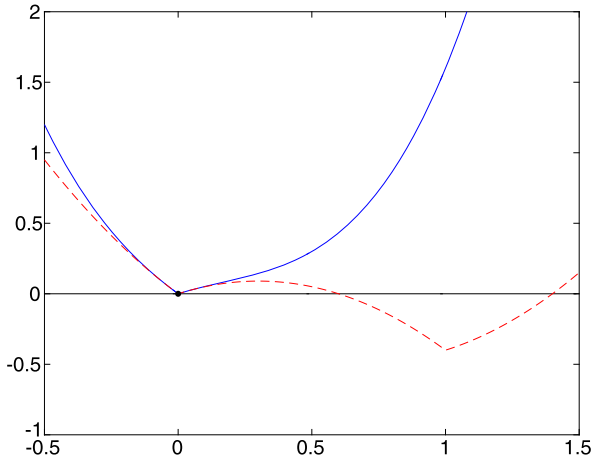


FIGURE 3.1
Functions $w(x)$ (unbroken) and $T_{w,2}(0, s) = T_{f,2}(0, s) + |T_{c,2}(0, s)|$ (dashed).

In the smooth ($h = 0$) case, Lemma 3.2 may be extended for unconstrained (i.e., $\mathcal{F} = \mathbb{R}^n$) twice-continuously differentiable f since then standard second-order optimality conditions at a global minimizer x_* of f imply that $T_{f,j}(x_*, d)$ is convex for $j = 1, 2$ and thus that $\phi^{\delta}_{f,1}(x_*) = \phi^{\delta}_{f,2}(x_*) = 0$. When constraints are present (i.e., $\mathcal{F} \subset \mathbb{R}^n$), unfortunately, this may require that we restrict δ. For example, the global minimizer of $f(x) = -(x - 1/3)^2 + 2/3x^3$ for $x \in [0, 1]$ lies at $x_* = 0$, but $T_{f,2}(x_*, d) = -(d - 1/3)^2$ which has its constrained global minimizer at $d = 1$ with $T_{f,2}(x_*, 1) < T_{f,2}(x_*, 0)$ and we would need $\delta \leq 2/3$ to ensure that $\phi^{\delta}_{f,2}(x_*) = 0$.

4. AN ADAPTIVE REGULARIZATION ALGORITHM FOR COMPOSITE OPTIMIZATION

We now consider an adaptive regularization algorithm to search for a (strong) (ε, δ)-approximate qth-order minimizer for problem (1.1), that is a point $x_k \in \mathcal{F}$ such that

$$\phi^{\delta}_{w,j}(x_k) \leq \varepsilon_j \frac{\delta_j^j}{j!} \quad \text{for } j \in \{1, \ldots, q\}, \tag{4.1}$$

where $\phi^{\delta}_{w,q}(x)$ is defined in (3.11). At each iteration, the algorithm seeks a feasible approximate minimizer of the (possibly nonsmooth) regularized model

$$\begin{aligned} m_k(s) &= T_{f,p}(x_k, s) + h\big(T_{c,p}(x_k, s)\big) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1} \\ &= T_{w,p}(x_k, s) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1} \end{aligned} \tag{4.2}$$

and this process is allowed to terminate whenever

$$m_k(s) \leq m_k(0) \tag{4.3}$$

and, for each $j \in \{1, \ldots, q\}$,

$$\phi^{\delta_{s,j}}_{m_k,j}(s) \leq \theta \varepsilon_j \frac{\delta_{s,j}^j}{j!} \tag{4.4}$$

for some $\theta \in (0, 1)$. Observe that $m_k(s)$ is bounded below since (3.6) ensures that the regularization term of degree $p + 1$ dominates for large steps. Obviously, the inclusion of h in the definition of the model (4.2) implicitly assumes that, as is common, the cost of evaluating h is small compared with that of evaluating f or c. It also implies that computing $\phi^{\delta_j}_{w,j}(x)$ and $\phi^{\delta_{s,j}}_{m_k,j}(s)$ is potentially more complicated than in the smooth case, although it does not impact the evaluation complexity of the algorithm because the model's approximate minimization does not involve evaluating f, c or any of their derivatives.

The rest of the algorithm, that we shall refer to as ARqpC, follows the standard pattern of adaptive regularization algorithms, and is stated on this page. As everywhere in this paper, we assume that $q \in \{1, \ldots, p\}$.

ARqpC algorithm for finding an (ε, δ)-approximate qth-order minimizer of the composite function w in (1.1)

Step 0: Initialization. An initial point x_0 and an initial regularization parameter $\sigma_0 > 0$ are given, as well as an accuracy level $\varepsilon \in (0, 1)^q$. The constants δ_0, θ, η_1, η_2, γ_1, γ_2, γ_3, and $\sigma_{\min}$ are also given and satisfy

$$\begin{aligned} &\theta \in (0, 1), \quad \delta_0 \in (0, 1]^q, \quad \sigma_{\min} \in (0, \sigma_0], \quad 0 < \eta_1 \leq \eta_2 < 1, \\ &\text{and} \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \end{aligned} \tag{4.5}$$

Compute $w(x_0)$ and set $k = 0$.

Step 1: Test for termination. Evaluate $\{\nabla^i_x f(x_k)\}_{i=1}^q$ and $\{\nabla^i_x c(x_k)\}_{i=1}^q$. If (4.1) holds with $\delta = \delta_k$, terminate with the approximate solution $x_\varepsilon = x_k$. Otherwise compute $\{\nabla^i_x f(x_k)\}_{i=q+1}^p$ and $\{\nabla^i_x c(x_k)\}_{i=q+1}^p$.

Step 2: Step calculation. Attempt to compute an approximate minimizer s_k of model $m_k(s)$ given in (4.2) such that $x_k + s_k \in \mathcal{F}$ and optimality radii $\delta_{s_k} \in (0, 1]^q$ exist such that (4.3) holds and (4.4) holds for $j \in \{1, \ldots, q\}$ and $s = s_k$. If no such step exists, terminate with the approximate solution $x_\varepsilon = x_k$.

Step 3: Acceptance of the trial point. Compute $w(x_k + s_k)$ and define

$$\rho_k = \frac{w(x_k) - w(x_k + s_k)}{w(x_k) - T_{w,p}(x_k, s)}. \tag{4.6}$$

If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \delta_{s_k}$; otherwise define $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k$.

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1\sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{if } \rho_k < \eta_1. \end{cases} \tag{4.7}$$

Increment k by one and go to Step 1 if $\rho_k \geq \eta_1$, or to Step 2 otherwise. ■

As expected, the ARqpC algorithm shows obvious similarities with that discussed in [12], but differs from it in significant ways. Beyond the fact that it now handles composite objective functions, the main one being that the termination criterion in Step 1 now tests for strong approximate minimizers, rather than weak ones.

As is standard for adaptive regularization algorithms, we say that an iteration is successful when $\rho_k \geq \eta_1$ (and $x_{k+1} = x_k + s_k$) and that it is unsuccessful otherwise. We denote by $\mathcal{S}_k$ the index set of all successful iterations from 0 to k, that is,

$$\mathcal{S}_k = \{j \in \{0, \ldots, k\} \mid \rho_j \geq \eta_1\},$$

and then obtain a well-known result ensuring that successful iterations up to iteration k do not amount to a vanishingly small proportion of these iterations.

Lemma 4.1. *The mechanism of the ARqpC algorithm guarantees that, if*

$$\sigma_k \leq \sigma_{\max}, \tag{4.8}$$

for some $\sigma_{\max} > 0$, *then*

$$k + 1 \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right). \tag{4.9}$$

Proof. See [5, THEOREM 2.4]. ■

We also have the following identity for the norm of the successive derivatives of the regularization term.

Lemma 4.2. *Let* s *be a vector of* $\mathbb{R}^n$. *Then*

$$\left\| \nabla_s^j \left(\|s\|^{p+1}\right) \right\| = \frac{(p+1)!}{(p-j+1)!} \|s\|^{p-j+1} \quad \text{for } j \in \{0, \ldots, p+1\}. \tag{4.10}$$

Moreover, $\nabla_s^j (\|s\|^{p+1})[d]^j$ *is a convex function in* d *for any* d *orthogonal to* s. *It is also convex for any multiple of* s *whenever* j *is even.*

Proof. See [12, LEMMA 2.4] with $\beta = 1$. ∎

As the conditions for accepting a pair (s_k, δ_s) in Step 2 are stronger than previously considered (in particular, they are stronger than those discussed in [12]), we must ensure that such acceptable pairs exist. We start by recalling a result discussed in [12] for the smooth case.

Lemma 4.3. *Suppose that*

$$\begin{array}{llll} (q = 1, & \mathcal{F} \text{ is convex}, & \text{and} & h \text{ is convex}), \text{ or} \\ (q = 2, & \mathcal{F} = \mathbb{R}^n & \text{and} & h = 0). \end{array} \tag{4.11}$$

Suppose in addition that $s_k^* \neq 0$ is a global minimizer of $m_k(s)$ for $x_k + s \in \mathcal{F}$. Then there exist a feasible neighborhood of s_k^* such that (4.3) and (4.4) hold for any s_k in this neighborhood with $\delta_s = 1$.

Proof. Consider first the composite case where $q = 1$. We have that

$$\begin{aligned} T_{m_k,1}(s_k^*, d) = T_{f,p}(x_k, s_k^*) + \nabla_s^1 T_{f,p}(x_k, s_k^*)[d] + h\big(T_{c,p}(x_k, s_k^*) + \nabla_s^1 T_{c,p}(x_k, s_k^*)[d]\big) \\ + \frac{\sigma_k}{(p+1)!}\big(\|s_k^*\|^{p+1} + \nabla_s^1 \|s_k^*\|^{p+1}[d]\big) \end{aligned}$$

is a convex function of d (since h is convex, all terms in the above right-hand side are). Suppose now that it has a feasible global minimizer d_* such that $T_{m_k,1}(s_k^*, d_*) < T_{m_k,1}(s_k^*, 0) = m_k(s_k^*)$. Since $\mathcal{F}$ is convex, $T_{m_k,1}(s_k^*, d') < m_k(s_k^*)$ for all d' in the segment $(0, d_*]$. But (3.5) implies that

$$\begin{aligned} \big\|\nabla_s^\ell T_{f,p}(x_k, s_k^*)\big\| &\leq \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!}\big\|\nabla_x^i f(x_k)\big\|\,\big\|s_k^*\big\|^{i-\ell} \leq \max_{j\in\{2,\ldots,p\}} L_{f,j-1} \sum_{i=\ell}^{p} \frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!}, \\ \big\|\nabla_s^\ell T_{c,p}(x_k, s_k^*)\big\| &\leq \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!}\big\|\nabla_x^i c(x_k)\big\|\,\big\|s_k^*\big\|^{i-\ell} \leq \max_{j\in\{2,\ldots,p\}} L_{c,j-1} \sum_{i=\ell}^{p} \frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!} \end{aligned}$$

and both must be bounded for s_k^* given. Thus, $T_{m_k,1}(s_k^*, d')$ approximates $m_k(s_k^* + d')$ arbitrarily well for small enough $\|d'\|$, and therefore $m_k(s_k^* + d') < m_k(s_k^*)$ for small enough $\|d'\|$, which is impossible since s_k^* is a global minimizer of $m_k(s)$. As a consequence $d = 0$ must be a global minimizer of $T_{m_k,1}(s_k^*, d)$. Thus $\phi^\delta_{m_k,1}(s_k^*) = 0$ for all $\delta > 0$, and in particular for $\delta = 1$, which, by continuity, yields the desired conclusion.

Consider now the case where $q = 2, h = 0$ and $\mathcal{F} = \mathbb{R}^n$. Suppose that $j = 1$ ($j = 2$). Then the jth order Taylor expansion of the model at s_k^* is a linear (positive semidefinite quadratic) polynomial, which is a convex function. As a consequence, we obtain as above that $\phi^\delta_{m_k,j}(s_k^*) = 0$ for all $\delta_{s,j} > 0$ and the conclusion then again follows. ∎

Alas, the example given at the end of Section 3 implies that δ_s may have to be chosen smaller than one for $q = 2$ and when h is nonzero, even if it is convex. Fortunately, the existence of a step is still guaranteed in general, even without assuming convexity of h. To state our result, we first define ξ to be an arbitrary constant in $(0, 1)$ independent of ε, which we specify later.

Lemma 4.4. *Let $\xi \in (0,1)$ and suppose that s_k^* is a global minimizer of $m_k(s)$ for $x_k + s \in \mathcal{F}$ such that $m_k(s_k^*) < m_k(0)$. Then there exists a pair $(\bar{s}, \delta_s)$ such that (4.3) and (4.4) hold. Moreover, one has that either $\|\bar{s}\| \geq \xi$ or (4.3) and (4.4) hold for $\bar{s}$ for all $\delta_{s,j}$ ($j \in \{1, \ldots, q\}$), for which*

$$0 < \delta_{s,j} \leq \frac{\theta}{q!(6L_w + 2\sigma_k)} \varepsilon_j. \tag{4.12}$$

Proof. We first need to show that a pair $(\bar{s}, \delta_s)$ satisfying (4.3) and (4.4) exists. Since $m_k(s_k^*) < m_k(0)$, we have that $s_k^* \neq 0$. By Taylor's theorem, we have that, for all d,

$$\begin{aligned} 0 &\leq m_k(s_k^* + d) - m_k(s_k^*) \\ &= \sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\left(\sum_{\ell=0}^{p} \frac{1}{\ell!} \nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell \right) - h\big(T_{c,p}(x_k, s_k^*)\big) \\ &\quad + \frac{\sigma_k}{(p+1)!} \left[\sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_s^\ell \big(\|s_k^*\|^{p+1}\big)[d]^\ell + \frac{1}{(p+1)!} \nabla_s^{p+1} \big(\|s_k^* + \tau d\|^{p+1}\big)[d]^{p+1} \right] \end{aligned} \tag{4.13}$$

for some $\tau \in (0,1)$. Using (4.10) in (4.13) and the subadditivity of h ensured by AS.3 then yields that, for any $j \in \{1, \ldots, q\}$ and all d,

$$\begin{aligned} &- \sum_{\ell=1}^{j} \frac{1}{\ell!} \nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\big(T_{c,p}(x_k, s_k^*)\big) - h\left(\sum_{\ell=0}^{j} \frac{1}{\ell!} \nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell \right) \\ &\quad - \frac{\sigma_k}{(p+1)!} \sum_{\ell=1}^{j} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell \\ &\leq \sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\left(\sum_{\ell=j+1}^{q} \frac{1}{\ell!} \nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell \right) \\ &\quad + \frac{\sigma_k}{(p+1)!} \left[\sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell + \|d\|^{p+1} \right]. \end{aligned} \tag{4.14}$$

Since $s_k^* \neq 0$, and using (3.6), we may then choose $\delta_{s,j} \in (0,1]$ such that, for every d with $\|d\| \leq \delta_{s,j}$, we have

$$\begin{aligned} &\sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\left(\sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell \right) \\ &\quad + \frac{\sigma_k}{(p+1)!} \left[\sum_{\ell=j+1}^{p} \frac{1}{\ell!} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell + \|d\|^{p+1} \right] \leq \frac{1}{2} \theta \varepsilon_j \frac{\delta_{s,j}^j}{j!}. \end{aligned} \tag{4.15}$$

As a consequence, we obtain that if $\delta_{s,j}$ is small enough to ensure (4.15), then (4.14) implies

$$\begin{aligned} &- \sum_{\ell=1}^{j} \frac{1}{\ell!} \nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\big(T_{c,p}(x_k, s_k^*)\big) - h\left(\sum_{\ell=0}^{j} \frac{1}{\ell!} \nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell \right) \\ &\quad - \frac{\sigma_k}{(p+1)!} \sum_{\ell=1}^{j} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell \leq \frac{1}{2} \theta \varepsilon_j \frac{\delta_{s,j}^j}{j!}. \end{aligned} \tag{4.16}$$

The fact that, by definition,

$$\phi_{m_k,j}^{\delta_{s,j}}(s) = \max\left[0, \max_{\|d\|\le\delta_{s,j}} \left\{ -\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s)[d]^\ell + h\big(T_{c,p}(x_k,s_k)\big) \right.\right.$$
$$\left.\left. - h\left(\sum_{\ell=0}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s)[d]^\ell\right) - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell \|s\|^{p+1}[d]^\ell \right\}\right], \tag{4.17}$$

continuity of $T_{f,p}(x_k, s)$ and $T_{c,p}(x_k, s)$ and their derivatives and the inequality $m_k(s_k^*) < m_k(0)$ then ensure the existence of a feasible neighborhood of $s_k^* \neq 0$ in which $\bar{s}$ can be chosen such that (4.3) and (4.4) hold for $s = \bar{s}$, concluding the first part of the proof.

To prove the second part, assume first that $\|s_k^*\| \ge 1$. We may then restrict the neighborhood of s_k^* in which $\bar{s}$ can be chosen enough to ensure that $\|\bar{s}\| \ge \xi$. Assume therefore that $\|s_k^*\| \le 1$. Remembering that, by definition and the triangle inequality,

$$\left\|\nabla_s^\ell T_{f,p}(x_k,s_k^*)\right\| \le \sum_{j=\ell}^{p}\frac{1}{(j-\ell)!}\left\|\nabla_x^j f(x_k)\right\|\left\|s_k^*\right\|^{j-\ell},$$
$$\left\|\nabla_s^\ell T_{c,p}(x_k,s_k^*)\right\| \le \sum_{j=\ell}^{p}\frac{1}{(j-\ell)!}\left\|\nabla_x^j c(x_k)\right\|\left\|s_k^*\right\|^{j-\ell},$$

for $\ell \in \{q+1, \ldots, p\}$, and thus, using (3.6), (3.5), and (4.10), we deduce that

$$\begin{aligned}
&\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k^*)[d]^\ell + h\left(\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k^*)[d]^\ell\right)\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p}\nabla_s^\ell\left\|s_k^*\right\|^{p+1}[d]^\ell\right]\\
&\le \sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k^*)[d]^\ell + L_{h,0}\left\|\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k^*)[d]^\ell\right\|\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p}\nabla_s^\ell\left\|s_k^*\right\|^{p+1}[d]^\ell\right]\\
&\le \sum_{\ell=j+1}^{p}\frac{\|d\|^\ell}{\ell!}\left[\sum_{i=\ell}^{p}\frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!}\left(\left\|\nabla_x^i f(x_k)\right\| + L_{h,0}\left\|\nabla_x^i c(x_k)\right\|\right) + \frac{\sigma_k\|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!}\right]\\
&\le \sum_{\ell=j+1}^{p}\frac{\|d\|^\ell}{\ell!}\left[L_w\sum_{i=\ell}^{p}\frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!} + \frac{\sigma_k\|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!}\right],
\end{aligned}$$

where L_w is defined in (3.7). We therefore obtain from (4.15) that any pair $(s_k^*, \delta_{s,j})$ satisfies (4.16) for $\|d\| \le \delta_{s,j}$ if

$$\sum_{\ell=j+1}^{p}\frac{\delta_{s,j}^\ell}{\ell!}\left[L_w\sum_{i=\ell}^{p}\frac{1}{(i-\ell)!}\left\|s_k^*\right\|^{i-\ell} + \frac{\sigma_k\|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!}\right] + \sigma_k\frac{\delta_{s,j}^{p+1}}{(p+1)!} \le \frac{1}{2}\theta\varepsilon_j\frac{\delta_{s,j}^j}{j!}. \tag{4.18}$$

which, because $\|s_k^*\| \leq 1$, is in turn ensured by the inequality

$$\sum_{\ell=j+1}^{p} \frac{\delta_{s,j}^{\ell}}{\ell!}\left[L_w \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} + \sigma_k\right] + \sigma_k \frac{\delta_{s,j}^{p+1}}{(p+1)!} \leq \frac{1}{2}\theta\varepsilon_j \frac{\delta_{s,j}^{j}}{j!}. \tag{4.19}$$

Observe now that, since $\delta_{s,j} \in [0,1]$, we have $\delta_{s,j}^{\ell} \leq \delta_{s,j}^{j+1}$ for $\ell \in \{j+1,\ldots,p\}$. Moreover, we have that,

$$\sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} \leq e < 3 \quad \big(\ell \in \{j+1,\ldots,p+1\}\big), \qquad \sum_{\ell=j+1}^{p+1} \frac{1}{\ell!} \leq e-1 < 2,$$

and therefore (4.19) is (safely) guaranteed by the condition

$$j!(6L_w + 2\sigma_k)\,\delta_{s,j} \leq \frac{1}{2}\theta\varepsilon_j, \tag{4.20}$$

which means that the pair (s_k^*, δ_s) satisfies (4.16) for all $j \in \{1,\ldots,q\}$ whenever

$$\delta_{s,j} \leq \frac{\frac{1}{2}\theta\varepsilon_j}{q!(6L_w + 2\sigma_k)} \stackrel{\text{def}}{=} \frac{1}{2}\delta_{\min,k}.$$

We may thus again invoke the continuity of the derivatives of m_k and (4.17) to deduce that there exists a neighborhood of s_k^* such that, for every $\overline{s}$ in this neighborhood, $m_k(\overline{s}) < m_k(0)$ and the pair $(\overline{s}, \delta_{\min,k})$ satisfies $\phi_{m_k,j}^{\delta_{\min,k}}(\overline{s}) \leq \theta\varepsilon_j \frac{\delta_{\min,k}^{j}}{j!}$, yielding the desired conclusion. ■

This lemma indicates that either the norm of the step is larger than ξ, or the range of acceptable $\delta_{s,j}$ is not too small in that any positive value at most equal to the right-hand side of (4.12) can be chosen. Thus any value larger than a fixed fraction (a half, say) of (4.12) is also acceptable. Such a value is, for instance, guaranteed if $\delta_{s,j}$ is chosen according to the technique described as Algorithm 4.1.

A detailed Step 2 for the ARqpC algorithm (Algorithm 4.1)

Step 2: Step calculation.

Step 2.1: Compute a descent step s_k such that

$$m_k(s_k) < m_k(0)$$

and either $\|s_k\| \geq 1$ or s_k is the global minimizer of $m_k(s)$ for $\|s\| \leq 1$. If no such step exists, terminate the ARqpC algorithm with the approximate solution $x_\varepsilon = x_k$.

Step 2.2: For $j \in \{1,\ldots,q\}$, set $\delta_{s,j} = 1$.

Step 2.3: If $\|s_k\| > 1$, return the pair (s_k, δ_s).

Step 2.4: For each $j \in \{3,\ldots,q\}$,

(i) compute the global minimum of $T_{m_k,j}(s_k, d)$ over all d such that $\|d\| \leq \delta_{s,j}$;

(ii) if

$$\phi^{\delta_{s,j}}_{m_k,j}(s_k) \leq \theta\varepsilon_j \frac{\delta^j_{s,j}}{j!}$$

consider the next value of j; else set $\delta_{s,j} = \frac{1}{2}\delta_{s,j}$ and return to Step 2.4(ii).

Step 2.5: Return the pair (s_k, δ_s). ■

Lemma 4.4 then ensures that this conceptual algorithm is well-defined (and, in particular, that the loop within Step 2.4 is finite for each j). We therefore assume, without loss of generality, that, if some constant $\sigma_{\max}$ is given such that $\sigma_k \leq \sigma_{\max}$ for all k, then the ARqpC algorithm ensures that

$$\delta_{s,j} \geq \kappa_{\delta,\min}\,\varepsilon_j \quad \text{with} \quad \kappa_{\delta,\min} \stackrel{\text{def}}{=} \frac{\theta}{2q!(6L_w + 2\sigma_{\max})} \in \left(0, \frac{1}{2}\right) \tag{4.21}$$

for $j \in \{1, \ldots, q\}$ whenever $\|s_k\| \leq \xi$.

We also need to establish that the possibility of termination in Step 2 of the ARqpC algorithm is a satisfactory outcome.

Lemma 4.5. *Termination cannot occur in Step 2 of the ARqpC algorithm if $q = 1$ and h is convex. In other cases, if the ARqpC algorithm terminates in Step 2 of iteration k with $x_\varepsilon = x_k$, then there exists a $\delta \in (0, 1]^q$ such that (4.1) holds for $x = x_\varepsilon$ and x_ε is a strong (ε, δ)-approximate qth-order-necessary minimizer.*

Proof. Given Lemma 4.4, if the algorithm terminates within Step 2, it must be because every (feasible) global minimizer s^*_k of $m_k(s)$ is such that $m_k(s^*_k) \geq m_k(0)$. In that case, $s^*_k = 0$ is one such global minimizer. If $q = 1$ and h is convex, Lemma 3.2 ensures that termination must have happened in Step 1, and termination in Step 2 is thus impossible. Otherwise, we have that, for any $j \in \{1, \ldots, q\}$ and all d with $x_k + d \in \mathcal{F}$,

$$\begin{aligned}
0 \leq m_k(d) - m_k(0) &= \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla^\ell_x f(x_k)[d]^\ell + \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla^\ell_x f(x_k)[d]^\ell \\
&\quad + h\left(c(x_k) + \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla^\ell_x c(x_k)[d]^\ell + \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla^\ell_x c(x_k)[d]^\ell\right) \\
&\quad + \frac{\sigma_k}{(p+1)!}\|d\|^{p+1} - h\big(c(x_k)\big) \\
&\leq \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla^\ell_x f(x_k)[d]^\ell + \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla^\ell_x f(x_k)[d]^\ell + h\left(\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla^\ell_x c(x_k)[d]^\ell\right) \\
&\quad + h\left(\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla^\ell_x c(x_k)[d]^\ell\right) + \frac{\sigma_k}{(p+1)!}\|d\|^{p+1},
\end{aligned}$$

where we used the subadditivity of h (ensured by AS.3) to derive the last inequality. Hence

$$-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell - h\left(\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right)$$
$$\leq \sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell + h\left(\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right) + \frac{\sigma_k}{(p+1)!}\|d\|^{p+1}.$$

Using (3.6), we may now choose each $\delta_j \in (0,1]$ for $j \in \{1,\ldots,q\}$ small enough to ensure that the absolute value of the last right-hand side is at most $\varepsilon_j \delta_{k,j}^j/j!$ for all d with $\|d\| \leq \delta_{k,j}$ and $x_k + d \in \mathcal{F}$, which, in view of (3.11), implies (4.1). ■

5. EVALUATION COMPLEXITY

To analyze the evaluation complexity of the ARqpC algorithm, we first derive the predicted decrease in the unregularized model from (4.2).

Lemma 5.1. *At every iteration k of the ARqpC algorithm, one has that*

$$w(x_k) - T_{w,p}(x_k, s_k) \geq \frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1}. \tag{5.1}$$

Proof. Immediate from (4.2) and (3.10), the fact that $m_k(0) = w(x_k)$ and (4.3). ■

We next derive the existence of an upper bound on the regularization parameter for the structured composite problem. The proof of this result hinges on the fact that, once the regularization parameter σ_k exceeds the relevant Lipschitz constant ($L_{w,p}$ here), there is no need to increase it any further because the model then provides an overestimation of the objective function.

Lemma 5.2. *Suppose that AS.1–AS.3 hold. Then, for all $k \geq 0$,*

$$\sigma_k \leq \sigma_{\max} \stackrel{\text{def}}{=} \max\left[\sigma_0, \frac{\gamma_3 L_{w,p}}{1-\eta_2}\right], \tag{5.2}$$

where $L_{w,p} = L_{f,p} + L_{h,0}L_{c,p}$.

Proof. Successively using (4.6), Theorem 3.1 applied to f and c, and (5.1), we deduce that, at iteration k,

$$\begin{aligned}
|\rho_k - 1| &= \left|\frac{w(x_k) - w(x_k + s_k)}{w(x_k) - T_{w,p}(x_k, s)} - 1\right| \\
&= \frac{|f(x_k + s_k) + h(c(x_k + s_k)) - T_{f,p}(x_k, s) - h(T_{c,p}(x_k, s))|}{w(x_k) - T_{w,p}(x_k, s)} \\
&\leq \frac{\frac{L_{f,p}\|s_k\|^{p+1}}{(p+1)!} + L_{h,0}\|c(x_k + s_k) - T_{c,p}(x_k, s)\|}{w(x_k) - T_{w,p}(x_k, s)} \\
&\leq \frac{\frac{L_{f,p}+L_{h,0}L_{c,p}}{(p+1)!}\|s_k\|^{p+1}}{\frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1}} = \frac{L_{f,p} + L_{h,0}L_{c,p}}{\sigma_k}.
\end{aligned}$$

Thus, if $\sigma_k \geq L_{w,p}/(1-\eta_2)$, then iteration k is successful, $x_{k+1} = x_k$, and (4.7) implies that $\sigma_{k+1} \leq \sigma_k$. The conclusion then follows from the mechanism of (4.7). ∎

We now establish an important inequality derived from our smoothness assumptions.

Lemma 5.3. *Suppose that AS.1–AS.3 hold. Suppose also that iteration k is successful and that the ARqpC algorithm does not terminate at iteration $k+1$. Then there exists a $j \in \{1,\dots,q\}$ such that*

$$(1-\theta)\,\varepsilon\,\frac{\delta_{k+1,j}^{j}}{j!} \leq (L_{w,p}+\sigma_{\max})\sum_{\ell=1}^{j}\frac{\delta_{k+1,j}^{\ell}}{\ell!}\|s_k\|^{p-\ell+1} + 2\frac{L_{h,0}L_{c,p}}{(p+1)!}\|s_k\|^{p+1}. \tag{5.3}$$

Proof. If the algorithm does not terminate at iteration $k+1$, there must exist a $j \in \{1,\dots,q\}$ such that (4.1) fails at order j at iteration $k+1$. Consider such a j and let d be the argument of the minimization in the definition of $\phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})$. Then $x_k + d \in \mathcal{F}$ and $\|d\| \leq \delta_{k+1,j} \leq 1$. The definition of $\phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})$ in (3.11) then gives that

$$\begin{aligned}
\varepsilon\frac{\delta_{k+1,j}^{j}}{j!} &< \phi_{w,j}^{\delta_{k+1,j}}(x_{k+1}) \\
&= -\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + h\big(c(x_{k+1})\big) - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_{k+1})[d]^\ell\right) \\
&= -\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + \sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell + h\big(c(x_{k+1})\big) \\
&\quad - h\big(T_{c,p}(x_k,s_k)\big) - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_{k+1})[d]^\ell\right) \\
&\quad + h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right) - \sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell \\
&\quad + h\big(T_{c,p}(x_k,s_k)\big) - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right) \\
&\quad - \sum_{\ell=1}^{j}\frac{\sigma_k\|s_k\|^{p-\ell+1}[d]^\ell}{\ell!(p-\ell+1)!} + \sum_{\ell=1}^{j}\frac{\sigma_k\|s_k\|^{p-\ell+1}[d]^\ell}{\ell!(p-\ell+1)!}.
\end{aligned} \tag{5.4}$$

Now, using Theorem 3.1 for $r = f$ yields

$$\begin{aligned}
&-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + \sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell \\
&\quad\leq \sum_{\ell=1}^{j}\frac{\delta_{k+1,j}^{\ell}}{\ell!}\left\|\nabla_x^\ell f(x_{k+1}) - \nabla_s^\ell T_{f,p}(x_k,s_k)\right\| \\
&\quad\leq L_{f,p}\sum_{\ell=1}^{j}\frac{\delta_{k+1,j}^{\ell}}{\ell!(p-\ell+1)!}\|s_k\|^{p-\ell+1}.
\end{aligned} \tag{5.5}$$

In the same spirit, also using AS.3 and applying Theorem 3.1 to c, we obtain

$$\begin{aligned}
&-h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_{k+1})[d]^\ell\right)+h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right)\\
&\quad\le L_{h,0}\left\|\sum_{\ell=0}^{j}\frac{1}{\ell!}\left[\nabla_x^\ell c(x_{k+1})-\nabla_s^\ell T_{c,p}(x_k,s_k)\right][d]^\ell\right\|\\
&\quad\le L_{h,0}\sum_{\ell=0}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!}\left\|\nabla_x^\ell c(x_{k+1})-\nabla_s^\ell T_{c,p}(x_k,s_k)\right\|\\
&\quad\le L_{h,0}L_{c,p}\sum_{\ell=0}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}\|s_k\|^{p-\ell+1}
\end{aligned}\tag{5.6}$$

and

$$h\big(c(x_{k+1})\big)-h\big(T_{c,p}(x_k,s_k)\big)\le L_{h,0}\left\|c(x_{k+1})-T_{c,p}(x_k,s_k)\right\|\le\frac{L_{h,0}L_{c,p}}{(p+1)!}\|s_k\|^{p+1}.\tag{5.7}$$

Because of Lemma 5.2 we also have that

$$\sum_{\ell=1}^{j}\frac{\sigma_k\|s_k\|^{p-\ell+1}\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}\le\sigma_{\max}\sum_{\ell=1}^{j}\frac{\|s_k\|^{p-\ell+1}\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}.\tag{5.8}$$

Moreover, in view of (4.2) and (4.4),

$$\begin{aligned}
&-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell+h\big(T_{c,p}(x_k,s_k)\big)-h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right)\\
&-\sum_{\ell=1}^{j}\frac{\sigma_k}{\ell!(p-\ell+1)!}\|s_k\|^{p-\ell+1}\delta_{k+1,j}^\ell\le\phi_{m_k,j}^{\delta_{s,j}}(s_k)=\theta\varepsilon\,\frac{\delta_{k+1,j}^j}{j!},
\end{aligned}\tag{5.9}$$

where the last equality is derived using $\delta_{s,j}=\delta_{k+1,j}$ if iteration k is successful. We may now substitute (5.5)–(5.9) into (5.4) and use the inequality $(p-\ell+1)!\ge 1$ to obtain (5.3). ■

Lemma 5.4. *Suppose that AS.1–AS.3 hold, that iteration k is successful, and that the ARqpC algorithm does not terminate at iteration $k+1$. Suppose also that the algorithm ensures, for each k, that either $\delta_{k+1,j}=1$ for $j\in\{1,\dots,q\}$ if (4.11) holds (as allowed by Lemma 4.3), or that (4.21) holds (as allowed by Lemma 4.4) otherwise. Then there exists a $j\in\{1,\dots,q\}$ such that*

$$\|s_k\|\ge\begin{cases}\left(\frac{1-\theta}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p-j+1}}\varepsilon_j^{\frac{1}{p-j+1}} & \text{if (4.11) holds,}\\ \left(\frac{(1-\theta)\kappa_{\delta,\min}^{j-1}}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p}}\varepsilon_j^{\frac{j}{p}} & \text{if (4.11) fails but } h=0,\\ \left(\frac{(1-\theta)\kappa_{\delta,\min}^{j}}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p+1}}\varepsilon_j^{\frac{j+1}{p+1}} & \text{if (4.11) fails and } h\ne 0,\end{cases}\tag{5.10}$$

where $\kappa_{\delta,\min}$ is defined in (4.21).

Proof. We now use our freedom to choose $\xi\in(0,1)$. Let

$$\xi\stackrel{\text{def}}{=}\left(\frac{1-\theta}{3q!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p-q+1}}=\min_{j\in\{1,\dots,q\}}\left(\frac{1-\theta}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p-j+1}}\in(0,1).$$

If $\|s_k\| \geq \xi$, then (5.10) clearly holds since $\varepsilon \leq 1$ and $\kappa_{\delta,\min} < 1$. We therefore assume that $\|s_k\| < \xi$. Because the algorithm has not terminated, Lemma 5.3 ensures that (5.3) holds for some $j \in \{1, \ldots, q\}$. It is easy to verify that this inequality is equivalent to

$$\alpha\, \varepsilon\, \delta_{k+1,j}^{j} \leq \|s_k\|^{p+1} \chi_j\left(\frac{\delta_{k+1,j}}{\|s_k\|}\right) + \beta \|s_k\|^{p+1} \tag{5.11}$$

where the function χ_j is defined in (2.5) and where we have set

$$\alpha = \frac{1-\theta}{j!(L_{w,p} + \sigma_{\max})} \quad \text{and} \quad \beta = \frac{2}{(p+1)!} \frac{L_{h,0} L_{c,p}}{L_{w,p} + \sigma_{\max}} \in [0, 1),$$

the last inclusion resulting from the definition of $L_{w,p}$ in Lemma 5.2. In particular, since $\chi_j(t) \leq 2t^j$ for $t \geq 1$ and $\beta < 1$, we have that, when $\|s_k\| \leq \delta_{k+1,j}$,

$$\alpha\, \varepsilon \leq 2\|s_k\|^{p+1}\left(\frac{1}{\|s_k\|}\right)^j + \left(\frac{\|s_k\|}{\delta_{k+1,j}}\right)^j \|s_k\|^{p-j+1} \leq 3\|s_k\|^{p-j+1}. \tag{5.12}$$

Suppose first that (4.11) hold. Then, from our assumptions, $\delta_{k+1,j} = 1$ and $\|s_k\| \leq \xi < 1 = \delta_{k+1,j}$. Thus (5.12) yields the first case of (5.10). Suppose now that (4.11) fails. Then our assumptions imply that (4.21) holds. If $\|s_k\| \leq \delta_{k+1,j}$, we may again deduce from (5.12) that the first case of (5.10) holds, which implies, because $\kappa_{\delta,\min} < 1$, that the second and third cases also hold. Consider therefore the case where $\|s_k\| > \delta_{k+1,j}$ and suppose first that $\beta = 0$. Then (5.11) and the fact that $\chi_j(t) < 2t$ for $t \in [0, 1]$ give that

$$\alpha\varepsilon\delta_{k+1,j}^{j} \leq 2\|s_k\|^{p+1}\left(\frac{\delta_{k+1,j}}{\|s_k\|}\right),$$

which, with (4.21), implies the second case of (5.10). Finally, if $\beta > 0$, (5.11), the bound $\beta \leq 1$, and $\chi_j(t) < 2$ for $t \in [0, 1]$ ensure that

$$\alpha\varepsilon\delta_{k+1,j}^{j} \leq 2\|s_k\|^{p+1} + \|s_k\|^{p+1},$$

the third case of (5.10) then follows from (4.21). ∎

Note that the proof of this lemma ensures the better lower bound given by the first case of (5.10) whenever $\|s_k\| \leq \delta_{k+1,j}$. Unfortunately, there is no guarantee that this inequality holds when (4.11) fails.

We may then derive our final evaluation complexity results. To make them clearer, we provide separate statements for the standard smooth and for the general composite cases.

Theorem 5.5 (Smooth case). *Suppose that AS.1 and AS.4 hold and that $h = 0$. Suppose also that the algorithm ensures, for each k, that either $\delta_{k+1,j} = 1$ for $j \in \{1, \ldots, q\}$ if (4.11) holds (as allowed by Lemma 4.3), or that (4.21) holds (as allowed by Lemma 4.4) otherwise.*

1. *Suppose that $\mathcal{F}$ is convex and $q = 1$ or that $\mathcal{F} = \mathbb{R}^n$ and $q = 2$. Then there exist positive constants $\kappa_{\text{ARqp}}^{s,1}$, $\kappa_{\text{ARqp}}^{a,1}$, and κ_{ARqp}^{c} such that, for any $\varepsilon \in (0, 1]^q$, the ARqpC algorithm requires at most*

$$\kappa_{\text{ARqp}}^{a,1} \frac{w(x_0) - w_{\text{low}}}{\min_{j\in\{1,\ldots,q\}} \varepsilon_j^{\frac{p+1}{p-j+1}}} + \kappa_{\text{ARqp}}^{c} = \mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}} \varepsilon_j^{-\frac{p+1}{p-j+1}}\Big) \tag{5.13}$$

evaluations of f and c, and at most

$$\kappa_{\text{ARqp}}^{s,1}\frac{w(x_0)-w_{\text{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{\frac{p+1}{p-j+1}}}+1=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-\frac{p+1}{p-j+1}}\Big) \tag{5.14}$$

evaluations of the derivatives of f of orders 1 to p to produce an iterate x_ε such that $\phi_{f,j}^1(x_\varepsilon)\le\varepsilon_j/j!$ for all $j\in\{1,\ldots,q\}$.

2. *Suppose that either $\mathcal{F}\subset\mathbb{R}^n$ and $q=2$, or that $\mathcal{F}$ is nonconvex or that $q>2$. Then there exist positive constants $\kappa_{\text{ARqp}}^{s,2}$, $\kappa_{\text{ARqp}}^{a,2}$, and κ_{ARqp}^{c} such that, for any $\varepsilon\in(0,1]^q$, the ARqpC algorithm requires at most*

$$\kappa_{\text{ARqp}}^{a,2}\frac{w(x_0)-w_{\text{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{\frac{j(p+1)}{p}}}+\kappa_{\text{ARqp}}^{c}=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-\frac{j(p+1)}{p}}\Big) \tag{5.15}$$

evaluations of f and c, and at most

$$\kappa_{\text{ARqp}}^{s,2}\frac{w(x_0)-w_{\text{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{\frac{j(p+1)}{p}}}+1=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-\frac{j(p+1)}{p}}\Big) \tag{5.16}$$

evaluations of the derivatives of f of orders 1 to p to produce an iterate x_ε such that $\phi_{f,j}^{\delta_\varepsilon}(x_\varepsilon)\le\varepsilon_j\,\delta_{\varepsilon,j}^j/j!$ for some $\delta_\varepsilon\in(0,1]^q$ and all $j\in\{1,\ldots,q\}$.

Theorem 5.6 (Composite case). *Suppose that AS.1–AS.4 hold. Suppose also that the algorithm ensures, for each k, that either $\delta_{k+1,j}=1$ for $j\in\{1,\ldots,q\}$ if (4.11) holds (as allowed by Lemma 4.3), or that (4.21) holds (as allowed by Lemma 4.4) otherwise.*

1. *Suppose that $\mathcal{F}$ is convex, $q=1$, and h is convex. Then there exist positive constants $\kappa_{\text{ARqpC}}^{s,1}$, $\kappa_{\text{ARqpC}}^{a,1}$, and $\kappa_{\text{ARqpC}}^{c}$ such that, for any $\varepsilon_1\in(0,1]$, the ARqpC algorithm requires at most*

$$\kappa_{\text{ARqpC}}^{a,1}\frac{w(x_0)-w_{\text{low}}}{\varepsilon_1^{\frac{p+1}{p}}}+\kappa_{\text{ARqpC}}^{c,1}=\mathcal{O}\big(\varepsilon_1^{-\frac{p+1}{p}}\big) \tag{5.17}$$

evaluations of f and c, and at most

$$\kappa_{\text{ARqpC}}^{s,1}\frac{w(x_0)-w_{\text{low}}}{\varepsilon_1^{\frac{p+1}{p}}}+1=\mathcal{O}\big(\varepsilon_1^{-\frac{p+1}{p}}\big) \tag{5.18}$$

evaluations of the derivatives of f and c of orders 1 to p to produce an iterate x_ε such that $\phi_{w,j}^1(x_\varepsilon)\le\varepsilon_1$ for all $j\in\{1,\ldots,q\}$.

2. *Suppose that $\mathcal{F}$ is nonconvex or that h is nonconvex, or that $q>1$. Then there exist positive constants $\kappa_{\text{ARqp}}^{s,2}$, $\kappa_{\text{ARqp}}^{a,2}$, and κ_{ARqp}^{c} such that, for any $\varepsilon\in(0,1]^q$, the ARqpC algorithm requires at most*

$$\kappa_{\text{ARqpC}}^{a,2}\frac{w(x_0)-w_{\text{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{j+1}}+\kappa_{\text{ARqpC}}^{c}=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-(j+1)}\Big) \tag{5.19}$$

evaluations of f and c, and at most

$$\kappa_{\text{ARqpC}}^{s,2}\frac{w(x_0)-w_{\text{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{j+1}}+1=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-(j+1)}\Big) \tag{5.20}$$

evaluations of the derivatives of f and c of orders 1 to p to produce an iterate x_ε such that $\phi^{\delta_\varepsilon}_{w,j}(x_\varepsilon) \le \varepsilon_j\, \delta^j_{\varepsilon,j}/j!$ for some $\delta_\varepsilon \in (0,1]^q$ and all $j \in \{1,\dots,q\}$.

Proof. We prove Theorems 5.5 and 5.6 together. At each successful iteration k of the ARqpC algorithm before termination, we have the guaranteed decrease

$$w(x_k) - w(x_{k+1}) \ge \eta_1 \big(T_{w,p}(x_k,0) - T_{w,p}(x_k,s_k)\big) \ge \frac{\eta_1 \sigma_{\min}}{(p+1)!}\, \|s_k\|^{p+1} \tag{5.21}$$

where we used (5.1) and (4.7). We now wish to substitute the bounds given by Lemma 5.4 in (5.21), and deduce that, for some $j \in \{1,\dots,q\}$,

$$w(x_k) - w(x_{k+1}) \ge \kappa^{-1} \varepsilon_j^\omega \tag{5.22}$$

where the definition of κ and ω depends on q and h. Specifically,

$$\kappa \stackrel{\text{def}}{=} \begin{cases} \kappa^{s,1}_{\text{ARqp}} = \kappa^{s,1}_{\text{ARqpC}} \stackrel{\text{def}}{=} \frac{(p+1)!}{\eta_1 \sigma_{\min}} \big(\frac{1-\theta}{3j!(L_{w,p}+\sigma_{\max})}\big)^{-\frac{p+1}{p-j+1}} \\ \quad \text{if } (q=1, h \text{ and } \mathcal{F} \text{ are convex}), \text{ and} \\ \quad \text{if } (q \in \{1,2\}, \mathcal{F} \text{ is convex and } h = 0), \\ \kappa^{s,2}_{\text{ARqp}} \stackrel{\text{def}}{=} \frac{(p+1)!}{\eta_1 \sigma_{\min}} \big(\frac{(1-\theta)\kappa^{j-1}_{\delta,\min}}{3j!(L_{w,p}+\sigma_{\max})}\big)^{-\frac{p+1}{p}} \\ \quad \text{if } h = 0 \text{ and} \\ \quad ((q=2 \text{ and } \mathcal{F} \subset \mathbb{R}^n) \text{ or } q > 2 \text{ or } \mathcal{F} \text{ is nonconvex}) \\ \kappa^{s,2}_{\text{ARqpC}} \stackrel{\text{def}}{=} \frac{(p+1)!}{\eta_1 \sigma_{\min}} \big(\frac{(1-\theta)\kappa^{j}_{\delta,\min}}{3j!(L_{w,p}+\sigma_{\max})}\big)^{-1} \\ \quad \text{if } h \neq 0 \text{ and } (q > 1 \text{ or } \mathcal{F} \text{ is nonconvex}), \end{cases}$$

where $\kappa_{\delta,\min}$ is given by (4.21), and

$$\omega \stackrel{\text{def}}{=} \begin{cases} \frac{p+1}{p-q+1} & \text{if } (q=1, h \text{ and } \mathcal{F} \text{ are convex}), \text{and} \\ & \text{if } (q=2, \mathcal{F} = \mathbb{R}^n \text{ and } h = 0), \\ \frac{q(p+1)}{p} & \text{if } h = 0 \text{ and} \\ & ((q=2 \text{ and } \mathcal{F} \subset \mathbb{R}^n) \text{ or } q > 2 \text{ or } \mathcal{F} \text{ is nonconvex}) \\ q+1 & \text{if } h \neq 0 \text{ and } (q > 1 \text{ or } \mathcal{F} \text{ is nonconvex}). \end{cases} \tag{5.23}$$

Thus, since $\{w(x_k)\}$ decreases monotonically,

$$w(x_0) - w(x_{k+1}) \ge \kappa^{-1} \min_{j \in \{1,\dots,q\}} \varepsilon_j^\omega\, |\mathcal{S}_k|.$$

Using AS.4, we conclude that

$$|\mathcal{S}_k| \le \kappa\, \frac{w(x_0) - w_{\text{low}}}{\min_{j \in \{1,\dots,q\}} \varepsilon_j^\omega} \tag{5.24}$$

until termination, bounding the number of successful iterations. Lemma 4.1 is then invoked to compute the upper bound on the total number of iterations, yielding the constants

$$\kappa^{a,1}_{\text{ARqp}} \stackrel{\text{def}}{=} \kappa^{s,1}_{\text{ARqp}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right), \quad \kappa^{a,2}_{\text{ARqp}} \stackrel{\text{def}}{=} \kappa^{s,2}_{\text{ARqp}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right),$$
$$\kappa^{a,1}_{\text{ARqpC}} \stackrel{\text{def}}{=} \kappa^{s,1}_{\text{ARqpC}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right), \quad \kappa^{a,2}_{\text{ARqpC}} \stackrel{\text{def}}{=} \kappa^{s,2}_{\text{ARqpC}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right),$$

and

$$\kappa^c_{\text{ARqp}} = \kappa^c_{\text{ARqpC}} \stackrel{\text{def}}{=} \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right),$$

where $\sigma_{\max} = \max[\sigma_0, \frac{\gamma_3 L_{w,p}}{1-\eta_2}]$ (see (5.2)). The desired conclusions then follow from the fact that each iteration involves one evaluation of f and each successful iteration one evaluation of its derivatives. ∎

For the standard smooth case, Theorem 5.5 provides the first results on the complexity of finding *strong* minimizers of arbitrary orders using adaptive regularization algorithms that we are aware of. By comparison, [12] provides similar results but for the convergence to *weak* minimizers (see (2.5)). Unsurprisingly, the worst-case complexity bounds for weak minimizers are better than those for strong ones: the $\mathcal{O}(\varepsilon^{-(p+1)/(p-q+1)})$ bound which we have derived for $q \in \{1,2\}$ then extends to any order q. Moreover, the full power of AS.1 is not needed for these results since it is sufficient to assume that $\nabla^p_x f(x)$ is Lipschitz continuous. It is interesting to note that the results for weak and strong approximate minimizers coincide for first and second order. The results of Theorem 5.5 may also be compared with the bound in $\mathcal{O}(\varepsilon^{-(q+1)})$ which was proved for trust-region methods in [11]. While these trust-region bounds do not depend on the degree of the model, those derived above for the ARqpC algorithm show that worst-case performance improves with p and is always better than that of trust-region methods. It is also interesting to note that the bound obtained in Theorem 5.5 for order q is identical to that which would be obtained for first-order but using ε^q instead of ε. This reflects the observation that, different from the weak approximate optimality, the very definition of strong approximate optimality in (2.4) requires very high accuracy on the (usually dominant) low order terms of the Taylor series while the requirement lessens as the order increases.

An interesting feature of the algorithm discussed in [12] is that computing and testing the value of $\phi^{\delta}_{m_k,j}(s_k)$ is unnecessary if the length of the step is large enough. The same feature can easily be introduced into the ARqpC algorithm. Specifically, we may redefine Step 2 to accept a step as soon as (4.3) holds and

$$\|s_k\| \geq \begin{cases} \varpi \min_{j\in\{1,\ldots,q\}} \varepsilon_j^{\frac{1}{p-q+1}} & \text{if } (q=1, h \text{ and } \mathcal{F} \text{ are convex}), \text{ and} \\ & \text{if } (q=2, \mathcal{F}=\mathbb{R}^n \text{ and } h=0), \\ \varpi \min_{j\in\{1,\ldots,q\}} \varepsilon_j^{\frac{q}{p}} & \text{if } h=0 \text{ and} \\ & ((q=2 \text{ and } \mathcal{F}\subset\mathbb{R}^n) \text{ or } q>2 \text{ or } \mathcal{F} \text{ is nonconvex}), \\ \varpi \min_{j\in\{1,\ldots,q\}} \varepsilon_j^{\frac{q+1}{p+1}} & \text{if } h\neq 0 \text{ and } (q>1 \text{ or } \mathcal{F} \text{ is nonconvex}), \end{cases}$$

for some $\varpi \in (\theta, 1]$. If these conditions fail, then one still needs to verify the requirements (4.3) and (4.4), as we have done previously. Given Lemma 5.1 and the proof of Theorems 5.5 and 5.6, it is easy to verify that this modification does not affect the conclusions of these complexity theorems, while potentially avoiding significant computations.

Existing complexity results for (possibly nonsmooth) composite problems are few [8, 13, 14, 20]. Theorem 5.6 provides, to the best of our knowledge, the first upper complexity bounds for optimality orders exceeding one, with the exception of [13] (but this paper requires

strong specific assumptions on $\mathcal{F}$). While equivalent to those of Theorem 5.5 for the standard case when $q = 1$, they are not as good and match those obtained for the trust-region methods when $q > 1$. They could be made identical in order of ε_j to those of Theorem 5.5 if one is ready to assume that $L_{h,0}L_{c,p}$ is sufficiently small (for instance, if c is a polynomial of degree less than p). In this case, the constant β in Lemma 5.11 will of the order of $\delta_{k+1,j}/\|s_k\|$, leading to the better bound.

6. SHARPNESS

We now show that the upper complexity bounds in Theorem 5.5 and the first part of Theorem 5.6 are sharp. Since it is sufficient for our purposes, we assume in this section that $\varepsilon_j = \varepsilon$ for all $j \in \{1, \ldots, q\}$.

We first consider a first class of problems, where the choice of $\delta_{k,j} = 1$ is allowed. Since it is proved in [12] that the order in ε given by the Theorem 5.5 is sharp for finding weak approximate minimizers for the standard (smooth) case, it is not surprising that this order is also sharp for the stronger concept of optimality whenever the same bound applies, that is when $q \in \{1, 2\}$. However, the ARqpC algorithm slightly differs from the algorithm discussed in [12]. Not only are the termination tests for the algorithm itself and those for the step computation weaker in [12], but the algorithm there makes a provision to avoid computing $\phi^{\delta}_{m_k,j}$ whenever the step is large enough, as discussed at the end of the last section. It is thus impossible to use the example of slow convergence provided in [12, SECTION 5.2] directly, but we now propose a variant that fits our present framework.

Theorem 6.1. *Suppose that $h = 0$ and that the choice $\delta_{k,j} = 1$ is possible (and made) for all k and all $j \in \{1, \ldots, q\}$. Then the ARqpC algorithm applied to minimize f may require*

$$\varepsilon^{-\frac{p+1}{p-q+1}}$$

iterations and evaluations of f and of its derivatives of order 1 up to p to produce a point x_ε such that $\phi^1_{w,q}(x_\varepsilon) \leq \varepsilon/j!$ for all $j \in \{1, \ldots, q\}$.

Proof. Our aim is to show that, for each choice of $p \geq 1$, there exists an objective function satisfying AS.1 and AS.4 such that obtaining a strong (ε, δ)-approximate qth-order-necessary minimizer may require at least $\varepsilon^{-(p+1)/(p-q+1)}$ evaluations of the objective function and its derivatives using the ARqpC algorithm. Also note that, in this context, $\phi^{\delta_j}_{w,q}(x) = \phi^{\delta_j}_{f,q}(x)$ and (4.1) reduces to (2.4).

Given a model degree $p \geq 1$ and an optimality order q, we also define the sequences $\{f_k^{(j)}\}$ for $j \in \{0, \ldots, p\}$ and $k \in \{0, \ldots, k_\varepsilon\}$ by

$$k_\varepsilon = \left\lceil \varepsilon^{-\frac{p+1}{p-q+1}} \right\rceil \tag{6.1}$$

and

$$\omega_k = \varepsilon \frac{k_\varepsilon - k}{k_\varepsilon} \in [0, \varepsilon], \tag{6.2}$$

as well as

$$f_k^{(j)} = 0 \quad \text{for } j \in \{1, \ldots, q-1\} \cup \{q+1, \ldots, p\} \quad \text{and} \quad f_k^{(q)} = -(\varepsilon + \omega_k) < 0.$$

Thus

$$T_{f,p}(x_k, s) = \sum_{j=0}^{p} \frac{f_k^{(j)}}{j!} s^j = f_k^{(0)} - (\varepsilon + \omega_k) \frac{s^q}{q!}. \tag{6.3}$$

We also set $\sigma_k = p!/(q-1)!$ for all $k \in \{0, \ldots, k_\varepsilon\}$ (we verify below that is acceptable). It is easy to verify using (6.3) that the model (4.2) is then globally minimized for

$$s_k = \left| f_k^{(q)} \right|^{\frac{1}{p-q+1}} = [\varepsilon + \omega_k]^{\frac{1}{p-q+1}} > \varepsilon^{\frac{1}{p-q+1}} \quad \big(k \in \{0, \ldots, k_\varepsilon\}\big). \tag{6.4}$$

We then assume that Step 2 of the ARqpC algorithm returns, for all $k \in \{0, \ldots, k_\varepsilon\}$, the step s_k given by (6.4) and the optimality radius $\delta_{k,j} = 1$ for $j \in \{1, \ldots, q\}$ (as allowed by our assumption). Thus implies that

$$\phi_{f,q}^{\delta_{k,q}}(x_k) = (\varepsilon + \omega_k) \frac{\delta_{k,q}^q}{q!}, \tag{6.5}$$

and therefore that

$$\omega_k \in (0, \varepsilon], \quad \phi_{f,j}^{\delta_{k,j}}(x_k) = 0 \quad (j = 1, \ldots, q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_k) > \varepsilon \frac{\delta_{k,q}^q}{q!} \tag{6.6}$$

(and (2.4) fails at x_k) for $k \in \{0, \ldots, k_\varepsilon - 1\}$, while

$$\omega_{k_\varepsilon} = 0, \quad \phi_{f,j}^{\delta_{k,j}}(x_{k_\varepsilon}) = 0 \quad (j = 1, \ldots, q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_{k_\varepsilon}) = \varepsilon \frac{\delta_{k,q}^q}{q!} \tag{6.7}$$

(and (2.4) holds at x_{k_ε}). The step (6.4) yields that

$$\begin{aligned} m_k(s_k) &= f_k^{(0)} - \frac{\varepsilon + \omega_k}{q!} [\varepsilon + \omega_k]^{\frac{q}{p-q+1}} + \frac{\sigma_k}{(p+1)!} [\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}} \\ &= f_k^{(0)} - \frac{\varepsilon + \omega_k}{q!} [\varepsilon + \omega_k]^{\frac{q}{p-q+1}} + \frac{1}{(p+1)(q-1)!} [\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}} \\ &= f_k^{(0)} - \zeta(q, p) [\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}} \end{aligned} \tag{6.8}$$

where

$$\zeta(q, p) \stackrel{\text{def}}{=} \frac{p-q+1}{(p+1)q!} \in (0, 1). \tag{6.9}$$

Thus $m_k(s_k) < m_k(0)$ and (4.3) holds. We then define

$$f_0^{(0)} = 2^{1+\frac{p+1}{p-q+1}} \quad \text{and} \quad f_{k+1}^{(0)} = f_k^{(0)} - \zeta(q, p)[\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}}, \tag{6.10}$$

which provides the identity

$$m_k(s_k) = f_{k+1}^{(0)} \tag{6.11}$$

(ensuring that iteration k is successful because $\rho_k = 1$ in (4.6) and thus that our choice of a constant σ_k is acceptable). In addition, using (6.2), (6.10), (6.6), (6.9) and the inequality $k_\varepsilon \leq 1 + \varepsilon^{-\frac{p+1}{p-q+1}}$ resulting from (6.1), gives that, for $k \in \{0, \ldots, k_\varepsilon\}$,

$$\begin{aligned} f_0^{(0)} \geq f_k^{(0)} &\geq f_0^{(0)} - k\zeta(q, p)[2\varepsilon]^{\frac{p+1}{p-q+1}} > f_0^{(0)} - k_\varepsilon \varepsilon^{\frac{p+1}{p-q+1}} 2^{\frac{p+1}{p-q+1}} \\ &\geq f_0^{(0)} - \big(1 + \varepsilon^{\frac{p+1}{p-q+1}}\big) 2^{\frac{p+1}{p-q+1}} \geq f_0^{(0)} - 2^{1+\frac{p+1}{p-q+1}}, \end{aligned}$$

and hence that

$$f_k^{(0)} \in (0, 2^{1+\frac{p+1}{p-q+1}}] \quad \text{for } k \in \{0, \ldots, k_\varepsilon\}. \tag{6.12}$$

We also set

$$x_0 = 0 \quad \text{and} \quad x_k = \sum_{i=0}^{k-1} s_i.$$

Then (6.11) and (4.2) give that

$$\left| f_{k+1}^{(0)} - T_{f,p}(x_k, s_k) \right| = \frac{1}{(p+1)(q-1)!} |s_k|^{p+1} \leq |s_k|^{p+1}. \tag{6.13}$$

Now note that, using (6.3) and the first equality in (6.4),

$$T_{f,p}^{(j)}(x_k, s_k) = \frac{f_k^{(q)}}{(q-j)!} s_k^{q-j} \, \delta_{[j \leq q]} = -\frac{1}{(q-j)!} s_k^{p-j+1} \, \delta_{[j \leq q]},$$

where $\delta_{[\cdot]}$ is the standard indicator function. We now see that, for $j \in \{1, \ldots, q-1\}$,

$$\left| f_{k+1}^{(j)} - T_{f,p}^{(j)}(x_k, s_k) \right| = \left| 0 - T_{f,p}^{(j)}(x_k, s_k) \right| \leq \frac{1}{(q-j)!} |s_k|^{p-j+1} \leq |s_k|^{p-j+1}, \tag{6.14}$$

while, for $j = q$, we have that

$$\left| f_{k+1}^{(q)} - T_{f,p}^{(q)}(x_k, s_k) \right| = \left| -s_k^{p-q+1} + s_k^{p-q+1} \right| = 0 \tag{6.15}$$

and, for $j \in \{q+1, \ldots, p\}$,

$$\left| f_{k+1}^{(j)} - T_{f,p}^{(j)}(x_k, s_k) \right| = |0 - 0| = 0. \tag{6.16}$$

Combining (6.13)–(6.16), we may then apply classical Hermite interpolation (see **[12, THEOREM 5.2]** with $\kappa_f = 1$), and deduce the existence of a p times continuously differentiable function f_{ARqpC} from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz continuous derivatives of order 0 to p (hence satisfying AS.1) which interpolates $\{f_k^{(j)}\}$ at $\{x_k\}$ for $k \in \{0, \ldots, k_\varepsilon\}$ and $j \in \{0, \ldots, p\}$. Moreover, (6.12), (6.3), (6.4), and the same Hermite interpolation theorem imply that $|f^{(j)}(x)|$ is bounded by a constant only depending on p and q, for all $x \in \mathbb{R}$ and $j \in \{0, \ldots, p\}$ (and thus AS.1 holds) and that f_{ARqpC} is bounded below (ensuring AS.4.) and that its range only depends on p and q. This concludes our proof. ∎

This immediately provides the following important corollary.

Corollary 6.2. *Suppose that $h = 0$ and that either $q = 1$ and $\mathcal{F}$ is convex, or $q = 2$ and $\mathcal{F} = \mathbb{R}^n$. Then the ARqpC algorithm applied to minimize f may require*

$$\varepsilon^{-\frac{p+1}{p-q+1}}$$

iterations and evaluations of f and of its derivatives of order 1 up to p to produce a point x_ε such that $\phi_{w,q}^1(x_\varepsilon) \leq \varepsilon / j!$ for all $j \in \{1, \ldots, q\}$.

Proof. We start by noting that, in both cases covered by our assumptions, Lemma 4.3 allows the choice $\delta_{k,j} = 1$ for all k and all $j \in \{1, \ldots, q\}$. We conclude by applying Theorem 6.1. ∎

It is then possible to derive a lower complexity bound for the simple composite case where h is nonzero but convex and $q = 1$.

Corollary 6.3. *Suppose that $q = 1$ and that h is convex. Then the ARqpC algorithm applied to minimize w may require*

$$\varepsilon^{-\frac{p+1}{p}}$$

iterations and evaluations of f and c and of their derivatives of order 1 up to p to produce a point x_ε such that $\phi^1_{w,1}(x_\varepsilon) \leq \varepsilon$.

Proof. It is enough to consider the unconstrained problem where $w = h(c(x))$ with $h(x) = |x|$ and c is the positive function f constructed in the proof of Theorem 6.1. ∎

We now turn to the high-order smooth case.

Theorem 6.4. *Suppose that $h = 0$ and that either $q > 2$, or $q = 2$ and $\mathcal{F} = \mathbb{R}^n$. If $\varepsilon \in (0, 1)$ is sufficiently small and if the ARqpC algorithm applied to minimize f allows the choice of an arbitrary $\delta_{k,j} > 0$ satisfying (4.21), it may then require*

$$\varepsilon^{-\frac{q(p+1)}{p}}$$

iterations and evaluations of f and of its derivatives of order 1 up to p to produce a point x_ε such that $\phi^{\delta_{\varepsilon,j}}_{f,j}(x_\varepsilon) \leq \varepsilon \delta^j_{\varepsilon,j}/j!$ for all $j \in \{1, \ldots, q\}$ and some $\delta_\varepsilon \in (0, 1]^q$.

Proof. As this is sufficient, we focus on the case where $\mathcal{F} = \mathbb{R}^n$. Our aim is now to show that, for each choice of $p \geq 1$ and $q > 2$, there exists an objective function satisfying AS.1 and AS.4 such that obtaining a strong (ε, δ)-approximate qth-order-necessary minimizer may require at least $\varepsilon^{-q(p+1)/p}$ evaluations of the objective function and its derivatives using the ARqpC algorithm. As in Theorem 6.1, we have to construct f such that it satisfies AS.1 and is globally bounded below, which then ensures AS.4. Again, we note that, in this context, $\phi^{\delta_j}_{f,q}(x) = \phi^{\delta_j}_{f,q}(x)$ and (4.1) reduces to (2.4).

Without loss of generality, we assume that $\varepsilon \leq \frac{1}{2}$. Given a model degree $p \geq 1$ and an optimality order $q > 2$, we set

$$k_\varepsilon = \left\lceil \varepsilon^{-\frac{q(p+1)}{p}} \right\rceil \tag{6.17}$$

and

$$\omega_k = \varepsilon^q \frac{k_\varepsilon - k}{k_\varepsilon} \in \left[0, \varepsilon^q\right] \quad \left(k \in \{0, \ldots, k_\varepsilon\}\right). \tag{6.18}$$

Moreover, for $j \in \{0, \ldots, p\}$ and each $k \in \{0, \ldots, k_\varepsilon\}$, we define the sequences $\{f^{(j)}_k\}$ by

$$f^{(1)}_k = -\frac{\varepsilon^q + \omega_k}{q!} < 0 \quad \text{and} \quad f^{(j)}_k = 0 \quad \text{for } j \in \{2, \ldots, p\}, \tag{6.19}$$

and therefore

$$T_{f,p}(x_k, s) = \sum_{j=0}^{p} \frac{f^{(j)}_k}{j!} s^j = f^{(0)}_k - \frac{\varepsilon^q + \omega_k}{q!} s. \tag{6.20}$$

This definition and the choice $\sigma_k = p!$ ($k \in \{0, \ldots, k_\varepsilon\}$) (we verify below that this is acceptable) then allow us to define the model (4.2) by

$$m_k(s) = f_k^{(0)} - \frac{\varepsilon^q + \omega_k}{q!} s + \frac{|s|^{p+1}}{p+1}. \tag{6.21}$$

We now assume that, for each k, Step 2 returns the model's global minimizer

$$s_k = \left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{1}{p}} \quad \big(k \in \{0, \ldots, k_\varepsilon\}\big) \tag{6.22}$$

and the optimality radius

$$\delta_{k,j} = \varepsilon \quad \big(j \in \{1, \ldots, q\}\big). \tag{6.23}$$

Indeed, a simple calculation shows that we may choose $\delta_{k,j}$ at least as large as

$$\delta_{k,j} = \frac{3|s_k|}{p-1} = \frac{3}{p-1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{1}{p}}. \tag{6.24}$$

which is clearly the case for (6.23) under our assumption on ε. Let us show that the above choice (6.24) is correct. Consider the model (6.21) and let $\beta = (\varepsilon^q + \omega_k)/q!$. We may then compute $T_{m_k,j}(s_k, \alpha s_k)$ the jth degree Taylor expansion of this model at s_k for $j \in \{1, \ldots, q\}$. Since $\nabla_s^1 m_k(s_k) = 0$, we obtain from Lemma 4.2 that

$$\begin{aligned} T_{m_k,j}(s_k, \alpha s_k) = \sum_{\ell=0}^{j} \frac{\nabla_s^\ell m_k(s_k)[\alpha s_k]^\ell}{\ell!} &= m_k(s_k) + \frac{\sigma_k}{(p+1)!} \sum_{\ell=2}^{j} \frac{\nabla_s^\ell (|s_k^*|^{p+1})[\alpha s_k]^\ell}{\ell!} \\ &= m_k(s_k) + \frac{\sigma_k}{(p+1)!} \sum_{\ell=2}^{j} \frac{\alpha^\ell \nabla_s^\ell (|s_k|^{p+1})[s_k]^\ell}{\ell!} \\ &= m_k(s_k) + \sigma_k \sum_{\ell=2}^{j} \frac{\alpha^\ell |s_k|^{p+1}}{\ell!(p+1-\ell)!}. \end{aligned}$$

Clearly, $T_{m_k,1}(s_k, \alpha s_k) = m_k(s_k)$ for all α because the standard first-order optimality condition at s_k gives that $\nabla_d^1 T_{m_k,1}(s_k, 0) = 0$. The second-order optimality condition implies that $T_{m_k,2}(s_k, \alpha s_k)$ is convex in α, but, given that α can be negative, approximations of degree larger than 2 are no longer convex for odd values of j. We are now interested in computing an upper bound on $\delta_{s_k,j}$ so that (4.4) holds and for odd j (and thus for all j). Consider the case where $j = 3$: choosing $\beta = 1$ (and thus $s_k^* = e_1$) as above, (4.4) then requires that, for all $|\alpha s_k| \le \delta_{s_k,3}$,

$$T_{m_k,3}(s_k, 0) - T_{m_k,3}(s_k^*, \alpha s_k) < \theta\varepsilon \frac{|\alpha s_k|^3}{6},$$

which is obviously satisfied for any $\delta_{s_k,3}$ smaller or equal to absolute value of the root $\alpha_{*,3}$ of the equation $T_{m_k,3}(s_k, 0) = T_{m_k,3}(s_k^*, \alpha s_k)$. Using the expression of $T_{m_k,3}(s_k^*, \alpha s_k)$ derived above, one verifies that

$$\alpha_{*,3} = -\frac{3p|s_k|^{p+1}}{p(p-1)|s_k|^{p+1}} = -\frac{3}{p-1}.$$

The cases $j = 5, 9, \ldots, p$ are less restrictive because the corresponding roots $\alpha_{*,j}$ are all smaller than $\alpha_{*,3}$. As a consequence, (4.4) holds for $j \in \{1, \ldots, q\}$ and $\delta_{k,j} = \frac{3|s_k|}{p-1}$.

Thus, from (6.24), (6.20) and (6.23),

$$\phi_{f,j}^{\delta_{k,j}}(x_k) = (\varepsilon^q + \omega_k)\frac{\varepsilon}{q!}$$

for $j \in \{1, \ldots, q\}$ and $k \in \{0, \ldots, k_\varepsilon\}$. Using (6.23), (6.17), and the fact that, for $j \in \{1, \ldots, q-1\}$,

$$\frac{\varepsilon^q + \omega_k}{q!} \leq \frac{2\varepsilon^q}{q!} \leq \frac{\varepsilon^j}{j!} = \frac{\delta_{k,j}^j}{j!} \tag{6.25}$$

when $q \geq 2$ and $\varepsilon \leq \frac{1}{2}$, we then obtain that

$$\phi_{f,j}^{\delta_{k,j}}(x_k) \leq \varepsilon\frac{\delta_{k,j}^j}{j!} \quad (j = 1 \ldots, q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_k) > \varepsilon\frac{\delta_{k,q}^q}{q!}$$

(and (2.4) fails at x_k) for $k \in \{0, \ldots, k_\varepsilon - 1\}$, while

$$\phi_{f,j}^{\delta_{k,j}}(x_{k_\varepsilon}) < \varepsilon\frac{\delta_{k,j}^j}{j!} \quad (j = 1 \ldots, q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_{k_\varepsilon}) = \varepsilon\frac{\delta_{k,q}^q}{q!}$$

(and (2.4) holds at x_{k_ε}). Now (6.21) and (6.22) give that

$$\begin{aligned} m_k(s_k) &= f_k^{(0)} - \frac{\varepsilon^q + \omega_k}{q!}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{1}{p}} + \frac{1}{p+1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{p+1}{p}} \\ &= f_k^{(0)} - \frac{p}{p+1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{p+1}{p}}. \end{aligned}$$

Thus $m_k(s_k) < m_k(0)$ and (4.3) holds. We then define

$$f_0^{(0)} = 2^{1+\frac{q(p+1)}{p}} \quad \text{and} \quad f_{k+1}^{(0)} = f_k^{(0)} - \frac{p}{p+1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{p+1}{p}}, \tag{6.26}$$

which provides the identity

$$m_k(s_k) = f_{k+1}^{(0)} \tag{6.27}$$

(ensuring that iteration k is successful because $\rho_k = 1$ in (4.6) and thus that our choice of a constant σ_k is acceptable). In addition, using (6.18), (6.26), and the inequality $k_\varepsilon \leq 1 + \varepsilon^{-q(p+1)/p}$ resulting from (6.17), (6.26) gives that, for $k \in \{0, \ldots, k_\varepsilon\}$,

$$\begin{aligned} f_0^{(0)} \geq f_k^{(0)} &\geq f_0^{(0)} - k[2\varepsilon]^{\frac{q(p+1)}{p}} \geq f_0^{(0)} - k_\varepsilon\varepsilon^{\frac{q(p+1)}{p}}2^{\frac{q(p+1)}{p}} \\ &\geq f_0^{(0)} - \left(1 + \varepsilon^{\frac{q(p+1)}{p}}\right)2^{\frac{q(p+1)}{p}} \geq f_0^{(0)} - 2^{1+\frac{q(p+1)}{p}}, \end{aligned}$$

and hence that

$$f_k^{(0)} \in \left[0, 2^{1+\frac{q(p+1)}{p}}\right] \quad \text{for } k \in \{0, \ldots, k_\varepsilon\}. \tag{6.28}$$

As in Theorem 6.1, we set $x_0 = 0$ and $x_k = \sum_{i=0}^{k-1} s_i$. Then (6.11) and (4.2) give that

$$\left|f_{k+1}^{(0)} - T_{f,p}(x_k, s_k)\right| = \frac{1}{p}|s_k|^{p+1}. \tag{6.29}$$

Using (6.20), we also see that

$$\left|f_{k+1}^{(1)} - T_{f,p}^{(1)}(x_k, s_k)\right| = \left|-\frac{(\varepsilon^q + \omega_{k+1})}{q!} + \frac{(\varepsilon^q + \omega_k)}{q!}\right| \leq |s_k|^p\left[1 - \frac{\varepsilon^q + \omega_{k+1}}{\varepsilon^q + \omega_k}\right] < |s_k|^p, \tag{6.30}$$

while, for $j \in \{2, \ldots, p\}$,

$$\left| f_{k+1}^{(j)} - T_{f,p}^{(j)}(x_k, s_k) \right| = |0 - 0| < |s_k|^{p-j+1}. \tag{6.31}$$

The proof is concluded as in Theorem 6.1. Combining (6.29)–(6.31), we may then apply classical Hermite interpolation (see [12, THEOREM 5.2] with $\kappa_f = 1$) and deduce the existence of a p-times continuously differentiable function f_{ARqpC} from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz continuous derivatives of order 0 to p (hence satisfying AS.1) which interpolates $\{f_k^{(j)}\}$ at $\{x_k\}$ for $k \in \{0, \ldots, k_\varepsilon\}$ and $j \in \{0, \ldots, p\}$. Moreover, the Hermite theorem, (6.19), and (6.22) also guarantee that $|f^{(j)}(x)|$ is bounded by a constant only depending on p and q, for all $x \in \mathbb{R}$ and $j \in \{0, \ldots, p\}$. As a consequence, AS.1, AS.2, and AS.4 hold. This concludes the proof. ■

Whether the bound (5.20) is sharp remains an open question at present.

7. INEXACT GLOBAL MINIMIZATION

We finally discuss the necessity of performing global minimization when calculating the (objective and model) optimality measures and, when relevant, the effect of performing such computations inexactly. We start by recalling that such minimization problems potentially occur in two parts of the algorithm: in Step 1 (for deciding termination) and in Step 2 (during the step computation).

Step computation. Consider the step computation first and remember that the ultimate purpose of Step 2 is to find a step s_k guaranteeing a sufficient decrease of the Taylor series at x_k, in that

$$T_{w,j}(x_k, 0) - T_{w,j}(x_k, s_k) \geq \kappa_{\text{decr}} \varepsilon_j^{\omega} \tag{7.1}$$

for some fixed $\kappa_{\text{decr}} > 0$ and $j \in \{1, \ldots, q\}$, where ω is defined in (5.23) (this argument is used in the proof of Theorems 5.5 and 5.6). Of course, if a step s_k that satisfies (7.1) for some given κ_{decr} can be found simply,[4] without resorting to global optimization, so much the better (and we may then choose $\delta_{k,j} = 1$ for $j \in \{1, \ldots, q\}$). In other cases, the decrease guarantee (7.1) is obtained in one of two possible ways: if $\|s_k\| \geq 1$ and given that $\varepsilon_j \in (0, 1]$, sufficient decrease follows from Lemma 5.1 with $\kappa_{\text{decr}} = \sigma_{\min}/(p+1)!$. Alternatively, if $\|s_k\| \leq 1$, we then have to enforce (4.4) for some $\delta_{s,j} \in (0, 1]$, and use the more complicated Lemma 5.4 to reach the desired conclusion. In our development, the constant 1 in the inequality $\|s_k\| \geq 1$ was chosen solely for simplicity of exposition, but can be replaced by any constant independent of k. In particular, it can be replaced by $\kappa_{\text{decr}}^{1/(p+1)} \varepsilon_{\min}^{\omega/(p+1)}$ where $\varepsilon_{\min} = \min_{j \in \{1,\ldots,q\}} \varepsilon_j$, so that sufficient decrease still immediately follows from Lemma 5.1 if

$$\|s_k\| \geq \kappa_{\text{decr}}^{1/(p+1)} \varepsilon_{\min}^{\omega/(p+1)}. \tag{7.2}$$

As a consequence, we see that performing any global optimization in Step 2 is only necessary whenever a descent step cannot be found that satisfies either (7.1) or (7.2). From a practical point of view, the failure of these two conditions could be considered as a reasonable ter-

4 Say, by applying some trusted local minimization method.

mination rule for small enough $\varepsilon_{\min}$, even if there is then no guarantee that the iterate x_k at which the algorithm appears to be stuck is an approximate minimizer.

If one now insists on true optimality, the details of Algorithm 4.1 become relevant. In this algorithm, the sole purpose of the global minimization in Step 2.1 is to ensure that Lemma 4.4 can be applied to guarantee finite termination of the loop within Step 2.2. Thus, if Step 2.1 cannot be performed exactly, it may happen that this loop does not terminate (even assuming feasibility of the additional global minimizations within the loop). A practical algorithm would terminate this loop if $\delta_{s,j}$ becomes too small or if a maximum number of inner iterations have been taken, returning a value of $\delta_{s,j}$ which is potentially too large for the computed step (compared to what would have resulted if global minimization had always been successful). This is also the outcome of Step 2.2 if the global minimizations involved within this step become too costly and the jth loop must be terminated prematurely. Thus, given that $\delta_{k+1} = \delta_s$ at successful iterations, we next have to consider what happens in Step 1 of iteration $k+1$ when one or more of the $\delta_{k+1,j}$ is too large. In this case, the definition of $\phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})$ (see (2.2)) implies that there might exist a move $d_{k+1,j}$ with $\|d_{k+1,j}\| \leq \delta_{k+1,j}$ such that

$$T_{w,j}(x_{k+1}, 0) - T_{w,j}(x_{k+1}, d_{k+1,j}) > \varepsilon_j \frac{\delta_{k+1,j}^j}{j!},$$

preventing termination even if x_{k+1} is a suitable (ε, δ)-approximate minimizer. This is obviously a serious problem from the point of view of bounding evaluation complexity, since the algorithm will continue and evaluate further, unnecessary, values of f, c, and their derivatives. Two possibilities may then occur. Either iteration $k+1$ is unsuccessful, σ_k increased causing a subsequent stepsize reduction and, if the behavior persists, forcing convergence to x_k, or it is successful,[5] yielding a further objective function reduction and allowing the algorithm to progress towards an alternative approximate minimizer with a lower objective function value. The complexity bound is maintained if (7.1) or (7.2) holds, or if an insufficient decrease only occurs at most a number of times independent of $\varepsilon_{\min}$. However, even if this is not the case and the complexity bound we have derived evaporates as a consequence, the fact that the algorithm moves on can be viewed as beneficial for the optimization process from a more global perspective.

Termination test. One also needs global minimization to compute the optimality measure $\phi_{w,j}^{\delta_{k,j}}(x_k)$ in Step 1. Clearly, the global optimization defining $\phi_{w,j}^{\delta_{k,j}}(x_k)$ in (2.2) may be terminated as soon as an approximate solution d is found such that

$$\phi_{w,j}^{\delta_{k,j}}(x_k) > \varepsilon_j \frac{\delta_{k,j}^j}{j!},$$

thereby avoiding a full-accuracy computation of the global minimizer. When far from the solution, we expect the optimality measure to be large, and hence such an approximate

5 As suggested by the fact that minimization in Step 2 of iteration $k+1$ may obviously be started from $x_{k+1} + d_{k+1,j}$, a point already providing descent on a good approximation of w.

solution d to be found quickly. Suppose now that the solution is approached, and that the minimization of $T_{w,j}(x_k, d)$ within the ball of radius $\delta_{k,j}$ can only be performed inexactly in that one can only find a move d such that

$$T_{w,j}(x_k, d) - T_{w,j}(x_k, d_*) \leq \varepsilon_{\phi,j} \frac{\delta_{k,j}^j}{j!}, \tag{7.3}$$

where d_* is the elusive constrained global minimizer and $\varepsilon_{\phi_j} \in (0, 1]$. Then the only effect of this computational constraint is to limit the achievable accuracy on the approximate minimizer by imposing that $\varepsilon_j \geq \varepsilon_{\phi,j}$. However, achieving (7.3) for small $\delta_{k,j}$ might also be too challenging: one is then left (as above) with the option of using a larger value of $\delta_{k,j}$, possibly missing the identification of x_k as an (ε, δ)-approximate minimizer, which potentially leads to an alternative better one but destroys the complexity guarantee.

To summarize this discussion, the need for global optimization in Steps 1 and 2.4 is driven by the desire to obtain a good evaluation complexity bound (by avoiding further evaluations if a suitable approximate minimizer has been found). The algorithm could still employ approximate calculations, but at the price of losing the complexity guarantee or limiting the achievable accuracy.

8. CONCLUSIONS AND PERSPECTIVES

We have presented an adaptive regularization algorithm for the minimization of nonconvex, nonsmooth composite functions, and proved bounds detailed in Table 1.1 on the evaluation complexity (as a function of accuracy) for composite and smooth problems and for arbitrary model degree and optimality orders.

These results complement the bound proved in [12] for weak approximate minimizers of inexpensively constrained smooth problems (third column of Table 1.1) by providing corresponding results for strong approximate minimizers. They also provide the first complexity results for the convergence to minimizers of order larger than one for (possibly nonsmooth and inexpensively constrained) composite ones.

The fact that high-order approximate minimizers for nonsmooth composite problems can be defined and computed in a quantifiable way opens up interesting possibilities. In particular, these results may be applied in the case of expensively-constrained optimization problems, where exact penalty functions result in composite subproblems of the type studied here.

REFERENCES

[1] A. A. Ahmadi and J. Zhang, Complexity aspects of local minima and related notions. *Adv. Math.* (2021), online.

[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** (2009), 183–202.

[3] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint, Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM J. Optim.* **29** (2019), no. 4, 2881–2915.

[4] E. G. Birgin, J.-L. Gardenghi, J. M. Martínez, and S. A. Santos, On the use of third-order models with fourth-order regularization for unconstrained optimization. *Optim. Lett.* **14** (2020), 815–838.

[5] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Math. Program., Ser. A* **163** (2017), no. 1, 359–368.

[6] A. M. Bruckner and E. Ostrow, Some function classes related to the class of convex functions. *Pacific J. Math.* **12** (1962), no. 4, 1203–1215.

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Math. Program., Ser. A* **130** (2011), no. 2, 295–319.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint, On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.* **21** (2011), no. 4, 1721–1739.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Improved worst-case evaluation complexity for potentially rank-deficient nonlinear least-Euclidean-norm problems using higher-order regularized models. Technical Report naXys-12-2015, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium, 2015.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients. *Optim. Methods Softw.* **6** (2017), no. 6, 1273–1298.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization. *Found. Comput. Math.* **18** (2018), no. 5, 1073–1107.

[12] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *SIAM J. Optim.* **30** (2020), no. 1, 513–541.

[13] X. Chen and Ph. L. Toint, High-order evaluation complexity for convexly-constrained optimization with non-Lipschitzian group sparsity terms. *Math. Program., Ser. A* **187** (2021), 47–78.

[14] X. Chen, Ph. L. Toint, and H. Wang, Partially separable convexly-constrained optimization with non-Lipschitzian singularities and its complexity. *SIAM J. Optim.* **29** (2019), 874–903.

[15] F. E. Curtis, D. P. Robinson, and M. Samadi, An inexact regularized Newton framework with a worst-case iteration complexity of $O(\varepsilon^{-3/2})$ for nonconvex optimization. *IMA J. Numer. Anal.* **00** (2018), 1–32.

[16] D. L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52** (2006), no. 4, 1289–1306.

[17] Z. Drezner and H. W. Hamacher, *Facility location: applications and theory*. Springer, Heidelberg, Berlin, New York, 2002.

[18] R. Fletcher, *Practical Methods of Optimization: Constrained Optimization*. J. Wiley and Sons, Chichester, England, 1981.

[19] N. I. M. Gould, T. Rees, and J. A. Scott, Convergence and evaluation-complexity analysis of a regularized tensor-Newton method for solving nonlinear least-squares problems. *Comput. Optim. Appl.* **73** (2019), no. 1, 1–35.

[20] S. Gratton, E. Simon, and Ph. L. Toint, An algorithm for the minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity. *Math. Program., Ser. A* **187** (2021), 1–24.

[21] P. C. Hansen, *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, Philadelphia, USA, 1998.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86** (1998), no. 11, 2278–2324.

[23] A. S. Lewis and S. J. Wright, A proximal method for composite minimization. *Math. Program., Ser. A* **158** (2016), 501–546.

[24] Yu. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance. *Math. Program., Ser. A* **108** (2006), no. 1, 177–205.

[25] C. W. Royer and S. J. Wright, Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM J. Optim.* **28** (2018), no. 2, 1448–1477.

[26] R. Tibshirani, Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58** (1996), no. 1, 267–288.

[27] N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan, Training deep neural networks with 8-bit floating point numbers. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7686–7695, 2018.

CORALIA CARTIS

Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK, coralia.cartis@maths.ox.ac.uk

NICHOLAS I. M. GOULD

Computational Mathematics Group, STFC-Rutherford Appleton Laboratory, Chilton OX11 0QX, UK, nick.gould@stfc.ac.uk

PHILIPPE L. TOINT

Namur Centre for Complex Systems (naXys), University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium, philippe.toint@unamur.be

AN OVERVIEW OF NONLINEAR OPTIMIZATION

YU-HONG DAI

ABSTRACT

Nonlinear optimization stems from calculus and becomes an independent subject due to the proposition of Karush–Kuhn–Tucker optimality conditions. The ever-growing realm of applications and the explosion in computing power is driving nonlinear optimization research in new and exciting directions. In this article, I shall give a brief overview of nonlinear optimization, mainly on unconstrained optimization, constrained optimization, and optimization with least constraint violation.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 90C30; Secondary 65K05, 90C06

KEYWORDS

unconstrained optimization, constrained optimization, optimization with least constraint violation, gradient method, conjugate gradient method, sequential quadratic programming method, interior-point method, augmented Lagrangian method of multipliers

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5290–5313
DOI 10.4171/ICM2022/72

1. INTRODUCTION

It is known that nonlinear optimization stems from calculus. Consider the unconstrained optimization problem

$$\min_{x\in\Re^n} \quad f(x), \tag{1.1}$$

where $f:\Re^n\to\Re$ is smooth and its gradient g is available. The calculus invented by Newton and Leibniz in the seventeenth century provided a necessary condition for a point x to be the optimal solution of (1.1), which is $\nabla f(x)=0$, i.e., the tangent line of f at x is horizontal. For equality constrained optimization, the necessary optimality condition is that the derivatives of the Lagrangian function with respect to the primal and dual variables are equal to zero, which was exposed by Lagrange in the eighteenth century. Nonlinear optimization became an independent subject when Karush [52] and Kuhn and Tucker [53] provided necessary optimality conditions for general optimization subjected to equality and inequality constraints,

$$\min \quad f(x) \tag{1.2}$$

$$\text{such that} \quad h(x)=0, \tag{1.3}$$

$$g(x)\geq 0, \tag{1.4}$$

where $f:\Re^n\to\Re$, $h:\Re^n\to\Re^{m_E}$, $g:\Re^n\to\Re^{m_I}$ are supposed to be twice continuously differentiable functions. The proposition of the Fletcher–Reeves conjugate gradient method [39] and the Davidon–Fletcher–Powell quasi-Newton method [33, 38] greatly promoted the development of nonlinear optimization.

This article shall give a brief overview of nonlinear optimization, mainly on unconstrained optimization, constrained optimization, and optimization with least constraint violation.

2. UNCONSTRAINED OPTIMIZATION

The design and analysis of numerical methods for unconstrained optimization is closely related to the unconstrained quadratic optimization

$$\min_{x\in\Re^n} \quad q(x):=\frac{1}{2}x^TAx-b^Tx, \tag{2.1}$$

where $b\in\Re^n$ and $A\in\Re^{n\times n}$ is symmetric and positive definite with eigenvalues $0<\lambda_1\leq\cdots\leq\lambda_n$. Fundamental methods for unconstrained optimization include gradient methods, conjugate gradient methods, quasi-Newton methods, Newton method, and derivative-free methods. We focus on two classes of first-order methods, gradient methods, and conjugate gradient methods, which are suitable for large-scale problems.

The gradient method can be dated back to Cauchy [9], and the first nonlinear conjugate gradient method is due to Fletcher and Reeves [39]. Driven by practical applications, various variants of the methods have been proposed for convex optimization, nonsmooth optimization, stochastic optimization, etc. For smooth optimization, the two classes of methods are significantly improved by asking their search directions to be close to the Newton or

quasi-Newton direction in some sense. Typical examples of the conjugate gradient method are the Dai–Yuan method [27], the Hager–Zhang method [46], and the Dai–Kou method [22]. For the gradient method, one milestone work is the Barzilai–Borwein (nonmonotone) gradient method, while another significant work is the Yuan stepsize [87], which leads to the proposition of the efficient Dai–Yuan (monotone) gradient method [30]. Interestingly enough, Huang et al. [51] found that it is possible to equip the Barzilai–Borwein method with the two-dimensional quadratic termination property.

2.1. Gradient methods

Gradient methods search along the negative gradient and are of the form

$$x_{k+1} = x_k - \alpha_k g_k, \tag{2.2}$$

where $g_k = \nabla f(x_k)$ and $\alpha_k > 0$ is the stepsize. Different choices of the stepsize α_k lead to different gradient methods. The steepest descent (SD) method, which is due to Cauchy [9], determines its stepsize by the exact line search, i.e.,

$$\alpha_k^{\mathrm{SD}} = \arg\min_{\alpha>0} f(x_k - \alpha g_k). \tag{2.3}$$

The SD method is shown to be Q-linearly convergent, but its performance is poor when the problem is ill-conditioned [1]. Specifically, the SD method will asymptotically tend to minimize the function in some two-dimensional subspace and produce zigzags [40]. If the dimension is greater than one, the SD stepsize (2.3) always tends to be long, and some shortened SD methods are proposed in [30].

One milestone work on the gradient method is due to Barzilai and Borwein [3]. Its basic idea is to ask the matrix $\alpha_k^{-1} I$ or $\alpha_k I$ have a certain quasi-Newton property. Then by minimizing $\|s_{k-1} - (\alpha_k^{-1} I) y_{k-1}\|$ or $\|(\alpha_k I) s_{k-1} - y_{k-1}\|$ with respect to α_k, where $s_{k-1} = x_k - x_{k-1}$, $y_{k-1} = g_k - g_{k-1}$ and $\|\cdot\|$ is the two-norm, two stepsizes are derived as

$$\alpha_k^{\mathrm{BB}\,1} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}, \quad \alpha_k^{\mathrm{BB}\,2} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}. \tag{2.4}$$

The stepsizes $\alpha_k^{\mathrm{BB}\,1}$ and $\alpha_k^{\mathrm{BB}\,2}$ are called long and short Barzilai–Borwein (BB) stepsizes, respectively, since $\alpha_k^{\mathrm{BB}\,1} \geq \alpha_k^{\mathrm{BB}\,2}$ if $s_{k-1}^T y_{k-1} > 0$. Despite its heavy nonmonotone behavior, the BB method performs significantly better than the SD method in practice; see, e.g., [36]. For unconstrained quadratic optimization, the BB method is proved to be R-superlinearly convergent if the dimension is two [3]. For general dimension, the BB method is globally convergent [73] and the convergence is R-linear [25]. An efficient extension of the BB method for unconstrained optimization is given in [74] by incorporating the Grippo–Lampariello–Lucidi (GLL) nonmonotone line search [45]. Interestingly enough, it is shown in [25] that the BB stepsize can asymptotically be accepted by the GLL nonmonotone line search when the iterate is close to the solution. This property is similar to the fact that the unit stepsize can asymptotically be accepted by Newton or quasi-Newton methods using the Armijo or Wolfe line search. Furthermore, efficient projected gradient methods based on BB-like methods and applications can be found in [4, 20, 63, 90], among many other references. The numerical

efficiency of the BB method over the SD method has stimulated many studies on the gradient method.

However, it is intriguing to provide theoretical evidence showing that the BB method performs much better than the SD method for high-dimensional problems. One possible angle is to relate the stepsize in the gradient method to the eigenvalues of the Hessian of the function. To this aim, consider the unconstrained quadratic optimization problem (2.1). In this case, by (2.1) and (2.2), we have that $g_{k+1} = (I - \alpha_k A)g_k$ for all $k \geq 1$. Then we see that the gradient method with constant stepsizes (i.e., $\alpha_k \equiv \alpha$ for some $\alpha > 0$) is equivalent to the shifted power method for computing some eigenvalue of the matrix A since

$$\frac{g_{k+1}}{\|g_{k+1}\|} = \frac{(I - \alpha A)^k g_1}{\|(I - \alpha A)^k g_1\|} = \frac{(A - \alpha^{-1} I)^k g_1}{\|(A - \alpha^{-1} I)^k g_1\|}. \tag{2.5}$$

For example, if $\alpha_k \equiv \frac{1}{2L}$, where L is the gradient Lipschitz constant, which was one choice in the early times [2], $\frac{g_{k+1}}{\|g_{k+1}\|}$ will tend to the eigenvector corresponding to the minimal eigenvalue of A provided the initial gradient g_1 has a nonzero component in this eigenvector. Another fact is the quadratic termination property of the gradient method, which was exposed by Lai [54]. To see this, notice that $g_{k+1} = \prod_{j=1}^{k}(I - \alpha_j A)g_1$ for $k \geq 1$. Then by the Hamilton–Caley theorem, we have that g_{n+1} vanishes if the set of stepsizes $\{\alpha_i : 1 \leq i \leq n\}$ coincides with the set of inverse eigenvalues of A, $\{\lambda_i^{-1} : 1 \leq i \leq n\}$. A natural corollary is as follows.

Lemma 2.1. *Consider the gradient method* (2.2) *for the unconstrained quadratic optimization problem* (2.1)*. Assume that the initial gradient g_1 has nonzero components in all eigenvectors of the matrix A. If the gradient method is R-superlinearly convergent, then, for each eigenvalue λ_i $(1 \leq i \leq n)$ of A, there exists a subsequence $\{\alpha_{k_i}\}$ such that* $\lim_{k_i \to \infty} \alpha_{k_i} = \lambda_i^{-1}$.

The above lemma provides us an insight about convergence properties of gradient methods. From the proof of the R-superlinear convergence of the BB method in the two-dimensional setting [3], it is easy to see that there do exist subsequences $\{\alpha_{k_1}\}$ and $\{\alpha_{k_2}\}$ such that they converge to the two inverse eigenvalues of the Hessian, respectively. Dai and Fletcher [19] observed this phenomenon for the BB method in the three-dimensional setting as well and showed that the BB method is likely to be R-superlinearly convergent in this case. It is also shown in [19] that the cyclic steepest descent method is likely to be R-superlinearly convergent for n-dimensional convex quadratic functions provided that $m \geq \frac{n+1}{2}$, where m is the cyclic time of the steepest descent stepsize.

Another significant addition to the gradient method is the Yuan stepsize [87], which is such that, if the previous and later steps use SD stepsizes, the gradient method can give the exact minimizer of a two-dimensional convex quadratic function. A variant of the Yuan stepsize is given by Dai and Yuan [30] as

$$\alpha_k^{\mathrm{DY}} = \frac{2}{\frac{1}{\alpha_{k-1}^{\mathrm{SD}}} + \frac{1}{\alpha_k^{\mathrm{SD}}} + \sqrt{\left(\frac{1}{\alpha_{k-1}^{\mathrm{SD}}} - \frac{1}{\alpha_k^{\mathrm{SD}}}\right)^2 + \frac{4\|g_k\|^2}{(\alpha_{k-1}^{\mathrm{SD}}\|g_{k-1}\|)^2}}}. \tag{2.6}$$

They also suggested the so-called Dai–Yuan gradient method (2.2) with

$$\alpha_k = \begin{cases} \alpha_k^{\mathrm{SD}}, & \text{if } \mathrm{mod}(k,4) = 0, 1, \\ \alpha_k^{\mathrm{DY}}, & \text{if } \mathrm{mod}(k,4) = 2, 3. \end{cases} \tag{2.7}$$

The Dai–Yuan gradient method is monotone since $\alpha_k^{\mathrm{DY}} \le \alpha_k^{\mathrm{SD}}$. This is the first monotone gradient method which can beat the BB nonmonotone gradient method for unconstrained quadratic optimization.

A recent progress in the gradient method is provided by Huang et al. [51], who introduced a new mechanism for the gradient method to achieve the two-dimensional quadratic termination property. Given $\nu_1(k), \nu_2(k) \in \{1, \ldots, k\}$ and some suitable functions ψ_1, ψ_2, ψ_3, ψ_4 satisfying $\psi_1(A)\psi_2(A) = \psi_3(A)\psi_4(A)$, they suggested calculating the stepsize α_k by solving the following quadratic equation:

$$\begin{aligned} &g_{\nu_1(k)}^T \psi_1(A)(I - \alpha_k A) g_k \cdot g_{\nu_2(k)}^T \psi_2(A)(I - \alpha_k A) g_k \\ &\quad = g_{\nu_1(k)}^T \psi_3(A)(I - \alpha_k A) g_k \cdot g_{\nu_2(k)}^T \psi_4(A)(I - \alpha_k A) g_k, \end{aligned} \tag{2.8}$$

and proved that the gradient method using any stepsize obtained from (2.8) and α_{k+2} in the form of $\frac{(A^\mu g_{k+i})^T (A^\mu g_{k+i})}{(A^\mu g_{k+i})^T A (A^\mu g_{k+i})}$ with $i = 1$ or 2 and μ being some real number achieves the two-dimensional quadratic termination property. Interestingly, the stepsize α_k^{DY} in (2.6) is a solution of equation (2.8) corresponding to $\nu_1(k) = k - 1$, $\nu_2(k) = k$, $\psi_1(A) = \psi_4(A) = (I - \alpha_{k-1} A)^{-1}$, and $\psi_2(A) = \psi_3(A) = I$.

To equip the BB method with the two-dimensional quadratic termination property, Huang et al. [51] chose $\nu_1(k) = k - 2$, $\nu_2(k) = k - 1$, $\psi_1(A) = (I - \alpha_{k-2} A)^{-1}$, $\psi_2(A) = (I - \alpha_{k-1} A)^{-1}$, $\psi_3(A) = \psi_1(A)\psi_2(A)$, and $\psi_4(A) = I$. Then by (2.8), they obtained the following novel stepsize:

$$\alpha_k^{\mathrm{HDL}} = \frac{2}{\frac{\phi_2}{\phi_3} + \sqrt{\left(\frac{\phi_2}{\phi_3}\right)^2 - 4\frac{\phi_1}{\phi_3}}}, \tag{2.9}$$

where

$$\frac{\phi_1}{\phi_3} = \frac{\alpha_{k-1}^{\mathrm{BB}\,2} - \alpha_k^{\mathrm{BB}\,2}}{\alpha_{k-1}^{\mathrm{BB}\,2} \alpha_k^{\mathrm{BB}\,2} (\alpha_{k-1}^{\mathrm{BB}\,1} - \alpha_k^{\mathrm{BB}\,1})}, \quad \frac{\phi_2}{\phi_3} = \frac{\alpha_{k-1}^{\mathrm{BB}\,1} \alpha_{k-1}^{\mathrm{BB}\,2} - \alpha_k^{\mathrm{BB}\,1} \alpha_k^{\mathrm{BB}\,2}}{\alpha_{k-1}^{\mathrm{BB}\,2} \alpha_k^{\mathrm{BB}\,2} (\alpha_{k-1}^{\mathrm{BB}\,1} - \alpha_k^{\mathrm{BB}\,1})}. \tag{2.10}$$

It is observed in [51] that the use of the stepsize α_k^{HDL} can make both the BB1 and BB2 methods achieve the two-dimensional quadratic termination property. The computation of α_k^{HDL} only involves the BB stepsizes in the previous two iterations and does not require exact line searches or the Hessian computation. Hence it can easily be extended for nonlinear optimization.

Based on the new stepsize α_k^{HDL} and the general framework in [92], an efficient gradient method for solving unconstrained optimization problem (2.1) is suggested in [51]. In particular, the method uses $\alpha_1 = \alpha_1^{\mathrm{SD}}$, $\alpha_2 = \alpha_2^{\mathrm{BB}\,1}$, and, for $k \ge 3$,

$$\alpha_k = \begin{cases} \min\{\alpha_{k-1}^{\mathrm{BB}\,2}, \alpha_k^{\mathrm{BB}\,2}, \alpha_k^{\mathrm{HDL}}\}, & \text{if } \alpha_k^{\mathrm{BB}\,2} / \alpha_k^{\mathrm{BB}\,1} < \tau_k, \\ \alpha_k^{\mathrm{BB}\,1}, & \text{otherwise}, \end{cases} \tag{2.11}$$

where $\tau_k > 0$ is chosen in some way. The method (2.11) appears to be much better than BB, Dai–Yuan, and some other recent gradient methods. With the projection technique, the method (2.11) was also extended in [51] to unconstrained optimization, box-constrained optimization, and singly linearly box-constrained optimization, and good numerical results were obtained.

There are still many questions about the gradient method to be investigated. Is it possible to provide more theoretical evidence showing the efficiency of the BB method for high-dimensional functions? What is the best choice of the stepsize in the gradient method? How to extend the existing efficient gradient methods to many other areas?

2.2. Conjugate gradient methods

Conjugate gradient methods are a class of important methods for solving (1.1). They are of the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2.12}$$

where α_k is the stepsize obtained by a line search and d_k is the search direction given by

$$d_k = -g_k + \beta_k d_{k-1}, \tag{2.13}$$

except for $d_1 = -g_1$. The scalar β_k is the so-called conjugate gradient parameter such that the method (2.12)–(2.13) reduces to the linear conjugate gradient method if the objective function is quadratic and the line search is exact.

For nonlinear functions, however, different formulae for the parameter β_k result in different conjugate gradient methods and their properties can be significantly different. To distinguish the linear conjugate gradient method, sometimes we call the conjugate gradient method for unconstrained optimization as the nonlinear conjugate gradient method. The work of Fletcher and Reeves [39] not only opened the door to the nonlinear conjugate gradient field but also greatly stimulated the study of nonlinear optimization. Four well-known formulae for β_k are called the Fletcher–Reeves (FR) [39], Dai–Yuan (DY) [27], Polak–Ribière–Polyak (PRP) [67,68], and Hestenes–Stiefel (HS) [50], which are given by

$$\begin{aligned} \beta_k^{\mathrm{FR}} &= \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \quad \beta_k^{\mathrm{DY}} = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}}, \\ \beta_k^{\mathrm{PRP}} &= \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}, \quad \beta_k^{\mathrm{HS}} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, \end{aligned} \tag{2.14}$$

respectively, where $y_{k-1} = g_k - g_{k-1}$ as before.

Since the exact line search is usually expensive and impractical, the strong Wolfe line search is often considered in the early convergence analysis and numerical implementation for nonlinear conjugate gradient methods. The strong Wolfe line search aims to find a stepsize $\alpha_k > 0$ satisfying

$$f(x_k + \alpha_k d_k) \le f(x_k) + \delta \alpha_k g_k^T d_k, \tag{2.15}$$

$$\left| g(x_k + \alpha_k d_k)^T d_k \right| \le -\sigma g_k^T d_k, \tag{2.16}$$

where $0 < \delta < \sigma < 1$. However, it was shown in [28] that even with strong Wolfe line searches, none of the FR, PRP, and HS methods can ensure the descent property of the search direction if the parameter σ is not properly chosen. If a descent search direction is not produced, a practical remedy is to restart the method along $-g_k$. This may degrade the efficiency of the method since the second-derivative information achieved along the previous search direction is discarded.

It is known that quasi-Newton methods often use the standard Wolfe line search, which aims to find a stepsize $\alpha_k > 0$ satisfying (2.15) and

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{2.17}$$

where $0 < \delta < \sigma < 1$. Dai and Yuan [27] were able to establish the descent property and global convergence of the DY method with the standard Wolfe line search under weak assumptions on the objective function.

Assumption 2.1. (i) The level set $\mathcal{L} = \{x \in \Re^n : f(x) \leq f(x_1)\}$ is bounded, where x_1 is the starting point; (ii) f is continuously differentiable in some neighborhood of $\mathcal{L}$ and its gradient is Lipschitz continuous.

Theorem 2.1 ([27]). *Suppose that f satisfies Assumption 2.1. Consider the sequence $\{x_k\}$ generated by the DY method (2.12)–(2.13) with $\beta_k = \beta_k^{\mathrm{DY}}$ and the standard Wolfe line search (2.15) and (2.17). Assume that $\|g_k\| \neq 0$ for all k. Then we have that $g_k^T d_k < 0$ for all k. Furthermore, the DY method converges in the sense that* $\liminf_{k\to+\infty} \|g_k\| = 0$.

It is noted in [27] that the DY formula can be rewritten as $\beta_k^{\mathrm{DY}} = \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}$. It is remarkable that the DY method has a certain self-adjusting property that is independent of the line search and the function convexity. The DY direction can also be used to restart optimization methods while guaranteeing the global convergence of the method (see [28]). Interestingly enough, Dai [16] provided another nonlinear conjugate gradient method which can ensure the descent property of the search direction without any line searches.

The following theorems provide general convergence results for nonlinear conjugate gradient methods with the strong Wolfe line search and the standard Wolfe line search, respectively. The results are very useful in the convergence analysis of various nonlinear conjugate gradient methods.

Theorem 2.2 ([21]). *Suppose that Assumption 2.1 holds. Consider any method of the form (2.12)–(2.13) with d_k satisfying $g_k^T d_k < 0$ and with the strong Wolfe line search (2.15) and (2.16). Then the method is globally convergent in the sense that* $\liminf_{k\to+\infty} \|g_k\| = 0$ *if* $\sum_{k\geq 1} \|d_k\|^{-2} = +\infty$.

Theorem 2.3 ([17]). *Suppose that Assumption 2.1 holds. Consider any method of the form (2.12)–(2.13) with d_k satisfying $g_k^T d_k < 0$ and with the standard Wolfe line search (2.15) and (2.17). Then the method is globally convergent in the sense that* $\liminf_{k\to+\infty} \|g_k\| = 0$ *if the scalar β_k is such that* $\sum_{k\geq 1} \prod_{j=2}^{k} \beta_j^{-2} = +\infty$.

Powell [71] found that the PRP method can automatically generate a search direction close to the steepest descent direction once a small step occurs, whereas the FR method may produce many tiny steps continuously. This explains why the PRP method sometimes performs much better than the FR method in practice. Nevertheless, Powell [71] showed that, even with exact line searches, the PRP method can cycle indefinitely without approaching a stationary point. To change this unbalanced state, Touati-Ahmed and Storey [79] proposed the hybrid conjugate gradient method, where

$$\beta_k^{\mathrm{FRPRP}} = \max\{0, \min\{\beta_k^{\mathrm{PRP}}, \beta_k^{\mathrm{FR}}\}\}. \tag{2.18}$$

Gilbert and Nocedal [42] modified the PRP method by setting

$$\beta_k^{\mathrm{PRP+}} = \max\left\{\frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}, 0\right\}. \tag{2.19}$$

They established the global convergence results for the FRPRP and PRP+ methods, but found that the two methods are not significantly more efficient than the PRP method itself. Nevertheless, Dai and Yuan [29] were able to extend the convergence theorem of the DY method, Theorem 2.1, to the following hybrid conjugate gradient method:

$$\beta_k^{\mathrm{DYHS}} = \max\{0, \min\{\beta_k^{\mathrm{HS}}, \beta_k^{\mathrm{DY}}\}\}, \tag{2.20}$$

and found that the DYHS method with the standard Wolfe line search performs much better than the PRP method using the strong Wolfe line search.

Since $y_{k-1} = As_{k-1} = \alpha_{k-1} Ad_{k-1}$ in case of unconstrained quadratic optimization, an equivalent expression of the conjugacy condition $d_k^T Ad_{k-1} = 0$ is $d_k^T y_{k-1} = 0$. For general functions, however, we have for quasi-Newton methods that $d_k = -B_k^{-1} g_k$, where the approximation matrix B_k satisfies the quasi-Newton equation $B_k s_{k-1} = y_{k-1}$. This hints us at the nonlinear conjugate gradient condition $d_k^T y_{k-1} = (-B_k^{-1} g_k)^T (B_k s_{k-1}) = -g_k^T s_{k-1}$. By introducing a scaling factor t, Dai and Liao [24] considered a nonlinear conjugacy condition $d_k^T y_{k-1} = -t g_k^T s_{k-1}$ and proposed the following family for conjugate gradient methods:

$$\beta_k^{\mathrm{DL}}(t) = \frac{g_k y_{k-1}}{d_{k-1}^T y_{k-1}} - t \frac{g_k s_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{2.21}$$

Although the descent property of the search direction is sufficient for establishing the convergence results, efficient conjugate gradient methods have been proposed such that the sufficient descent condition

$$g_k^T d_k \le -c\|g_k\|^2 \tag{2.22}$$

holds for some constant $c > 0$ and all $k \ge 1$. Specifically, Hager and Zhang [46] proposed a family of conjugate gradient methods, where

$$\beta_{k+1}^{\mathrm{HZ}} = \frac{g_{k+1}^T y_k}{d_k^T y_k} - \lambda_k \frac{\|y_k\|^2}{d_k^T y_k} \frac{g_{k+1}^T d_k}{d_k^T y_k}, \tag{2.23}$$

with $\lambda_k > \bar{\lambda} > \frac{1}{4}$, and they preferred the choice $\lambda_k = 2$. By introducing a suitable truncation of β_k and the approximate Wolfe line search, they built a conjugate gradient software, called

CG_DESCENT [47], which performs better than PRP+ of Gilbert and Nocedal. By observing that the loss of orthogonality in the sequence of gradients caused by numerical error might slow down the convergence of conjugate gradient methods, Hager and Zhang [48] updated CG_ DESCENT to Version 6.8 by combining the limited memory technique.

By projecting the search direction of the self-scaling memoryless BFGS method, which was proposed by Perry [66] and Shanno [77], into the one-dimensional manifold $\mathcal{S}_k = \text{Span}\{-g_k + \beta d_{k-1}\}$, Dai and Kou [22] proposed a family of conjugate gradient methods, where

$$\beta_k(\tau_k) = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - \left\{\tau_{k-1} + \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} - \frac{s_{k-1}^T y_{k-1}}{\|s_{k-1}\|^2}\right\} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{2.24}$$

Then by choosing $\tau_{k-1} = \frac{s_{k-1}^T y_{k-1}}{\|s_{k-1}\|^2}$, they recommended the formula

$$\beta_k^{\text{DK}} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \tag{2.25}$$

which is such that the sufficient descent condition (2.22) holds with $c = \frac{3}{4}$. The software CGOPT was then developed in [22] based on the Dai–Kou method and an improved Wolfe line search. Furthermore, CGOPT was updated in [60] to Version 2.0, which consists of standard CG iterations and subspace iterations and is a strong competitor of CG_DESCENT.

Despite significant progresses, we feel there is still much more room to seek for the best nonlinear conjugate gradient algorithms. For example, Yuan and Stoer [88] first presented the subspace minimization conjugate gradient method by determining the search direction via the subproblem $\min\{g_k^T d + \frac{1}{2} d^T B_k d : d \in \text{Span}\{g_k, d_{k-1}\}\}$. Following this line, Dai and Kou [23] approximated the term $g_k^T B_k g_k$ by $\frac{3}{2} \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \|g_k\|^2$ and presented an efficient Barzilai–Borwein conjugate gradient method.

3. CONSTRAINED OPTIMIZATION

An intuitive way to deal with constrained optimization problems is to transform them into unconstrained optimization problems via penalty functions or indicator functions. Nowadays, there are many classes of numerical methods and software for constrained optimization; see, e.g., [65, 89]. Sequential quadratic programming methods and interior-point methods are two classes of very efficient numerical methods for constrained optimization among many others. In addition, augmented Lagrangian methods of multipliers also received a lot of attention since they form the base of alternating direction method of multipliers (see [5]), which can deal with large-scale structured problems. In this section, we shall briefly review some of our recent contributions to the three classes of methods.

3.1. Sequential quadratic programming methods

The sequential quadratic programming (SQP) method, also called Wilson–Han–Powell method, is one of the most effective methods for constrained optimization and can be viewed as a natural extension of Newton and quasi-Newton methods. Its basic idea is

to transform the original problem into a sequence of quadratic program (QP) subproblems. After solving each QP subproblem, we wish the full SQP-step to be a superlinearly convergent step; by combining some criterion, we evaluate whether to accept this full step and introduce some remedy if necessary. Based on the used criterion, SQP methods can roughly be classified into two categories. One is penalty-type methods, whose main feature is to use some penalty function. The other is penalty-free methods, which do not use any penalty parameters, e.g., filter methods [37], the methods without any penalty function or a filter [44].

However, two possible difficulties may arise in SQP methods. One is that the QP subproblem may be inconsistent. The other is how to avoid the Maratos effect [61] since the full SQP-step may lead to an increase in both the objective function and the constraint violation even when the iteration is arbitrarily close to a regular minimizer.

Various techniques are available for dealing with inconsistency of the QP subproblem. Early such works include the scaling technique by Powell [70] and the Sl_1QP method by Fletcher [35]. Spellucct [78] introduced some slack variables for dealing with inconsistent subproblems. Liu and Yuan [58] provided a robust SQP method by solving an unconstrained piecewise quadratic subproblem and a QP subproblem at each iteration. Fabien [34] solved a relaxed, strictly convex, QP subproblem if the constraints are inconsistent.

For the Maratos effect, Chen et al. [13] gave the following formal definition.

Definition 3.1. Let x^* and $v(\|c(x)\|)$ be a solution and a measurement of the constraint violation of (1.2)–(1.4), respectively. Given a sequence $\{x_k\}$ which converges to x^* and a sequence of full SQP-steps $\{d_k\}$, we say that the Maratos effect happens if (i) $\lim_{k\to+\infty} \|x_k + d_k - x^*\|/\|x_k - x^*\| = 0$; (ii) $f(x_k + d_k) > f(x_k)$ and $v(\|c(x_k + d_k)\|) > v(\|c(x_k)\|)$.

When the Maratos effect happens, the full SQP-step may not be accepted since it makes both the objective function and the constraint violation worse. In fact, in this case, we see that the pair $(\|h(x_k)\|, f(x_k))$ dominates the pair $(\|h(x_k + d_k)\|, f(x_k + d_k))$ even if $x_k + d_k$ is much closer to x^* than x_k and hence $x_k + d_k$ will not be accepted by the filter method initially proposed by Fletcher and Leyffer [37]. This is also the case for many other globally convergent penalty-type and penalty-free-type algorithms. For example, if l_1 and l_∞ exact penalty functions are used, the full trial step d_k will be rejected as well since the value $f(x) + \sigma\|h(x)\|_p$ ($\sigma > 0$, $p = 1, \infty$) becomes worse.

Several approaches have been proposed for avoiding the Maratos effect, including nonmonotone line search strategies [11], second order correction step [41, 62], and the use of differentiable exact penalty functions [72]. The computation of second-order correction steps may be cumbersome, and the nonmonotone framework will complicate the algorithmic implementation. Another approach of avoiding the Maratos effect is to utilize the Lagrangian function value instead of the objective function value. Such an idea can be found in Ulbrich [80], who proposed a trust-region filter-SQP method by introducing the Lagrangian function value in the filter.

For efficiency evidence of using the Lagrangian function value in avoiding the Maratos effect, Chen et al. [13] provided the following basic result (for simplicity, it is assumed that $m_I = 0$).

Theorem 3.1. *Suppose that* (x^*, λ^*) *is a KKT pair of problem* (1.2)–(1.3)*, at which the second-order sufficient conditions and the linear independence constraint qualification hold. Assume that* $v(\|h(x)\|)$ *is a measurement of constraint violation of the problem,* $\lambda(x)$ *is a Lipschitz continuous multipliers function, and* $P_k(B_k - \nabla^2_{xx} L(x_k, \lambda(x_k)))d_k = o(\|d_k\|)$*, where* $\{x_k\}$ *converges to* x^**,* B_k *is the approximation of* $\nabla^2_{xx} L(x_k, \lambda(x_k))$ *in the QP subproblem,* P_k *is an orthogonal projection matrix from* $\mathcal{R}^n$ *to the null space of* A_k^T*, and* d_k *is the full SQP-step. If* $v(\|h(x_k + d_k)\|) > v(\|h(x_k)\|)$*, then there must exist some constant* $b_0 > 0$ *such that* $L(x_k + d_k, \lambda(x_k + d_k)) \leq L(x_k, \lambda(x_k)) - b_0\|d_k\|^2$.

The above theorem indicates that, when the Maratos effect happens, there must be a sufficient decrease in the Lagrangian function. Thus we see that the Lagrangian function value can play an important role. In this case, we can prove that Fletcher's differentiable exact penalty function is decreasing as well.

Furthermore, Chen et al. [12] proposed a penalty-free trust-region method with the Lagrangian function value without using feasibility restoration phase. Chen et al. [13] presented a line search penalty-free SQP method for equality constrained optimization with the Lagrangian function value. Thus with the use of the Lagrangian function value, one would expect SQP methods to control possible erratic behavior in a better manner and share the rapid convergence of Newton-like methods. More researches are required on the use of Lagrangian function value in SQP methods for general nonlinear optimization.

3.2. Interior-point methods

Interior-point methods have been among the most efficient methods for continuous optimization, see, e.g., Ye [84], Byrd et al. [8], Vanderbei and Shanno [81], Wächter and Biegler [83], Liu and Yuan [59], Curtis [15], and Gould et al. [43]. These methods are iterative and require every iterate to be an interior point. The numerical efficiency and polynomial computational complexity of interior-point methods for linear programming made a lot of researchers to be interested again in interior-point methods for nonlinear optimization.

However, Wächter and Biegler [82] noticed that many line-search interior-point methods for nonlinear optimization may fail to find a feasible point of a single-variable nonlinear and nonconvex problem, even though the problem is well posed. In addition, the algorithmic framework of interior-point methods for nonlinear optimization often includes an inner-loop and an outer-loop, in which the inner-loop finds an approximate solution of a logarithmic barrier subproblem and the outer-loop focuses on the update of the barrier parameter. This framework is distinct from that of interior-point methods for linear programming (which reduces the barrier parameter at each iteration) and is believed to be ineffective sometimes.

Below we shall describe a primal–dual interior-point relaxation method with nice properties. The method was recently introduced in [56].

In order to avoid requiring feasible interior-point iterates, traditional interior-point methods for problem (1.2)–(1.4) introduce slack variables for inequality constraints and solve

the logarithmic barrier subproblem

$$\min \quad f(x) - \mu \sum_{i=1}^{m_I} \ln z_i \tag{3.1}$$

$$\text{such that} \quad h(x) = 0, \tag{3.2}$$

$$g(x) - z = 0, \tag{3.3}$$

where $\mu > 0$ is a barrier parameter and $z_i > 0$ is the ith component of z.

Noting that the subproblem (3.1)–(3.3) is an equality constrained optimization problem, we reformulate it as another constrained optimization by using the Hestenes–Powell augmented Lagrangian:

$$\min_{x,z} \quad f(x) - \mu \sum_{i=1}^{m_I} \ln z_i - v^T \big(g(x) - z\big) + \frac{1}{2}\rho \big\| \big(g(x) - z\big) \big\|^2 \tag{3.4}$$

$$\text{such that} \quad h(x) = 0, \tag{3.5}$$

where $v \in \Re^{m_I}$ is an estimate of the Lagrange multipliers associated with the original inequality constraints, $\mu > 0$ is a barrier parameter, and $\rho > 0$ is a penalty parameter. Since the objective function in (3.4) is strictly convex with respect to z, the unique minimizer of z can be derived with the expression

$$z_i = \frac{1}{2\rho}\Big[\sqrt{\big(v_i - \rho g_i(x)\big)^2 + 4\rho\mu} - \big(v_i - \rho g_i(x)\big)\Big], \quad i = 1, \ldots, m_I. \tag{3.6}$$

The preceding expression depends on the primal variable vector x and the dual variable vector v, thus can be taken as a function of (x, v). For simplicity, corresponding to (3.6), denote

$$y_i = \frac{1}{2\rho}\Big[\sqrt{\big(v_i - \rho g_i(x)\big)^2 + 4\rho\mu} + \big(v_i - \rho g_i(x)\big)\Big], \quad i = 1, \ldots, m_I. \tag{3.7}$$

Substituting (3.6) for z in the objective function in (3.4) and maximizing the derived function with respect to v, since it is a strictly concave function of v, the subproblem (3.4)–(3.5) can be reformulated as

$$\min_x \max_v \quad f(x) - \mu \sum_{i=1}^{m_I} \ln z_i + \frac{1}{2}\rho \|y\|^2 - \frac{1}{2\rho}\|v\|^2 \tag{3.8}$$

$$\text{such that} \quad h(x) = 0, \tag{3.9}$$

where both z and y are real-valued functions of $x \in \Re^n$ and $v \in \Re^{m_I}$ defined by (3.6) and (3.7).

Although the subproblem (3.8)–(3.9) is originated from the logarithmic barrier subproblem (3.1)–(3.3), it is different from the latter in that z_i is not a primal variable but a positive function of primal and dual variables. Thus, the requirement that the primal and dual iterates are interior-points is relieved. Our primal–dual interior-point relaxation method is proposed to solve the subproblem (3.8)–(3.9) approximately. In particular, the barrier and penalty parameters are updated adaptively during the iterative process.

We firstly describe the relation between the logarithmic barrier subproblem (3.1)–(3.3) and its augmented Lagrangian reformulation (3.8)–(3.9).

Theorem 3.2 ([56]). *Suppose $\mu > 0$ and $\rho > 0$. Then $(x^*, v^*) \in \Re^n \times \Re^{m_I}$ is a local solution of the constrained minimax problem (3.8)–(3.9) if and only if x^* is a local solution of the logarithmic-barrier subproblem (3.1)–(3.3) and $g_i(x^*) > 0$, $v_i^* = \mu/g_i(x^*)$ for all $i = 1, \ldots, m_I$.*

Then we show the relation between the original problem (1.2)–(1.4) and the augmented Lagrangian reformulation (3.8)–(3.9).

Theorem 3.3 ([56]). *Given $\rho > 0$. Let z be defined by (3.6). The point (x^*, u^*, v^*) is a KKT triple of the original problem (1.2)–(1.4) if and only if (x^*, u^*, v^*) and μ^* satisfy the system*

$$\mu = 0, \tag{3.10}$$

$$\nabla f(x) - \nabla h(x)u - \nabla g(x)v = 0, \tag{3.11}$$

$$h(x) = 0, \tag{3.12}$$

$$g(x) - z = 0. \tag{3.13}$$

It should be noted that equations (3.11)–(3.13) are the KKT conditions of the subproblem (3.8)–(3.9). Moreover, for all $\mu > 0$ and $i = 1, \ldots, m_I$, both z_i and y_i are twice continuously differentiable with respect to x and v. Thus the subproblem (3.8)–(3.9) can be thought as a *smoothing* problem of the original problem (1.2)–(1.4) in the sense that the system (3.11)–(3.13) is a smoothing system of the KKT conditions of the original problem. Letting the merit function $\phi_{(\mu,\rho)}(x, u, v)$ be the square of l_2 residuals of the system (3.11)–(3.13), the preceding system (3.10)–(3.13) can be further reformulated as the system

$$\mu + \gamma\phi_{(\mu,\rho)}(x, u, v) = 0, \tag{3.14}$$

$$\nabla f(x) - \nabla h(x)u - \nabla g(x)v = 0, \tag{3.15}$$

$$h(x) = 0, \tag{3.16}$$

$$g(x) - z = 0, \tag{3.17}$$

where $\gamma \in (0, 1)$ is a scalar. The two systems (3.10)–(3.13) and (3.14)–(3.17) are equivalent, but the connection between the parameter μ and the KKT residual $\phi_{(\mu,\rho)}(x, u, v)$ is enhanced in (3.14) which requires that μ vanishes with $\phi_{(\mu,\rho)}(x, u, v)$.

Then by sequentially solving the linearized system of the system (3.14)–(3.17) and using the merit function $\phi_{(\mu,\rho)}(x, u, v)$, an efficient primal–dual interior-point relaxation method was provided in [55]. Under suitable assumptions, the new method is proved to have strong global convergence and rapid local convergence [55,56]. In particular, [26] shows that some variant of this method is capable of rapidly detecting the infeasibility of nonlinear optimization. Numerical experiments demonstrate that the new method not only is efficient for well-posed feasible problems, but also is applicable for some feasible problems without LICQ or MFCQ and some infeasible problems.

The new method is robust in the following three aspects. Firstly, the new method does not require any primal or dual iterate to be an interior-point but prompts the iterate to

be an interior-point, which is quite different from most of the globally convergent interior-point methods in the literature. Secondly, the new method uses a single-loop framework and updates the barrier parameter adaptively, which is similar to that of interior-point methods for linear programming. Thirdly, the new method has strong global convergence and is capable of rapidly detecting the infeasibility.

For convex and linear programming, our primal–dual interior point relaxation method provides an intermediate approach between the simplex method and the interior-point method. In addition, we admit the components of $g(x)$ and v to be zero during the iterative process and thus μ can be zero when the solution is obtained. Based on these observations, we may expect our relaxation method to give a solution with high accuracy and to avoid the ill-conditioning phenomenon of interior-point methods, and improve the performance of interior-point methods for large scale problems. Some future topics include its extension for nonlinear semidefinite programming and its complexity when applied for linear programming. An efficient extension of the method has been given for convex quadratic programming [91]. More researches and software-building are expected to go along this line.

3.3. Augmented Lagrangian method of multipliers

The augmented Lagrangian method of multipliers (ALM) was initially proposed by Hestenes [49] and Powell [69] for solving nonlinear optimization with only equality constraints. The ALM minimizes the Hestenes–Powell augmented Lagrangian approximately and circularly with update of multipliers and has been attracting extensive attention in the community. It was generalized by Rockafellar [76] to solve optimization problems with inequality constraints. Many ALMs have been proposed for various optimization problems.

Consider the general nonlinear optimization problem (1.2)–(1.4). By introducing some slack variables z_i $(i = 1, \ldots, m_I)$ for the inequality constraints, the problem can equivalently be transformed to that with general equality constraints and nonnegative constraints

$$\begin{aligned} \min \quad & f(x) && (3.18)\\ \text{such that} \quad & h(x) = 0, && (3.19)\\ & g(x) - z = 0, && (3.20)\\ & z \geq 0. && (3.21) \end{aligned}$$

Using the augmented Lagrangian on equality constraints, problem (3.18)–(3.21) is reformulated as a nonlinear program with only nonnegative constraints:

$$\begin{aligned} \min_{x,z} \quad & f(x) - u^T h(x) - v^T\big(g(x) - z\big) + \frac{1}{2}\rho\big(\big\|h(x)\big\|^2 + \big\|g(x) - z\big\|^2\big) && (3.22)\\ \text{such that} \quad & z \geq 0, && (3.23) \end{aligned}$$

where $u \in \Re^{m_E}$ and $v \in \Re^{m_I}$ are the estimates of Lagrange multipliers and $\rho > 0$ is the penalty parameter. Thanks to the strict convexity of the objective function with respect to z, we may explicitly get the optimal z, yielding an equivalent unconstrained optimization sub-

problem of (3.22)–(3.23),

$$\min_{x} \quad f(x) - u^T h(x) + \frac{1}{2}\rho\|h(x)\|^2 + \sum_{i=1}^{m_I} \phi\big(g_i(x), v_i; \rho\big), \tag{3.24}$$

where $\phi(g_i(x), v_i; \rho)$ equals to $-v_i g_i(x) + \frac{1}{2}\rho g_i(x)^2$ if $g_i(x) \le v_i/\rho$ and $-\frac{1}{2}v_i^2/\rho$ otherwise. Unfortunately, the function ϕ in (3.24) is in general discontinuous in the second derivative with respect to x. Some other unconstrained reformulation is based on the optimization (3.4)–(3.5) in the form

$$\min_{x,z} \quad f(x) - \mu \sum_{i=1}^{m_I} \ln z_i - u^T h(x) + \frac{1}{2}\rho\|h(x)\|^2 - v^T\big(g(x) - z\big) + \frac{1}{2}\rho\big\|\big(g(x) - z\big)\big\|^2, \tag{3.25}$$

where both x and z are primal variables, u and v are dual estimates, and z should be an interior-point.

Originated from solving the augmented Lagrangian reformulation of problem (3.8)–(3.9), the new ALM, proposed by Liu et al. [57], solves the following problem approximately and circularly with update of multipliers u and v:

$$\min_{x} \quad f(x) - u^T h(x) + \frac{1}{2}\rho\|h(x)\|^2 + \sum_{i=1}^{m_I} \psi\big(g_i(x), v_i; \mu, \rho\big), \tag{3.26}$$

where $\psi(g_i(x), v_i; \mu, \rho) = -\mu \ln z_i + \frac{1}{2}\rho y_i^2 - \frac{1}{2\rho}v_i^2$ and z_i and y_i are defined by (3.6) and (3.7). A detailed description of the new ALM is given in Algorithm 1. It is a generalization of the classical Hestenes–Powell augmented Lagrangian and a combination of the augmented Lagrangian and the interior-point technique.

Algorithm 1: A new ALM for problem (1.2)–(1.4) [57]

1 Given (x_0, u_0, v_0), and μ_0, ρ_0. Let $k := 0$.
2 **while** $\mu_k > \epsilon$ *or* $\phi_{(\mu_k,\rho_k)}(x_k, u_k, v_k) > \epsilon$ **do**
3 Compute x_{k+1} to be an approximate solution of problem (3.26) with the initial point x_k.
4 Update u_k by $u_{k+1} = u_k - \rho_k h(x_k)$.
5 Update v_k by $v_{k+1} = \rho_k y(x_{k+1}, v_k; \mu_k, \rho_k)$.
6 Update $\rho_{k+1} \ge 2\rho_k$ if $\|z(x_{k+1}, v_{k+1}; \mu_k, \rho_k) - c(x_{k+1})\|$ is not small.
7 Update $\mu_{k+1} \le 0.5\mu_k$ if $\|z(x_{k+1}, v_{k+1}; \mu_k, \rho_k) - c(x_{k+1})\|$ is small.
8 Let $k := k + 1$.
9 **end**

Liu et al. [57] proved that the new ALM is of strong global convergence, rapid infeasibility detection, and shares the same convergence rate to the KKT point as the Hestenes–Powell augmented Lagrangian for optimization with equality constraints.

Although the subproblem (3.26) is similar to the augmented Lagrangian counterpart (3.24) and the interior-point counterpart (3.25) in appearance that all of them are unconstrained optimization and first-order smooth, but it is essentially distinct from the latter two subproblems in the following aspects.

Firstly, the function ψ in (3.26) has one more parameter μ than ϕ in (3.24) and is always twice continuously differentiable with respect to x provided g is twice continuously differentiable and v holds fixed, while ϕ in (3.24) has discontinuous second derivative with respect to x. The problem (3.25) has the same property with (3.26). Secondly, the subproblems (3.24) and (3.26) are convex if the original problem (1.2)–(1.4) is convex, while the subproblem (3.25) can be nonconvex even though the original problem is convex. Thirdly, unlike subproblem (3.25), the subproblems (3.24) and (3.26) do not require any primal or dual variable to be positive. Moreover, $\psi(g_i(x), v_i; \mu, \rho)$ is well defined for every $x \in \Re^n$ and $v \in \Re^{m_I}$, while (3.25) requests $z > 0$ and $v > 0$.

To summarize, the new ALM can deal with optimization problems with inequality constraints and shares the same convergence rate to the KKT point as the Hestenes–Powell augmented Lagrangian for optimization problems with equality constraints. As the new ALM has nice properties, more researches are expected along this line.

4. OPTIMIZATION WITH LEAST CONSTRAINT VIOLATION

The theory and algorithms for constrained optimization usually assume the feasibility of the optimization problem. If the constraints are inconsistent, several numerical algorithms have been proposed to find infeasible stationary points, which have nothing to do with the objective function; see, e.g., Byrd et al. [7], Burke et al. [6], and Dai et al. [26]. However, there are important optimization problems, which may be either feasible or infeasible and whose objective function is wished to be minimized with the least constraint violation even if they are infeasible. A typical example comes from rocket trajectory optimal control, where the fuel is minimized with the aim of landing at a target point and subjected to other constraints. If landing at the target is not possible, we might wish to minimize the distance between the real landing point and the target and thereafter optimize the required fuel. Hence we are led to optimization problems with least constraint violation.

For optimization with possible inconsistent constraints, we prove that the minimization problem with least constraint violation is equivalent to a Lipschitz equality constrained optimization problem. To this aim, consider the nonlinear optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{such that} \quad & Ax = b, \\ & g_i(x) \geq 0,\ i = 1, \ldots, p, \end{aligned} \tag{4.1}$$

where $f : \Re^n \to \Re$ is smooth and g_i $(i = 1, \ldots, p)$ are differentiable concave functions. In this case, the optimization problem with the least constraint violation can be expressed as

$$\begin{aligned} \min \quad & f(x) \\ \text{such that} \quad & A^T(Ax - b) + \mathcal{J}g(x)^T\big[g(x)\big]_- = 0. \end{aligned} \tag{4.2}$$

Define $H(x, y) = A^T A - \sum_{j=1}^{p} y_j \nabla^2 g_j(x)$. For $y^* = [-g(x^*)]_+$ and $z^* = [g(x^*)]_+$, define $\alpha^* = \{i : y_i^* > 0\}$, $\beta^* = \{i : y_i^* = z_i^* = 0\}$, $\gamma^* = \{i : z_i^* > 0\}$. Then we are able to give an elegant necessary optimality condition from the classical optimality theory of Lipschitz continuous optimization.

Theorem 4.1 ([31]). *Let (x^*, y^*) be a local minimizer of problem (4.2). Suppose that the matrix $H(x^*, y^*) + \mathcal{J}g_{\alpha^*}(x^*)^T \mathcal{J}g_{\alpha^*}(x^*)$ is positive definite. Then there exist $\lambda^* \in \Re^n$ and $[v_b]_{\beta^*} \in \Re^{|\beta^*|}$ satisfying $[v_b]_i \in [0, 1]$, $i \in \beta^*$ such that*

$$\begin{aligned} \nabla f(x^*) + \big[H(x^*, y^*) + \mathcal{J}g_{\alpha^*}(x^*)^T \mathcal{J}g_{\alpha^*}(x^*) \\ + \mathcal{J}g_{\beta^*}(x^*)^T \mathrm{Diag}\big([v_b]_{\beta^*}\big)\mathcal{J}g_{\beta^*}(x^*)\big]\lambda^* = 0. \end{aligned} \tag{4.3}$$

Dai and Zhang [31] found that the penalty method can be used for solving optimization problems with least constraint violation. Chiche and Gilbert [14] proved that the augmented Lagrangian method of multipliers (ALM) can deal with an infeasible convex quadratic optimization problem. Is the ALM still valid for general convex optimization with the least constraint violation?

To this aim, consider the following convex constrained optimization problem:

$$\text{(P)} \qquad \begin{aligned} \min \quad & f(x) \\ \text{such that} \quad & g(x) \in \mathcal{K}, \end{aligned} \tag{4.4}$$

where $f : \Re^n \to \Re$, $g : \Re^n \to \mathcal{Y}$, $\mathcal{K} \subset \mathcal{Y}$ is a nonempty closed convex set, and $\mathcal{Y}$ is a finite-dimensional Hilbert space. We analyze the dual of the problem with the least constraint violation. By introducing a vector $y \in \mathcal{Y}$, problem (4.4) is equivalently expressed as

$$\begin{aligned} \min \quad & f(x) \\ \text{such that} \quad & g(x) = y, \\ & y \in \mathcal{K}. \end{aligned} \tag{4.5}$$

For a given $s \in \mathcal{Y}$, the shifted problem is defined as

$$\text{P}(s) \qquad \begin{aligned} \min \quad & f(x) \\ \text{such that} \quad & g(x) + s \in \mathcal{K}. \end{aligned} \tag{4.6}$$

Here we call s a shift. The set of feasible shifts, denoted as $\mathcal{S}$, is defined by

$$\mathcal{S} := \big\{s \in \mathcal{Y} : \text{there exists some } x \in \Re^n \text{ such that } g(x) + s \in \mathcal{K}\big\}. \tag{4.7}$$

Define the smallest norm shift by $\bar{s} = \arg\min\{\frac{1}{2}\|s\|^2 : s \in \mathcal{S}\}$. If $\mathcal{S}$ is closed, then $\bar{s}$ can be achieved, i.e., $\bar{s} \in \mathcal{S}$. In this case, the optimization problem with the least constraint violation is expressed as follows:

$$\text{P}(\bar{s}) \qquad \begin{aligned} \min \quad & f(x) \\ \text{such that} \quad & g(x) + \bar{s} \in \mathcal{K}. \end{aligned} \tag{4.8}$$

Now we shall present the properties of the ALM for problem (4.5), which was provided by Dai and Zhang [32]. The Lagrangian of problem (4.5), denoted by l, is defined by $l(x, y, \lambda) = f(x) + \lambda^T(g(x) - y)$. The augmented Lagrangian function of problem (4.5), denoted by l_r, is defined by

$$l_r(x, y, \lambda) = f(x) + \lambda^T\big(g(x) - y\big) + \frac{r}{2}\big\|g(x) - y\big\|^2. \tag{4.9}$$

The dual function $\theta : \mathcal{Y} \to \overline{\Re}$ associated with problem (4.5) is

$$\theta(\lambda) := -\inf_{x\in\Re^n, y\in K} l(x, y, \lambda). \tag{4.10}$$

Denote by D and D(s) the conjugate dual problems of P and P(s), respectively. Then problems D and D(s) are expressed as follows:

$$\text{(D)} \quad \max_{\lambda}\big[-\theta(\lambda)\big], \quad \big(\text{D}(s)\big) \quad \max_{\lambda}\big[s^T\lambda - \theta(\lambda)\big]. \tag{4.11}$$

The following proposition reveals that the solution set of the dual problem, if nonempty, is unbounded when $\bar{s} \neq 0$.

Proposition 4.1 ([32]). *Assume that* $\bar{s} \neq 0$, $\operatorname{val}\text{P}(\bar{s}) \in \Re$, v *is lower semicontinuous at* $\bar{s}$ *and* $\operatorname{Sol}\text{D}(\bar{s}) \neq \emptyset$. *Then* $\operatorname{Sol}\text{D}(\bar{s})$ *is unbounded with* $-\bar{s} \in [\operatorname{Sol}\text{D}(\bar{s})]^\infty$.

For the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the ALM for solving problem (4.5), defining $s^k = y^k - g(x^k)$, we are able to prove the following theorem.

Theorem 4.2 ([32]). *Assume that* $\bar{s} \neq 0$, $\operatorname{val}\text{P}(\bar{s}) \in \Re$, v *is lower semicontinuous at* $\bar{s}$ *and* $\operatorname{Sol}\text{D}(\bar{s}) \neq \emptyset$. *Assume also that* $\{r_k\}$ *has a positive lower bound and* $\{(x^k, y^k)\}$ *has an accumulation point. Then we have that* (i) $s^k \to \bar{s}$; (ii) $\{\lambda^k\}$ *diverges;* (iii) *for every* $\varepsilon > 0$, *there exists an index* k *large enough such that* (x^k, y^k) *satisfies* ε*-approximate optimality conditions of problem* P($\bar{s}$) *in terms of the augmented Lagrangian.*

The above theorem shows that the ALM can deal with convex optimization with least constraint violation. Studies on the theory and algorithms for optimization with least constraint violation are clearly required.

5. SOME DISCUSSIONS

Due to limited space, this article only reviewed some numerical methods for general nonlinear optimization. An early good review on unconstrained optimization is given by Nocedal [64], where two open questions about quasi-Newton methods were summarized. One is whether the DFP method with the Wolfe line search converges for uniformly convex functions. The other is whether the BFGS method with the Wolfe line search converges for general nonlinear functions. A negative answer of the second open question has been known (see, e.g., Dai [18]). Although Yuan [86] made a significant progress on the first open question, we do not know its answer, yet. The infimum of the Q-order of the convergence of quasi-Newton methods is only one [85]. The work of Rodomanov and Nesterov [75] stimulated research interests on this topic again.

Previous studies for constrained optimization usually assumed the feasibility of the optimization problem. This article described optimality conditions for optimization with least constraint violation and the ability of the ALM method to deal with such a problem. Independently of the work [31], Censor et al. [10] proposed a data-compatibility approach for the problem and presented some theoretical analysis. However, more researches are clearly required along this line.

The development of nonlinear optimization influences many research directions in optimization such as matrix optimization, sparse optimization, and nonsmooth optimization. There is still much to do in extending nonlinear optimization methods for minimax optimization, which arises from both modern machine learning and tradition research areas.

ACKNOWLEDGMENTS

The author is very grateful to Professor Ya-xiang Yuan for his long-time guidance and help, and to his collaborators for their valuable discussions and cooperations.

FUNDING

This work was partially supported by the NSFC grants (nos. 12021001, 11991021, 11991020, 11631013, and 11971372) and the Strategic Priority Research Program of Chinese Academy of Sciences (no. XDA27000000).

REFERENCES

[1] H. Akaike, On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Statist. Math.* **11** (1959), no. 1, 1–16.

[2] L. Armijo, Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific J. Math.* **16** (1966), no. 1, 1–3.

[3] J. Barzilai and J. M. Borwein, Two-point step size gradient methods. *IMA J. Numer. Anal.* **8** (1988), no. 1, 141–148.

[4] E. G. Birgin, J. M. Martínez, and M. Raydan, Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim.* **10** (2000), no. 4, 1196–1211.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** (2011), no. 1, 1–122.

[6] J. V. Burke, F. E. Curtis, and H. Wang, A sequential quadratic optimization algorithm with rapid infeasibility detection. *SIAM J. Optim.* **24** (2014), no. 2, 839–872.

[7] R. H. Byrd, F. E. Curtis, and J. Nocedal, Infeasibility detection and SQP methods for nonlinear optimization. *SIAM J. Optim.* **20** (2010), no. 5, 2281–2299.

[8] R. H. Byrd, M. E. Hribar, and J. Nocedal, An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optim.* **9** (1999), no. 4, 877–900.

[9] A. Cauchy, Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris* **25** (1847), 536–538.

[10] Y. Censor, M. Zaknoon, and A. J. Zaslavski, Data-compatibility of algorithms. 2019, arXiv:1911.11389.

[11] R. Chamberlain, M. J. D. Powell, C. Lemarechal, and H. Pedersen, The watchdog technique for forcing convergence in algorithms for constrained optimization. In *Algorithms for constrained minimization of smooth nonlinear functions*, edited by A. G. Buckley and J. Goffin, pp. 1–17, Springer, 1982.

[12] Z. W. Chen, Y. H. Dai, and J. Y. Liu, A penalty-free method with superlinear convergence for equality constrained optimization. *Comput. Optim. Appl.* **76** (2020), no. 3, 801–833.

[13] Z. W. Chen, Y. H. Dai, and T. Y. Zhang, *A line search penalty-free SQP method for equality constrained optimization without Maratos effect*. Tech. rep., AMSS, Chinese Academy of Sciences, Beijing, China, 2021.

[14] A. Chiche and J. C. Gilbert, How the augmented Lagrangian algorithm can deal with an infeasible convex quadratic optimization problem. *J. Convex Anal.* **23** (2016), no. 2, 425–459.

[15] F. E. Curtis, A penalty-interior-point algorithm for nonlinear constrained optimization. *Math. Program. Comput.* **4** (2012), no. 2, 181–209.

[16] Y. H. Dai, A nonmonotone conjugate gradient algorithm for unconstrained optimization. *J. Syst. Sci. Complex.* **15** (2002), 139–145.

[17] Y. H. Dai, Convergence analysis of nonlinear conjugate gradient methods. In *Optimization and regularization for computational inverse problems and applications*, edited by Y. Wang, C. Yang, and A. G. Yagola, pp. 157–181, Springer, 2010.

[18] Y. H. Dai, A perfect example for the BFGS method. *Math. Program.* **138** (2013), no. 1, 501–530.

[19] Y. H. Dai and R. Fletcher, On the asymptotic behaviour of some new gradient methods. *Math. Program.* **13** (2005), no. 3, 541–559.

[20] Y. H. Dai and R. Fletcher, New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds. *Math. Program.* **106** (2006), no. 3, 403–421.

[21] Y. H. Dai, J. Han, G. Liu, D. Sun, H. Yin, and Y. X. Yuan, Convergence properties of nonlinear conjugate gradient methods. *SIAM J. Optim.* **10** (1999), no. 2, 345–358.

[22] Y. H. Dai and C. X. Kou, A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. *SIAM J. Optim.* **23** (2013), no. 1, 296–320.

[23] Y. H. Dai and C. X. Kou, A Barzilai–Borwein conjugate gradient method. *Sci. China Math.* **59** (2016), no. 8, 1511–1524.

[24] Y. H. Dai and L. Z. Liao, New conjugacy conditions and related nonlinear conjugate gradient methods. *Appl. Math. Optim.* **43** (2001), no. 1, 87–101.

[25] Y. H. Dai and L. Z. Liao, *R*-linear convergence of the Barzilai and Borwein gradient method. *IMA J. Numer. Anal.* **22** (2002), no. 1, 1–10.

[26] Y. H. Dai, X. W. Liu, and J. Sun, A primal–dual interior-point method capable of rapidly detecting infeasibility for nonlinear programs. *J. Ind. Manag. Optim.* **16** (2020), no. 2, 1009–1035.

[27] Y. H. Dai and Y. X. Yuan, A nonlinear conjugate gradient with a strong global convergence property. *SIAM J. Optim.* **10** (1999), no. 1, 177–182.

[28] Y. H. Dai and Y. X. Yuan, *Nonlinear conjugate gradient methods*. Shanghai Scientific & Technical Publishers, Shanghai, 2000 (in Chinese).

[29] Y. H. Dai and Y. X. Yuan, An efficient hybrid conjugate gradient method for unconstrained optimization. *Ann. Oper. Res.* **103** (2001), no. 1, 33–47.

[30] Y. H. Dai and Y. X. Yuan, Analysis of monotone gradient methods. *J. Ind. Manag. Optim.* **1** (2005), no. 2, 181–192.

[31] Y. H. Dai and L. W. Zhang, Optimization with least constraint violation. *CSIAM Trans. Appl. Math.* **2** (2021), no. 3, 551–584.

[32] Y. H. Dai and L. W. Zhang, *The augmented Lagrangian method can approximately solve convex optimization with least constraint violation*. Tech. rep., AMSS, Chinese Academy of Sciences, Beijing, China, 2021.

[33] W. C. Davidon, Variable metric methods for minimization. *SIAM J. Optim.* **1** (1991), no. 1, 1–17.

[34] B. C. Fabien, Implementation of a robust SQP algorithm. *Optim. Methods Softw.* **23** (2008), no. 6, 827–846.

[35] R. Fletcher, *Practical methods of optimization*, 2nd edition. John Wiley, Chichester, 1987.

[36] R. Fletcher, On the Barzilai–Borwein method. In *Optimization and control with applications*, edited by L. Qi, K. Teo, and X. Yang, pp. 235–256, Springer, 2005.

[37] R. Fletcher and S. Leyffer, Nonlinear programming without a penalty function. *Math. Program.* **91** (2002), no. 2, 239–269.

[38] R. Fletcher and M. J. D. Powell, A rapidly convergent descent method for minimization. *Comput. J.* **6** (1963), no. 2, 163–168.

[39] R. Fletcher and C. M. Reeves, Function minimization by conjugate gradients. *Comput. J.* **7** (1964), no. 2, 149–154.

[40] G. E. Forsythe, On the asymptotic directions of the s-dimensional optimum gradient method. *Numer. Math.* **11** (1968), no. 1, 57–76.

[41] M. Fukushima, A successive quadratic programming algorithm with global and superlinear convergence properties. *Math. Program.* **35** (1986), no. 3, 253–264.

[42] J. C. Gilbert and J. Nocedal, Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optim.* **2** (1992), no. 1, 21–42.

[43] N. I. Gould, D. Orban and P. L. Toint, An interior-point ℓ_1-penalty method for nonlinear optimization. In *Numerical analysis and optimization*, edited by M. Al-Baali, L. Grandinetti, A. Purnama, Springer Proc. Math. Stat. 134, Springer, Berlin, Cham, 2015.

[44] N. I. Gould and P. L. Toint, Nonlinear programming without a penalty function or a filter. *Math. Program.* **122** (2010), no. 1, 155–196.

[45] L. Grippo, F. Lampariello, and S. Lucidi, A nonmonotone line search technique for Newton's method. *SIAM J. Numer. Anal.* **23** (1986), no. 1, 707–716.

[46] W. W. Hager and H. Zhang, A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Optim.* **16** (2005), no. 1, 170–192.

[47] W. W. Hager and H. C. Zhang, Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent. *ACM Trans. Math. Software* **32** (2006), no. 1, 113–137.

[48] W. W. Hager and H. C. Zhang, The limited memory conjugate gradient method. *SIAM J. Optim.* **23** (2013), no. 4, 2150–2168.

[49] M. R. Hestenes, Multiplier and gradient methods. *J. Optim. Theory Appl.* **4** (1969), no. 5, 303–320.

[50] M. R. Hestenes and E. L. Stiefel, Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.* **49** (1952), no. 6, 409–436.

[51] Y. K. Huang, Y. H. Dai, and X. W. Liu, Equipping the Barzilai–Borwein method with the two dimensional quadratic termination property. *SIAM J. Optim.* **31** (2021), no. 4, 3068–3096.

[52] W. Karush, *Minima of functions of several variables with inequalities as side constraints*. M. Sc. thesis, University of Chicago, 1939.

[53] H. W. Kuhn and A. Tucker, Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, edited by J. Neyman, pp. 481–492, University of California Press, Berkeley, CA, 1951.

[54] Y. L. Lai, Some properties of the steepest descent method. *Acta Math. Appl. Sin.* **4** (1981), no. 2, 106–116 (in Chinese).

[55] X. W. Liu and Y. H. Dai, A globally convergent primal-dual interior-point relaxation method for nonlinear programs. *Math. Comp.* **89** (2020), no. 323, 1301–1329.

[56] X. W. Liu, Y. H. Dai, and Y. K. Huang, A primal–dual interior-point relaxation method with adaptively updating barrier for nonlinear programs. 2020, arXiv:2007.10803.

[57] X. W. Liu, Y. H. Dai, Y. K. Huang, and J. Sun, A novel augmented Lagrangian method of multipliers for optimization with general inequality constraints. 2021, arXiv:2106.15044.

[58] X. W. Liu and Y. X. Yuan, A robust algorithm for optimization with general equality and inequality constraints. *SIAM J. Sci. Comput.* **22** (2000), no. 2, 517–534.

[59] X. W. Liu and Y. X. Yuan, A null-space primal-dual interior-point algorithm for nonlinear optimization with nice convergence properties. *Math. Program.* **123** (2010), no. 1, 163–193.

[60] Z. X. Liu, H. W. Liu, and Y. H. Dai, An improved Dai–Kou conjugate gradient algorithm for unconstrained optimization. *Comput. Optim. Appl.* **75** (2020), no. 1, 145–167.

[61] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*. PhD Thesis, University of London, 1978.

[62] D. Q. Mayne and E. Polak, A surperlinearly convergent algorithm for constrained optimization problems. In *Algorithms for constrained minimization of smooth nonlinear functions*, edited by A. G. Buckley and J. L. Goffin, pp. 45–61, Springer, 1982.

[63] M. Mu, H. Xu, and W. Duan, A kind of initial errors related to "spring predictability barrier" for El Ninõ events in Zebiak–Cane mode. *Geophys. Res. Lett.* **34** (2007), L03709.

[64] J. Nocedal, Theory of algorithms for unconstrained optimization. *Acta Numer.* **1** (1991), 199–242.

[65] J. Nocedal and S. Wright, *Numerical optimization*. Springer, New York, 2006.

[66] A. Perry, *A class of conjugate gradient algorithms with a two-step variable metric memory*. Tech. rep., Northwestern University, Center for Mathematical Studies in Economics and Management Science, 1977.

[67] E. Polak and G. Ribiere, Note sur la convergence de méthodes de directions conjuguées. *ESAIM Math. Model. Numer. Anal.* **3** (1969), no. R1, 35–43.

[68] B. T. Polyak, The conjugate gradient method in extreme problems. *USSR Comput. Math. Math. Phys.* **9** (1969), no. 4, 94–112.

[69] M. J. D. Powell, A method for nonlinear constraints in minimization problems. In *Optimization*, edited by R. Fletcher, pp. 283–298, Academic Press, London, 1969.

[70] M. J. D. Powell, A fast algorithm for nonlinearly constrained optimization calculations. In *Numerical analysis*, edited by G. A. Watson, pp. 144–157, Springer, Berlin, 1977.

[71] M. J. D. Powell, Nonconvex minimization calculations and the conjugate gradient method. In *Numerical analysis*, edited by D. F. Griffiths, pp. 122–141, Lecture Notes in Math. 1066, Springer, Berlin, Heidelberg, 1984.

[72] M. J. D. Powell and Y. X. Yuan, A trust region algorithm for equality constrained optimization. *Math. Program.* **49** (1991), no. 1, 1189–211.

[73] M. Raydan, On the Barzilai and Borwein choice of steplength for the gradient method. *IMA J. Numer. Anal.* **13** (1993), no. 3, 321–326.

[74] M. Raydan, The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM J. Optim.* **7** (1997), no. 1, 26–33.

[75] A. Rodomanov and Y. Nesterov, Greedy quasi-Newton methods with explicit superlinear convergence. *SIAM J. Optim.* **31** (2021), no. 1, 785–811.

[76] R. T. Rockafellar, A dual approach to solving nonlinear programming problems by unconstrained optimization. *Math. Program.* **5** (1973), no. 1, 354–373.

[77] D. F. Shanno, On the convergence of a new conjugate gradient algorithm. *SIAM J. Numer. Anal.* **15** (1978), no. 6, 1247–1257.

[78] P. Spellucct, A new technique for inconsistent QP problems in the SQP method. *Math. Methods Oper. Res.* **47** (1998), no. 3, 355–400.

[79] D. Touati-Ahmed and C. Storey, Efficient hybrid conjugate gradient techniques. *J. Optim. Theory Appl.* **64** (1990), no. 2, 379–397.

[80] S. Ulbrich, On the superlinear local convergence of a filter-SQP method. *Math. Program.* **100** (2004), no. 1, 217–245.

[81] R. J. Vanderbei and D. F. Shanno, An interior-point algorithm for nonconvex nonlinear programming. *Comput. Optim. Appl.* **13** (1999), no. 1, 231–252.

[82] A. Wächter and L. T. Biegler, Failure of global convergence for a class of interior point methods for nonlinear programming. *Math. Program.* **88** (2000), no. 3, 565–574.

[83] A. Wächter and L. T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.* **106** (2006), no. 1, 25–57.

[84] Y. Ye, *Interior-point algorithm: theory and analysis*. Wiley & Sons, New York, 1997.

[85] Y. X. Yuan, On the least Q-order of convergence of variable metric algorithms. *IMA J. Numer. Anal.* **4** (1984), no. 2, 233–239.

[86] Y. X. Yuan, Convergence of DFP algorithm. *Sci. China Ser. A* **38** (1995), no. 11, 1281–1294.

[87] Y. X. Yuan, A new stepsize for the steepest descent method. *J. Comput. Math.* **24** (2006), no. 2, 149–156.

[88] Y. X. Yuan and J. Stoer, A subspace study on conjugate gradient algorithms. *Z. Angew. Math. Mech.* **75** (1995), no. 1, 69–77.

[89] Y. X. Yuan and W. Y. Sun, *Optimization theories and methods*. Science Press, Beijing, 1997 (in Chinese).

[90] L. Zanni, T. Serafini, G. Zanghirati, K. P. Bennett, and E. Parrado-Hernández, Parallel software for training large scale support vector machines on multiprocessor systems. *J. Mach. Learn. Res.* **7** (2006), no. 54, 1467–1492.

[91] R. Zhang, X. W. Liu, and Y. H. Dai, *Solving convex quadratic programming by a primal-dual interior-point relaxation method*. Tech. rep., AMSS, Chinese Academy of Sciences, Beijing, China, 2021.

[92] B. Zhou, L. Gao, and Y. H. Dai, Gradient methods with adaptive step-sizes. *Comput. Optim. Appl.* **35** (2006), no. 1, 69–86.

YU-HONG DAI

AMSS, Chinese Academy of Sciences, Beijing 100190, China, and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, dyh@lsec.cc.ac.cn

CONTROL THEORY OF STOCHASTIC DISTRIBUTED PARAMETER SYSTEMS: RECENT PROGRESS AND OPEN PROBLEMS

QI LÜ

ABSTRACT

In recent years, important progresses have been made in the control theory for stochastic distributed parameter control systems (SDPSs for short). However, the theory is far from being complete. The primary difficulty is that many effective tools and methods for deterministic distributed parameter control systems and stochastic finite-dimensional control systems do not work anymore for SDPSs. One has to develop new mathematical tools, such as stochastic transposition method and stochastic Carleman estimate, even for some very simple SDPSs. The objectives of this paper are to provide some new results, to show some new phenomena, to explain the new difficulties, and to present some new methods for the control theory of SDPSs. We mainly focus on our works for the controllability for stochastic hyperbolic equations, and the Pontryagin-type maximum principle for controlled stochastic evolution equations. At last, a number of open questions and future directions of research are given.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 60H15; Secondary 93B05, 93E20, 49K27

KEYWORDS

Stochastic distributed parameter control systems, controllability, optimal control, Pontryagin type maximum principle

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5314–5338
DOI 10.4171/ICM2022/57

1. INTRODUCTION

Control theory was founded by N. Wiener in 1948. It is an interdiscipline among mathematics, engineering, and computer science. The early works in this field were mainly concerned with deterministic finite-dimensional control systems. Motivated by applications, numerous mathematicians and engineers put great effort to study control theory for more complex systems, such as distributed parameter control systems (typically governed by partial differential equations), stochastic finite-dimensional control systems (governed by stochastic differential equations), and SDPSs (typically governed by stochastic partial differential equations, SPDEs for short). These studies provide a rich source of complex mathematical problems, which have fundamental impact on the development of many areas in mathematics.

It is very surprising that the control theory for SDPSs is still in its infancy though it has been studied for around 60 years. Compared with other directions in mathematical control theory (including control theory for deterministic and stochastic finite-dimensional systems and that for distributed parameter systems), many aspects of control theory for SDPSs are much less understood or even still unknown. Nevertheless, one cannot, by no means, ignore its importance. On the one hand, the world is full of uncertainties. They enter the system through noise in sensing/actuation, external disturbances affecting the underlying system, and uncertain dynamics in the system (parameter errors, unmodeled effects, etc.). For lots of significant physical and biological systems, these uncertainties cannot be ignored, and the systems should be governed by SPDEs (e.g., [19]). This leads to a major requirement for the study of the control theory of SDPSs (e.g., [11, 21, 41]). On the other hand, control theory for deterministic finite-dimensional control systems is relatively mature now, and there is a huge list of publications for distributed parameter control systems and stochastic finite-dimensional control systems. The study of SDPSs is a natural development of the mathematical control theory. Then, what slows the pace of the control theory of SDPSs? In my opinion, it lies in the fact that the complexity of SDPSs introduces extreme difficulties. Firstly, the formulation of the control problems for SDPSs may differ from those for distributed parameter control systems or stochastic finite-dimensional control systems. Secondly, many powerful methods and tools developed for the latter two systems mentioned above cannot work for SDPSs. Thirdly, people know very little about SPDEs although much progress has been made in recent years. As a result, new notions and mathematical tools are required, even for some very simple SDPSs. We will demonstrate this by illustrative examples in Sections 2 and 3.

The most fundamental problem in control theory is to modify the behavior of the system by means of suitable "control" actions in an "optimal" way. This leads to the formation of *controllability* and *optimal control problems*. Roughly speaking, controllability involves finding one way to steer the state of the system to a desired target from a given starting point. Optimal control concerns finding the "best way," according to a given cost criterion, to achieve the desired goal. In this paper, we mainly focus on some recent progress on these two topics for SDPSs. We do not attempt to cover the whole field of these topics,

which is virtually hopeless. Rather, with admitted bias, we choose subjects that are undergoing rapid change and require new approaches to meet the challenges and opportunities. No attempt will be made to provide an exhaustive list of all the papers in the corresponding topics, which would only tend to make the narrative very disjoint.

Although we will deal with SDPSs, it is helpful to introduce some fundamental ideas in a simpler setting, i.e., for finite-dimensional deterministic control systems. It can also help readers see the essential differences between the deterministic and stochastic problems.

Let $T > 0$. Consider the following control system:

$$\begin{cases} y_t(t) = Ay(t) + Bu(t), \quad \text{a.e. } t \in [0, T], \\ y(0) = y_0, \end{cases} \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ $(n, m \in \mathbb{N})$, y is the state, and $u \in L^2(0, T; \mathbb{R}^m)$ is the control.

Definition 1.1. The control system (1.1) is called *exactly controllable* at time T if for any $y_0, y_1 \in \mathbb{R}^n$, there is a control $u \in L^2(0, T; \mathbb{R}^m)$ such that the corresponding state y to (1.1) satisfies $y(T) = y_1$.

Remark 1.1. Definition 1.1 can be easily extended to more general control systems, for which the requirement $y(T) = y_1$ may be too restrictive and has to be relaxed. This leads to the notions of approximate/null/partial controllability, and so on.

The exact controllability problem of (1.1) can be regarded as a two-point boundary value problem. However, it is clearly ill-posed and cannot be solved by the classical well-posedness theory of ODEs. To study it, people introduce the adjoint equation of (1.1):[1]

$$\begin{cases} z_t(t) = -A^\top z(t), \quad t \in [0, T], \\ z(T) = z_T \in \mathbb{R}^n, \end{cases} \tag{1.2}$$

and prove the following result:

Theorem 1.1. *The system* (1.1) *is exactly controllable at time* T *if and only if solutions to* (1.2) *satisfy*

$$|z_T|^2_{\mathbb{R}^n} \leq \mathcal{C} \int_0^T \left|B^\top z(t)\right|^2_{\mathbb{R}^m} dt, \quad \forall z_T \in \mathbb{R}^n. \tag{1.3}$$

Here and henceforth, unless otherwise stated, we shall write $\mathcal{C}$ for a generic positive constant, which may vary from one place to another.

Remark 1.2. The inequality (1.3) is called an *observability estimate* for (1.2). Roughly speaking, it concerns whether the solution of (1.3) can be fully determined from the observation $B^\top z(t)$, $t \in [0, T]$. Usually, $B^\top$ is not of full row rank. Hence, one cannot solve for z_T from $B^\top z_T$ directly. In such a case, we do our observation on a time interval $[0, T]$. Besides the connection with controllability, observability has its own interest in control theory.

---

1 For any matrix D, denote by $D^\top$ the transpose of D.

Remark 1.3. Whether inequality (1.3) holds or not depends on A and B, where A decides the type of the control system and B reflects the way we control the system. A sufficient and necessary condition for (1.3) is that (A, B) fulfills the Kalman rank condition (e.g., [18]).

By Theorem 1.1, the controllability problem of (1.1) is reduced to an *a priori estimate* of its adjoint equation. This idea is greatly extended to different kinds of control systems. Most of the controllability results for linear control systems are proved by establishing suitable observability estimates for their adjoint equations (e.g., [16, 21, 40, 49, 50]). However, it is much more complicated to study the controllability problems for SDPSs. Indeed, as we will explain in Section 2, we have to handle the observability for backward SPDEs. Moreover, since one may put controls on both drift and diffusion terms in SDPSs (as we shall see in Section 2, sometimes it is necessary to introduce controls in such a way), the controls will affect each other. Further, compared with distributed parameter control systems, some new and unexpected phenomena are found for controllability problems of SDPSs:

(1) One may need stronger conditions to get the approximate controllability than the null controllability for SDPSs (e.g., [26]).

(2) Two controls are needed to get the exact controllability of stochastic Schrödinger equations and stochastic transport equations (e.g., [27, 29]).

(3) The approximate/null controllability may be sensitive with respect to small perturbations of lower order terms (e.g., [8, 23]).

(4) To get the exact controllability, the control may be very irregular (e.g., [31]).

(5) The reachable set is very "small" if there is no control in the diffusion term (e.g., [47]).

(6) A stochastic hyperbolic equation is not exactly controllable with controls acting on the whole domain where the equation evolves on (e.g., [37]).

Generally speaking, the controllability properties for different SDPSs are drastically different. Consequently, when studying controllability problems of SDPSs, we should consider concrete models of SDPSs. There are two prototypical equations needed to be understood first: the stochastic hyperbolic equation and the stochastic parabolic equation. Due to the limitation of space, we will focus on the former which possesses sufficient complexity to permit exposition of a wide variety of interesting questions and differs from the controllability of deterministic hyperbolic equations essentially. Readers are referred to [23, 26, 40, 45] and the references therein for controllability of the latter equation.

Next, we present a typical optimal control problem. Fix a suitable function $f : [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and a nonempty subset U of $\mathbb{R}^m$. Let

$$\mathcal{U} \triangleq \big\{u : [0, T] \to U \mid u \text{ is Lebesgue measurable}\big\}.$$

Consider the following control system:

$$\begin{cases} y_t(t) = f\big(t, y(t), u(t)\big), \quad \text{a.e. } t \in [0, T], \\ y(0) = y_0, \end{cases} \tag{1.4}$$

with a cost functional

$$\mathcal{J}(u) = \int_0^T g\big(t, y(t), u(t)\big)dt + h\big(y(T)\big), \quad u(\cdot) \in \mathcal{U}. \tag{1.5}$$

Here $y_0 \in \mathbb{R}^n$, y is the state, and u is the control, valued in $\mathbb{R}^n$ and U, respectively; g and h are suitable functions. The optimal control problem is as follows:

Problem (DOP). Find a $\bar{u} \in \mathcal{U}$ such that

$$\mathcal{J}(\bar{u}) = \inf_{u \in \mathcal{U}} \mathcal{J}(u). \tag{1.6}$$

Any control $\bar{u} \in \mathcal{U}$ satisfying (1.6) is called an *optimal control*, and the corresponding state, denoted by $\bar{y}$, is called an *optimal state*, and $(\bar{y}, \bar{u})$ is called an *optimal pair*.

Problem (DOP) can be regarded as an infinite-dimensional optimization problem. A principal approach to solve it is to derive necessary conditions satisfied by optimal solutions. Nevertheless, since $\mathcal{U}$ may be quite general, the classical variation technique cannot be applied to Problem (DOP) directly. In [43], L. S. Pontryagin's group employed the spike variation to derive the so-called *Pontryagin's Maximum Principle*, which states a necessary condition that any optimal pair must satisfy:

Theorem 1.2. *Let $(\bar{y}, \bar{u})$ be an optimal pair for Problem (DOP). Then, for a.e. $t \in [0, T]$,*

$$\mathbb{H}\big(t, \bar{y}(t), \bar{u}(t), z(t)\big) = \max_{u \in U} \mathbb{H}\big(t, \bar{y}(t), u, z(t)\big), \tag{1.7}$$

where $z : [0, T] \to \mathbb{R}^n$ solves

$$\begin{cases} z_t(t) = -f_y\big(t, \bar{y}(t), \bar{u}(t)\big)^\top z(t) + g_y\big(t, \bar{y}(t), \bar{u}(t)\big), \quad \text{a.e. } t \in [0, T], \\ z(T) = -h_y\big(\bar{y}(T)\big) \end{cases} \tag{1.8}$$

and

$$\mathbb{H}(t, y, u, p) \triangleq \big\langle p, f(t, y, u)\big\rangle_{\mathbb{R}^n} - g(t, y, u), \quad (t, y, u, p) \in [0, T] \times \mathbb{R}^n \times U \times \mathbb{R}^n.$$

The significance of Theorem 1.2 lies in that the infinite-dimensional optimization problem (1.6) is reduced to the finite-dimensional optimization problem (1.7) (in the pointwise sense). Particularly, in many cases, U is a finite set and (1.7) itself allows people to construct the optimal control.

Compared with Problem (DOP), there are new essential difficulties in establishing Pontryagin-type Maximum Principle for optimal control problems of SDPSs. The primary one is the well-posedness of the adjoint equation (a generalization of (1.8)), which is an operator-valued backward stochastic evolution equation. There is no suitable stochastic integration theory for general operator-valued stochastic processes. Hence, that equation cannot be understood as a stochastic integral equation and does not admit a mild or a weak solution.

To overcome this difficulty, we introduce a new notion, i.e., relaxed transposition solution and employ the stochastic transposition method to prove the well-posedness of that equation. More details are provided in Section 3.

In this paper, we consider control problems for SDPSs governed by Itô-type SPDEs. The system is completely observable (meaning that the controller is able to observe the system state completely) and the noise is a one-dimensional standard Brownian motion. For the optimal control problem, the cost functional is an integral over a deterministic time interval. The reasons for these settings are that we would like to show readers some fundamental structure and properties of control problems for SDPSs in a clean and clear way, and avoid technicalities caused by more complicated models.

The rest of this paper consists of three parts. The first (resp. second) one is devoted to controllability (resp. optimal control) problems for SDPSs. At last, in the third part, we provide some open problems for control theory of SDPSs.

2. EXACT CONTROLLABILITY OF STOCHASTIC HYPERBOLIC EQUATIONS

For the readers' convenience, we first recall some basic notations. Let $T > 0$ and $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$ (with $\mathbf{F} = \{\mathcal{F}_t\}_{t\geq 0}$ being a filtration) be a complete filtered probability space. Denote by $\mathbb{F}$ the progressive σ-field (in $[0, T] \times \Omega$) with respect to $\mathbf{F}$. Let $\mathbf{X}$ be a Banach space. For any $p, q \in [1, \infty)$, write $L^p_{\mathcal{F}_t}(\Omega; \mathbf{X}) \triangleq L^p(\Omega, \mathcal{F}_t, \mathbb{P}; \mathbf{X})$ ($t \in [0, T]$), and define

$$L^q_{\mathbb{F}}\big(0, T; L^p(\Omega; \mathbf{X})\big) \triangleq \left\{\varphi : (0, T) \times \Omega \to \mathbf{X} \mid \varphi(\cdot) \text{ is } \mathbf{F}\text{-adapted and } \int_0^T \big(\mathbb{E}|\varphi(t)|^p_{\mathbf{X}}\big)^{\frac{q}{p}} dt < \infty\right\}.$$

Similarly, for $1 \leq p < \infty$, we may also define $L^\infty_{\mathbb{F}}(0, T; L^p(\Omega; \mathbf{X}))$, $L^p_{\mathbb{F}}(0, T; L^\infty(\Omega; \mathbf{X}))$, and $L^\infty_{\mathbb{F}}(0, T; L^\infty(\Omega; \mathbf{X}))$. In the sequel, we shall simply denote $L^p_{\mathbb{F}}(\Omega; L^p(0, T; \mathbf{X})\big) \equiv L^p_{\mathbb{F}}\big(0, T; L^p(\Omega; \mathbf{X})\big)$ by $L^p_{\mathbb{F}}(0, T; \mathbf{X})$. For any $p \in [1, \infty)$, set

$$C_{\mathbb{F}}\big([0, T]; L^p(\Omega; \mathbf{X})\big) \triangleq \big\{\varphi : [0, T] \times \Omega \to \mathbf{X} \mid \varphi \text{ is } \mathbf{F}\text{-adapted and } \varphi : [0, T] \to L^p_{\mathcal{F}_T}(\Omega; \mathbf{X}) \text{ is continuous}\big\}.$$

Similarly, for any $k \in \mathbb{N}$, one can define the Banach space $C^k_{\mathbb{F}}([0, T]; L^p(\Omega; \mathbf{X}))$. Also, we write $D_{\mathbb{F}}([0, T]; L^p(\Omega; \mathbf{X}))$ for the Banach space of all X-valued, $\mathbf{F}$-adapted, stochastic processes X which are càdlàg in $L^p_{\mathcal{F}_T}(\Omega; X)$ and $|X|_{L^\infty_{\mathbb{F}}(0,T;L^p(\Omega;\mathbf{X}))} < \infty$, with the norm inherited from $L^\infty_{\mathbb{F}}(0, T; L^p(\Omega; \mathbf{X}))$.

Throughout this section, we assume that there is a 1-dimensional standard Brownian motion $W(\cdot)$ on $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$ and $\mathbf{F}$ is the natural filtration generated by $W(\cdot)$.

Let $G \subset \mathbb{R}^n$ ($n \in \mathbb{N}$) be a bounded domain with a C^2 boundary Γ. Let $\Gamma_0 \subset \Gamma$ be a nonempty subset satisfying suitable assumptions to be given later. Set $Q = (0, T) \times G$, $\Sigma = (0, T) \times \Gamma$, and $\Sigma_0 = (0, T) \times \Gamma_0$. Let $(a^{jk})_{1\leq j,k\leq n} \in C^3(\overline{G}; \mathbb{R}^{n\times n})$ be such that

$a^{jk} = a^{kj}$ $(j,k = 1,2,\dots,n)$ and, for some constant $s_0 > 0$,

$$\sum_{j,k=1}^{n} a^{jk}(x)\xi^j\xi^k \geq s_0|\xi|^2, \quad \forall (x,\xi) \triangleq \big(x,\xi^1,\dots,\xi^n\big) \in G \times \mathbb{R}^n.$$

Fix $a_1 \in L^\infty_{\mathbb{F}}(0,T;W^{1,\infty}(G;\mathbb{R}^n))$, $a_2, a_3, a_4 \in L^\infty_{\mathbb{F}}(0,T;L^\infty(G))$, and $a_5 \in L^\infty_{\mathbb{F}}(0,T;W_0^{1,\infty}(G))$.

2.1. Formulation of the problem

Consider the following controlled stochastic hyperbolic equation:

$$\begin{cases} dy_t - \displaystyle\sum_{j,k=1}^{n} (a^{jk}y_{x_j})_{x_k}dt = (a_1\cdot\nabla y + a_2 y + f)dt + (a_3 y + g)dW(t) & \text{in } Q, \\ y = h & \text{on } \Sigma, \\ y(0) = y_0,\ y_t(0) = y_1 & \text{in } G, \end{cases} \tag{2.1}$$

where the initial data $(y_0, y_1) \in L^2(G) \times H^{-1}(G)$, (y, y_t) is the state, and $f, g \in L^\infty_{\mathbb{F}}(0,T;H^{-1}(G))$ and $h \in L^2_{\mathbb{F}}(0,T;L^2(\Gamma))$ are three controls. As we shall see in Section 2.2, equation (2.1) admits a unique *transposition solution*

$$y \in C_{\mathbb{F}}\big([0,T];L^2\big(\Omega;L^2(G)\big)\big) \cap C^1_{\mathbb{F}}\big([0,T];L^2\big(\Omega;H^{-1}(G)\big)\big).$$

Inspired by the definition of the exact controllability of deterministic hyperbolic equations and stochastic differential equations, we introduce the following notion.

Definition 2.1. We say that the control system (2.1) is exactly controllable at time T if for any $(y_0, y_1) \in L^2(G) \times H^{-1}(G)$ and $(y_0', y_1') \in L^2_{\mathcal{F}_T}(\Omega;L^2(G)) \times L^2_{\mathcal{F}_T}(\Omega;H^{-1}(G))$, one can find controls $(f,g,h) \in L^2_{\mathbb{F}}(0,T;H^{-1}(G)) \times L^2_{\mathbb{F}}(0,T;H^{-1}(G)) \times L^2_{\mathbb{F}}(0,T;L^2(\Gamma))$ such that the corresponding state y to (2.1) satisfies that $(y(T), y_t(T)) = (y_0', y_1')$ a.s.

Remark 2.1. Compared with Definition 1.1, Definition 2.1 looks much more complex. This is due to the complexity of the control system. The two definitions share the same spirit, that is, using controls to steer the state of the system to the desired destination. Here and in what follows, we use adapted stochastic processes as controls according to two reasons:

(1) In stochastic control systems, "uncertainty" is critical, i.e., there is some possible variations in the system's behavior. The controls have to take different possibilities into account.

(2) We cannot use information from the future. Thus, the control at time t has to be measurable with respect to the σ-algebra $\mathcal{F}_t$, which reflects the information we can obtain at time t.

Three controls are applied in (2.1). One may expect the exact controllability to be correct. However, surprisingly enough, we have the following negative result.

Theorem 2.1 ([37, THEOREM 2.1]). *The system* (2.1) *is not exactly controllable for any* $T > 0$.

Remark 2.2. Both Theorem 2.1 and Theorem 2.2 below are negative results, which have their own interests. Indeed, one aspect of control theory that is particularly important is the exploration of fundamental limits of the control ability for a given control system, since trade-offs between the cost we pay for controls and the performance of the behavior of the system will be the primary design challenge for a control system.

The controls we introduce into (2.1) are the strongest possible ones. Theorem 2.1 shows that the controllability property of stochastic hyperbolic equations differs significantly from the well-known controllability property for deterministic hyperbolic equations (e.g., [50]). Motivated by this, we consider the following refined version of controlled stochastic hyperbolic equation:

$$\begin{cases} dy = \hat{y}dt + (a_4 y + f)dW(t) & \text{in } Q, \\ d\hat{y} - \sum_{j,k=1}^{n} (a^{jk} y_{x_j})_{x_k} dt = (a_1 \cdot \nabla y + a_2 y + a_5 g)dt + (a_3 y + g)dW(t) & \text{in } Q, \\ y = \chi_{\Sigma_0} h & \text{on } \Sigma, \\ y(0) = y_0,\ \hat{y}(0) = \hat{y}_0 & \text{in } G. \end{cases} \tag{2.2}$$

Here $(y_0, \hat{y}_0) \in L^2(G) \times H^{-1}(G)$, $(y, \hat{y})$ is the state, and $f \in L^2_{\mathbb{F}}(0, T; L^2(G))$, $g \in L^2_{\mathbb{F}}(0, T; H^{-1}(G))$, and $h \in L^2_{\mathbb{F}}(0, T; L^2(\Gamma_0))$ are controls. As we shall see in Section 2.2, the system (2.2) admits a unique *transposition solution* $(y, \hat{y}) \in C_{\mathbb{F}}([0, T]; L^2(\Omega; L^2(G))) \times C_{\mathbb{F}}([0, T]; L^2(\Omega; H^{-1}(G)))$. Readers are referred to [37] for the derivation of (2.2).

Remark 2.3. Usually, if we put a control in the diffusion term, it may affect the drift term in one way or another. Here we assume that the effect is in the form of "$a_5 g dt$" as that in the second equation of (2.2). One may consider a more general case, say, by adding a term like "$a_6 f dt$" (in which $a_6 \in L^\infty_{\mathbb{F}}(0, T; L^\infty(G))$) into the first equation of (2.2). However, except for $n = 1$, the corresponding controllability problem is still unsolved (e.g., [39]).

Definition 2.2. We say that the system (2.2) is exactly controllable at time T if for any $(y_0, \hat{y}_0) \in L^2(G) \times H^{-1}(G)$ and $(y_1, \hat{y}_1) \in L^2_{\mathcal{F}_T}(\Omega; L^2(G)) \times L^2_{\mathcal{F}_T}(\Omega; H^{-1}(G))$, one can find controls $(f, g, h) \in L^2_{\mathbb{F}}(0, T; L^2(G)) \times L^2_{\mathbb{F}}(0, T; H^{-1}(G)) \times L^2_{\mathbb{F}}(0, T; L^2(\Gamma_0))$ such that the corresponding solution $(y, \hat{y})$ to (2.2) satisfies that $(y(T), \hat{y}(T)) = (y_1, \hat{y}_1)$.

Under some assumptions, we can show that (2.2) is exactly controllable (see Theorem 2.3). Hence, from the viewpoint of controllability, (2.2) is a more reasonable model than (2.1).

2.2. Well-posedness of stochastic hyperbolic equations with boundary controls

Both (2.1) and (2.2) are SPDEs with nonhomogeneous boundary values. They may not have weak or mild solutions. Therefore, as the deterministic case (e.g., [22]), solutions

to them are understood in the sense of a transposition solution. To this end, we need the following backward stochastic hyperbolic equation:

$$\begin{cases} dz = \hat{z}dt + Z dW(t) & \text{in } Q_\tau, \\ d\hat{z} - \sum_{j,k=1}^{n} (a^{jk} z_{x_j})_{x_k} dt = (b_1 \cdot \nabla z + b_2 z + b_3 Z + b_4 \hat{Z})dt + \hat{Z} dW(t) & \text{in } Q_\tau, \\ z = 0 & \text{on } \Sigma_\tau, \\ z(\tau) = z^\tau,\ \hat{z}(\tau) = \hat{z}^\tau & \text{in } G, \end{cases} \tag{2.3}$$

where $\tau \in (0, T]$, $Q_\tau \triangleq (0, \tau) \times G$, $\Sigma_\tau \triangleq (0, \tau) \times \Gamma$, $(z^\tau, \hat{z}^\tau) \in L^2_{\mathcal{F}_\tau}(\Omega; H_0^1(G) \times L^2(G))$, $b_1 \in L^\infty_{\mathbb{F}}(0, T; W^{1,\infty}(G; \mathbb{R}^n))$, and $b_i \in L^\infty_{\mathbb{F}}(0, T; L^\infty(G))$ $(i = 2, 3, 4)$.

For any $(z^\tau, \hat{z}^\tau) \in L^2_{\mathcal{F}_\tau}(\Omega; H_0^1(G)) \times L^2_{\mathcal{F}_\tau}(\Omega; L^2(G))$, the system (2.3) admits a unique solution $(z, Z, \hat{z}, \hat{Z}) \in C_{\mathbb{F}}([0, \tau]; H_0^1(G)) \times L^2_{\mathbb{F}}(0, \tau; H_0^1(G)) \times C_{\mathbb{F}}([0, \tau]; L^2(G)) \times L^2_{\mathbb{F}}(0, \tau; L^2(G))$ (e.g., [40, THEOREM 4.10]), which satisfies the following hidden regularity:

Proposition 2.1 ([37, PROPOSITION 3.1]). *The solution* $(z, \hat{z}, Z, \hat{Z})$ *to* (2.3) *satisfies* $\frac{\partial z}{\partial \nu}|_\Gamma \in L^2_{\mathbb{F}}(0, \tau; L^2(\Gamma))$ *and*

$$\left| \frac{\partial z}{\partial \nu} \right|_{L^2_{\mathbb{F}}(0,\tau;L^2(\Gamma))} \le \mathcal{C}\big(|z^\tau|_{L^2_{\mathcal{F}_\tau}(\Omega;H_0^1(G))} + |\hat{z}^\tau|_{L^2_{\mathcal{F}_\tau}(\Omega;L^2(G))} \big), \tag{2.4}$$

where the constant $\mathcal{C}$ *is independent of* τ *and* $(z^\tau, \hat{z}^\tau) \in L^2_{\mathcal{F}_\tau}(\Omega; H_0^1(G)) \times L^2_{\mathcal{F}_\tau}(\Omega; L^2(G))$.

Definition 2.3. A stochastic process $y \in C_{\mathbb{F}}([0, T]; L^2(\Omega; L^2(G))) \cap C^1_{\mathbb{F}}([0, T]; L^2(\Omega; H^{-1}(G)))$ is called a *transposition solution* to (2.1) if for any $\tau \in (0, T]$ and $(z^\tau, \hat{z}^\tau) \in L^2_{\mathcal{F}_\tau}(\Omega; H_0^1(G)) \times L^2_{\mathcal{F}_\tau}(\Omega; L^2(G))$, it holds that

$$\begin{aligned} &\mathbb{E}\langle y_t(\tau), z^\tau \rangle_{H^{-1}(G), H_0^1(G)} - \mathbb{E}\langle y(\tau), \hat{z}^\tau \rangle_{L^2(G)} - \langle \hat{y}_0, z(0) \rangle_{H^{-1}(G), H_0^1(G)} + \langle y_0, \hat{z}(0) \rangle_{L^2(G)} \\ &\quad = \mathbb{E} \int_0^\tau \langle f, z \rangle_{H^{-1}(G), H_0^1(G)} dt + \mathbb{E} \int_0^\tau \langle g, Z \rangle_{H^{-1}(G), H_0^1(G)} dt - \mathbb{E} \int_0^\tau \int_{\Gamma_0} h \frac{\partial z}{\partial \nu} d\Gamma ds, \end{aligned}$$

where $(z, \hat{z}, Z, \hat{Z})$ solves (2.3) with $b_1 = -a_1$, $b_2 = -\mathrm{div} a_1 + a_2$, $b_3 = a_3$, and $b_4 = 0$.

A pair of stochastic processes $(y, \hat{y}) \in C_{\mathbb{F}}([0, T]; L^2(\Omega; L^2(G))) \times C_{\mathbb{F}}([0, T]; L^2(\Omega; H^{-1}(G)))$ is called a *transposition solution* to (2.2) if for any $\tau \in (0, T]$ and $(z^\tau, \hat{z}^\tau) \in L^2_{\mathcal{F}_\tau}(\Omega; H_0^1(G)) \times L^2_{\mathcal{F}_\tau}(\Omega; L^2(G))$, it holds that

$$\begin{aligned} &\mathbb{E}\langle \hat{y}(\tau), z^\tau \rangle_{H^{-1}(G), H_0^1(G)} - \mathbb{E}\langle y(\tau), \hat{z}^\tau \rangle_{L^2(G)} - \langle \hat{y}_0, z(0) \rangle_{H^{-1}(G), H_0^1(G)} + \langle y_0, \hat{z}(0) \rangle_{L^2(G)} \\ &\quad = -\mathbb{E} \int_0^\tau \langle f, \hat{Z} \rangle_{L^2(G)} dt + \mathbb{E} \int_0^\tau \langle g, a_5 z + Z \rangle_{H^{-1}(G), H_0^1(G)} dt - \mathbb{E} \int_0^\tau \int_{\Gamma_0} h \frac{\partial z}{\partial \nu} d\Gamma ds, \end{aligned}$$

where $(z, \hat{z}, Z, \hat{Z})$ solves (2.3) with $b_1 = -a_1$, $b_2 = -\mathrm{div} a_1 + a_2$, $b_3 = a_3$, and $b_4 = -a_4$.

Remark 2.4. By Proposition 2.1, the term "$\mathbb{E} \int_0^\tau \int_{\Gamma_0} h \frac{\partial z}{\partial \nu} d\Gamma ds$" makes sense. The above definitions of transposition solutions to (2.1) and (2.2) are the generalization of the transposition solution to deterministic hyperbolic equation (e.g., [22]).

Proposition 2.2 ([37, PROPOSITIONS 4.1 AND 4.2]). *The system* (2.1) *(*resp. (2.2)*) admits a unique transposition solution* y *(resp.* $(y, \hat{y})$*).*

2.3. The controllability results

We have introduced three controls (f, g, and h) in the system (2.2). At first glance, it seems unreasonable that especially the controls f and g in the diffusion terms of (2.2) are acting on the whole domain G. One may ask whether localized controls are enough or the boundary control can be dropped. However, the answers are "NO."

Theorem 2.2 ([37, THEOREM 2.3]). *For any open subset* Γ_0 *of* Γ *and open subset* G_0 *of* G*, the system* (2.2) *is not exactly controllable at any time* $T > 0$*, provided that one of the following three conditions is satisfied:*

(1) $a_4 \in C_{\mathbb{F}}([0,T]; L^\infty(\Omega; L^\infty(G)))$, $G \setminus \overline{G_0} \neq \emptyset$, *and* f *is supported in* G_0*;*

(2) $a_3 \in C_{\mathbb{F}}([0,T]; L^\infty(\Omega; L^\infty(G)))$, $G \setminus \overline{G_0} \neq \emptyset$, *and* g *is supported in* G_0*;*

(3) $h = 0$.

To get a positive controllability result for the system (2.2), the time T should be large enough due to the finite propagation speed of solutions to stochastic hyperbolic equations. On the other hand, noting that the deterministic wave equation is a special case of (2.2), by [2], we see that exact controllability of (2.2) is impossible without conditions on Γ_0 and $(a^{jk})_{1\le j,k\le n}$. Hence, to continue, we introduce the following assumptions:

Condition 2.1. *There exists a positive function* $\varphi(\cdot) \in C^3(\overline{G})$ *satisfying the following:*

(1) *For some constant* $\mu_0 > 0$ *and all* $(x, \xi^1, \dots, \xi^n) \in \overline{G} \times \mathbb{R}^n$,

$$\sum_{j,k=1}^{n} \sum_{j',k'=1}^{n} \left[2a^{jk'}(a^{j'k}\varphi_{x_{j'}})_{x_{k'}} - a^{jk}_{x_{k'}} a^{j'k'}\varphi_{x_{j'}}\right]\xi^j\xi^k \ge \mu_0 \sum_{j,k=1}^{n} a^{jk}\xi^j\xi^k.$$

(2) *The function* $\varphi(\cdot)$ *has no critical point in* $\overline{G}$*, i.e.,* $|\nabla\varphi(x)| > 0$ *for* $x \in \overline{G}$.

We shall choose the set Γ_0 as follows:

$$\Gamma_0 \triangleq \left\{ x \in \Gamma \;\middle|\; \sum_{j,k=1}^{n} a^{jk}\varphi_{x_j}(x)\nu^k(x) > 0 \right\}.$$

Also, write

$$R_1 \triangleq \sqrt{\max_{x\in\overline{G}} \varphi(x)}, \quad R_0 \triangleq \sqrt{\min_{x\in\overline{G}} \varphi(x)}.$$

Clearly, if $\varphi(\cdot)$ satisfies Condition 2.1, then for any given constants $\alpha \ge 1$ and $\beta \in \mathbb{R}$, so does $\tilde{\varphi} = \alpha\varphi + \beta$ with μ_0 replaced by $\alpha\mu_0$. Therefore we may choose φ, μ_0, c_0, c_1 and T such that

Condition 2.2. *The following inequalities hold:*

(1) $\frac{1}{4} \sum_{j,k=1}^{n} a^{jk}(x)\varphi_{x_j}(x)\varphi_{x_k}(x) \ge R_1^2$, $\forall x \in \overline{G}$;

(2) $T > T_0 \triangleq 2R_1$;

(3) $(\frac{2R_1}{T})^2 < c_1 < \frac{2R_1}{T}$;

(4) $\mu_0 - 4c_1 - c_0 > c_0 + 2R_1(1 + |a_5|^2_{L^\infty_{\mathbb{F}}(0,T;L^\infty(G))})$.

Remark 2.5. As we have explained before Condition 2.2, this condition can always be satisfied. We put it here merely to emphasize the relationship among c_0, c_1, μ_0 and T.

Remark 2.6. To ensure that (4) in Condition 2.2 holds, c_1 and T depend on $|a_5|_{L^\infty_{\mathbb{F}}(0,T;L^\infty(G))}$. This seems to be reasonable because a_5 stands for the effect of the control in the diffusion term to the drift term. One needs time to get rid of such an effect. Nevertheless, this does not happen when $n = 1$ (e.g., [39]).

The exact controllability result for the system (2.2) is stated as follows:

Theorem 2.3 ([37, THEOREM 2.2]). *System* (2.2) *is exactly controllable at time* T *if Conditions* 2.1 *and* 2.2 *hold.*

Remark 2.7. Although it is necessary to put controls f and g on the whole domain G, one may suspect that Theorem 2.3 is trivial and give a possible "proof" of Theorem 2.3 as follows: Choosing $f = -a_4 y$ and $g = -a_3 y$, the system (2.2) becomes

$$\begin{cases} dy = \hat{y}dt & \text{in } Q, \\ d\hat{y} - \sum_{j,k=1}^{n} (a^{jk} y_{x_j})_{x_k} dt = (a_1 \cdot \nabla y + a_2 y - a_5 a_3 y)dt & \text{in } Q, \\ y = \chi_{\Sigma_0} h & \text{on } \Sigma, \\ y(0) = y_0,\ \hat{y}(0) = \hat{y}_0 & \text{in } G. \end{cases} \tag{2.5}$$

This is a hyperbolic equation with random coefficients. If one regards the sample point ω as a parameter, then for every given $\omega \in \Omega$, there is a control $h(\cdot, \cdot, \omega)$ such that the solution to (2.5) fulfills $(y(T, x, \omega), \hat{y}(T, x, \omega)) = (y_1(x, \omega), \hat{y}_1(x, \omega))$. However, it is unclear whether the control constructed in this way is adapted to the filtration $\mathbf{F}$ or not. If it is not the case, then to determine the value of the control at present, one needs to use information from the future, which is meaningless in the stochastic framework.

In order to prove Theorem 2.3, by a standard duality argument, it suffices to establish the following observability estimate for the adjoint equation (2.3).

Theorem 2.4. *Under the assumptions of Theorem* 2.3, *all solutions to equation* (2.3) *with* $\tau = T$ *satisfy*

$$\begin{aligned} &\left|\left(z^T, \hat{z}^T\right)\right|_{L^2_{\mathcal{F}_T}(\Omega; H^1_0(G) \times L^2(G))} \\ &\quad \le \mathcal{C}\left(\left| \frac{\partial z}{\partial \nu} \right|_{L^2_{\mathbb{F}}(0,T;L^2(\Gamma_0))} + |a_5 z + Z|^2_{L^2_{\mathbb{F}}(0,T;H^1_0(G))} + |\hat{Z}|^2_{L^2_{\mathbb{F}}(0,T;L^2(G))} \right). \end{aligned}$$

Remark 2.8. Although Theorem 2.4 is much more complex than Theorem 1.1, it has the same features in common with Theorem 1.1, that is, a solution of an equation can be fully determined by a suitable observation of the solution.

Remark 2.9. The proof of Theorem 2.4 is almost the same as that of [37, THEOREM 7.1]. We do not provide the explicit dependence of the constant $\mathcal{C}$ on the observation time T and the coefficients b_i $(1 \le i \le 4)$. Interested readers are referred to [37].

2.4. Carleman estimate

Theorem 2.4 is an observability estimate of equation (2.3). Generally speaking, there are three main approaches to establish the observability estimate for multidimensional deterministic hyperbolic equations.

The first is the multiplier techniques (e.g., [21]). Two key points for applying this method are the time reversibility of the equation and the time independence of the coefficients. Equation (2.3) does not fulfill the second property above.

The second approach is based on the microlocal analysis (e.g., [2]), which gives a sharp sufficient condition, i.e., the *Geometric Control Condition*, for the observability estimate of hyperbolic equations. It is interesting to generalize this method to study the observability estimate of equation (2.3).

The last one is the global Carleman estimate (e.g., [15, 49]). It has been generalized to study the observability estimate for stochastic hyperbolic equations recently (e.g., [28, 34, 48, 49]). Theorem 2.3 is also proved likewise. The key is the following identity.

Lemma 2.1 ([37, LEMMA 6.1]). *Let z be an $H^2(\mathbb{R}^n)$-valued Itô process and $\hat{z}$ be an $L^2(\mathbb{R}^n)$-valued Itô process such that for some $Z \in L^2_{\mathbb{F}}(0,T;H^1(\mathbb{R}^n))$, $dz = \hat{z}dt + ZdW(t)$ in $(0,T)\times\mathbb{R}^n$. Let $\ell, \Psi \in C^2((0,T)\times\mathbb{R}^n)$. Set $\theta = e^{\ell}$, $v = \theta z$ and $\hat{v} = \theta\hat{z} + \ell_t v$. Then, for a.e. $x \in \mathbb{R}^n$,*

$$
\begin{aligned}
&\theta\left(-2\ell_t\hat{v} + 2\sum_{j,k=1}^n a^{jk}\ell_{x_j}v_{x_k} + \Psi v\right)\left[d\hat{z} - \sum_{j,k=1}^n \left(a^{jk}z_{x_j}\right)_{x_k}dt\right] \\
&\quad + \sum_{j,k=1}^n\left[\sum_{j',k'=1}^n \left(2a^{jk}a^{j'k'}\ell_{x_{j'}}v_{x_j}v_{x_{k'}} - a^{jk}a^{j'k'}\ell_{x_j}v_{x_{j'}}v_{x_{k'}}\right)\right. \\
&\quad \left. - 2\ell_t a^{jk}v_{x_j}\hat{v} + a^{jk}\ell_{x_j}\hat{v}^2 + \Psi a^{jk}v_{x_j}v - \frac{\Psi_{x_j}}{2}a^{jk}v^2 - \mathcal{A}a^{jk}\ell_{x_j}v^2\right]_{x_k} \\
&\quad + d\left[\ell_t\sum_{j,k=1}^n a^{jk}v_{x_j}v_{x_k} + \ell_t\hat{v}^2 - 2\sum_{j,k=1}^n a^{jk}\ell_{x_j}v_{x_k}\hat{v} - \Psi v\hat{v} + \left(\mathcal{A}\ell_t + \frac{\Psi_t}{2}\right)v^2\right] \\
&= \left\{\left[\ell_{tt} + \sum_{j,k=1}^n \left(a^{jk}\ell_{x_j}\right)_{x_k} - \Psi\right]\hat{v}^2 + \sum_{j,k=1}^n c^{jk}v_{x_j}v_{x_k} + \mathcal{B}v^2\right. \\
&\quad \left. - 2\sum_{j,k=1}^n\left[\left(a^{jk}\ell_{x_k}\right)_t + a^{jk}\ell_{tx_k}\right]v_{x_j}\hat{v} + \left(-2\ell_t\hat{v} + 2\sum_{j,k=1}^n a^{jk}\ell_{x_j}v_{x_k} + \Psi v\right)^2\right\}dt
\end{aligned}
$$

$$
\begin{aligned}
&+\ell_t(d\hat v)^2-2\sum_{j,k=1}^n a^{jk}\ell_{x_j}dv_{x_k}d\hat v-\Psi dvd\hat v+\ell_t\sum_{j,k=1}^n a^{jk}(dv_{x_j})(dv_{x_k})\\
&+\mathcal{A}\ell_t(dv)^2-\Bigg\{\theta\bigg(-2\ell_t\hat v+2\sum_{j,k=1}^n a^{jk}\ell_{x_j}v_{x_k}+\Psi v\bigg)\ell_t Z\\
&-\bigg[2\sum_{j,k=1}^n a^{jk}(\theta Z)_{x_k}\ell_{x_j}\hat v-\theta\Psi_t vZ+\theta\Psi\hat vZ\bigg]\\
&+2\bigg[\sum_{j,k=1}^n a^{jk}v_{x_j}(\theta Z)_{x_k}+\theta\mathcal{A}vZ\bigg]\ell_t\Bigg\}dW(t),\quad a.s.,
\end{aligned}
\tag{2.6}
$$

where $(dv)^2$ *and* $(d\hat v)^2$ *denote the quadratic variation processes of* v *and* $\hat v$*, respectively, and*

$$
\begin{cases}
c^{jk}\triangleq\big(a^{jk}\ell_t\big)_t+\displaystyle\sum_{j',k'=1}^n\big[2a^{jk'}\big(a^{j'k}\ell_{x_{j'}}\big)_{x_{k'}}-\big(a^{jk}a^{j'k'}\ell_{x_{j'}}\big)_{x_{k'}}\big]+\Psi a^{jk},\\
\mathcal{A}\triangleq(\ell_t^2-\ell_{tt})-\displaystyle\sum_{j,k=1}^n\big[a^{jk}\ell_{x_j}\ell_{x_k}-\big(a^{jk}\ell_{x_j}\big)_{x_k}\big]-\Psi,\\
\mathcal{B}\triangleq\mathcal{A}\Psi+(\mathcal{A}\ell_t)_t-\displaystyle\sum_{j,k=1}^n\big(\mathcal{A}a^{jk}\ell_{x_j}\big)_{x_k}+\frac12\bigg[\Psi_{tt}-\sum_{j,k=1}^n\big(a^{jk}\Psi_{x_j}\big)_{x_k}\bigg].
\end{cases}
$$

Remark 2.10. The derivation of (2.6) requires a fairly complex but elementary computation. Identities in the spirit of (2.6) are widely used to solve observability problems for deterministic and stochastic PDEs (e.g., [14,15,39,40]).

Choosing $\ell(t,x)=\lambda[\varphi(x)-c_1(t-\frac{T}{2})^2]$ and $\Psi=\ell_{tt}+\sum_{j,k=1}^n(a^{jk}\ell_{x_j})_{x_k}-c_0\lambda$ in (2.6), integrating (2.6) in Q and taking the mathematical expectation, after some technical computations, one can prove Theorem 2.3.

The above not only gives a sketch of the proof of Theorem 2.3, but also presents a methodology of getting the observability estimates for SPDEs and backward SPDEs: indeed, one has to establish a suitable pointwise identity and choose a suitable weight function. Almost all observability estimates for SPDEs and backward SPDEs are obtained in this way (e.g., [14,27–29,34,39,40,45,48,49]). That said, we do not mean that the proofs of these observability estimates are similar; rather we want to emphasize the common ground in the idea of the proofs.

3. PONTRYAGIN-TYPE STOCHASTIC MAXIMUM PRINCIPLE AND STOCHASTIC TRANSPOSITION METHOD

This section is devoted to the Pontryagin-type stochastic maximum principle (PMP for short) for optimal control problems of semilinear SDPSs. There is a long history for the study of this topic. We refer to [3] for a pioneering result and to [17,44] and the references therein for subsequent results. These works addressed three special cases: (1) the diffusion

term does not depend on the control variable; (2) U is convex; (3) the second-order derivatives of g and h with respect to y in (3.2) below are Hilbert–Schmidt operator-valued. On the one hand, under the first two assumptions (resp. the third assumption), the PMP and their proofs are similar to those of the distributed parameter control systems (resp. stochastic finite-dimensional control systems). On the other hand, when one puts a control in the drift term, it will affect the diffusion term, i.e., the control could influence the scale of uncertainty. Hence, it is important to study PMP for SDPSs with control-dependent diffusion terms and nonconvex control domains. This was done in [33] (some generalizations were given in [35, 36, 39]).

3.1. Formulation of the optimal control problem

Unlike in Section 2, we will formulate our system in an abstract framework. Throughout this section, $T > 0$, $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$ (with $\mathbf{F} \triangleq \{\mathcal{F}_t\}_{t\in[0,T]}$) is a fixed filtered probability space satisfying the usual conditions, on which a 1-dimensional standard Brownian motion $W(\cdot)$ is defined, and H is a separable Hilbert space. Denote by $\mathbb{F}$ the progressive σ-field (in $[0, T] \times \Omega$) with respect to $\mathbf{F}$.

Let A be a linear operator (with the domain $D(A) \subset H$), which generates a C_0-semigroup $\{S(t)\}_{t\geq 0}$ on H. Denote by A^* the adjoint operator of A, which generates the adjoint C_0-semigroup of $\{S(t)\}_{t\geq 0}$. Let U be a separable metric space. Put

$$\mathcal{U}[0, T] \triangleq \big\{u : [0, T] \times \Omega \to U \mid u \text{ is } \mathbf{F}\text{-adapted}\big\}.$$

We assume the following condition:

(A1). *The maps $a, b : [0, T] \times H \times U \to H$ satisfy (for $\varphi = a, b$):* (i) *for any $(y, u) \in H \times U$, $\varphi(\cdot, y, u) : [0, T] \to H$ is Lebesgue measurable;* (ii) *for any $(t, y) \in [0, T] \times H$, $\varphi(t, y, \cdot) : U \to H$ is continuous; and* (iii) *there is a constant $\mathcal{C}_L > 0$ such that*

$$\begin{cases} \big|\varphi(t, y_1, u) - \varphi(t, y_2, u)\big|_H \leq \mathcal{C}_L|y_1 - y_2|_H, \\ \big|\varphi(t, 0, u)\big|_H \leq \mathcal{C}_L, \end{cases} \quad \forall (t, y_1, y_2, u) \in [0, T] \times H \times H \times U.$$

Consider the following controlled stochastic evolution equation:

$$\begin{cases} dy(t) = \big[Ay(t) + a(t, y, u)\big]dt + b(t, y, u)dW(t), \quad \text{a.e. } t \in (0, T], \\ y(0) = \eta, \end{cases} \tag{3.1}$$

where $u \in \mathcal{U}[0, T]$ is control, y is state, and $\eta \in L^8_{\mathcal{F}_0}(\Omega; H)$. The control system (3.1) admits a unique mild solution $y \in C_{\mathbb{F}}([0, T]; L^8(\Omega; H))$ (e.g., [40, THEOREM 3.13]).

Remark 3.1. In (3.1), the diffusion term depends on the control. This means that the control could influence the scale of uncertainty (as is indeed the case in many practical systems, especially in the system of mesoscopic scale). In such a setting, the stochastic problems essentially differ from the deterministic ones.

Also, we need the following condition:

(A2). *The maps* $g(\cdot,\cdot,\cdot):[0,T]\times H\times U\to\mathbb{R}$ *and* $h(\cdot):H\to\mathbb{R}$ *satisfy:* (i) *for any* $(y,u)\in H\times U$, $g(\cdot,y,u):[0,T]\to\mathbb{R}$ *is Lebesgue measurable;* (ii) *for any* $(t,y)\in[0,T]\times H$, $g(t,y,\cdot):U\to\mathbb{R}$ *is continuous; and* (iii) *there is a constant* $\mathcal{C}_L>0$ *such that*

$$\begin{cases}\big|g(t,y_1,u)-g(t,y_2,u)\big|_H+\big|h(y_1)-h(y_2)\big|_H\le\mathcal{C}_L|y_1-y_2|_H,\\ \big|g(t,0,u)\big|_H+\big|h(0)\big|_H\le\mathcal{C}_L,\\ \qquad\forall(t,y_1,y_2,u)\in[0,T]\times H\times H\times U.\end{cases}$$

Define a cost functional $\mathcal{J}(\cdot)$ (for the control system (3.1)) as follows:

$$\mathcal{J}(u)\triangleq\mathbb{E}\bigg[\int_0^T g\big(t,y(t),u(t)\big)dt+h\big(y(T)\big)\bigg],\quad\forall u\in\mathcal{U}[0,T],\tag{3.2}$$

where y is the state of (3.1) corresponding to u. Consider an optimal control problem:

Problem (OP). *Find a* $\bar{u}\in\mathcal{U}[0,T]$ *such that*

$$\mathcal{J}(\bar{u})=\inf_{u\in\mathcal{U}[0,T]}\mathcal{J}(u).\tag{3.3}$$

Any $\bar{u}$ satisfying (3.3) is called an *optimal control*. The corresponding state $\bar{y}$ is called an *optimal state*, and $(\bar{y},\bar{u})$ is called an *optimal pair*.

3.2. Transposition solution and relaxed transposition solution to backward stochastic evolution equation

We first recall that the key idea in the proof of Theorem 1.2 is as follows: One first perturbs an optimal control by means of the spike variation, then considers the first-order term in a sort of Taylor expansion with respect to this perturbation. By sending the perturbation to zero, one obtains a kind of variational inequality. The Pontryagin's maximum principle then follows from a duality argument. When applying this idea to study PMP for Problem (OP), one encounters an essential difficulty, which, roughly speaking, is that the Itô stochastic integral $\int_t^{t+\varepsilon}rdW(s)$ is only of order $\sqrt{\varepsilon}$ (rather than ε as with the Lebesgue integral). To overcome this difficulty, we should study both the first and second order terms in the Taylor expansion of the spike variation. In such case, inspired by [42], we need to introduce two adjoint equations. The first is

$$\begin{cases}dz=-A^*zdt+F(t,z,Z)dt+ZdW(t)\quad\text{in }[0,T),\\ z(T)=z_T.\end{cases}\tag{3.4}$$

In (3.4), $F:[0,T]\times H\times H\to H$ is Lebesgue measurable with respect to t and Lipschitz continuous with respect to z and Z.

Neither the usual natural filtration condition nor the quasi-left continuity is assumed for the filtration $\mathbf{F}$, and the operator A is only assumed to generate a general C_0-semigroup. Hence, equation (3.4) may not have a weak or mild solution. Similar to equation (2.1), we

should introduce new notion of solution to (3.4). To this end, consider the following stochastic evolution equation:

$$\begin{cases} d\varphi = (A\varphi + \psi)ds + \tilde{\psi}\, dW(s) \quad \text{in } (t, T], \\ \varphi(t) = \zeta, \end{cases} \tag{3.5}$$

where $t \in [0, T)$, $\psi \in L^1_{\mathbb{F}}(t, T; L^2(\Omega; H))$, $\tilde{\psi} \in L^2_{\mathbb{F}}(t, T; H)$, and $\zeta \in L^2_{\mathcal{F}_t}(\Omega; H)$. Equation (3.5) admits a unique (mild) solution $\varphi \in C_{\mathbb{F}}([t, T]; L^2(\Omega; H))$ (e.g., [40, THEOREM 3.13]).

Definition 3.1. We call $(z, Z) \in D_{\mathbb{F}}([0, T]; L^2(\Omega; H)) \times L^2_{\mathbb{F}}(0, T; H)$ a transposition solution to (3.4) if for any $t \in [0, T]$, $\psi \in L^1_{\mathbb{F}}(t, T; L^2(\Omega; H))$, $\tilde{\psi} \in L^2_{\mathbb{F}}(t, T; H)$, $\zeta \in L^2_{\mathcal{F}_t}(\Omega; H)$, and the corresponding solution $\varphi \in C_{\mathbb{F}}([t, T]; L^2(\Omega; H))$ to (3.5), it holds that

$$\begin{aligned} &\mathbb{E}\langle \varphi(T), z_T\rangle_H - \mathbb{E}\int_t^T \big\langle \varphi(s), f\big(s, z(s), Z(s)\big)\big\rangle_H ds \\ &\quad = \mathbb{E}\langle \zeta, z(t)\rangle_H + \mathbb{E}\int_t^T \langle \psi(s), z(s)\rangle_H ds + \mathbb{E}\int_t^T \langle \tilde{\psi}(s), Z(s)\rangle_H ds. \end{aligned} \tag{3.6}$$

Remark 3.2. On the one hand, if (3.4) admits a strong solution $(z, Z) \in [C_{\mathbb{F}}([0, T]; L^2(\Omega; H)) \cap L^2_{\mathbb{F}}(0, T; D(A))] \times L^2_{\mathbb{F}}(0, T; H)$, then, we can get (3.6) by Itô's formula (e.g., [40, THEOREM 2.142]). On the other hand, (3.6) can be used to get the PMP for Problem (OP). These are the reasons for introducing Definition 3.1. The main idea of this definition is to interpret the solution to a less understood equation by means of another well-understood one.

Theorem 3.1 ([33, THEOREM 3.1]). *Equation* (3.4) *has a unique transposition solution* (z, Z) *and*

$$\begin{aligned} &\big|(z, Z)\big|_{D_{\mathbb{F}}([0,T];L^2(\Omega;H))\times L^2_{\mathbb{F}}(0,T;H)} \\ &\quad \le \mathcal{C}\big(\big|F(\cdot, 0, 0)\big|_{L^1_{\mathbb{F}}(0,T;L^2(\Omega;H))} + |z_T|_{L^2_{\mathcal{F}_T}(\Omega;H)}\big). \end{aligned}$$

The proof of Theorem 3.1 is based on a Riesz-type representation theorem obtained in [31].

The second adjoint equation is[2]

$$\begin{cases} dP = \big[-(A^* + J^*)P - P(A + J) - K^*PK - (K^*Q + QK) + F\big]dt + Q\, dW(t) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in } [0, T), \\ P(T) = P_T. \end{cases} \tag{3.7}$$

where $F \in L^1_{\mathbb{F}}(0, T; L^2(\Omega; \mathcal{L}(H)))$, $P_T \in L^2_{\mathcal{F}_T}(\Omega; \mathcal{L}(H))$, and $J, K \in L^4_{\mathbb{F}}(0, T; L^\infty(\Omega; \mathcal{L}(H)))$.

2 In this paper, for any operator-valued process R, we denote by R^* its pointwise dual operator-valued process.

Equation (3.7), as written, is a rather formidable operator-valued backward stochastic evolution equation. When $H = \mathbb{R}^n$, (3.7) is an $\mathbb{R}^{n\times n}$ matrix-valued backward stochastic differential equation, and therefore, the desired well-posedness follows from that of an $\mathbb{R}^{n^2}$ (vector)-valued backward stochastic differential equation. However, in the infinite-dimensional setting, although $\mathcal{L}(H)$ is still a Banach space, it is neither reflexive nor separable even if H itself is separable. There exists no stochastic integration/evolution equation theory that can be employed to treat the well-posedness of (3.7) even if the filtration $\mathbf{F}$ is generated by $W(\cdot)$ (e.g., [46]). Hence, we should employ the stochastic transposition method again and define the solution to (3.7) in the transposition sense. To this end, we need the following stochastic evolution equation:

$$\begin{cases} d\varphi = (A+J)\varphi ds + \psi ds + K\varphi dW(s) + \tilde{\psi} dW(s) \quad \text{in } (t,T], \\ \varphi(t) = \xi. \end{cases} \tag{3.8}$$

Here $\xi \in L^4_{\mathcal{F}_t}(\Omega;H)$ and $\psi, \tilde{\psi} \in L^2_{\mathbb{F}}(t,T;L^4(\Omega;H))$. Also, we should introduce the solution space for (3.7). Write

$$\begin{aligned} \mathcal{P}[0,T] \triangleq \big\{ & P : [0,T]\times\Omega \to \mathcal{L}(H) \mid |P|_{\mathcal{L}(H)} \in L^\infty_{\mathbb{F}}\big(0,T;L^2(\Omega)\big) \text{ and for every} \\ & t \in [0,T] \text{ and } \xi \in L^4_{\mathcal{F}_t}(\Omega;H),\ P\xi \in D_{\mathbb{F}}\big([t,T];L^{\frac{4}{3}}(\Omega;H)\big) \text{ and} \\ & |P\xi|_{D_{\mathbb{F}}([t,T];L^{\frac{4}{3}}(\Omega;H))} \le \mathcal{C}|\xi|_{L^4_{\mathcal{F}_t}(\Omega;H)} \big\} \end{aligned}$$

and

$$\begin{aligned} \mathcal{Q}[0,T] \triangleq \big\{ & \big(Q^{(\cdot)}, \hat{Q}^{(\cdot)}\big) \mid \text{for any } t \in [0,T], \text{ both } Q^{(t)} \text{ and } \hat{Q}^{(t)} \text{ are bounded linear} \\ & \text{operators from } L^4_{\mathcal{F}_t}(\Omega;H) \times L^2_{\mathbb{F}}\big(t,T;L^4(\Omega;H)\big) \times L^2_{\mathbb{F}}\big(t,T;L^4(\Omega;H)\big) \text{ to} \\ & L^2_{\mathbb{F}}\big(t,T;L^{\frac{4}{3}}(\Omega;H)\big) \text{ and } Q^{(t)}(0,0,\cdot)^* = \hat{Q}^{(t)}(0,0,\cdot) \big\}. \end{aligned}$$

Definition 3.2. We call $(P(\cdot), Q^{(\cdot)}, \hat{Q}^{(\cdot)}) \in \mathcal{P}[0,T]\times\mathcal{Q}[0,T]$ a relaxed transposition solution to (3.7) if for any $t \in [0,T]$, $\xi_1, \xi_2 \in L^4_{\mathcal{F}_t}(\Omega;H)$, and $\psi_1, \psi_2, \tilde{\psi}_1, \tilde{\psi}_2 \in L^2_{\mathbb{F}}(t,T;L^4(\Omega;H))$, it holds that

$$\begin{aligned} & \mathbb{E}\big\langle P_T\varphi_1(T), \varphi_2(T)\big\rangle_H - \mathbb{E}\int_t^T \big\langle F(s)\varphi_1(s), \varphi_2(s)\big\rangle_H ds \\ & = \mathbb{E}\big\langle P(t)\xi_1, \xi_2\big\rangle_H + \mathbb{E}\int_t^T \big\langle P(s)\psi_1(s), \varphi_2(s)\big\rangle_H ds + \mathbb{E}\int_t^T \big\langle P(s)\varphi_1(s), \psi_2(s)\big\rangle_H ds \\ & \quad + \mathbb{E}\int_t^T \big\langle P(s)K(s)\varphi_1(s), \tilde{\psi}_2(s)\big\rangle_H ds + \mathbb{E}\int_t^T \big\langle P(s)\tilde{\psi}_1(s), K(s)\varphi_2(s) + \tilde{\psi}_2(s)\big\rangle_H ds \\ & \quad + \mathbb{E}\int_t^T \big\langle \tilde{\psi}_1(s), \hat{Q}^{(t)}(\xi_2,\psi_2,\tilde{\psi}_2)(s)\big\rangle_H ds + \mathbb{E}\int_t^T \big\langle Q^{(t)}(\xi_1,\psi_1,\tilde{\psi}_1)(s), \tilde{\psi}_2(s)\big\rangle_H ds, \end{aligned}$$

Here, for $j = 1,2$, φ_j solves (3.8) with ξ, ψ, and $\tilde{\psi}$ replaced by ξ_j, ψ_j, and $\tilde{\psi}_j$, respectively.

Remark 3.3. Due to the very weak characterization of Q, a relaxed transposition solution is more like a half-measure rather than the natural solution to (3.7). We believe that a more suitable definition should be as follows:

Let $\hat{\mathcal{Q}}[0,T] \triangleq \big\{ Q : [0,T]\times\Omega \to \mathcal{L}(H) \mid |Q|_{\mathcal{L}(H)} \in L^2_{\mathbb{F}}\big(0,T;L^2(\Omega)\big)\big\}$.

We call $(P(\cdot), Q(\cdot)) \in \mathcal{P}[0,T] \times \hat{\mathcal{Q}}[0,T]$ a *transposition solution* to (3.7) if for any $t \in [0,T]$, $\xi_1, \xi_2 \in L^4_{\mathcal{F}_t}(\Omega; H)$, and $\psi_1, \psi_2, \tilde{\psi}_1, \tilde{\psi}_2 \in L^2_{\mathbb{F}}(t, T; L^4(\Omega; H))$, it holds that

$$\begin{aligned}
&\mathbb{E}\big\langle P_T\varphi_1(T), \varphi_2(T)\big\rangle_H - \mathbb{E}\int_t^T \big\langle F(s)\varphi_1(s), \varphi_2(s)\big\rangle_H ds \\
&\quad = \mathbb{E}\big\langle P(t)\xi_1, \xi_2\big\rangle_H + \mathbb{E}\int_t^T \big\langle P(s)\psi_1(s), \varphi_2(s)\big\rangle_H ds + \mathbb{E}\int_t^T \big\langle P(s)\varphi_1(s), \psi_2(s)\big\rangle_H ds \\
&\qquad + \mathbb{E}\int_t^T \big\langle P(s)K(s)\varphi_1(s), \tilde{\psi}_2(s)\big\rangle_H ds + \mathbb{E}\int_t^T \big\langle P(s)\tilde{\psi}_1(s), K(s)\varphi_2(s) + \tilde{\psi}_2(s)\big\rangle_H ds \\
&\qquad + \mathbb{E}\int_t^T \big\langle Q(s)\tilde{\psi}_1(s), \varphi_2(s)\big\rangle_H ds + \mathbb{E}\int_t^T \big\langle Q(s)\varphi_1(s), \tilde{\psi}_2(s)\big\rangle_H ds.
\end{aligned}$$

Here, for $j = 1, 2$, φ_j solves (3.8) with ξ, ψ, and $\tilde{\psi}$ replaced by ξ_j, ψ_j, and $\tilde{\psi}_j$, respectively. If (3.7) admits a transposition solution, then it has a relaxed transposition solution (e.g., [40, REMARK 12.11]). Until now, we have no idea how to prove the existence of a transposition solution to (3.7). In such a case, sometimes, we introduce another kind of solution, namely, the V-transposition solution to (3.7), as a substitute (e.g., [12, 32, 38, 39]).

Remark 3.4. Only the first term P of the solution to (3.7) appears in the PMP for Problem (OP). Nevertheless, the characterization of Q has its own interest. On the one hand, Q is used to get higher-order necessary conditions and to solve operator-valued backward stochastic Riccati equations (e.g., [12, 32, 38, 39]). On the other hand, the information about the whole solution helps us understand the first part of the solution.

Theorem 3.2 ([33, THEOREM 6.1]). *Suppose that $L^2_{\mathcal{F}_T}(\Omega; \mathbb{R})$ is separable. Then equation (3.7) admits a unique relaxed transposition solution $(P(\cdot), Q^{(\cdot)}, \hat{Q}^{(\cdot)})$. Furthermore,*

$$|P|_{\mathcal{P}[0,T]} + \big|\big(Q^{(\cdot)}, \hat{Q}^{(\cdot)}\big)\big|_{\mathcal{Q}[0,T]} \leq \mathcal{C}\big(|F|_{L^1_{\mathbb{F}}(0,T;L^2(\Omega;\mathcal{L}(H)))} + |P_T|_{L^2_{\mathcal{F}_T}(\Omega;\mathcal{L}(H))}\big).$$

3.3. Pontryagin-type maximum principle

Let us assume a further condition:

(A3). *For any $(t,u) \in [0,T] \times U$, the maps $a(t,\cdot,u)$, $b(t,\cdot,u)$, $g(t,\cdot,u)$, and $h(\cdot)$ are C^2, such that for $\varphi = a, b$, and $\psi = g, h$, $\varphi_x(t,x,\cdot)$, $\psi_x(t,x,\cdot)$, $\varphi_{xx}(t,x,\cdot)$, and $\psi_{xx}(t,x,\cdot)$ are continuous for any $(t,x) \in [0,T] \times H$. Moreover, there exists a constant $\mathcal{C}_L > 0$ such that*

$$\begin{cases} |\varphi_x(t,x,u)|_{\mathcal{L}(H)} + |\psi_x(t,x,u)|_H \leq \mathcal{C}_L, \\ |\varphi_{xx}(t,x,u)|_{\mathcal{L}(H,H;H)} + |\psi_{xx}(t,x,u)|_{\mathcal{L}(H)} \leq \mathcal{C}_L, \end{cases} \quad \forall (t,x,u) \in [0,T] \times H \times U.$$

Remark 3.5. Condition (A3) is a little restrictive. When the C_0-semigroup $\{S(t)\}_{t\geq 0}$ enjoys some smoothing effect, it can be relaxed (e.g., [37]). Due to (A3), Theorem 3.3 cannot be applied to stochastic linear quadratic optimal control problems for SDPSs directly. Nevertheless, following the proof of Theorem 3.3, we can get the PMP for that problem (e.g., [35]).

Let $\mathbb{H}(t,x,\rho,k_1,k_2) \triangleq \big\langle k_1, a(t,x,\rho)\big\rangle_H + \big\langle k_2, b(t,x,\rho)\big\rangle_H - g(t,x,\rho)$ for $(t,x,\rho, k_1,k_2) \in [0,T] \times H \times U \times H \times H$.

Theorem 3.3. *Suppose that $L^2_{\mathcal{F}_T}(\Omega;\mathbb{R})$ is separable and (A1)–(A3) hold. Let $(\bar{y}(\cdot),\bar{u}(\cdot))$ be an optimal pair of Problem (OP), $(z(\cdot),Z(\cdot))$ be the transposition solution to (3.4) with $F(t,\,z,\,Z) = -a_y(t,\,\bar{y}(t),\,\bar{u}(t))^*z - b_y(t,\,\bar{y}(t),\,\bar{u}(t))^*Z + g_y(t,\,\bar{y}(t),\,\bar{u}(t))$, $z_T = -h_y(\bar{y}(T))$, and $(P(\cdot),Q^{(\cdot)},\hat{Q}^{(\cdot)})$ be the relaxed transposition solution to (3.7) with*

$$\begin{cases} P_T = -h_{yy}(\bar{y}(T)), & J(t) = a_y(t,\bar{y}(t),\bar{u}(t)),\\ K(t) = b_y(t,\bar{y}(t),\bar{u}(t)), & F(t) = -\mathbb{H}_{yy}(t,\bar{y}(t),\bar{u}(t),z(t),Z(t)). \end{cases}$$

Then, for a.e. $(t,\omega)\in[0,T]\times\Omega$ and for all $\rho\in U$,

$$\begin{aligned}&\mathbb{H}\big(t,\bar{y}(t),\bar{u}(t),z(t),Z(t)\big) - \mathbb{H}\big(t,\bar{y}(t),\rho,z(t),Z(t)\big)\\ &\quad - \frac{1}{2}\big\langle P(t)\big[b\big(t,\bar{y}(t),\bar{u}(t)\big) - b\big(t,\bar{y}(t),\rho\big)\big], b\big(t,\bar{y}(t),\bar{u}(t)\big) - b\big(t,\bar{y}(t),\rho\big)\big\rangle_H \geq 0.\end{aligned}$$

Remark 3.6. Compared with Theorem 1.2, the main difference in Theorem 3.3 is the appearance of the term P. This reflects that, in the stochastic situation, the controller has to balance the scale of control and the degree of uncertainty if the control affects the volatility of the system. If b is independent of u, then we do not need P and one adjoint equation, say (3.4), is enough to get the PMP for Problem (OP) (e.g., **[40, THEOREM 12.4]**).

PMP is a necessary condition for optimal controls, which gives a minimum qualification for the candidates of optimal controls. It is natural to ask whether it is also sufficient. To this end, let us introduce the following assumption.

(A4). *The control domain U is a convex subset with a nonempty interior of a separable Hilbert space $\tilde{H}$. The maps a, b, and g are locally Lipschitz in u, and their derivatives in x are continuous in (x,u).*

Theorem 3.4. *Suppose the assumptions of Theorem 3.3 and (A4) hold. Let $u\in\mathcal{U}[0,T]$ and y be the corresponding state of (3.1). Let (z,Z) be the transposition solution to (3.4) with $F(t,\,z,\,Z) = -a_y(t,\,y(t),\,u(t))^*z - b_y(t,\,y(t),\,u(t))^*Z + g_y(t,\,y(t),\,u(t))$, $z_T = -h_y(y(T))$, and $(P(\cdot),Q^{(\cdot)},\hat{Q}^{(\cdot)})$ be the relaxed transposition solution to (3.7) with*

$$\begin{cases} P_T = -h_{yy}(y(T)), & J(t) = a_y(t,y(t),u(t)),\\ K(t) = b_y(t,y(t),u(t)), & F(t) = -\mathbb{H}_{yy}(t,y(t),u(t),z(t),Z(t)). \end{cases}$$

Suppose that $h(\cdot)$ is convex, $\mathbb{H}(t,\cdot,\cdot,z(t),Z(t))$ is concave for all $t\in[0,T]$ a.s., and

$$\begin{aligned}&\mathbb{H}\big(t,y(t),u(t),z(t),Z(t)\big) - \mathbb{H}\big(t,y(t),\rho,z(t),Z(t)\big)\\ &\quad - \frac{1}{2}\big\langle P(t)\big[b\big(t,y(t),u(t)\big) - b\big(t,y(t),\rho\big)\big], b\big(t,y(t),u(t)\big) - b\big(t,y(t),\rho\big)\big\rangle_H \geq 0\end{aligned}$$

for all $\rho\in U$, then $(y(\cdot),u(\cdot))$ is an optimal pair of Problem (OP).

4. OPEN PROBLEMS

SDPSs offers challenges and opportunities for the study of the mathematical control theory. There are many interesting problems in this topic. Some of them are listed below,

which is by no means an exhaustive list and only reflects our research taste. We believe that new mathematical results and even fundamentally new approaches will be required.

(1) Null and approximate controllability of stochastic hyperbolic equations. We have shown that the system (2.1) is not exactly controllable for any $T > 0$ and $\Gamma_0 \subset \Gamma$. It is natural to ask whether it is null/approximately controllable. Of course, for these problems, fewer controls should be employed. The difficulty to do that lies in proving suitable observability estimate of equation (2.3), in which Z and $\hat{Z}$ do not appear in the right-hand side.

(2) Exact controllability for stochastic wave-like equations with more regular controls. Is the system (2.2) exactly controllable when $g \in L^2_{\mathbb{F}}(0,T;L^2(G))$? The desired controllability is equivalent to the following observability estimate:

$$\begin{aligned}&\left|\left(z^T,\hat{z}^T\right)\right|_{L^2_{\mathcal{F}_T}(\Omega;H_0^1(G))\times L^2_{\mathcal{F}_T}(\Omega;L^2(G))}\\&\quad\le \mathcal{C}\left(\left|\frac{\partial z}{\partial \nu}\right|_{L^2_{\mathbb{F}}(0,T;L^2(\Gamma_0))}+|a_5 z+Z|_{L^2_{\mathbb{F}}(0,T;L^2(G))}+|\hat{Z}|_{L^2_{\mathbb{F}}(0,T;L^2(G))}\right),\end{aligned}\tag{4.1}$$

where $(z, Z, \hat{z}, \hat{Z})$ is the solution to (2.3) with $\tau = T$ and final datum $(z^T, \hat{z}^T)$. But one cannot mimic the method in [37] to prove (4.1).

(3) Null/approximate controllability for stochastic parabolic equations with one control. One needs two controls to get the null/approximate controllability for stochastic parabolic equations (e.g., [45]). We believe that one control is enough. However, except for some special cases (e.g., [23,26]), we have no idea on how to prove that.

(4) The cost for the approximate controllability for SDPSs. It is shown in [45] that stochastic parabolic equations are approximately controllable. But it does not give any estimate for the cost of the control. Can one generalize the results in [10] to stochastic parabolic equations? Furthermore, it deserves to study the cost of the approximate controllability for general SDPSs.

(5) Controllability for semilinear SDPSs. In [9], based on sharp estimates on the dependence of controls for the underlying linear equation perturbed by a potential and fixed point arguments, it was proved that semilinear parabolic and hyperbolic equations are null controllable with nonlinearities that grow slower than $s\log(s)^{\frac{3}{2}}$. Whether such results can be obtained for semilinear stochastic parabolic/hyperbolic equations is open. On the other hand, for nonlinearities growing at infinity as $s\log(s)^p$ with $p > 2$, one cannot get the null controllability due to the blow-up of solutions. However, this does not exclude controllability for some particular classes of nonlinear terms (e.g., [7]). More generally, there are lots of interesting results for controllability of semilinear distributed parameter systems (e.g., [6]). So a systematic study of controllability problems for semilinear SDPSs deserves attention.

(6) Stabilization of SDPSs. Stabilization for distributed parameter control systems is a well-studied area. In recent years, some progresses were obtained for SDPSs (e.g., [1,4]). However, this problem is far from being well understood. For example, as far as we know, there is no result for the stabilization of stochastic hyperbolic equations with localized damping.

(7) Optimal control problems for SDPSs with endpoint/state constraints. For some special constraints, such as $y(T)$ belonging to some nonempty open subset of $L^2_{\mathcal{F}_T}(\Omega; H)$, one can use the Ekeland variational principle to establish a Pontryagin-type maximum principle with nontrivial Lagrange multipliers. Nevertheless, for the general case, one does need some further conditions to obtain nontrivial results. For deterministic optimal control problems, people introduce the so-called finite codimensionality condition to guarantee the nontriviality of the Lagrange multiplier (e.g., [20, 25]). There are some attempts to generalize this condition to the stochastic framework (e.g., [24]). Another way is to use some tools from the set-valued analysis (e.g., [12]). However, the existing results are still not satisfactory so far.

(8) Well-posedness of (3.7) in the sense of transposition solution. It would be quite important for some optimal control problems to prove that equation (3.7) admits a unique transposition solution. So far this is only done for a very special case (e.g., [33, THEOREM 4.1]).

(9) Higher-order necessary conditions for optimal controls. Similar to calculus, in addition to the first-order necessary conditions (PMP), sometimes higher-order necessary conditions should be established to distinguish optimal controls from the candidates satisfying the first-order necessary conditions trivially. Some results in this direction for SDPSs can be found in [12, 13, 32]. However, these results were obtained only under very strong assumptions which should be relaxed. To this end, we believe one should first show the existence of a transposition solution to equation (3.7).

(10) Existence of optimal controls. We have discussed the necessary conditions for optimal controls without proving the existence of an optimal control, which is a very difficult problem. There are two general approaches available to study it. One is to prove the verification theorem, the other is to show that a minimizing sequence of controls is compact. Both methods have not been developed well for SDPSs. Except for some trivial cases, such as

- U is a closed and convex subset of a reflective Banach space V, and the functionals g and h are convex and for some $\delta, \mu > 0$,
$$g(x,u,t) \geq \delta|u|_V - \mu, \quad h(x) \geq -\mu, \quad \forall (x,u,t) \in H \times V \times [0,T];$$
- U is a closed, convex and bounded subset of a reflective Banach space V, and the functionals g and h are convex;

there is no further result for that problem.

(11) The relationship between PMP and dynamic programming for SDPSs. PMP and dynamic programming serve as two of the most important tools in solving optimal control problems. Both of them provide some necessary conditions for optimal controls. There should exist a basic link between them. This link is established for finite dimensional stochastic control systems (e.g., [47]). A possible relationship unavoidably involves the derivatives of the value functions, which could be nonsmooth in even very simple cases (e.g., [5]).

(12) The connection between controllability and optimal control. The survey divides itself naturally into two parts—controllability and optimal control. There should be a close

relationship between these two topics. Some initial findings are given in [24], in which a new link between (finite-codimensional exact) controllability and optimal control problems for SDPSs with endpoint state constraints is presented. However, lots of things are to be done, which are by no means easy tasks.

(13) Numerics of the controllability and optimal control problems for SDPSs. By generalizing J.-L. Lions' HUM (e.g., [16]), one can find the numerical solution to controllability problems of SDPSs by solving suitable adjoint equations numerically (e.g., [40, SECTION 7.4]). On the other hand, by Theorem 3.4, one can obtain an optimal control by solving suitable forward–backward stochastic evolution equation. Unfortunately, the numerical approximation of the equations mentioned above can be quite cumbersome. We refer the readers to [30] and references therein for some recent works on this. There are lots of things to be done.

(14) What can we benefit from the uncertainty? From Sections 2 and 3, we see that the uncertainty in SDPSs places many disadvantages for controlling the systems. Nevertheless, sometimes, surprisingly, it provides advantages (e.g., [34, 39]). What can we benefit from the uncertainty in SDPSs is far from being understood. We believe that the study for that problem will lead to new insights into uncertainty.

ACKNOWLEDGMENTS

I would like to express my sincerely gratitude to my mentor, Xu Zhang, who introduced me to this interesting area and provided so many fruitful collaborations and suggestions that have been extremely influential on most of my research work. Also, I have been influenced enormously by five scholars: Jean Michel Coron, Hélène Frankowska, Gengsheng Wang, Jiongmin Yong, and Enrique Zuazua, with whom I had the privilege of working, and I take this opportunity to pay them my highest respect. At last, I would like to mention some of my colleagues, in particular, Xiaoyu Fu, Xu Liu, Jan van Neerven, Kim Dang Phung, Tianxiao Wang, Zhongqi Yin, and Haisen Zhang, with whom I had the opportunity to develop part of the theory and learn so many things.

FUNDING

This work was partially supported by NSF of China under grants 12025105, 11971334, 11931011, by the Chang Jiang Scholars Program from the Ministry of Education of the People's Republic of China, and by the Science Development Project of Sichuan University under grants 2020SCUNL101 and 2020SCUNL201.

REFERENCES

[1] V. Barbu and G. Da Prato, Internal stabilization by noise of the Navier–Stokes equation. *SIAM J. Control Optim.* **49** (2011), no. 1, 1–20.

[2] C. Bardos, G. Lebeau, and J. Rauch, Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary. *SIAM J. Control Optim.* **30** (1992), no. 5, 1024–1065.

[3] A. Bensoussan, Stochastic maximum principle for distributed parameter systems. *J. Franklin Inst.* **315** (1983), no. 5–6, 387–406.

[4] S. K. Biswas and N. U. Ahmed, Stabilization of systems governed by the wave equation in the presence of distributed white noise. *IEEE Trans. Automat. Control* **30** (1985), no. 10, 1043–1045.

[5] L. Chen and Q. Lü, Relationships between the Maximum Principle and Dynamic Programming for infinite dimensional stochastic control systems. 2021, arXiv:2112.14636.

[6] J.-M. Coron, *Control and nonlinearity*. American Mathematical Society, Providence, RI, 2007.

[7] J.-M. Coron and E. Trélat, Global steady-state controllability of one-dimensional semilinear heat equations. *SIAM J. Control Optim.* **43** (2004), no. 2, 549–569.

[8] F. Dou and Q. Lü, Partial approximate controllability for linear stochastic control systems. *SIAM J. Control Optim.* **57** (2019), no. 2, 1209–1229.

[9] T. Duyckaerts, X. Zhang, and E. Zuazua, On the optimality of the observability inequalities for parabolic and hyperbolic systems with potentials. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **25** (2008), no. 1, 1–41.

[10] E. Fernández-Cara and E. Zuazua, The cost of approximate controllability for heat equations: the linear case. *Adv. Differential Equations* **5** (2000), no. 4–6, 465–514.

[11] W. H. Fleming (ed.), Future directions in control theory: a mathematical perspective, SIAM, Philadelphia, PA, 1988.

[12] H. Frankowska and Q. Lü, First and second order necessary optimality conditions for controlled stochastic evolution equations with control and state constraints. *J. Differential Equations* **268** (2020), no. 6, 2949–3015.

[13] H. Frankowska and X. Zhang, Necessary conditions for stochastic optimal control problems in infinite dimensions. *Stochastic Process. Appl.* **130** (2020), no. 7, 4081–4103.

[14] X. Fu and X. Liu, A weighted identity for stochastic partial differential operators and its applications. *J. Differential Equations* **262** (2017), no. 6, 3551–3582.

[15] X. Fu, Q. Lü, and X. Zhang, *Carleman estimates for second order partial differential operators and applications. A unified approach*. Springer, Cham, 2019.

[16] R. Glowinski, J.-L. Lions, and J. He, *Exact and approximate controllability for distributed parameter systems. A numerical approach*. Cambridge University Press, Cambridge, 2008.

[17] Y. Hu and S. Peng, Maximum principle for semilinear stochastic evolution control systems. *Stoch. Stoch. Rep.* **33** (1990), no. 3–4, 159–180.

[18] R. E. Kalman, Contributions to the theory of optimal control. *Bol. Soc. Mat. Mex.* **5** (1960), 102–119.

[19] P. Kotelenez, *Stochastic ordinary and stochastic PDEs. Transition from microscopic to macroscopic equations*. Springer, New York, 2008.

[20] X. Li and J. Yong, *Optimal control theory for infinite-dimensional systems*. Birkhäuser Boston, Inc., Boston, MA, 1995.

[21] J.-L. Lions, Exact controllability, stabilization and perturbations for distributed systems. *SIAM Rev.* **30** (1988), no. 1, 1–68.

[22] J.-L. Lions and E. Magenes, *Non-homogeneous boundary value problems and applications. I*. Springer, New York–Heidelberg, 1972.

[23] X. Liu, Controllability of some coupled stochastic parabolic systems with fractional order spatial differential operators by one control in the drift. *SIAM J. Control Optim.* **52** (2014), no. 2, 836–860.

[24] X. Liu, Q. Lü, H. Zhang, and X. Zhang, Finite codimensionality technique in optimization and optimal control problems. 2021, arXiv:2102.00652.

[25] X. Liu, Q. Lü, and X. Zhang, Finite codimensional controllability and optimal control problems with endpoint state constraints. *J. Math. Pures Appl.* **138** (2020), 164–203.

[26] Q. Lü, Some results on the controllability of forward stochastic parabolic equations with control on the drift. *J. Funct. Anal.* **260** (2011), no. 3, 832–851.

[27] Q. Lü, Exact controllability for stochastic Schrödinger equations. *J. Differential Equations* **255** (2013), no. 8, 2484–2504.

[28] Q. Lü, Observability estimate and state observation problems for stochastic hyperbolic equations. *Inverse Probl.* **29** (2013), no. 9, 095011, 22 pp.

[29] Q. Lü, Exact controllability for stochastic transport equations. *SIAM J. Control Optim.* **52** (2014), no. 1, 397–419.

[30] Q. Lü, P. Wang, Y. Wang, and X. Zhang, Numerics for stochastic distributed parameter control systems: a finite transposition method. 2021, arXiv:2104.02964.

[31] Q. Lü, J. Yong, and X. Zhang, Representation of Itô integrals by Lebesgue/Bochner integrals. *J. Eur. Math. Soc. (JEMS)* **14** (2012), no. 6, 1795–1823 (Erratum: *J. Eur. Math. Soc. (JEMS)* **20** (2018), no. 1, 259–260).

[32] Q. Lü, H. Zhang, and X. Zhang, Second order necessary conditions for optimal control problems of stochastic evolution equations. *SIAM J. Control Optim.* **59** (2021), no. 4, 2924–2954.

[33] Q. Lü and X. Zhang, *General Pontryagin-type stochastic maximum principle and backward stochastic evolution equations in infinite dimensions*. Springer, Cham, 2014.

[34] Q. Lü and X. Zhang, Global uniqueness for an inverse stochastic hyperbolic problem with three unknowns. *Comm. Pure Appl. Math.* **68** (2015), no. 6, 948–963.

[35] Q. Lü and X. Zhang, Transposition method for backward stochastic evolution equations revisited, and its application. *Math. Control Relat. Fields* **5** (2015), no. 3, 529–555.

[36] Q. Lü and X. Zhang, Operator-valued backward stochastic Lyapunov equations in infinite dimensions, and its application. *Math. Control Relat. Fields* **8** (2018), no. 1, 337–381.

[37] Q. Lü and X. Zhang, Exact controllability for a refined stochastic wave equation. 2019, arXiv:1901.06074.

[38] Q. Lü and X. Zhang, Optimal feedback for stochastic linear quadratic control and backward stochastic Riccati equations in infinite dimensions. *Mem. Amer. Math. Soc.* (accepted), arXiv:1901.00978.

[39] Q. Lü and X. Zhang, Control theory for stochastic distributed parameter systems, an engineering perspective. *Annu. Rev. Control* **51** (2021), 268–330.

[40] Q. Lü and X. Zhang, *Mathematical control theory for stochastic partial differential equations*. Springer, Switzerland AG, 2021.

[41] R. M. Murray (ed.), *Control in an information rich world*. SIAM, Philadelphia, PA, 2003.

[42] S. Peng, A general stochastic maximum principle for optimal control problems. *SIAM J. Control Optim.* **28** (1990), 966–979.

[43] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *Mathematical theory of optimal processes*. Wiley, New York, 1962.

[44] S. Tang and X. Li, Maximum principle for optimal control of distributed parameter stochastic systems with random jumps. In *Differential equations, dynamical systems, and control science*, pp. 867–890, Dekker, New York, 1994.

[45] S. Tang and X. Zhang, Null controllability for forward and backward stochastic parabolic equations. *SIAM J. Control Optim.* **48** (2009), no. 4, 2191–2216.

[46] J. M. A. M. van Neerven, M. C. Veraar, and L. W. Weis, Stochastic integration in UMD Banach spaces. *Ann. Probab.* **35** (2007), no. 4, 1438–1478.

[47] J. Yong and X. Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*. Springer, New York, 1999.

[48] X. Zhang, Carleman and observability estimates for stochastic wave equations. *SIAM J. Math. Anal.* **40** (2008), no. 2, 851–868.

[49] X. Zhang, A unified controllability/observability theory for some stochastic and deterministic partial differential equations. In *Proceedings of the International Congress of Mathematicians. IV*, pp. 3008–3034, Hindustan Book Agency, New Delhi, 2010.

[50] E. Zuazua, Controllability and observability of partial differential equations: some results and open problems. In *Handbook of differential equations: evolutionary differential equations. 3*, pp. 527–621, Elsevier Science, 2006.

QI LÜ

School of Mathematics, Sichuan University, Chengdu 610065, Sichuan Province, China, lu@scu.edu.cn

INDEPENDENT LEARNING IN STOCHASTIC GAMES

ASUMAN OZDAGLAR, MUHAMMED O. SAYIN, AND KAIQING ZHANG

ABSTRACT

Reinforcement learning (RL) has recently achieved tremendous successes in many artificial intelligence applications. Many of the forefront applications of RL involve *multiple agents*, e.g., playing chess and Go games, autonomous driving, and robotics. Unfortunately, the framework upon which classical RL builds is inappropriate for multiagent learning, as it assumes an agent's environment is stationary and does not take into account the adaptivity of other agents. In this review paper, we present the model of *stochastic games* [69] for multiagent learning in *dynamic* environments. We focus on the development of *simple* and *independent* learning dynamics for stochastic games: each agent is myopic and chooses best-response type actions to other agents' strategy without any coordination with her opponent. There has been limited progress on developing convergent best-response type independent learning dynamics for stochastic games. We present our recently proposed simple and independent learning dynamics that guarantee convergence in zero-sum stochastic games, together with a review of other contemporaneous algorithms for dynamic multiagent learning in this setting. Along the way, we also reexamine some classical results from both the game theory and RL literature, to situate both the conceptual contributions of our independent learning dynamics, and the mathematical novelties of our analysis. We hope this review paper serves as an impetus for the resurgence of studying independent and natural learning dynamics in game theory, for the more challenging settings with a dynamic environment.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 91A15; Secondary 91A25, 68T05

KEYWORDS

Stochastic games, learning in games, reinforcement learning

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5340–5373
DOI 10.4171/ICM2022/152

1. INTRODUCTION

Reinforcement learning (RL) in which autonomous agents make decisions in unknown dynamic environments has emerged as the backbone of many artificial intelligence (AI) problems. The frontier of many AI systems emerges in *multiagent* settings, including playing games such as chess and Go [73,74], robotic manipulation with multiple connected arms [30], autonomous vehicle control in dynamic traffic and automated warehouses or production facilities [68,86]. Further advances in these problems critically depend on developing stable and agent incentive-compatible learning dynamics in multiagent environment. Unfortunately, the mathematical framework upon which classical RL depends on is inadequate for multiagent learning, since it assumes an agent's environment is stationary and does not contain any adaptive agents.

The topic of multiagent learning has a long history in game theory, almost as long as the discipline itself. One of the most studied models of learning in games is *fictitious play*, introduced by Brown [14], with first rigorous convergence analysis presented by Robinson [59] for its discrete-time variant and for finite two-player zero-sum games. See also [27,28,33,49,51,70] and others for the analysis of fictitious play. In fictitious play, each agent is myopic (i.e., she does not take into account the fact that her current action will have an impact on the future actions of other players[1]), and therefore chooses a best response to the opponent's strategy, which she estimates to be the empirical distribution of past play. Despite extensive study on learning in repeated play of static complete-information games (also referred to as strategic- or normal-form games) and the importance of the issues, there is limited progress on multiagent learning in dynamic environments (where the environments evolve over time). The key challenge is to estimate the decision rules of other agents that in turn *adapt* their behavior to changing nonstationary environments.

In this review paper, we first present stochastic games, first introduced in [69], as a model for representing dynamic multiagent interactions in Section 3.[2] Stochastic games extend strategic-form games to dynamic settings where the environment changes with players' decisions. They also extend single-agent Markov decision problems (Markov decision processes) to competitive situations with more than one decision-maker. Developing simple and independent learning rules, e.g., the fictitious-play/best-response type dynamics, for stochastic games has been an open question for some time in the literature (see [19,20,78] for some negative nonconvergent results due to nonstationarity).

In the second part of the paper in Section 4, we present recently proposed simple and independent learning rules from [63,64], and show their convergence for zero-sum stochastic games. Crucially, these rules are based on fictitious play-type dynamics and, unlike earlier works, do not require coordination between agents, leading to fully decentralized and independent multiagent learning dynamics. We combine ideas from *g*ame theory and *R*L in developing these learning rules, and consider three different settings: *model-based* setting

1 Hereafter, we use *player* and *agent* interchangeably.

2 The preliminary information on strategic-form games and learning in strategic-form games with repeated play are provided in Section 2.

where players know their payoff functions, transition probabilities of the underlying stochastic games, and observe opponent's actions; *model-free* setting where players do not know payoff functions and transition probabilities but can still observe the opponent's actions; and the *minimal information* setting where players do not even observe opponent's actions. In all three settings, the players do not know the opponent's objective, i.e., they do not possess the knowledge that the underlying game is zero-sum. In the minimal-information setting, the players may not even know the existence of an opponent.

In Section 5, we have also reviewed several other algorithms/learning dynamics, and their convergence results for multiagent learning in stochastic games. We cover both results from the game theory literature that typically assumes knowledge of the model of the players' payoff functions, and the transition probabilities of the underlying stochastic games, and also from the RL literature which posit learning dynamics that perform updates without knowing the transition probabilities. Most of these update rules typically involve coordination and computationally intensive steps for the players. These algorithms can be viewed more as ones for *computing* the Nash equilibrium of the stochastic games, as opposed to natural learning dynamics that would be adopted by self-interested agents interested in maximizing their own payoffs given their inferences (as captured in our learning dynamics). Finally, we conclude the paper with open questions on independent learning in stochastic games in Section 6.

2. PRELIMINARIES: STRATEGIC-FORM GAMES

A two-player strategic-form game can be characterized by a tuple $\langle A^1, A^2, r^1, r^2 \rangle$, in which

- the *finite* set of actions that player i can take is denoted by A^i,
- the *payoff function* of player i is denoted by $r^i : A \to \mathbb{R}$, where $A := A^1 \times A^2$.[3]

Each player i takes an action from her action set A^i *simultaneously* and receives the payoff $r^i(a^1, a^2)$.

We let players choose a mixed strategy to randomize their actions independently. For example, $\pi^i : A^i \to [0, 1]$ denotes the mixed strategy of player i such that $\pi^i(a^i)$ corresponds to the probability that player i plays a^i. Note that we have $\sum_{a^i \in A^i} \pi^i(a^i) = 1$ by its definition.

We represent the strategy profile and action profile of the players by $\pi = (\pi^1, \pi^2)$ and $a = (a^1, a^2)$, respectively. Under the strategy profile π, the expected payoff of player i is defined by

$$U^i(\pi) := \mathbb{E}_{a \sim \pi}\{r^i(a)\}.$$

3 We can generalize the definition to arbitrary number of players in a rather straightforward way.

Note that the expected payoff of player i is affected by the strategy of the opponent. We next introduce the Nash equilibrium where players do not have any (or large enough) incentive to change their strategies unilaterally.

Definition 2.1 ($(\varepsilon$-)Nash equilibrium). A strategy profile π_* is a mixed-strategy ε-Nash equilibrium with $\varepsilon \geq 0$ if we have

$$U^1\left(\pi_*^1, \pi_*^2\right) \geq U^1\left(\pi^1, \pi_*^2\right) - \varepsilon, \quad \text{for all } \pi^1, \tag{2.1a}$$

$$U^2\left(\pi_*^1, \pi_*^2\right) \geq U^2\left(\pi_*^1, \pi^2\right) - \varepsilon, \quad \text{for all } \pi^2. \tag{2.1b}$$

Furthermore, π_* is a mixed-strategy Nash equilibrium if (2.1) holds with $\varepsilon = 0$.

The following is the classical existence result for any strategic-form game (e.g., see [4, THEOREM 3.2]).

Theorem 2.2 (Existence of an equilibrium in strategic-form games). *In strategic-form games (with finitely many players and finitely many actions), a mixed-strategy equilibrium always exists.*

The key question is whether an equilibrium can be realized or not in the interaction of self-interested decision-makers. In general, finding the best strategy against another decision-maker is not a well-defined optimization problem because the best strategy that reflects the viewpoint of the individual depends on the opponent's strategy. Therefore, players are generally not able to compute their best strategy beforehand. When there exists a unique equilibrium, we can expect the players to identify their equilibrium strategies as a result of an introspective thinking process. For example, what would the opponent choose? What would the opponent have chosen if she knew I am considering what she would pick while choosing my strategy? And so on. However, many empirical analyses suggest that an equilibrium would not typically be realized in one shot even with such reasoning (see, e.g., [29]).

It is instructive to consider the following well-known example: Consider a game played among $n > 1$ students. The teacher asks the students to pick a number between 0 and 100, and submit it within a closed envelope. The winner will be the one who chooses the number closest to the two-thirds of the average of all numbers picked. It can be seen that the unique equilibrium is the strategy profile where every player chooses 0. We would expect the students to pick 0 as a result of an introspective thinking process, however, empirical studies show that they typically pick numbers other than zero such that their average ends up around 30, with its two-thirds around 20 [53]. This results in players who have selected 0 by strategizing their actions introspectively losing the game. However, if the game is played repeatedly with players observing chosen actions, each player will have a tendency to pick numbers closer to the winning number (or its two-thirds if they notice that others can also have such a tendency to pick the number closest to the winning one). This results in convergence to the equilibrium play along repeated play of the game, even when the players have not engaged in any forward-looking strategy.

Many games have multiple equilibria which makes coordination and selection through introspective thinking challenging. On the other hand, empirical studies suggest even in strategic situations equipped with multiple equilibria, individual agents reach an equilibrium as long as they *engage with each other multiple times and receive feedback* to revise their strategies [29].

In the following, we review the canonical models of learning with multiple agents through repeated interactions.

2.1. Learning in strategic-form games with repeated play

Suppose that players know the primitives of the game, i.e., $\langle A^1, A^2, r^1, r^2\rangle$. If players knew the opponent's strategy, computation of the best strategy is a simple optimization problem where they pick one of the maxima among linearly ordered finitely many elements. However, players do not know the opponent's strategy. When they play the same game repeatedly and observe the opponent's actions in these games, they have a chance to reason about what the opponent would play in the next repetition of the game. Therefore, they can estimate the opponent's strategy based on the history of the play. However, the opponent is not necessarily playing according to a stationary strategy since she is also a strategic decision-maker who can adapt her strategy according to her best interest.

Fictitious play is a simple and stylist learning dynamic where players (erroneously) assume that the opponent plays according to a stationary strategy.[4] This assumption lets players form a belief on the opponent's strategy based on the history of the play, e.g., the empirical distribution of the actions taken. Then, the players can adapt their strategies based on the belief constructed.

Fictitious play, since its first introduction by [14], has become the most appealing best-response type learning dynamics in game theory. Formally, at iteration k, player i maintains a *belief* on the opponent's strategy, denoted by $\hat{\pi}_k^{-i} \in \Delta(A^{-i})$.[5] For example, the belief can correspond to the empirical average of the actions taken in the past. Note that we can view an action a^i as a deterministic strategy in which the action is played with probability 1, i.e., $a^i \in \Delta(A^i)$ with slight abuse of notation. Then, the empirical average is given by

$$\hat{\pi}_{k+1}^{-i} = \frac{1}{k+1}\sum_{\kappa=0}^{k} a_\kappa^{-i}. \tag{2.2}$$

The belief $\hat{\pi}_{k+1}^{-i}$ can be computed iteratively using bounded memory according to

$$\hat{\pi}_{k+1}^{-i} = \hat{\pi}_k^{-i} + \frac{1}{k+1}\cdot\left(a_k^{-i} - \hat{\pi}_k^{-i}\right), \tag{2.3}$$

with arbitrary initialization $\hat{\pi}_0^{-i} \in \Delta(A^{-i})$. In other words, players do not have to remember every action taken by the opponent in the past. Moreover, player i selects her action following

$$a_k^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim\hat{\pi}_k^{-i}}\left\{r^i(a^1, a^2)\right\}, \tag{2.4}$$

---

4 It is called *fictitious play* because [14] introduced it as an introspective thinking process that a player can play by herself.

5 We represent the probability simplex over a set A by $\Delta(A)$.

with an arbitrary tie-breaking rule, playing a greedy best-response to the belief she maintains on opponent's strategy.

We say that fictitious play dynamics *converge to an equilibrium* if beliefs formed converge to a Nash equilibrium when all players follow the fictitious play dynamics (2.3)–(2.4). We also say that a class of games has *fictitious play property* if fictitious play converges in every game of that class. The following theorem is about two important classes of games from two extremes of the game spectrum: two-player zero-sum strategic-form games, where $r^1(a) + r^2(a) = 0$ for all $a \in A$, and n-player identical-interest strategic-form games, where there exists a common payoff function $r : A \to \mathbb{R}$ such that $r^i(a) = r(a)$ for all $a \in A$ and for each player i.

Theorem 2.3 (Fictitious play property of zero-sum and identical-interest games).

- *The two-player zero-sum strategic-form games have fictitious play property* **[59]**.
- *The n-player identical-interest strategic-form games have fictitious play property* **[51]**.

As an alternative to the insightful proofs in **[59]** and **[51]**, we can establish a connection between fictitious play and continuous-time best response dynamics to characterize its convergence properties. For example, **[31]** provided a proof for the continuous-time best-response dynamics in zero-sum strategic-form games through a Lyapunov function formulation. This convergence result also implies the convergence of fictitious play in repeated play of the same zero-sum strategic-form game. We next briefly describe the approach in **[31]** to convergence analysis for continuous-time best-response dynamics.

In continuous-time best response dynamics, the strategies (π^1, π^2) evolve according to the following differential inclusion:

$$\frac{d\pi^i}{dt} + \pi^i \in \underset{a^i \in A^i}{\operatorname{argmax}} \mathbb{E}_{a^{-i} \sim \pi^{-i}} \left\{ r^i(a^1, a^2) \right\} \tag{2.5}$$

for $i = 1, 2$. We highlight the resemblance between (2.3) and (2.5) because we can view (2.5) as the limiting flow of (2.3) as $1/(k+1) \to 0$. Note also that there exists an absolutely continuous solution to this differential inclusion **[31]**. To characterize the convergence properties of this flow, **[31]** showed that the function

$$V(\pi) = \sum_{i=1,2} \left(\max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}} \left\{ r^i(a^1, a^2) \right\} - \mathbb{E}_{a \sim \pi} \left\{ r^i(a) \right\} \right) \tag{2.6}$$

is a Lyapunov function when $r^1(a) + r^2(a) = 0$ for all $a \in A$.[6] This yields that $V(\pi(t)) \geq V(\pi(t'))$ for all $t' > t$ and $V(\pi(t)) > V(\pi(t'))$ if $V(\pi(t)) > 0$. Correspondingly, we have $V(\pi(t)) \to 0$ as $t \to \infty$. This implies that the continuous-time best response dynamics converge to the equilibrium of the zero-sum game. Since the terms in parentheses

6 Note that $\mathbb{E}_{a \sim \pi}\{r^1(a)\} + \mathbb{E}_{a \sim \pi}\{r^2(a)\} = 0$ when $r^1(a) + r^2(a) = 0$ for all $a \in A$.

in (2.6) are nonnegative, $V(\pi) = 0$ yields that they are equal to zero for each $i = 1, 2$, which is indeed the definition of the Nash equilibrium.

Generally, the convergence of the limiting flow would not lead to the convergence of the discrete-time update. However, based on tools from differential inclusion approximation theory [5], the existence of such a Lyapunov function yields that the fictitious play dynamics converge to an equilibrium since its linear interpolation after certain transformation of the time axis can be viewed as a perturbed solution to the differential inclusion (2.5) with asymptotically negligible perturbation while the existence of Lyapunov function yields that any such perturbed solution also converges to the zero-set of the Lyapunov function, i.e., $\{\pi : V(\pi) = 0\}$.

The fictitious play dynamics enjoy the following desired properties [29]: (i) The dynamics do not require knowledge of the underlying game's class, e.g., the opponent's payoff function, and is not specific to any specific class of games; (ii) Players attain the best-response performance against an opponent following an asymptotically stationary strategy, i.e., the learning dynamics is rational; (iii) If the dynamics converge, it must converge to an equilibrium of the underlying game.

Unfortunately, there exist strategic-form games that do not have fictitious play property as shown by [70] through a counterexample. The classes of strategic-form games with fictitious play property have been studied extensively, e.g., see [6, 7, 48–52, 59, 65]. Variants of fictitious play, including smoothed fictitious play [27] and weakened fictitious play [81] have also been studied extensively. However, all these studies focus on the repeated play of *the same* strategic-form game at every stage. There are very limited results on dynamic games where players interact repeatedly while the game played at a stage (called *stage-game*) evolves with their actions. Note that players need to consider the impact of their actions in their future payoffs as in dynamic programming or optimal control when they have utilities defined over infinite horizon.

In the next section, we introduce stochastic games, a special (and important class) of dynamic games where the stage-games evolve over infinite horizon based on the current actions of players.

3. STOCHASTIC GAMES

Stochastic games (also known as *Markov* games), since their first introduction by Shapley [69], have been widely used as a canonical model for dynamic multiagent interactions (e.g., see the surveys [16, 89]). At each time $k = 0, 1, \ldots$, players play a stage game that corresponds to a particular state of a multistate environment. The stage games evolve stochastically according to the transition probabilities of the states controlled jointly by the actions of both players. The players receive a payoff which is some aggregate of the stage payoffs; a typical model is to assume the players receive a discounted sum of stage payoffs over an infinite horizon.

Formally, a two-player stochastic game is characterized by a tuple $\langle S, A^1, A^2, r^1, r^2, p, \gamma \rangle$, in which:

- The *finite* set of states is denoted by S.
- The *finite* set of actions that player i can take at any state is denoted by A^i.[7]
- The *stage payoff function* of player i is denoted by $r^i : S \times A \to \mathbb{R}$, where $A = A^1 \times A^2$.
- For any pair of states (s, s') and action profile $a \in A$, we define $p(s'|s, a)$ as the *transition probability* from s to s' given action profile a.
- The players also discount the impact of future payoff in their utility with the discount factor $\gamma \in [0, 1)$.

The objective of player i is to maximize the expected sum of discounted stage-payoffs collected over infinite horizon, given by

$$\mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k)\right\}, \tag{3.1}$$

where $a_k \in A$ denotes the action profile played at stage k, $\{s_0 \sim p_o, s_{k+1} \sim p(\cdot \mid s_k, a_k), k \geq 0\}$ is a stochastic process representing the state at each stage k and $p_o \in \Delta(S)$ is the initial state distribution. The expectation is taken with respect to randomness due to stochastic state transitions and actions mixed independently by the players.

The players can play an *infinite* sequence of (mixed) actions. When they have perfect recall, they can mix their actions independently according to a *behavioral strategy* in which the probability of an action is taken depends on the history of states and action profiles, e.g., $h_k = \{s_0, a_0, s_1, a_1, \ldots, s_{k-1}, a_{k-1}, s_k\}$ at stage k. This results in an infinite-dimensional strategy space, and therefore, the universal result for the existence of an equilibrium, Theorem 2.2, does not apply here. On the other hand, stochastic games can also be viewed as a generalization of Markov decision processes (MDPs) to multiagent cases since state transition probabilities depend only on the current state and current action profile of players. Behavioral strategies that depend only on the final state of the history (which corresponds to the current state) are known as *Markov strategies*. Furthermore, we call a Markov strategy by a *stationary strategy* if it does not depend on the stage, e.g., see [71, SECTION 6.2]. In (discounted) MDPs, there always exists an optimal strategy that is stationary, e.g., see [23]. Shapley [69] showed that this can be generalized to two-player zero-sum stochastic games.

We denote the stationary mixed strategy of player i by $\pi^i : S \to \Delta(A^i)$, implying that she takes actions according to the mixed strategy specific to state s, i.e., $\pi^i(s) \in \Delta(A^i)$. We represent the strategy profile of players by $\pi := \{\pi^1, \pi^2\}$. Correspondingly, the expected discounted sum of stage payoffs of player i under the strategy profile π is defined by

$$U^i(\pi) := \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r^i(s_k, a_k)\right\}, \tag{3.2}$$

7 The formulation can be generalized to the case where the action spaces depend on state in a rather straightforward way.

where $a_k \sim \pi(s_k)$, and the expectation is taken with respect to the all randomness. We next introduce the Nash equilibrium (more specifically, Markov perfect equilibrium [44,45]) where players do not gain any utility improvement by unilateral changes in their *stationary* strategies regardless of the initial state, e.g., see [71, SECTION 6.2].

Definition 3.1 (Stationary (ε-)Nash equilibrium). We say that a stationary strategy profile π is a stationary mixed-strategy ε-Nash equilibrium with $\varepsilon \geq 0$ if we have

$$U^1(\pi^1, \pi^2) \geq U^1(\bar{\pi}^1, \pi^2) - \varepsilon \quad \text{for all } \bar{\pi}^1, \tag{3.3a}$$

$$U^2(\pi^1, \pi^2) \geq U^2(\pi^1, \bar{\pi}^2) - \varepsilon \quad \text{for all } \bar{\pi}^2. \tag{3.3b}$$

We say that π is a stationary mixed-strategy Nash equilibrium if (3.3) holds with $\varepsilon = 0$.

We next state an important existence result for discounted stochastic games.

Theorem 3.2 (Existence of a stationary equilibrium in stochastic games [24]). *In stochastic games (with finitely many players, states, and actions, and discount factor $\gamma \in [0, 1)$), a stationary mixed-strategy equilibrium always exists.*

The proof for two-player zero-sum stochastic games is shown by Shapley [69] while its generalization to n-player general-sum stochastic games is proven by Fink [24] and Takahashi [77] concurrently. Shapley [69] had also presented an iterative algorithm to compute the unique equilibrium value of a two-player zero-sum stochastic game. To describe the algorithm, let us first note that in a zero-sum strategic-form game, there always exists a unique equilibrium *value* for the players (though there may exist multiple equilibria). For example, given a zero-sum strategic-form game $\langle A^1, A^2, u^1, u^2 \rangle$, we denote the equilibrium values of player 1 and player 2, respectively, by

$$\operatorname{val}^1[u^1] = \max_{\pi^1 \in \Delta(A^1)} \min_{\pi^2 \in \Delta(A^2)} \mathbb{E}_{a \sim (\pi^1, \pi^2)}\{u^1(a)\}, \tag{3.4}$$

$$\operatorname{val}^2[u^2] = \max_{\pi^2 \in \Delta(A^2)} \min_{\pi^1 \in \Delta(A^1)} \mathbb{E}_{a \sim (\pi^1, \pi^2)}\{u^2(a)\}. \tag{3.5}$$

It is instructive to examine the following thought experiment. Imagine that players are at the edge of the infinite horizon. Then the players' continuation payoff would be determined by the stage game at state s since there would not be any future stages to consider. The unique equilibrium values they would get would be $\operatorname{val}^i[r^i(s, \cdot)]$. Then, at the stage just before the last one, they would have played the strategic-form game $\langle A^1, A^2, Q^1(s, \cdot), Q^2(s, \cdot) \rangle$ at state s, where

$$Q^i(s, \cdot) = r^i(s, \cdot) + \gamma \sum_{s' \in S} p(s'|s, \cdot) \operatorname{val}^i[r^i(s', \cdot)]. \tag{3.6}$$

Shapley [69] showed that if we follow this *backward induction*, we can always compute the equilibrium values associated with a stationary equilibrium. To this end, he introduced the operator $\mathcal{T}^i$ defined by

$$(\mathcal{T}^i v^i)(s) := \operatorname{val}^i\left[r^i(s, \cdot) + \gamma \sum_{s' \in S} p(s|s, \cdot) v^i(s') \right], \quad \forall s \in S, \tag{3.7}$$

which is a contraction with respect to the ℓ_∞-norm when $\gamma \in (0,1)$ since val^i is a nonexpansive mapping, i.e.,

$$\left|\mathrm{val}^i(u^i) - \mathrm{val}^i(\tilde{u}^i)\right| \leq \max_{a \in A} \left|u^i(a) - \tilde{u}^i(a)\right|,$$

for any $u^i : A \to \mathbb{R}$ and $\tilde{u}^i : A \to \mathbb{R}$, similar to the maximum function in the Bellman operator. Therefore, the iteration

$$v^i_{(n+1)} = \mathcal{T}^i v^i_{(n)}, \quad \forall n \geq 0, \tag{3.8}$$

starting from arbitrary $v^i_{(0)}$ converges to the unique fixed point of the operator. Further inspection of the fixed point reveals that it is indeed the equilibrium values of states associated with some stationary equilibrium of the underlying two-player zero-sum stochastic game. There does not exist a counterpart of this iteration for the computation of equilibrium values in general-sum stochastic games, since the value of a game is not uniquely defined for general-sum stochastic games, and involves a fixed point operation, which is hard to compute at each stage of an algorithm. However, Shapley's iteration is still a powerful method to compute equilibrium values in a two-player zero-sum stochastic game.

In the following section, we examine whether a stationary equilibrium would be realized as a consequence of nonequilibrium adaptation of learning agents as in Section 2.1 but now for stochastic games instead of repeated play of the same strategic-form game.

4. LEARNING IN STOCHASTIC GAMES

Fictitious play dynamics is a best-response type learning dynamics where each player aims to take the best response against the opponent by learning the opponent's strategy based on the history of the play. This stylist learning dynamic can be generalized to stochastic games as players (again erroneously) assume that the opponent plays according to a *stationary* strategy (which depends only on the current state). Hence, they can again form a belief on the opponent's stationary strategy based on the history of the play. Particularly, they can form a belief on the opponent's mixed strategy specific to a state based on the actions taken at that state only due to the stationarity assumption on the opponent's strategy. Given that belief on the opponent's strategy, players can also compute the value of each state-action pair based on *backward induction* since their actions determine both the stage payoff and the continuation payoff by determining the state transitions. Therefore, they essentially play an *auxiliary stage-game* at each stage specific to the current state, which can be represented by $\mathcal{G}_s := \langle A^1, A^2, Q^1(s,\cdot), Q^2(s,\cdot)\rangle$, where the payoff or the *Q-function*, $Q^i(s,\cdot) : A \to \mathbb{R}$ is determined according to the backward induction given π^{-i} the belief of player i about player $-i$'s strategy, and therefore, it satisfies the following fixed-point equation:

$$Q^i(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p\big(s'|s,a\big) \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}(s')}\big\{Q^i(s', a^1, a^2)\big\}. \tag{4.1}$$

For notational convenience, we also define the *value function* $v^i : S \to \mathbb{R}$ by

$$v^i(s) := \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi^{-i}(s)}\big\{Q^i(s, a^1, a^2)\big\}. \tag{4.2}$$

At each stage k, player i has a belief on player $-i$'s strategy, which we denote by $\hat{\pi}_k^{-i}$. Player i also forms a belief on the payoff function for the auxiliary game, or the Q-function, denoted by $\hat{Q}_k^i$. Let s be the current state of the stochastic game. Then, player i selects her action a_k^i according to

$$a_k^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \hat{\pi}_k^{-i}(s)} \{\hat{Q}_k^i(s, a^1, a^2)\}. \tag{4.3}$$

Observing the opponent's action a_k^{-i}, player i forms her belief on player $-i$'s strategy for the current state s as a weighted empirical average, which can be constructed iteratively as

$$\hat{\pi}_{k+1}^{-i}(s) = \hat{\pi}_k^{-i}(s) + \alpha_{c_k(s)}\big(a_k^{-i} - \hat{\pi}_k^{-i}(s)\big). \tag{4.4}$$

Here $\alpha_c \in [0, 1]$ is a step size and it vanishes with $c_k(s)$ indicating the number of visits to state s rather than time. Note that if there was a single state, $c_k(s)$ would correspond to the time, i.e., $c_k(s) = k$, as in the classical fictitious play. The update (4.4) can also be viewed as taking a convex combination of the current belief $\hat{\pi}_k^2(s)$ and the observed action a_k^2 while the step size $\alpha_{c_k(s)}$ is the (vanishing) weight of the action observed. Vanishing step size as a function of the number of visits implies that, the players give less weight to their current belief than the observed action by using a large step size if that state has not been visited many times. This means that the players will still give less weight to their current belief even at later stages if the specific state has not been visited many times, and indicating, they have not been able to strengthen their belief enough to rely more on it.

Simultaneously, player i updates her belief on her own Q-function for the current state s according to

$$\hat{Q}_{k+1}^i(s,a) = \hat{Q}_k^i(s,a) + \beta_{c_k(s)}\left(r^i(s,a) + \gamma \sum_{s' \in S} p\big(s'|s,a\big)\hat{v}_k^i(s') - \hat{Q}_k^i(s,a)\right), \quad \forall a \in A, \tag{4.5}$$

where we define $\hat{v}_k^i : S \to \mathbb{R}$ as the value function estimate given by

$$\hat{v}_k^i(s) = \max_{a^i} \mathbb{E}_{a_k^{-i} \sim \hat{\pi}_k^{-i}(s)} \{\hat{Q}_k^i(s, a^1, a^2)\}, \tag{4.6}$$

and $\beta_c \in [0, 1]$ is another step size that also vanishes with $c_k(s)$. Similar to (4.4), the update of the belief on the Q-function (4.5) can be viewed as a convex combination of the current belief $\hat{Q}_k^i(s,a)$ and the new observation $r^i(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)\hat{v}_k^i(s')$. Such vanishing step size again implies that the players are relying on their beliefs more if they have had many chances to strengthen them.

The *key feature* of this learning dynamic is that the players update their beliefs on their Q-functions at a slower timescale than the update of their beliefs on the opponent strategy. This is consistent with the literature on evolutionary game theory [22, 62] (which postulates players' choices to be more dynamic than changes in their preferences) since we can view Q-functions in auxiliary games as slowly evolving player preferences. Particularly, the two-timescale learning framework implies that the players take smaller and smaller steps at (4.5) than the steps at (4.4) such that the ratio of the step sizes, β_c/α_c, goes to zero with the number of visits to the associated state. Note that this implies that β_c goes to zero faster

than α_c does, implying slower update of the Q-function estimate compared to the opponent's strategy estimate. This weakens the dependence between evolving beliefs on opponent strategy and Q-function.

We say that this *two-timescale fictitious play dynamics converge to an equilibrium* if beliefs on opponent strategies converge to a Nash equilibrium which associates with the auxiliary games while the beliefs on Q-functions converge to the Q-functions for a stationary equilibrium of the underlying stochastic game. Particularly, given an equilibrium π_*, the associated Q-function of player i satisfies

$$Q^i(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p\left(s'|s,a\right) \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \pi_*^{-i}(s)} \left\{ Q^i(s', a^1, a^2) \right\}, \quad \forall (s,a) \in S \times A.$$

Recall that players are playing a dynamically evolving auxiliary game at each state repeatedly, but update their beliefs on the Q-functions and opponent strategies only when that state is visited. Therefore, the players are updating their beliefs on the opponent strategy and Q-function specific to that state only during these visits. Hence, we make the following assumption ensuring that players have sufficient time to revise and improve their beliefs specific to a state.

Assumption 4.1 (Markov chain). Each state is visited infinitely often.

Stochastic games reduce to the repeated play of the same strategic-form game if there exists only one state and the discount factor is zero. Correspondingly, Assumption 4.1 always holds in such a case. However, when there are multiple states, Assumption 4.1 does not necessarily hold, e.g., since some states can be absorbing by preventing transitions to others. In the following, we exemplify four Markov chain configurations with different generality:

- *Case* (i) The probability of transition between any pair of states is positive for *any* action profile. This condition is also known as *irreducible stochastic games* [40].
- *Case* (ii) The probability of transition between any pair of states is positive for *at least one* action profile. Case (ii) includes Case (i) as a special case.[8]
- *Case* (iii) There is positive probability that any state can be reached from any state within a finite number of stages for any sequence of action profiles taken during these stages. Case (iii) includes Case (i) as a special case but not necessarily Case (ii).
- *Case* (iv) There is positive probability that any state can be reached from any state within a finite number of stages for at least one sequence of action pro-

8 Another possibility in between Cases (i) and (ii) is that the probability of transition between any pair of states is positive for at least one action of one player and any action of the opponent. In other words, the opponent cannot prevent the game to transit from any state to any state.

files taken during these stages. Case (iv) includes Cases (ii) and (iii) as special cases.[9]

Note that Assumption 4.1 holds under Case (iii) but not necessarily under Case (ii) or (iv).

Recall that in the classical fictitious play, the beliefs on opponent strategy are formed by the empirical average of the actions taken by the opponent. The players can also form their beliefs as a weighted average of the actions while the weights may give more (or less) importance to recent ones depending on the player's preferences, e.g., as in (4.4). In other words, we let α_c take values other than $1/(c+1)$ for $c = 0, 1, \ldots$ Furthermore, the two-timescale learning scheme imposes that β_c/α_c goes to zero as c goes to infinity. In the following, we specify conditions on step sizes that are sufficient to ensure convergence of the two-timescale fictitious play in two-player zero-sum stochastic games under Assumption 4.1.

Assumption 4.2 (Step sizes). The step sizes $\{\alpha_c\}$ and $\{\beta_c\}$ satisfy the following conditions:

(a) They vanish at a slow enough rate such that

$$\sum_{c\geq 0} \alpha_c = \sum_{c\geq 0} \beta_c = \infty$$

while $\alpha_c \to 0$ and $\beta_c \to 0$ as $c \to \infty$.

(b) They vanish at two separate timescales such that

$$\lim_{c\to\infty} \frac{\beta_c}{\alpha_c} = 0.$$

The following theorem shows that the two-timescale fictitious play converges in two-player zero-sum stochastic games under these assumptions.

Theorem 4.3 ([63]). *Given a two-player zero-sum stochastic game, suppose that players follow the two-timescale fictitious play dynamics* (4.4) *and* (4.5)*. Under Assumptions* 4.1 *and* 4.2*, we have*

$$(\hat{\pi}_k^1, \hat{\pi}_k^2) \to (\pi_*^1, \pi_*^2) \quad \text{and} \quad (\hat{Q}_k^1, \hat{Q}_k^2) \to (Q_*^1, Q_*^2), \quad \text{with probability } 1, \tag{4.7}$$

as $k \to \infty$ for some stationary equilibrium $\pi_* = (\pi_*^1, \pi_*^2)$ of the underlying stochastic game and (Q_*^1, Q_*^2) denote the associated Q-functions.

Before delving into the technical details of the proof, it is instructive to compare the two-timescale fictitious play with both the classical fictitious play and the Shapley's iteration. For example, the update of $\hat{\pi}_k^{-i}$, described in (4.4), differs from the classical fictitious play dynamics (2.3) since the auxiliary game depends on the belief $\hat{Q}_k^i$ while the belief (and therefore the payoffs of the auxiliary games) evolves in time with new observations, quite

9 Another possibility in between Cases (iii) and (iv) is that there is positive probability that any state can be reached from any state within a finite number of stages for at least one sequence of actions of one player and for any sequence of actions taken by the opponent during these stages. In other words, the opponent cannot prevent the player to reach any state from any state within a finite number of stages.

contrary to the classical scheme (2.4). In general, this constitutes a challenge in directly adopting the convergence analysis for the classical scheme to stochastic games. However, the two-timescale learning scheme weakens this coupling, enabling us to characterize the asymptotic behavior specific to a state separately from the dynamics in other states as if $(\hat{Q}_k^1(s,\cdot), \hat{Q}_k^2(s,\cdot))$ is stationary.

Moreover, even with the two-timescale learning scheme, we still face a challenge in directly adopting the convergence analysis of fictitious play specific to zero-sum games, e.g., [31, 59]. Particularly, players form beliefs on their Q-functions independently based on the backward induction that they will always look for maximizing their utility against the opponent strategy. Due to this independent update, the auxiliary games can deviate from the zero-sum structure even though the underlying game is zero-sum. Hence we do not necessarily have $\hat{Q}_k^1(s,a) + \hat{Q}_k^2(s,a) = 0$ for all $a \in A$ and for each $s \in S$. This poses an important challenge in the analysis since an arbitrary general-sum game does not necessarily have fictitious play property in general.

Next, we compare the two-timescale fictitious play with Shapley's value iteration. We can list the differences between the update of $\hat{Q}_k^i$, described in (4.5), and the Shapley's iteration (3.8) as follows:

- The Shapley's iteration is over the value functions, however, it can be turned into an iteration over the Q-functions with the operator
$$(\mathcal{F}^i Q^i)(s,a) = r^i(s,a) + \gamma \sum_{s' \in S} p\big(s'|s,a\big)\, \mathrm{val}^i\big(Q^i(s',\cdot)\big), \quad \forall (s,a) \in S \times A, \tag{4.8}$$
as derived in [76]. The transformed iteration is given by $Q^i_{(n+1)} = \mathcal{F}^i Q^i_{(n)}$ starting from arbitrary $Q^i_{(0)}$. Furthermore, the Shapley's iteration does not involve a step size, however, a step size can be included if we view $Q^i_{(n+1)} = \mathcal{F}^i Q^i_{(n)}$ as the one
$$Q^i_{(n+1)} = Q^i_{(n)} + \beta_{(n)}(\mathcal{F}^i Q^i_{(n)} - Q^i_{(n)}) \tag{4.9}$$
with the step size $\beta_{(n)} = 1$ for all n.

- The Shapley's iteration updates the value function at every state at each stage while (4.5) takes place only when the state is visited. Therefore, we face the asynchronous update challenge in the convergence analysis of (4.5) together with (4.4), which can take place only when the associated state is visited. To address this, we can resort to the asynchronous stochastic approximation methods, e.g., see [80] (also upcoming Theorem 4.5).

- More importantly, the convergence of the Shapley's iteration benefits from the contraction property of the operator (3.7) (or its transformed version (4.8)) based on the nonexpansive mapping $\mathrm{val}^i(\cdot)$. However, in the update (4.5), we have
$$\hat{v}_k^i(s) = \max_{a^i \in A^i} \mathbb{E}_{a^{-i} \sim \hat{\pi}^{-i}(s)} \big\{ \hat{Q}^i(s, a^1, a^2) \big\}$$
rather than $\mathrm{val}^i(\hat{Q}^i(s,\cdot))$, which need not lead to a contraction.

The proof of Theorem 4.3 follows from exploiting the two-timescale learning scheme to analyze the evolution of the beliefs on opponent strategies specific to a state in isolation as if the beliefs on Q-functions are stationary and then showing that $\hat{v}_k^i(s)$ tracks $\mathrm{val}^i(\hat{Q}^i(s,\cdot))$ while addressing the deviation from the zero-sum structure via a novel Lyapunov function construction. The two-timescale learning scheme yields that the limiting flow of the dynamics specific to a state is given by

$$\frac{d\pi^i(s)}{dt} + \pi_s^i \in \operatorname*{argmax}_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a^1,a^2)\}, \tag{4.10}$$

$$\frac{dQ^i(s,a)}{dt} = 0, \tag{4.11}$$

for all $(s,a) \in S \times A$ and $i = 1,2$. The function (2.6) presented in [31] for continuous-time best response dynamics in zero-sum games is no longer a valid Lyapunov function since $\sum_{i=1,2} Q^i(s,a)$ is not necessarily zero for all s and a. Therefore, we modify this function to characterize the asymptotic behavior of this flow in terms of the deviation from the zero-sum structure, e.g., $\max_{a\in A} |\sum_{i=1,2} Q^i(s,a)|$. The new function is defined by

$$V_*\big(\pi(s), Q(s,\cdot)\big) := \left(\sum_{i=1,2} \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a^1,a^2)\} - \lambda \max_{a\in A}\left|\sum_{i=1,2} Q^i(s,a)\right|\right)_+,$$

where λ is a fixed scalar satisfying $\lambda \in (1, 1/\gamma)$. The lower bound on λ plays a role in its validity as a Lyapunov function when $\max_{a\in A} |\sum_{i=1,2} Q^i(s,a)| \neq 0$ while the upper bound will play a role later when we focus on the evolution of $\sum_{i=1,2} Q^i(s,a)$ to show that the sum converges to zero, i.e., the auxiliary stage games become zero-sum, almost surely.

Note that $V_*(\cdot)$ reduces to $V(\cdot)$, described in (2.6), if $\sum_{i=1,2} Q^i(s,a) = 0$ for all $a \in A$. Furthermore, it is a valid Lyapunov function for any $Q^1(s,\cdot)$ and $Q^2(s,\cdot)$ since we have

$$\begin{aligned}\frac{d}{dt}&\left(\sum_{i=1,2} \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a)\}\right)\\ &= \sum_{i=1,2} Q^i(s,a_*) - \sum_{i=1,2} \max_{a^i\in A^i} \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a)\},\end{aligned} \tag{4.12}$$

where $a_* = (a_*^1, a_*^2)$ are the maximizing actions in (4.10), and we always have

$$\sum_{i=1,2} Q^i(s,a_*) < \lambda \max_{a\in A}\left|\sum_{i=1,2} Q^i(s,a)\right|$$

if it is not zero-sum, since $\lambda > 1$. In other words, the term inside $(\cdot)_*$ in the new Lyapunov function always decreases along the flow when it is nonnegative and cannot be positive once it becomes nonpositive.

If we let $\bar{v}_k := \hat{v}_k^1 + \hat{v}_k^2$ and $\bar{Q}_k := \hat{Q}_k^1 + \hat{Q}_k^2$, the new Lyapunov function yields that

$$\left(\bar{v}_k(s) - \lambda \max_{a\in A} |\bar{Q}_k(s,a)|\right)_+ \to 0 \tag{4.13}$$

as $k \to \infty$ for each $s \in S$. On the other hand, we always have $\bar{v}_k(s) \geq -\lambda \max_{a \in A} |\bar{Q}_k(s,a)|$ by the definition of $\hat{v}_k^i$. These bounds imply that $\bar{Q}_k(s,a) \to 0$, and therefore $\bar{v}_k(s) \to 0$ for all $(s,a) \in S \times A$, because the evolution of $\bar{Q}_k$ for the current state s is given by

$$\bar{Q}_{k+1}(s,a) = \bar{Q}_k(s,a) + \beta_{c_k(s)}\left(\gamma \sum_{s' \in S} p(s'|s,a)\bar{v}_k(s') - \bar{Q}_k(s,a)\right), \quad \forall a \in A \tag{4.14}$$

by (4.5), while the upper bound on λ ensures that $\lambda\gamma \in (0,1)$, and therefore, $\bar{Q}_k(s,a)$ contracts at each stage until it converges to zero for all $s \in S$ and $a \in A$. The asynchronous update and the asymptotic upper bound on $\bar{v}_k$, as described in (4.13), constitute a technical challenge to draw this conclusion, however, they can be addressed via asynchronous stochastic approximation methods, e.g., see [80].

Furthermore, the saddle point equilibrium yields

$$\max_{a^1 \in A^1} \mathbb{E}_{a^2 \sim \hat{\pi}_k^2(s)}\{\hat{Q}_k^1(s,a)\} \geq \mathrm{val}^1\big(\hat{Q}_k^1(s,\cdot)\big) \geq \min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\hat{Q}_k^1(s,a)\}, \tag{4.15}$$

and the right-hand side is bounded from below by

$$\min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\hat{Q}_k^1(s,a)\} \geq \min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{-\hat{Q}_k^2(s,a)\} + \min_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\bar{Q}_k(s,a)\} \tag{4.16}$$

$$\geq -\max_{a^2 \in A^2} \mathbb{E}_{a^1 \sim \hat{\pi}_k^1(s)}\{\hat{Q}_k^2(s,a)\} - \max_{a \in A} |\bar{Q}_k(s,a)|. \tag{4.17}$$

These bounds lead to

$$0 \leq \hat{v}_k^i(s) - \mathrm{val}^i\big(\hat{Q}_k^i(s,\cdot)\big) \leq \bar{v}_k(s) + \max_{a \in A} |\bar{Q}_k(s,a)|, \tag{4.18}$$

Since the right-hand side goes to zero as $k \to \infty$, we have that $\hat{v}_k^i(s)$ tracks $\mathrm{val}^i(\hat{Q}_k^i(s,\cdot))$. Based on this tracking result, the update of $\hat{Q}_k^i$ can be viewed as an asynchronous version of the iteration

$$Q_{(n+1)}^i = Q_{(n)}^i + \beta_{(n)}\big(\mathcal{F}^i Q_{(n)}^i + \epsilon_{(n)}^i - Q_{(n)}^i\big), \tag{4.19}$$

where the tracking error $\epsilon_{(n)}^i$ is asymptotically negligible almost surely and the operator $\mathcal{F}$, as described in (4.8), is a contraction similar to the Shapley's operator, described in (3.7). This completes the sketch of the proof for Theorem 4.3.

4.1. Model-free learning in stochastic games

We next consider scenarios where players do not know the transition probabilities and their own stage payoff function, however, they can still observe their stage payoffs (associated with the current action profile), the opponent's action, and the current state visited. Therefore, the players can still form beliefs on opponent strategy and their Q-functions.

The update of the belief on opponent strategy does not depend on the model knowledge. Therefore, the players can update their beliefs $\hat{\pi}_k^{-i}$ as in (4.4) also in the model-free case. However, the update of $\hat{Q}_k^i$ necessitates the model knowledge by depending on the stage payoff function and transition probabilities. The same challenge arises also in model-free solution of Markov decision processes (MDPs)—a *single* player version of stochastic games.

For example, Q-learning algorithm, introduced by [82], can be viewed as a model-free version of the value iteration in MDPs and the update rule is given by

$$\hat{q}_{k+1}(s,a) = \hat{q}_k(s,a) + \beta_k(s,a)\Big(r_k + \gamma \max_{\tilde{a}\in A} \hat{q}_k(\tilde{s},\tilde{a}) - \hat{q}_k(s,a)\Big), \tag{4.20}$$

where the triple $(s,a,\tilde{s})$ denotes the current state s, current action a, and the next state $\tilde{s}$, respectively, the payoff r_k corresponds to the payoff received, i.e., $r_k = r(s,a)$, and $\beta_k(s,a) \in [0,1]$ is a step size specific to the state-action pair (s,a). The entries corresponding to the pairs $(s',a') \neq (s,a)$ do not get updated, i.e., $\hat{q}_{k+1}(s',a') = \hat{q}_k(s',a')$.

Watkins and Dayan [82] provided an ingenious (direct) proof for the almost sure convergence of Q-learning algorithm. Alternatively, it is also instructive to establish a connection between Q-learning algorithm and the classical value iteration to characterize its convergence properties. For example, the differences between them can be listed as follows:

- In Q-learning, agents use the value function estimate for the next state $\tilde{s}$, i.e., $\hat{v}^i_k(\tilde{s})$, in place of the expected continuation payoff $\sum_{s'\in S} p(s'|s,a)\hat{v}^i_k(s')$. This way, they can sample from the state transition probabilities associated with the current state–action pair by observing the state transitions. Correspondingly, this update takes place only after the environment transitions to the next state.
- The update can take place only for the current state–action pair because the agent can sample only from the transition probabilities associated with the current state-action pair by letting the environment do the experimentation.

Therefore, the Q-learning algorithm can be viewed as an asynchronous Q-function version of the value iteration

$$\hat{q}_{k+1} = \hat{q}_k + \beta_k(\mathcal{F}_o\hat{q}_k + \omega_{k+1} - \hat{q}_k), \tag{4.21}$$

where the Q-function version of the Bellman operator is given by

$$(\mathcal{F}_o\hat{q}_k)(s,a) = r(s,a) + \gamma \sum_{s'\in S} p\big(s'|s,a\big) \max_{a'\in A} \hat{q}_k(s',a') \tag{4.22}$$

and the stochastic approximation error ω_{k+1} is defined by

$$\omega_{k+1}(s,a) := \gamma\left(\max_{\tilde{a}\in A} \hat{q}_k(\tilde{s},\tilde{a}) - \sum_{s'\in S} p\big(s'|s,a\big) \max_{a'\in A} \hat{q}_k(s',a')\right), \tag{4.23}$$

with $\tilde{s}$ denoting the next state at stage k. Note that (4.21) turns into an asynchronous update if $\beta_k(s,a)$ is just zero when $\hat{q}_k(s,a)$ is not updated. Though these error terms $\{\omega_k\}_{k>0}$ do not form an independent sequence, they form a finite-variance martingale difference sequence conditioned on the history of parameters. The following well-known result shows that the weighted sum of such martingale difference sequences vanishes asymptotically almost surely.

Lemma 4.4 ([58]). *Let $\{\mathcal{F}_k\}_{k\geq 0}$ be an increasing sequence of σ-fields. Given a sequence $\{\omega_k\}_{k\geq 0}$, suppose that ω_{k-1} is $\mathcal{F}_k$-measurable random variable satisfying $\mathbb{E}[\omega_k|\mathcal{F}_k] = 0$ and $\mathbb{E}[\omega_k^2|\mathcal{F}_k] \leq K$ for some K. Then, the sequence $\{W_k\}_{k\geq 0}$ evolving according to*

$$W_{k+1} = (1-\alpha_k)W_k + \alpha_k\omega_k, \tag{4.24}$$

vanishes to zero asymptotically almost surely, i.e., $\lim_{k\to\infty} W_k = 0$ *with probability* 1*, provided that* $\alpha_k \in [0, 1]$ *is a vanishing step size that is* $\mathcal{F}_k$*-measurable, square-summable* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ *while* $\sum_{k=0}^{\infty} \alpha_k = \infty$ *with probability* 1.

This is a powerful result to characterize the convergence properties of stochastic approximation algorithms having the structure

$$x_{k+1} = x_k + \alpha_k \big(F(x_k) - x_k + \omega_k\big)$$

where x_k is an n-dimensional vector, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz function, $\alpha_k \in [0, 1]$ is a step size, and ω_k is a stochastic approximation error term forming a finite-variance martingale difference sequence conditioned on the history of parameters. Note that every entry of the vector x_k gets updated synchronously. If we also have that the iterate is bounded, we can characterize the convergence properties of this discrete-time update based on its limiting ordinary differential equation via a Lyapunov function formulation [11]. If the entries do not get updated synchronously, the asynchronous update challenge can be addressed based the *averaging techniques* [38]. In the case of Q-learning, this corresponds to assuming that different state–action pairs occur at well-defined average frequencies, which can be a restriction in practical applications [80]. Instead, [80] showed that we do not need such an assumption if the mapping F has a contraction-like property based on the asynchronous convergence theory [8, 9].

Theorem 4.5 ([80]). *Given an MDP, let an agent follow the* Q*-learning algorithm, described in* (4.20)*, with vanishing step sizes* $\beta_k(s, a) \in [0, 1]$ *satisfying* $\sum_{k\geq 0} \beta_k(s, a) = \infty$ *and* $\sum_{k\geq 0} \beta_k(s, a)^2 < \infty$ *for each* $(s, a) \in S \times A$*. Suppose that the entries corresponding to each* (s, a) *gets updated infinitely often. Then, we have*

$$\hat{q}_k(s, a) \to q_*(s, a), \quad \text{with probability } 1, \tag{4.25}$$

for each $(s, a) \in S \times A$*, as* $k \to \infty$*, where* q_* *is the unique* Q*-function solving the MDP.*

Tsitsiklis [80] considered a more general case where agents receive random payoffs. In general, such randomness can result in unbounded parameters. However, this is not the case for Q-learning algorithm, i.e., the iterates in the Q-learning algorithm remains bounded. Furthermore, the boundedness of the iterates plays a crucial role in the proof of Theorem 4.5. Particularly, consider the deviation between the iterate $\hat{q}_k$ and the unique solution q_*, i.e., $\tilde{q}_k = \hat{q}_k - q_*$, which evolves according to

$$\tilde{q}_{k+1} = \tilde{q}_k + \beta_k(\mathcal{F}_0\tilde{q}_k + \omega_{k+1} - \tilde{q}_k) \tag{4.26}$$

by (4.21) and since $\mathcal{F}_0 q_* = q_*$. Boundedness of the iterates $\hat{q}_k$ yields that $\tilde{q}_k$ is also bounded. For example, let $|\tilde{q}_k(s, a)| \leq D$ for all (s, a) and k. Furthermore, by the contraction property of $\mathcal{F}_0$ with respect to the maximum norm, we have

$$\max_{(s,a)} \big|(\mathcal{F}_0\tilde{q}_k)(s, a)\big| \leq \gamma \max_{(s,a)} \big|\tilde{q}_k(s, a)\big|.$$

Therefore, we can show that the absolute value of new iterates are bounded from above by

$$\big|\tilde{q}_k(s, a)\big| \leq Y_k(s, a) + W_{k+1}(s, a), \tag{4.27}$$

where $\{Y_k(s,a)\}_{k\geq 0}$ and $\{W_{k+1}(s,a)\}_{k\geq 0}$ are two sequences evolving, respectively, according to

$$Y_{k+1}(s,a) = \big(1-\beta_k(s,a)\big)D + \beta_k(s,a)\gamma D \tag{4.28}$$

starting from $Y_0 = D$, and

$$W_{k+1}(s,a) = \big(1-\beta_k(s,a)\big)W_k(s,a) + \beta_k(s,a)\omega_k(s,a), \tag{4.29}$$

starting from $W_1(s,a) = 0$ for all (s,a). For each (s,a), the sequence $\{Y_k(s,a)\}_{k\geq 0}$ converges to γD while $\{W_{k+1}(s,a)\}_{k\geq 0}$ converges to zero with probability 1 by Lemma 4.4 due to the assumptions on the step size and the infinitely often update of every entry. Letting $k\to\infty$ for both sides of (4.27), we obtain that the shifted iterates are asymptotically bounded from above by γD. This yields that there exists a stage where the iterates remain bounded from above by $(\gamma+\epsilon)D$ where $\epsilon > 0$ is sufficiently small such that $\gamma+\epsilon < 1$. By following the same lines, we can find a smaller asymptotic bound on the iterates. Therefore, we can induce that the shifted iterates converge to zero and the iterates converge to the solution of the MDP even with the asynchronous update.

Similar to the generalization of the value iteration to Q-learning for model-free solutions, [42] generalized the Shapley's iteration to *Minimax-Q learning* to compute equilibrium values in two-player zero-sum stochastic games in a model-free way. The update rule is given by

$$\hat{Q}^i_{k+1}(s,a) = \hat{Q}^i_k(s,a) + \beta_k(s,a)\big(r^i_k + \gamma\,\mathrm{val}^i\big[\hat{Q}^i_k(\tilde{s})\big] - \hat{Q}^i_k(s,a)\big), \tag{4.30}$$

for the current state s, current action profile a, and next state $\tilde{s}$ with a step size $\beta_k(s,a)\in[0,1]$ vanishing sufficiently slow such that $\sum_{k\geq 0}\beta_k(s,a) = \infty$ and $\sum_{k\geq 0}\beta_k(s,a)^2 < \infty$ with probability 1. The payoff r^i_k corresponds to the payoff received for the current state and action profile, i.e., $r^i_k = r^i(s,a)$. The Minimax-Q algorithm converges to the equilibrium Q-functions of the underlying two-player zero-sum stochastic game almost surely if every state and action profile occur infinitely often.

In model-free methods, the assumption that every state–action pair occur infinitely often can be restrictive for practical applications. A remedy to this challenge is that agents explore at random instances by taking any action with uniform probability. Such random exploration results in that every state-action pair gets realized infinitely often if every state is visited infinitely often. Indeed, random exploration will also yield that each state gets visited infinitely often if there is always positive probability that any state is reachable from any state within a finite number of stages for at least one sequence of actions taken during these stages. This corresponds to Case (iv) described in Section 4.

In the model-free two-timescale fictitious play, players play the best response in the auxiliary game with probability $(1-\epsilon)$ while experimenting with probability ϵ by playing any action with uniform probability. They still update the belief on the opponent strategy as in (4.4). Furthermore, they update their beliefs on the Q-function for the current state s,

current action profile a and next state s' triple (s, a, s') according to

$$\hat{Q}^1_{k+1}(s,a) = \hat{Q}^1_k(s,a) + \beta_{c_k(s,a)}\Big(r^1_k + \gamma \max_{a^1 \in A} \mathbb{E}_{a^2 \sim \hat{\pi}^2_k(s')}\{\hat{Q}^1_k(s', a^1, a^2)\} - \hat{Q}^1_k(s,a)\Big), \tag{4.31}$$

where $\beta_{c_k(s,a)} \in [0,1]$ is a step size vanishing with the number of times (s,a) is realized and the payoff r^1_k corresponds to the payoff associated with the current state s and action profile a, i.e., $r^1_k = r^1(s,a)$.

Recall that the two-timescale learning scheme plays an important role in the convergence of the dynamics. Particularly, the step size α_c used in the update of the belief $\hat{\pi}^{-i}_k(s)$ goes to zero slower than the step size β_c used in the update of the belief $\hat{Q}^i_k(s,\cdot)$. Since both step size depend on the number of visits to the associated state, the assumption that $\beta_c/\alpha_c \to 0$ as $c \to \infty$ is sufficient to ensure this timescale separation. However, in the model-free case, the asynchronous update of $\hat{Q}^i_k(s,a)$ for different action profiles can undermine this timescale separation because the step size β_c specific to the update of $\hat{Q}^i_k(s,a)$ depends the number of times the state and action profile (s,a), i.e., $c_k(s,a)$, is realized. Therefore, we make the following assumption ensuring that the step size in the update of $\hat{Q}^i_k(s,a)$ vanishes still faster than the step size in the update of $\hat{\pi}^{-i}_k(s)$ as long as $c_k(s,a)$ is comparable with $c_k(s)$, i.e., $\liminf_{k\to\infty} c_k(s,a)/c_k(s) > 0$ with probability 1.

Assumption 4.6 (Step sizes). The step sizes $\{\alpha_c\}$ and $\{\beta_c\}$ satisfy the following conditions:

(a) They vanish at a slow enough rate such that

$$\sum_{c\geq 0}\alpha_c = \sum_{c\geq 0}\beta_c = \infty, \quad \text{and} \quad \sum_{c\geq 0}\alpha_c^2 < \infty, \quad \sum_{c\geq 0}\beta_c^2 < \infty$$

while $\alpha_c \to 0$ and $\beta_c \to 0$ as $c \to \infty$.[10]

(b) The sequence $\{\beta_c\}_{c\geq 0}$ is monotonically decreasing. For any $m \in (0,1]$, we have[11]

$$\lim_{c\to\infty} \frac{\beta_{\lfloor mc \rfloor}}{\alpha_c} = 0.$$

When we have $\liminf_{k\to\infty} c_k(s,a)/c_k(s) > 0$ with probability 1 for all (s,a), the second part of Assumption 4.6 ensures that $\lim_{k\to\infty} \frac{\beta_{c_k(s,a)}}{\alpha_{c_k(s)}} = 0$ with probability 1 for all (s,a). Indeed, Assumptions 4.2 and 4.6 are satisfied for the usual (vanishing) step sizes such as

$$\alpha_c = \frac{1}{(c+1)^{\rho_\alpha}} \quad \text{and} \quad \beta_c = \frac{1}{(c+1)^{\rho_\beta}},$$

where $1/2 < \rho_\alpha < \rho_\beta \leq 1$.

---

10 We have the additional assumption that the step size β_c is square summable to ensure that the stochastic approximation error terms have finite variance conditioned on the history of the parameters.

11 Perkins and Leslie **[54]** made a similar assumption that $\sup_c \frac{\beta_{\lfloor mc \rfloor}}{\beta_c} < M$ for all $m \in (0,1)$ and $\frac{\beta_c}{\alpha_c} \to 0$ for two-timescale asynchronous stochastic approximation.

When players do random experimentation in the model-free case, they do not take the best response with certain probability. Therefore, we do not have convergence to an exact equilibrium as in the model-based case. However, the players still converge to a near equilibrium of the game with linear dependence on the experimentation probability and the following theorem provides an upper bound on this approximation error.

Theorem 4.7 ([63]). *Given a two-player zero-sum stochastic game, suppose that players follow the model-free two-timescale fictitious play dynamics with experimentation probability $\epsilon > 0$. Under Assumptions 4.1 and 4.6, we have*

$$\limsup_{k\to\infty} \left| v_k^i(s) - v^i(s) \right| \leq \epsilon D \frac{1+\gamma}{\gamma(1-\gamma)^2}, \tag{4.32}$$

$$\limsup_{k\to\infty} \max_{a\in A} \left| \hat{Q}_k^i(s,a) - Q^i(s,a) \right| \leq \epsilon D \frac{1+\gamma}{(1-\gamma)^2}, \tag{4.33}$$

with probability 1, *where* $D = \frac{1}{1-\gamma} \sum_i \max_{(s,a)} |r^i(s,a)|$, *where* v_*^i *and* Q_*^i *denote, respectively, the value function and* Q*-function of player* i *for some stationary equilibrium of the stochastic game.*

Even though the random experimentation can prevent convergence to an exact equilibrium, it provides an advantage for the applicability of this near-convergence result because every state gets visited infinitely often, and therefore, Assumption 4.1 holds, if the underlying Markov chain satisfies Case (iv), i.e., there is positive probability that any state can be reached from any state within a finite number of stages for at least one sequence of action profiles taken during these stages.

The dynamics can converge to an exact equilibrium also in the model-free case if players let the experimentation probability vanish at certain rate. However, there are technical details that can limit the applicability of the result for Case (iv).

4.2. Radically uncoupled learning in stochastic games

Finally, we consider minimal-information scenarios where players do not even observe the opponent's actions in the model-free case. Each player can still observe its own stage payoff received and the current state visited. The players also do not know the opponent's action set. Indeed, they may even be oblivious to the presence of an opponent. The learning dynamics under such minimal information case is known as *radically uncoupled learning* in the learning in games literature, e.g., see [25].

Without observing the opponent's actions and knowing her action space, players are not able to form beliefs on opponent strategy as in the fictitious play. This challenge is present also in the repeated play of the same strategic-form game. For example, consider the strategic-form game $\langle A^1, A^2, r^1, r^2 \rangle$ and define $q^i : A^i \to \mathbb{R}$ by

$$q^i(a^i) := \mathbb{E}_{a^{-i}\sim\pi^{-i}} \left\{ r^i(a^1, a^2) \right\}, \quad \forall a^i \in A^i \tag{4.34}$$

given the opponent's strategy π^{-i}. Then, the computation of the best response is a simple optimization problem for player i, given by

$$a^i_* \in \operatorname*{argmax}_{a^i \in A^i} q^i(a^i).$$

Player i would be able to compute her best response a^i_* even when she does not know the opponent strategy π^{-i} and her payoff function r^i if she knew the function $q^i(\cdot)$. Hence, the question is whether the computation of $q^i(\cdot)$ can be achieved without observing the opponent's action.

Suppose that players are playing the same strategic-form game repeatedly and player i makes the forward induction that the opponent will play as how he has played in the past similar to the fictitious play dynamics. If that were the case, i.e., the opponent were playing according to a stationary strategy π^{-i}, then at each stage the payoff received by player i would be the realized payoff $r^i(a^1, a^2)$, where $a^{-i} \sim \pi^{-i}$ and a^i is the current action she has taken. Correspondingly, player i can form a belief about $q^i(a^i)$ for all $a^i \in A^i$ and update $q^i(\cdot)$ associated with the current action based on the payoff she received. For example, let $\hat{q}^i_k$, a^i_k and r^i_k denote, respectively, the belief of player i on q^i, her current action and the current payoff she received. Similar to the update of the belief on opponent's strategy, the update of $\hat{q}^i_k$ is given by

$$\hat{q}^i_{k+1}(a^i) = \begin{cases} \hat{q}^i_k(a^i) + \alpha_k(a^i)(r^i_k - \hat{q}^i_k(a^i)) & \text{if } a^i = a^i_k, \\ \hat{q}^i_k(a^i) & \text{otherwise,} \end{cases}$$

where $\alpha_k(a^i) \in [0, 1]$ is a vanishing step size specific to the action a^i. However, this results in an asynchronous update of $\hat{q}_k$ for different actions quite contrary to the synchronous belief update (2.3) in the fictitious play. There is no guarantee that it would converge to an equilibrium even in the zero-sum case. On the other hand, such an asynchrony issue is not present and the update turns out to be synchronous in expectation if players take *smoothed* best response while normalizing the step size by the probability of the current action taken [39].

Given $\hat{q}^i_k$, the smoothed best response $\overline{\mathrm{BR}}^i_k \in \Delta(A^i)$ is given by

$$\overline{\mathrm{BR}}^i_k := \operatorname*{argmax}_{\mu^i \in \Delta(A^i)} \left(\mathbb{E}_{a^i \sim \mu^i} \{\hat{q}^i_k(a^i)\} + \tau v^i(\mu^i) \right), \tag{4.35}$$

where $v^i : \Delta(A^i) \to \mathbb{R}$ is a smooth and strictly concave function whose gradient is unbounded at the boundary of the simplex $\Delta(A^i)$ [29]. The temperature parameter $\tau > 0$ controls the amount of perturbation on the smoothed best response. Note that the smooth perturbation ensures that there always exists a unique maximizer in (4.35). Since players take smoothed best response rather than best response, we use an equilibrium concept different from the Nash equilibrium. This new definition is known as Nash distribution or quantal response equilibrium [46].

Definition 4.8 (Nash distribution). We say that a strategy profile π_* is a Nash distribution if we have

$$\pi^i_* = \operatorname*{argmax}_{\mu^i \in \Delta(A^i)} \left(\mathbb{E}_{(a^i, a^{-i}) \sim (\mu^i, \pi^{-i}_*)} \{r^i_k(a)\} + \tau v^i(\mu^i) \right) \tag{4.36}$$

for each i.

An example to the smooth function is $v^i(\mu^i) := -\mathbb{E}_{a^i\sim\mu^i}\{\log(\mu^i(a^i))\}$, also known as the entropy [34], and the associated smoothed best response has the following analytical form:

$$\overline{\mathrm{BR}}_k^i(a^i) = \frac{\exp(\hat{q}_k^i(a^i)/\tau)}{\sum_{\tilde{a}^i\in A^i}\exp(\hat{q}_k^i(\tilde{a}^i)/\tau)},$$

which is positive for all $a^i \in A^i$.

When player i takes her action according to the smoothed best response $\overline{\mathrm{BR}}_k^i$, any action will be taken with some positive probability $\overline{\mathrm{BR}}_k^i(a^i) > 0$. Hence she can update her belief according to

$$\hat{q}_{k+1}^i(a^i) = \begin{cases} \hat{q}_k^i(a^i) + \overline{\mathrm{BR}}_k^i(a^i)^{-1}\alpha_k(r_k^i - \hat{q}_k^i(a^i)) & \text{if } a^i = a_k^i, \\ \hat{q}_k^i(a^i) & \text{otherwise,} \end{cases} \tag{4.37}$$

where $\alpha_k \in (0,1)$ is a step size vanishing with k and not specific to any action. This asynchronous update rule, also known as *individual Q-learning*, turns out to be synchronous in the expectation. Particularly, the new update rule is given by

$$\hat{q}_{k+1}^i(a^i) = \hat{q}_k^i(a^i) + \alpha_k\big(\mathbb{E}_{a^{-i}\sim\overline{\mathrm{BR}}_k^{-i}}\{r^i(a^1,a^2)\} - \hat{q}_k^i(a^i) + \omega_k^i(a^i)\big), \quad \forall a^i \in A^i, \tag{4.38}$$

and $\omega_k^i(a^i)$ is the stochastic approximation error defined by

$$\omega_k^i(a^i) := \mathbf{1}_{\{a^i=a_k^i\}}\overline{\mathrm{BR}}_k^i(a^i)^{-1}\big(r_k^i - \hat{q}_k^i(a^i)\big) - \mathbb{E}_{a\sim\overline{\mathrm{BR}}_k}\big\{\mathbf{1}_{\{a^i=a_k^i\}}\overline{\mathrm{BR}}_k^i(a^i)^{-1}\big(r_k^i - \hat{q}_k^i(a^i)\big)\big\},$$

where $\overline{\mathrm{BR}}_k = (\overline{\mathrm{BR}}_k^1, \overline{\mathrm{BR}}_k^2)$, because we have

$$\mathbb{E}_{a\sim\overline{\mathrm{BR}}_k}\big\{\mathbf{1}_{\{a^i=a_k^i\}}\overline{\mathrm{BR}}_k^i(a^i)^{-1}\big(r_k^i - \hat{q}_k^i(a^i)\big)\big\} = \mathbb{E}_{a^{-i}\sim\overline{\mathrm{BR}}_k^{-i}}\{r^i(a^1,a^2)\} - \hat{q}_k^i(a^i).$$

Furthermore, the stochastic approximation error term forms a martingale difference sequence conditioned on the history of iterates while the *boundedness* of the iterates ensure that it has finite variance. Therefore, we can invoke Lemma 4.4 to characterize the convergence properties of (4.38)—a rewritten version of (4.37) with the stochastic approximation term ω_k^i.

Theorem 4.9 ([39]). *In two-player zero-sum (or identical-payoff) strategic-form games played repeatedly, if both player follows the individual Q-learning algorithm, described in (4.37), then their estimate $\hat{q}_k^i$ converges to q_*^i for all $a^i \in A^i$ satisfying*

$$q_*^i(a^i) = \mathbb{E}_{a^{-i}\sim\pi_*^{-i}}\{r^i(a^1,a^2)\}$$

for some Nash distribution $\pi_* = (\pi_*^1, \pi_*^2)$ under the assumption that the iterates remain bounded. Correspondingly, their smoothed best response also converges to π^*.

Recall that in stochastic games, players are playing an *auxiliary stage-game* specific to the current state $\mathcal{G}_s = \langle A^1, A^2, Q^1(s,\cdot), Q^2(s,\cdot)\rangle$, where Q^i satisfies (4.1). Therefore, in the minimal information case, each player i can form a belief about the associated

$$q^i(s,a^i) := \mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}\{Q^i(s,a^1,a^2)\},$$

which is now specific to state s contrary to (4.34), and update it based on the stage payoffs received as in the individual Q-learning dynamics. We can view q^i as the local Q-function

since it is defined over individual actions rather than action profiles. We denote player i's belief on q^i by $\hat{q}^i_k$. Let s be the current state of the stochastic game. Then, player i selects her action a^i_k according to smoothed best response

$$\overline{\mathrm{BR}}^i_k(s,\cdot) = \operatorname*{argmax}_{\mu^i \in \Delta(A^i)} \big(\mathbb{E}_{a^i\sim\mu^i}\{\hat{q}^i_k(s,a^i)\} + \tau v^i(\mu^i)\big),$$

i.e., $a^i_k \sim \overline{\mathrm{BR}}^i_k(s,\cdot)$. The smoothed best response depends only on the belief on the local Q-function, i.e., $\hat{q}^i_k(s,\cdot)$. Observing the stage reward r^i_k and the next state s', player i can update her belief according to

$$\hat{q}^i_{k+1}(s,a^i) = \begin{cases} \hat{q}^i_k(s,a^i) + \overline{\mathrm{BR}}^i_k(s,a^i)^{-1}\alpha_{c_k(s)}(r^i_k + \gamma \hat{v}^i_k(s') - \hat{q}^i_k(s,a^i)) & \text{if } a^i = a^i_k, \\ \hat{q}^i_k(s,a^i) & \text{otherwise,} \end{cases} \tag{4.39}$$

where $\alpha_c \in (0,1)$ is a vanishing step size and recall that $c_k(s)$ denotes the number of visits to state s until and including stage k. The update (4.39) differs from (4.37) due to the additional term $\gamma\hat{v}^i_k(s')$ corresponding to an unbiased estimate of the continuation payoff in the model-free case. Due to this additional term, the individual Q-learning dynamics in auxiliary stage-games specific to each state are coupled with each other. A two-timescale learning framework can weaken this coupling if players estimate $\hat{v}^i_k$ at a slower timescale according to

$$\hat{v}^i_{k+1}(s) = \hat{v}^i_k(s) + \beta_{c_k(s)}\big(\mathbb{E}_{a^i\sim\overline{\mathrm{BR}}^i_k(s,\cdot)}\{\hat{q}^i_k(s,a^i)\} - \hat{v}^i_k(s)\big), \tag{4.40}$$

where $\beta_c \in (0,1)$ is a vanishing step size that goes to zero faster than α_c, rather than $\hat{v}^i_k(s) = \mathbb{E}_{a^i\sim\overline{\mathrm{BR}}^i_k(s,\cdot)}\{\hat{q}^i_k(s,a^i)\}$.

This decentralized Q-learning dynamics, described in (4.39) and (4.40), have convergence properties similar to the two-timescale fictitious play even in this minimal information case. Furthermore, random exploration is inherent in the smoothed best response. Therefore, Assumption 4.1 holds if the underlying Markov chain satisfies Case (iv). However, due to the smoothed best response, the dynamics does not necessarily converge to an exact Nash equilibrium.

Theorem 4.10 ([64]). *Given a two-player zero-sum stochastic game, suppose that players follow the decentralized Q-learning dynamics. In addition to Assumptions 4.1 and 4.6, we assume that $\sum_{c\geq 0}\alpha_c^2 < \infty$ and the iterates are bounded. Let Q^i_* and v^i_* denote the unique equilibrium Q-function and value function of player i. Then, we have*

$$\limsup_{k\to\infty} \big|\hat{v}^i_k(s) - v^i_*(s)\big| \leq \tau \log\big(\big|A^1\big|\big|A^2\big|\big) g(\gamma), \tag{4.41}$$

for all $(i,s) \in \{1,2\}\times S$, with probability 1, where $g(\lambda) := \frac{2+\lambda-\lambda\gamma}{(1-\lambda\gamma)(1-\gamma)}$ with some $\lambda \in (1, 1/\gamma)$.

Furthermore, let $\hat{\pi}^i_k(s) \in \Delta(A^i)$ be the weighted time-average of the smoothed best response updated as

$$\hat{\pi}^i_{k+1}(s) = \hat{\pi}^i_k(s) + \mathbf{1}_{\{s=s_k\}}\alpha_{c_k(s)}\big(\overline{\mathrm{BR}}^i_k(s,\cdot) - \hat{\pi}^i_k(s)\big).$$

Then, we have

$$\limsup_{k\to\infty} \max_{a^i \in A^i} \mathbb{E}_{a^{-i}\sim \hat{\pi}_k^{-i}(s)} \{Q_*^i(s,a)\} - v_{\pi_*}^i(s)| \le \tau \log(|A^1||A^2|)h(\gamma), \tag{4.42}$$

for all $(i,s) \in \{1,2\} \times S$, *with probability* 1, *where* $h(\gamma) := g(\gamma)(1+\gamma) - 1$. *In other words, these weighted-average strategies converge to near Nash equilibrium strategies of the stochastic game.*

The iterates would be bounded inherently if players update the local Q-function (4.37) by thresholding the step $\overline{\mathrm{BR}}_k^i(a^i)^{-1}\alpha_{c_k(s)}$ from above by 1. Furthermore, the dynamics could converge to an exact equilibrium if players let their temperature parameter $\tau > 0$ vanishes over time at a certain rate, e.g., see [64]. With vanishing temperature, Assumption 4.1 holds if the underlying Markov chain satisfies Case (iii).

5. OTHER LEARNING ALGORITHMS

Previous sections have focused on a detailed description of best-response/fictitious-play type learning dynamics, together with Q-learning dynamics, for stochastic games. In this section, we summarize several other algorithms in the learning in games literature, with a focus on independent/decentralized learning for stochastic games (also belonging to the area of *multiagent reinforcement learning* in the machine learning literature).

5.1. Classical algorithms

For stochastic games, other than Q-learning-type algorithms presented in Section 4.1, [10] also established the asymptotic convergence of an actor–critic algorithm to a weaker notion of generalized Nash equilibrium. Another early work [13] proposed R-MAX, an optimism-based RL algorithm for average-reward two-player zero-sum stochastic games, with polynomial time convergence guarantees. However, convergence to the actual Nash equilibrium is not guaranteed from the regret definition in the paper.

For strategic-form games, besides fictitious play, several other *decentralized* learning dynamics have also been thoroughly studied. A particular example is the *no-regret* learning algorithms[12] from the online learning literature. It is a folklore theorem that: If both players of a game use some no-regret learning dynamics to adapt their strategies to their opponent's strategies, then the time-average strategies of the players constitute a Nash equilibrium of the zero-sum strategic-form game [18, 61]. Popular no-regret dynamics include multiplicative weights update [26, 41], online gradient descent [91], and their generalizations [47, 67]. These no-regret learning dynamics are *uncoupled* in that a player's dynamics does not explicitly rely on the payoffs of other players [32]. They are also posited to be a rational model of players' rational behavior [60, 75]. In addition, [39] proposed individual Q-learning, a fully decentralized learning dynamics where each player's update rule requires no observation of the opponent's actions, with convergence to the Nash equilibrium *distribution* of

12 See [18] for formal definitions and results of no-regret learning.

certain two-player games. Notably, these decentralized learning dynamics are only known to be effective for strategic-form games.

5.2. Multiagent reinforcement learning

There has been a flurry of recent works on multiagent RL in stochastic games with focuses on *nonasymptotic* performance guarantees. The authors of [56,57] proposed batch RL algorithms to find an approximate Nash equilibrium using approximate dynamic programming analysis. Wei et al. [83] studied *online* RL, where only one of the player is controlled, and develops the UCSG algorithm with sublinear regret guarantees that improves the results in [13], though still without guarantees of finding the Nash equilibrium. Subsequently, [72] provided near-optimal sample complexity for solving *turn-based* two-player zero-sum finite stochastic games, when a generative model that enables sampling from any state–action pair is available. Under the same setting, the near-optimal sample complexity for general two-player zero-sum finite stochastic games was then established in [87]. Without a generative model, [2, 85] presented optimistic value iteration-based RL algorithms for two-player zero-sum stochastic games, with efficient exploration of the environment, and finite-time regret guarantees. The two players need some coordination to perform the algorithms, and the focus in these two works is the *finite-horizon episodic* setting. Later, [3] and [43] provided tighter regret bounds for the same setting, with model-free and model-based RL methods, respectively. Liu et al. [43] has also studied the general-sum setting, with finite-sample guarantees for finding the Nash equilibrium, assuming some computation oracle for finding the equilibrium of general-sum strategic-form games at each iteration. Contemporaneously, [35, 37] studied multiagent RL with *function approximation* in finite-horizon episodic zero-sum stochastic games, with also the optimism principle and regret guarantees.

In addition, *policy-based* RL algorithms have also been developed for solving stochastic games. The authors of [15, 88] developed double-loop policy gradient methods for solving zero-sum linear quadratic dynamic games, a special case of zero-sum stochastic games with linear transition dynamics and quadratic cost functions, with convergence guarantees to the Nash equilibrium. Later, [90] also studied double-loop policy gradient methods for zero-sum stochastic games with general function approximation. Note that these double-loop algorithms are not symmetric in that they require one of the players to wait for the opponent to update her policy parameter multiple steps while updating her own policy for one step, which necessarily requires some coordination between players. Finally, [66] developed an Explore–Improve–Supervise approach, which combines ideas from Monte Carlo Tree Search and Nearest Neighbors methods, to find the approximate Nash equilibrium value of *continuous-space* turn-based zero-sum stochastic games. The two players are coordinated to learn the minimax value jointly.

Notably, as minimax Q-learning, these multiagent RL algorithms are mostly focused on the *computational* aspect of learning in stochastic games: compute the Nash equilibrium without knowing the model, using possibly as few samples as possible. Certain level of coordination among the players is either explicitly or implicitly assumed when implementing these algorithms, even for the zero-sum setting where the players compete against each

other. For human-like self-interested players, these update rules may not be sufficiently rational and natural to execute. Indeed, as per [12], a preferable multiagent RL algorithm should be both *rational* and *convergent*: a rational algorithm ensures that the iterates converge to the opponent's best-response if the opponent converges to a stationary policy; while a convergent algorithm ensures convergence to some equilibrium if all the agents apply the learning dynamics. In general, a rational algorithm, in which each player *adapts* to the (possibly nonstationary) behavior of other players and uses only *local* information she observes without the aid of any central coordinator, does not lead to the equilibrium of the game. In fact, investigating whether a game-theoretical equilibrium can be realized as a result of nonequilibrium adaptation dynamics is the core topic in the literature of *learning in games* [29]. These multiagent RL works have thus motivated our study of independent learning dynamics presented in Section 4.

5.3. Decentralized learning in stochastic games

Decentralized learning in stochastic games has attracted increasing research interest lately. In [1], decentralized Q-learning has been proposed for *weakly acyclic* stochastic games, which include stochastic teams (identical-interest stochastic games) as a special case. The update rule for each player does not need to observe the opponent players' actions, and is even oblivious to the presence of other players. However, the players are implicitly coordinated to explore every multiple iterations (in the exploration phase) without changing their policies, in order to create a stationary environment for each player. The key feature of the update rule is to restrict player strategies to stationary pure strategies. Since there are only finitely many stationary pure strategy, players can create a huge-game matrix for each stationary pure strategy and a pure-strategy equilibrium always exists when this huge-game is weakly acyclic with respect to best response. However, in the model-free case, players do not know the payoffs of this huge-game and the two-phase update rule addresses this challenge. Perolat et al. [55] developed actor–critic-type learning dynamics that are decentralized and of fictitious-play type, where the value functions are estimated at a faster timescale (in the critic step), and the policy is improved at a slower one (in the actor step). Nonetheless, the learning dynamics only applies to a special class of stochastic games with a "multistage" structure, in which each state can only be visited once. In [21], an independent policy gradient method was investigated for zero-sum stochastic games with convergence rate analysis, where two players use *asymmetric* stepsizes in their updates with one updates faster than the other. This implicitly requires some coordination between players to determine who shall update faster. Contemporaneously, [79] studied *online* RL in unknown stochastic games, where only one player is controlled and the update rule is fully decentralized. The work focused on the efficient *exploration* aspect of multiagent RL, by establishing the regret[13] guarantees of the proposed update rule. The work considered only the finite-horizon episodic setting, and it

13 The regret defined in [79] is weaker than the normal one with the *best-in-hindsight* comparator. See [79, SECT. 2] for a detailed comparison.

is also unclear if the learning dynamics converge to any equilibrium when all players apply it.[14]

With symmetric and decentralized learning dynamics, [17, 40, 84] are, to the best our knowledge, the latest efforts on learning in stochastic games. Leslie et al. [40] studied *continuous-time* best-response dynamics for zero-sum stochastic games, with a *two-timescale* update rule: at the slower timescale, a single continuation payoff (common among the players) is updated, representing time average of auxiliary game payoffs up to time k; at the faster timescale, each player updates its strategy in the direction of its best response to opponent's current strategy in the auxiliary game. The common continuation payoff update ensures that the auxiliary game is always zero-sum, allowing the use of the techniques for the strategic-form game setting [31]. The dynamics update the mixed strategies at every state at every time. Alternatively, the work also considered a continuous-time embedding of the *actual play* of the stochastic game where game transitions according to a controlled continuous-time Markov chain. Both [84] and [17] studied the genuine infinite-horizon discounted zero-sum stochastic games, and provided *last-iterate* convergence rate guarantees to approximate Nash equilibrium. To this end, [84] developed an optimistic variant of gradient descent-ascent update rule; while [17] focused on the *entropy-regularized* stochastic games, and advocated the use of policy extragradient methods. Though theoretically strong and appealing, these update rules assume either exact access or sufficiently accurate estimates of the continuation payoffs under instantaneous joint strategies and/or the instantaneous strategy of the opponent. In particular, to obtain finite-time bounds, the players are coordinated to interact multiple steps to estimate the continuation payoffs in the learning setting [84].

By and large, ever since the introduction of fictitious play [14] and stochastic games [69], it remains a long-standing problem whether an equilibrium in a stochastic game can be realized as an outcome of some natural and decentralized nonequilibrium adaptation, e.g., fictitious play (except the contemporaneous work [40] with some continuous-time embeddings). Hence, our solutions in Section 4 serve as an initial attempt towards settling the argument positively.

6. CONCLUSIONS AND OPEN PROBLEMS

In this review paper, we introduced multiagent dynamic learning in stochastic games, an increasingly active research area where artificial intelligence, specifically reinforcement learning, meets game theory. We have presented the fundamentals and background of the problem, followed by our recent advances in this direction, with a focus on studying *independent learning* dynamics. We believe our work has opened up fruitful directions for future research, on developing more natural and rational multiagent learning dynamics for

14 The same update rule with different stepsize and bonus choices and a certified policy technique, however, can return a non-Markovian approximate Nash equilibrium policy pair in the zero-sum setting; see [3], and the very recent and more complete treatment [36], for more details.

stochastic games. In particular, several future/ongoing research directions include: (1) establishing convergence guarantees of our independent learning dynamics for other stochastic games, e.g., identical-interest ones; (2) establishing nonasymptotic convergence guarantees of our learning dynamics, or other independent learning dynamics, for stochastic games; (3) developing natural learning dynamics that also account for the large state–action spaces in practical stochastic games, e.g., via function approximation techniques.

FUNDING

A. Ozdaglar and K. Zhang were supported by DSTA grant 031017-00016 and ARO Project W911NF1810407.

REFERENCES

[1] G. Arslan and S. Yüksel, Decentralized Q-learning for stochastic teams and games. *IEEE Trans. Automat. Control* **62** (2017), no. 4, 1545–1558.

[2] Y. Bai and C. Jin, Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pp. 551–560, PMLR, 2020.

[3] Y. Bai, C. Jin, and T. Yu, Near-optimal reinforcement learning with self-play. In *Advances in neural information processing systems 33*, pp. 2159–2170, Curran Associates, Inc., 2020.

[4] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory. 2nd edn.* Classics Appl. Math., SIAM, 1999.

[5] M. Benaim, J. Hofbauer, and S. Sorin, Stochastic approximations and differential inclusions. *SIAM J. Control Optim.* **44** (2005), no. 1, 328–348.

[6] U. Berger, Fictitious play in $2 \times n$ games. *J. Econom. Theory* **120** (2005), no. 2, 139–154.

[7] U. Berger, Learning in games with strategic complementarities revisited. *J. Econom. Theory* **143** (2008), no. 1, 292–301.

[8] D. P. Bertsekas, Distributed dynamic programming. *IEEE Trans. Automat. Control* **AC-27** (1982), 610–616.

[9] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice Hall, New Jersey, 1989.

[10] V. S. Borkar, Reinforcement learning in Markovian evolutionary games. *Adv. Complex Syst.* **5** (2002), no. 01, 55–72.

[11] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Hindustan Book Agency, 2008.

[12] M. Bowling and M. Veloso, Rational and convergent learning in stochastic games. In *Proceedings 17th international joint conference on artificial intelligence*, pp. 1021–1026, Morgan Kaufmann Publishers Inc., 2001.

[13] R. I. Brafman and M. Tennenholtz, R-MAX—A general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.* **3** (2002), 213–231.

[14] G. W. Brown, Iterative solution of games by fictitious play. In *Activity analysis of production and allocation*, pp.374–376, Cowles Commission Monograph 13, Wiley, New York, 1951.

[15] J. Bu, L. J. Ratliff, and M. Mesbahi, Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. 2019, arXiv:1911.04672.

[16] L. Busoniu, R. Babuska, and B. D. Schutter, A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **38** (2008), no. 2, 156–172.

[17] S. Cen, Y. Wei, and Y. Chi, Fast policy extragradient methods for competitive games with entropy regularization. 2021, arXiv:2105.15186.

[18] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[19] C. Claus and C. Boutilier, The dynamics of reinforcement learning in cooperative multiagent systems. In *Conference on artificial intelligence*, pp. 746–752, American Association for Artificial Intelligence, 1998.

[20] A. Condon, On algorithms for simple stochastic games. In *Advances in computational complexity theory 13*, pp. 51–72, American Mathematical Society, 1990.

[21] C. Daskalakis, D. J. Foster, and N. Golowich, Independent policy gradient methods for competitive reinforcement learning. In *Advances in neural information processing systems*, Curran Associates, Inc., 2020.

[22] J. C. Ely and O. Yilankaya, Nash equilibrium and the evolution of preferences. *J. Econom. Theory* **97** (2001), no. 2, 255–272.

[23] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer, 2012.

[24] A. M. Fink, Equilibrium in stochastic n-person game. *J. Sci. Hiroshima Univ., Ser. A-I* **28** (1964), 89–93.

[25] D. P. Foster and H. P. Young, Regret testing: learning to play Nash equilibrium without knowing you have an opponent. *Theor. Econ.* **1** (2006), 341–367.

[26] Y. Freund and R. E. Schapire, Adaptive game playing using multiplicative weights. *Games Econom. Behav.* **29** (1999), no. 1–2, 79–103.

[27] D. Fudenberg and D. M. Kreps, Learning mixed equilibria. *Games Econom. Behav.* **5** (1993), no. 3, 320–367.

[28] D. Fudenberg and D. K. Levine, Consistency and cautious fictitious play. *J. Econom. Dynam. Control* **19** (1995), no. 5–7, 1065–1089.

[29] D. Fudenberg and D. K. Levine, *The theory of learning in games. 2*. MIT Press, 1998.

[30] S. Gu, E. Holly, T. Lillicrap, and S. Levine, Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE international conference on robotics and automation*, pp. 3389–3396, IEEE, 2017.

[31] C. Harris, On the rate of convergence of continuous-time fictitious play. *Games Econom. Behav.* **22** (1998), 238–259.

[32] S. Hart and A. Mas-Colell, Uncoupled dynamics do not lead to Nash equilibrium. *Am. Econ. Rev.* **93** (2003), no. 5, 1830–1836.

[33] J. Hofbauer and W. H. Sandholm, On the global convergence of stochastic fictitious play. *Econometrica* **70** (2002), no. 6, 2265–2294.

[34] J. Hofbauer and W. H. Sandholm, On the global convergence of stochastic fictitious play. *Econometrica* **70** (2002), 2265–2294.

[35] B. Huang, J. D. Lee, Z. Wang, and Z. Yang, Towards general function approximation in zero-sum Markov games. 2021, arXiv:2107.14702.

[36] C. Jin, Q. Liu, Y. Wang, and T. Yu, V-learning—a simple, efficient, decentralized algorithm for multiagent RL. 2021, arXiv:2110.14555.

[37] C. Jin, Q. Liu, and T. Yu, The power of exploiter: provable multi-agent RL in large state spaces. 2021, arXiv:2106.03352.

[38] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*. Springer, 1978.

[39] D. S. Leslie and E. J. Collins, Individual Q-learning in normal form games. *SIAM J. Control Optim.* **44** (2005), no. 2, 495–514.

[40] D. S. Leslie, S. Perkins, and Z. Xu, Best-response dynamics in zero-sum stochastic games. *J. Econom. Theory* **189** (2020).

[41] N. Littlestone and M. K. Warmuth, The weighted majority algorithm. *Inform. and Comput.* **108** (1994), no. 2, 212–261.

[42] M. L. Littman, Markov games as a framework for multi-agent reinforcement learning. In *International conference on machine learning*, pp. 157–163, Morgan Kaufmann Publishers Inc., 1994.

[43] Q. Liu, T. Yu, Y. Bai, and C. Jin, A sharp analysis of model-based reinforcement learning with self-play. In *International conference on machine learning*, pp. 7001–7010, PMLR, 2021.

[44] E. Maskin and J. Tirole, A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs. *Econometrica* (1988), 549–569.

[45] E. Maskin and J. Tirole, A theory of dynamic oligopoly, II: Price competition, kinked demand curves, and Edgeworth cycles. *Econometrica* (1988), 571–599.

[46] R. McKelvey and T. Palfrey, Quantal response equilibria for normal form games. *Games Econom. Behav.* **10** (1995), 6–38.

[47] B. McMahan, Follow-the-regularized-leader and mirror descent: equivalence theorems and l_1 regularization. In *International conference on artificial intelligence and statistics*, pp. 525–533, PMLR, 2011.

[48] P. Milgrom and J. Roberts, Adaptive and sophisticated learning in normal form games. *Games Econom. Behav.* **3** (1991), 82–100.

[49] K. Miyasawa, On the convergence of the learning process in a 2×2 non-zero-sum game. *Economic Research Program, Princeton University, Research Memorandum* **33** (1961).

[50] D. Monderer and A. Sela, A 2×2 game without the fictitious play property. *Games Econom. Behav.* **14** (1996), 144–148.

[51] D. Monderer and L. Shapley, Fictitious play property for games with identical interests. *Games Econom. Behav.* **68** (1996), 258–265.

[52] D. Monderer and L. Shapley, Potential games. *Games Econom. Behav.* **14** (1996), 124–143.

[53] R. Nagel, Unraveling in guessing games: An experimental study. *Am. Econ. Rev.* **5** (1995), 1313–1326.

[54] S. Perkins and D. S. Leslie, Asynchronous stochastic approximation with differential inclusions. *Stoch. Syst.* **2** (2012), no. 2, 409–446.

[55] J. Pérolat, B. Piot, and O. Pietquin, Actor–critic fictitious play in simultaneous move multistage games. In *International conference on artificial intelligence and statistics*, pp. 919–928, PMLR, 2018.

[56] J. Pérolat, B. Scherrer, B. Piot, and O. Pietquin, Approximate dynamic programming for two-player zero-sum Markov games. In *International conference on machine learning*, pp. 1321–1329, PMLR, 2015.

[57] J. Pérolat, F. Strub, B. Piot, and O. Pietquin, Learning Nash Equilibrium for General-Sum Markov Games from Batch Data. In *International conference on artificial intelligence and statistics*, pp. 232–241, PMLR, 2017.

[58] B. T. Poljak and Y. Z. Tsypkin, Pseudogradient adaptation and training algorithms. *Autom. Remote Control* **12** (1973), 83–94.

[59] J. Robinson, An iterative method of solving a game. *Ann. of Math.* (1951), 296–301.

[60] T. Roughgarden, Intrinsic robustness of the price of anarchy. In *ACM symposium on theory of computing*, pp. 513–522, Association for Computing Machinery, 2009.

[61] T. Roughgarden, Algorithmic game theory. *Commun. ACM* **53** (2010), no. 7, 78–86.

[62] W. H. Sandholm, Preference evolution, two-speed dynamics, and rapid social change. *Rev. Econ. Dyn.* **4** (2001), no. 3, 637–679.

[63] M. O. Sayin, F. Parise, and A. Ozdaglar, Fictitious play in zero-sum stochastic games. 2020, arXiv:2010.04223.

[64] M. O. Sayin, K. Zhang, D. S. Leslie, T. Başar, and A. Ozdaglar, Decentralized Q-learning in zero-sum markov games. In *Thirty-fifth conference on neural information processing systems*, 2021.

[65] A. Sela, Fictitious play in "one-against-all" multi-player games. *Econom. Theory* **14** (1999), 635–651.

[66] D. Shah, V. Somani, Q. Xie, and Z. Xu, On reinforcement learning for turn-based zero-sum Markov games. 2020, arXiv:2002.10620.

[67] S. Shalev-Shwartz, Online learning and online convex optimization. *Found. Trends Mach. Learn.* **4** (2011), no. 2, 107–194.

[68] S. Shalev-Shwartz, S. Shammah, and A. Shashua, Safe, multi-agent, reinforcement learning for autonomous driving. 2016, arXiv:1610.03295.

[69] L. S. Shapley, Stochastic games. *Proc. Natl. Acad. Sci. USA* **39** (1953), no. 10, 1095–1100.

[70] L. S. Shapley, Some topics in two-person games. *Adv. Game Theory* **52** (1964), 1–29.

[71] Y. Shoham and K. Leyton-Brown, *Multiagent systems: algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[72] A. Sidford, M. Wang, L. Yang, and Y. Ye, Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International conference on artificial intelligence and statistics*, pp. 2992–3002, PMLR, 2020.

[73] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search. *Nature* **529** (2016), no. 7587, 484–489.

[74] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of Go without human knowledge. *Nature* **550** (2017), no. 7676, 354–359.

[75] V. Syrgkanis and E. Tardos, Composable and efficient mechanisms. In *ACM symposium on theory of computing*, pp. 211–220, Association for Computing Machinery, 2013.

[76] C. Szepesvári and M. L. Littman, A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Comput.* **11** (1999), no. 8, 2017–2060.

[77] M. Takahashi, Equilibrium points of stochastic non-cooperative n-person games. *J. Sci. Hiroshima Univ., Ser. A-I* **28** (1964), 95–99.

[78] M. Tan, Multi-agent reinforcement learning: independent vs. cooperative agents. In *International conference on machine learning*, pp. 330–337, PMLR, 1993.

[79] Y. Tian, Y. Wang, T. Yu, and S. Sra, Online learning in unknown Markov games. In *International conference on machine learning*, pp. 10279–10288, PMLR, 2021.

[80] J. N. Tsitsiklis, Asynchronous stochastic approximation and Q-learning. *Mach. Learn.* **16** (1994), 185–202.

[81] B. Van der Genugten, A weakened form of fictitious play in two-person zero-sum games. *Int. Game Theory Rev.* **2** (2000), no. 4, 307–328.

[82] C. J. C. H. Watkins and P. Dayan, Q-learning. *Mach. Learn.* **8** (1992), no. 3, 279–292.

[83] C.-Y. Wei, Y.-T. Hong, and C.-J. Lu, Online reinforcement learning in stochastic games. In *Advances in neural information processing systems*, pp. 4987–4997, Curran Associates, Inc., 2017.

[84] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on learning theory 134*, pp. 4259–4299, PMLR, 2021.

[85] Q. Xie, Y. Chen, Z. Wang, and Z. Yang, Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pp. 3674–3682, PMLR, 2020.

[86] Y. Yang, J. Li, and L. Peng, Multi-robot path planning based on a deep reinforcement learning DQN algorithm. *CAAI Trans. Intell. Technol.* **5** (2020), no. 3, 177–183.

[87] K. Zhang, S. Kakade, T. Başar, and L. Yang, Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity. In *Advances in neural information processing systems 33*, pp. 1166–1178, Curran Associates, Inc., 2020.

[88] K. Zhang, Z. Yang, and T. Başar, Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in neural information processing systems*, pp. 11598–11610, Curran Associates, Inc., 2019.

[89] K. Zhang, Z. Yang, and T. Başar, Multi-agent reinforcement learning: a selective overview of theories and algorithms. In *Handbook of reinforcement learning and control*, pp. 321–384, Stud. Syst. Decis. Control. Springer, 2021.

[90] Y. Zhao, Y. Tian, J. D. Lee, and S. S. Du, Provably efficient policy gradient methods for two-player zero-sum Markov games. 2021, arXiv:2102.08903.

[91] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent. In *International conference on machine learning*, pp. 928–936, PMLR, 2003.

ASUMAN OZDAGLAR
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, asuman@mit.edu

MUHAMMED O. SAYIN
Electrical and Electronics Engineering Department in Bilkent University, Ankara, Turkey, sayin@ee.bilkent.edu.tr

KAIQING ZHANG
LIDS and CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA, kaiqing@mit.edu

REACHABLE STATES FOR INFINITE-DIMENSIONAL LINEAR SYSTEMS: OLD AND NEW

MARIUS TUCSNAK

ABSTRACT

This work describes some recent results on the reachable spaces for infinite dimensional linear time invariant systems. The focus is on systems described by the constant coefficients heat equation, when the question is shown to be intimately connected to the theory of Hilbert spaces of analytic functions.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 93B03; Secondary 93B05, 35K08, 30H20

KEYWORDS

Reachable space, null controllability, Bergman spaces, smooth inputs, control cost

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5374–5395
DOI 10.4171/ICM2022/12

1. INTRODUCTION

Determining the reachable states of a controlled dynamical system is a major question in control theory. The set formed by these states measures our capability of acting on a system and provides important information for safety verifications. This fundamental question is well understood for linear finite-dimensional systems but much less is known in an infinite-dimensional context (namely for systems governed by partial differential equations). Most of the known results concern the case when the system is exactly controllable, which means, as reminded below, that the reachable state coincides with the state space of the system. When the reachable space is a strict subspace of the state space, its description is generally far from being complete. Note that for infinite-dimensional systems, as recalled below, the reachable space also serves to define the main controllability types in a precise and condensed manner.

The present work aims at describing some of the major advances in this field, with focus on those involving interactions with complex and harmonic analysis techniques. With no claim of exhaustiveness, we first briefly discuss some of the interactions which are by now classical (such as those based on Ingham–Beurling-type theorems) and then we describe recent advances involving various complex analysis techniques, such as the theory of reproducing kernel Hilbert spaces (namely of Bergman type) or separation of singularities for spaces of holomorphic functions.

The study of the reachable space and of the controllability of finite-dimensional linear control systems have been set at the center of control theory by the works of R. Kalman in the 1960s (see, for instance, [20]). Controllability theory for infinite-dimensional linear control systems emerged soon after. Among the early contributors we mention D. L. Russell, H. Fattorini, T. Seidman, A. V. Balakrishnan, R. Triggiani, W. Littman, and J.-L. Lions. The latter gave the field an enormous impact with his book [26], which opened the way to fascinating interactions of controllability theory with various fields of analysis.

The related question of the study of the *reachable space of infinite-dimensional linear control systems*, namely those governed by partial differential equations, has been initiated, as far as we know, by the papers of Russell [31] and Fattorini and Russell [11]. In these famous papers the authors provide relevant information on the reachable space of systems described by hyperbolic and parabolic partial differential equations in one space dimension controlled from the boundary.

The techniques generally employed for one-dimensional wave or Euler–Bernoulli plate equations are quite close to those used for the corresponding controllability problems, in particular Ingham–Beurling-type theorems, and they often provide full characterizations of the reachable space. To give the reader a flavor of the techniques used for systems describing one-dimensional elastic structures, we give an abstract result in Section 3 and an illustrating example in Section 4. The situation is much more complicated for the wave equation in several space dimensions where (with the exception of the exactly controllable case) characterizing the reachable spaces is essentially an open question.

On the other hand, determining the reachable states for systems described by the heat equation with boundary control is an extremely challenging question, on which major advances have been obtained within the last years. Indeed, due to the smoothing effect of the heat kernel, the reachable states are expected to be very smooth functions. However, since the control functions are in general only in L^2, the characterization of the reachable space, even in apparently very simple situations, is a difficult question, solved only very recently. To be more precise, consider the system

$$\begin{cases} \dfrac{\partial \theta}{\partial t}(t,x) = \dfrac{\partial^2 \theta}{\partial x^2}(t,x), & t \geqslant 0,\ x \in (0,\pi), \\ \theta(t,0) = u_0(t), \quad \theta(t,\pi) = u_\pi(t), & t \in [0,\infty), \\ \theta(0,x) = 0, & x \in (0,\pi), \end{cases} \tag{1.1}$$

which models the heat propagation in a rod of length π, controlled by prescribing the temperature at both ends. It is well known that for every $u_0, u_\pi \in L^2[0,\infty)$, problem (1.1) admits a unique solution θ and that the restriction of this function to $(0,\infty)\times(0,\pi)$ is an analytic function. The *input-to-state maps* (briefly, input maps) $(\Phi_\tau^{\text{heat}})_{\tau\geqslant 0}$ are defined by

$$\Phi_\tau^{\text{heat}} \begin{bmatrix} u_0 \\ u_\pi \end{bmatrix} = \theta(\tau,\cdot) \quad \big(\tau \geqslant 0,\ u_0, u_\pi \in L^2[0,\tau]\big). \tag{1.2}$$

Determining the reachable space at instant τ of the system determined by the 1D *heat equation with Dirichlet boundary control consists in determining* $\operatorname{Ran}\Phi_\tau^{\text{heat}}$.

The first result on this space goes back to [11], where it is shown that the functions which extend holomorphically to a horizontal strip containing $[0,\pi]$ and vanishing, together with all their derivatives of even order, at $x=0$ and $x=\pi$, belong to $\operatorname{Ran}\Phi_\tau^{\text{heat}}$. The fact that some other types of functions (like polynomials), not necessarily vanishing at the extremities of the considered interval, are in the reachable space has been remarked in a series of papers published in the 1980s (see, for instance, Schmidt [36] and the references therein). A significant advance towards such a characterization was reported only in 2016, in the work by Martin, Rosier, and Rouchon [27], where it was shown that any function which can be extended to a holomorphic map in a disk centered in $\frac{\pi}{2}$ and of diameter $\pi e^{(2e)^{-1}}$ lies in the reachable space. This result has been further improved in Dardé and Ervedoza [8], where it has been shown that any function which can be extended to a holomorphic one in a neighborhood of the square D defined by

$$D = \big\{ s = x + iy \in \mathbb{C} \mid |y| < x \text{ and } |y| < \pi - x \big\} \tag{1.3}$$

lies in the reachable space.

On the other hand, it is not difficult to check (see, for instance, [27, THEOREM 1]) that if $\psi \in \operatorname{Ran}\Phi_\tau^{\text{heat}}$ then ψ can be extended to a function holomorphic in D, so that the assertion in [8] suggests that the reachable space could in this case be connected to a classical space of holomorphic functions defined on D. This has been confirmed by a series of recent papers (see, Hartmann, Kellay, and Tucsnak [13], Normand, Kellay, and Tucsnak [21], Orsoni [29], and Hartmann and Orsoni [14]) which led to a full characterization of this space to be described in Section 6.

2. SOME BACKGROUND ON WELL-POSED LINEAR CONTROL SYSTEMS

The concept of a well-posed linear system, introduced in Salamon [35] and further developed in Weiss [44], plays an important role in control theory for infinite-dimensional systems. We briefly recall below some basic facts about these systems, including the definition of the reachable space and the three main controllability types.

Let U (the input space) and X (the state space) be Hilbert spaces (possibly infinite-dimensional). The spaces U and X will be constantly identified with their duals and, if there is no risk of confusion, the inner product and norm in these spaces will be simply denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively.

From a system-theoretic viewpoint, the simplest way to define a linear well-posed time-invariant system in a possibly infinite-dimensional setting is to introduce families of operators satisfying the properties in the definition below.

Definition 2.1. Let U and X be Hilbert spaces. A *well-posed linear control system* with input space U and state space X is a couple $\Sigma = (\mathbb{T}, \Phi)$ of families of operators such that

(1) $\mathbb{T} = (\mathbb{T}_t)_{t \geqslant 0}$ is an operator semigroup on X, i.e.,

- $\mathbb{T}_t \in \mathcal{L}(X)$ for every $t \geqslant 0$,
- $\mathbb{T}_0 \psi = \psi$ for every $\psi \in X$,
- $\mathbb{T}_{t+\tau} = \mathbb{T}_t \mathbb{T}_\tau$ $(t, \tau \geqslant 0)$,
- $\lim_{t \to 0+} \mathbb{T}_t \psi = \psi$ $(\psi \in X)$;

(2) For every $t \geqslant 0$, we have $\Phi_t \in \mathcal{L}(L^2([0, \infty); U), X)$ and

$$\Phi_{\tau+t}(u \underset{\tau}{\diamondsuit} v) = \mathbb{T}_t \Phi_\tau u + \Phi_t v \quad (t, \tau \geqslant 0), \tag{2.1}$$

where the *τ-concatenation* of two signals u and v, denoted by $u \underset{\tau}{\diamondsuit} v$, is the function

$$u \underset{\tau}{\diamondsuit} v = \begin{cases} u(t) & \text{for } t \in [0, \tau), \\ v(t - \tau) & \text{for } t \geqslant \tau. \end{cases} \tag{2.2}$$

It can be shown that the above properties imply that the map

$$(t, u) \mapsto \Phi_t u,$$

is continuous from $[0, \infty) \times L^2([0, \infty); U)$ to X.

Let $A : \mathcal{D}(A) \to X$ be the generator of $\mathbb{T} = (\mathbb{T}_t)_{t \geqslant 0}$ on X. We denote by $\mathbb{T}^*$ the adjoint semigroup, which is generated by the adjoint of A^* of A. The operator domain $\mathcal{D}(A)$, when endowed with norm

$$\|\varphi\|_{X_1}^2 = \|\varphi\|^2 + \|A\varphi\|^2 \quad (\varphi \in X_1), \tag{2.3}$$

is a Hilbert space. This Hilbert space is denoted by X_1. Similarly, we denote by X_1^d the Hilbert space obtained by endowing $\mathcal{D}(A^*)$ with the norm

$$\|\varphi\|_{X_1^d}^2 = \|\varphi\|^2 + \left\|A^* \varphi\right\|^2 \quad \left(\varphi \in X_1^d\right). \tag{2.4}$$

Let X_{-1} be the dual of X_1^d with respect to the pivot space X, so that $X_1 \subset X \subset X_{-1}$ with continuous and dense embeddings. Note that, for each $k \in \{-1, 1\}$, the original semigroup $\mathbb{T}$ has a restriction (or an extension) to X_k that is the image of $\mathbb{T}$ through the unitary operator $(\beta I - A)^{-k} \in \mathcal{L}(X, X_k)$, where $\beta \in \rho(A)$ (the resolvent set of A). We refer to [41, REMARK 2.10.5] for a proof of the last statement. This restriction (extension) will be still denoted by $\mathbb{T}$.

An important consequence of Definition 2.1 is (see, for instance, [44]) that there exists a unique $B \in \mathcal{L}(U, X_{-1})$, called the *control operator* of Σ, such that

$$\Phi_\tau u = \int_0^\tau \mathbb{T}_{\tau-\sigma} B u(\sigma)\, d\sigma \quad \big(\tau \geqslant 0,\ u \in L^2\big([0,\infty); U\big)\big). \tag{2.5}$$

Notice that in the above formula, $\mathbb{T}$ acts on X_{-1} and the integration is carried out in X_{-1}. The operator B can be found by

$$B\mathrm{v} = \lim_{\tau \to 0} \frac{1}{\tau} \Phi_\tau(\chi \cdot \mathrm{v}) \quad (\mathrm{v} \in U), \tag{2.6}$$

where χ denotes the characteristic function of the interval $[0, 1]$. We mention that it follows from the above definitions that if $(\mathbb{T}, \Phi)$ is a well-posed control system then for all $u \in L^2([0,\infty); U)$, $t \mapsto \Phi_t u$ is a continuous function from $[0,\infty)$ to X.

From the above facts, it follows that a well-posed control system can alternatively be described by a pair (A, B), where $A : \mathcal{D}(A) \to X$ generates a C^0-semigroup $\mathbb{T}$ on X and $B \in \mathcal{L}(U, X_{-1})$ is an *admissible control operator for* $\mathbb{T}$. This latter property means that for some $t > 0$, the operator Φ_t defined by (2.5) has its range contained in X. We refer to Tucsnak and Weiss [41, SECTIONS 4 AND 5] for more material on this concept.

We also recall (see, for instance, [41, PROPOSITION 4.2.5]) that the families $\mathbb{T}$ and Φ can also be seen as the solution operators for the initial value problem

$$\dot{z}(t) = Az(t) + Bu(t), \quad z(0) = z_0, \tag{2.7}$$

in the following sense:

Proposition 2.1. *Let $\tau > 0$. Then for every $z_0 \in X$ and every $u \in L^2([0,\tau]; U)$, the initial value problem (2.7) has a unique solution*

$$z \in C\big([0,\tau]; X\big) \cap H^1\big((0,\tau); X_{-1}\big).$$

This solution is given by

$$z(t) = \mathbb{T}_t z_0 + \Phi_t u \quad \big(t \in [0,\tau]\big). \tag{2.8}$$

In most of the remaining part of this work, we describe a well-posed control system either by a couple $(\mathbb{T}, \Phi)$ as in Definition 2.1 or by a couple (A, B), where A is the generator of $\mathbb{T}$ and B is the unique operator in $\mathcal{L}(U, X_{-1})$ satisfying (2.5).

Given a well-posed control system $\Sigma = (\mathbb{T}, \Phi)$ and $\tau > 0$, *the reachable space in time τ of* Σ is defined as the range $\mathrm{Ran}\,\Phi_\tau$ of the operator Φ_τ. This space can be endowed with the norm induced from $L^2([0,\tau]; U)$, which is

$$\|\eta\|_{\mathrm{Ran}\,\Phi_\tau} = \inf_{\substack{u \in L^2([0,\tau];U) \\ \Phi_\tau u = \eta}} \|u\|_{L^2([0,\tau];U)} \quad (\eta \in \mathrm{Ran}\,\Phi_\tau). \tag{2.9}$$

Remark 2.1. From the above construction of the reachable space, it easily follows (see, for instance, Saitoh and Sawano [34, THEOREM 2.36]) that, when endowed with the norm (2.9), Ran Φ_τ becomes a Hilbert space, isomorphic to the orthogonal complement in $L^2([0,\tau];U)$ of Ker Φ_τ.

Remark 2.2. We obviously have that Φ_τ is onto from $L^2([0,\tau];U)$ onto Ran Φ_τ. Moreover, we have

$$\|\Phi_\tau\|_{\mathcal{L}(L^2([0,\tau];U),\operatorname{Ran}\Phi_\tau)} = 1. \tag{2.10}$$

Indeed, we clearly have that

$$\|\Phi_\tau\|_{\mathcal{L}(L^2([0,\tau];U),\operatorname{Ran}\Phi_\tau)} \leqslant 1.$$

Moreover, if $\eta \in \operatorname{Ran}\Phi_\tau \setminus \{0\}$ there exists a sequence $(u_n)_{n\geqslant 0}$ in $(L^2([0,\tau];U)\setminus\{0\})^{\mathbb{N}}$ such that $\Phi_\tau u_n = \eta$ for every $n \in \mathbb{N}$ and $\|u_n\|_{L^2([0,\tau];U)} \to \|\eta\|_{\operatorname{Ran}\Phi_\tau}$ as $n \to \infty$. We thus have that

$$\lim_{n\to\infty} \frac{\|\Phi_\tau u_n\|_{\operatorname{Ran}\Phi_\tau}}{\|u_n\|_{L^2([0,\tau];U)}} = 1,$$

and, consequently, we have (2.10).

If the spaces U and X are finite-dimensional then there exists $A \in \mathcal{L}(X)$ such that $\mathbb{T}_t = \exp(tA)$ for every $t \geqslant 0$ and $B \in \mathcal{L}(U,X)$. In this case the following result, known as the *Kalman rank condition* for controllability, holds:

Proposition 2.2. *If U and X are finite-dimensional then we have, for every $\tau > 0$,*

$$\operatorname{Ran}\Phi_\tau = \operatorname{Ran}\begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix}. \tag{2.11}$$

Remark 2.3. From Proposition 2.2, it follows in particular that for finite-dimensional systems the reachable space does not depend on the time horizon $\tau > 0$. Moreover, it is not difficult to check (see, for instance, Normand, Kellay, and Tucsnak [21]) that Proposition 2.2 implies that Ran Φ_τ coincides with the range of the restriction of Φ_τ to signals which can be extended to entire functions from $\mathbb{C}$ to U.

Unlike the finite-dimensional case, for general well-posed linear control systems, there is no simple characterization of the reachable space in terms of the operators A and B. Moreover, this space depends in general on τ and, for most systems described by partial differential equations, we have only a small amount of information on the reachable space. Another difference with respect to the finite-dimensional case is that the range Ran Φ_τ^∞ of the restriction of Φ_τ to a smaller space (such as $L^\infty([0,\tau];U)$) is in general a strict subset of Ran Φ_τ.

The concept of reachable space appears, in particular, in the definition of the main three controllability concepts used in the infinite-dimensional system theory.

Definition 2.2. Let $\tau > 0$ and let the pair $(\mathbb{T}, \Phi)$ define a well-posed control LTI system.

- The pair $(\mathbb{T}, \Phi)$ is *exactly controllable in time* τ if Ran $\Phi_\tau = X$.

- $(\mathbb{T}, \Phi)$ is *approximately controllable in time* τ if $\operatorname{Ran} \Phi_\tau$ is dense in X.
- The pair $(\mathbb{T}, \Phi)$ is *null-controllable in time* τ if $\operatorname{Ran} \Phi_\tau \supset \operatorname{Ran} \mathbb{T}_\tau$.

From the above definition, we see that for systems which are approximately controllable in some time $\tau > 0$ we can define the dual $(\operatorname{Ran} \Phi_\tau)'$ of $\operatorname{Ran} \Phi_\tau$ with respect to the pivot space X (we refer to Tucsnak and Weiss [41, SECTION 2.9] for the general definition of this concept). More precisely:

Definition 2.3. Let $\Sigma = (\mathbb{T}, \Phi)$ be approximately controllable in time τ and let $(\operatorname{Ran} \Phi_\tau)'$ be the dual of $\operatorname{Ran} \Phi_\tau$ with respect to the pivot space X, so that we have

$$\operatorname{Ran} \Phi_\tau \subset X \subset (\operatorname{Ran} \Phi_\tau)',$$

with continuous and dense inclusions.

The dual $\Phi'_\tau \in \mathcal{L}((\operatorname{Ran} \Phi_\tau)', L^2([0,\tau]; U))$ of the operator Φ_τ introduced in (2.5) is defined by

$$\langle \Phi_\tau u, \eta \rangle_{\operatorname{Ran} \Phi_\tau, (\operatorname{Ran} \Phi_\tau)'} = \big\langle u, (\Phi_\tau)' \eta \big\rangle_{L^2([0,\tau];U)},$$

for every $u \in L^2([0,\tau]; U)$ and $\eta \in (\operatorname{Ran} \Phi_\tau)'$.

It can be easily checked that the norm in the space $\operatorname{Ran} \Phi_\tau$ can be characterized as follows:

Proposition 2.3. *Assume that* (A, B) *is approximately controllable in some time* $\tau > 0$. *Then*

$$\|\eta\|_{(\operatorname{Ran} \Phi_\tau)'} = \|\Phi_\tau^* \eta\|_{L^2([0,\tau];U)} \quad (\eta \in X), \tag{2.12}$$

where $\Phi_\tau^* \in \mathcal{L}(X, L^2([0,\tau]; U))$ *is the adjoint of* Φ_τ *defined by*

$$\langle \Phi_\tau u, \eta \rangle_X = \big\langle u, \Phi_\tau^* \eta \big\rangle_{L^2([0,\tau];U)} \quad \big(u \in L^2\big([0,\tau]; U\big),\ \eta \in X\big).$$

Note that the fact that the right-hand side of (2.12) defines a norm follows from the fact that $\operatorname{Ran} \Phi_\tau$ is dense in X.

A direct consequence of Proposition 2.3 is the following characterization of $(\operatorname{Ran} \Phi_\tau)'$:

Proposition 2.4. *If* (A, B) *is approximately controllable in time* $\tau > 0$ *then* $(\operatorname{Ran} \Phi_\tau)'$ *coincides with the completion of* X *with respect to the norm* $\eta \mapsto \|\Phi_\tau^* \eta\|_{L^2([0,\tau];U)}$.

By combining the above result with a classical duality argument (see, for, instance, [41, PROPOSITION 4.4.1]), we obtain:

Corollary 2.1. *If* (A, B) *is approximately controllable in time* $\tau > 0$ *then* $(\operatorname{Ran} \Phi_\tau)'$ *coincides with the completion of* $\mathcal{D}(A^*)$ *with respect to the norm* $\eta \mapsto (\int_0^\tau \|B^* \mathbb{T}_t^* \eta\|^2 \, \mathrm{d}t)^{\frac{1}{2}}$.

As already mentioned, in the infinite-dimensional case the reachable space generally depends on the time horizon τ. However, as precisely stated below, there exists an important class of infinite-dimensional systems for which the reachable space is independent of the time horizon.

Proposition 2.5. *If the well-posed linear control system* $(\mathbb{T}, \Phi)$ *is null-controllable in any positive time then* $\operatorname{Ran} \Phi_\tau$ *does not depend on* $\tau > 0$.

Following the ideas of [37], a very short proof of the above result is provided in [21].

3. SINGLE INPUT SYSTEMS WITH SKEW-ADJOINT GENERATOR

In this section we consider, for the sake of simplicity, a class of systems which can be seen as a "toy model" for many linear control problems involving the dynamics of flexible structures. In fact, our abstract result in Theorem 3.1 below can be directly applied only to problems in one space dimension. Nevertheless, estimates similar to the inequality in Theorem 3.2 below can be used when tackling some problems in several space dimensions, at least in particular geometric configurations (see, for instance, Allibert [2], Jaffard [17], Jaffard and Micu [18], or Komornik and Loreti [22]). The situation is much more complicated, requiring different techniques, in several space dimensions and with arbitrary shapes of the domain filled by the elastic structure, see, for instance, Avdonin, Belishev, and Ivanov [3].

Let $A : \mathcal{D}(A) \to X$ be a skew-adjoint operator, with nonempty resolvent set $\rho(A)$ and with compact resolvents. We denote by $(\phi_k)_{k\in\mathbb{Z}^*}$ an orthonormal basis of X consisting of eigenvectors of A. For every $k \in \mathbb{Z}^*$, we denote by $i\lambda_k$ the eigenvalue associated to the eigenvector ϕ_k, so that λ_k is real for all $k \in \mathbb{Z}^*$. Without loss of generality, we can assume that $\lambda_1 \geqslant \lambda_{-1}$ and

$$\lambda_{n+1} - \lambda_n \geqslant 0 \quad \big(n \in \mathbb{Z}^* \setminus \{-1\}\big). \tag{3.1}$$

According to Stone's theorem, the operator A generates a strongly continuous group of unitary operators on X. This group, denoted by $\mathbb{T} = (\mathbb{T}_t)_{t\in\mathbb{R}}$, is described by the formula

$$\mathbb{T}_t \psi = \sum_{k\in\mathbb{Z}^*} \langle \psi, \phi_k \rangle \exp(i\lambda_k t)\phi_k \quad (t \in \mathbb{R},\ \psi \in X). \tag{3.2}$$

Assume that the control space U is one-dimensional (i.e., that $U = \mathbb{C}$) and that the control operator $B \in \mathcal{L}(U; X_{-1})$ is given by

$$B\mathrm{u} = \mathrm{u}b \quad (\mathrm{u} \in U), \tag{3.3}$$

with b a fixed element of X_{-1}, where, as mentioned in Section 2, X_{-1} is the dual of $\mathcal{D}(A^*)$ with respect to the pivot space X. For b as above and $\psi \in \mathcal{D}(A)$, the notation $\langle b, \psi \rangle$ stands for the duality product of b and ψ. For every $k \in \mathbb{N}$, we set

$$b_k := \langle b, \phi_k \rangle. \tag{3.4}$$

The main result in this section is:

Theorem 3.1. *Let A be a skew-adjoint operator with compact resolvents on X with spectrum $\sigma(A) = i\Lambda$, where $\Lambda = (\lambda_k)_{k\in\mathbb{Z}^*}$ is a* regular *sequence of real numbers, i.e., with*

$$\gamma_1 := \inf_{\substack{n\in\mathbb{Z}^* \\ n\neq -1}} |\lambda_{n+1} - \lambda_n| > 0. \tag{3.5}$$

Moreover, assume that there exist $p \in \mathbb{N}$ *and* $\gamma_p > 0$ *such that*

$$\gamma_p := \inf_{\substack{n\in\mathbb{Z}^* \\ n\neq -p}} \left(\frac{\lambda_{n+p} - \lambda_n}{p} \right) > 0. \tag{3.6}$$

Finally, suppose that the sequence (b_k) *defined in* (3.4) *is bounded and that* $b_k \neq 0$ *for every* $k \in \mathbb{Z}^*$. *Then for every* $\tau > \frac{2\pi}{\gamma_p}$, *the input map* Φ_τ *of the system* (A, B) *(with* B *defined in* (3.3)*) satisfies*

$$\operatorname{Ran} \Phi_\tau = \left\{ \eta \in X \;\middle|\; \sum_{k\in\mathbb{Z}^*} |b_k|^{-2} \big|\langle \eta, \phi_k \rangle\big|^2 < \infty \right\}. \tag{3.7}$$

Remark 3.1. The assumption that $b_k \neq 0$ for every $k \in \mathbb{Z}^*$ is not essential. Indeed, it is not difficult to check that for every $b \neq 0$ we have that $\operatorname{Ran} \Phi_\tau$ is contained in the closed span $\tilde{X}$ of the set $\{\varphi_k \mid b_k \neq 0\}$. Consequently, we can apply Theorem 3.6 to the restriction of our original system to $\tilde{X}$ and obtain that

$$\operatorname{Ran} \Phi_\tau = \left\{ \eta \in \tilde{X} \;\middle|\; \sum_{\substack{k\in\mathbb{Z}^* \\ b_k\neq 0}} |b_k|^{-2} \big|\langle \eta, \phi_k \rangle\big|^2 < \infty \right\}.$$

The proof of Theorem 3.1 is based on a class of results playing, more generally, an important role in the study of reachability questions for the 1D elastic structures. More precisely, we refer here to several inequalities coming from the theory of nonharmonic Fourier series, introduced in Ingham [16]. In particular, we use below the following generalization of Parseval's inequality:

Proposition 3.1 (Ingham, 1936). *Let* $\Lambda = (\lambda_n)_{n\in\mathbb{Z}^*}$ *be a real sequence satisfying* (3.5). *Then for any interval* I *with length* $|I|$ *there exists a constant* c, *depending on* $|I|$ *and* γ_1, *such that*

$$\int_I \left| \sum_{n\in\mathbb{Z}^*} a_n \exp(i\lambda_n t) \right|^2 \mathrm{d}t \leqslant c \sum_{n\in\mathbb{Z}^*} |a_n|^2,$$

for any sequence $(a_n) \in \ell^2(\mathbb{Z}^*, \mathbb{C})$.

It is not difficult to check that the proposition above implies the following admissibility result for (3.3) (note that the result below can also be seen as a particular case of the admissibility conditions given in Ho and Russell [15] and Weiss [43]).

Proposition 3.2. *Let* A *be a skew-adjoint operator with compact resolvents on* X *with spectrum* $\sigma(A) = i\Lambda$, *where* $\Lambda = (\lambda_k)_{k\in\mathbb{Z}^*}$ *satisfies* (3.5). *Assume that* $b \in X_{-1}$ *is such that for every* $k \in \mathbb{Z}^*$, *the number* b_k *defined in* (3.4) *is nonzero. Moreover, suppose that* $\sup_{k\in\mathbb{N}} |b_k| < \infty$ *(recall that the sequence* (b_k) *has been defined in* (3.4)*). Then* B *defined by* (3.3) *is an admissible control operator for* $\mathbb{T}$.

The main analytical tool in the proof of Theorem 3.1 is a lower bound for exponential sums, in the spirit of classical inequalities of Ingham [16], Beurling [5], and Kahane [19]. We give below the quantitative version proved in Tenenbaum and Tucsnak [39], making the dependency of the involved constants explicit in terms of various parameters.

Theorem 3.2. *Let $\Lambda = (\lambda_n)_{n\in\mathbb{Z}^*}$ be a real sequence satisfying (3.5) and (3.6). Then, for any $\gamma \in (0, \gamma_p)$ and interval I with length $|I| \geqslant \frac{2\pi}{\gamma}$, there exists a constant $\kappa = \kappa(\gamma_1) > 0$ such that, writing $\varepsilon := \frac{1}{2}\{1/\gamma - 1/\gamma_p\}$, we have*

$$\int_I \left| \sum_{n\in\mathbb{Z}^*} a_n \exp(i\lambda_n t) \right|^2 \mathrm{d}t \geqslant \frac{\kappa \varepsilon^{5p+2}}{p^{12p}} \sum_{n\in\mathbb{Z}^*} |a_n|^2$$

for any sequence $(a_n) \in \ell^2(\mathbb{Z}^*, \mathbb{C})$.

We are now in a position to prove Theorem 3.1.

Proof of Theorem 3.1. It is not difficult to check that our standing assumptions imply that the system (A, B) is approximately controllable in time τ. Thus, according to Corollary 2.1, it suffices to identify the completion of $\mathcal{D}(A^*) = \mathcal{D}(A)$ with respect to the norm

$$\eta \mapsto \left(\int_0^\tau \left\| B^* \mathbb{T}_t^* \eta \right\|^2 \mathrm{d}t \right)^{\frac{1}{2}}. \tag{3.8}$$

After some simple calculations, we obtain that

$$B^* \mathbb{T}_t^* \eta = \sum_{k\in\mathbb{Z}^*} b_k \langle \eta, \phi_k \rangle \exp(-i\lambda_k t) \quad \big(t \geqslant 0,\ \eta \in \mathcal{D}(A^*)\big).$$

By combining Proposition 3.1 and Theorem 3.2, it follows that the norm defined in (3.8) is equivalent to the norm

$$\eta \mapsto \left(\sum_{k\in\mathbb{Z}^*} |b_k|^2 \big|\langle \eta, \phi_k \rangle\big|^2 \right)^{\frac{1}{2}}. \tag{3.9}$$

We can thus use Corollary 2.1 to conclude that the dual $(\operatorname{Ran} \Phi_\tau)'$ of $\operatorname{Ran} \Phi_\tau$ with respect to the pivot space X is the completion of $\mathcal{D}(A)$ with respect to the norm defined in (3.9).

On the other hand, the completion of $\mathcal{D}(A)$ with respect to the norm defined in (3.9) clearly coincides with the dual with respect to the pivot space X of the space

$$\left\{ \eta \in X \;\middle|\; \sum_{k\in\mathbb{Z}^*} |b_k|^{-2} \big|\langle \eta, \phi_k \rangle\big|^2 < \infty \right\},$$

so that we obtain the conclusion (3.7). ∎

4. AN EXAMPLE COMING FROM ELASTICITY

In this section we show how the abstract result in Theorem 3.1 can be applied to determine the reachable space of a system describing the vibrations of an Euler–Bernoulli beam with piezoelectric actuator. More precisely, we consider the initial and boundary value problem modeling the vibrations of an Euler–Bernoulli beam which is subject to the action of a piezoelectric actuator. Most of the results in this section appear, using a different terminology, in Tucsnak [40].

If we suppose that the beam is hinged at both ends and that the actuator is excited in a manner so as to produce pure bending moments, the model for the controlled beam can be written as (see, for instance, Crawley [7] or Destuynder et al. [10]):

$$\ddot{w}(t,x) + \frac{\partial^4 w}{\partial x^4}(t,x) = u(t)\frac{\mathrm{d}}{\mathrm{d}x}\big[\delta_b(x) - \delta_a(x)\big] \quad (0 < x < \pi,\ t > 0), \tag{4.1}$$

$$w(t,0) = w(t,\pi) = 0, \quad \frac{\partial^2 w}{\partial x^2}(t,0) = \frac{\partial^2 w}{\partial x^2}(t,\pi) = 0 \quad (t \geqslant 0), \tag{4.2}$$

$$w(0,x) = 0, \quad \dot{w}(0,x) = 0 \quad (0 < x < \pi). \tag{4.3}$$

In the equations above, w represents the transverse deflection of the beam, $a, b \in (0,\pi)$ stand for the ends of the actuator, and δ_y is the Dirac mass at the point y. Moreover, $\dot{w}$, $\ddot{w}$ denote the partial derivatives of w with respect to time. The control is the function u representing the time variation of the voltage applied to the actuator.

It is easily seen that equations (4.1)–(4.3) can be written, using the standard notation for Sobolev spaces, using a second-order abstract form in the space $H = H^{-1}(0,\pi)$. More precisely, the system (4.1)–(4.3) can be rephrased as

$$\ddot{w}(t) + A_0^2 w(t) = B_0 u(t) \quad (t > 0), \tag{4.4}$$

$$w(0) = 0, \quad \dot{w}(0) = 0, \tag{4.5}$$

where A_0 is the Dirichlet Laplacian on $(0,\pi)$ defined by

$$\mathcal{D}(A_0) = H_0^1(0,\pi), \tag{4.6}$$

$$A_0\varphi = -\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2} \quad \big(\varphi \in \mathcal{D}(A_0)\big), \tag{4.7}$$

and the operator B_0 is defined by

$$B_0 \mathrm{u} = \mathrm{u}\frac{\mathrm{d}}{\mathrm{d}x}(\delta_b - \delta_a) \quad (u \in \mathbb{C}). \tag{4.8}$$

We first recall the following well-posedness result from [40]:

Proposition 4.1. *Equations* (4.1)–(4.3) *determine a well-posed control system with state space* $X = \mathcal{D}(A_0) \times H$ *and control space* $U = \mathbb{C}$. *The corresponding semigroup generator and control operator are defined by*

$$\mathcal{D}(A) = \mathcal{D}\big(A_0^2\big) \times \mathcal{D}(A_0), \quad A = \begin{bmatrix} 0 & \mathbb{I} \\ -A_0^2 & 0 \end{bmatrix}, \tag{4.9}$$

respectively

$$B = \begin{bmatrix} 0 \\ B_0 \end{bmatrix}, \tag{4.10}$$

where the operators A_0 *and* B_0 *have been defined in* (4.6)–(4.8).

Let Φ^{beam} be the input maps associated to the well-posed system from Proposition 4.1, defined by

$$\Phi_\tau^{\mathrm{beam}} u = \begin{bmatrix} w(\tau,\cdot) \\ \dot{w}(\tau,\cdot) \end{bmatrix} \quad \big(u \in L^2\big([0,\tau];U\big)\big).$$

The main result in this section is

Proposition 4.2. *For* $\left[\begin{smallmatrix} f \\ g \end{smallmatrix}\right] \in X$, *we define the Fourier coefficients* (a_n) *and* (b_n) *by*

$$f(x) = \sum_{n=1}^{\infty} c_n \sin(nx), \quad g(x) = \sum_{n=1}^{\infty} n^2 d_n \sin(nx), \tag{4.11}$$

with (nc_n) *and* (nd_n) *in* $\ell^2(\mathbb{N}, \mathbb{C})$. *Moreover, assume that*

$$\frac{a+b}{\pi}, \ \frac{a-b}{\pi} \in \mathbb{R} - \mathbb{Q}. \tag{4.12}$$

Then for every $\tau > 0$, *we have that* $\left[\begin{smallmatrix} f \\ g \end{smallmatrix}\right] \in X$ *lies in* $\operatorname{Ran} \Phi_\tau^{\text{beam}}$ *if and only if*

$$\sum_{n\in\mathbb{N}} n^2 \sin^{-2}\left[\frac{n(a+b)}{2}\right] \sin^{-2}\left[\frac{n(a-b)}{2}\right] \left(|c_n|^2 + |d_n|^2\right) < \infty. \tag{4.13}$$

Proof. It is known that the operators A_0 and A_0^2, where A_0 has been defined in (4.6) and (4.7), are self-adjoint and positive on H (see, for instance, **[41, SECTIONS 3.3 AND 3.4]**). From this it follows that the operator A defined in (4.9) is skew-adjoint on $X = \mathcal{D}(A_0) \times H$, see **[41, SECTION 3.7]**. Moreover, we know from Proposition 4.1 that B is an admissible control operator for the unitary group $\mathbb{T}$ generated by A, so that the system (A, B) is eligible for the application of Theorem 3.1. To check that all the assumptions in this theorem are satisfied, let

$$\varphi_k(x) = k\sqrt{\frac{2}{\pi}} \sin(kx) \quad \big(k \in \mathbb{N}, \ x \in (0, \pi)\big).$$

It is easily seen that $(\varphi_k)_{k\in\mathbb{N}}$ is an orthonormal basis of H comprising eigenvectors of A_0 with corresponding eigenvalues $(\lambda_k^2)_{k\in\mathbb{N}}$, where $\lambda_k = k^2$ for every $k \in \mathbb{N}$. This enables us, according to **[41, PROPOSITION 3.7.7]**, to construct an orthonormal basis in X consisting of eigenvectors of A. More precisely, for every $k \in \mathbb{N}$, we set $\varphi_{-k} = -\varphi_k$ and $\lambda_{-k} = -\lambda_k$, and

$$\phi_k = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{i\lambda_k}\varphi_k \\ \varphi_k \end{bmatrix}. \tag{4.14}$$

Then for every $k \in \mathbb{Z}^*$, we have that ϕ_k is an eigenvector of A corresponding to the eigenvalue $i\lambda_k$ and $(\phi_k)_{k\in\mathbb{Z}^*}$ is an orthonormal basis in X.

Let us note at this stage that the sequence $(\lambda_k)_{k\in\mathbb{Z}^*}$ obviously satisfies assumption (3.5) from Theorem 3.1 and that for every $\gamma > 0$ there exists $p \in \mathbb{N}$ such that $(\lambda_k)_{k\in\mathbb{Z}^*}$ satisfies assumption (3.6) with $\gamma_p > \gamma$.

To compute the coefficients (b_k) defined in (3.4), we note that from (4.10) and (4.14) it follows that

$$\langle B\mathrm{u}, \phi_k\rangle_{X_{-1},X_1} = \frac{1}{\sqrt{2}} \langle B_0 u, \varphi_k\rangle_{[\mathcal{D}(A_0)]',\mathcal{D}(A_0)},$$

where $[\mathcal{D}(A_0)]'$ is the dual of $\mathcal{D}(A_0)$ with respect to the pivot space H. Recalling that $H = H^{-1}(0, \pi)$ and using (4.8), it follows that, for every $u \in \mathbb{C}$ and $k \in \mathbb{Z}^*$, we have

$$\langle B\mathrm{u}, \phi_k\rangle_{X_{-1},X_1} = \frac{1}{\sqrt{2}} \mathrm{u} \left(\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(A_0^{-1}\varphi_k\right)\right]_{x=a} - \left[\frac{\mathrm{d}}{\mathrm{d}x}\left(A_0^{-1}\varphi_k\right)\right]_{x=b} \right),$$

so that

$$b_k = \frac{1}{\sqrt{2}}\left(\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(A_0^{-1}\varphi_k\right)\right]_{x=a} - \left[\frac{\mathrm{d}}{\mathrm{d}x}\left(A_0^{-1}\varphi_k\right)\right]_{x=b}\right) \quad (k \in \mathbb{Z}^*).$$

After some simple calculations, we obtain that

$$b_k = \frac{1}{\sqrt{\pi}}\big(\cos(ka) - \cos(kb)\big) = \frac{2}{\sqrt{\pi}}\sin\left[\frac{k(b+a)}{2}\right]\sin\left[\frac{k(b-a)}{2}\right] \quad \big(k \in \mathbb{Z}^*\big).$$

From the above formula, it follows that the sequence (b_k) is bounded and, recalling (4.12), that $b_k \neq 0$ for every $k \in \mathbb{Z}^*$.

We have thus checked all the assumptions of Theorem 3.1. Applying this theorem to the system described by the operators of A and B defined in this section, it this follows that $\left[\begin{smallmatrix} f \\ g \end{smallmatrix}\right] \in X$ indeed belongs to the reachable space of the considered system iff (4.13) holds. ■

Remark 4.1. The result in Proposition 4.2 can be combined with some simple diophantine approximation results to obtain more explicit information on $\operatorname{Ran}\Phi_\tau^{\mathrm{beam}}$. Some of these properties are:

- There exist no locations a and b for which the system is exactly controllable. Indeed, from (4.13) it follows that the system (A, B) is exactly controllable iff the sequences $(|\sin[\frac{n(a\pm b)}{2}]|)_{n\in\mathbb{N}}$ are bounded away from zero. Or, using the continuous fraction approximation of real numbers, it is easy to check (see [40]) that there are no real numbers a and b with this property.
- The largest reachable spaces are obtained when $\frac{a\pm b}{\pi}$ can be "badly" approximated by rational numbers. In particular, if $\frac{a\pm b}{\pi}$ are quadratic irrationals (i.e., solutions of a second-order equation with integer coefficients), then

$$\operatorname{Ran}\Phi_\tau^{\mathrm{beam}} \supset \mathcal{D}(A).$$

- On the other hand, choosing a and b such that $\frac{a\pm b}{\pi}$ can be well approximated by rational numbers, the reachable space diminishes. We think, in particular, of Liouville numbers (see Valiron [42]). More precisely, for every $m \in \mathbb{N}$, there exist $a, b \in (0, \pi)$ such that $\frac{a\pm b}{\pi} \notin \mathbb{Q}$ and $\mathcal{D}(A^m)$ contains states which are not reachable.

5. THE HEAT EQUATION ON A HALF-LINE

The properties of the system we consider in this section strongly contrast those encountered in the finite-dimensional context. We just mention here that its reachable space depends on time and that the system is approximately controllable but not null-controllable. The results presented in this section are not new, but we chose to describe them in detail for two reasons. Firstly, the study of the reachable space of this system brought in new techniques in control theory for infinite-dimensional systems, essentially coming from the theory of reproducing kernel Hilbert spaces. Secondly, as it has been very recently discovered, these results have an important role in characterizing the reachable space for the controlled heat equation on a bounded interval, as it will be shown in Section 6.

Consider the initial and boundary value problem

$$\begin{cases} \dfrac{\partial v}{\partial t}(t,x) = \dfrac{\partial^2 v}{\partial x^2}(t,x) & (t \geqslant 0,\ x \in (0,\infty)), \\ v(t,0) = u_0(t), & (t \in [0,\infty)), \\ v(0,x) = 0 & (x \in (0,\infty)), \end{cases} \tag{5.1}$$

and the associated input maps $(\Phi_\tau^{\text{left}})_{\tau>0}$ defined by

$$\Phi_\tau^{\text{left}} u = v(\tau,\cdot) \quad \big(u \in L^2[0,\infty), \tau > 0\big). \tag{5.2}$$

As far as we know, the first paper with explicit control-theoretic purposes tackling the system described by the first two equations in (5.1) is Micu and Zuazua [28]. The main results in [28] assert that the first two equations in (5.1) determine a well-posed control system in appropriate spaces and that this system is not null-controllable in any time $\tau > 0$ (concerning this last assertion, we also refer to Dardé and Ervedoza [9] for an elegant proof and extensions to related PDE systems). Combining the above mentioned lack of controllability property with Proposition 2.5 suggests that Ran Φ_τ^{left} depends on the time τ. This dependence was, in fact, already made explicit in a series of papers driven by complex analysis motivations, see Aikawa et al. [1] and Saitoh [32,33]. These results came to the attention of the control-theoretic community only very recently, when they became an important ingredient in proving the main results in [13].

Before stating some of the main results from [1] and [33], we first recall some definitions concerning Bergman spaces. More precisely, for $\Omega \subset \mathbb{C}$ an open set and $\omega \in C(\Omega)$, with $|\omega(x)| > 0$ for every $x \in \Omega$, the *Bergman space on* Ω *with weight* ω, denoted $A^2(\Omega,\omega)$ is formed by all the functions f holomorphic on Ω such that $f\sqrt{|\omega|}$ is in $L^2(\Omega)$. For $\omega = 1$, this space is simply denoted by $A^2(\Omega)$. Note that $A^2(\Omega,\omega)$ becomes a Hilbert space when endowed with the norm

$$\|\psi\|^2_{A^2(\Omega,\omega)} = \int_\Omega \big|\psi(x+iy)\big|^2 \big|\omega(x+iy)\big|\,\mathrm{d}x\,\mathrm{d}y.$$

We also recall (see, for instance, [6, SECTION 4.1]) that the input maps defined in (5.2) can be alternatively described by the integral formula

$$\big(\Phi_\tau^{\text{left}} u\big)(x) = -\int_0^\tau \frac{\partial \kappa}{\partial x}(\tau-\sigma,x)u(\sigma)\,\mathrm{d}\sigma \quad \big(u \in L^2[0,\infty), \tau > 0, x \in (0,\pi)\big), \tag{5.3}$$

where

$$\kappa(t,x) = \sqrt{\frac{1}{\pi t}}\exp\left(-\frac{x^2}{4t}\right) \quad (t > 0,\ x \in \mathbb{R}) \tag{5.4}$$

is the heat kernel on $\mathbb{R}$.

For each $\tau > 0$, the range of the input map Φ_τ^{left} defined in (5.2) has been completely described in [32] as an appropriate subspace of the space of functions continuous on $(0,\pi)$ and which can be extended to a function which is holomorphic on the set Δ defined by

$$\Delta = \left\{ s \in \mathbb{C} \;\middle|\; -\frac{\pi}{4} < \arg s < \frac{\pi}{4} \right\}. \tag{5.5}$$

The precise description given in [32] of this space involves the sum of a two Hilbert spaces of holomorphic functions defined on Δ, one of them being of Bergman type. To avoid extra notational complexity, we choose to omit the precise statement of this result and to focus on the characterization of the range of the restriction of Φ_τ^{left} to the space of inputs $u(t) = \sqrt{t}\, f(t)$ with $f \in L^2[0,\tau]$. Recalling (5.3) and (5.4), this means that for every $\tau > 0$ we focus on the range of the operator defined by

$$(P_\tau f)(x) = \int_0^\tau \frac{x \exp(-\frac{x^2}{4(\tau-\sigma)})}{2\sqrt{\pi}(\tau-\sigma)^{\frac{3}{2}}} f(\sigma)\sqrt{\sigma}\, \mathrm{d}\sigma \quad \big(f \in L^2[0,\tau],\ x \in (0,\pi)\big). \tag{5.6}$$

We are now in a position to state the main result in [1].

Theorem 5.1. *For every $\tau > 0$, the operator P_τ defined in (5.6) is an isometry from $L^2[0,\tau]$ onto $A^2(\Delta, \omega_{0,\tau})$, where Δ has been defined in (5.5) and*

$$\omega_{0,\delta}(s) = \frac{\exp(\frac{\mathrm{Re}(s^2)}{2\delta})}{\delta} \quad (\delta > 0,\ s \in \Delta). \tag{5.7}$$

The proof of Theorem 5.1 is a very nice application of the theory of linear operators in reproducing kernel spaces, as described, for instance, in [34]. More precisely, the main steps of the proof from [1] are:

- remarking that, by elementary calculus, if in the definition (5.3) of Φ_τ^{left} we replace $x \in (0,\pi)$ by $s \in \Delta$ then the right-hand side of (5.3) defines a function which is holomorphic on Δ;
- using general results on the range of integral operators on RKHS and appropriate integrations, deduce that Ran P_τ is an RKHS of holomorphic functions on Δ whose kernel is

$$K_\tau(s,\overline{w}) = \exp\left(-\frac{s^2+\overline{w}^2}{4\tau}\right)\frac{4s\overline{w}}{\pi(s^2+\overline{w}^2)^2}; \tag{5.8}$$

- finally, remarking that the kernel K_τ in (5.8) coincides with the reproducing kernel of $A^2(\Delta, \omega_{0,\tau})$.

6. THE HEAT EQUATION ON AN INTERVAL

In this section we come back to equations in (1.1), already briefly discussed in Section 1. More precisely, we describe below very recent advances which have lead to several equivalent characterizations of the system described by the first two equations in (1.1). In other words, our aim is to characterize the range of the operator Φ_τ^{heat} introduced in (1.2). We continue using the notation introduced in Section 5, namely for Bergman spaces (possibly weighted).

We first state the following remarkably simple characterization, proved in [14] and confirming a conjecture formulated in [13]:

Theorem 6.1. *Let $\tau > 0$ and let Φ_τ^{heat} be the input map introduced in (1.2). Then*

$$\operatorname{Ran} \Phi_\tau^{\text{heat}} = A^2(D), \tag{6.1}$$

where D is the square introduced in (1.3).

It is quite natural to postpone the discussion of the main steps of the proof of Theorem 6.1 to the end of this section. Indeed, this proof is based, in particular, on another characterization of $\operatorname{Ran} \Phi_\tau^{\text{heat}}$, as a sum of two Bergman spaces on two symmetric infinite sectors. Besides being used in the proof of Theorem 6.1, this type of characterization is of independent interest.

To state these results, we need some notation. We first introduce the set

$$\tilde{\Delta} = \pi - \Delta, \tag{6.2}$$

where Δ has been defined in (5.5), and the weight function

$$\omega_{\pi,\delta}(\tilde{s}) = \frac{\exp(\frac{\operatorname{Re}[(\pi-\tilde{s})^2]}{2\delta})}{\delta} \quad (\delta > 0,\ \tilde{s} \in \tilde{\Delta}). \tag{6.3}$$

Note that

$$\omega_{0,\delta}(s) = \omega_{\pi,\delta}(\pi - s) \quad (s \in \Delta), \tag{6.4}$$

where $\omega_{0,\delta}$ is the weight introduced in (5.7).

We also introduce the space X_δ defined for every $\delta > 0$ by

$$X_\delta = \left\{ \psi \in C(0,\pi) \;\middle|\; \begin{array}{l} \exists \varphi_0 \in A^2(\Delta, \omega_{0,\delta}) \\ \exists \varphi_\pi \in A^2(\tilde{\Delta}, \omega_{\pi,\delta}) \end{array},\ \psi = \varphi_0 + \varphi_\pi \text{ on } (0,\pi) \right\}, \tag{6.5}$$

which is endowed with the norm

$$\|\varphi\|_\delta = \inf \left\{ \|\varphi_0\|_{A^2(\Delta,\omega_{0,\delta})} + \|\varphi_\pi\|_{A^2(\tilde{\Delta},\omega_{\pi,\delta})} \;\middle|\; \begin{array}{c} \varphi_0 + \varphi_\pi = \varphi \\ \varphi_0 \in A^2(\Delta, \omega_{0,\delta}) \\ \varphi_\pi \in A^2(\tilde{\Delta}, \omega_{\pi,\delta}) \end{array} \right\}. \tag{6.6}$$

We are now in a position to formulate the main result in [21].

Theorem 6.2. *With the above notation, for every $\tau, \delta > 0$, we have*

$$\operatorname{Ran} \Phi_\tau^{\text{heat}} = X_\delta = A^2(\Delta) + A^2(\tilde{\Delta}). \tag{6.7}$$

Let us mention that the equality $\operatorname{Ran} \Phi_\tau^{\text{heat}} = A^2(\Delta) + A^2(\tilde{\Delta})$ has been obtained independently in [29].

We briefly describe below the main steps of the proof of Theorem 6.2.

- We first remark that Φ_τ^{heat} is the sum of a series of integral operators involving the heat kernel. More precisely, we have

$$\left(\Phi_\tau \begin{bmatrix} u_0 \\ u_\pi \end{bmatrix}\right)(x) = \int_0^\tau \frac{\partial K_0}{\partial x}(\tau - \sigma, x) u_0(\sigma)\, d\sigma + \int_0^\tau \frac{\partial K_\pi}{\partial x}(\tau - \sigma, x) u_\pi(\sigma)\, d\sigma$$
$$\big(\tau > 0,\ u_0, u_\pi \in L^2[0,\tau],\ x \in (0,\pi)\big), \tag{6.8}$$

where

$$K_0(\sigma, x) = -\sqrt{\frac{1}{\pi\sigma}} \sum_{m\in\mathbb{Z}} \exp\left(-\frac{(x+2m\pi)^2}{4\sigma}\right) \quad \big(\sigma > 0,\ x \in [0,\pi]\big), \tag{6.9}$$

$$K_\pi(\sigma, x) = K_\pi(\sigma, \pi - x) \quad \big(\sigma > 0, x \in [0,\pi]\big). \tag{6.10}$$

Formula (6.8) can be derived, using symmetry considerations, from (5.3). An alternative proof is proposed in **[13]** by combining the Fourier series expression of the solution of (1.1) and the Poisson summation formula.

- The second step consists in remarking that Ran Φ_τ^{heat} coincides with the range of the map (still defined on $(L^2[0,\tau])^2$)

$$\begin{bmatrix} f \\ g \end{bmatrix} \mapsto \Phi_\tau^{\text{heat}} \begin{bmatrix} \sqrt{t}\, f \\ \sqrt{t}\, g \end{bmatrix} \quad \big(f, g \in L^2[0,\tau]\big).$$

This can be easily proved using the fact that the considered system is null-controllable in any positive time, see **[21, PROPOSITION 3.2]**.

- For the third step, we first prove that from (6.8) it follows that for every $f, g \in L^2[0,\tau]$, we have

$$\Phi_\tau^{\text{heat}} \begin{bmatrix} \sqrt{t}\, f \\ \sqrt{t}\, g \end{bmatrix} = P_\tau f + Q_\tau g + R_\tau \begin{bmatrix} f \\ g \end{bmatrix},$$

where P_τ has been defined in (5.6),

$$(Q_\tau g)(x) = (P_\tau g)(\pi - x) \quad \big(x \in (0,\pi)\big),$$

and R_τ is an operator whose norm tends to zero when $\tau \to 0+$. In other words, Φ_τ^{heat} decomposes into the sum of the input maps of the system describing the boundary controlled heat equation on $[0,\infty)$ and $(-\infty,\pi]$, respectively (for which the ranges are known from the previous section) and a remainder term R_τ which becomes "negligible" for small τ.
Combined with Theorem 5.1, this fact implies, recalling (6.5), that

$$\text{Ran}\, \Phi_\tau^{\text{heat}} = X_\tau \quad (\tau > 0).$$

- The last step of the proof consists in showing that

$$X_\tau = A^2(\Delta) + A^2(\tilde{\Delta}) \quad (\tau > 0).$$

This can be accomplished by combining Proposition 2.5 and the construction of appropriate multipliers (see **[21]** for details).

We end this section by coming back to the proof of Theorem 6.1. In view of Theorem 6.2, the conclusion of Theorem 6.1 is equivalent to the equality

$$A^2(D) = A^2(\Delta) + A^2(\tilde{\Delta}). \tag{6.11}$$

This a question which is part of a class of problems with a quite long history in complex analysis: the *separation of singularities* for holomorphic functions. A general formulation of this type of problems is: denoting by $\mathrm{Hol}(\mathcal{O})$ the space of holomorphic functions on an open set $\mathcal{O} \subset \mathbb{C}$ (in particular, the Banach space of analytic functions) and given two open sets $\Omega_1, \Omega_2 \subset \mathbb{C}$ with $\Omega_1 \cap \Omega_2 \neq \emptyset$, is it true that $\mathrm{Hol}(\Omega_1 \cap \Omega_2) = \mathrm{Hol}(\Omega_1) + \mathrm{Hol}(\Omega_2)$? We refer to [14] for detailed historical information on this issue, mentioning here just that [14] is the first work considering this question in a Bergman space context. Moreover, using a methodology involving sophisticated analytical techniques, like Hörmander-type L^p-estimates for the solution of the $\bar{\partial}$ equation, the main results in [14] assert that the separation of singularities for Bergman spaces holds in a geometrical context more general than that in (6.11).

7. CONCLUSIONS, REMARKS, AND OPEN QUESTIONS

This work gives an overview, far from being exhaustive, of the applications of complex and harmonic analysis methods in the study of the reachable space of infinite-dimensional systems. In most of the presented results, the analytical tools appearing in the previous sections have been developed for purposes having a priori nothing to do with the infinite-dimensional system theory. This is the case, for instance, for the Ingham–Beurling–Kahane-type inequalities appearing in Section 3, which began to be applied in controllability and reachability questions only several decades after their publication. The situation is similar for the methods coming from the theories of RKHS and spaces of analytic functions, namely those described in Sections 5 and 6: their penetration in the control-theoretic community took place 20 years after their first publication. An important fact is that these interactions raised new problems and allowed significant progress in the concerned fields of analysis. The separation of singularities for Bergman spaces, briefly discussed in Section 6, is a remarkable example illustrating these mutual interactions.

We conclude this work by briefly describing some open questions which are, at least in the author's opinion, of major interest in the infinite-dimensional system theory.

7.1. Time reversible systems

We think here of linear control systems described by the wave, Schrödinger, or Euler–Bernoulli equations. As already mentioned, the characterization of the reachable space of these systems is quite well understood in the case of one space dimension, but essentially open in several space dimensions. Taking the example of a system described by the wave equation in a bounded domain in $\mathbb{R}^n$ ($n \geq 2$), we should mention the famous paper by Bardos, Lebeau, and Rauch [4], where it is proved that the exact controllability (in sufficiently large time) holds iff the control support satisfies the so-called *geometric optics* condition. On the other hand, using a duality argument and Holmgren's uniqueness theorem, it is not difficult to see that if the control support is an arbitrary open subset of the boundary, then the system is approximately controllable, again in sufficiently large time. As far as we know, the question of characterizing the reachable space when the control support does not

satisfy the geometric optics condition is essentially open and it seems an extremely challenging one. Some information on these spaces can be found in [18], where the wave equation holds in a rectangular domain, or in [3]. Possible tools for tackling a more general geometry can be found in Lebeau [25], Robbiano [30], or Laurent and Léautaud [23].

7.2. Systems described by parabolic equations

As described in Section 6, the reachable space for systems described by the constant coefficient heat equation on a bounded interval has been recently completely characterized in terms of Hilbert spaces of analytic functions. However, for systems described by variable coefficient parabolic equations, even in one space dimension, many natural questions are still open. We think, in particular, of the sharp identification of the domain of analyticity of the reachable space when all the coefficients of the parabolic equation are entire functions of the space variable, see Laurent and Rosier [24] for several remarkable results in this direction.

Coming back to the system described by the one-dimensional constant coefficient heat equations, it would be important to understand the action of the heat semigroup on the reachable space. In particular, is the semigroup obtained by restricting the heat semigroup to the reachable space strongly continuous on Ran Φ_τ (when endowed with the norm defined in (2.9))? A positive answer to this question would be a good departure point in studying the robustness of the reachable space with respect to various perturbations (linear or nonlinear), in the vein of the corresponding theory for exactly controllable systems.

Finally, let us briefly discuss the state-of-the-art for systems described by the constant coefficient heat equation in several space dimensions. An early result in this direction has been provided in Fernández-Cara and Zuazua [12], where it is shown that a class of functions which are holomorphic in an appropriate infinite strip are in the reachable space. A very recent and important contribution to this question has been recently brought in a work by Strohmaier and Waters [38]. In this work, assuming that the spatial domain is a ball and that the control acts on the whole boundary, the authors provide detailed information on the reachable space, similar to that obtained in [8] for systems described by the one-dimensional heat equation. As far as we know, with the exception of the above-mentioned situation, the study of the reachable space described by boundary-controlled parabolic equation in $\mathbb{R}^n$, with $n \geqslant 2$, is a widely open question.

ACKNOWLEDGMENTS

The author thanks S. Ervedoza for numerous and fruitful discussions on the topics discussed in this paper.

FUNDING

This work was partially supported by the Research Chair ACIDDS of the Excellency Initiative (IdEX) of University of Bordeaux.

REFERENCES

[1] H. Aikawa, N. Hayashi, and S. Saitoh, The Bergman space on a sector and the heat equation. *Complex Var. Theory Appl.* **15** (1990), no. 1, 27–36.

[2] B. Allibert, Analytic controllability of the wave equation over a cylinder. *ESAIM Control Optim. Calc. Var.* **4** (1999), 177–207.

[3] S. A. Avdonin, M. I. Belishev, and S. A. Ivanov, Controllability in a filled domain for the multidimensional wave equation with a singular boundary control. *J. Math. Sci.* **83** (1997), 165–174.

[4] C. Bardos, G. Lebeau, and J. Rauch, Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary. *SIAM J. Control Optim.* **30** (1992), 1024–1065.

[5] A. Beurling, Mittag-Leffler lectures on harmonic analysis. In *The collected works of Arne Beurling*, p. xx+389, Contemp. Mathematicians 2, Birkhäuser Boston Inc., Boston, MA, 1989.

[6] J. R. Cannon, *The one-dimensional heat equation. 23*. Cambridge University Press, 1984.

[7] E. Crawley and E. Anderson, Detailed models for piezoceramic actuation of beams, AIAA conf. Tech. rep., paper 89-1388-CP, 1989.

[8] J. Darde and S. Ervedoza, On the reachable set for the one-dimensional heat equation. *SIAM J. Control Optim.* **56** (2018), no. 3, 1692–1715.

[9] J. Dardé and S. Ervedoza, Backward uniqueness results for some parabolic equations in an infinite rod. *Math. Control Relat. Fields* **9** (2019), no. 4, 673–696.

[10] P. Destuynder, I. Legrain, L. Castel, and N. Richard, Theoretical, numerical and experimental discussion on the use of piezoelectric devices for control–structure interaction. *Eur. J. Mech. A Solids* **11** (1992), no. 2, 181–213.

[11] H. O. Fattorini and D. L. Russell, Exact controllability theorems for linear parabolic equations in one space dimension. *Arch. Ration. Mech. Anal.* **43** (1971), 272–292.

[12] E. Fernández-Cara and E. Zuazua, The cost of approximate controllability for heat equations: the linear case. *Adv. Differential Equations* **5** (2000), no. 4–6, 465–514.

[13] A. Hartmann, K. Kellay, and M. Tucsnak, From the reachable space of the heat equation to Hilbert spaces of holomorphic functions. *J. Eur. Math. Soc. (JEMS)* **22** (2020), no. 10, 3417–3440.

[14] A. Hartmann and M.-A. Orsoni, Separation of singularities for the Bergman space and application to control theory. *J. Math. Pures Appl.* **150** (2021), 181–201.

[15] L. F. Ho and D. L. Russell, Admissible input elements for systems in Hilbert space and a Carleson measure criterion. *SIAM J. Control Optim.* **21** (1983), no. 4, 614–640.

[16] A. E. Ingham, Some trigonometrical inequalities with applications to the theory of series. *Math. Z.* **41** (1936), no. 1, 367–379.

[17] S. Jaffard, Contrôle interne exact des vibrations d'une plaque rectangulaire (internal exact control for the vibrations of a rectangular plate). *Port. Math.* **47** (1990), no. 4, 423–429.

[18] S. Jaffard and S. Micu, Estimates of the constants in generalized Ingham's inequality and applications to the control of the wave equation. *Asymptot. Anal.* **28** (2001), no. 3–4, 181–214.

[19] J.-P. Kahane, Pseudo-périodicité et séries de Fourier lacunaires. *Ann. Sci. Éc. Norm. Supér. (3)* **79** (1962), 93–150.

[20] R. E. Kalman, Mathematical description of linear dynamical systems. *J. Soc. Indust. Appl. Math. Ser. A Control* **1** (1963), no. 2, 152–192.

[21] K. Kellay, T. Normand, and M. Tucsnak, Sharp reachability results for the heat equation in one space dimension. *Anal. PDE* (to appear) (2022). Hal preprint: hal-02302165.

[22] V. Komornik and P. Loreti, *Fourier series in control theory*. Monogr. Math., Springer, New York, 2005.

[23] C. Laurent and M. Léautaud, Quantitative unique continuation for operators with partially analytic coefficients. application to approximate control for waves. *J. Eur. Math. Soc. (JEMS)* **21** (2018), no. 4, 957–1069.

[24] C. Laurent and L. Rosier, Exact controllability of nonlinear heat equations in spaces of analytic functions. 2018, arXiv:1812.06637.

[25] G. Lebeau, Control for hyperbolic equations. *J. Équ. Dériv. Partielles* (1992), 1–24.

[26] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome 1*. Rech. Math. Appl. 8, Masson, Paris, 1988.

[27] P. Martin, L. Rosier, and P. Rouchon, On the reachable states for the boundary control of the heat equation. *Appl. Math. Res. Express. AMRX* (2016), no. 2, 181–216.

[28] S. Micu and E. Zuazua, On the lack of null-controllability of the heat equation on the half-line. *Trans. Amer. Math. Soc.* **353** (2001), no. 4, 1635–1659.

[29] M.-A. Orsoni, Reachable states and holomorphic function spaces for the 1-D heat equation. *J. Funct. Anal.* **280** (2021), no. 7, 108852.

[30] L. Robbiano, Fonction de coût et contrôle des solutions des équations hyperboliques. *Asymptot. Anal.* **10** (1995), no. 2, 95–115.

[31] D. L. Russell, A unified boundary controllability theory for hyperbolic and parabolic partial differential equations. *Stud. Appl. Math.* **52** (1973), 189–211.

[32] S. Saitoh, Isometrical identities and inverse formulas in the one-dimensional heat equation. *Appl. Anal.* **40** (1991), no. 2–3, 139–149.

[33] S. Saitoh, *Integral transforms, reproducing kernels and their applications*. Pitman Res. Notes Math. Ser. 369, Longman, Harlow, 1997.

[34] S. Saitoh and Y. Sawano, *Theory of reproducing kernels and applications*. Springer, 2016.

[35] D. Salamon, Infinite-dimensional linear systems with unbounded control and observation: a functional analytic approach. *Trans. Amer. Math. Soc.* **300** (1987), no. 2, 383–431.

[36] E. J. P. G. Schmidt, Even more states reachable by boundary control for the heat equation. *SIAM J. Control Optim.* **24** (1986), no. 6, 1319–1322.

[37] T. I. Seidman, Time-invariance of the reachable set for linear control problems. *J. Math. Anal. Appl.* **72** (1979), no. 1, 17–20.

[38] A. Strohmaier and A. Waters, Analytic properties of heat equation solutions and reachable sets. 2020, arXiv:2006.05762.

[39] G. Tenenbaum and M. Tucsnak, Fast and strongly localized observation for the Schrödinger equation. *Trans. Amer. Math. Soc.* **361** (2009), no. 2, 951–977.

[40] M. Tucsnak, Regularity and exact controllability for a beam with piezoelectric actuator. *SIAM J. Control Optim.* **34** (1996), no. 3, 922–930.

[41] M. Tucsnak and G. Weiss, *Observation and control for operator semigroups*. Birkhäuser Adv. Texts. Basl. Lehrb. [Birkhäuser Advanced Texts: Basel Textbooks], Birkhäuser, Basel, 2009.

[42] G. Valiron, *Théorie des fonctions*. Masson, Paris, 1966.

[43] G. Weiss, Admissibility of input elements for diagonal semigroups on l^2. *Systems Control Lett.* **10** (1988), no. 1, 79–82.

[44] G. Weiss, Admissibility of unbounded control operators. *SIAM J. Control Optim.* **27** (1989), no. 3, 527–545.

MARIUS TUCSNAK

Institut de Mathématiques de Bordeaux, Université de Bordeaux, 351, Cours de la Libération – F 33 405 Talence, France, marius.tucsnak@u-bordeaux.fr

17. STATISTICS AND DATA ANALYSIS

GRADIENT DESCENT ON INFINITELY WIDE NEURAL NETWORKS: GLOBAL CONVERGENCE AND GENERALIZATION

FRANCIS BACH AND LÉNAÏC CHIZAT

ABSTRACT

Many supervised machine learning methods are naturally cast as optimization problems. For prediction models which are linear in their parameters, this often leads to convex problems for which many mathematical guarantees exist. Models which are nonlinear in their parameters such as neural networks lead to nonconvex optimization problems for which guarantees are harder to obtain. In this paper, we consider two-layer neural networks with homogeneous activation functions where the number of hidden neurons tends to infinity, and show how qualitative convergence guarantees may be derived.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68T07; Secondary 49Q22, 90C30, 62G08

KEYWORDS

Machine learning, neural networks, gradient descent, gradient flow, optimal transport

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5398–5419
DOI 10.4171/ICM2022/121

1. INTRODUCTION

In the past 20 years, data in all their forms have played an increasing role: in personal lives, with various forms of multimedia and social networks, in the economic sector where most industries monitor all of their processes and aim at making data-driven decisions, and in sciences, where data-based research is having more and more impact, both in fields which are traditionally data-driven such as medicine and biology, but also in humanities.

This proliferation of data leads to a need for automatic processing, with striking recent progress in some perception tasks where humans excel, such as image recognition or natural language processing. These advances in artificial intelligence were fueled by the combination of three factors: (1) massive data to learn from, such as millions of labeled images, (2) increased computing resources to treat this data, and (3) continued scientific progress in algorithms.

Machine learning is one of the scientific disciplines that have made this progress possible, by blending statistics and optimization to design algorithms with theoretical generalization guarantees. The goal of this paper is to highlight our recent progress and to present a few open mathematical problems.

2. SUPERVISED LEARNING

In this paper, we will focus on the supervised machine learning problem, where we are being given n pairs of observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, for example, images ($\mathcal{X}$ is then the set of all possible images), with a set of labels ($\mathcal{Y}$ is then a finite set, which we will assume to be a subset of $\mathbb{R}$ for simplicity). The goal is to be able to predict a new output $y \in \mathcal{Y}$, given a previously unobserved input $x \in \mathcal{X}$.

Following the traditional statistical *M-estimation* framework [45], this can be performed by considering prediction functions $x \mapsto h(x, \theta) \in \mathbb{R}$, parameterized by $\theta \in \mathbb{R}^d$. The vector θ is then estimated through regularized empirical risk minimization, that is, by solving

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, h(x_i, \theta)\big) + \lambda \Omega(\theta), \tag{2.1}$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is a loss function, and $\Omega : \mathbb{R}^d \to \mathbb{R}$ is a regularization term that avoids overfitting (that is, learning a carbon copy of the observed data that does not generalize well to unseen data).

Typical loss functions are the square loss $\ell(y_i, h(x_i, \theta)) = \frac{1}{2}(y_i - h(x_i, \theta))^2$ for regression problems, and the logistic loss $\ell(y_i, h(x_i, \theta)) = \log(1 + \exp(-y_i h(x_i, \theta)))$ for binary classification where $\mathcal{Y} = \{-1, 1\}$. In this paper, we will always assume that the loss function is continuously twice differentiable and convex with respect to the second variable. This applies to a wide variety of output spaces beyond regression and binary classification (see [36] and references therein).

When the predictor depends linearly on the parameters, typical regularizers are the squared Euclidean norm $\Omega(\theta) = \frac{1}{2}\|\theta\|_2^2$ or the ℓ_1-norm $\Omega(\theta) = \|\theta\|_1$, which both lead to

improved generalization performance, with the ℓ_1-norm providing additional variable selection benefits [14].

2.1. Statistics and optimization

The optimization problem in equation (2.1) leads naturally to two sets of questions, which are often treated separately. Given that some minimizer $\hat{\theta}$ is obtained (no matter how), how does the corresponding prediction function generalize to unseen data? This is a statistical question that requires assumptions on the link between the observed data (usually called the "training data"), and the unseen data (usually called the "testing data"). It is typical to assume that the training and testing data are sampled independently and identically from the same fixed distribution. Then a series of theoretical guarantees applies, based on various probabilistic concentration inequalities (see, e.g., [31]).

The second question is how to obtain an approximate minimizer $\hat{\theta}$, which is an optimization problem, regardless on the relevance of $\hat{\theta}$ on unseen data (see, e.g., [6]). For high-dimensional problems where d is large (up to millions or billions), classical gradient-based algorithms are preferred because of their simplicity, efficiency, robustness, and favorable convergence properties. The most classical one is gradient descent, which is an iterative algorithm with iteration

$$\theta_k = \theta_{k-1} - \gamma \nabla \mathcal{R}(\theta_{k-1}),$$

where $\mathcal{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, h(x_i, \theta)) + \lambda\Omega(\theta)$ is the objective function in equation (2.1), and $\gamma > 0$ the step-size.

In this paper, where we aim at tackling high-dimensional problems, we will often consider the two problems of optimization and statistical estimation jointly.

2.2. Linear predictors and convex optimization

In many applications, a prediction function which is linear in the parameter θ is sufficient for good predictive performance, that is, we can write

$$h(x, \theta) = \theta^\top \Phi(x)$$

for some function $\Phi : \mathcal{X} \to \mathbb{R}^d$, which is often called a "feature function." For simplicity, we have assumed finite-dimensional features, but infinite-dimensional features can also be considered, with a specific computational argument to allow finite-dimensional computations through reproducing kernel Hilbert spaces (see, e.g., [40] and references therein).

Given a convex loss function, the optimization problem is convex and gradient descent on the objective function, together with its stochastic extensions, has led to a number of efficient algorithms with strong generalization guarantees of convergence towards the *global* optimum of the objective function [6]. For example, for the square or logistic loss, if the feature function is bounded in ℓ_2-norm by R for all observations, and for the squared Euclidean norm $\Omega(\theta) = \frac{1}{2}\|\theta\|_2^2$, bounds on the number of iterations to reach a certain precision ε (difference between the candidate function value $\mathcal{R}(\theta)$ and the minimal value) can be obtained:

- For gradient descent, $\frac{R^2}{\lambda}\log\frac{1}{\varepsilon}$ iterations are needed, but each iteration has a running time complexity of $O(nd)$, because the d-dimensional gradients of the n functions $\theta \mapsto \ell(y_i, h(x_i, \theta))$, $i = 1, \dots, n$, are needed.
- For stochastic gradient descent, with iteration $\theta_k = \theta_{k-1} - \gamma\nabla\ell(y_{i(k)}, h(x_{i(k)}, \theta_{k-1}))$, with $i(k) \in \{1, \dots, n\}$ taken uniformly at random, the number of iterations is at most $\frac{R^2}{\lambda}\frac{1}{\varepsilon}$. We lose the logarithmic dependence, but each iteration has complexity $O(d)$, which can be a substantial gain when n is large.
- More recent algorithms based on variance reduction can achieve an overall complexity proportional to $(n + \frac{R^2}{\lambda})\log\frac{1}{\varepsilon}$, thus with an exponential convergence rate at low iteration cost (see [16] and references therein).

In summary, for linear models, algorithms come with strong performance guarantees that reasonably match their empirical behavior. As shown below, nonlinear models exhibit more difficulties.

2.3. Neural networks and nonconvex optimization

In many other application areas, in particular in multimedia processing, linear predictors have been superseded by nonlinear predictors, with neural networks being the most classical example (see [15]). A vanilla neural network is a prediction function of the form

$$h(x, \theta) = \theta_s^\top \sigma(\theta_{s-1}^\top \sigma\big(\cdots \theta_2^\top \sigma(\theta_1^\top x)\big),$$

where the function $\sigma : \mathbb{R} \to \mathbb{R}$ is taken component-wise, with the classical examples being the sigmoid function $\sigma(t) = (1 + \exp(-t))^{-1}$ and the "rectified linear unit" (ReLU), $\sigma(t) = t_+ = \max\{t, 0\}$. The matrices $\theta_1, \dots, \theta_s$ are called weight matrices. The simplest nonlinear predictor is for $s = 2$, and will be the main subject of study in this paper. See Figure 1 for an illustration.

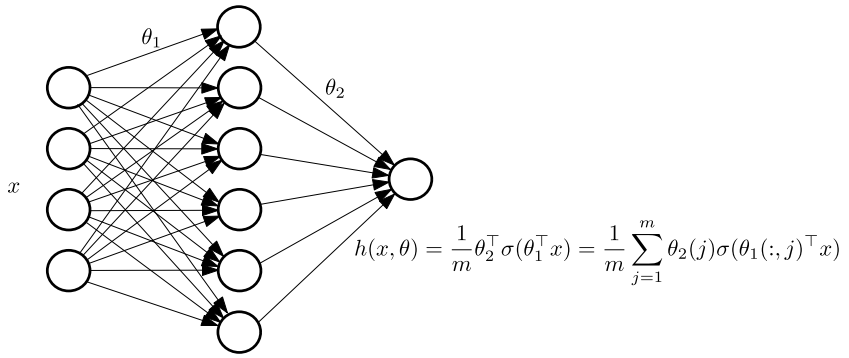


FIGURE 1
Neural network with a single hidden layer, with an input weight matrix $\theta_1 \in \mathbb{R}^{d\times m}$ and a output weight vector $\theta_2 \in \mathbb{R}^m$.

The main difficulty is that now the optimization problem in equation (2.1) is not convex anymore, and gradient descent can converge to stationary points that are not global minima. Theoretical guarantees can be obtained regarding the decay of the norm of the gradient of the objective function, or convergence to a local minimizer may be ensured [21, 25], but this does not exclude bad local minima, and global quantitative convergence guarantees can only be obtained with exponential dependence in dimension for the class of (potentially nonconvex) functions of a given regularity [33].

An extra difficulty is related to the number of hidden neurons, also referred to as the *width* of the network (equal to the size of θ_2 when $s = 2$), which is often very large in practice, which poses both statistical and optimization issues. We will see that this is precisely this overparameterization that allows obtaining qualitative global convergence guarantees.

3. MEAN FIELD LIMIT OF OVERPARAMETERIZED ONE-HIDDEN LAYER NEURAL NETWORKS

We now tackle the study of neural networks with one infinitely wide hidden layer. They are also referred to as (wide) two-layer neural networks, because they have two layers of weights. We first rescale the prediction function by $1/m$ (which can be obtained by rescaling θ_2 by $1/m$), and express it explicitly as an empirical average, namely

$$h(x, \theta) = \frac{1}{m}\theta_2^\top \sigma(x^\top \theta_1) = \frac{1}{m}\sum_{j=1}^{m} \theta_2(j) \cdot \sigma\big[x^\top \theta_1(\cdot, j)\big],$$

where $\theta_2(j) \in \mathbb{R}$ is the output weight associated to neuron j, and $\theta_1(\cdot, j) \in \mathbb{R}^d$ the corresponding vector of input weights. The key observation is that the prediction function $x \mapsto h(x, \theta)$ is the average of m prediction functions $x \mapsto \theta_2(j) \cdot \sigma[\theta_1(\cdot, j)^\top x]$, for $j = 1, \dots, m$, with *no sharing of the parameters* (which is not true if extra layers of hidden neurons are added).

In order to highlight this parameter separability, we define

$$w_j = \big[\theta_2(j), \theta_1(\cdot, j)\big] \in \mathbb{R}^{d+1},$$

the set of weights associated to the hidden neuron j, and consider

$$\Psi(w) : x \mapsto w(1) \cdot \sigma\big[x^\top w(2, \dots, d+1)\big],$$

so that the prediction function $x \mapsto h(\cdot, w_1, \dots, w_m)$, parameterized by $w_1, \dots, w_m$, is now

$$h(\cdot, w_1, \dots, w_m) = \frac{1}{m}\sum_{j=1}^{m} \Psi(w_j). \tag{3.1}$$

The empirical risk is of the form

$$R(h) = \mathbb{E}\big[\ell\big(y, h(x)\big)\big],$$

which is convex in h for convex loss functions (even for neural networks), but typically nonconvex in w. Note that the resulting problem of minimizing a convex function $R(h)$ for $h = \frac{1}{m}\sum_{j=1}^{m} \Psi(w_j)$ applies beyond neural networks, for example, for sparse deconvolution [7].

3.1. Reformulation with probability measures

We now define by $\mathcal{P}(\mathcal{W})$ the set of probability measures on $\mathcal{W} = \mathbb{R}^{d+1}$. We can rewrite equation (3.1) as

$$h = \int_{\mathcal{W}} \Psi(w) d\mu(w),$$

with $\mu = \frac{1}{m}\sum_{j=1}^{m} \delta_{w_j}$ being the average of Dirac measures at each $w_1, \dots, w_m$. Following a physics analogy, we will refer to each w_j as a *particle*. When the number m of particles grows, the empirical measure $\frac{1}{m}\sum_{j=1}^{m} \delta_{w_j}$ may converge in distribution to a probability measure with a density, often referred to as a *mean field* limit. Our main reformulation will thus be to consider an optimization problem over probability measures.

The optimization problem we are faced with is equivalent to

$$\inf_{\mu \in \mathcal{P}(\mathcal{W})} R\left(\int_{\mathcal{W}} \Psi(w) d\mu(w) \right), \tag{3.2}$$

with the constraint that μ is an average of m Dirac measures. In this paper, following a long line of work in statistics and signal processing [5, 23], we consider the optimization problem *without this constraint*, and relate optimization algorithms for finite but large m (thus acting on $W = (w_1, \dots, w_m)$ in $\mathcal{W}^m$) to a well-defined algorithm in $\mathcal{P}(\mathcal{W})$.

Note that we now have a convex optimization problem, with a convex objective in μ over a convex set (all probability measures). However, it is still an infinite-dimensional space that requires dedicated finite-dimensional algorithms. In this paper we focus on gradient descent on w, which corresponds to standard practice in neural networks (e.g., back-propagation). For algorithms based on classical convex optimization algorithms such as the Frank–Wolfe algorithm, see [4].

3.2. From gradient descent to gradient flow

Our general goal is to study the gradient descent recursion on $W = (w_1, \dots, w_m) \in \mathcal{W}^m$, defined as

$$W_k = W_{k-1} - \gamma m \nabla G(W_{k-1}), \tag{3.3}$$

with

$$G(W) = R\big(h(\cdot, w_1, \dots, w_m)\big) = R\left(\frac{1}{m} \sum_{j=1}^{m} \Psi(w_j) \right).$$

In the context of neural networks, this is exactly the back-propagation algorithm. We include the factor m in the step-size to obtain a well-defined limit when m tends to infinity (see Section 3.3).

For convenience in the analysis, we look at the limit when the step-size γ goes to zero. If we consider a function $V : \mathbb{R} \to \mathcal{W}^m$, with values $V(k\gamma) = W_k$ at $t = k\gamma$, and we interpolate linearly between these points, then we obtain exactly the standard Euler discretization of the ordinary differential equation (ODE) [44],

$$\dot{V} = -m \nabla G(V). \tag{3.4}$$

This gradient flow will be our main focus in this paper. As highlighted above, and with extra regularity assumptions, it is the limit of the gradient recursion in equation (3.3) for vanishing step-sizes γ. Moreover, under appropriate conditions, stochastic gradient descent, where we only observe an unbiased noisy version of the gradient, also leads in the limit $\gamma \to 0$ to the same ODE [24]. This allows applying our results to probability distributions of the data (x, y) which are not the observed empirical distribution, but the unseen test distribution, where the stochastic gradients come from the gradient of the loss from a single observation.

Three questions now emerge:

(1) What is the limit (if any) of the gradient flow in equation (3.4) when the number of particles m gets large?

(2) Where can the gradient flow converge to?

(3) Can we ensure a good generalization performance when the number of parameters grows unbounded?

In this paper, we will focus primarily in the next sections on the first two questions, and tackle the third question in Section 5.

3.3. Wasserstein gradient flow

Above, we have described a general framework where we want to minimize a function F defined on probability measures,

$$F(\mu) = R\bigg(\int_{\mathcal{W}} \Psi(w) d\mu(w) \bigg), \tag{3.5}$$

with an algorithm minimizing $G(w_1, \dots, w_m) = R(\frac{1}{m} \sum_{j=1}^m \Psi(w_j))$ through the gradient flow $\dot{W} = -m \nabla G(W)$, with $W = (w_1, \dots, w_m)$.

As shown in a series of works concerned with the infinite-width limit of two-layer neural networks [8, 30, 35, 38, 41], this converges to a well-defined mathematical object called a Wasserstein gradient flow [2]. This is a gradient flow derived from the Wasserstein metric on the set of probability measures, which is defined as [39]

$$W_2(\mu, \nu)^2 = \inf_{\gamma \in \Pi(\mu, \nu)} \int \|v - w\|_2^2 d\gamma(v, w),$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathcal{W} \times \mathcal{W}$ with marginals μ and ν. In a nutshell, the gradient flow is defined as the limit when γ tends to zero of the extension of the following discrete time dynamics:

$$\mu(t + \gamma) = \inf_{\nu \in \mathcal{P}(\mathcal{W})} F(\nu) + \frac{1}{2\gamma} W_2\big(\mu(t), \nu\big)^2.$$

When applying such a definition in a Euclidean space with the Euclidean metric, we recover the usual gradient flow $\dot{\mu} = -\nabla F(\mu)$, but here with the Wasserstein metric, this defines a specific flow on the set of measures. When the initial measure is a weighted sum of Diracs, this is exactly asymptotically (when $\gamma \to 0$) equivalent to backpropagation. When initialized

with an arbitrary probability measure, we obtain a partial differential equation (PDE), satisfied in the sense of distributions. Moreover, when the sum of Diracs converges in distribution to some measure, the flow converges to the solution of the PDE. More precisely, assuming $\Psi : \mathbb{R}^{d+1} \to \mathcal{F}$, where $\mathcal{F}$ is a Hilbert space, and $\nabla R(h) \in \mathcal{F}$ the gradient of R, we consider the *mean potential*

$$J(w|\mu) = \left\langle \Psi(w), \nabla R\left(\int_{\mathcal{W}} \Psi(v) d\mu(v) \right) \right\rangle. \tag{3.6}$$

The PDE is then the classical continuity equation

$$\partial_t \mu_t(w) = \operatorname{div}\big(\mu_t(w) \nabla J(w|\mu_t)\big), \tag{3.7}$$

which is understood in the sense of distributions. The following result formalizes this behavior (see [8] for details and a more general statement).

Theorem 1. *Assume that $R : \mathcal{F} \to [0, +\infty[$ and $\Psi : \mathcal{W} = \mathbb{R}^{d+1} \to \mathcal{F}$ are (Fréchet) differentiable with Lipschitz differentials, and that R is Lipschitz on its sublevel sets. Consider a sequence of initial weights $(w_j(0))_{j \geq 1}$ contained in a compact subset of $\mathcal{W}$ and let $\mu_{t,m} := \frac{1}{m} \sum_{j=1}^{m} w_j(t)$ where $(w_1(t), \dots, w_m(t))$ solves the ODE (3.4). If $\mu_{0,m}$ weakly converges to some $\mu_0 \in \mathcal{P}(\mathcal{W})$ then $\mu_{t,m}$ weakly converges to μ_t where $(\mu_t)_{t \geq 0}$ is the unique weakly continuous solution to (3.7) initialized with μ_0.*

In the following section, we will study the solution of this PDE (i.e., the Wasserstein gradient flow), interpreting it as the limit of the gradient flow in equation (3.4), when the number of particles m tends to infinity.

4. GLOBAL CONVERGENCE

We consider the Wasserstein gradient flow defined above, which leads to the PDE in equation (3.7). Our goal is to understand when we can expect that when $t \to \infty$, μ_t converges to a global minimum of F defined in equation (3.5). Obtaining a global convergence result is not out of the question because F is a convex functional defined on the convex set of probability measures. However, it is nontrivial because with our choice of the Wasserstein geometry on measures, which allows an approximation through particles, the flow has some stationary points which are not the global optimum (see the examples in Section 4.4).

We start with an informal general result without technical assumptions before stating a formal simplified result.

4.1. Informal result

In order to avoid too many technicalities, we first consider an informal theorem in this paper and refer to [8] for a detailed set of technical assumptions (in particular smoothness assumptions). This leads to the informal theorem:

Theorem 2 (Informal). *If the support of the initial distribution includes all directions in $\mathbb{R}^{d+1}$, and if the function Ψ is positively 2-homogeneous then, if the Wasserstein gradient flow weakly converges to a distribution, it can only be to a global optimum of F.*

In [8] another version of this result that allows for *partial* homogeneity (e.g., with respect to a subset of variables) of degree 1 is proven, at the cost of a more technical assumption on the initialization. For neural networks, we have $\Psi(w_j)(x) = m\theta_2(j) \cdot \sigma[\theta_1(\cdot, j)^\top x]$, and this more general version applies. For the classical ReLU activation function $u \mapsto \max\{0, u\}$, we get a positively 2-homogeneous function, as required in the previous statement. A simple way to spread all directions is to initialize neural network weights from Gaussian distributions, which is standard in applications [15].

From qualitative to quantitative results? Our result states that for infinitely many particles, we can only converge to a global optimum (note that we cannot show that the flow always converges). However, it is only a qualitative result in comparison with what is known for convex optimization problems in Section 2.2:

- This is only for $m = +\infty$, and we cannot provide an estimation of the number of particles needed to approximate the mean field regime that is not exponential in t (see such results, e.g., in [28]).
- We cannot provide an estimate of the performance as the function of time, that would provide an upper bound on the running time complexity.

Moreover, our result does not apply beyond a single hidden layer, and understanding the nonlinear infinite-width limits for deeper networks is an important research area [3, 12, 13, 19, 34, 42, 48].

From informal to formal results. Beyond the lack of quantitative guarantees, obtaining a formal result requires regularity and compactness assumptions which are not satisfied for the classical ReLU activation function $u \mapsto \max\{0, u\}$, which is not differentiable at zero (a similar result can be obtained in this case but under stronger assumptions on the data distribution and the initialization [9, 47]). In the next section, we will consider a simplified formal result, with a detailed proof.

4.2. Simplified formal result

In order to state a precise result, we will cast the flow on probability measures on $\mathcal{W} = \mathbb{R}^{d+1}$ to a flow on measures on the unit sphere

$$S^d = \big\{w \in \mathbb{R}^{d+1}, \|w\|_2 = 1\big\}.$$

This is possible when the function Ψ is positively 2-homogeneous on $\mathcal{W} = \mathbb{R}^{d+1}$, that is, such that $\Psi(\lambda w) = \lambda^2 \Psi(w)$ for $\lambda > 0$. We can use homogeneity by reparameterizing each particle w_j in polar coordinates as

$$w_j = r_j \eta_j, \quad \text{with } r_j \in \mathbb{R} \text{ and } \eta_j \in S^d.$$

Using homogenetity, we have a prediction function

$$h = \frac{1}{m} \sum_{j=1}^{m} \Psi(w_j) = \frac{1}{m} \sum_{j=1}^{m} r_j^2 \Psi(\eta_j).$$

Moreover, the function J defined in equation (3.6) is also 2-homogeneous, and its gradient then 1-homogeneous. The flow from equation (3.4), can be written as

$$\dot{w}_j = -\nabla J(w_j|\mu), \quad \text{with } \mu = \frac{1}{m}\sum_{i=1}^{m} \delta_{w_j}.$$

A short calculation shows that the flow

$$\begin{cases} \dot{r}_j = -2r_j J(\eta_j|\nu), \\ \dot{\eta}_j = -\nabla_{\mathcal{S}} J(\eta_j|\nu) = (I - \eta_j \eta_j^\top)\nabla J(\eta_j|\nu), \end{cases} \quad \text{with } \nu = \frac{1}{m}\sum_{i=1}^{m} r_j^2 \delta_{\eta_j}, \tag{4.1}$$

where $\nabla_{\mathcal{S}}$ denotes the gradient of functions defined on the sphere $\mathcal{S}^d$, leads to exactly the same dynamics. Indeed, by homogeneity of Ψ, the two definitions of μ and ν (through the w_j's, or the η_j's and r_j's) lead to the same functions $J(\cdot|\mu)$ and $J(\cdot|\nu)$, and we get

$$\begin{aligned} \dot{w}_j &= \dot{r}_j \eta_j + r_j \dot{\eta}_j = -2r_j J(\eta_j|\nu)\eta_j - r_j(I - \eta_j\eta_j^\top)\nabla J(\eta_j|\nu) \\ &= -r_j \nabla J(\eta_j|\nu) - r_j\big[2J(\eta_j|\nu) - \eta_j^\top \nabla J(\eta_j|\nu)\big]\eta_j \\ &= -\nabla J(w_j|\mu), \end{aligned}$$

because $w \mapsto \nabla J(w|\nu)$ is 1-homogeneous, and we have used the Euler identity for the 2-homogeneous function $w \mapsto J(w|\nu) = J(w|\mu)$.

Moreover, the flow defined in equation (4.1) is such that η_j remains on the sphere $\mathcal{S}^d$. We will study this flow under the assumption that the function Ψ is sufficiently regular, which excludes ReLU neural networks, but makes the proof easier (see more details in [7]).

We first derive a PDE analogous to equation (3.7). We consider a smooth test function $f : \mathcal{S}^d \to \mathbb{R}$, and the quantity

$$a = \int_{\mathcal{S}^d} f(\eta) d\nu(\eta) = \frac{1}{m}\sum_{j=1}^{m} r_j^2 f(\eta_j).$$

We have

$$\begin{aligned} \dot{a} &= \frac{1}{m}\sum_{j=1}^{m} 2r_j \dot{r}_j f(\eta_j) + \frac{1}{m}\sum_{j=1}^{m} r_j^2 \nabla_{\mathcal{S}} f(\eta_j)^\top \dot{\eta}_j \\ &= -\frac{1}{m}\sum_{j=1}^{m} 4r_j^2 J(\eta_j|\nu) f(\eta_j) - \frac{1}{m}\sum_{j=1}^{m} r_j^2 \nabla_{\mathcal{S}} f(\eta_j)^\top \nabla_{\mathcal{S}} J(\eta_j|\nu) \\ &= -4\int_{\mathcal{S}^d} f(\eta) J(\eta|\nu) d\nu(\eta) - \int_{\mathcal{S}^d} \nabla_{\mathcal{S}} f(\eta)^\top \nabla_{\mathcal{S}} J(\eta|\nu) d\nu(\eta). \end{aligned} \tag{4.2}$$

This shows that we have the PDE for the density ν_t at time t

$$\partial_t \nu_t(\eta) = -4\nu_t(\eta) J(\eta|\nu_t) + \operatorname{div}_{\mathcal{S}}\big(\nu_t(\eta)\nabla_{\mathcal{S}} J(\eta|\nu_t)\big) \tag{4.3}$$

satisfied in the sense of distributions (see e.g. [39, PROP. 4.2]). We can now state our main result.

Theorem 3. *Assume the function $\Psi : \mathcal{S}^d \to \mathcal{F}$ is d-times continuously differentiable. Assume ν_0 is a nonnegative measure on the sphere $\mathcal{S}^d$ with finite mass and full support. Then the flow defined in (4.3) is well defined for all $t \geqslant 0$. Moreover, if ν_t converges weakly to some limit ν_∞, then ν_∞ is a global minimum of the function $\nu \mapsto F(\nu) = R(\int_{\mathcal{S}^d} \Psi(\eta) d\nu(\eta))$ over the set of nonnegative measures.*

4.3. Proof of Theorem 3

The global optimality conditions for minimizing the convex functional F is that on the support of ν_∞ then $J(\eta|\nu_\infty) = 0$, while on the entire sphere $J(\eta|\nu_\infty) \geqslant 0$. The proof, adapted from [7], then goes as follows:

- The existence and uniqueness of the flow $(\nu_t)_{t\geq 0}$ can be proved by using the equivalence with a Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ in $\mathcal{P}(\mathbb{R}^{d+1})$ and the theory of Wasserstein gradient flows [2]. As a matter of fact, $(\nu_t)_{t\geq 0}$ is itself a gradient flow for a certain metric between nonnegative measures, that is, in a certain sense, the inf-convolution between the Wasserstein and Hellinger metrics, see the discussion in [7].

- The flow ν_t has a full support at all time t. This can be deduced from the representation of the solutions to equation (4.3) as

$$\nu_t = X(t,\cdot)_\#\left(\nu_0 \exp\left(-4\int_0^t J\big(X(s,\cdot)|\nu_s\big)ds\right)\right),$$

  where $X : [0,+\infty[\times \mathcal{S}^d \to \mathcal{S}^d$ is the flow associated to the time-dependent vector field $-\nabla_{\mathcal{S}} J(\cdot \mid \nu_t)$, i.e., it satisfies $X(0,\eta) = \eta$ and $\frac{d}{dt}X(t,\eta) = -\nabla_{\mathcal{S}} J(X(t,\eta)|\nu_t)$ for all $\eta \in \mathcal{S}^d$, see, e.g., [27]. Under our regularity assumptions, standard stability results for ODEs guarantee that at all time t, $X(t,\cdot)$ is a diffeomorphism of the sphere. Thus ν_t is the image measure (this is what the "sharp" notation stands for) by a diffeomorphism of a measure of the form $\nu_0 \exp(\dots)$ which has full support and thus ν_t has full support.

- We assume that the flow converges to some measure ν_∞ (which could be singular). From equation (4.1), this imposes by stationarity of ν_∞ that $J(\eta|\nu_\infty) = 0$ on the support of ν_∞, but nothing is imposed beyond the support of ν_∞ (and we need nonnegativity of $J(\eta|\nu_\infty)$ for all $\eta \in \mathcal{S}^d$).
  In order to show that $\min_{\eta\in\mathcal{S}^d} J(\eta|\nu_\infty) \geqslant 0$, we assume that it is strictly negative and will get a contradiction. We first need a $v < 0$ such that $v > \min_{\eta\in\mathcal{S}^d} J(\eta|\nu_\infty)$, and the gradient $\nabla_{\mathcal{S}} J(\eta|\nu_\infty)$ does not vanish on the v-level-set $\{\eta \in \mathcal{S}^d, J(\eta \mid \nu_\infty) = v\}$ of $J(\cdot|\nu_\infty)$. Such a v exists because of Morse–Sard lemma which applies because under our assumptions, $J(\cdot|\nu)$ is d-times continuously differentiable for any finite nonnegative measure ν.
  We then consider the set $K = \{\eta \in \mathcal{S}^d, J(\eta \mid \nu_\infty) \leqslant v\}$, which has some boundary ∂K, such that the gradient $\nabla_{\mathcal{S}} J(\eta|\nu_\infty)$ has strictly positive dot-product with an outward normal vector to the level set at $\eta \in \partial K$.
  Since ν_t converges weakly to ν_∞, there exists $t_0 > 0$ such that for all $t \geqslant t_0$, $\sup_{\eta\in K} J(\eta|\nu_t) < v/2$, while on the boundary $\nabla_{\mathcal{S}} J(\eta|\nu_\infty)$ has nonnegative dot-product with an outward normal vector. This means that for all $t > t_0$, applying equation (4.2) to the indicator function of K, if $a_t = \nu_t(K)$,

$$a'(t) \geqslant -4 \sup_{\eta\in K} J(\eta|\nu_t) a(t).$$

By the previous point, $a(t_0) > 0$ and thus, by Grönwall's lemma, $a(t)$ diverges, which is a contradiction with the convergence of ν_t to ν_∞.

4.4. Experiments

In order to illustrate[1] the global convergence result from earlier sections, we consider a supervised learning problem on $\mathbb{R}^2$, with Gaussian input data x, and output data given by a "teacher" neural network

$$y = \sum_{j=1}^{m_0} \theta_2(j) \max\{\theta_1(:, j)^\top x, 0\}$$

for some finite m_0 and weights θ_1 and θ_2. We consider $R(h)$ the expected square loss and stochastic gradient descent with fresh new samples (x_i, y_i) and a small step-size.

We consider several number m of hidden neurons, to assess when the original neurons can be recovered. In Figure 2, for large m (e.g., $m = 100$ or $m = 1000$), all learned neurons converge to the neurons that generated the function which is in accordance with our main global convergence result (note that in general, recovering the neurons of the teacher is not a necessary condition for optimality, but it is always sufficient), while for $m = 5 > m_0$, where the global optimum will lead to perfect estimation, we may not recover the global optimum with a gradient flow. An interesting open question is to characterize mathematically the case $m = 20$, where we obtain the global optimum with moderate m.

In Figure 3, we consider several random initializations and random "teacher" networks and compute the generalization performance of the neural network after optimization. We see that for large m, good performance is achieved, while when m is too small, local minima remain problematic. This experiment suggests that the probability of global convergence quickly tends to 1 as m increases beyond m_0 in this setting, even in moderately high dimension.

5. GENERALIZATION GUARANTEES AND IMPLICIT BIAS FOR OVERPARAMETERIZED MODELS

A shown above, overparameterization—which takes the form of a large number of hidden neurons in our context—is a blessing for optimization, as it allows ensuring convergence to a global minimizer. When stochastic gradient descent with fresh observations at each iteration is used, then the predictor will converge to the optimal predictor (that is, it will minimize the performance on unseen data), but will do so potentially at a slow speed, and with the need for many observations. In this context, overparameterization does not lead to overfitting, but may rather underfit.

In practice, several passes over a finite amount of data (n observations) are used, and then overparameterization can in principle lead to overfitting. Indeed, among all predictors that will perfectly predict the training data, some will generalize, some will not. In this

1 The code to reproduce Figures 2 and 3 is available on the webpage https://github.com/lchizat/2021-exp-ICM.

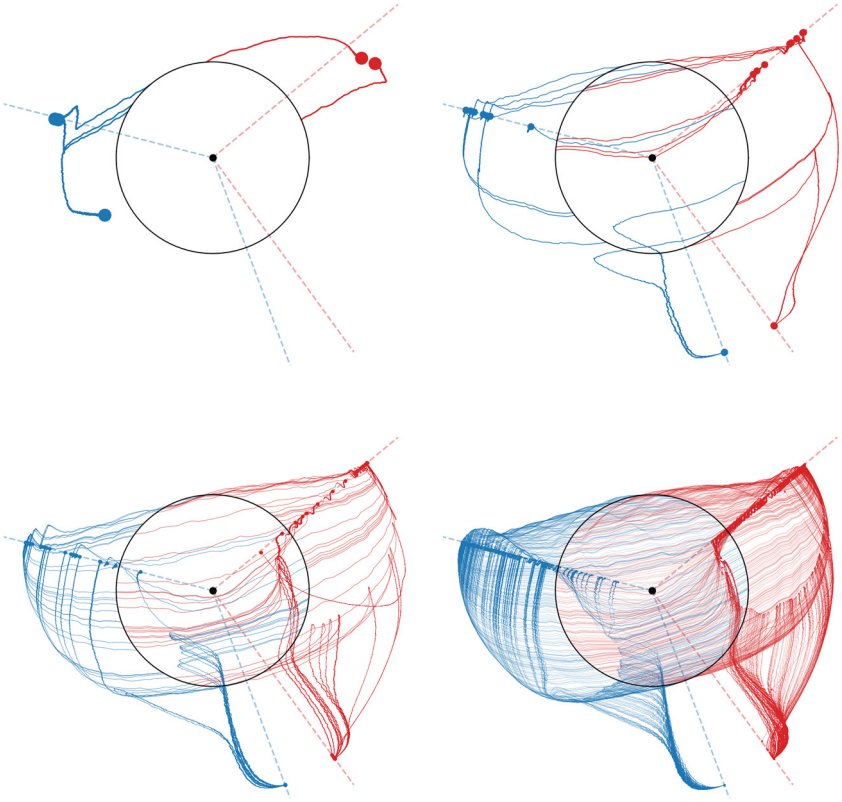


FIGURE 2

Gradient flow on a two-layer ReLU neural network with $m = 5$, $m = 20$, $m = 100$, and $m = 1000$, respectively. The position of the particles is given by $|\theta_2(j)| \cdot \theta_1(\cdot, j)$ and the color depends on the sign of $\theta_2(j)$. The dashed directions represent the neurons of the network that generates the data distribution (with $m_0 = 4$). The unit circle, where the particles are initialized, is plotted in black and the radial axis is scaled by tanh to improve readability.

section, we show that the predictor obtained after convergence of the gradient flow can in certain cases be characterized precisely.

To obtain the simplest result, following [17, 18, 43], this will be done for binary classification problems with the logistic loss. We will first review the implicit bias for linear models before considering neural networks.

5.1. Implicit bias for linear logistic regression

In this section, we consider a linear model $h(x, \theta) = \theta^\top \Phi(x)$ and we consider the minimization of the unregularized empirical risk with the logistic loss, that is,

$$\mathcal{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\bigl(1 + \exp\bigl(-y_i \theta^\top \Phi(x_i)\bigr)\bigr). \tag{5.1}$$

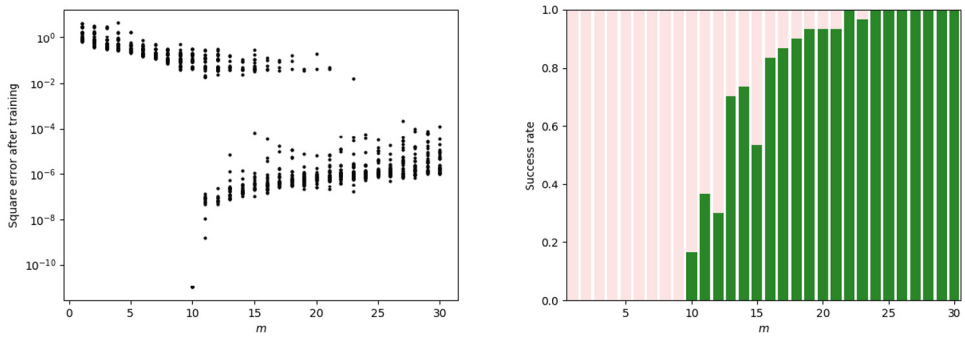


FIGURE 3

SGD on the square loss in the "teacher–student" setting (10^4 iterations, batch size 100, learning rate 0.005, $d = 100$, the teacher has $m_0 = 10$ neurons): (left) risk (expected square loss) after training as a function of m over 30 random repetitions; (right) success rate as a function of m over 30 repetitions (success means that the risk after training is below 10^{-3}).

We consider a *separable* problem where there exists a linear function in $\Phi(x)$, $\theta^\top \Phi(x)$ such that $y_i \theta^\top \Phi(x_i) > 0$ for all $i \in \{1, \dots, n\}$. By rescaling, we may equivalently assume that there exists $\theta \in \mathbb{R}^d$ such that

$$\forall i \in \{1, \dots, n\}, \quad y_i \theta^\top \Phi(x_i) \geqslant 1.$$

This means that the objective function in equation (5.1) has an infimal value of zero, which is not attained for any θ, since it is strictly positive. However, taking any θ that separates the data as above, it holds that $\mathcal{R}(t\theta)$ converges towards 0 as t tends to infinity. There are thus in general an infinite number of directions towards which θ can tend to reach zero risk.

It turns out that gradient descent selects a particular one: the iterate of gradient descent will diverge, but its direction (that is, the element of the sphere it is proportional to) will converge [43] to the direction of a *maximum margin classifier* defined as [46] a solution to

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_2^2 \quad \text{subject to } \forall i \in \{1, \dots, n\}, \; y_i \theta^\top \Phi(x_i) \geqslant 1. \tag{5.2}$$

The optimization problem above has a nice geometric interpretation (see Figure 4). These classifiers with a large margin has been shown to have favorable generalization guarantees in a wide range of contexts [22].

5.2. Extension to two-layer neural networks

We will now extend this convergence of gradient descent to a minimum norm classifier beyond linear models. We consider the minimization of the logistic loss

$$\frac{1}{n} \sum_{i=1}^{n} \log\bigl(1 + \exp\bigl(-y_i h(x_i)\bigr)\bigr),$$

where $h(x) = \frac{1}{m} \sum_{j=1}^{m} \theta_2(j) \max\{\theta_1(:, j)^\top x, 0\}$ is a two-layer neural network. We will consider two regimes: (1) where only the output weights $\theta_2(j)$, $j = 1, \dots, m$ are optimized, and (2) where all weights are optimized. In these two situations, we will let the width m

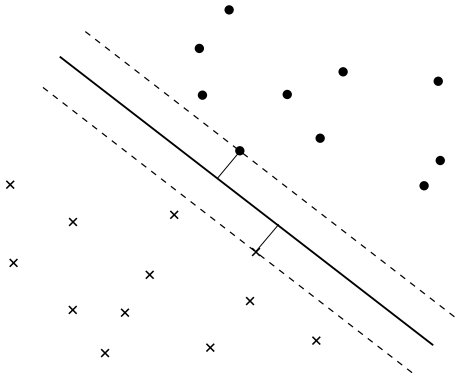


FIGURE 4

Geometric interpretation of equation (5.2) with a linearly separable binary classification problem in two dimensions (with each observation represented by one of the two labels $\times$ or $\bullet$): among all separating hyperplanes going through zero, that with the largest minimal distance from observations to the hyperplane will be selected.

go to infinity and consider the infinite-dimensional resulting flows. As shown in the previous section, when they converge, these flows converge to the global optimum of the objective function. But in the separable classification setting, the functions h should diverge. We essentially characterize towards which directions they diverge, by identifying the norms that are implicitly minimized [9].

5.3. Kernel regime

In this section, we consider random input weights $\theta_1(:, j)$, sampled from the uniform distribution on the sphere, and kept fixed throughout the optimization procedure. In other words, we only run the gradient flow with respect to the output weights $\theta_2 \in \mathbb{R}^m$.

Since the model is a linear model with feature vectors in m dimensions with components

$$\Phi(x)_j = \frac{1}{\sqrt{m}} \max\{\theta_1(:, j)^\top x, 0\},$$

we can apply directly the result above from [43], and the resulting classifier will minimize implicitly $\|\theta_2\|_2^2$, that is, the direction of θ_2 will tend to a maximum margin direction.

In order to study the situation when the number of features m tends to infinity, it is classical within statistics and machine learning to consider the kernel function $\hat{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined as

$$\hat{k}(x, x') = \Phi(x)^\top \Phi(x') = \frac{1}{m} \sum_{j=1}^{m} \max\{\theta_1(:, j)^\top x, 0\} \max\{\theta_1(:, j)^\top x', 0\}.$$

When m tends to infinity, the law of large number implies that $\hat{k}(x, x')$ tends to

$$k(x, x') = \mathbb{E}\big[\max\{\eta^\top x, 0\} \max\{\eta^\top x', 0\}\big],$$

for η uniformly distributed on the sphere.

Thus, we should expect that in the overparameterized regime, the predictor behaves like predictors associated with the limiting kernel function [32,37]. It turns out that the kernel k can be computed in closed form [11], and that the reproducing kernel Hilbert space (RKHS) functional norm $\|\cdot\|$ associated to the kernel k is well understood (see below for a formula that defines it). In particular, this norm is infinite unless the function is at least $d/2$-times differentiable [4], and thus very smooth in high dimension (this is to be contrasted with the fact that each individual neuron leads to a nonsmooth function). We thus expect smooth decision boundaries at convergence (see experiments below). This leads to the following result (see details in [9]):

Theorem 4 (Informal). *When $m, t \to +\infty$ (limits can be interchanged), the predictor associated to the gradient flow converges (up to normalization) to the function in the RKHS that separates the data with minimum* RKHS norm $\|\cdot\|$*, that is, the solution to*

$$\min_f \|f\|^2 \quad \textit{subject to } \forall i \in \{1,\dots,n\},\ y_i f(x_i) \geqslant 1.$$

Note that the minimum RKHS norm function can also be found by using the finite-dimensional representation $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ and minimizing $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)$ under the margin constraint, which is a finite-dimensional convex optimization problem.

A striking phenomenon is the absence of catastrophic overfitting, where the observed data are perfectly classified but with a very irregular function that would potentially not generalize well. Despite the strong overparameterization, the classifier selected by gradient descent can be shown to generalize through classical results from maximum margin estimation. See [29] for a related result where the performance as a function of m, and not only for infinite m, is considered in special settings. We will see a similar behavior when optimizing the two layers, but with a different functional norm.

5.4. Feature learning regime

We now consider the minimization with respect to both input and output weights. This will correspond to another functional norm that will not anymore be an RKHS norm, and will allow for more adaptivity, where the learned function can exhibit finer behaviors.

We first provide an alternative formulation of the RKHS norm as [4]

$$\|f\|^2 = \inf_{a(\cdot)} \int_{\mathcal{S}^{d-1}} |a(\eta)|^2 d\tau(\eta) \quad \text{such that } f(x) = \int_{\mathcal{S}^{d-1}} (\eta^\top x)_+ a(\eta) d\tau(\eta),$$

where the infimum is taken over all square-integrable functions on the sphere $\mathcal{S}^{d-1}$, and τ is the uniform probability measure on the sphere. This formulation highlights that functions in the RKHS combine infinitely many neurons.

We can then define the alternative *variation norm* [23] as

$$\Omega(f) = \inf_{a(\cdot)} \int_{\mathcal{S}^{d-1}} |a(\eta)| d\tau(\eta) \quad \text{such that } f(x) = \int_{\mathcal{S}^{d-1}} (\eta^\top x)_+ a(\eta) d\tau(\eta),$$

where the infimum is now taken over all integrable functions on $\mathcal{S}^{d-1}$. Going from squared L_2-norms to L_1-norms enlarges the space by adding nonsmooth functions. For example,

a single neuron corresponds to $a(\cdot)d\tau(\cdot)$ tending to a Dirac measure at a certain point, and thus has a finite variation norm.

This leads to the following result (see the details and full set of assumptions in [9]).

Theorem 5 (Informal). *When $m, t \to +\infty$, if the predictor associated to the gradient flow converges (up to normalization), then the limit is the function that separates the data with minimum* variation norm $\Omega(f)$*, that is, the solution to*

$$\min_f \Omega(f) \quad \textit{subject to } \forall i \in \{1, \dots, n\},\ y_i f(x_i) \geqslant 1.$$

Compared to the RKHS norm result, there is no known finite-dimensional convex optimization algorithms to efficiently obtain the minimum variation norm algorithm. Moreover, the choice of an L_1-norm has a sparsity-inducing effect, where the optimal $a(\cdot)d\tau(\cdot)$ will often corresponds to singular measure supported by a finite number of elements of the sphere. These elements can be seen as features learned by the algorithm: neural networks are considered as methods that learn representations of the data, and we provide here a justification with a single hidden layer. Such feature learning can be shown to lead to improved prediction performance in a series of classical situations, such as when the optimal function only depends on a few of the d original variables [4, 9].

5.5. Experiments

In this section, we consider a large ReLU network with $m = 1000$ hidden units, and compare the implicit bias and statistical performances of training both layers, which leads to a max margin classifier with the variation norm, versus the output layer, which leads to max margin classifier in the RKHS norm. These experiments are reproduced from [9].

Setting. Our data distribution is supported on $[-1/2, 1/2]^d$ and is generated as follows. In dimension $d = 2$, the distribution of input variables is a mixture of k^2 uniform distributions on disks of radius $1/(3k-1)$ on a uniform 2-dimensional grid with step $3/(3k-1)$, see Figure 6(a) for an illustration with $k = 3$. In dimension larger than 2, all other coordinates follow a uniform distribution on $[-1/2, 1/2]$. Each cluster is then randomly assigned a class in $\{-1, +1\}$.

Low dimensional illustrations. Figure 5 illustrates the differences in the implicit biases when $d = 2$. It represents a sampled training set and the resulting decision boundary between the two classes for 4 examples. The variation norm max-margin classifier is nonsmooth and piecewise affine, which comes from the fact that the L_1-norm favors sparse solutions. In contrast, the max-margin classifier for the RKHS norm has a smooth decision boundary, which is typical of learning in a RKHS.

Performance. In higher dimensions, we observe the superiority of training both layers by plotting the test error versus m or d on Figures 6(b) and 6(c). We ran 20 independent experiments with $k = 3$ and show with a thick line the average of the test error $\mathbb{P}(yf(x) < 0)$ after training. For each m, we ran 30 experiments using fresh random samples from the same data distribution.

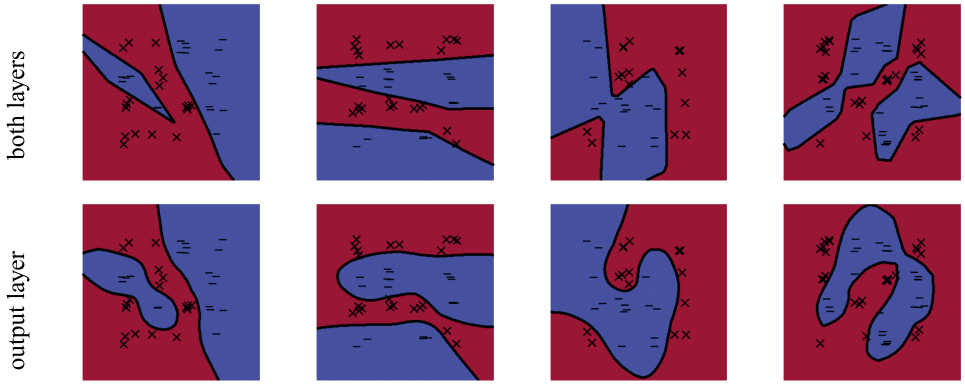
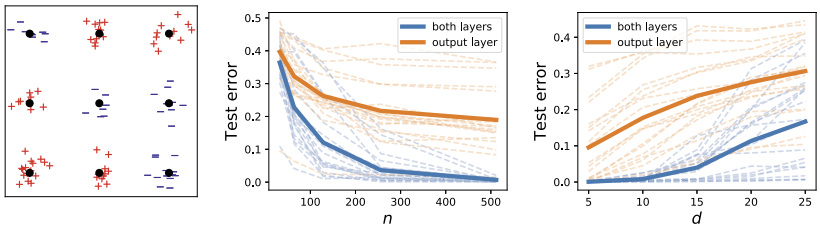


FIGURE 5

Comparison of the implicit bias of training (top) both layers versus (bottom) only the output layer for wide two-layer ReLU networks with $d = 2$ and for 4 different random training sets.



(a) Projected distribution (b) Test error vs. n (c) Test error vs. d

FIGURE 6

(a) Projection of the data distribution on the two first dimensions, (b) test error as a function of n with $d = 15$, and (c) test error as a function of d with $n = 256$.

6. DISCUSSION

In this paper, we have presented qualitative convergence guarantees for infinitely-wide two layer neural networks. These were obtained with a precise scaling—in the number of neurons—of the prediction function, the initialization and the step-size used in the gradient flow. With those scalings, the mean-field limit exhibits feature learning capabilities, as illustrated in binary classification where precise functional spaces could be used to analyze where optimization converges to. However, this limit currently does not lead to quantitative guarantees regarding the number of neurons or the convergence time, and obtaining such guarantees remains an open problem. This is an active area of research with, in particular, recent results concerning the local convergence [1, 7, 49] or global convergence under strong assumption on the data [26]. Moreover, extending this analysis to more than a single hidden layer or convolutional networks remains difficult.

Different scalings lead to different behaviors [10]. In particular, there is a scaling for which the limit behaves as a kernel method (even though all layers are trained, and not just the output layer) leading to another RKHS norm with a larger space than that from Section 5.3,

see [20]. While not leading to representation learning, extensions to deeper networks are possible with this scaling and provide one of few optimization and statistical guarantees for these models. Some recent progress has been made in the categorization of the various possible scalings for deep networks [48], and this emerging general picture calls for a large theoretical effort to understand the asymptotic behaviors of wide neural networks.

FUNDING

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support the European Research Council (grant SEQUOIA 724063).

REFERENCES

[1] S. Akiyama and T. Suzuki, On learnability via gradient method for two-layer ReLU neural networks in teacher-student setting. 2021, arXiv:2106.06251.

[2] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2008.

[3] D. Araújo, R. I. Oliveira, and D. Yukimura, A mean-field limit for certain deep neural networks. 2019, arXiv:1906.00193.

[4] F. Bach, Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.* **18** (2017), no. 1, 629–681.

[5] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** (1993), no. 3, 930–945.

[6] S. Bubeck, Convex optimization: algorithms and complexity. *Found. Trends Mach. Learn.* **8** (2015), no. 3–4, 231–357.

[7] L. Chizat, Sparse optimization on measures with over-parameterized gradient descent. *Math. Program.* (2021), 1–46.

[8] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport. *Adv. Neural Inf. Process. Syst.* **31** (2018), 3036–3046.

[9] L. Chizat and F. Bach, Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pp. 1305–1338, PMLR, 2020.

[10] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming. In *Advances in neural information processing systems*, pp. 2937–2947, Curran Associates, Inc., 2019.

[11] Y. Cho and L. K. Saul, Kernel methods for deep learning. In *Advances in neural information processing systems*, pp. 342–350, Curran Associates, Inc., 2009.

[12] W. E and S. Wojtowytsch, On the Banach spaces associated with multi-layer ReLU networks: Function representation, approximation theory and gradient descent dynamics. 2020, arXiv:2007.15623.

[13] C. Fang, J. Lee, P. Yang, and T. Zhang, Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pp. 1887–1936, PMLR, 2021.

[14] C. Giraud, *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[16] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik, Variance-reduced methods for machine learning. *Proc. IEEE* **108** (2020), no. 11, 1968–1983.

[17] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, Characterizing implicit bias in terms of optimization geometry. In *International conference on machine learning*, pp. 1832–1841, PMLR, 2018.

[18] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, Implicit regularization in matrix factorization. In *Advances in neural information processing systems*, pp. 6151–6159, Curran Associates, Inc., 2017.

[19] B. Hanin and M. Nica, Finite depth and width corrections to the neural tangent kernel. In *International conference on learning representations*, 2019.

[20] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, Curran Associates, Inc., 2018.

[21] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732, PMLR, 2017.

[22] V. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.* **30** (2002), no. 1, 1–50.

[23] V. Kurkova and M. Sanguineti, Bounds on rates of variable-basis and neural-network approximation. *IEEE Trans. Inf. Theory* **47** (2001), no. 6, 2659–2665.

[24] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications. 2nd edn.* Springer, 2003.

[25] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257, PMLR, 2016.

[26] Y. Li, T. Ma, and H. R. Zhang, Learning over-parametrized two-layer neural networks beyond NTK. In *Conference on learning theory*, pp. 2613–2682, PMLR, 2020.

[27] S. Maniglia, Probabilistic representation and uniqueness results for measure-valued solutions of transport equations. *J. Math. Pures Appl.* **87** (2007), no. 6, 601–626.

[28] S. Mei, T. Misiakiewicz, and A. Montanari, Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464, PMLR, 2019.

[29] S. Mei and A. Montanari, The generalization error of random features regression: precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.* (2019).

[30] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci.* **115** (2018), no. 33, E7665–E7671.

[31] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT Press, 2018.

[32] R. M. Neal, *Bayesian learning for neural networks*. Ph.D. thesis, University of Toronto, 1995.

[33] Y. Nesterov, *Lectures on convex optimization*. Springer Optim. Appl. 137, Springer, 2018.

[34] P.-M. Nguyen and H. T. Pham, A rigorous framework for the mean field limit of multilayer neural networks. Tech. Rep. 2020, arXiv:2001.11443.

[35] A. Nitanda and T. Suzuki, Stochastic particle gradient descent for infinite ensembles. 2017, arXiv:1712.05438.

[36] A. Nowak-Vila, F. Bach, and A. Rudi, A general theory for structured prediction with smooth convex surrogates. 2019, arXiv:1902.01958.

[37] A. Rahimi and B. Recht, Random features for large-scale kernel machines. *Adv. Neural Inf. Process. Syst.* **20** (2007), 1177–1184.

[38] G. M. Rotskoff and E. Vanden-Eijnden, Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Adv. Neural Inf. Process. Syst.* **31** (2018) 7146–7155.

[39] F. Santambrogio, *Optimal transport for applied mathematicians*. Springer, 2015.

[40] B. Schölkopf and A. J. Smola, *Learning with kernels*. MIT Press, 2001.

[41] J. Sirignano and K. Spiliopoulos, Mean field analysis of neural networks: a law of large numbers. *SIAM J. Appl. Math.* **80** (2020), no. 2, 725–752.

[42] J. Sirignano and K. Spiliopoulos, Mean field analysis of deep neural networks. *Math. Oper. Res.* (2021).

[43] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** (2018), no. 1, 2822–2878.

[44] E. Suli and D. F. Mayers, *An introduction to numerical analysis*. Cambridge University Press, 2003.

[45] A. W. Van der Vaart, *Asymptotic statistics*. Camb. Ser. Stat. Probab. Math. 3, Cambridge University Press, 2000.

[46] V. N. Vapnik and A. Y. Chervonenkis, On a perceptron class. *Avtomat. i Telemekh.* **25** (1964), no. 1, 112–120.

[47] S. Wojtowytsch, On the convergence of gradient descent training for two-layer ReLU-networks in the mean field regime. 2020, arXiv:2005.13530.

[48] G. Yang and E. J. Hu, Feature learning in infinite-width neural networks. 2020, arXiv:2011.14522.

[49] M. Zhou, R. Ge, and C. Jin, A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on learning theory*, pp. 4577–4632, PMLR, 2021.

FRANCIS BACH
Inria & Ecole Normale Supérieure, PSL Research University, Paris, France,
francis.bach@inria.fr

LÉNAÏC CHIZAT
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland,
lenaic.chizat@epfl.ch

ON MATHEMATICAL MODELING IN IMAGE RECONSTRUCTION AND BEYOND

BIN DONG

ABSTRACT

Imaging has been playing a vital role in the development of natural sciences. Advances in sensory, information, and computer technologies have further extended the scope of influence of imaging, making digital images an essential component of our daily lives. Image reconstruction is one of the most fundamental problems in imaging. For the past three decades, we have witnessed phenomenal developments of mathematical models and algorithms in image reconstruction. In this paper, we will first review some progress of the two prevailing mathematical approaches, i.e., the wavelet frame-based and PDE-based approaches, for image reconstruction. We shall discuss the connections between the two approaches and the implications and impact of the connections. Furthermore, we will review how the studies of the links between the two approaches lead us to a mathematical understanding of deep convolutional neural networks, which has led to further developments in modeling and algorithmic design in deep learning and new applications of machine learning in scientific computing.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 65D18; Secondary 68U10, 65T60, 68T05, 65M32

KEYWORDS

Convolutional neural networks, deep learning, machine learning, partial differential equations, total variation, wavelets, wavelet frames

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5420–5449
DOI 10.4171/ICM2022/2

1. INTRODUCTION

The development of natural sciences has been heavily relying on visual examinations. Through observations on natural phenomenons made by our naked eyes or via instruments such as cameras, microscopes, telescopes, etc., scientists make a scientific hypothesis on the underlying principles hidden in the phenomenon, and they later conduct more experiments or resort to mathematical deductions to further verify their hypothesis. Therefore, images play a central role since they can accurately record the phenomenon of interest and be further processed and analyzed by algorithms to assist human decision-making. In the past few decades, we are experiencing rapid advances in information technology, which contribute significantly to the exponential growth of data. Digital images are of no doubt one of the essential components of data. Advanced computer technology has made it possible to apply some of the most sophisticated developments in mathematics and machine learning to the design and implementation of efficient algorithms to process and analyze image data. As a result, the impact of images has now gone far beyond natural sciences. Image processing and analysis techniques are now widely adopted in engineering, medicine, technical disciplines, and social media, and digital images have become an essential element of our daily lives.

Among all tasks within the scope of computer vision, image reconstruction, such as image denoising, deblurring, inpainting, medical imaging, etc., is one of the most fundamental ones. Its objective is to obtain high-quality reconstructions of images that are corrupted in various ways during the process of acquisition, storage, and communication, and enable us to see crucial but subtle objects that reside in the images. Mathematics has been the main driven force in the advancement of image reconstruction for the past few decades [7, 33, 53]. Conversely, image reconstruction also brings to mathematics new challenging problems and fascinating applications that gave birth to new mathematical tools, whose application has even gone beyond the scope of image reconstruction.

Image reconstruction can be formulated as the following inverse problem:

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{u} + \boldsymbol{\eta}. \tag{1.1}$$

Here, $\boldsymbol{A}$ is a linear operator corresponding to the imaging process. For example, $\boldsymbol{A}$ is an identity operator for image denoising; a convolution operator for image deblurring; a restriction operator for image inpainting [13]; a subsampled Fourier transform for magnetic resonance imaging (MRI) [19]; a subsampled Radon transform for X-ray-based computed tomography (CT) [22]. Variable $\boldsymbol{u}$ is the unknown image to be reconstructed, and $\boldsymbol{f}$ is the measurements that are contaminated by additive noise $\boldsymbol{\eta}$ with known or partially known statistics, e.g., Gaussian, Laplacian, Poisson, etc. The main challenge in solving the linear inverse problem (1.1) is the ill-posedness of the problem. A naive inversion of $\boldsymbol{A}$, such as pseudoinversion or via Tikhonov regularization [154], may result in a reconstructed image with amplified noise and smeared-out edges.

Many existing image reconstruction models and algorithms are transformation-based. One of the earliest transforms was the Fourier transform, which is effective on signals that are smooth and sinusoidal-like. However, the Fourier transform is not adequate on images with multiple localized frequency components. Windowed Fourier transforms [72]

were introduced to overcome the poor spatial localization of the Fourier transform. However, the high-frequency coefficients in the transform domain are not ideally sparse for images due to the fixed time-frequency resolution of the windowed Fourier transforms. This is why wavelets and wavelet frames are much more effective for images than Fourier or windowed Fourier transforms because of their varied time-frequency resolution, which enables them to provide a better sparse approximation to piecewise smooth functions [45, 51, 110].

Another influential class of methods for image reconstruction that have been developed through a rather different path from wavelets is the PDE-based approach [33, 119, 136], which includes variational and (nonlinear) PDE methods. The basic idea of variational methods is to characterize images as functions living in a certain function space, such as the BV space [115, 131] (space of functions with bounded variations), and an energy functional is designed according to the function space assumption. PDE methods, on the other hand, often take the observed low-quality image or a coarsely reconstructed image as the initialization and enhance it by evolving a carefully designed nonlinear PDE that conducts smoothing in homogeneous regions and edge-preservation or enhancement near edges [120, 123].

The two approaches have been developing independently for decades. Although studies were showing the links between the two approaches [84, 148] using specific models and algorithms, their general connections were still unknown. Later in [24, 26, 42, 52], fundamental connections between wavelet frame-based approach and variational methods were established. Connections of wavelet frame-based models to the total variation model were established in [24], to the Mumford–Shah model were established in [26], and to some more general variational models such as the total generalized variation model [18] were established in [42, 52]. On the other hand, [49] established a generic connection between iterative wavelet frame shrinkage and general nonlinear evolution PDEs. We showed that wavelet frame shrinkage algorithms could be viewed as discrete approximations of nonlinear evolution PDEs. Such connection led to new understandings of both the wavelet frame- and PDE-based approach and expanded the scope of applications for both. The series of papers [24, 26, 42, 49, 52] essentially merged the two seemingly unrelated areas: wavelet frame-based and PDE-based approach for image reconstruction, and gave birth to many new image reconstruction models and algorithms.

For the past decade, the landscape of research and technological development of image reconstruction and computer vision is experiencing a significant transformation due to the advances in machine learning, especially deep learning [71, 91, 145]. A new set of models call the convolutional neural networks (CNNs) [65, 92] were introduced, where the AlexNet [89], U-Net [130], ResNet [77], and DenseNet [79] are well-known examples. Most CNNs have millions to billions of parameters that are trained (or optimized) on large data sets via stochastic algorithms. One remarkable property of deep neural networks (DNNs) in general is that they can well approximate nonlinear functions in high-dimensional spaces without suffering from the curse of dimensionality [36, 104, 114, 142–144, 163, 164, 170]. CNNs were first shown to be extremely effective in image classification [77, 89]. They were later adopted in image reconstruction and significantly advanced its state-of-the-art (see, e.g., [38, 113, 156, 159, 172]).

Why CNNs perform so well in practice and where their capability boundary locates is arguably the biggest mystery in deep learning for the moment. One possible way of unraveling such mystery, at least for image reconstruction, is to explore the connection between CNNs and mathematical models we now have a systematic understanding of. More importantly, what do CNNs do differently to outperform these mathematical models significantly, and can we combine the wisdom from both sides? Answering these questions can bring new insights into CNN models and further extend the scope of their applications.

Let $\mathcal{F}$ be an image reconstruction operator for the problem (1.1) that takes a coarse reconstruction of the image as input and the reconstructed image as output. For both wavelet frame-based and PDE-based models, this mapping $\mathcal{F}$ is a discrete dynamical system. As shown by [49], most of these discrete dynamical systems are various discrete approximations to differential equations. CNNs, on the other hand, are formed by consecutive compositions of relatively simple functions, which makes them discrete dynamical systems as well. We use $\mathcal{F}_{\boldsymbol{\Theta}}$ to denote a CNN, which is a parameterized dynamical system. One apparent difference between $\mathcal{F}$ and $\mathcal{F}_{\boldsymbol{\Theta}}$ is that the former is entirely design-based using human knowledge while the latter has minimal human design and its actual form mostly relies on a large number of parameters $\boldsymbol{\Theta}$ that are optimized through empirical risk minimization. The dynamics $\mathcal{F}$ and $\mathcal{F}_{\boldsymbol{\Theta}}$ are two extremes of modeling where the former advocates human knowledge, which grants solid theoretical foundations and adequate interpretability, while the latter promotes data-driven modeling which can extract features and principles from data that may be unknown to humans to better assist in decision making. However, in practice, neither extreme is ideal, which is especially the case in science, economics, and medicine. In these disciplines, interpretability is mostly required. Also, we have some knowledge to describe a particular phenomenon but still largely not enough, and we have observational or simulation data but limited in quantity. Therefore, we need to balance between the two extremes depending on the specific application of interest. Finding connections between $\mathcal{F}$ and $\mathcal{F}_{\boldsymbol{\Theta}}$ may better assist us in this regard.

This motivated us to study connections between CNNs and differential equations. From the standpoint of dynamical systems, we explored the structural similarities between numerical differential equations and CNNs in [101,102,107]. In [107], we showed that not only ResNet could be viewed as a forward-Euler approximation to differential equations as first pointed out by [74,162], but many other CNNs with bypass structures (or skip connections) can also be viewed as a discrete approximation of differential equations. Furthermore, [107] was the first to draw connections between residual-type CNNs with random perturbations and stochastic differential equations (SDEs). In fact, [107] suggested numerical ODEs/SDEs as a systematic framework for designing CNNs for image classification. In [101, 102], we were among the earliest to explore the structural similarity between CNNs and numerical PDEs. The key to such structural similarity is also the key to the connections between wavelet frame-based and PDE-based approaches for image reconstruction. By exploiting such structural similarity, we proposed a set of new CNNs called PDE-Nets, which can estimate the analytical form of (time-dependent) PDEs from observed dynamical data with minor prior knowledge on the underlying mechanism that drives the dynamics. Once trained,

the PDE-Net also serves as a simulator that can generate more dynamical data accurately and efficiently.

This paper will review the development of the wavelet frame-based and PDE-based approaches for image reconstruction. We shall discuss the connections between the two approaches and demonstrate how the connections lead to new models for image reconstruction. Furthermore, we will show how these theoretical studies inspired our exploration of structural similarities between differential equations and CNNs. These findings lead to further developments in modeling and algorithmic design in deep learning and new applications of machine learning in scientific computing.

2. WAVELET FRAME-BASED APPROACH FOR IMAGE RECONSTRUCTION

We start with a brief introduction to the concept of wavelet frame transform in a discrete setting. The interested readers should consult [45, 46, 128, 129] for theories of frames and wavelet frames, [51, 140] for a short survey on the theory and applications of frames, and [53] for a more detailed survey.

In the discrete setting, let an image $\boldsymbol{f}$ be a d-dimensional array. We denote by $\mathcal{I}_d = \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_d}$ the set of all d-dimensional images. We denote the d-dimensional fast $(L+1)$-level wavelet frame transform/decomposition with filters $\{\boldsymbol{q}^{(0)}, \boldsymbol{q}^{(1)}, \ldots, \boldsymbol{q}^{(r)}\}$ (see, e.g., [53]) by

$$\boldsymbol{W}\boldsymbol{u} = \left\{\boldsymbol{W}_{\ell,l}\boldsymbol{u} : (\ell, l) \in \mathbb{B}\right\}, \quad \boldsymbol{u} \in \mathcal{I}_d, \tag{2.1}$$

where $\mathbb{B} = \{(\ell, l) : 1 \le \ell \le r,\ 0 \le l \le L\} \cup \{(0, L)\}$. The wavelet frame coefficients $\boldsymbol{W}_{\ell,l}\boldsymbol{u} \in \mathcal{I}_d$ are computed by $\boldsymbol{W}_{\ell,l}\boldsymbol{u} = \boldsymbol{q}_{\ell,l}[-\cdot] \circledast \boldsymbol{u}$, where $\circledast$ denotes the convolution operator with a certain boundary condition, e.g., periodic boundary condition, and $\boldsymbol{q}_{\ell,l}$ is defined as

$$\boldsymbol{q}_{\ell,l} = \check{\boldsymbol{q}}_{\ell,l} \circledast \check{\boldsymbol{q}}_{l-1,0} \circledast \cdots \circledast \check{\boldsymbol{q}}_{0,0} \quad \text{with} \quad \check{\boldsymbol{q}}_{\ell,l}[\boldsymbol{k}] = \begin{cases} \boldsymbol{q}_\ell[2^{-l}\boldsymbol{k}], & \boldsymbol{k} \in 2^l\mathbb{Z}^d, \\ 0, & \boldsymbol{k} \notin 2^l\mathbb{Z}^d. \end{cases} \tag{2.2}$$

Similarly, we can define $\tilde{\boldsymbol{W}}\boldsymbol{u}$ and $\tilde{\boldsymbol{W}}_{\ell,l}\boldsymbol{u}$ given a set of dual filters $\{\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}_1, \ldots, \tilde{\boldsymbol{q}}_r\}$. We denote the inverse wavelet frame transform (or wavelet frame reconstruction) as $\tilde{\boldsymbol{W}}^\top$, which is the adjoint operator of $\tilde{\boldsymbol{W}}$. When the primal filters $\{\boldsymbol{p}, \boldsymbol{q}^{(1)}, \ldots, \boldsymbol{q}^{(r)}\}$ and dual filters $\{\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}_1, \ldots, \tilde{\boldsymbol{q}}_r\}$ satisfy the extension principles [128, 129], we have the perfect reconstruction formula

$$\boldsymbol{u} = \tilde{\boldsymbol{W}}^\top \boldsymbol{W}\boldsymbol{u}, \quad \text{for all } \boldsymbol{u} \in \mathcal{I}_d.$$

In particular, when the dual filters are the same as the primal filters with the extension principle satisfied, $\boldsymbol{W}$ is the transform associated to a tight frame system, and we simply have that

$$\boldsymbol{u} = \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{u}, \quad \text{for all } \boldsymbol{u} \in \mathcal{I}_d. \tag{2.3}$$

For simplicity, we will mostly focus our discussions on the case $d = 2$.

Two simple but useful examples of filters for univariate tight frame systems, i.e., Haar and piecewise linear tight frame system, constructed from B-splines [129] are given as follows.

Example 2.1. Filters of B-spline tight frame systems.

(1) *Haar*. Let $\boldsymbol{p} = \frac{1}{2}[1, 1]$ be the refinement mask of the piecewise constant B-spline $B_1(x) = 1$ for $x \in [0, 1]$ and 0 otherwise. Define $\boldsymbol{q}_1 = \frac{1}{2}[1, -1]$.

(2) *Piecewise linear*. Let $\boldsymbol{p} = \frac{1}{4}[1, 2, 1]$ be the refinement mask of the piecewise linear B-spline $B_2(x) = \max(1 - |x|, 0)$. Define $\boldsymbol{q}_1 = \frac{\sqrt{2}}{4}[1, 0, -1]$ and $\boldsymbol{q}_2 = \frac{1}{4}[-1, 2, -1]$.

The key to the success of wavelet frames in image reconstruction is their capability to provide a sparse approximation to images. In other words, the high-frequency band $\mathbb{B}\backslash\{(0, L)\}$ of the wavelet frame transform $\boldsymbol{W}\boldsymbol{u}$ of a typical image $\boldsymbol{u}$ is sparse. Large (in magnitude) wavelet frame coefficients encode image features such as edges, while the coefficients are small in smooth regions. This is mainly due to the short support and high order of vanishing moments of wavelet frames that make them behave like differential operators (we will come back to this in Section 4).

Wavelet frame-based image reconstruction started from the seminal work [32]. The basic idea is as follows: Consider the linear inverse problem (1.1). After an initial reconstruction of the image $\boldsymbol{u}$, edges might be blurred, and noise is still present in the image. Since a clean image should be sparse in the wavelet frame domain, one of the simplest ways to sharpen the image and remove noise at the same time is to set small high-frequency coefficients to zero. When we reconstruct the image using the processed wavelet frame coefficients, it will no longer be consistent with the data, i.e., $\boldsymbol{A}\boldsymbol{u}$ may be far away from $\boldsymbol{f}$. The simplest way to correct it is by moving $\boldsymbol{u}$ closer to the hyperplane $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$. Then, we iterate this procedure till convergence. This leads to a wavelet frame-based iterative algorithm, which was later analyzed by [23] and revealed its relation to the following wavelet frame-based balanced model:

$$\min_{\boldsymbol{d}} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{W}^{\top}\boldsymbol{d} - \boldsymbol{f}\right\|_2^2 + \frac{\kappa}{2}\left\|(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^{\top})\boldsymbol{d}\right\|_2^2 + \|\boldsymbol{\lambda} \cdot \boldsymbol{d}\|_1. \tag{2.4}$$

The balanced model also takes the analysis model [25, 55, 147]

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|_2^2 + \|\boldsymbol{\lambda} \cdot \boldsymbol{W}\boldsymbol{u}\|_1, \tag{2.5}$$

and the synthesis model [47, 59, 60, 63, 64]

$$\min_{\boldsymbol{d}} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{W}^{\top}\boldsymbol{d} - \boldsymbol{f}\right\|_2^2 + \|\boldsymbol{\lambda} \cdot \boldsymbol{d}\|_1 \tag{2.6}$$

as special cases. The balanced, analysis and synthesis models (and their variants) are among the most commonly used models in image reconstruction.

The objective functions in (2.4)–(2.6) are all convex, and can be efficiently optimized by convex optimization algorithms. For example, both the balanced and synthesis models can be solved efficiently by proximal forward–backward splitting (PFBS) [20, 35, 44,

[122, 155] and can be further accelerated by Nesterov's approach [10, 141]. The analysis model can be solved efficiently using the alternating direction method of multipliers (ADMM) [16, 25, 66, 68, 69] and the primal dual hybrid gradient (PDHG) method [30, 58, 176].

3. PDE-BASED APPROACH FOR IMAGE RECONSTRUCTION

In the past few decades, many variational and PDE models have been proposed with success in different tasks in image reconstruction. In this section, we shall refer to them both as the PDE-based approach. Successful examples of the PDE-based approach include the total variation (TV) model [131], total generalized variation model [18], Mumford–Shah model [115], shock-filter [120], Perona–Malik (PM) equation [123], anisotropic diffusion models [161], fluid dynamics model [12], etc. In this section, we will recall the TV model and the PM equation.

Regularization is crucial in solving ill-posed inverse problems. In 1963, Tikhonov proposed the so-called Tikhonov regularization [154] that penalizes the H^1 seminorm of the image to be reconstructed. Tikhonov regularization can effectively remove noise while it smears out important image features such as edges as well. This is essentially because H^1 is not an appropriate function space to model images. It has such a strong regularity requirement that functions with jump discontinuities are not allowed in the function space. To overcome such drawbacks, Rudin, Osher, and Fatemi proposed the refined TV model that penalizes the total variation of the function to be reconstructed so that jump discontinuities can be well-preserved and noise can be adequately removed. This is because the BV space is large enough to include functions with discontinuities but not too large, so that noise is still excluded.

Now, we first recall the definition of TV and the BV space. Let $\Omega \subset \mathbb{R}^2$ be an open set and $u \in L_1(\Omega)$. Then, the total variation of u is defined as

$$\mathrm{TV}(u) := \sup \left\{ \int_\Omega u \operatorname{div} v \, \mathrm{d}x : v \in C_c^1(\Omega, \mathbb{R}^2),\ \|v\|_{L_\infty(\Omega)} \le 1 \right\}, \tag{3.1}$$

where $C_c^1(\Omega, \mathbb{R}^2)$ is the space of all compactly supported continuously differentiable functions on Ω. Another convenient notation for the TV of a function u is $\mathrm{TV}(u) = \int_\Omega |Du(x)| \mathrm{d}x$, where Du is the distributional derivative of u. Intuitively speaking, the TV of a function u records the total amount of fluctuation of the function on domain Ω. If u is differentiable, then $\mathrm{TV}(u) = \int_\Omega |\nabla u(x)| \, \mathrm{d}x$. We define the BV space as

$$\mathrm{BV}(\Omega) = \{u \in L_1(\Omega) : \mathrm{TV}(u) < +\infty\}.$$

We now consider the function version of the image reconstruction problem (1.1), namely

$$f = Au + \eta.$$

We use nonbold characters to denote functions and linear operators in contrast to the bold characters that denote arrays and matrices. Then, the TV model for image reconstruction

reads as follows:

$$\min_{u\in \mathrm{BV}} \mathrm{TV}(u) + \frac{\lambda}{2}\int_{\Omega} \big(Au(x) - f(x)\big)^2 \mathrm{d}x, \tag{3.2}$$

where $\lambda > 0$ is a preselected hyperparameter that balances the amount of regularization from the first term and data consistency from the second term. Ways of solving the TV model include solving the associated Euler–Lagrange equation or the gradient flow, or we can discretize the model first and then use a convex optimization algorithm (e.g., one of those described in the previous section). Note that (3.2) is similar in form to (2.5). The difference is that in (2.5), we penalize the ℓ_1-norm of the wavelet frame transform of $\boldsymbol{u}$, while in (3.2) we penalize the L_1-norm of Du. This is an indication that (3.2) and (2.5) may be closely related. For convenience, we shall call the variational model (3.2) and its variants as differential operator-based analysis model, and (2.5) the wavelet frame-based analysis model.

In contrast to variational models for image reconstruction, designing PDE models is less restrictive and more intuitive to incorporate local geometric structures of images in the design. The scale-space theory tells us that using PDEs to model image reconstruction is a reasonable option. Let us use a set of nonlinear operators $\{\boldsymbol{T}_t\}_{t\geq 0}$ with $u(t,x) = (T_t u_0)(x)$ to denote the flow of image reconstruction starting from an initial estimation $u_0(x)$. If the set of operators satisfies certain axioms, such as recursivity, regularity, locality, translation invariance, etc., then there exists a second-order nonlinear evolution PDE such that $u(t,x)$ is its viscosity solution [3]. The PM equation is one of the well-known PDE models that are effective in image reconstruction (originally for image denoising but can be extended to general image reconstruction problems). It imposes a different amount of diffusion, even backward diffusion, in different regions of the images depending on local regularity and the orientation of edges. In the following, we will recall the idea of the original design of the PM equation for image denoising. Interested readers should consult [7,123] for more details.

Given an observed noisy image $u_0(x)$, the PM equation takes the following form:

$$\begin{cases} u_t = \mathrm{div}\big(g\big(|\nabla u|^2\big)\nabla u\big), & \text{on } (0,T)\times\Omega, \\ \frac{\partial u}{\partial \vec{n}}(t,x) = 0, & \text{on } (0,T)\times\partial\Omega, \\ u(0,x) = u_0(x), & \text{on } \Omega, \end{cases}$$

where the diffusivity function g is a scalar function satisfying

$$\begin{cases} g : [0,\infty) \mapsto (0,\infty) \text{ is monotonically decreasing;} \\ g(0) = 1; \qquad g(x) \to 0 \quad \text{as} \quad x \to \infty; \\ g(x) + 2xg'(x) > 0 \quad \text{for } x \leq K; \qquad g(x) + 2xg'(x) < 0 \quad \text{for } x > K. \end{cases} \tag{3.3}$$

The specific design of the diffusivity function g is to impose not only a spatially variant diffusion, but also different amount of diffusion in different directions at any given location. Commonly used examples of the diffusivity function g include

$$g(s) = e^{-\frac{s}{2\sigma^2}}, \quad \text{or} \quad g(s) = \frac{1}{1 + s^p/\lambda^2}, \quad p > \frac{1}{2},\ \lambda > 0.$$

From (3.3), we can see that $g(|\nabla u|^2)$ is relatively large at smooth regions of the image u where $|\nabla u|$ is relatively small. Thus, the PM equation applies stronger smoothing

in smooth regions of the image. In contrast, $|\nabla u|$ is relatively large near edges, and hence $g(|\nabla u|^2)$ is relatively small. Then, the PM equation applies less smoothing near edges which can reduce the amount of blurring. On the other hand, if we decompose the PM equation along the tangential and normal direction of the level sets of u, we can rewrite the original PM equation as

$$u_t = g\big(|\nabla u|^2\big)u_{TT} + \tilde{g}\big(|\nabla u|^2\big)u_{NN},$$

with

$$\tilde{g}(x) = g(x) + 2xg'(x), \quad N = \frac{\nabla u}{|\nabla u|}, \quad \text{and} \quad T = N^{\perp}, \quad |T| = 1.$$

Here, T and N are two unit vector fields that record, respectively, the tangential and normal directions of the level sets of function u. Further, u_{TT} and u_{NN} are the second-order derivatives along the tangential direction T and normal direction N, respectively. We can see from (3.3) that the PM equation imposes forward diffusion along the tangential direction to remove noise, while imposing backward diffusion along the normal direction near edges for enhancement. This, however, makes the PM equation an ill-posed PDE. This problem was later resolved by [28] where a modification of the PM equation was proposed and analyzed.

4. CONNECTIONS BETWEEN WAVELET FRAME-BASED AND PDE-BASED APPROACHES

In this section, we will summarize the main findings from the work [24] that established the connections between the differential operator-based and wavelet frame-based analysis models, and the work [49] that established the connections between nonlinear evolution PDEs and iterative wavelet frame-based shrinkage algorithms. Extensions of these results can be found in [26, 42, 52].

4.1. Wavelet frame transform and differential operators

Wavelet frame transform is a collection of convolution operators with both low- and high-pass filters. For a given multiresolution analysis (MRA) based wavelet frame system, the low-pass filters are associated with the refinable functions, while the high-pass filters are associated with wavelet functions. Key properties of both refinable and wavelet functions, such as smoothness and vanishing moments, can be characterized by their associated filters. The key observation that eventually leads to the connections between wavelet frame and PDE-based approaches is the link between vanishing moments of wavelet functions and differential operators in discrete and continuum settings. This observation was first made in [24] and was further exploited in [26, 42, 49, 52].

For a high-pass filter $\boldsymbol{q}$, let $\widehat{\boldsymbol{q}}(\boldsymbol{\omega}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \boldsymbol{q}[\boldsymbol{k}]e^{-i\boldsymbol{k}\boldsymbol{\omega}}$ be its two-scale symbol. Throughout this paper, for a multiindex $\boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \mathbb{Z}_+^2$ and $\boldsymbol{\omega} \in \mathbb{R}^2$, write

$$\boldsymbol{\alpha}! = \alpha_1!\alpha_2!, \quad |\boldsymbol{\alpha}| = \alpha_1 + \alpha_2, \quad \boldsymbol{D}_{\boldsymbol{\alpha}} = \frac{\partial^{\boldsymbol{\alpha}}}{\partial \boldsymbol{\omega}^{\boldsymbol{\alpha}}} = \frac{\partial^{\alpha_1+\alpha_2}}{\partial \omega_2^{\alpha_2} \partial \omega_1^{\alpha_1}}.$$

We say that $\boldsymbol{q}$ (and $\widehat{\boldsymbol{q}}(\boldsymbol{\omega})$) has vanishing moments of order $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, where $\boldsymbol{\alpha} \in \mathbb{Z}_+^2$, provided that

$$\sum_{\boldsymbol{k}\in\mathbb{Z}^2} \boldsymbol{k}^{\boldsymbol{\beta}} \boldsymbol{q}[\boldsymbol{k}] = i^{|\boldsymbol{\beta}|} \frac{\partial^{\boldsymbol{\beta}}}{\partial \boldsymbol{\omega}^{\boldsymbol{\beta}}} \widehat{\boldsymbol{q}}(\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=\boldsymbol{0}} = 0 \tag{4.1}$$

for all $\boldsymbol{\beta} \in \mathbb{Z}_+^2$ with $|\boldsymbol{\beta}| < |\boldsymbol{\alpha}|$ and for all $\boldsymbol{\beta} \in \mathbb{Z}_+^2$ with $|\boldsymbol{\beta}| = |\boldsymbol{\alpha}|$ but $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$. We say that $\boldsymbol{q}$ has total vanishing moments of order K with $K \in \mathbb{Z}_+$ if (4.1) holds for all $\boldsymbol{\beta} \in \mathbb{Z}_+^2$ with $|\boldsymbol{\beta}| < K$. Suppose $K \geq 1$. If (4.1) holds for all $\boldsymbol{\beta} \in \mathbb{Z}_+^2$ with $|\boldsymbol{\beta}| < K$ except for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ with certain $\boldsymbol{\beta}_0 \in \mathbb{Z}_+^2$ and $|\boldsymbol{\beta}_0| = J < K$, then we say that $\boldsymbol{q}$ has total vanishing moments of order $K\backslash\{J+1\}$.

To have a better understanding of the concept of vanishing moments, let us look at one example. Let $\widehat{\boldsymbol{q}}_1(\boldsymbol{\omega}) = e^{i\omega_1} - e^{-i\omega_1}$, which is the first high-pass filter of the piecewise linear B-spline tight wavelet frame system in Example 2.1. Then,

$$\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = 0, \quad \frac{\partial}{\partial\omega_1}\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = 2i \neq 0, \quad \frac{\partial}{\partial\omega_2}\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = 0.$$

Thus $\widehat{\boldsymbol{q}}_1(\boldsymbol{\omega})$ has vanishing moments of order $(1,0)$. In addition, we have

$$\frac{\partial^2}{\partial\omega_1^2}\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = 0, \quad \frac{\partial^2}{\partial\omega_1\partial\omega_2}\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = 0, \quad \frac{\partial^2}{\partial\omega_2^2}\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = 0.$$

Therefore, $\boldsymbol{q}_1$ has total vanishing moments of order $3\backslash\{|(1,0)|+1\}$, or $3\backslash\{2\}$ (it does not have total vanishing moments of order $4\backslash\{2\}$ since $\frac{\partial^3}{\partial\omega_1^3}\widehat{\boldsymbol{q}}_1(\boldsymbol{0}) = -2i \neq 0$).

The following proposition from [49] describes the relation between the vanishing moments of high-pass filters and finite difference approximations of differential operators. This proposition was later applied to the work of PDE-Net [101,102] that explores and exploits structure similarities between deep convolutional neural networks and numerical PDEs.

Proposition 4.1. *Let $\boldsymbol{q}$ be a high-pass filter with vanishing moments of order $\boldsymbol{\alpha} \in \mathbb{Z}_+^2$. Then for a smooth function $F(\boldsymbol{x})$ on $\mathbb{R}^2$, we have*

$$\frac{1}{\varepsilon^{|\boldsymbol{\alpha}|}} \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \boldsymbol{q}[\boldsymbol{k}] F(\boldsymbol{x} + \varepsilon\boldsymbol{k}) = C_{\boldsymbol{\alpha}} \frac{\partial^{\boldsymbol{\alpha}}}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}}} F(\boldsymbol{x}) + O(\varepsilon),$$

where $C_{\boldsymbol{\alpha}}$ is the constant defined by

$$C_{\boldsymbol{\alpha}} = \frac{1}{\boldsymbol{\alpha}!} \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \boldsymbol{k}^{\boldsymbol{\alpha}} \boldsymbol{q}[\boldsymbol{k}] = \frac{i^{|\boldsymbol{\alpha}|}}{\boldsymbol{\alpha}!} \frac{\partial^{\boldsymbol{\alpha}}}{\partial \boldsymbol{\omega}^{\boldsymbol{\alpha}}} \widehat{\boldsymbol{q}}(\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=\boldsymbol{0}}.$$

If, in addition, $\boldsymbol{q}$ has total vanishing moments of order $K\backslash\{|\boldsymbol{\alpha}|+1\}$ for some $K > |\boldsymbol{\alpha}|$, then

$$\frac{1}{\varepsilon^{|\boldsymbol{\alpha}|}} \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \boldsymbol{q}[\boldsymbol{k}] F(\boldsymbol{x} + \varepsilon\boldsymbol{k}) = C_{\boldsymbol{\alpha}} \frac{\partial^{\boldsymbol{\alpha}}}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}}} F(\boldsymbol{x}) + O(\varepsilon^{K-|\boldsymbol{\alpha}|}).$$

Similar results can be written in terms of wavelet frame functions which is given by the following proposition of [42]. Note that a version of the same result for B-splines wavelet frames was proposed earlier in [24].

Proposition 4.2. *Let a tensor product wavelet frame function $\psi_{\boldsymbol{\alpha}} \in L_2(\mathbb{R}^2)$ have vanishing moments of order $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq s$, and let* $\operatorname{supp}(\psi_{\boldsymbol{\alpha}}) = [a_1, a_2] \times [b_1, b_2]$. *Then, there exists a unique $\varphi_{\boldsymbol{\alpha}} \in L_2(\mathbb{R}^2)$ such that $\varphi_{\boldsymbol{\alpha}}$ is differentiable up to order $\boldsymbol{\alpha}$ a.e.,*

$$c_{\boldsymbol{\alpha}} = \int_{\mathbb{R}^2} \varphi_{\boldsymbol{\alpha}} \neq 0 \quad \text{and} \quad \psi_{\boldsymbol{\alpha}} = \partial^{\boldsymbol{\alpha}} \varphi_{\boldsymbol{\alpha}}.$$

Furthermore, for $n \in \mathbb{N}$ and $\boldsymbol{k} \in \mathbb{Z}^2$ with $\operatorname{supp}(\psi_{\boldsymbol{\alpha},n-1,\boldsymbol{k}}) \subseteq \overline{\Omega}$, *we have*

$$\langle u, \psi_{\boldsymbol{\alpha},n-1,\boldsymbol{k}} \rangle = (-1)^{|\boldsymbol{\alpha}|} 2^{|\boldsymbol{\alpha}|(1-n)} \langle \partial^{\boldsymbol{\alpha}} u, \varphi_{\boldsymbol{\alpha},n-1,\boldsymbol{k}} \rangle$$

for every u belonging to the Sobolev space $W_1^s(\Omega)$. Here,

$$\psi_{\boldsymbol{\alpha},n-1,\boldsymbol{k}} = 2^{n-2} \psi_{\boldsymbol{\alpha}}\left(2^{n-1} \cdot - \boldsymbol{k}/2\right)$$

and $\varphi_{\boldsymbol{\alpha},n-1,\boldsymbol{k}}$ is defined similarly.

Note that Proposition 4.1 is more convenient to use in addressing the connections between wavelet frame shrinkage algorithm and nonlinear evolution PDEs. In contrast, Proposition 4.2 is more convenient to use in addressing the connections between wavelet frame-based and differential operator-based analysis models.

4.2. Connections between wavelet frame-based analysis model and TV model

The wavelet frame-based analysis model considered by [24] is given as

$$\inf_{u \in W_1^s(\Omega)} E_n(u) := \nu \|\boldsymbol{\lambda}_n \cdot \boldsymbol{W} \boldsymbol{T}_n u\|_1 + \frac{1}{2} \|\boldsymbol{A}_n \boldsymbol{T}_n u - \boldsymbol{T}_n f\|_2^2, \tag{4.2}$$

and the differential operator-based analysis model is given as

$$\inf_{u \in W_1^s(\Omega)} E(u) := \nu \|\boldsymbol{D} u\|_1 + \frac{1}{2} \|Au - f\|_{L_2(\Omega)}^2. \tag{4.3}$$

Here, $\boldsymbol{W}$ denotes the wavelet frame transform defined by (2.1) and (2.2), $\boldsymbol{T}_n$ is the sampling operator generated by the refinable function corresponding to the underlying wavelet frame system, $\boldsymbol{A}_n$ is a discrete approximation of the operator A, $\boldsymbol{D}$ is a certain linear differential operator with highest order s (e.g., for the TV model, $\boldsymbol{D} = \nabla$ and $s = 1$). We denote by $W_p^r(\Omega)$ the Sobolev space with functions whose rth order weak derivatives belong to $L_p(\Omega)$ and which is equipped with the norm $\|f\|_{W_p^r(\Omega)} := \sum_{|\boldsymbol{\alpha}| \leq r} \|D_{\boldsymbol{\alpha}} f\|_p$.

From the form of E_n and E, we can see a similarity between the two functionals. It was proved in [24] that E_n converges to E pointwise on $W_1^s(\Omega)$. However, since we are interested in the (approximated) minimizers of these functionals, pointwise convergence does not guarantee a relation between their associated (approximated) minimizers. Therefore, Γ-convergence [17] was used in [24] to draw a connection between the problems $\min_u E_n(u)$ and $\min_u E(u)$. We first recall the definition of Γ-convergence.

Definition 4.1. Given $E_n(u) : W_1^s(\Omega) \mapsto \bar{\mathbb{R}}$ and $E(u) : W_1^s(\Omega) \mapsto \bar{\mathbb{R}}$, we say that E_n Γ-converges to E if:

(i) for every sequence $u_n \to u$ in $W_1^s(\Omega)$, $E(u) \leq \liminf_{n\to\infty} E_n(u_n)$;

(ii) for every $u \in W_1^s(\Omega)$, there is a sequence $u_n \to u$ in $W_1^s(\Omega)$, such that $E(u) \geq \limsup_{n\to\infty} E_n(u_n)$.

Then, based on the link between wavelet frame transform and differential operators given by Proposition 4.2, the main result of [24] is given as follows:

Theorem 4.1. *Given the variational problem* (4.3)*, there exists a set of coefficients* $\boldsymbol{\lambda}_n$*, such that the functional* E_n *of the problem* (4.2) Γ*-converges to the functional* E *of the problem* (4.3) *in* $W_1^s(\Omega)$*. Let* $u_n^\star$ *be an* ε*-optimal solution to the problem* (4.2)*, i.e.,* $E_n(u_n^\star) \leq \inf_u E_n(u) + \epsilon$ ($\epsilon > 0$)*. We have that*

$$\limsup_{n\to\infty} E_n(u_n^\star) \leq \inf_u E(u) + \epsilon,$$

and any cluster point of $\{u_n^\star\}_n$ *is an* ε*-optimal solution to the problem* (4.3)*.*

Theorem 4.1 goes beyond the theoretical justifications of the linkage of (4.2) and (4.3). Since the differential operator-based analysis model (4.3) has strong geometric interpretations, this connection brings geometric interpretations to the wavelet frame-based approach (4.2) as well. This also leads to even wider applications of the wavelet frame-based approach, e.g., image segmentation [27, 48, 99] and 3D surface reconstruction [50]. Conversely, the theorem also grants a new perspective of sparse approximation to the PDE-based approach supplementing its current function space perspective. On the other hand, not only the wavelet frame-based analysis model can be viewed as a discrete approximation of the differential operator-based analysis model, but such discretization can also be superior to standard finite difference discretization commonly used in PDE-based methods. Taking the Haar wavelet frame-based analysis model as an example, its regularization term has the property of 45-degree rotation invariance. In contrast, the standard finite difference discretization for TV regularization does not have such an invariance. This enables Haar wavelet frame-based analysis model to generate better reconstructed images than the TV model with the standard discretization.

4.3. Connections between wavelet shrinkage algorithms and nonlinear evolutional PDEs

In [49], general connections between wavelet frame shrinkage algorithms and nonlinear evolution PDEs (e.g., PM equation, shock-filters, anisotropic diffusions) were established. The links between the two approaches provide new and inspiring interpretations of themselves that enable us to derive new PDE models and (better) wavelet frame shrinkage algorithms for image restoration. Here, we will recall some of the main results from [49].

Let $\boldsymbol{d} := \boldsymbol{W}\boldsymbol{u}$ be the wavelet frame transform of $\boldsymbol{u}$, $\tilde{\boldsymbol{W}}^\top \boldsymbol{d}$ be the inverse wavelet frame transform defined by (2.1) and (2.2) with the corresponding filters satisfying the extension principles [128,129]. Then, we have $\tilde{\boldsymbol{W}}^\top \boldsymbol{W} = \boldsymbol{I}$. For simplicity, we only consider 1-level wavelet frame transform. Given wavelet frame coefficients $\boldsymbol{d} = \{d_{\ell,\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{Z}^2, 0 \leq \ell \leq L\}$ and a threshold $\boldsymbol{\lambda}(\boldsymbol{d}) = \{\lambda_{\ell,\boldsymbol{n}}(\boldsymbol{d}) : \boldsymbol{n} \in \mathbb{Z}^2, 0 \leq \ell \leq L\}$, the shrinkage operator $\boldsymbol{S}_{\boldsymbol{\lambda}}(\boldsymbol{d})$ is defined as follows:

$$\boldsymbol{S}_{\boldsymbol{\lambda}}(\boldsymbol{d}) = \big\{S_{\lambda_{\ell,\boldsymbol{n}}(\boldsymbol{d})}(d_{\ell,\boldsymbol{n}}) = d_{\ell,\boldsymbol{n}}\big(1 - \lambda_{\ell,\boldsymbol{n}}(\boldsymbol{d})\big) : \boldsymbol{n} \in \mathbb{Z}^2, 0 \leq \ell \leq L\big\}. \tag{4.4}$$

Two well-known examples of the shrinkage operator (4.4) are the isotropic and anisotropic soft-thresholding operators [24, 49, 54].

Given the shrinkage operator $\boldsymbol{S}_{\boldsymbol{\lambda}}$, a generic wavelet frame shrinkage algorithm takes the form

$$\boldsymbol{u}^k = \tilde{\boldsymbol{W}}^\top \boldsymbol{S}_{\boldsymbol{\lambda}^{k-1}}(\boldsymbol{W}\boldsymbol{u}^{k-1}), \quad k = 1, 2, \dots \tag{4.5}$$

Note that, for simplicity, we have dropped the term of data fidelity. More general versions of the algorithm can be found in [49]. Now, consider the following nonlinear evolution PDE:

$$u_t = \sum_{\ell=1}^{L} \frac{\partial^{\boldsymbol{\alpha}_\ell}}{\partial x^{\boldsymbol{\alpha}_\ell}} \Phi_\ell(\boldsymbol{D}u, u), \quad \boldsymbol{D} = \left(\frac{\partial^{\boldsymbol{\beta}_1}}{\partial x^{\boldsymbol{\beta}_1}}, \dots, \frac{\partial^{\boldsymbol{\beta}_L}}{\partial x^{\boldsymbol{\beta}_L}} \right). \tag{4.6}$$

The PDE (4.6) is defined on $\mathbb{R}^2$, and $|\boldsymbol{\alpha}_\ell|, |\boldsymbol{\beta}_\ell| \geq 0$, $1 \leq \ell \leq L$. Thus, it covers most nonlinear parabolic and hyperbolic equations that we use for image reconstruction.

One key results of [49] can be summarized as follows: Given a PDE that takes the form (4.6), then we can construction wavelet frame transforms $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$, and a shrinkage operator $\boldsymbol{S}_{\boldsymbol{\lambda}}$ such that the wavelet frame shrinkage algorithm (4.5) is an approximation of the PDE (4.6). When the PDE (4.6) is a well-posed anisotropic diffusion, the discrete solution obtained from (4.5) converges to the solution of the PDE. This result is a consequence of Proposition 4.1.

Let us consider a simple example. Consider the PDE

$$u_t = \frac{\partial \Phi_1}{\partial x_1}\left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, u \right) + \frac{\partial \Phi_2}{\partial x_2}\left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, u \right).$$

Let $\boldsymbol{W}_\ell$, $\ell = 1, 2$, be the Haar wavelet frame transform corresponding to the first two high-frequency bands. By Proposition 4.1, we have the following discretization of the above PDE:

$$\begin{aligned} \boldsymbol{u}^k = \boldsymbol{u}^{k-1} &- \tau\tilde{\gamma}_1 \boldsymbol{W}_1^\top \Phi_1(\gamma_1 \boldsymbol{W}_1 \boldsymbol{u}^{k-1}, \gamma_2 \boldsymbol{W}_2 \boldsymbol{u}^{k-1}, \boldsymbol{u}^{k-1}) \\ &- \tau\tilde{\gamma}_2 \boldsymbol{W}_2^\top \Phi_2(\gamma_1 \boldsymbol{W}_1 \boldsymbol{u}^{k-1}, \gamma_2 \boldsymbol{W}_2 \boldsymbol{u}^{k-1}, \boldsymbol{u}^{k-1}), \end{aligned}$$

with parameters γ_ℓ and $\tilde{\gamma}_\ell$ being properly chosen such that $\gamma_\ell \boldsymbol{W}_\ell \approx \frac{\partial}{\partial x_\ell}$ and $\tilde{\gamma}_\ell \boldsymbol{W}_\ell^\top \approx \frac{\partial}{\partial x_\ell}$. On the other hand, the iterative algorithm (4.5) can be rewritten as

$$\begin{aligned} \boldsymbol{u}^k = \boldsymbol{u}^{k-1} &- \boldsymbol{W}_1^\top \big[\boldsymbol{W}_1 \boldsymbol{u}^{k-1} - \mathcal{S}_1(\boldsymbol{W}_1 \boldsymbol{u}^{k-1}, \boldsymbol{W}_2 \boldsymbol{u}^{k-1}, \boldsymbol{u}^{k-1}) \big] \\ &- \boldsymbol{W}_2^\top \big[\boldsymbol{W}_2 \boldsymbol{u}^{k-1} - \mathcal{S}_2(\boldsymbol{W}_1 \boldsymbol{u}^{k-1}, \boldsymbol{W}_2 \boldsymbol{u}^{k-1}, \boldsymbol{u}^{k-1}) \big]. \end{aligned}$$

Comparing the above two iterative formulas, we can see that if we define the operator $\mathcal{S} = \{S^\ell : \ell = 1, 2\}$ as

$$S^\ell(\xi_1, \xi_2, \zeta) := \xi_\ell - \tau\tilde{\gamma}_\ell \Phi_\ell(\xi_1, \xi_2, \zeta) = \xi_\ell \big(1 - \tau\tilde{\gamma}_\ell \Phi_\ell(\xi_1, \xi_2, \zeta)/\xi_\ell \big), \quad \xi_\ell, \zeta \in \mathbb{R},$$

(whenever $\Phi_\ell(\xi_1, \xi_2, \zeta)/\xi_\ell$ is well defined), then there is an exact correspondence between the two iterative formulas. Note that the threshold level in the original definition (4.4) is given by $\tau\tilde{\gamma}_\ell \Phi_\ell(\xi_1, \xi_2, \zeta)/\xi_\ell$. In particular, when

$$\Phi_\ell\left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, u \right) = g_\ell\big(|\nabla u|^2, u \big) \frac{\partial u}{\partial x_\ell},$$

we have

$$S^{\ell}(\xi_1,\xi_2,\zeta)=\xi_{\ell}\big(1-\tau\tilde{\gamma}\ell g_{\ell}(\xi_1^2+\xi_2^2,\zeta)\big).$$

It is interesting to observe that the threshold level given by $\tau\tilde{\gamma}\ell g_{\ell}(\xi_1^2+\xi_2^2,\zeta)$ is proportional to the diffusivity g_{ℓ}.

Other than showing that the wavelet frame shrinkage algorithms can be viewed as a discrete approximation of PDEs, [49] also presented examples of new PDEs that can be derived from wavelet frame shrinkage algorithms. Conversely, new wavelet shrinkage algorithms that better exploit local image geometry can also be derived. Here, we recall one such example.

Consider the accelerated wavelet frame shrinkage algorithm [93, 116, 117]

$$\begin{aligned}\boldsymbol{u}^k=(I-\mu\boldsymbol{A}^{\top}\boldsymbol{A})\boldsymbol{W}^{\top}\boldsymbol{S}_{\boldsymbol{\alpha}^{k-1}}\big((1+\gamma^{k-1})\boldsymbol{W}\boldsymbol{u}^{k-1}-\gamma^{k-1}\boldsymbol{W}\boldsymbol{u}^{k-2}\big)\\+\mu\boldsymbol{A}^{\top}\boldsymbol{f},\quad k=1,2,\ldots\end{aligned}\tag{4.7}$$

When we properly choose the wavelet frame transform $\boldsymbol{W}$ and the parameters μ and γ^k, the iterative algorithm (4.7) leads to the following PDE:

$$\begin{aligned}u_{tt}+Cu_t=\sum_{\ell=1}^{L}(-1)^{1+|\boldsymbol{\beta}_{\ell}|}\frac{\partial^{\boldsymbol{\beta}_{\ell}}}{\partial\boldsymbol{x}^{\boldsymbol{\beta}_{\ell}}}\left[g_{\ell}\left(u,\frac{\partial^{\boldsymbol{\beta}_1}u}{\partial\boldsymbol{x}^{\boldsymbol{\beta}_1}},\ldots,\frac{\partial^{\boldsymbol{\beta}_L}u}{\partial\boldsymbol{x}^{\boldsymbol{\beta}_L}}\right)\frac{\partial^{\boldsymbol{\beta}_{\ell}}}{\partial\boldsymbol{x}^{\boldsymbol{\beta}_{\ell}}}u\right]\\-\kappa A^{\top}(Au-f).\end{aligned}\tag{4.8}$$

What makes equation (4.8) interesting is the presence of both u_t and u_{tt}. The term u_t makes the PDE parabolic-like so that the first term on the right-hand side regularizes the solution u; the term u_{tt} makes the PDE hyperbolic-like so that the evolution of u is accelerated. The idea of using a hyperbolic equation to speed up convergence was proposed in [111] for sparse signal reconstruction from noisy, blurry observations. Furthermore, related findings was also given by [149, 150]. It also inspired more recent studies in machine learning that established connections between numerical ODEs and CNNs [107].

5. GOING BEYOND IMAGE RECONSTRUCTION

Differential equations, especially partial differential equations (PDEs), play a prominent role in physics, chemistry, biology, economics, engineering, etc., to describe the governing laws underlying virtually every physical, technical, or biological process. The application of differential equations in image reconstruction and computer vision is a relatively new field that started around 1990. In Section 4, we have unified the prevailing models in image reconstruction, from which we can see that most effective image reconstruction algorithms are various discrete approximations of differential equations. In this section, we shall bridge the design of certain types of CNNs with numerical differential equations. More specifically, the bridge between numerical ODEs/SDEs and CNNs was established by [107] and the bridge between numerical PDEs and CNNs was established by [101, 102]. In this line of work, we regard CNNs as a discrete dynamical system, and the flow of features from the very first layer to the last layer of the CNNs is the underlying dynamical process. We argue that different

numerical schemes of differential equations lead to different architectures of CNNs, which inherit certain properties from the differential equations. By connecting CNNs with numerical differential equations, we can bring in tools from applied mathematics and physics to shed light on the interpretability of CNNs; and we can also bring in tools from deep learning to further advance not only image reconstruction but also a much broader field of scientific computing.

5.1. ODE-Nets: exploring structural similarity between numerical ODEs and CNNs

One of the central tasks in deep learning is designing effective deep architectures with strong generalization potential and are easy to train. The first ultra-deep CNN is the ResNet [77] where skip connections were introduced to keep feature maps in different layers on the same scale and to avoid gradient vanishing. Structures other than the skip connections of the ResNet were also introduced to prevent gradient vanishing, such as the dense connections [79] and fractal path [90].

Observe that each residual block of ResNet can be written as

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t F(\boldsymbol{u}^k, t_k), \tag{5.1}$$

where k is the index for the residual block F. As first suggested by [74, 162], each residual block of ResNet is one step of the forward-Euler discretization of the ODE $\dot{u} = F(u, t)$. In [107], we further showed that many state-of-the-art deep network architectures, such as PolyNet [174], FractalNet [90], and RevNet [70], which can be considered as different discretizations of ODEs. From the perspective of [107], the success of these networks is mainly due to their ability to efficiently approximate dynamical systems.

Taking PolyNet as an example, a PolyInception module was introduced in each residual block. The PolyInception model includes polynomial compositions that can be described as

$$(I + F + F^2) \cdot x = x + F(x) + F\big(F(x)\big).$$

We observed in [107] that PolyInception model can be interpreted as an approximation to one step of the backward-Euler (implicit) scheme, $\boldsymbol{u}^{k+1} = (I - \Delta t F)^{-1}\boldsymbol{u}^k$. Indeed, we can formally rewrite $(I - \Delta t F)^{-1}$ as

$$I + \Delta t F + (\Delta t F)^2 + \cdots + (\Delta t F)^k + \cdots.$$

Therefore, the architecture of PolyNet can be viewed as an approximation to the backward-Euler scheme solving the ODE $\boldsymbol{u}_t = F(\boldsymbol{u}, t)$. Note that the implicit scheme allows a larger step size [6], which in turn allows fewer residual blocks.

Furthermore, for residual-type networks with random perturbations, such as ResNet with shake-shake regularization [67] and stochastic depth [80], it was shown by [107] that these networks can be viewed as weak approximations [118] to certain SDEs, which links the

training of such networks with mean-field stochastic optimal control

$$\min \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{P}}\left(\mathbb{E}\left(\ell(X_T,\boldsymbol{y})+\int_0^T r(s,X_s,\theta_s)\right)ds\right)$$
$$\text{s.t. } dX_t = f(X_t,\theta_t)dt + g(X_t,\theta_t)dB_t, \quad X_0 = \boldsymbol{x},$$

where $\ell(\cdot,\cdot)$ is a certain loss function measuring the distance between the two input arguments, $r(\cdot,\cdot,\cdot)$ is a running cost that regularizes the dynamics and $\mathcal{P}$ is the distribution of the data. Note that the SDE and stochastic optimal control perspective on ResNet with dropout [146] was later proposed by [151].

In [107], we argued that we could exploit numerical ODEs to design new residual-type CNNs with state-of-the-art classification accuracy. Here, we shall call these deep residual-type CNNs inspired by numerical schemes of ODEs as ODE-Nets. As an example, we proposed to use a linear two-step scheme for ODEs to design a new ODE-Net, called LM-ResNet, as follows:

$$\boldsymbol{u}^{k+1} = (1-\alpha_k)\boldsymbol{u}^k + \alpha_k\boldsymbol{u}^{k-1} + F(\boldsymbol{u}^k, t_k), \quad \alpha_k \in \mathbb{R}. \tag{5.2}$$

The difference between the LM-ResNet (5.2) and the original ResNet (5.1) is revealed by the modified equation analysis [160]. Modified equations are commonly used in numerical analysis to describe numerical behaviors of numerical schemes. The modified equations of the ResNet and the LM-ResNet are as follows:

$$\begin{cases} \dot{u}^k + \frac{\Delta t}{2}\ddot{u}^k = F(u^k,t_k), & \text{ResNet;} \\ (1+\alpha_k)\dot{u}^k + (1-\alpha_k)\frac{\Delta t}{2}\ddot{u}^k = F(u^k,t_k), & \text{LM-ResNet.} \end{cases} \tag{5.3}$$

Here, $u^k = u(t_k)$ and similarly for $\dot{u}^k$ and $\ddot{u}^k$. Comparing the two modified equations in (5.3), we can see that when $\alpha_k \le 0$, the second-order term $\ddot{u}$ of the modified equation of LM-ResNet is bigger than that of the original ResNet. Note that the term $\ddot{u}$ represents acceleration which leads to acceleration of the convergence of $\boldsymbol{u}^k$ when $F = -\nabla G$, which was observed earlier for $F(u)$ taking a particular form in (4.8). This was our original motivation to select (5.2) among numerous other numerical ODE schemes, since we believed the depth of the corresponding ODE-Net could be reduced compared to the original ResNet because of the acceleration mechanisms induced by the term $\ddot{u}^k$. It turned out that it was indeed the case, and LM-ResNet managed to reduce the depth of the original ResNet (the versions with stochastic perturbations as well) by a factor of 2–10 without hurting classification accuracy. This was empirically validated on image classification benchmarks CIFAR10/100 and ImageNet.

The bridge between numerical schemes and architectures of neural networks can not only inspire various designs of ODE-Nets [34, 94, 108, 177], concepts from the numerical analysis can also be introduced to enforce the ODE-Nets to satisfy certain desired properties. For example, [41] utilized a symplectic scheme to enforce the learned network to preserve the physic structure, and [173] boosted the stability and adversarial robustness of ResNet through stability analysis on the underlying dynamical system. The bridge also inspired the work on the neural ODE [37, 97] in which ODEs and SDEs were used as machine learning models, and they have achieved huge success in generative modeling. The validity of

using dynamical systems as machine learning models was provided by [96] where the universal approximation property of these models was established. This line of research also inspired more applications of ODE-Nets in time series prediction [87] and physical system identification [73, 127, 168]. By regarding training ResNet as an optimal control problem, [95] discovered that the BP-based optimization algorithm could be viewed as an iterative solver for the maximal principle of the optimal control problem. Based on this observation, [98, 171] designed new accelerated training algorithms for ODE-Nets inspired by the theory of optimal control. Although the structural similarity between numerical ODEs and CNNs is mostly formal, theoretical analysis regarding the depth limit of ODE-Nets has become a vibrant and fast-moving field of research [43, 106, 121, 125, 153].

5.2. PDE-Nets: exploring structural similarity between numerical PDEs and CNNs

The original motivation of the work PDE-Nets [101, 102] was to design transparent CNNs to uncover hidden PDE models from observed dynamical data with minor prior knowledge on the mechanisms of the dynamics and to perform accurate predictions at the same time. Learning PDEs from observation or measurement data is a typical task in inverse problems in which machine learning methods have recently attracted tremendous attention [5]. However, existing CNNs designed for computer vision tasks primarily emphasize prediction accuracy. They are generally considered black-boxes and cannot reveal the hidden PDE model that drives the dynamical data. Therefore, we need to carefully design the CNN by exploring the structure similarity between numerical PDEs and CNNs.

Assume that the PDE to be uncovered takes the following generic form:

$$u_t = F(u, \boldsymbol{D}u), \quad x \in \Omega \subset \mathbb{R}^2,\ t \in [0, T],$$

where $\boldsymbol{D}$ was defined in (4.6). In a nutshell, PDE-Nets are designed as feedforward networks by discretizing the above PDE using forward-Euler (or any other temporal discretization) in time and finite-difference in space. The forward-Euler approximation of the temporal derivative makes PDE-Nets residual-type neural networks. As has been extensively discussed in Section 4, the finite-difference approximation to the differential operator $\boldsymbol{D}$ can realized by convolutions with properly chosen convolution kernels (i.e., filters). In fact, not only finite-difference approximations can be realized by convolutions, any discretization of $\boldsymbol{D}$ based on an approximation of u using translation-invariant basis functions can also be realized by convolutions [39]. The nonlinear response function F is approximated by a symbolic neural network denoted as SymNet (or a regular DNN as in [102] that is more expressive but less interpretable). Let $\boldsymbol{u}^{k+1}$ be the predicted value at time $t_k + \Delta t$ based on $\boldsymbol{u}^k$. Then, the PDE-Nets take the following dynamical form:

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \Delta t \cdot \mathrm{SymNet}_m^n(Q_{00}\boldsymbol{u}^k, Q_{01}\boldsymbol{u}^k, Q_{10}\boldsymbol{u}^k, \dots), \quad k = 0, 1, \dots, K-1. \tag{5.4}$$

Here, the operators $\{Q_{ij}\}$ denote convolution operators with the underlying filters denoted by $\boldsymbol{q}_{ij}$, i.e., $Q_{ij}u = \boldsymbol{q}_{ij} \circledast \boldsymbol{u}$. The operators Q_{10}, Q_{01}, Q_{11}, etc., approximate differential operators, i.e., $Q_{ij}u \approx \frac{\partial^{i+j}\boldsymbol{u}}{\partial^i x \partial^j y}$. In particular, Q_{00} is a certain averaging operator. The symbolic

neural network SymNet_m^n is introduced to approximate the multivariate nonlinear response function F. The design of SymNet_m^n is motivated by [135]. It can accurately estimate function F that is formed or can be well approximated by multivariate polynomials. Details on SymNet_m^n and its properties can be found in [101]. All the parameters of the SymNet and the filters $\{\boldsymbol{q}_{ij}\}$ are jointly learned from data.

A key difference from existing works (e.g., [14,21,100,133,137,138,167]) on discovering PDE models from observation data prior to [101,102] is that the filters corresponding to the specific finite-difference approximations to $\boldsymbol{D}$ are learned jointly with the estimation of the nonlinear response function F. The benefits of doing such joint learning in both system identification and prediction were empirically demonstrated in [101]. More importantly, in order to grant desired interpretability to the PDE-Nets, proper constraints are enforced on the filters. These constraints are motivated from Proposition 4.1 which we now elaborate.

In [101,102], the moment matrix associated to a given filter $\boldsymbol{q}$ was introduced to easily enforce constraints on the filter during training. Recall that the moment matrix $M(\boldsymbol{q})$ of an $N \times N$ filter $\boldsymbol{q}$ is defined by

$$M(\boldsymbol{q}) = (m_{i,j})_{N\times N}, \tag{5.5}$$

where

$$m_{i,j} = \frac{1}{i!j!} \sum_{k_1,k_2=-\frac{N-1}{2}}^{\frac{N-1}{2}} k_1^i k_2^j \boldsymbol{q}[k_1,k_2], \quad i,j = 0,1,\ldots,N-1. \tag{5.6}$$

Then, by examining (5.6), (4.1), and Proposition 4.1, it is not hard to see that, with a properly chosen N, filter $\boldsymbol{q}$ can be designed to approximate any differential operator with prescribed order of accuracy by imposing constraints on $M(q)$.

For example, if we want to approximate $\frac{\partial u}{\partial x}$ (up to a constant) by convolution $\boldsymbol{q} \circledast \boldsymbol{u}$ where $\boldsymbol{q}$ is a 3×3 filter and $\boldsymbol{u}$ is the evaluation of u on a regular grid, we can consider the following constrains on $M(q)$:

$$\begin{pmatrix} 0 & 0 & \star \\ 1 & \star & \star \\ \star & \star & \star \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & \star \\ 0 & \star & \star \end{pmatrix}. \tag{5.7}$$

Here, $\star$ means no constraint on the corresponding entry which allows one degree of freedom for learning. The constraints described by the moment matrix on the left of (5.7) guarantee that the approximation accuracy is at least of first order, and that on the right guarantees an approximation of at least second order. In particular, when all entries of $M(q)$ are constrained, the corresponding filter is uniquely determined. In the PDE-Nets, all filters are learned subjected to partial constraints on their associated moment matrices. Similar ideas on learning constrained filters to approximate differential operators were later used in [9] to design data-driven solvers for PDEs, and in [31] to design data-driven discretizations for total variations. A more extended discussion on the connections between numerical PDEs and neural networks was given in [2].

Exploiting the links between PDEs and CNNs has become a popular line of research that has led to many new designs of CNN models for machine learning and computer vision

tasks [4, 76, 113, 134, 152, 175]. It can also be used to improve the efficiency of CNNs [57]. On the other hand, this line of research also drives the development of data-driven modeling in scientific computing including efficient solvers for PDEs [9, 11, 39, 40, 85, 105, 158], model reduction of complex systems [109, 112, 126, 165, 166, 169], system identification from observation or simulation data [8, 15, 40, 75, 81, 83, 103, 124, 132], control of physical systems [78, 157], inverse problems [5, 61, 62, 86], and applications in seismology [88]. In addition, building PDE models on unstructured data for machine learning and scientific computing tasks is now an emerging branch of research [1, 29, 56, 82, 139].

FUNDING

This work was partially supported by the National Natural Science Foundation of China (grant No. 12090022), Beijing Natural Science Foundation (grant No. 180001) and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] F. Alet, A. K. Jeewajee, M. B. Villalonga, A. Rodriguez, T. Lozano-Perez, and L. Kaelbling, Graph element networks: adaptive, structured computation and memory. In *International Conference on Machine Learning*, pp. 212–222, Proceedings of Machine Learning Research, 2019.

[2] T. Alt, K. Schrader, M. Augustin, P. Peter, and J. Weickert, Connections between numerical algorithms for PDEs and neural networks. 2021, arXiv:2107.14742.

[3] L. Alvarez, F. Guichard, P. Lions, and J. Morel, Axioms and fundamental equations of image processing. *Arch. Ration. Mech. Anal.* **123** (1993), no. 3, 199–257.

[4] S. Arridge and A. Hauptmann, Networks for nonlinear diffusion problems in imaging. *J. Math. Imaging Vision* **62** (2020), no. 3, 471–487.

[5] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, Solving inverse problems using data-driven models. *Acta Numer.* **28** (2019), 1–174.

[6] U. M. Ascher and L. R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM: Society for Industrial and Applied Mathematics, 1997.

[7] G. Aubert and P. Kornprobst, *Mathematical problems in image processing: partial differential equations and the calculus of variations*. Appl. Math. Sci. 147, Springer, New York, 2006.

[8] I. Ayed, E. de Bézenac, A. Pajot, J. Brajard, and P. Gallinari, Learning dynamical systems from partial observations. 2019, arXiv:1902.11136.

[9] Y. Bar-Sinai, S. Hoyer, J. Hickey, and M. P. Brenner, Learning data-driven discretizations for partial differential equations. *Proc. Natl. Acad. Sci. USA* **116** (2019), no. 31, 15344–15349.

[10] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** (2009), no. 1, 183–202.

[11] C. Beck, S. Becker, P. Cheridito, A. Jentzen, and A. Neufeld, Deep splitting method for parabolic PDEs. 2019, arXiv:1907.03452

[12] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, Navier–Stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. I–I, 1, IEEE, 2001.

[13] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 417–424, ACM Press/Addison-Wesley Publishing Co., 2000.

[14] J. Bongard and H. Lipson, Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **104** (2007), no. 24, 9943–9948.

[15] G.-J. Both, S. Choudhury, P. Sens, and R. Kusters, Deepmod: deep learning for model discovery in noisy data. *J. Comput. Phys.* **428** (2021), 109985.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Signal Process.* **3** (2011), no. 1, 1–122.

[17] A. Braides, *Gamma-convergence for beginners*, Oxford Lecture Ser. Math. Appl. 22, Oxford University Press, 2002.

[18] K. Bredies, K. Kunisch, and T. Pock, Total generalized variation. *SIAM J. Imaging Sci.* **3** (2010), 492.

[19] R. W. Brown, E. M. Haacke, Y.-C. N. Cheng, M. R. Thompson, and R. Venkatesan, *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[20] R. E. Bruck Jr, On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *J. Math. Anal. Appl.* **61** (1977), no. 1, 159–164.

[21] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* (2016), 201517384.

[22] T. M. Buzug, *Computed tomography: from photon statistics to modern cone-beam CT*. Springer, Berlin, 2008.

[23] J. Cai, R. Chan, and Z. Shen, A framelet-based image inpainting algorithm. *Appl. Comput. Harmon. Anal.* **24** (2008), no. 2, 131–149.

[24] J. Cai, B. Dong, S. Osher, and Z. Shen, Image restorations: total variation, wavelet frames and beyond. *J. Amer. Math. Soc.* **25** (2012), no. 4, 1033–1089.

[25] J. Cai, S. Osher, and Z. Shen, Split Bregman methods and frame based image restoration. *Multiscale Model. Simul.* **8** (2009), no. 2, 337–369.

[26] J.-F. Cai, B. Dong, and Z. Shen, Image restoration: a wavelet frame based model for piecewise smooth functions and beyond. *Appl. Comput. Harmon. Anal.* **41** (2016), no. 1, 94–138.

[27] X. Cai, R. Chan, S. Morigi, and F. Sgallari, Vessel segmentation in medical imaging using a tight-frame–based algorithm. *SIAM J. Imaging Sci.* **6** (2013), no. 1, 464–486.

[28] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.* **29** (1992), no. 1, 182–193.

[29] B. P. Chamberlain, J. Rowbottom, M. Gorinova, S. Webb, E. Rossi, and M. M. Bronstein, GRAND: graph neural diffusion. 2021, arXiv:2106.10934.

[30] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40** (2011), no. 1, 120–145.

[31] A. Chambolle and T. Pock, Learning consistent discretizations of the total variation. *SIAM J. Imaging Sci.* **14** (2021), no. 2, 778–813.

[32] R. Chan, T. Chan, L. Shen, and Z. Shen, Wavelet algorithms for high-resolution image reconstruction. *SIAM J. Sci. Comput.* **24** (2003), no. 4, 1408–1432.

[33] T. F. Chan and J. Shen, *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*. SIAM, 2005.

[34] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, and E. Holtham, Reversible architectures for arbitrarily deep residual neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

[35] G. H. Chen and R. Rockafellar, Convergence rates in forward–backward splitting. *SIAM J. Optim.* **7** (1997), no. 2, 421–444.

[36] L. Chen and C. Wu, A note on the expressive power of deep rectified linear unit networks in high-dimensional spaces. *Math. Methods Appl. Sci.* **42** (2019), no. 9, 3400–3404.

[37] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6572–6583, 2018.

[38] T. Chen, X. Chen, W. Chen, H. Heaton, J. Liu, Z. Wang, and W. Yin, Learning to optimize: a primer and a benchmark. 2021, arXiv:2103.12828.

[39] Y. Chen, B. Dong, and J. Xu, Meta-MgNet: Meta multigrid networks for solving parameterized partial differential equations. 2020, arXiv:2010.14088.

[40] Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart, Solving and learning nonlinear PDEs with Gaussian processes. 2021, arXiv:2103.12959.

[41] Z. Chen, J. Zhang, M. Arjovsky, and L. Bottou, Symplectic recurrent neural networks. 2019, arXiv:1909.13334.

[42] J. K. Choi, B. Dong, and X. Zhang, An edge driven wavelet frame model for image restoration. *Appl. Comput. Harmon. Anal.* **48** (2020), no. 3, 993–1029.

[43] A. Cohen, R. Cont, A. Rossier, and R. Xu, Scaling properties of deep residual networks. 2021, arXiv:2105.12245.

[44] P. Combettes and V. Wajs, Signal recovery by proximal forward–backward splitting. *Multiscale Model. Simul.* **4** (2006), no. 4, 1168–1200.

[45] I. Daubechies, *Ten lectures on wavelets*. CBMS-NSF Lecture Notes, SIAM, 61, Society for Industrial and Applied Mathematics, 1992.

[46] I. Daubechies, B. Han, A. Ron, and Z. Shen, Framelets: MRA-based constructions of wavelet frames. *Appl. Comput. Harmon. Anal.* **14** (2003), no. 1, 1–46.

[47] I. Daubechies, G. Teschke, and L. Vese, Iteratively solving linear inverse problems under general convex constraints. *Inverse Probl. Imaging* **1** (2007), no. 1, 29.

[48] B. Dong, A. Chien, and Z. Shen, Frame based segmentation for medical images. *Commun. Math. Sci.* **9** (2010), no. 2, 551–559.

[49] B. Dong, Q. Jiang, and Z. Shen, Image restoration: wavelet frame shrinkage, nonlinear evolution PDEs, and beyond. *Multiscale Model. Simul.* **15** (2017), no. 1, 606–660.

[50] B. Dong and Z. Shen, Frame based surface reconstruction from unorganized points. *J. Comput. Phys.* **230** (2011), 8247–8255.

[51] B. Dong and Z. Shen, Image restoration: a data-driven perspective. In *Proceedings of the International Congress of Industrial and Applied Mathematics (ICIAM)*, pp. 65–108, Citeseer, 2015.

[52] B. Dong, Z. Shen, and P. Xie, Image restoration: a general wavelet frame based model and its asymptotic analysis. *SIAM J. Math. Anal.* **49** (2017), no. 1, 421–445.

[53] B. Dong, Z. Shen et al., *MRA-based wavelet frames and applications*. Summer Program on "The Mathematics of Image Processing". IAS Lect. Notes Ser. 19, Park City Mathematics Institute, 2010.

[54] D. L. Donoho, De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **41** (1995), no. 3, 613–627.

[55] M. Elad, J. Starck, P. Querre, and D. Donoho, Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmon. Anal.* **19** (2005), no. 3, 340–358.

[56] M. Eliasof, E. Haber, and E. Treister, PDE-GCN: novel architectures for graph neural networks motivated by partial differential equations. 2021, arXiv:2108.01938.

[57] J. Ephrath, M. Eliasof, L. Ruthotto, E. Haber, and E. Treister, Leanconvnets: low-cost yet effective convolutional neural networks. *IEEE J. Sel. Top. Signal Process.* **14** (2020), no. 4, 894–904.

[58] E. Esser, X. Zhang, and T. F. Chan, A general framework for a class of first order primal–dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.* **3** (2010), no. 4, 1015–1046.

[59] M. Fadili and J. Starck, Sparse representations and Bayesian image inpainting. *Proc. SPARS* **5** (2005).

[60] M. Fadili, J. Starck, and F. Murtagh, Inpainting and zooming using sparse representations. *Comput. J.* **52** (2009), no. 1, 64.

[61] Y. Fan and L. Ying, Solving inverse wave scattering with deep learning. 2019, arXiv:1911.13202.

[62] Y. Fan and L. Ying, Solving electrical impedance tomography with deep learning. *J. Comput. Phys.* **404** (2020), 109119.

[63] M. Figueiredo and R. Nowak, An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.* **12** (2003), no. 8, 906–916.

[64] M. Figueiredo and R. Nowak, A bound optimization approach to wavelet-based image deconvolution. In *IEEE International Conference on Image Processing, 2005. ICIP 2005*, pp. II–782, 2, IEEE, 2005.

[65] K. Fukushima, Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernet.* **36** (1980), no. 4, 193–202.

[66] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* **2** (1976), no. 1, 17–40.

[67] X. Gastaldi, Shake-shake regularization. 2017, arXiv:1705.07485.

[68] R. Glowinski and A. Marroco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Rev. Fr. Autom. Inform. Rech. Opér., Anal. Numér.* **9** (1975), no. R2, 41–76.

[69] T. Goldstein and S. Osher, The split Bregman method for l_1-regularized problems. *SIAM J. Imaging Sci.* **2** (2009), no. 2, 323–343.

[70] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, The reversible residual network: backpropagation without storing activations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2211–2221, 2017.

[71] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, Cambridge, 2016.

[72] K. Gröchenig, *Foundations of time-frequency analysis*. Birkhäuser, 2001.

[73] V. L. Guen and N. Thome, Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020.

[74] E. Haber and L. Ruthotto, Stable architectures for deep neural networks. *Inverse Probl.* **34** (2017), no. 1, 014004.

[75] J. Han, C. Ma, Z. Ma, and E. Weinan, Uniformly accurate machine learning-based hydrodynamic models for kinetic equations. *Proc. Natl. Acad. Sci. USA* **116** (2019), no. 44, 21983–21991.

[76] J. He and J. Xu, Mgnet: a unified framework of multigrid and convolutional neural network. *Sci. China Math.* **62** (2019), no. 7, 1331–1354.

[77] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[78] P. Holl, N. Thuerey, and V. Koltun, Learning to control PDEs with differentiable physics. 2020, arXiv:2001.07457

[79] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

[80] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, Deep networks with stochastic depth. In *European Conference on Computer Vision*, pp. 646–661, Springer, 2016.

[81] J. Huang, Z. Ma, Y. Zhou, and W.-A. Yong, Learning thermodynamically stable and galilean invariant partial differential equations for non-equilibrium flows. *J. Non-Equilib. Thermodyn.* (2021).

[82] V. Iakovlev, M. Heinonen, and H. Lähdesmäki, Learning continuous-time PDEs from sparse data with graph neural networks. 2020, arXiv:2006.08956.

[83] J. Jia and A. R. Benson, Neural jump stochastic differential equations. *Adv. Neural Inf. Process. Syst.* **32** (2019), 9847–9858.

[84] Q. Jiang, Correspondence between frame shrinkage and high order nonlinear diffusion. *Appl. Numer. Math.* (2011).

[85] Y. Khoo, J. Lu, and L. Ying, Solving parametric PDE problems with artificial neural networks. *European J. Appl. Math.* **32** (2021), no. 3, 421–435.

[86] Y. Khoo and L. Ying, Switchnet: a neural network model for forward and inverse scattering problems. *SIAM J. Sci. Comput.* **41** (2019), no. 5, A3182–A3201.

[87] P. Kidger, J. Morrill, J. Foster, and T. Lyons, Neural controlled differential equations for irregular time series. 2020, arXiv:2005.08926.

[88] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft, Machine learning in seismology: turning data into insights. *Seismol. Res. Lett.* **90** (2019), no. 1, 3–14.

[89] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25** (2012), 1097–1105.

[90] G. Larsson, M. Maire, and G. Shakhnarovich, FractalNet: ultra-deep neural networks without residuals. 2016, arXiv:1605.07648.

[91] Y. Lecun, Y. Bengio, and G. Hinton, Deep learning. *Nature* **521** (2015), 436–444.

[92] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **2** (1989).

[93] M. Li, Z. Fan, H. Ji, and Z. Shen, Wavelet frame based algorithm for 3D reconstruction in electron microscopy. *SIAM J. Sci. Comput.* **36** (2014), no. 1, B45–B69.

[94] M. Li, L. He, and Z. Lin, Implicit Euler skip connections: enhancing adversarial robustness via numerical stability. In *International Conference on Machine Learning*, pp. 5874–5883, Proceedings of Machine Learning Research, 2020.

[95] Q. Li, L. Chen, and C. Tai, Maximum principle based algorithms for deep learning. *J. Mach. Learn. Res.* **18** (2018), 1–29.

[96] Q. Li, T. Lin, and Z. Shen, Deep learning via dynamical systems: an approximation perspective. *J. Eur. Math. Soc. (JEMS)* (to appear).

[97] X. Li, T.-K. L. Wong, R. T. Chen, and D. Duvenaud, Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882, Proceedings of Machine Learning Research, 2020.

[98] G.-H. Liu, T. Chen, and E. A. Theodorou, Differential dynamic programming neural optimizer. 2020, arXiv:2002.08809.

[99] J. Liu, X. Zhang, B. Dong, Z. Shen, and L. Gu, A wavelet frame method with shape prior for ultrasound video segmentation. *SIAM J. Imaging Sci.* **9** (2016), no. 2, 495–519.

[100] R. Liu, Z. Lin, W. Zhang, and Z. Su, Learning PDEs for image restoration via optimal control. In *European Conference on Computer Vision*, pp. 115–128, Springer, 2010.

[101] Z. Long, Y. Lu, and B. Dong, PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. *J. Comput. Phys.* (2019), 108925.

[102] Z. Long, Y. Lu, X. Ma, and B. Dong, PDE-Net: Learning PDEs from data. In *International Conference on Machine Learning*, pp. 3214–3222, 2018.

[103] F. Lu, M. Maggioni, and S. Tang, Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. 2019, arXiv:1910.04832.

[104] J. Lu, Z. Shen, H. Yang, and S. Zhang, Deep network approximation for smooth functions. 2020, arXiv:2001.03040.

[105] L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis, Deepxde: a deep learning library for solving differential equations. *SIAM Rev.* **63** (2021), no. 1, 208–228.

[106] Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying, A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pp. 6426–6436, Proceedings of Machine Learning Research, 2020.

[107] Y. Lu, A. Zhong, Q. Li, and B. Dong, Beyond finite layer neural networks: bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, pp. 3276–3285, 2018.

[108] Z. Luo, Z. Sun, W. Zhou, and S.-i. Kamata, Rethinking ResNets: improved stacking strategies with high order schemes. 2021, arXiv:2103.15244.

[109] M. Lutter, C. Ritter, and J. Peters, Deep Lagrangian networks: using physics as model prior for deep learning. 2019, arXiv:1907.04490.

[110] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Academic Press, 2008.

[111] Y. Mao, S. Osher, and B. Dong, A nonlinear PDE-based method for sparse deconvolution. *Multiscale Model. Simul.* **8** (2010), no. 3.

[112] A. Mohan, D. Daniel, M. Chertkov, and D. Livescu, Compressed convolutional LSTM: an efficient deep learning framework to model high fidelity 3D turbulence. 2019, arXiv:1903.00033.

[113] V. Monga, Y. Li, and Y. C. Eldar, Algorithm unrolling: interpretable, efficient deep learning for signal and image processing. *IEEE Signal Process. Mag.* **38** (2021), no. 2, 18–44.

[114] H. Montanelli, H. Yang, and Q. Du, Deep ReLu networks overcome the curse of dimensionality for bandlimited functions. 2019, arXiv:1903.00735.

[115] D. Mumford and J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.* **42** (1989), no. 5, 577–685.

[116] Y. Nesterov, A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Sov. Math., Dokl.* **27** (1983), no. 2, 372–376.

[117] Y. Nesterov, On an approach to the construction of optimal methods for minimizing smooth convex functions. *Èkon. Mat. Metody* **24** (1988), no. 3, 509–517.

[118] B. Oksendal, *Stochastic differential equations: an introduction with applications*. Springer, Berlin, 2013.

[119] S. Osher and R. Fedkiw, *Level set methods and dynamic implicit surfaces*. Appl. Math. Sci. 153, Springer, New York, 2003.

[120] S. Osher and L. Rudin, Feature-oriented image enhancement using shock filters. *SIAM J. Numer. Anal.* **27** (1990), no. 4, 919–940.

[121] K. Ott, P. Katiyar, P. Hennig, and M. Tiemann, ResNet after all: neural ODEs and their numerical solution. In *International Conference on Learning Representations*, 2020.

[122] G. B. Passty, Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.* **72** (1979), 383–290.

[123] P. Perona and J. Malik, Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** (1990), no. 7, 629–639.

[124] T. Qin, K. Wu, and D. Xiu, Data driven governing equations approximation using deep neural networks. *J. Comput. Phys.* **395** (2019), 620–635.

[125] A. F. Queiruga, N. B. Erichson, D. Taylor, and M. W. Mahoney, Continuous-in-depth neural networks. 2020, arXiv:2008.02389.

[126] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, A. Ramadhan, and A. Edelman, Universal differential equations for scientific machine learning. 2020, arXiv:2001.04385.

[127] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Multistep neural networks for data-driven discovery of nonlinear dynamical systems. 2018, arXiv:1801.01236.

[128] A. Ron and Z. Shen, Affine systems in $L_2(\mathbb{R}^d)$ II: dual systems. *J. Fourier Anal. Appl.* **3** (1997), no. 5, 617–638.

[129] A. Ron and Z. Shen, Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator. *J. Funct. Anal.* **148** (1997), no. 2, 408–447.

[130] O. Ronneberger, P. Fischer, and T. Brox, U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[131] L. Rudin, S. Osher, and E. Fatemi, Nonlinear total variation based noise removal algorithms. *Phys. D* **60** (1992), 259–268.

[132] S. Rudy, A. Alla, S. L. Brunton, and J. N. Kutz, Data-driven identification of parametric partial differential equations. *SIAM J. Appl. Dyn. Syst.* **18** (2019), no. 2, 643–660.

[133] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Data-driven discovery of partial differential equations. *Sci. Adv.* **3** (2017), no. 4, e1602614.

[134] L. Ruthotto and E. Haber, Deep neural networks motivated by partial differential equations. *J. Math. Imaging Vision* **62** (2020), no. 3, 352–364.

[135] S. Sahoo, C. Lampert, and G. Martius, Learning equations for extrapolation and control. In *International Conference on Machine Learning*, pp. 4442–4450, Proceedings of Machine Learning Research, 2018.

[136] G. Sapiro, *Geometric partial differential equations and image analysis*. Cambridge University Press, 2001.

[137] H. Schaeffer, Learning partial differential equations via data discovery and sparse optimization. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **473** (2017), no. 2197, 20160446.

[138] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data. *Science* **324** (2009), no. 5923, 81–85.

[139] S. Seo, C. Meng, and Y. Liu, Physics-aware difference graph networks for sparsely-observed dynamics. In *International Conference on Learning Representations*, 2019.

[140] Z. Shen, Wavelet frames and image restorations. In *Proceedings of the International Congress of Mathematicians*, Vol. 4, pp. 2834–2863, World Scientific, 2010.

[141] Z. Shen, K.-C. Toh, and S. Yun, An accelerated proximal gradient algorithm for frame-based image restoration via the balanced approach. *SIAM J. Imaging Sci.* **4** (2011), no. 2, 573–596.

[142] Z. Shen, H. Yang, and S. Zhang, Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Comput.* **33** (2021), no. 4, 1005–1036.

[143] Z. Shen, H. Yang, and S. Zhang, Neural network approximation: three hidden layers are enough. *Neural Netw.* **141** (2021), 160–173.

[144] J. W. Siegel and J. Xu, Optimal approximation rates and metric entropy of ReLU^k and cosine networks. 2021, arXiv:2101.12365.

[145] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., Mastering the game of Go with deep neural networks and tree search. *Nature* **529** (2016), no. 7587, 484–489.

[146] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** (2014), no. 1, 1929–1958.

[147] J. Starck, M. Elad, and D. Donoho, Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Process.* **14** (2005), no. 10, 1570–1582.

[148] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk, On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides. *SIAM J. Numer. Anal.* (2005), 686–713.

[149] W. Su, S. Boyd, and E. Candes, A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Adv. Neural Inf. Process. Syst.* **27** (2014), 2510–2518.

[150] W. Su, S. Boyd, and E. J. Candes, A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res.* **17** (2016), no. 153, 1–43.

[151] Q. Sun, Y. Tao, and Q. Du, Stochastic training of residual networks: a differential equation viewpoint. 2018, arXiv:1812.00174.

[152] Y. Sun, L. Zhang, and H. Schaeffer, Neupde: neural network based ordinary and partial differential equations for modeling time-dependent data. In *Mathematical and Scientific Machine Learning*, pp. 352–372, Proceedings of Machine Learning Research, 2020.

[153] M. Thorpe and Y. van Gennip, Deep limits of residual neural networks. 2018, arXiv:1810.11741.

[154] A. Tikhonov, V. Arsenin, and F. John, *Solutions of ill-posed problems*. VH Winston, Washington, DC, 1977.

[155] P. Tseng, Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.* **29** (1991), no. 1, 119–138.

[156] H. Wang, Y. Wu, M. Li, Q. Zhao, and D. Meng, A survey on rain removal from video and single image. 2019, arXiv:1909.08326.

[157] W. Wang, S. Axelrod, and R. Gómez-Bombarelli, Differentiable molecular simulations for control and learning. 2020, arXiv:2003.00868.

[158] Y. Wang, Learning to discretize: solving 1D scalar conservation laws via deep reinforcement learning. *Commun. Comput. Phys.* **28** (2020), no. 5, 2158–2179.

[159] Z. Wang, J. Chen, and S. C. Hoi, Deep learning for image super-resolution: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43** (2020), 3365–3387.

[160] R. F. Warming and B. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods. *J. Comput. Phys.* **14** (1974), no. 2, 159–179.

[161] J. Weickert, *Anisotropic diffusion in image processing*. Teubner, Stuttgart, 1998.

[162] E. Weinan, A proposal on machine learning via dynamical systems. *Commun. Math. Stat.* **5** (2017), no. 1, 1–11.

[163] E. Weinan, C. Ma, and L. Wu, A priori estimates of the population risk for two-layer neural networks. *Commun. Math. Sci.* **17** (2019), no. 5, 1407–1425.

[164] E. Weinan and Q. Wang, Exponential convergence of the deep neural network approximation for analytic functions. *Sci. China Math.* **61** (2018), no. 10, 1733–1740.

[165] S. Wiewel, M. Becher, and N. Thuerey, Latent space physics: towards learning the temporal evolution of fluid flow. *Comput. Graph. Forum* **38** (2019), 71–82. Wiley Online Library.

[166] J.-L. Wu, K. Kashinath, A. Albert, D. Chirila, H. Xiao et al., Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems. *J. Comput. Phys.* **406** (2020), 109209.

[167] Z. Wu and R. Zhang, Learning physics by data for the motion of a sphere falling in a non-Newtonian fluid. *Commun. Nonlinear Sci. Numer. Simul.* **67** (2019), 577–593.

[168] X. Xie, F. Bao, T. Maier, and C. Webster, Analytic continuation of noisy data using Adams Bashforth residual neural network. *Discrete Contin. Dyn. Syst. Ser. S* (2021). DOI 10.3934/dcdss.2021088.

[169] Y. Xie, E. Franz, M. Chu, and N. Thuerey, tempoGAN: a temporally coherent, volumetric GAN for super-resolution fluid flow. *ACM Trans. Graph.* **37** (2018), no. 4, 1–15.

[170] D. Yarotsky and A. Zhevnerchuk, The phase diagram of approximation rates for deep neural networks. 2019, arXiv:1906.09477.

[171] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong, You only propagate once: accelerating adversarial training via maximal principle. *Adv. Neural Inf. Process. Syst.* **32** (2019), 227–238.

[172] H.-M. Zhang and B. Dong, A review on deep learning in medical image reconstruction. *J. Oper. Res. Soc. China* (2020), 1–30.

[173] J. Zhang, B. Han, L. Wynter, L. K. Hsiang, and M. Kankanhalli, Towards robust ResNet: a small step but a giant leap. In *28th International Joint Conference on Artificial Intelligence*, 2019.

[174] X. Zhang, Z. Li, C. Change Loy, and D. Lin, Polynet: a pursuit of structural diversity in very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 718–726, 2017.

[175] X. Zhang, J. Liu, Y. Lu, and B. Dong, Dynamically unfolding recurrent restorer: a moving endpoint control method for image restoration. In *International Conference on Learning Representations*, 2019.

[176] M. Zhu and T. Chan, An efficient primal–dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report* **34** (2008).

[177] M. Zhu, B. Chang, and C. Fu, Convolutional neural networks combined with Runge–Kutta methods. 2018, arXiv:1802.08831.

BIN DONG

Beijing International Center for Mathematical Research, National Engineering Laboratory for Big Data Analysis and Applications, National Biomedical Imaging Center, Institute for Artificial Intelligence, Peking University, Beijing, China, dongbin@math.pku.edu.cn

THEORY OF GRAPH NEURAL NETWORKS: REPRESENTATION AND LEARNING

STEFANIE JEGELKA

ABSTRACT

Graph Neural Networks (GNNs), neural network architectures targeted to learning representations of graphs, have become a popular learning model for prediction tasks on nodes, graphs and configurations of points, with wide success in practice. This article summarizes a selection of emerging theoretical results on approximation and learning properties of widely used message passing GNNs and higher-order GNNs, focusing on representation, generalization, and extrapolation. Along the way, it summarizes broad mathematical connections.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68T05; Secondary 05C62, 05C60, 05C90, 68Q32, 68R10

KEYWORDS

Machine learning, graph representation learning, graph isomorphism, approximation theory, learning theory

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5450–5476
DOI 10.4171/ICM2022/162

1. INTRODUCTION

There has been growing interest in solving machine learning tasks when the input data is given in form of a graph $G = (V, E, X, W)$ from a set of attributed graphs $\mathcal{G}$, where $X \in \mathbb{R}^{d \times |V|}$ contains vectorial attributes for each node, and $W \in \mathbb{R}^{d_w \times |E|}$ contains attributes for each edge (X and W may be empty). Examples include predictions in social networks, recommender systems and link prediction (given two nodes, predict an edge), property prediction of molecules, prediction of drug interactions, traffic prediction, forecasting physics simulations, and learning combinatorial optimization algorithms for hard problems. These examples use two types of task: (1) given a graph G, predict a label $F(G)$; (2) given a graph G and node $v \in V(G)$, predict a node label $f(v)$. An edge may be similarly predicted, but from two nodes instead of one.

Solving these tasks demands a sufficiently rich *embedding* of the graph or each node that captures structural properties as well as the attribute information. While graph embeddings have been a widely studied topic, including spectral embeddings and graph kernels, recently, *Graph Neural Networks (GNNs)* [36,37,39,49,65,83] have emerged as an empirically broadly successful model class that, as opposed to, e.g., spectral embeddings, allows adapting the embedding to the task at hand, generalizes to other graphs of the same input type, and incorporates attributes. Due to space limits, this survey focuses on the popular message passing (spatial) GNNs, formally defined below, and their rich mathematical connections, with an excursion into higher-order GNNs.

When *learning* a GNN, we observe N i.i.d. samples $\mathcal{D} = \{G_i, y_i\}_{i=1}^N \in (\mathcal{G} \times \mathcal{Y})^N$ drawn from an underlying distribution $\mathcal{P}$ on $\mathcal{G} \times \mathcal{Y}$. The *labels* y_i are often given by an unknown *target function* $g(G_i)$, and observed with or without i.i.d. noise. Given a (convex) loss function $\ell : \mathcal{G} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures prediction error, i.e., mismatch of y and $F(G)$, such as the squared loss or cross-entropy, we aim to estimate a model F from our GNN model class $\mathcal{F}$ to minimize the expected loss (*population risk*) $\mathcal{R}(F)$:

$$\min_{F \in \mathcal{F}} \mathbb{E}_{(G,y)\sim\mathcal{P}}\big[\ell\big(G, y, F(G)\big)\big] \equiv \min_{F \in \mathcal{F}} \mathcal{R}(F). \tag{1.1}$$

When analyzing this quantity, three main questions become important:

1. Representational power (Section 2). Which target functions g can be approximated well by a GNN model class $\mathcal{F}$? Answers to this question relate to graph isomorphism testing, approximation theory for neural networks, local algorithms and representing invariance/equivariance under permutations.

2. Generalization (Section 3). Even with sufficient approximation power, we can only estimate a function $\hat{F} \in \mathcal{F}$ from the data sample $\mathcal{D}$. The common learning or *training* procedure is to instead minimize the *empirical risk* $\widehat{\mathcal{R}}(F)$:

$$\hat{F} \in \arg\min_{F \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \ell\big(G_i, y_i, F(G_i)\big) \equiv \arg\min_{F \in \mathcal{F}} \widehat{\mathcal{R}}(F). \tag{1.2}$$

Generalization asks how well $\hat{F}$ is performing according to the population risk, i.e., $\widehat{\mathcal{R}}(\hat{F})$, as a function of N and model properties. Good generalization may demand explicit (e.g., via penalties) or implicit regularization (e.g., via the optimization algorithm, typically variants

of stochastic gradient descent). Hence, generalization analyses involve the complexity of the model class $\mathcal{F}$, the target function, the data and the optimization procedure.

3. Generalization under distribution shifts (Section 4). In practice, a learned model $\hat{F}$ is often deployed on data from a distribution $\mathcal{Q} \neq \mathcal{P}$, e.g., graphs of different size, degree or attribute ranges so that for instance $\operatorname{supp}(Q) \supset \operatorname{supp}(P)$. In which cases can we expect successful *extrapolation* to $\mathcal{Q}$? This depends on the structure of the graphs and the task, formalizable via graph limits, local structures and algorithmic structures, e.g., dynamic programming.

Beyond these topics, GNNs have close connections to graph signal processing as learnable filters, to geometric learning and probabilistic inference.

1.1. Graph Neural Networks (GNNs)

Message passing graph neural networks (MPNNs) follow an iterative scheme [36, 37, 39, 49, 65, 83]. Throughout, they maintain a representation (embedding) $h_v^{(t)} \in \mathbb{R}^{d_t}$ for each node $v \in V$. In each iteration t, we update each node v's embedding $h_v^{(k)}$ as a function of its neighbors' embeddings and possible edge attributes:

$$h_v^{(0)} = x_v, \quad \forall v \in V, \tag{1.3}$$

$$m_v^{(t)} = f_{\text{Agg}}^{(t)}\big(h_v^{(t-1)}, \{\!\{h_u^{(t-1)}, w(u,v) \mid u \in \mathcal{N}(v)\}\!\}\big), \quad 1 \leq t < T \quad \text{(Aggregate)}, \tag{1.4}$$

$$h_v^{(t)} = f_{\text{Up}}\big(h_v^{(t)}, m_v^{(t)}\big) \quad \text{(Update)}. \tag{1.5}$$

The final node representation $f(v) = h_v^{(T)}, \forall v \in V$ is the last iterate, possibly concatenated with a linear classifier. Here, $\mathcal{N}(v) \subset V$ denotes the neighborhood of $v \in V$, and $\{\!\{\cdot\}\!\}$ a multiset. Via the updates, $h_v^{(t)}$ encodes the t-hop neighborhood of node v, i.e., the subgraph of all nodes reachable from v within t steps. The number of iterations T is also termed the GNN *depth*, and one iteration may be viewed as a layer.

The *aggregation function* $f_{\text{Agg}}^{(t)} : \mathbb{R}^{d_{t-1}} \to \mathbb{R}^{d_t}$ plays a major role and is shared by all nodes within an iteration. It is a nonlinear function of the form

$$f_{\text{Agg}}^{(t)}\big(h_v^{(t-1)}, \{\!\{h_u^{(t-1)}, w(u,v) \mid u \in \mathcal{N}(v)\}\!\}\big) = \phi_1^{(t)}\Bigg(\sum_{u \in \mathcal{N}(v)} \phi_2^{(t)}\big(h_u^{(t)}, h_v^{(t)}, w(u,v)\big)\Bigg). \tag{1.6}$$

The sum may also be replaced by an average, degree-normalized sum or coordinate-wise min or max. In the most general form, the functions ϕ_1, ϕ_2 are implemented as *multilayer perceptrons (MLPs)*, neural networks that alternate linear transformations and coordinate-wise nonlinear activations such as the ReLU ($\sigma(a) = \max\{a, 0\}$) or sigmoid function:

$$\text{MLP}(h; \theta) = \sigma\big(W^{(M)} \cdots \sigma\big(W^{(2)}\sigma(W^{(1)}h + b^{(1)}) + b^{(2)}\big) \cdots + b^{(M)}\big). \tag{1.7}$$

The learnable parameters θ of the MLP are the weight matrices $W^{(j)}$ and bias vectors $b^{(j)}$. The update f_{Up} is typically a weighted combination with learnable weight matrices:

$$f_{\text{Up}}\big(h_v^{(t)}, m_v^{(t)}\big) = \sigma\big(W_1^{(t)}h_v^{(t)} + W_2^{(t)}m_v^{(t)}\big) \quad \text{or} \quad f_{\text{Up}}\big(h_v^{(t)}, m_v^{(t)}\big) = m_v^{(t)}. \tag{1.8}$$

Finally, if a graph-level prediction is desired, all node representations can be aggregated by a permutation invariant *readout* function

$$F(G) = f_{\text{Read}}\big(\{\!\{h_v^{(T)} \mid v \in V\}\!\}\big). \tag{1.9}$$

Here, we assume the readout has the form (1.6) or is a simple sum or average. Typically, all parameters are learned jointly via stochastic gradient descent minimizing the empirical risk.

Throughout this article, $n = |V|$ denotes the number of nodes and N the number of training data points.

Permutation invariance. An important property of GNNs is permutation invariance of the graph, and equivariance of the node representations. Let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix of a graph $G \in \mathcal{G}$, and $X \in \mathbb{R}^{n\times d}$ its node features. Permutation invariance/equivariance means that for all permutation matrices $P \in \mathbb{R}^{n\times n}$ and all $G \in \mathcal{G}$:

$$F(PAP^\top, PX) = F(A, X) \tag{1.10}$$

$$f(PAP^\top, PX, v) = f(A, X, v) \tag{1.11}$$

2. REPRESENTATIONAL POWER OF GNNS

For functions on graphs, representational power has mainly been studied in terms of graph isomorphism: which graphs a GNN can distinguish. Via variations of the Stone–Weierstrass theorem, these results yield universal approximation results. Other works bound the ability of GNNs to compute specific polynomials of the adjacency matrix and to distinguish graphons [28, 60]. Observed limitations of MPNNs have inspired higher-order GNNs (Section 2.3). Moreover, if all node attributes are unique, then analogies to local algorithms yield algorithmic approximation results and lower bounds (Section 2.2).

2.1. GNNs and graph isomorphism testing

A standard characterization of the discriminative power of GNNs is via the hierarchy of the *Weisfeiler–Leman (WL)* algorithm for graph isomorphism testing, also known as color refinement or vertex classification [75], which was inspired by the work of Weisfeiler and Leman [93, 94]. The WL algorithm does not entirely solve the graph isomorphism problem, but its power has been widely studied.

A *labeled* graph is a graph endowed with a node coloring $l : V(G) \to \Sigma$ for some sufficiently large alphabet Σ. Given a labeled graph (G, l), the 1-dimensional WL algorithm (1-WL) iteratively computes a node coloring $c_l^{(t)} : V(G) \to \Sigma$ for some sufficiently large alphabet Σ. Starting with $c_l^{(0)}$ in iteration $t = 0$, in iteration $t > 0$ it sets for all $v \in V$,

$$c_l^{(t)}(v) = \text{Hash}\big(c_l^{t-1}(v), \{\!\{c_l^{t-1}(u) \mid u \in \mathcal{N}(v)\}\!\}\big), \tag{2.1}$$

where Hash is an injective map from the input pair to Σ, i.e., it assigns a unique color to each neighborhood pattern. To compare two graphs G, G', the algorithm compares the multisets $\{\!\{c_l^{(t)}(v) \mid v \in V(G)\}\!\}$ and $\{\!\{c_l^{(t)}(u) \mid u \in V(G')\}\!\}$ in each iteration. If the sets differ, then it

determines that $G \neq G'$. Otherwise, it terminates when the number of colors in iteration t and $t-1$ are the same, which occurs after at most $\max\{|V(G)|, |V(G')|\}$ iterations.

The computational analogy between the 1-WL algorithm and MPNNs is obvious. Since the WL algorithm uniquely colors each neighborhood, the coloring $c_l^{(t)}(v)$ always *refines* the coloring $h_v^{(t)}$ from a GNN.

Theorem 1 ([66, 95]). *If for two graphs G, G' a message passing GNN outputs $f_G(G) \neq f_G(G')$, then the 1-WL algorithm will determine that $G \neq G'$.*

For any t, there exists an MPNN such that $c_l^{(t)} \equiv h^{(t)}$. A sufficient condition is that the aggregate, update, and readout operations are injective multiset functions.

GNNs that use the degree for normalization in the aggregation [49] can be equivalent to the 1-WL agorithm too, but with one more iteration in the WL algorithm [35].

2.1.1. Representing multiset functions

Theorem 1 demands the neighbor aggregation f_{Agg} to be an injective multiset function. Theorem 2 shows how to universally approximate multiset functions.

Theorem 2 ([92, 95]). *Any multiset function G on a countable domain can be expressed as*

$$G(S) = \phi_1\Big(\sum_{s \in S} \phi_2(s)\Big), \tag{2.2}$$

where $\phi_1 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ and $\phi_2 : \mathbb{R}^{d_2} \to \mathbb{R}$ are nonlinear functions.

The proof idea is to show that there exists an injective function $\sum_{s \in S} \phi(s)$. The above result is an extension of a universal approximation result for set functions [72, 73, 101], and suggests a neural network model for sets where ϕ_1, ϕ_2 are approximated by MLPs. The Graph Isomorphism Network (GIN) [95] implements this sum decomposition in the aggregation function to ensure the ability to express injective operations.

Here, the latent dimension d_2 plays a role. Proofs for countable domains use a discontinuous mapping ϕ_1 into a fixed-dimensional space, whereas MLPs universally approximate *continuous* functions [25]. Continuous set functions on $\mathbb{R}^{\leq M}$ (i.e., $|S| \leq M$) can be sum-decomposed as above with continuous ϕ_1, ϕ_2 and latent dimension at least $d_2 = M$. The dimension is a necessary and sufficient condition for universal approximation [92]. For GNNs, this means d_2 must be at least the maximum degree $\deg(G)$ of the input graph G.

2.1.2. Implications for graph distinction

Theorem 1 allows directly transferring any known result for 1-WL to MPNNs. For instance, 1-WL succeeds in distinguishing graphs sampled uniformly from all graphs on n nodes with high probability, and failure probability going to zero as $n \to \infty$ [8, 9]. 1-WL can also distinguish any nonisomorphic pair of trees [42]. It fails for regular graphs, as all node colors will be the same. The graphs that 1-WL can distinguish from any nonisomorphic graph can be recognized in quasilinear time [6]. See also [6, 18, 48] for more detailed results on the expressive power of variants of the WL algorithm.

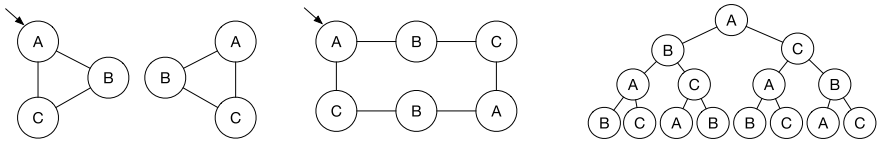


FIGURE 1

Graphs G_1 (left, 2 connected components) and G_2 (middle) with node attributes indicated by letters. The computation tree rooted at the node with arrow (right) agrees in both graphs, and likewise for the other nodes. Hence, 1-WL and MPNNs cannot distinguish G_1 and G_2. Figure adapted from **[33]**.

2.1.3. Computation trees and structural graph properties

To further illustrate the implications of GNNs' discriminative power, we look at some specific examples. The maximum information contained in any embedding $h_v^{(t)}$ can be characterized by a computation tree $\mathcal{T}(h_v^{(t)})$, i.e., an "unrolling" of the message passing procedure. 1-WL essentially colors computation trees. The tree $\mathcal{T}(h_v^{(t)})$ is constructed recursively: let $\mathcal{T}(h_v^{(0)}) = x_v$ for all $v \in V$. For $t > 0$, construct a root with label x_v and, for any $u \in \mathcal{N}(v)$ construct a child subtree $\mathcal{T}(h_u^{(t-1)})$. Figure 1 illustrates an example.

Proposition 1. *If for two nodes $u \neq v$, we have $\mathcal{T}(h_v^{(t)}) = \mathcal{T}(h_u^{(t)})$, then $h_v^{(t)} = h_u^{(t)}$.*

Comparing computation trees directly implies that MPNNs cannot distinguish regular graphs. It also shows further limitations with practical impact (Fig. 1), in particular for learning combinatorial algorithms, and for predicting properties of molecules, where functional groups are of key importance. We say a class of models $\mathcal{F}$ *decides* a graph property if there exists an $F \in \mathcal{F}$ such that for any two G, G' that differ in the property, we obtain $F(G) \neq F(G')$.

Proposition 2. *MPNNs cannot decide girth, circumference, diameter, radius, existence of a conjoint cycle, total number of cycles, and existence of a k-clique* **[33]**. *MPNNs cannot count induced (attributed) subgraphs for any connected pattern of 3 or more nodes, except star-shaped patterns* **[22]**.

Motivated by these limitations, generalizations of GNNs were proposed that provably increase their representational power. Two main directions are to (1) introduce node IDs (Section 2.2), and (2) use higher-order functions that act on tuples of nodes (Section 2.3).

2.2. Node IDs, local algorithms, combinatorial optimization, and lower bounds

The major weaknesses of MPNNs arise from their inability to identify nodes as the origin of specific messages. Hence, MPNNs can be strengthened by making nodes more distinguishable. The gained representational power follows from connections with local algorithms, where the input graph defines both the computational problem and the network topology of a distributed system: each node $v \in V$ is a local machine and generates a local output, and all nodes execute the same algorithm, without faults.

Approximation algorithms. Sato et al. [80] achieve a partial node distinction by transferring the idea of a *port numbering* from local algorithms. Edges incident to each node are numbered as outgoing ports. In each round, each node simultaneously sends a message to each port, but the messages can differ across ports:

$$m_v^{(t)} = f_{\text{Agg}}^{(t)}\big(\{\!\{\big(\text{port}(u,v), \text{port}(v,u), h_u^{(t-1)}\big) \mid u \in \mathcal{N}(v)\}\!\}\big). \tag{2.3}$$

Permutation invariance, though, is not immediate. This corresponds to the vector–vector consistent (VV_C) model for local algorithms [41]. The VV_C analogy allows transferring results on representing approximation algorithms. CPNGNN is a specific VV_C GNN model.

Theorem 3 ([80]). *There exists a CPNGNN that can compute a* $(\deg(G)+1)$*-approximation for the minimum dominating set problem, a CPNGNN that can compute a 2-approximation for the minimum vertex cover problem, but no CPNGNN can do better. No CPNGNN can compute a constant-factor approximation for the maximum matching problem.*

Adding a weak vertex 2-coloring leads to further results. Despite the increased power compared to MPNN, CPNGNNs retain most limitations of Proposition 2 [33].

A more powerful alternative is to endow nodes with fully unique identifiers [59, 81]. For example, augmenting the GIN model (a maximally expressive MPNN) [95] with random node identifiers yields a model that can decide subgraphs that MPNN and CPNGNN cannot [81]. This model can further achieve better approximation results for minimum dominating set ($(H(\deg(G)+1)+\varepsilon)$, where H is the harmonic number) and maximum matching $(1+\varepsilon)$.

Turing completeness. Analogies to local algorithms imply that MPNNs with unique node IDs are *Turing complete*, i.e., they can compute any function that a Turing machine can compute, including graph isomorphism. In particular, the proof shows an equivalence to the Turing universal LOCAL model from distributed computing [3, 57, 69].

Theorem 4 ([59]). *If* f_{Up} *and* f_{Agg} *are Turing complete functions and the GNN gets unique node IDs, then GNN and LOCAL are equivalent. For any MPNN* F *there exists a local algorithm* $\mathcal{A}$ *of the same depth, such that* $F(G) = \mathcal{A}(G)$*, and vice versa.*

Corollary 1 ([59]). *Under the conditions in Theorem 4, if the GNN depth (number of iterations) is at least* $\text{diameter}(G)$ *and the width is unbounded, then MPNNs can compute any Turing computable function over connected attributed graphs.*

Lower bounds. The *width* of a GNN refers to the dimensionality of the embeddings $h_v^{(t)}$. For bounded size, GNNs lose computational power. Via analogies to the CONGEST model [70], which bounds message sizes, one can transfer results on decision, optimization and estimation problems on graphs. These lead to lower bounds on the product of depth and width of the GNN. Here, the nodes do not have access to a random generator.

Theorem 5 ([59]). *If a problem cannot be solved in less than* T *rounds in CONGEST using messages of at most* b *bits, then it cannot be solved by an MPNN of width* $w \leq (b - \log_2 n)/p = O(b/\log n)$ *and depth* T*, where* $p = \Theta(n)$*.*

Theorem 5 directly implies lower bounds for solving combinatorial problems, e.g., $Tw = \Omega(n/\log n)$ for cycle detection and computing diameter, and $T\sqrt{w} = \Omega(\sqrt{n}/\log n)$ for minimum spanning tree, minimum cut, and shortest path [59].

Moreover, we can transfer ideas from communication complexity. The *communication capacity* c_f of an MPNN f (with unique node IDs) is the maximum number of symbols that the MPNN can transmit between any two disjoint sets $V_1, V_2 \subset V$ of nodes when viewed as a communication network: $c_f \leq \mathrm{cut}(V_1, V_2) \sum_{t=1}^{T} \min\{m_t, w_t\} + \sum_{t=1}^{T} \gamma_t$, where T is the GNN depth, w_t the width of layer t, m_t the size of the messages, and γ_t the size of a global state that is maintained. The communication capacity of the MPNN must be at least $c_f = \Omega(n)$ to distinguish all trees, and $c_f = \Omega(n^2)$ to distinguish all graphs [58]. By relating discrimination and function approximation (Section 2.4), these results have implications for function approximation, too.

Random node IDs. While unique node IDs are powerful in theory, in many practical examples the input graphs do not have unique IDs. An alternative is to assign random node IDs [1, 27]. This can still yield GNNs that are essentially permutation invariant: while their outputs are random, the outputs for different graphs are still sufficiently separated [1]. This leads to a probabilistic universal approximation result:

Theorem 6 ([1]). *Let $h : \mathcal{G} \to \mathbb{R}$ be a permutation invariant function on graphs of size $n \geq 1$. Then for all $\varepsilon, \delta > 0$ there exists an MPNN F with access to a global readout and with random node IDs such that for every $G \in \mathcal{G}$ it holds that* $\Pr(|F(G) - h(G)| \leq \varepsilon) \geq 1 - \delta$.

The proof builds on a result by [10] that states that any logical sentence in FOC_2 can be expressed by the addressed GNN. The logic considered here is a fragment of first-order (FO) predicate logic that allows to incorporate counting quantifiers of the form $\exists^{\geq k} x\, \psi(x)$, i.e., there are at least k elements x satisfying ψ, but is restricted to two variables. FOC_2 is tightly linked with the 1-WL test: for any nodes $u, v \in V$ in any graph, 1-WL colors u and v the same if and only if they are classified the same by all FOC_2 classifiers [19].

2.3. Higher-order GNNs

Instead of adding unique node IDs, one may increase the expressive power of GNNs by encoding subsets of V that are larger than the single nodes used in MPNNs. Three such directions are: (1) neural network versions of higher-dimensional WL algorithms, (2) (non)linear equivariant operations, and (3) recursion. Other strategies that could not be covered here use, e.g., simplicial and cell complexes [16, 17] or augment node attributes with topological information (e.g., persistent homology) [102].

Most of these GNNs act on k-tuples $s \in V^k$, and may be written in a unified form via tensors $H^{(t)} \in R^{n^k \times d_t}$, where the first k coordinates index the tuple, and $H^{(t)}_{s,:} \in \mathbb{R}^{d_t}$ is the representation of tuple s in layer t. For MPNNs, which use node and edge information, $H^{(0)} \in \mathbb{R}^{n \times n \times (d+1)}$. The first d channels of $H^{(0)}$ encode the node attributes: $H^{(0)}_{v,v,1:d} = x_v$ and $H^{(0)}_{u,v,1:d} = 0$ for $u \neq v$. The final channel captures the adjacency matrix A of the graph:

$H^{(0)}_{:,:,(d+1)} = A$. Node embeddings are computed by a permutation equivariant network:

$$f(G) = m \circ S_E \circ F^{(T)} \circ \cdots \circ F^{(1)} \circ \text{SHAPE}(G), \tag{2.4}$$

where $m : \mathbb{R}^{d_T} \to \mathbb{R}^{d_{\text{out}}}$ is an MLP that is applied to each representation h_v^T separately, $S_E : \mathbb{R}^{n^k \times d_T} \to \mathbb{R}^{n \times d_T}$ is a reduction $S_E(H)_{v,:} = \sum_{s \in V^k : s_1 = v} H_{s,:}$, and each layer $F^{(t)} : \mathbb{R}^{n^k \times d_{t-1}} \to \mathbb{R}^{n^k \times d_t}$ is a message passing (aggregation and update) operation for MPNNs, and will be defined for higher-order networks. The first operation shapes the input into the correct tensor form, if needed. For a graph embedding, we switch to a reduction $S_I : \mathbb{R}^{n^k \times d_T} \to \mathbb{R}^{d_T}$, $S_I(H) = \sum_{s \in V^k} H_{s,:}$ and apply the MLP m to the resulting vector: $F(G) = m \circ S_I \circ F^{(T)} \circ \cdots \circ F^{(1)} \circ \text{SHAPE(G)}$. The GNNs differ in their layers $F^{(t)}$.

2.3.1. Higher-order WL networks

Extending analogies of MPNNs and the 1-WL algorithm [66, 95], the first class of higher-order GNNs imitates versions of the k-dimensional WL algorithm. The k-WL algorithms are defined on k-tuples of nodes, and different versions differ in their aggregation and definition of neighborhood. In iteration 0, the *k-WL* algorithm labels each k-tuple $s \in V^k$ by a unique ID for its isomorphism type. Then it aggregates over neighborhoods $\mathcal{N}_i^{\text{WL}}(s) = \{(s_1, s_2, \ldots, s_{i-1}, v, s_{i+1}, \ldots, s_k) \mid \forall v \in V\}$ for $1 \leq i \leq k$:

$$c_i^{(t)}(s) = \left\{\!\!\left\{ c^{(t-1)}(s') \mid s' \in \mathcal{N}_i^{\text{WL}}(s) \right\}\!\!\right\}, \quad 1 \leq i \leq k,\ s \in V^k, \tag{2.5}$$

$$c^{(t)}(s) = \text{Hash}\big(c^{(t-1)}(s), c_1^{(t)}(s), c_2^{(t)}(s), \ldots, c_k^{(t)}(s)\big) \quad \forall s \in V^k. \tag{2.6}$$

For two graphs G, G' the k-WL algorithm then decides "not isomorphic" if $\{\!\{c^{(t)}(s) \mid s \in V(G)^k\}\!\} \neq \{\!\{c^{(t)}(s') \mid s' \in V(G')^k\}\!\}$ for some t, and returns "maybe isomorphic" otherwise. Like 1-WL, k-WL decides "not isomorphic" only if $G \not\cong G'$. The *Folklore k-WL* algorithm (k-FWL) differs in its update rule, which "swaps" the order of the aggregation steps [19]:

$$c_u^{(t)}(s) = (c^{(t-1)}_{(u,s_2,\ldots,s_k)}, c^{(t-1)}_{(s_1,u,s_3,\ldots,s_k)}, \ldots, c^{(t-1)}_{(s_1,\ldots,s_{k-1},u)}) \quad \forall u \in V, s \in V^k, \tag{2.7}$$

$$c^{(t)}(s) = \text{Hash}\big(c^{(t-1)}(s), \{\!\{c_u^{(t)}(s) \mid u \in V\}\!\}\big) \quad \forall s \in V^k. \tag{2.8}$$

The 1-WL and 2-WL test are equivalent, and for $k \geq 2$, $(k+1)$-WL can distinguish strictly more graphs than k-WL [19]. The k-FWL is as powerful as the $(k+1)$-WL for $k \geq 2$ [38].

Set-WL GNN. Since computations on k-tuples are expensive, [66] consider a GNN that corresponds to a set version of a k-WL algorithm. For any *set* $S \subseteq V$ with $|S| = k$, let $\mathcal{N}^{\text{set}}(S) = \{T \subset V, |T| = k \mid |S \cap T| = k - 1\}$. The set-based WL test (k-SWL) then updates as

$$c^{(t)}(S) = \text{Hash}\big(c^{(t-1)}(S), \{\!\{c^{(t-1)}(T) \mid T \in \mathcal{N}^{\text{set}}(S)\}\!\}\big); \tag{2.9}$$

its GNN analogue uses the aggregation and update (cf. equations (1.6) and (1.8))

$$h_S^{(t+1)} = \sigma\Big(W_1^{(t)} h_S^{(t)} + \sum\nolimits_{T \in \mathcal{N}^{\text{set}}(S)} W_2^{(t)} h_T^{(t)}\Big), \tag{2.10}$$

where σ is a coordinatewise nonlinearity (e.g., sigmoid or ReLU). This family of GNNs is equivalent in power to the k-SWL test [66] (Theorem 8). For computational efficiency, a local version restricts the neighborhood of S to sets T such that the nodes $\{u, v\} = S \Delta T$ in the symmetric difference are connected in the graph. This local version is weaker [1].

Folklore WL GNN. In analogy to the k-FWL algorithm, Maron et al. [62] define k-FGNNs with aggregations

$$h_s^{(t+1)} = f_{\mathrm{Up}}^{(t+1)}\left(h_s^{(t)}, \sum_{v\in V}\prod_{i=1}^{k} f_i^{(t+1)}\left(h^{(t)}_{(s_1,\ldots,s_{i-1},v,s_{i+1},\ldots,s_k)}\right)\right). \tag{2.11}$$

For $k = 2$, this model can be implemented via matrix multiplications. The input to the aggregation, for all pairs of nodes simultaneously, is a tensor $H \in \mathbb{R}^{n\times n\times d_t}$, with $H_{(u,v),:} = h_{(u,v)}$. The initial $H^{(0)} \in \mathbb{R}^{n\times n\times(d+1)}$ is defined as in the beginning of Section 2.3.

To compute the aggregation layer, first, we apply three MLPs $m_1, m_2 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ and $m_3 : \mathbb{R}^{d_1} \to \mathbb{R}^{d_3}$ to each embedding $h_{(u,v)}$ in H: $m_l(H)_{(u,v),:} = m_l(H_{(u,v),:})$ for $1 \le l \le 3$. Then one computes an intermediate representation $H' \in \mathbb{R}^{n\times n\times d_2}$ by multiplying matching "slices" of the outputs of m_1, m_2: $H'_{:,:,i} = m_1(H)_{:,:,i} \cdot m_2(H)_{:,:,i}$. The final output of the aggregation is the concatenation $(m_3(H), H') \in \mathbb{R}^{n\times n\times(d_2+d_3)}$. A variation of this model, a low-rank global *attention* model, was shown to relate attention and the 2-FWL algorithm via *algorithmic alignment*, which we discuss in Section 3.3 [71]. Attention in neural networks introduces learned pair-wise weights in the aggregation function.

The family of k-FGNNs is a class of nonlinear equivariant networks, and is equivalent in power to the k-FWL test and the $(k+1)$-WL test [7, 62] (Theorem 8).

2.3.2. Linear equivariant layers

While the models discussed so far rely on message passing, the GNN definition (2.4) only requires permutation equivariant or invariant operations in each layer. The k-linear (equivariant) GNNs (k-LEGNNs), introduced in [63], allow more general linear equivariant operations. In k-LEGNNs, each layer $F^{(t)} = \sigma \circ L^{(t)} : \mathbb{R}^{n^k\times d_{t-1}} \to \mathbb{R}^{n^k\times d_t}$ is a concatenation of a linear equivariant function $L^{(t)}$ and a coordinatewise nonlinear activation function. The function σ may also be replaced with a nonlinear function $f_1^{(t)} : \mathbb{R}^{d_{t+1/2}} \to \mathbb{R}^{d_{t+1}}$ (an MLP) applied separately to each tuple embedding $L^{(t)}(H^{(t-1)})_{s,:}$.

Characterizations of equivariant functions or networks were studied in [40, 52, 53, 74]. Maron et al. [63] explicitly characterize all invariant and equivariant linear layers, and show that the vector space of linear invariant or equivariant functions $f : \mathbb{R}^{n^k} \to \mathbb{R}^{n^\ell}$ has dimension $\mathrm{b}(k)$ and $\mathrm{b}(k+\ell)$, respectively, where $\mathrm{b}(k)$ is the kth Bell number. When including multiple channels and bias terms, one obtains the following bounds.

Theorem 7 ([63]). *The space of invariant (equivariant) linear layers $\mathbb{R}^{n^k\times d} \to \mathbb{R}^{d'}$ ($\mathbb{R}^{n^k\times d} \to \mathbb{R}^{n^k\times d'}$) has dimension $dd'\mathrm{b}(k) + d'$ (for equivariant, $dd'\mathrm{b}(2k) + d'\mathrm{b}(k)$).*

The GNN model uses one parameter (coefficient) for each basis tensor. Importantly, the number of parameters is independent of the number of nodes. The proof for identifying the basis tensors sets up a fixed point equation with Kronecker products of any permutation matrix that any equivariant tensor must satisfy. The solutions to these equations are defined by equivalence classes of multiindices in $[n]^k$. Each equivalence class is represented by a partition γ of $[k]$, e.g., $\gamma = \{\{1\},\{2,3\}\}$ includes all multiindices (i_1, i_2, i_3) where $i_1 \neq i_2, i_3$ and $i_2 = i_3$. The basis tensors $B^\gamma \in \{0,1\}^{n^k}$ are then such that $B^\gamma_s = 1$ if and only if $s \in \gamma$.

Linear equivariant GNNs of order k (k-LEGNNs) parameterized with the full basis are as discriminative as the k-WL algorithm [62] (Theorem 8). To achieve this discriminative power, each entry $H^{(0)}_{s,:}$ in the input tensor encodes an initial coloring of the isomorphism type of the subgraph indexed by the k-tuple s.

2.3.3. Summary of representational power via WL

The following theorem summarizes equivalence results between the GNNs discussed so far and variants of the WL test. Following [7], we here use equivalence relations, as they suffice for universal approximation in Section 2.4. For a set $\mathcal{F}$ of functions defined on $\mathcal{G}$, define an equivalence relation ρ via the joint discriminative power of all $F \in \mathcal{F}$, i.e., for any $G, G' \in \mathcal{G}$:

$$(G, G') \in \rho(\mathcal{F}) \iff \forall F \in \mathcal{F},\ F(G) = F(G'). \tag{2.12}$$

Theorem 8. *The above GNN families have the following equivalences:*

$$\rho(\mathit{MGNN}) = \rho(2\text{-}\mathit{WL}) \quad [95], \tag{2.13}$$

$$\rho(k\text{-}\mathit{set\text{-}GNN}) = \rho(k\text{-}\mathit{SWL}) \quad [66], \tag{2.14}$$

$$\rho(k\text{-}\mathit{LEGNN}) = \rho(k\text{-}\mathit{WL}) \quad [34, 63], \tag{2.15}$$

$$\rho(k\text{-}\mathit{FGNN}) = \rho((k+1)\text{-}\mathit{WL}) \quad [7, 62]. \tag{2.16}$$

Analogous results hold for equivariant models (for node representations), with the exception of equality (2.15), which becomes an inclusion: $\rho(k\text{-LEGNN}_E) \subseteq \rho(k\text{-WL}_E)$ [7].

2.3.4. Relational pooling

One option to obtain nonlinear permutation invariant functions is to average permutation-sensitive functions over the permutation group Π_n. Murphy et al. [67, 68] propose such a model, inspired by joint exchangeability of random variables [2, 29]. Concretely, if $A \in \mathbb{R}^{n \times n}$ denotes the adjacency matrix of the input graph G and $X \in \mathbb{R}^{n \times d}$ the matrix of node attributes, then

$$F_{\mathrm{RP}}(G) = \frac{1}{n!} \sum_{\pi \in \Pi_n} g(A_{\pi,\pi}, X_\pi) = g(\pi \cdot H^{(0)}), \tag{2.17}$$

where X_π is X with permuted rows, and $H^{(0)}$ is the tensor combining adjacency matrix and node attributes. Here, g is any permutation-sensitive function, and may be modeled via various nonlinear function approximators, e.g., neural networks such as fully connected networks (MLPs), recurrent neural networks or a combination of a convolutional network applied to A and an MLP applied to X. In particular, this model allows implementing graph isomorphism testing via node IDs (cf. Section 2.2) if g is a universal approximator [68]. For instance, node IDs may be permuted over nodes and concatenated with the node attributes:

$$F_{\mathrm{RP}}(G) = \frac{1}{n!} \sum_{\pi \in \Pi_n} \big(A_{\pi,\pi}, [X_\pi, I_n]\big) = \frac{1}{n!} \sum_{\pi \in \Pi_n} g\big(A, \big[X, (I_n)_\pi\big]\big), \tag{2.18}$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. If g is an MPNN, the resulting model is strictly more powerful than the 1-WL test and hence g by itself.

The drawback of the *Relational Pooling* (2.17) is its computational intractability. Various approximations have been considered, e.g., defining canonical orders, stochastic approximations, and applying g to all possible k-subsets of V. In the latter case, increasing k strictly increases the expressive power. *Local Relational Pooling* is a variant that applies relational pooling to the k-hop subgraphs centered at each node, and then aggregates the results. This operation provably allows to identify and count subgraphs of size up to k [22].

2.3.5. Recursion

A general strategy for encoding a graph is to encode a collection of subgraphs, and then aggregate these encodings. When doing so, an important bit of information are node correspondences across subgraphs [15, 86]. Otherwise, this process includes the reconstruction hypothesis [46, 88], i.e., the question whether any graph G can be reconstructed from the collection of its subgraphs $G \setminus \{v\}$, for all v in G.

Indeed, the expressive power of such a model depends on the set of subgraphs, the type of subgraph encodings and the aggregation. Tahmasebi et al. [86] show that recursion can be a powerful tool: instead of iterative message passing or layering, a *recursive* application of the above subgraph embedding step, even with a simple set aggregation like (1.6), can enable a GNN that can count any bounded-size subgraphs, as opposed to MPNNs (Proposition 2).

Let $\mathcal{N}_r(v)$ be the r-hop neighborhood of v in G. *Recursive neighborhood pooling (RNP)* encodes *intersections* of such neighborhoods of different radii. Given an input graph G with node attributes $\{h_u^{\text{in}}\}_{u \in V(G)}$ and a sequence $(r_1, \dots, r_t)$ of radii, RNP recursively encodes the node-deleted r_1-neighborhoods $G_v = \mathcal{N}_{r_1}(v) \setminus \{v\}$ of all nodes $v \in V$ after marking the deletion in augmented representations h_u^{aug}, $u \in V$. It then combines the results, and returns node representations of all nodes. I.e., for each node $v \in V$, it computes G_v and

$$h_u^{\text{aug}} = \big(h_u^{\text{in}}, \mathbf{1}\big[(u,v) \in E(G_v)\big]\big) \quad \forall u \in V(G_v), \tag{2.19}$$

$$\big\{\!\big\{h'_{v,u}\big\}\!\big\}_{u \in G_v} \leftarrow \text{RNP-GNN}\big(G_v, \big\{\!\big\{h_u^{\text{aug}}\big\}\!\big\}_{u \in G_v}, (r_2, r_3, \dots, r_t)\big), \quad \text{(recursion)} \tag{2.20}$$

$$\text{return} \qquad h_v^{\text{out}} = f_{\text{Agg}}^{(t)}\big(h_v^{\text{in}}, \big\{\!\big\{h'_{v,u}\big\}\!\big\}_{u \in G_v}\big), \quad \forall v \in V. \tag{2.21}$$

If the sequence of radii is empty (base case), then the algorithm returns the input attributes h_u^{in}. In contrast to *iterative* message passing, the encoded subgraphs here correspond to intersections of local neighborhoods. Together with the node deletions and markings that retain node correspondences, this maintains more structural information. If the radii sequence dominates a covering sequence for a subgraph H of interest, then, with appropriate parameters, RNP can count the induced and noninduced subgraphs of G isomorphic to H [86]. The computational cost is $O(n^k)$ for recursion depth k, and better for very sparse graphs, in line with computational lower bounds.

2.4. Universal approximation

Distinguishing given graphs is closely tied to approximating continuous functions on graphs. In early work, Scarselli et al. [82] take a fixed point view and show a universal approximation result for infinite-depth MPNNs whose layers are contraction operators, for

functions on equivalence classes defined by computation trees. Dehmamy et al. [28] analyze the ability of GNNs to compute polynomial functions of the adjacency matrix.

Later works derive universal approximation results for graph and permutation-equivariant functions from graph discrimination results via extensions of the Stone–Weierstrass theorem [7, 23, 47, 64]. For instance, H-invariant networks (for a permutation group H) can universally approximate H-invariant polynomials [64], which in turn can universally approximate any invariant function [98]. Keriven and Peyré [47] do not fix the size of the graph and show that shallow equivariant networks can, with a single set of parameters, well approximate a function on graphs of varying size. Both constructions involve very large tensors.

More generally, the Stone–Weierstrass theorem (for symmetries) allows translating Theorem 8 into universal approximation results. Let $\mathcal{C}_I(\mathcal{X}, \mathcal{Y})$ be the set of invariant continuous functions from $\mathcal{X}$ to $\mathcal{Y}$. Then a class $\mathcal{F}$ of GNNs is *universal* if its closure $\overline{\mathcal{F}}$ (in uniform norm) on a compact set K is the entire $\mathcal{C}_I(K, \mathbb{R}^p)$.

Theorem 9 ([7]). *Let $K_{disc} \subseteq \mathcal{G}_n \times \mathbb{R}^{d_0 \times n}$, $K \subseteq \mathbb{R}^{d_0 \times n}$ be compact sets, where $\mathcal{G}_n$ is the set of all unweighted graphs on n nodes. Then*

$$\overline{MGNN} = \left\{ f \in \mathcal{C}_I\left(K_{\text{disc}}, \mathbb{R}^p\right) : \rho(2\text{-}WL) \subseteq \rho(f) \right\}, \tag{2.22}$$

$$\overline{k\text{-}LEGNN} = \left\{ f \in \mathcal{C}_I(K, \mathbb{R}^p) : \rho(k\text{-}WL) \subseteq \rho(f) \right\}, \tag{2.23}$$

$$\overline{k\text{-}FGNN} = \left\{ f \in \mathcal{C}_I(K, \mathbb{R}^p) : \rho((k+1)\text{-}WL) \subseteq \rho(f) \right\}. \tag{2.24}$$

Analogous relations hold for equivariant functions, except for

$$\overline{k\text{-}LEGNN_E} = \left\{ f \in \mathcal{C}_E(K, \mathbb{R}^{n \times p}) : \rho(k\text{-}LEGNN_E) \subseteq \rho(f) \right\},$$

which is a superset of $\{ f \in \mathcal{C}_E(K, \mathbb{R}^{n \times p}) : \rho(k\text{-}WL_E) \subseteq \rho(f) \}$.

3. GENERALIZATION

Beyond approximation power, a second important question in machine learning is generalization. *Generalization* asks how well the estimated function $\hat{F}$ is performing according to the population risk, i.e., $\mathcal{R}(\hat{F})$, as a function of the number of data points N and model properties. Good generalization may demand explicit (e.g., via a penalty term) or implicit regularization (e.g., via the optimization algorithm). Hence, generalization analyses involve aspects of the complexity of the model class $\mathcal{F}$, the target function we aim to learn, the data and the optimization procedure. This is particularly challenging for neural networks, due to the nested functional form and the nonconvexity of the empirical risk.

A classic learning theoretic perspective bounds the *generalization gap* $\mathcal{R}(\hat{F}) - \widehat{\mathcal{R}}(\hat{F})$ via the complexity of the model class $\mathcal{F}$ (Section 3.1). These approaches do not take into account possible implicit regularization via the optimization procedure. One possibility to do so is via the *Neural Tangent Kernel* approximation (Section 3.2). Finally, for more complex, structured target functions, e.g., algorithms or physics simulations, one may want to also consider the structure of the target task. One such option is *Algorithmic Alignment*

(Section 3.3). Another strategy for obtaining generalization bounds is via *algorithmic stability*, the condition that, if one data point is replaced, the outcome of the learning algorithm does not change much. This strategy led to some early bounds for spectral GNNs [91].

3.1. Generalization bounds via complexity of the model class

Vapnik–Chervonenkis dimension. The first GNN generalization bound was based on bounding the Vapnik–Chervonenkis (VC) dimension [89] of the GNN function class $\mathcal{F}$. The *VC dimension* of $\mathcal{F}$ expresses the maximum size of a set of data points such that for any binary labeling of the data, some GNN in $\mathcal{F}$ can perfectly fit, i.e., *shatter*, the set. The VC dimension directly leads to a bound on the generalization gap. Here, we only state the results for sigmoid activation functions.

Theorem 10 ([84]). *The VC dimension of GNNs with p parameters, H hidden neurons (in the MLP) and input graphs of size n is $O(p^2H^2n^2)$.*

Strictly speaking, Theorem 10 is for node classification with one hidden layer in the aggregation function MLPs. The VC dimension directly yields a bound on the generalization gap: for a class $\mathcal{F}$ with VC dimension D, with probability $1-\delta$, it holds that $\mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f}) \leq O(\sqrt{\frac{D}{N}\log\frac{N}{D}}) + \sqrt{\frac{1}{2N}\log\frac{1}{\delta}}$. Interestingly, in these bounds, GNNs are a generalization of recurrent neural networks [84]. The VC dimension bounds for GNNs are the same as for recurrent neural networks [50]; for fully connected MLPs, they are missing the factor n^2 [45].

Rademacher complexity. Bounds that are in many cases tighter can be obtained via Rademacher complexity. The *empirical Rademacher complexity* $\widehat{\mathfrak{R}}_S(\mathcal{F})$ of a function class $\mathcal{F}$ measures how well it can fit "noise" in the form of uniform random variables $\sigma = (\sigma_1, \ldots, \sigma_N)$ in $\{-1, +1\}$: $\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma[\sup_{F\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \sigma_i F(x_i)]$, for a fixed data sample $S = \{x_1, \ldots, x_N\}$. Similarly to VC dimension, $\widehat{\mathfrak{R}}_S(\mathcal{F})$ provides a bound on the probability of error under the full data distribution: $\mathbb{P}[\text{error}(F)] \leq \widehat{\mathcal{R}}(F) + 2\widehat{\mathfrak{R}}_S(\mathcal{J}) + 3\sqrt{\frac{\log(2/\delta)}{2N}}$, where $\mathcal{J}$ is the class of functions $F \in \mathcal{F}$ concatenated with the loss. Garg et al. [33] analyze a GNN that applies a logistic linear binary classifier at each node, averages these predictions for a graph-level prediction, and uses a *mean field update* [26]: $h_v^t = \phi(W_1 x_v + W_2\rho(\sum_{u\in N(v)} g(h_u^{t-1})))$, where ϕ, ρ, g are nonlinear functions with bounded Lipschitz constant that are zero at zero (e.g., tanh), and $\|W_1\|_F, \|W_2\|_F \leq B$. The logistic predictor outputs a "probability" for the label 1, and is evaluated by a margin loss function that gives a (scaled) penalty if the "probability" of the correct label is below a threshold ($\frac{\gamma+1}{2}$).

Theorem 11 ([33]). *Let $\mathcal{C}$ be the product of the Lipschitz constants of ϕ, ρ, g, and B; T the number of GNN iterations; w the dimension of the embeddings h_v^t, and d the maximum branching factor in the computation tree. Then the generalization gap of the GNN can be bounded as: $\tilde{O}(\frac{wd}{\sqrt{N}\gamma})$ for $\mathcal{C} < 1/d$, $\tilde{O}(\frac{wdT}{\sqrt{N}\gamma})$ for $\mathcal{C} = 1/d$, and $\tilde{O}(\frac{wd\sqrt{wT}}{\sqrt{N}\gamma})$ for $\mathcal{C} > 1/d$.*

The factor d is equal to $\max_{v\in G}\deg(v) - 1$. For recurrent neural networks, the same bounds hold, but with $d = 1$ [21]: a sequence is a tree with branching factor 1. In comparison,

for the VC bounds in this setting, with $H = w$, $n > d$ and p is the size of the matrices W (about w^2), we obtain a generalization bound of $\tilde{O}(w^3 n/\sqrt{N})$, ignoring log factors. Later work tightens the bounds in Theorem 11 by using a *PAC-Bayesian* approach [56].

3.2. Generalization bounds via the Neural Tangent Kernel

Infinitely-wide neural networks can be related to kernel learning techniques [4, 5, 31, 32, 43]. Du et al. [30] extend this analysis to a broad class of GNNs. The main idea underlying the *Neural Tangent Kernel (NTK)* is to approximate a neural network $F(\theta, G)$ with a kernel derived from the training dynamics. Assume we fit $F(\theta, G)$ with the squared loss $L(\theta) = \sum_{i=1}^{N} \ell(F(\theta, G_i), y_i) = \frac{1}{2}(F(\theta, G_i) - y_i)^2$, where $\theta \in \mathbb{R}^m$ collects all parameters of the network. If we optimize with gradient descent with infinitesimally small step size, i.e., $\frac{d\theta(t)}{dt} = -\nabla L(\theta(t))$, then the network outputs $u(t) = (F(\theta(t), G_i))_{i=1}^{N}$ follow the dynamics

$$\frac{du}{dt} = -H(t)\big(u(t) - \mathbf{y}\big), \quad \text{where } H(t)_{ij} = \left\langle \frac{\partial F(\theta(t), G_i)}{\partial \theta}, \frac{\partial F(\theta(t), G_j)}{\partial \theta} \right\rangle. \tag{3.1}$$

Here, $\mathbf{y} = (y_i)_{i=1}^{N}$. If θ is sufficiently large (i.e., the network sufficiently wide), then it was shown that the matrix $H(t) \in \mathbb{R}^{N \times N}$ remains approximately constant as a function of t. In this case, the neural network becomes approximately a kernel regression [85]. If the parameters $\theta(0)$ are initialized as i.i.d. Gaussian, then the matrix $H(0)$ converges to a deterministic kernel matrix $\tilde{H}$, the *Neural Tangent Kernel*, with closed form regression solution $F_{\tilde{H}}(G)$. Given this approximation, one may analyze generalization via kernel learning theory.

Theorem 12 ([11]). *Given N i.i.d. training data points and any loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, 1]$ that is 1-Lipschitz in the first argument with $\ell(y, y) = 0$, with probability $1 - \delta$ the population risk of the Graph Neural Tangent predictor is bounded as*

$$\mathcal{R}(F_{\tilde{H}}) = O\left(\frac{1}{N}\sqrt{\mathbf{y}^\top \tilde{H}^{-1}\mathbf{y} \cdot \operatorname{tr}(\tilde{H})} + \sqrt{\frac{1}{N}\log(1/\delta)}\right).$$

In contrast to the results in Section 3.1, the complexity measure $\mathbf{y}^\top \tilde{H}^{-1}\mathbf{y}$ of the target function is data-dependent. If the target function to be learned follows a simple GNN structure with a polynomial, then this bound can be polynomial:

Theorem 13 ([30]). *Let $\bar{h}_v = c_v \sum_{u \in \mathcal{N}(v) \cup \{v\}} h_u$. If the labels y_i, $1 \le i \le N$, satisfy*

$$y_i = \alpha_1 \sum_{v \in V(G_i)} \beta_1^\top \bar{h}_v + \sum_{l=1}^{\infty} \alpha_{2l} \sum_{v \in V} \big(\beta_{2l}^\top \bar{h}_v\big)^{2l}$$

for $\alpha_k \in \mathbb{R}$, $\beta_k \in \mathbb{R}^d$, then $\mathbf{y}^\top \tilde{H}^{-1}\mathbf{y} \le 2|\alpha_1| \cdot \|\beta_1\|_2 + \sum_{l=1}^{\infty} \sqrt{2\pi}(2l-1)|\alpha_{2l}| \cdot \|\beta_{2l}\|_2^{2l}$. With $n = \max_i V(G_i)$, we have $\operatorname{tr}(\tilde{H}) = O(n^2 N)$.

3.3. Generalization via algorithmic alignment

The Graph NTK analysis shows a polynomial sample complexity if the function to be learned is close to the computational structure of the GNN, in a simple way. While this applies to mainly simpler learning tasks, the idea of an "alignment" of computational

structure carries further. Recently, there has been growing interest in learning scientific tasks, e.g., given a set of particles or planets along with their location, mass and velocity, predict the next state of the system [12,78,79], and in "algorithmic reasoning," e.g., learning to solve combinatorial optimization problems in particular over graphs [20]. In such cases, the target function corresponds to an algorithm, e.g., dynamic programming.

While many neural network architectures have the power to represent such tasks, empirically, they do not learn them equally well from data. In particular, GNNs perform well here, i.e., their architecture encodes suitable *inductive biases* [13,96]. As a concrete example, consider the Shortest Path problem. The computational structure of MPNNs matches that of the Bellman–Ford (BF) algorithm [14] very well: both "algorithms" iterate, and in each iteration t, update the state as a function of the neighboring nodes and edge weights $w(u,v)$:

$$\begin{aligned} &\text{(BF)}\ \ d[t][v] = \min_{u\in\mathcal{N}(v)} d[t-1][u] + w(u,v), \\ &\text{(GNN)}\ \ h_v^t = \sum_{u\in\mathcal{N}(v)} \mathrm{MLP}\big(h_u^{t-1}, h_v^{t-1}, w(u,v)\big). \end{aligned} \tag{3.2}$$

Hence, the GNN can simulate the BF algorithm if it uses sufficiently many iterations, and if the aggregation function approximates the BF state update. Intuitively, this is a much simpler function to learn than the full algorithm as a black box, i.e., the GNN encodes much of the algorithmic structure, sparsity and invariances in the architecture. More generally, MPNNs match the structure of many dynamic programs in an analogous way [96].

The NTK results formalize simplicity by a small function norm in the RKHS associated with the Graph NTK; this can become complicated with more complex tasks and multiple layers. To quantify *structural* match, Xu et al. [96] define *algorithmic alignment* by viewing a neural network as a structured arrangement of learnable modules – in a GNN, the (MLPs in the) aggregation functions – and define complexity via sample complexity of those modules in a PAC-learning framework. Sample complexity in PAC learning is defined as follows: We are given a data sample $\{(x_i, y_i)\}_{i=1}^N$ drawn i.i.d. from a distribution $\mathcal{P}$ that satisfies $y_i = g(x_i)$ for an underlying target function g. Let $f = \mathcal{A}(\{x_i, y_i\}_{i=1}^N)$ be the function output by a learning algorithm $\mathcal{A}$. For a fixed error ε and failure probability $1-\delta$, the function g is (N,ε,δ)-*PAC learnable* with $\mathcal{A}$ if

$$\mathbb{P}_{x\sim\mathcal{P}}\big[\big|f(x) - g(x)\big| < \varepsilon\big] \geq 1-\delta. \tag{3.3}$$

The *sample complexity* $\mathcal{C}_{\mathcal{A}}(g,\varepsilon,\delta)$ is the smallest N so that g is (N,ε,δ)-learnable with $\mathcal{A}$.

Definition 1 (Algorithmic alignment [96]). Let g be a target function and $\mathcal{N}$ a neural network with M modules $\mathcal{N}_i$. The module functions $f_1,\dots,f_M$ generate g for $\mathcal{N}$ if, by replacing $\mathcal{N}_i$ with f_i, the network $\mathcal{N}$ simulates g. Then $\mathcal{N}$ (N,ε,δ)-*algorithmically aligns* with g if (1) $f_1,\dots,f_M$ generate g and (2) there are learning algorithms $\mathcal{A}_i$ for the $\mathcal{N}_i$'s such that $M\cdot\max_i C_{\mathcal{A}_i}(f_i,\varepsilon,\delta) \leq N$.

Algorithmic alignment resembles Kolmogorov complexity [51]. Thus, it can be hard to obtain the optimal alignment between a neural network and an algorithm. But, *any* algorithmic alignment yields a bound, and any with acceptable sample complexity may suffice.

The complexity of the MLP modules in GNNs may be measured with a variety of techniques. One option is the NTK framework. The module-based bounds then resemble the polynomial bound in Theorem 13, since both are extensions of [5]. However, here, the bounds are applied at a module level, and not for the entire GNN as a unit. Theorem 14 translates these bounds, in a simplified setting, into sample complexity bounds for the full network.

Theorem 14 ([96]). *Fix ε and δ. Suppose $\{(G_i, y_i)\}_{i=1}^N \sim \mathcal{P}$, where $|V(G_i)| < n$, and $y_i = g(G_i)$ for some g. Suppose $\mathcal{N}_1, \ldots, \mathcal{N}_M$ are network $\mathcal{N}$'s MLP modules in sequential order of processing. Suppose $\mathcal{N}$ and g (N, ε, δ)-algorithmically align via functions $f_1, \ldots, f_M$ for a constant M. Under the following assumptions, g is $(N, O(\varepsilon), O(\delta))$-learnable by $\mathcal{N}$.*

(a) Sequential learning. *We train $\mathcal{N}_i$'s sequentially: $\mathcal{N}_1$ has input samples $\{\hat{x}_i^{(1)}, f_1(\hat{x}_i^{(1)})\}_{i=1}^N$, with $\hat{x}_i^{(1)}$ obtained from G_i. For $j > 1$, the input $\hat{x}_i^{(j)}$ for $\mathcal{N}_j$ are the outputs of the previous modules, but labels are generated by the correct functions $f_{j-1}, \ldots, f_1$ on $\hat{x}_i^{(1)}$.*

(b) Algorithm stability. *Let $\mathcal{A}$ be the learning algorithm for the $\mathcal{N}_i$'s, $f = \mathcal{A}(\{x_i, y_i\}_{i=1}^N)$, and $\hat{f} = \mathcal{A}(\{\hat{x}_i, y_i\}_{i=1}^N)$. For any x, $\|f(x) - \hat{f}(x)\| \leq L_0 \cdot \max_i \|x_i - \hat{x}_i\|$, for some $L_0 < \infty$.*

(c) Lipschitzness. *The learned functions $\hat{f}_j$ satisfy $\|\hat{f}_j(x) - \hat{f}_j(\hat{x})\| \leq L_1 \|x - \hat{x}\|$, for some $L_1 < \infty$.*

The big O notation here hides factors including the Lipschitz constants, number of modules, and graph size. When measuring module complexity via the NTK, Theorem 14, e.g., indeed yields a gap between fully connected networks and GNNs in simple cases [96], supporting empirical results. While some works use sequential training [90], empirically, better alignment improves learning and generalization in practice even with more common "end-to-end" training, i.e., optimizing all parameters simultaneously [13,96].

At a general level, these alignment results indicate that it is not only possible to learn combinatorial algorithms and physical reasoning tasks with machine learning, but how, in turn, incorporating expert knowledge, e.g., in algorithmic techniques or physics, into the design of the learning method can improve sample efficiency.

4. EXTRAPOLATION

Section 3 summarizes results for in-distribution generalization, i.e., how well a learned model performs on data from the same distribution $\mathcal{P}$ as the training data. Yet, in many practical scenarios, a model is applied to data from a different distribution. A strong case of such a distribution shift is *extrapolation*. It considers the expected loss $\mathbb{E}_{(G,y)\sim\mathcal{Q}}[\ell(G, y, F(G))]$ under a distribution $\mathcal{Q}$ with different support, e.g., $\text{supp}(\mathcal{Q}) \supset \text{supp}(\mathcal{P})$. For graphs, $\mathcal{Q}$ may entail graphs of different sizes, different degrees, or with node attributes in different ranges from the training graphs. As no data has been observed in the new domain parts, extrapolation can be ill-defined without stronger assumptions on the

task and model class. What assumptions are sufficient? Theoretical results on extrapolation assume the graphs have sufficient structural similarity and/or the model class is sufficiently restricted to extrapolate accurately. Empirically, while extrapolation has been difficult, several works achieve GNN extrapolation in tasks like predicting the time evolution of physical systems [12], learning graph algorithms [90], and solving equations [54].

Structural similarity of graphs. One possibility to guarantee successful extrapolation to larger graphs is to assume sufficient structural similarity between the graphs in $\mathcal{P}$ and $\mathcal{Q}$, in particular, structural properties that matter for the GNN family under consideration. For spectral GNNs, which learn functions of the graph Laplacian, this assumption has been formalized as the graphs arising from the same underlying topological space, manifold or graphon. Under such conditions, spectral GNNs – with conditions on the employed filters – can generalize to larger graphs [55,76,77].

For message passing GNNs, whose representations rely on computation trees as local structures (Section 2.1), an agreement in the distributions of the computation trees in the graphs sampled from $\mathcal{P}$ and $\mathcal{Q}$ is necessary [99]. This is violated, for instance, if the degree distribution is a function of the graph size, as is the case for random graphs under the Erdős–Rényi or Preferential Attachment models. The computation tree of depth t rooted at a node v corresponds to the color $c^{(t)}(v)$ assigned by the 1-WL algorithm.

Theorem 15 ([99]). *Let $\mathcal{P}$ and $\mathcal{Q}$ be finitely supported distributions of graphs. Let $\mathcal{P}^t$ be the distribution of colors $c^{(t)}(v)$ over $\mathcal{P}$ and similarly $\mathcal{Q}^t$ for $\mathcal{Q}$. Assume that any graph in $\mathcal{Q}$ contains a node with a color in $\mathcal{Q}^t \setminus \mathcal{P}^t$. Then, for any graph regression task solvable by a GNN with depth t there exists a GNN with depth at most $t+3$ that perfectly solves the task on $\mathcal{P}$ and predicts an answer with arbitrarily large error on all graphs from $\mathcal{Q}$.*

The proof exploits the fact that GNN predictions on nodes only depend on the associated computation tree, and that a sufficiently flexible GNN (depth at least $t+2$ layers and width $\max\{(\max \deg(G)+1)^t \cdot |C|, 2\sqrt{|P|}\}$, where the max degree refers to any graph in the support, $|C|$ is the finite number of possible input node attributes and P the set of colors encountered in graphs in the support) can assign arbitrary target labels to any computation tree [66,99]. That is, the available information allows for multiple local minima of the empirical risk. A similar result can be shown for node prediction tasks.

Conditions on the GNN. If one cannot guarantee sufficient structural similarity of the input graphs, then further restrictions on the GNN model can enable extrapolation to different graph sizes, structures and ranges of input node attributes. If there are no training observations in a certain range of attributes or local structures, then the predictions of the learned model depend on the *inductive biases* induced by the model architecture, loss function and training algorithm. In other words, which, out of multiple fitting functions (minima), a model will choose, depends on these biases.

Xu et al. [97] analyze such biases to obtain conditions on the GNN for extrapolation. Taking the perspective of algorithmic alignment (Section 3.3), they first analyze how individual module functions, i.e., the MLPs in the aggregation function of a GNN, extrapolate,

and then transfer this to the entire GNN. The aggregation functions enter the extrapolation regime, e.g., if the node attributes, node degrees or computation trees are different in $\mathcal{Q}$, as they determine the inputs to the aggregations. The following theorem states that, away from $\text{supp}(\mathcal{P})$, MLPs implement directionally linear functions.

Theorem 16 ([97]). *Suppose we train a two-layer MLP $f : \mathbb{R}^d \to \mathbb{R}$ with ReLU activation functions with squared loss in the NTK regime. For any direction $v \in \mathbb{R}^d$, let $x_0 = tv$. As $t \to \infty$, $f(x_0 + hv) - f(x_0) \to \beta_v \cdot h$ for any $h > 0$, where β_v is a constant linear coefficient. Moreover, given $\varepsilon > 0$, for $t = O(\frac{1}{\varepsilon})$, we have $|\frac{f(x_0+hv)-f(x_0)}{h} - \beta_v| < \varepsilon$.*

The linear function and the constant terms in the convergence rate depend on the training data and the direction v. The proof of Theorem 16 relies on the fact that a neural network in the NTK regime learns a minimum-norm interpolation function [4, 5, 43]. Although Theorem 16 uses a simplified setting of a wide 2-layer network, similar results hold empirically for more general MLPs [97].

To appreciate the implications of this result in the context of GNNs, consider the example of Shortest Path in equation (3.2). For the aggregation function to mimic the Bellman–Ford algorithm, the MLP must approximate a nonlinear function. But, in the extrapolation regime, it implements a linear function and therefore is expected to not approximate Bellman–Ford well any more. Indeed, empirical works that successfully extrapolate GNNs for Shortest Path use a different aggregation function of the form [13, 90]

$$h_u^{(t)} = \min_{v \in \mathcal{N}(u)} \text{MLP}^{(t)}\big(h_u^{(t-1)}, h_v^{(t-1)}, w_{(v,u)}\big). \tag{4.1}$$

Here, the nonlinear parts do not need to be learned, allowing to extrapolate with a linear learned MLP. More generally, the directionally linear extrapolation suggests that the (1) architecture or (2) input encoding should be set up such that the target function can be approximated by MLPs learning linear functions (*linear algorithmic alignment*). An example for (2) may be found in forecasting physical systems, e.g., predicting the evolution of n objects in a gravitational system, and the node (object) attributes are mass, location, and velocity at time t. The position of an object at time $t + 1$ is a nonlinear function of the attributes of the other objects. When encoding the nonlinear function as transformed edge attributes, the function to be learned becomes linear. Indeed, many empirical works that successfully extrapolate implement the idea of linear algorithmic alignment [24, 44, 61, 87, 97, 100].

Finally, the geometry of the training data also plays an important role. Xu et al. [97] show empirical results and initial theoretical results for learning max-degree, suggesting that, even with linear algorithmic alignment, sufficient diversity in the training data is needed to identify the correct linear functions.

For the case when the target test distribution $\mathcal{Q}$ is known, Yehudai et al. [99] propose approaches for combining elements of $\mathcal{P}$ and $\mathcal{Q}$ to enhance the range of the data seen by the GNN.

5. CONCLUSION

This survey covered three main topics in understanding GNNs: representation, generalization, and extrapolation. As GNNs are an active research area, many results could not be covered. For example, we focused on MPNNs and main ideas for higher-order GNNs, but neglected spectral GNNs, which closely relate to ideas in graph signal processing. Other emergent topics include adversarial robustness, optimization behavior of the empirical risk and its improvements, and computational scalability and approximations. Moreover, GNNs have a rich set of mathematical connections, a selection of which was summarized here.

For function approximation, the limitations of MPNNs motivated powerful higher-order GNNs. However, these are still computationally expensive. What efficiency is theoretically possible? Moreover, most applications may not require full graph isomorphism power, or k-WL power for large k. What other measures make sense? Do they allow better and sharper complexity results? Initial works consider, e.g., subgraph counting [22, 86].

The generalization results so far need to use simplifications in the analysis. To what extent can they be relaxed? Do more specific tasks or graph classes allow sharper results? Which modifications of GNNs would allow them to generalize better, and how do higher-order GNNs generalize? Similar questions pertain to extrapolation and reliability under distribution shifts, a topic that has been studied even less than GNN generalization.

In general, revealing further mathematical connections may enable the design of richer models and enable a more thorough understanding of GNNs' learning abilities and limitations, and potential improvements.

ACKNOWLEDGMENTS

The author would like to thank Keyulu Xu, Derek Lim, Behrooz Tahmasebi, Vikas Garg, Tommi Jaakkola, Andreas Loukas, Joan Bruna, and Yusu Wang for many discussions on the theory of GNNs, and Jingling Li, Mozhi Zhang, Simon Du, Ken-ichi Kawarabayashi, Weihua Hu, and Jure Leskovec for collaborations.

FUNDING

This work was partially supported by NSF CAREER award 1553284, SCALE MoDL award 2134108, and CCF-2112665 (TILOS AI Research Institute).

REFERENCES

[1] R. Abboud, I. I. Ceylan, M. Grohe, and T. Lukasiewicz, The surprising power of graph neural networks with random node initialization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[2] D. J. Aldous, Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** (1981), no. 4, 581–598.

[3] D. Anglouin, Local and global properties in networks of processors. In *Symposium on Theory of Computing (STOC)*, 1980.

[4] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Int. Conference on Machine Learning (ICML)*, 2019.

[6] V. Arvind, J. Köbler, G. Rattan, and O. Verbitsky, On the power of color refinement. In *International Symposium on Fundamentals of Computation Theory (FCT)*, pp. 339–350, Springer International Publishing, 2015.

[7] W. Azizian and M. Lelarge, Expressive power of invariant and equivariant graph neural networks. In *Int. Conf. on Learning Representations (ICLR)*, 2021.

[8] L. Babai, P. Erdős, and S. M. Selkow, Random graph isomorphism. *SIAM J. Comput.* **9** (1980), no. 3, 628–635.

[9] L. Babai and L. Kučera, Canonical labelling of graphs in linear average time. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 1979.

[10] P. Barceló, E. V. Kostylev, M. Monet, J. Pérez, J. L. Reutter, and J. P. Silva, The logical expressiveness of graph neural networks. In *Int. Conf. on Learning Representations (ICLR)*, 2020.

[11] P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** (2002), 463–482.

[12] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu, Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4502–4510, 2016.

[13] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, Relational inductive biases, deep learning, and graph networks. 2018, arXiv:1806.01261v3.

[14] R. Bellman, On a routing problem. *Quart. Appl. Math.* **16** (1958), 87–90.

[15] B. Bevilacqua, F. Frasca, D. Lim, B. Srinivasan, C. Cai, G. Balamurugan, M. M. Bronstein, and H. Maron, Equivariant Subgraph Aggregation Networks. 2021, arXiv:2110.02910v2.

[16] C. Bodnar, F. Frasca, N. Otter, Y. Guang Wang, P. Liò, G. Montúfar, and M. Bronstein, Weisfeiler and Lehman go cellular: CW networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] C. Bodnar, F. Frasca, N. Otter, Y. Guang Wang, P. Liò, G. Montúfar, and M. Bronstein, Weisfeiler and Lehman go topological: Message passing simplicial networks. In *Int. Conference on Machine Learning (ICML)*, 2021.

[18] J.-Y. Cai, M. Fürer, and N. Immerman, An optimal lower bound on the number of variables for graph identification. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 1989.

[19] J.-Y. Cai, M. Fürer, and N. Immerman, An optimal lower bound on the number of variables for graph identification. *Combinatorica* **12** (1992), no. 4, 389–410.

[20] Q. Cappart, D. Chételat, E. Khalil, A. Lodi, C. Morris, and P. Veličković, Combinatorial optimization and reasoning with graph neural networks. 2021, arXiv:2102.09544.

[21] M. Chen, X. Li, and T. Zhao, On generalization bounds of a family of recurrent neural networks. In *Proc. Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[22] Z. Chen, L. Chen, S. Villar, and J. Bruna, Can graph neural networks count substructures? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] Z. Chen, S. Villar, L. Chen, and J. Bruna, On the equivalence between graph isomorphism testing and function approximation with GNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[24] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** (1989), no. 4, 303–314.

[26] H. Dai, B. Dai, and L. Song, Discriminative embeddings of latent variable models for structured data. In *Int. Conference on Machine Learning (ICML)*, 2016.

[27] G. Dasoulas, L. Dos Santos, K. Scaman, and A. Virmaux, Coloring graph neural networks for node disambiguation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[28] N. Dehmamy, A. L. Barabási, and R. Yu, Understanding the representation power of graph neural networks in learning graph topology. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[29] P. Diaconis and S. Janson, Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* **VII** (2008), 33–61.

[30] S. S. Du, K. Hou, R. R. Salakhutdinov, B. Poczos, R. Wang, and K. Xu, Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[31] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, Gradient descent finds global minima of deep neural networks. In *Int. Conference on Machine Learning (ICML)*, 2019.

[32] S. S. Du, X. Zhai, B. Poczos, and A. Singh, Gradient descent provably optimizes over-parameterized neural networks. In *Int. Conf. on Learning Representations (ICLR)*, 2019.

[33] V. K. Garg, S. Jegelka, and T. Jaakkola, Generalization and representational limits of graph neural networks. In *Int. Conference on Machine Learning (ICML)*, 2020.

[34] F. Geerts, The expressive power of kth-order invariant graph networks. 2020, arXiv:2007.12035.

[35] F. Geerts, F. Mazowiecki, and G. A. Pérez, Let's agree to degree: Comparing graph convolutional networks in the message-passing framework. In *Int. Conference on Machine Learning (ICML)*, 2021.

[36] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, Neural message passing for quantum chemistry. In *Int. Conference on Machine Learning (ICML)*, 2017.

[37] M. Gori, G. Monfardini, and F. Scarselli, A new model for learning in graph domains. In *International Joint Conference on Neural Networks (IJCNN)*, 2005.

[38] M. Grohe and M. Otto, Pebble games and linear equations. *J. Symbolic Logic* **80** (2015), no. 3, 797–844.

[39] W. L. Hamilton, R. Ying, and J. Leskovec, Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[40] J. S. Hartford, D. R. Graham, K. Leyton-Brown, and S. Ravanbakhsh, Deep models of interactions across sets. In *Int. Conference on Machine Learning (ICML)*, 2018.

[41] L. Hella, M. Järvisalo, A. Kuusisto, J. Laurinharju, T. Lempiäinen, K. Luosto, J. Suomela, and J. Virtema, Weak models of distributed computing, with connections to modal logic. In *ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 185–194, Association for Computing Machinery New York, NY, United States, 2012.

[42] N. Immerman and E. S. Lander, Describing graphs: a first-order approach to graph canonization. In *Complexity theory retrospective*, pp. 59–81, Springer, 1990.

[43] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[44] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F. Li, C. L. Zitnick, and R. Girshick, Inferring and executing programs for visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] M. Karpinski and A. Macintyre, Polynomial bounds for the VC dimension of sigmoidal and general Pfaffian networks. *J. Comput. System Sci.* **54** (1997), no. 1, 169–176.

[46] P. Kelly, A congruence theorem for trees. *Pacific J. Math.* **7** (1957), 961–968.

[47] N. Keriven and G. Peyré, Universal invariant and equivariant graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[48] S. Kiefer, P. Schweitzer, and E. Selman, Graphs identified by logics with counting. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, 2015.

[49] T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks. In *Int. Conf. on Learning Representations (ICLR)*, 2017.

[50] P. Koiran and E. D. Sontag, Vapnik–Chervonenkis dimension of recurrent neural networks. In *European Conference on Computational Learning Theory*, pp. 223–237, 1997.

[51] A. N. Kolmogorov, On tables of random numbers. *Theoret. Comput. Sci.* **207** (1998), no. 2, 387–395.

[52] R. Kondor, H. Truong Son, H. Pan, B. M. Anderson, and S. Trivedi, Covariant compositional networks for learning graphs. In *International Conference on Learning Representations (ICLR) – Workshop Track*, 2018.

[53] R. Kondor and S. Trivedi, On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Int. Conference on Machine Learning (ICML)*, 2018.

[54] G. Lample and F. Charton, Deep learning for symbolic mathematics. In *Int. Conf. on Learning Representations (ICLR)*, 2020.

[55] R. Levie, W. Huang, L. Bucci, M. M. Bronstein, and G. Kutyniok, Transferability of spectral graph convolutional neural networks. *J. Mach. Learn. Res.* (2021).

[56] R. Liao, R. Urtasun, and R. Zemel, A PAC-Bayesian approach to generalization bounds for graph neural networks. In *Int. Conf. on Learning Representations (ICLR)*, 2021.

[57] N. Linial, Locality in distributed graph algorithms. *SIAM J. Comput.* **21** (1992), no. 1, 193–201.

[58] A. Loukas, How hard is to distinguish graphs with graph neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[59] A. Loukas, What graph neural networks cannot learn: depth vs width. In *Int. Conf. on Learning Representations (ICLR)*, 2020.

[60] A. Magner, M. Baranwal, and A. O. Hero, The power of graph convolutional networks to distinguish random graph models. In *IEEE International Symposium on Information Theory (ISIT)*, 2020.

[61] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Int. Conf. on Learning Representations (ICLR)*, 2019.

[62] H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman, Provably powerful graph networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[63] H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman, Invariant and equivariant graph networks. In *Int. Conf. on Learning Representations (ICLR)*, 2019.

[64] H. Maron, E. Fetaya, N. Segol, and Y. Lipman, On the universality of invariant networks. In *Int. Conference on Machine Learning (ICML)*, 2019.

[65] C. Merkwirth and T. Lengauer, Automatic generation of complementary descriptors with molecular graph networks. *J. Chem. Inf. Model.* **45** (2005), no. 5, 1159–1168.

[66] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, Weisfeiler and Leman go neural: Higher-order graph neural networks. In *Proc. AAAI Conference on Artificial Intelligence*, 2019.

[67] R. L. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro, Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *Int. Conf. on Learning Representations (ICLR)*, 2019.

[68] R. L. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro, Relational pooling for graph representations. In *Int. Conference on Machine Learning (ICML)*, 2019.

[69] M. Naor and L. J. Stockmeyer, What can be computed locally? In *Symposium on Theory of Computing (STOC)*, 1993.

[70] D. Peleg, *Distributed Computing: A Locality-Sensitive Approach*. Society for Industrial and Applied Mathematics, 2000.

[71] O. Puny, H. Ben-Hamu, and Y. Lipman, From graph low-rank global attention to 2-FWL approximation. In *Int. Conference on Machine Learning (ICML)*, 2020.

[72] C. R. Qi, H. Su, K. Mo, and L. G. PointNet, Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[73] S. Ravanbakhsh, J. Schneider, and B. Poczos, Deep Learning with Sets and Point Clouds. 2016, arXiv:1611.04500v3.

[74] S. Ravanbakhsh, J. Schneider, and B. Póczos, Equivariance through parameter-sharing. In *Int. Conference on Machine Learning (ICML)*, 2017.

[75] R. C. Read and D. G. Corneil, The graph isomorphism disease. *J. Graph Theory* **1** (1977), 339–363.

[76] L. Ruiz, L. F. O. Chamon, and A. Ribeiro, Graphon neural networks and the transferability of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[77] L. Ruiz, F. Gama, and A. Ribeiro, Graph neural networks: architectures, stability and transferability. *Proc. IEEE* **109** (2021), 660–682.

[78] A. Santoro, F. Hill, D. Barrett, A. Morcos, and T. Lillicrap, Measuring abstract reasoning in neural networks. In *Int. Conference on Machine Learning (ICML)*, pp. 4477–4486, 2018.

[79] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Oascanu, P. Battaglia, and T. Lillicrap, A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[80] R. Sato, M. Yamada, and H. Kashima, Approximation ratios of graph neural networks for combinatorial problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[81] R. Sato, M. Yamada, and H. Kashima, Random features strengthen graph neural networks. In *SIAM International Conference on Data Mining (SDM)*, 2021.

[82] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, Computational capabilities of graph neural networks. *IEEE Trans. Neural Netw.* **20** (2009), no. 1, 81–102.

[83] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, The graph neural network model. *IEEE Trans. Neural Netw.* **20** (2009), no. 1, 61–80.

[84] F. Scarselli, A. C. Tsoi, and M. Hagenbuchner, The Vapnik–Chervonenkis dimension of graph and recursive neural networks. *Neural Netw.* **108** (2018), 248–259.

[85] B. Schölkopf and A. Smola, *Learning with kernels*. Adapt. Comput. Mach. Learn., MIT Press, 2001.

[86] B. Tahmasebi, D. Lim, and S. Jegelka, Counting substructures with higher-order graph neural networks: possibility and impossibility results. 2021, arXiv:2012.03174v2.

[87] A. Trask, F. Hill, S. E. Reed, J. Rae, C. Dyer, and P. Blunsom, Neural arithmetic logic units. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[88] S. M. Ulam, *A collection of mathematical problems*. Interscience Publishers, 1960.

[89] V. N. Vapnik and A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** (1971), no. 2, 264–280.

[90] P. Velickovic, R. Ying, M. Padovano, R. Hadsell, and C. Blundell, Neural execution of graph algorithms. In *Int. Conf. on Learning Representations (ICLR)*, 2020.

[91] S. Verma and Z.-L. Zhang, Stability and generalization of graph convolutional neural networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1539–1548, 2019.

[92] E. Wagstaff, F. B. Fuchs, M. Engelcke, I. Posner, and M. Osborne, On the limitations of representing functions on sets. In *Int. Conference on Machine Learning (ICML)*, 2019.

[93] B. Weisfeiler, *On construction and identification of graphs*. Springer, 1976.

[94] B. Weisfeiler and A. A. Leman, A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia* **2** (1968), no. 9, 12–16.

[95] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, How powerful are graph neural networks? In *Int. Conf. on Learning Representations (ICLR)*, 2019.

[96] K. Xu, J. Li, M. Zhang, S. Du, K. Kawarabayashi, and S. Jegelka, What can neural networks reason about? In *Int. Conf. on Learning Representations (ICLR)*, 2020.

[97] K. Xu, M. Zhang, J. Li, S. Du, K. Kawarabayashi, and S. Jegelka, How neural networks extrapolate: From feedforward to graph neural networks. In *Int. Conf. on Learning Representations (ICLR)*, 2021.

[98] D. Yarotsky, Universal approximations of invariant maps by neural networks. *Constr. Approx.* (2021).

[99] G. Yehudai, E. Fetaya, E. Meirom, G. Chechik, and H. Maron, From local structures to size generalization in graph neural networks. In *Int. Conference on Machine Learning (ICML)*, 2021.

[100] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[101] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, Deep sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[102] Q. Zhao, Z. Ye, C. Chen, and Y. Wang, Persistence enhanced graph neural network. In *Proc. Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

STEFANIE JEGELKA

Department of EECS, MIT, Cambridge, USA, stefje@mit.edu

THEORY OF ADAPTIVE ESTIMATION

OLEG V. LEPSKI

ABSTRACT

The paper is an introduction to the modern theory of adaptive estimation. We introduce a universal estimation procedure based on a random choice from collections of estimators satisfying a few very general assumptions. In the framework of an abstract statistical model, we present an upper bound for the risk of the proposed estimator (ℓ-oracle inequality). The basic technical tools here are a commutativity property of some operators and upper functions for positive random functionals. Since the obtained result is not related to a particular observation scheme, many conclusions for various problems in different statistical models can be derived from the single ℓ-oracle inequality.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 62G05; Secondary 62G20, 62G07, 62G08

KEYWORDS

Abstract statistical experiment, minimax approach, adaptive estimation, oracle inequality, selection rule.

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5478–5498
DOI 10.4171/ICM2022/15

1. INTRODUCTION

Let $(V^{(n)}, \mathfrak{A}^{(n)}, \mathbb{P}_f^{(n)}, f \in \mathfrak{F})$, $n \in \mathbb{N}^*$, be a family of statistical experiments generated by observation $X^{(n)}$. It means that $X^{(n)}$ is a $V^{(n)}$-valued random variable defined on some probability space, and the probability law of $X^{(n)}$ belongs to the family $(\mathbb{P}_f^{(n)}, f \in \mathfrak{F})$. Since the probability space on which $X^{(n)}$ is defined will play no role in the sequel, we will just assume its existence.

Furthermore, in this paper:

- $(\mathcal{D}, \mathfrak{D}, \mu)$ is a measurable space;
- $\mathfrak{F}$ is a set of functions $f : \mathcal{D} \to \mathbb{R}$. Typical examples of set $\mathfrak{F}$ are functional spaces, e.g., $\mathfrak{F} = \mathbb{L}_2(\mathbb{R}^d)$, $\mathbb{C}_b(\mathbb{R}^d)$, the set of all measurable real functions, etc.;
- $G : \mathfrak{F} \to \mathfrak{S}$, where $\mathfrak{S}$ is a set endowed with semimetric ℓ.

The goal is to estimate $G(f)$, $f \in \mathfrak{F}$, from observation $X^{(n)}$. By an estimator we mean any $X^{(n)}$-measurable $\mathfrak{S}$-valued mapping. The accuracy of an estimator $\tilde{G}$ is measured by the ℓ-risk

$$\mathcal{R}_n^{(\ell)}\big[\tilde{G}; G(f)\big] = \big(\mathbb{E}_f^{(n)}\big[\ell\big(\tilde{G}, G(f)\big)\big]^q\big)^{\frac{1}{q}}. \tag{1.1}$$

Here and later, $\mathbb{E}_f^{(n)}$ denotes the mathematical expectation with respect to the probability measure $\mathbb{P}_f^{(n)}$ and the number $q \geq 1$ is supposed to be fixed. Recall that for any $X^{(n)}$-measurable map $T : V^{(n)} \to \mathbb{R}$,

$$\mathbb{E}_f^{(n)}[T] = \int_{V^{(n)}} T(v)\mathbb{P}_f^{(n)}(\mathrm{d}v).$$

1.1. Examples of models

In these notes we will consider the following statistical models.

Density model. Let $\mathfrak{P}(\mathcal{D}, \mu)$ denote the set of all probability densities with respect to measure μ defined on $\mathcal{D}$ and let $\mathfrak{F} \subseteq \mathfrak{P}(\mathcal{D}, \mu)$.

Then the statistical experiment is generated by the observation $X^{(n)} = (X_1, \dots, X_n)$, $n \in \mathbb{N}^*$, where $X_i, i \in \mathbb{N}^*$, are i.i.d. random vectors possessing unknown density $f \in \mathfrak{F}$.

White Gaussian Noise Model. Let $\mathfrak{F} = \mathbb{L}_2(\mathcal{D}, \mu)$. Put $\tilde{\mathfrak{D}} = \{B \in \mathfrak{D} : \mu(B) < \infty\}$ and let $(W(B), B \in \tilde{\mathfrak{D}})$ be the white noise with intensity μ.

Consider the statistical model generated by the observation $X^{(n)} = \{X_n(g), g \in \mathbb{L}_2(\mathcal{D}, \mu)\}$ where

$$X_n(g) = \int_{\mathcal{D}} f(t)g(t)\mu(\mathrm{d}t) + n^{-1/2}\int_{\mathcal{D}} g(t)W(\mathrm{d}t). \tag{1.2}$$

Recall also that for any $g \in \mathbb{L}_2(\mathcal{D}, \mu)$,

$$X_n(g) \backsim \mathcal{N}\big(\langle g, f\rangle, n^{-1}\langle g, g\rangle\big), \tag{1.3}$$

where $\langle \cdot, \cdot \rangle$ is the inner product of $\mathbb{L}_2(\mathcal{D}, \mu)$ and $\mathcal{N}(\cdot, \cdot)$ denotes the normal law on $\mathbb{R}$.

1.2. Examples of estimation targets G

Global estimation $G(f) = f$. The goal is to estimate the entire function f. Here $\mathfrak{S} = \mathfrak{F}$, and the accuracy of estimation is usually measured by the $\mathbb{L}_p$-risk on $D \subseteq \mathcal{D}$, i.e., $\ell(g_1, g_2) = \|g_1 - g_2\|_{p,D}$, $1 \le p \le \infty$, where

$$\|g\|_{p,D}^p = \int_D |g|^p \mu(\mathrm{d}t), \quad p \in [1, \infty), \quad \|g\|_{\infty,D} = \sup_{t \in D} \big|g(t)\big|.$$

Pointwise estimation $G(f) = f(t_0), t_0 \in D$. Here $\mathfrak{S} = \mathbb{R}^1$ and $\ell(a,b) = |a-b|, a, b \in \mathbb{R}$, and $D \subseteq \mathcal{D}$. We present this estimation problem separately from the below-discussed problems of estimation of functionals because it is often used in order to recover the underlying function itself.

Estimation of functionals. Here $\mathfrak{S} = \mathbb{R}^1$ and $\ell(a,b) = |a-b|$, $a, b \in \mathbb{R}$, and $D \subseteq \mathcal{D}$. One can consider

- Estimation of a derivative at a given point, $G(f) = f^{(k)}(t_0)$, $t_0 \in D$, $k \in \mathbb{N}^*$;
- Estimation of norms, $G(f) = \|f\|_{p,D}$, $1 \le p \le \infty$;
- Estimation of extreme points, $G(f) = \arg\max_{t \in D} f(t)$;
- Estimation of regular functionals, for example, $G(f) = \int_D f^s(t)\mathrm{d}t$, $s \in \mathbb{N}^*$.

2. MINIMAX ADAPTIVE ESTIMATION

Let $\mathbb{F}$ be a given subset of $\mathfrak{F}$. For any estimator $\tilde{G}_n$, define its *maximal risk* on $\mathbb{F}$ by

$$\mathcal{R}_n^{(\ell)}[\tilde{G}_n; \mathbb{F}] = \sup_{f \in \mathbb{F}} \mathcal{R}_n^{(\ell)}\big[\tilde{G}_n; G(f)\big]$$

and the *minimax risk* on $\mathbb{F}$ is given by

$$\phi_n(\mathbb{F}) := \inf_{\tilde{G}_n} \mathcal{R}_n^{(\ell)}[\tilde{G}_n; \mathbb{F}], \tag{2.1}$$

where the infimum is always taken over all possible estimators. An estimator whose maximal risk is proportional to $\phi_n(\mathbb{F})$ is called a minimax on $\mathbb{F}$.

Let $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\}$ be the collection of subsets of $\mathfrak{F}$, where ϑ is a nuisance parameter which may have very complicated structure (see the examples below). Without further mentioning, we will consider only scales of functional classes for which a minimax on $\mathbb{F}_\vartheta$ estimator (usually depending on ϑ) exists for any $\vartheta \in \Theta$.

The problem of adaptive estimation can be formulated as follows: *Is it possible to construct a single estimator $\hat{G}_n$ which is simultaneously minimax on each class $\mathbb{F}_\vartheta$, $\vartheta \in \Theta$, i.e., such that*

$$\limsup_{n \to \infty} \phi_n^{-1}(\mathbb{F}_\vartheta) \mathcal{R}_n^{(\ell)}[\hat{G}_n; \mathbb{F}_\vartheta] < \infty, \quad \forall \vartheta \in \Theta?$$

We refer to this question as *the problem of minimax adaptive estimation over the scale of classes* $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\}$. If such an estimator exists, we will call it optimally-adaptive, or rate-adaptive.

The first adaptive results were obtained in [14]. Starting from this pioneering paper, a variety of adaptive methods were proposed in different statistical models such as density and spectral density estimation, nonparametric regression, deconvolution model, inverse problems, and many others. The interested reader can find a very detailed overview of this topic in [31]. Here we only mention several methods allowing one to construct optimally-adaptive estimators:

- Extension of Efroimovich–Pinsker method [11,12];
- Lepski method [27] and its extension, namely Goldenshluger–Lepski method [18];
- Unbiased risk minimization [20,21];
- Wavelet thresholding [10];
- Model selection [1,2];
- Aggregation of estimators [3,15,23,37,42,43];
- Exponential weights [9,36,40];
- Risk hull method [7];
- Blockwise Stein method [4,8,39].

We will discuss existence of optimally-adaptive estimators in details later. Now let us provide some example of scales of functional classes over which the adaptation is studied.

2.1. Scales of functional classes

2.1.1. Classes of smooth functions

Let $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_d)$ denote the canonical basis of $\mathbb{R}^d$, $d \in \mathbb{N}^*$. For a function $T : \mathbb{R}^d \to \mathbb{R}^1$ and real number $u \in \mathbb{R}$, the first-order difference operator with step size u in the direction of the variable x_j is defined by $\Delta_{u,j} T(x) = T(x + u\boldsymbol{e}_j) - T(x)$, $j = 1, \dots, d$. By induction, the kth-order difference operator is

$$\Delta_{u,j}^k T(x) = \Delta_{u,j} \Delta_{u,j}^{k-1} T(x) = \sum_{l=1}^{k} (-1)^{l+k} \binom{k}{l} \Delta_{ul,j} T(x).$$

Definition 2.1. For given vectors $\vec{\beta} = (\beta_1, \dots, \beta_d) \in (0, \infty)^d$, $\vec{r} = (r_1, \dots, r_d) \in [1, \infty]^d$, and $\vec{L} = (L_1, \dots, L_d) \in (0, \infty)^d$, a function $T : \mathbb{R}^d \to \mathbb{R}^1$ is said to belong to anisotropic Nikolskii's class $\mathbb{N}_{\vec{r},d}(\vec{\beta}, \vec{L})$ if $\|T\|_{r_j} \le L_j$ for all $j = 1, \dots, d$, and there exist natural numbers $k_j > \beta_j$ such that

$$\left\| \Delta_{u,j}^{k_j} T \right\|_{r_j} \le L_j |u|^{\beta_j}, \quad \forall u \in \mathbb{R}, \ \forall j = 1, \dots, d.$$

Let $\mathfrak{F} = \bigcup_{q \ge 1}^{d} \mathbb{L}_q(\mathbb{R}^d)$ and

$$\mathbb{F}_\vartheta = \mathbb{N}_{\vec{r},d}(\vec{\beta}, \vec{L}), \quad \vartheta = (\vec{\beta}, \vec{r}, \vec{L}) \in \Theta \subseteq (0, \infty)^d \times [1, \infty]^d \times (0, \infty)^d,$$

where $\mathbb{N}_{\vec{r},d}(\vec{\beta}, \vec{L})$ is the anisotropic Nikolskii's class of functions on $\mathbb{R}^d$, $d \ge 1$.

2.1.2. Functional classes with structure

Structural models are usually used in estimation of multivariate functions in order to improve estimation accuracy and to overcome the curse of the dimensionality.

Single index structure. Let $\mathfrak{F} = \bigcup_{q\geq 1}^{d} \mathbb{L}_q(\mathbb{R}^d)$ and let $\mathbb{S}^{d-1}$, $d \geq 2$, denote the unit sphere in $\mathbb{R}^d$. Let also $\mathbb{N}_{r,1}(\beta, L)$, $r \geq 1$, $\beta > 0$, $L > 0$ be the Nikolskii's class of functions on $\mathbb{R}^1$.

For any $\mathcal{S} \subseteq \mathbb{S}^{d-1}$ and any $r \geq 1$, $\beta > 0$, $L > 0$, introduce the following functional class:

$$\mathcal{F}_r^{\text{single}}(\beta, L, \mathcal{S}) = \left\{ f : \mathbb{R}^d \to \mathbb{R}^1 : f(\cdot) = F\left(\omega^{\intercal}\cdot\right), F \in \mathbb{N}_{r,1}(\beta, L), \omega \in \mathcal{S} \right\}.$$

The adaptive estimation over the collection

$$\mathbb{F}_\vartheta = \mathcal{F}_r^{\text{single}}(\beta, L, \mathcal{S}), \quad \vartheta = (\beta, r, L, \mathcal{S}) \in \Theta \subseteq (0,\infty) \times [1,\infty] \times (0,\infty) \times \mathbb{S}^{d-1}$$

is called the estimation under the single-index constraint.

Additive structure. Let as previously $\mathfrak{F} = \bigcup_{q\geq 1}^{d} \mathbb{L}_q(\mathbb{R}^d)$, $d \geq 2$, and let $\mathbb{N}_{r,1}(\beta, L)$, $r \geq 1$, $\beta > 0$, $L > 0$ denote the Nikolskii's class of functions on $\mathbb{R}^1$.

For any $r \geq 1$, $\beta > 0$, $L > 0$, introduce the following functional class:

$$\mathcal{F}_r^{\text{additive}}(\beta, L, \mathcal{S}) = \left\{ f : \mathbb{R}^d \to \mathbb{R}^1 : f(x) = \sum_{k=1}^{d} F_k(x_k), F_k \in \mathbb{N}_{r,1}(\beta, L) \right\}.$$

The adaptive estimation over the collection

$$\mathbb{F}_\vartheta = \mathcal{F}_r^{\text{additive}}(\beta, L), \quad \vartheta = (\beta, r, L) \in \Theta \subseteq (0,\infty) \times [1,\infty] \times (0,\infty)$$

is called the estimation under the additive constraint.

The functional classes introduced above are considered in the framework of Gaussian White Noise Model or, more generally, in nonparametric regression context.

Hypothesis of independence. The functional classes introduced below are used in the Density Model. Let $\mathcal{D} = \mathbb{R}^d$, $d \geq 2$, μ be the Lebesgue measure and recall that $\mathfrak{F} \subseteq \mathfrak{P}(\mathcal{D}, \mu)$. At last, let $\mathcal{I}_d$ be the set of all subsets of $\{1, \dots, d\}$.

For any $I \in \mathcal{I}_d$ and any $x \in \mathbb{R}^d$, denote $x_I = \{x_i \in \mathbb{R}, j \in I\}$, $\bar{I} = \{1, \dots, d\} \setminus I$, and set for any density $f \in \mathfrak{F}$,

$$f_I(x_I) = \int_{\mathbb{R}^{\bar{I}}} f(x) \mathrm{d}x_{\bar{I}}, \quad x_I \in \mathbb{R}^{|I|}.$$

If we denote the coordinates of the random vector X_i by $X_{i,1}, \dots, X_{i,d}$, we can assert that f_I is the marginal density of the random vector $X_{i,I} := (X_{i,j}, j \in I)$ for any $i = 1, \dots, n$. The latter is true because X_i, $i = 1, \dots, n$, are identically distributed.

Let Π denote the set of all partitions of $\{1, \dots, d\}$. The independence hypothesis supposes that there exists a partition $\mathcal{P}$ such that the random vectors $X_{1,I}$, $I \in \mathcal{P}$, are mutually independent, meaning that

$$f(x) = \prod_{I \in \mathcal{P}} f_I(x_I), \quad \forall x \in \mathbb{R}^d.$$

For given vectors $\vec{\beta} = (\beta_1, \dots, \beta_d) \in (0, \infty)^d$, $\vec{r} = (r_1, \dots, r_d) \in [1, \infty]^d$, $\vec{L} = (L_1, \dots, L_d) \in (0, \infty)^d$ and a given partition $\mathcal{P} \in \Pi$, introduce the following functional class:

$$\mathcal{F}^{\text{indep}}_{\vec{r}}(\vec{\beta}, \vec{L}, \mathcal{P}) = \left\{ f : \mathbb{R}^d \to \mathbb{R}_+ : f(x) = \prod_{I \in \mathcal{P}} f_I(x_I),\, f_I \in \mathbb{N}_{r_I, |I|}(\beta_I, L_I), I \in \mathcal{P} \right\}.$$

The adaptive estimation over the collection

$$\mathbb{F}_\vartheta = \mathcal{F}^{\text{indep}}_{\vec{r}}(\vec{\beta}, \vec{L}, \mathcal{P}), \quad \vartheta = (\vec{\beta}, \vec{r}, \vec{L}) \in \Theta \subseteq (0, \infty)^d \times [1, \infty]^d \times (0, \infty)^d \times \Pi$$

is called the estimation under hypothesis of independence.

2.2. Existence of adaptive estimators. Fundamental problem

It is well-known that optimally-adaptive estimators do not always exist, see [5,13,26,28]. Formally, the nonexistence of optimally-adaptive estimator means that

$$\liminf_{n\to\infty} \inf_{\tilde{G}_n} \sup_{\vartheta \in \{\vartheta_1, \vartheta_2\}} \phi_n^{-1}(\mathbb{F}_\vartheta) \mathcal{R}_n^{(\ell)}[\tilde{G}_n; \mathbb{F}_\vartheta] = \infty, \quad \forall \vartheta_1, \vartheta_2 \in \Theta. \tag{2.2}$$

Indeed, since a minimax estimator on $\mathbb{F}_\vartheta$ exists for any $\vartheta \in \Theta$, we can assert that

$$0 < \liminf_{n\to\infty} \inf_{\tilde{G}_n} \phi_n^{-1}(\mathbb{F}_\vartheta) \mathcal{R}_n^{(\ell)}[\tilde{G}_n; \mathbb{F}_\vartheta] < \infty, \quad \forall \vartheta \in \Theta.$$

The latter result means that the optimal (from the minimax point of view) family of normalizations $\{\phi_n(\mathbb{F}_\vartheta), \vartheta \in \Theta\}$ is attainable for each value ϑ, while (2.2) shows that this family is unattainable by any estimation procedure simultaneously for any couple of elements from Θ. This, in its turn, implies that optimally-adaptive over the scale $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\}$ does not exist.

However, the question of constructing a single estimator for all values of the nuisance parameter $\vartheta \in \Theta$ remains relevant. Hence, if (2.2) holds, we need to find an attainable family of normalization and to prove its optimality. The realization of this program dates back to [27] where the notion of *adaptive rate of convergence* was introduced. Nowadays there exist several definitions of adaptive rate of convergence and corresponding to this notion criteria of optimality, see [25,27,38,41]. Here we present the simplest definition of the adaptive rate which is the following.

Definition 2.2. A normalization family $\{\psi_n(\mathbb{F}_\vartheta), \vartheta \in \Theta\}$ is called an adaptive rate of convergence over collection of functional classes $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\}$ if

$$\liminf_{n\to\infty} \inf_{\tilde{G}_n} \sup_{\vartheta \in \{\vartheta_1, \vartheta_2\}} \psi_n^{-1}(\mathbb{F}_\vartheta) \mathcal{R}_n^{(\ell)}[\tilde{G}_n; \mathbb{F}_\vartheta] > 0, \quad \forall \vartheta_1, \vartheta_2 \in \Theta, \tag{2.3}$$

and there exists an estimator $\hat{G}_n$ such that

$$\limsup_{n\to\infty} \sup_{\vartheta \in \{\vartheta_1, \vartheta_2\}} \psi_n^{-1}(\mathbb{F}_\vartheta) \mathcal{R}_n^{(\ell)}[\hat{G}_n; \mathbb{F}_\vartheta] < \infty, \quad \forall \vartheta_1, \vartheta_2 \in \Theta. \tag{2.4}$$

The sequence $\sup_{\vartheta \in \Theta}[\psi_n(\vartheta)/\varphi_n(\vartheta)]$ is called the price to pay for adaptation, and the estimator $\hat{G}_n$ is called an adaptive estimator.

Note that (2.4) is equivalent to

$$\limsup_{n\to\infty} \psi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}[\hat{G}_n;\mathbb{F}_\vartheta] < \infty, \quad \forall \vartheta \in \Theta$$

and, therefore, if (2.4) is fulfilled for any $n \in \mathbb{N}^*$ with

$$\psi_n(\vartheta) = c(\vartheta)\phi_n(\vartheta), \quad c(\vartheta) < \infty, \quad \forall \vartheta \in \Theta,$$

then one can assert that $\hat{G}_n$ is an *optimally-adaptive estimator*.

Example 2.3. Consider univariate model (1.2), where $\mathcal{D} = [0, 1]$ and μ is the Lebesgue measure. Let also $\mathbb{F}_\vartheta = \mathbb{N}_{\infty,1}(\beta, L)$, $\vartheta = (\beta, L)$, be the collection of Nikolskii's classes with $r = \infty$ (Hölder's classes). Let $b, \mathcal{L} > 0$ be arbitrary but a priori chosen numbers, and let $\Theta = (0, b] \times (0, \mathcal{L}]$. The goal is to estimate $G(f) = f(a)$ where $a \in (0, 1)$ is a given point.

The minimax rate of convergence for this problem is given by

$$\phi_n\big(\mathbb{N}_{\infty,1}(\beta, L)\big) = (L^{\frac{1}{\beta}}/n)^{\frac{\beta}{2\beta+1}},$$

while the adaptive rate of convergence is given, see [26], by

$$\psi_n\big(\mathbb{N}_{\infty,1}(\beta, L)\big) = \big(L^{\frac{1}{\beta}} \ln(n)/n\big)^{\frac{\beta}{2\beta+1}}.$$

We conclude that optimally-adaptive estimators do not exist in this estimation problem.

The most challenging problem of the adaptive theory is to understand how the existence/nonexistence of optimally-adaptive estimators depends on the statistical model, underlying estimation problem (mapping G), loss functional ℓ, and the collection of considered classes. An attempt to provide such classification was undertaken in [27, 28], but the sufficient conditions found there for both the existence and nonexistence of optimally-adaptive estimators turned out to be too restrictive.

Problem. Find necessary and sufficient conditions of the existence of optimally-adaptive estimators, i.e., the existence of an estimator $\hat{G}_n$ satisfying the following property:

$$\limsup_{n\to\infty} \phi_n^{-1}(\mathbb{F}_\vartheta)\mathcal{R}_n^{(\ell)}[\hat{G}_n;\mathbb{F}_\vartheta] < \infty, \quad \forall \vartheta \in \Theta.$$

This problem stated in [27] 30 years ago remains unsolved.

It is important to realize that answers to the formulated problem may be different even if the statistical model and the collection of functional classes are the same and estimation problems have "similar nature."

Example 2.4. Consider the univariate model (1.2), where $\mathcal{D} = [0, 1]$ and μ is the Lebesgue measure. Let also $\mathbb{F}_\vartheta = \mathbb{N}_{\infty,1}(\beta, L)$, $\vartheta = (\beta, L)$, be the collection of Nikolskii's classes with $r = \infty$ (Hölder's classes). Let $b, \mathcal{L} > 0$ be arbitrary but a priori chosen numbers, and let $\Theta = (0, b] \times (0, \mathcal{L}]$. Set

$$G_\infty(f) = \|f\|_{\infty,[0,1]}, \quad G_2(f) = \|f\|_{2,[0,1]}.$$

The optimally-adaptive estimator of $G_\infty(\cdot)$, was constructed in [29]. On the other hand, there is no optimally-adaptive estimator for $G_2(\cdot)$, see [6].

2.3. Adaptive estimation via the oracle approach

Let $\mathcal{G} = \{\hat{G}_\mathfrak{h}, \mathfrak{h} \in \mathfrak{H}\}$ be a family of estimators built from the observation $X^{(n)}$. The goal is to propose a data-driven (based on $X^{(n)}$) selection procedure from the collection $\mathcal{G}$ and establish for it an ℓ-oracle inequality.

More precisely, we want to construct an $\mathfrak{H}$-valued random element $\hat{\mathfrak{h}}$ completely determined by the observation $X^{(n)}$ and to prove that for any $n \geq 1$,

$$\mathcal{R}_n^{(\ell)}\big[\hat{G}_{\hat{\mathfrak{h}}}; G(f)\big] \leq \inf_{\mathfrak{h}\in T} U_n^{(\ell)}(f, \mathfrak{h}) + r_n, \quad \forall f \in \mathfrak{F}. \tag{2.5}$$

We call (2.5) an *ℓ-oracle inequality*. Here $r_n \to 0$, $n \to \infty$ is a given sequence which may depend on $\mathfrak{F}$ and the family of estimators $\mathcal{G}$ only. As to the quantity $U_n^{(\ell)}(\cdot,\cdot)$, it is explicitly expressed, and for some particular problems one can prove inequality (2.5) with

$$U_n^{(\ell)}(f, \mathfrak{h}) = C\,\mathcal{R}_n^{(\ell)}\big[\hat{G}_\mathfrak{h}; G(f)\big], \tag{2.6}$$

where C is a constant which may depend on $\mathfrak{F}$ and the family of estimators $\mathcal{G}$ only.

Historically, inequality (2.5) with $U_n^{(\ell)}(\cdot,\cdot)$ given in (2.6) was called the oracle inequality. The latter means that the "oracle" knowing the true parameter f can construct the estimator $\hat{G}_{\mathfrak{h}(f)}$ which provides the minimal over the collection $\mathcal{G}$ risk for any $f \in \mathfrak{F}$, that is,

$$\mathfrak{h}(f) : \mathcal{R}_n^{(\ell)}\big[\hat{G}_{\mathfrak{h}(f)}; G(f)\big] = \inf_{\mathfrak{h}\in\mathfrak{H}} \mathcal{R}_n^{(\ell)}\big[\hat{G}_\mathfrak{h}; G(f)\big].$$

Since $\mathfrak{h}(f)$ depends on unknown f, the estimator $\hat{G}_{\mathfrak{h}(f)}$, called oracle estimator, is not an estimator in the usual sense and, therefore, cannot be used. The goal is to construct the estimator $\hat{G}_{\hat{\mathfrak{h}}}$ which "mimics" the oracle one.

It is worth noting that the ℓ-oracle inequality with $U_n^{(\ell)}(\cdot,\cdot)$ given in (2.6) is not always available, and this is the reason why we deal with a more general definition given by (2.5).

The important remark is that inequality (2.5) provides a very simple criterion allowing one to assert that the selected estimator $\hat{G}_{\hat{\mathfrak{h}}}$ is optimally-adaptive, or adaptive with respect to the scale of functional classes $\{\mathbb{F}_\vartheta, \vartheta \in \Theta\}$. Indeed, let us assume that

(i) $r_n \leq C \inf_{\vartheta\in\Theta} \phi_n(\mathbb{F}_\vartheta)$ for some $C > 0$ (verified for all known problems);

(ii) $\exists \vartheta \mapsto \mathfrak{h}(\vartheta)$ and $c(\vartheta) > 0$ such that

$$\sup_{f\in\mathbb{F}_\vartheta} U_n^{(\ell)}\big(f, \mathfrak{h}(\vartheta)\big) \leq c(\vartheta)\phi_n(\mathbb{F}_\vartheta), \quad \forall \vartheta \in \Theta.$$

Hence we deduce from (2.5) that, for any $\vartheta \in \Theta$,

$$\sup_{f\in\mathbb{F}_\vartheta} \mathcal{R}_n^{(\ell)}\big[\hat{G}_{\hat{\mathfrak{h}}}; G(f)\big] \leq \sup_{f\in\mathbb{F}_\vartheta} U_n^{(\ell)}\big(f, \mathfrak{h}(\vartheta)\big) + r_n \leq \big(c(\vartheta) + C\big)\phi_n(\mathbb{F}_\vartheta),$$

and, therefore, we can assert that $\hat{G}_{\hat{\mathfrak{h}}}$ is optimally-adaptive. If (i) and (ii) hold with $\psi_n(\mathbb{F}_\vartheta)$ instead of $\phi_n(\mathbb{F}_\vartheta)$, where $\psi_n(\mathbb{F}_\vartheta)$ is the adaptive rate of convergence, we can state that $\hat{G}_{\hat{\mathfrak{h}}}$ is an adaptive estimator.

3. UNIVERSAL SELECTION RULE AND ℓ-ORACLE INEQUALITY

Our objective now is to propose a data-driven selection rule from a family of estimators satisfying few very general assumptions and to establish for it an ℓ-oracle inequality (2.5). It is important to emphasize that we provide an explicit expression of the functional $U_n^{(\ell)}(\cdot,\cdot)$ that allows us to derive various adaptive results from the unique oracle inequality. The proposed approach can be viewed as a generalization of several estimation procedures developed by the author and his collaborators during last 20 years, see [16–19,22,24,30,31,34].

3.1. Assumptions

Let $\mathfrak{H}_n$, $n \in \mathbb{N}^*$, be a sequence of countable subsets of $\mathfrak{H}$. Let $\{\hat{G}_\mathfrak{h}, \mathfrak{h} \in \mathfrak{H}\}$ and $\{\hat{G}_{\mathfrak{h},\eta}, \mathfrak{h}, \eta \in \mathfrak{H}\}$ be the families of $X^{(n)}$-measurable $\mathfrak{S}$-valued mappings possessing the properties formulated below. Both $\hat{G}_h$ and $\hat{G}_{\mathfrak{h},\eta}$ usually depend on n, but we will omit this dependence for the sake of simplicity of notations.

Let $\varepsilon_n \to 0$, $n \to \infty$, and δ_n, $n \to \infty$, be two given sequences. Suppose there exist collections of $\mathfrak{S}$-valued functionals $\{\Lambda_\mathfrak{h}(f), \mathfrak{h} \in \mathfrak{H}\}$, $\{\Lambda_{\mathfrak{h},\eta}(f), \mathfrak{h}, \eta \in \mathfrak{H}\}$, and a collection of *positive* $X^{(n)}$-measurable random variables $\Psi_n = \{\Psi_n(\mathfrak{h}), \mathfrak{h} \in \mathfrak{H}\}$ for which the following conditions hold (the functionals $\Lambda_\mathfrak{h}$ and $\Lambda_{\mathfrak{h},\eta}$ may depend on n (not necessarily) but we will omit this dependence in the notations):

($\mathbf{A}^{\text{permute}}$) *For any* $f \in \mathfrak{F}$ *and* $n \geq 1$,

$$\text{either} \quad \text{(i)} \quad \hat{G}_{\mathfrak{h},\eta}(f) = \hat{G}_{\eta,\mathfrak{h}}(f), \quad \forall \eta, \mathfrak{h} \in \mathfrak{H};$$
$$\text{or} \quad \text{(ii)} \quad \sup_{\mathfrak{h},\eta \in \mathfrak{H}_n} \ell\big(\Lambda_{\mathfrak{h},\eta}(f), \Lambda_{\eta,\mathfrak{h}}(f)\big) \leq \delta_n.$$

($\mathbf{A}^{\text{upper}}$) *For any* $f \in \mathfrak{F}$ *and* $n \geq 1$,

$$\text{(i)} \quad \mathbb{E}_f^{(n)}\Big(\sup_{\mathfrak{h} \in \mathfrak{H}_n} \big[\ell\big(\hat{G}_\mathfrak{h}, \Lambda_\mathfrak{h}(f)\big) - \Psi_n(\mathfrak{h})\big]_+^q \Big) \leq \varepsilon_n^q;$$
$$\text{(ii)} \quad \mathbb{E}_f^{(n)}\Big(\sup_{\mathfrak{h},\eta \in \mathfrak{H}_n} \big[\ell\big(\hat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}(f)\big) - \big\{\Psi_n(\mathfrak{h}) \wedge \Psi_n(\eta)\big\}\big]_+^q \Big) \leq \varepsilon_n^q.$$

Some remarks are in order.

1) Assumption ($\mathbf{A}^{\text{permute}}$)(i) was called in [18] the *commutativity property*. The selection rule presented in the next section was proposed in [33] and an ℓ-oracle inequality was established under Assumptions ($\mathbf{A}^{\text{upper}}$)(i) and ($\mathbf{A}^{\text{permute}}$). However, it turned out that for some estimator collections Assumption ($\mathbf{A}^{\text{permute}}$)(i) is not verified. So our main objective is to prove the same (up to absolute constants) ℓ-oracle inequality under assumptions ($\mathbf{A}^{\text{permute}}$)(ii) and ($\mathbf{A}^{\text{upper}}$).

2) For many statistical models and problems,

$$\Lambda_\mathfrak{h}(f) = \mathbb{E}_f^{(n)}(\hat{G}_\mathfrak{h}), \quad \Lambda_{\mathfrak{h},\eta}(f) = \mathbb{E}_f^{(n)}(\hat{G}_{\mathfrak{h},\eta}).$$

In this case $\ell(\hat{G}_\mathfrak{h}, \Lambda_\mathfrak{h}(f))$ and $\ell(\hat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}(f))$ can be viewed as stochastic errors related to the estimators $\hat{G}_\mathfrak{h}$ and $\hat{G}_{\mathfrak{h},\eta}$, respectively. Hence, following the terminology used in [32], we can say that $\{\Psi_n(\mathfrak{h}), \mathfrak{h} \in \mathfrak{H}\}$ and $\{\Psi_n(\mathfrak{h}) \wedge \Psi_n(\eta), \mathfrak{h}, \eta \in \mathfrak{H}\}$ are upper functions of level ε_n for the collection of corresponding stochastic errors. Often the collection $\{\Psi_n(\mathfrak{h}), \mathfrak{h} \in \mathfrak{H}\}$ is

not random. This is typically the case when a statistical problem is studied in the framework of white Gaussian noise or regression model.

3) We consider countable $\mathfrak{S}_n$ in order not to discuss of the measurability of the supremum inside the mathematical expectation appearing in Assumption ($\mathbf{A}^{\mathrm{upper}}$). The theory developed in the next section remains valid for any parameter set over which the corresponding supremum is $X^{(n)}$-measurable.

3.2. Universal selection rule and corresponding ℓ-oracle inequality

Our objective is to propose the selection rule from an arbitrary collection $\mathcal{G}(\mathfrak{S}_n) = \{\hat{G}_h, \mathfrak{h} \in \mathfrak{S}_n\}$ satisfying hypotheses ($\mathbf{A}^{\mathrm{permute}}$) and ($\mathbf{A}^{\mathrm{upper}}$), and establish for it the ℓ-oracle inequality (2.5).

Define, for any $\mathfrak{h} \in \mathfrak{S}_n$,

$$\hat{R}_n(\mathfrak{h}) = \sup_{\eta \in \mathfrak{S}_n} \big[\ell(\hat{G}_\eta, \hat{G}_{\mathfrak{h},\eta}) - 2\Psi_n(\eta)\big]_+.$$

Let $\hat{\mathfrak{h}}^{(n)} \in \mathfrak{S}_n$ be an arbitrary $X^{(n)}$-measurable random element satisfying

$$\hat{R}_n\big(\hat{\mathfrak{h}}^{(n)}\big) + 2\Psi_n\big(\hat{\mathfrak{h}}^{(n)}\big) \le \inf_{\mathfrak{h} \in \mathfrak{S}_n} \big\{\hat{R}_n(\mathfrak{h}) + 2\Psi_n(\mathfrak{h})\big\} + \varepsilon_n.$$

Our final estimator is $\hat{G}_{\hat{\mathfrak{h}}^{(n)}}$. In order to bound from above its risk, introduce the following notation: for any $f \in \mathbb{F}$, $\mathfrak{h} \in \mathfrak{S}_n$ and $n \ge 1$,

$$\mathcal{B}^{(n)}(f,\mathfrak{h}) = \ell\big(\Lambda_{\mathfrak{h}}(f), G(f)\big) + 2 \sup_{\eta \in \mathfrak{S}_n} \ell\big(\Lambda_{\mathfrak{h},\eta}(f), \Lambda_\eta(f)\big),$$

$$\psi_n(f,\mathfrak{h}) = \big[\mathbb{E}_f^{(n)}\big\{\Psi_n^q(\mathfrak{h})\big\}\big]^{\frac{1}{q}}.$$

Theorem 3.1 ([33]). *Let* ($\mathbf{A}^{\mathrm{permute}}$)(i) *and* ($\mathbf{A}^{\mathrm{upper}}$) *be fulfilled. Then, for any* $f \in \mathfrak{F}$ *and* $n \ge 1$,

$$\mathcal{R}_n^{(\ell)}\big[\hat{G}_{\hat{\mathfrak{h}}^{(n)}}; G(f)\big] \le \inf_{\mathfrak{h} \in \mathfrak{S}_n} \big\{\mathcal{B}^{(n)}(f,\mathfrak{h}) + 5\psi_n(f,\mathfrak{h})\big\} + 6\varepsilon_n.$$

Thus, the ℓ-oracle inequality is established with $r_n = 6\varepsilon_n$ and

$$U_n^{(\ell)}(f,\mathfrak{h}) = \mathcal{B}^{(n)}(f,\mathfrak{h}) + 5\psi_n(f,\mathfrak{h}).$$

Our goal now is to prove the following result.

Theorem 3.2. *Let* ($\mathbf{A}^{\mathrm{permute}}$)(ii) *and* ($\mathbf{A}^{\mathrm{upper}}$) *be fulfilled. Then, for any* $f \in \mathfrak{F}$ *and* $n \ge 1$,

$$\mathcal{R}_n^{(\ell)}\big[\hat{G}_{\hat{\mathfrak{h}}^{(n)}}; G(f)\big] \le \inf_{\mathfrak{h} \in \mathfrak{S}_n} \big\{\mathcal{B}^{(n)}(f,\mathfrak{h}) + 9\psi_n(f,\mathfrak{h})\big\} + 10\varepsilon_n + \delta_n.$$

Thus, the ℓ-oracle inequality is established with $r_n = 10\varepsilon_n + \delta_n$ and

$$U_n^{(\ell)}(f,\mathfrak{h}) = \mathcal{B}^{(n)}(f,\mathfrak{h}) + 9\psi_n(f,\mathfrak{h}).$$

Proof. We break the proof into three short steps and, for the simplicity of notations, we will write $\hat{\mathfrak{h}}$ instead of $\hat{\mathfrak{h}}^{(n)}$. Set

$$\xi_1 = \sup_{\eta \in \mathfrak{S}_n} \big[\ell(\hat{G}_\eta, \Lambda_\eta) - \Psi_n(\eta)\big]_+, \quad \xi_2 = \sup_{\mathfrak{h},\eta \in \mathfrak{S}_n} \big[\ell(\hat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}) - \big\{\Psi_n(\mathfrak{h}) \wedge \Psi_n(\eta)\big\}\big]_+.$$

1) Our first goal is to prove that for any $\mathfrak{h}, \eta \in \mathfrak{H}_n$,

$$\ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\mathfrak{h},\eta}) \le \hat{R}_n(\eta) + 6\Psi_n(\mathfrak{h}) + 2\xi_1 + 2\xi_2 + \delta_n. \tag{3.1}$$

Indeed, the following chain od inequalities is obtained from the triangle inequality:

$$\begin{aligned}\ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\mathfrak{h},\eta}) &\le \ell(\hat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}) + \ell(\Lambda_{\mathfrak{h}}, \hat{G}_{\mathfrak{h},\eta}) \\ &\le \ell(\hat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}) + \ell(\Lambda_{\mathfrak{h}}, \Lambda_{\mathfrak{h},\eta}) + \ell(\hat{G}_{\mathfrak{h},\eta}, \Lambda_{\mathfrak{h},\eta}) \\ &\le \ell(\Lambda_{\mathfrak{h}}, \Lambda_{\mathfrak{h},\eta}) + 2\Psi_n(\mathfrak{h}) + \xi_1 + \xi_2. \end{aligned} \tag{3.2}$$

Similarly, taking into account $(\mathbf{A}^{\text{permute}})$(ii), we get

$$\begin{aligned}\ell(\Lambda_{\mathfrak{h}}, \Lambda_{\mathfrak{h},\eta}) &\le \ell(\Lambda_{\mathfrak{h}}, \Lambda_{\eta,\mathfrak{h}}) + \delta_n \\ &\le \ell(\hat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}) + \ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\eta,\mathfrak{h}}) + \ell(\hat{G}_{\eta,\mathfrak{h}}, \Lambda_{\eta,\mathfrak{h}}) + \delta_n \\ &\le \ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\eta,\mathfrak{h}}) + 2\Psi_n(\mathfrak{h}) + \xi_1 + \xi_2 + \delta_n. \end{aligned} \tag{3.3}$$

It remains to note that in view of the definition of $\hat{R}_n(\cdot)$,

$$\ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\eta,\mathfrak{h}}) \le 2\Psi_n(\mathfrak{h}) + \big[\ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\eta,\mathfrak{h}}) - 2\Psi_n(\mathfrak{h})\big]_+ \le 2\Psi_n(\mathfrak{h}) + \hat{R}_n(\eta).$$

This, together with (3.2) and (3.3), implies (3.1).

2) Let $\mathfrak{h} \in \mathfrak{H}_n$ be fixed. We have in view of the definition of $\hat{R}_n(\cdot)$ that

$$\ell(\hat{G}_{\hat{\mathfrak{h}}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) \le 2\Psi_n(\hat{\mathfrak{h}}) + \big[\ell(\hat{G}_{\hat{\mathfrak{h}}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) - 2\Psi_n(\hat{\mathfrak{h}})\big]_+ \le 2\Psi_n(\hat{\mathfrak{h}}) + \hat{R}_n(\mathfrak{h}). \tag{3.4}$$

Here we have also used that $\hat{\mathfrak{h}} \in \mathfrak{H}_n$ by its definition.

Applying (3.1) with $\eta = \hat{\mathfrak{h}}$, we obtain

$$\ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) \le \hat{R}_n(\hat{\mathfrak{h}}) + 6\Psi_n(\mathfrak{h}) + 2\xi_1 + 2\xi_2 + \delta_n. \tag{3.5}$$

We get from (3.4), (3.5), and the definition of $\hat{\mathfrak{h}}$ that

$$\begin{aligned}\ell(\hat{G}_{\hat{\mathfrak{h}}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) + \ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) &\le \hat{R}_n(\hat{\mathfrak{h}}) + 2\Psi_n(\hat{\mathfrak{h}}) + \hat{R}_n(\mathfrak{h}) + 6\Psi_n(\mathfrak{h}) + 2\xi_1 + 2\xi_2 + \delta_n \\ &\le 2\hat{R}_n(\mathfrak{h}) + 8\Psi_n(\mathfrak{h}) + 2\xi_1 + 2\xi_2 + \varepsilon_n + \delta_n. \end{aligned} \tag{3.6}$$

3) We have, in view of the triangle inequality, for any $\mathfrak{h} \in \mathfrak{H}_n$ that

$$\hat{R}_n(\mathfrak{h}) \le \sup_{\eta \in \mathfrak{H}_n} \ell\big(\Lambda_{\mathfrak{h},\eta}(f), \Lambda_\eta(f)\big) + \xi_1 + \xi_2. \tag{3.7}$$

Thus, we obtain from (3.6) and (3.7), for any $\mathfrak{h} \in \mathfrak{H}_n$,

$$\begin{aligned}&\ell(\hat{G}_{\hat{\mathfrak{h}}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) + \ell(\hat{G}_{\mathfrak{h}}, \hat{G}_{\mathfrak{h},\hat{\mathfrak{h}}}) \\ &\quad\le 2 \sup_{\eta \in \mathfrak{H}_n} \ell\big(\Lambda_{\mathfrak{h},\eta}(f), \Lambda_\eta(f)\big) + 8\Psi_n(\mathfrak{h}) + 4\xi_1 + 4\xi_2 + \varepsilon_n + \delta_n. \end{aligned} \tag{3.8}$$

Obviously, for any $\mathfrak{h} \in \mathfrak{H}_n$,

$$\ell\big(\hat{G}_{\mathfrak{h}}, G(f)\big) \le \ell\big(\Lambda_{\mathfrak{h}}(f), G(f)\big) + \Psi_n(\mathfrak{h}) + \xi_1.$$

By the triangle inequality, this yields, together with (3.8), for any $\mathfrak{h} \in \mathfrak{H}_n$ that

$$\ell\big(\hat{G}_{\hat{h}}, G(f)\big) \le \mathcal{B}^{(n)}(f, \mathfrak{h}) + 9\Psi_n(\mathfrak{h}) + 5\xi_1 + 4\xi_2 + \varepsilon_n + \delta_n, \quad \forall f \in \mathfrak{F}.$$

Taking into account Assumption ($\mathbf{A}^{\text{upper}}$), we get for any $\mathfrak{h} \in \mathfrak{H}_n$ and any $f \in \mathfrak{F}$,

$$\{\mathbb{E}_f^{(n)}[\ell(\hat{G}_{\hat{\mathfrak{h}}}, G(f))]^q\}^{\frac{1}{q}} \le \mathcal{B}^{(n)}(f, \mathfrak{h}) + 9\psi_n(f, \mathfrak{h}) + 10\varepsilon_n + \delta_n.$$

Noting that the left-hand side of the obtained inequality is independent of $\mathfrak{h}$, we come to the assertion of the theorem. ■

We finish this section with simple, but very useful (in minimax and minimax adaptive estimation) consequence of Theorems 3.1–3.2.

For any $\mathbb{F} \subseteq \mathfrak{F}$, set

$$\gamma_n(\mathbb{F}) = \inf_{\mathfrak{h} \in \mathfrak{H}} \sup_{f \in \mathbb{F}} [\mathcal{B}^{(n)}(f, \mathfrak{h}) + \psi_n(f, \mathfrak{h})].$$

The quantity $\gamma_n(\mathbb{F})$ is often called the *bias–variance tradeoff*.

Corollary 1. *Let* ($\mathbf{A}^{\text{upper}}$) *be fulfilled. Assume also that either* ($\mathbf{A}^{\text{permute}}$)(i) *holds or* ($\mathbf{A}^{\text{permute}}$)(ii) *is verified with* $\delta_n = \varepsilon_n$. *Then, for any* $\mathbb{F} \subseteq \mathfrak{F}$ *and* $n \ge 1$,

$$\mathcal{R}_n^{(\ell)}[\hat{G}_{\hat{\mathfrak{h}}^{(n)}}; \mathbb{F}] \le 9\gamma_n(\mathbb{F}) + 11\varepsilon_n.$$

The proof of the corollary is elementary and can be omitted.

4. EXAMPLES OF ESTIMATOR COLLECTIONS SATISFYING ASSUMPTION ($\mathbf{A}^{\text{permute}}$)

4.1. Estimator collections in the density model

First example. Let $\mathcal{D} = \mathbb{R}^d$, $d \ge 1$, and μ be the Lebesgue measure. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a function from $\mathbb{L}_1(\mathbb{R}^d)$ and $\int_{\mathbb{R}} K = 1$. Let $\mathfrak{H} \subseteq (0, 1]^d$, and define for any $\mathfrak{h} = (\mathfrak{h}_1, \dots, \mathfrak{h}_d) \in \mathfrak{H}$,

$$K_{\mathfrak{h}}(t) = V_{\mathfrak{h}}^{-1} K(t_1/\mathfrak{h}_1, \dots, t_d/\mathfrak{h}_d), \quad t \in \mathbb{R}^d, \quad V_{\mathfrak{h}} = \prod_{j=1}^{d} \mathfrak{h}_j. \tag{4.1}$$

Introduce the following estimator collection:

$$\mathcal{G} = \left\{ \hat{G}_{\mathfrak{h}}(x) = n^{-1} \sum_{i=1}^{n} K_{\mathfrak{h}}(X_i - x), x \in \mathbb{R}^d, \mathfrak{h} \in \mathfrak{H} \right\}. \tag{4.2}$$

The estimator $\hat{G}_{\mathfrak{h}}(\cdot)$ is called the kernel estimator with bandwidth $\mathfrak{h}$. Kernel estimators are used in estimating the underlying density at a given point, as well as in estimating the entire f. Also, they are used as a building block for constructing estimators of many functionals of density mentioned in Section 1.2. Selection from the family $\mathcal{G}$, usually referred to as bandwidth selection, is one of the central problems in nonparametric density estimation.

For any $\mathfrak{h} \in \mathfrak{H}$, set

$$\Lambda_{\mathfrak{h}}(f, \cdot) = \mathbb{E}_f^{(n)}[\hat{G}_{\mathfrak{h}}(\cdot)] = \int_{\mathcal{D}} K_{\mathfrak{h}}(t - \cdot) f(t) \mathrm{d}t$$

and consider two possible constructions of the collection $\hat{G}_{\mathfrak{h},\eta}(\cdot)$, $\mathfrak{h}, \eta \in \mathfrak{H}$.

Construction based on the convolution product. Define $K_{\mathfrak{h},\eta} : \mathbb{R}^d \to \mathbb{R}$ by

$$K_{\mathfrak{h},\eta}(\cdot) = \int_{\mathbb{R}^d} K_\eta(\cdot - t) K_{\mathfrak{h}}(t) \mathrm{d}t =: [K_{\mathfrak{h}} * K_\eta](\cdot)$$

and set

$$\hat{G}_{\mathfrak{h},\eta}(\cdot) = n^{-1} \sum_{i=1}^{n} K_{\mathfrak{h},\eta}(X_i - \cdot), \quad \Lambda_{\mathfrak{h},\eta}(f,\cdot) = \mathbb{E}_f^{(n)}\big[\hat{G}_{\mathfrak{h},\eta}(\cdot)\big].$$

Since obviously $K_{\mathfrak{h},\eta} \equiv K_{\eta,\mathfrak{h}}$, we can assert that $\hat{G}_{\mathfrak{h},\eta} \equiv \hat{G}_{\eta,\mathfrak{h}}$ and, therefore, Assumptions $(\mathbf{A}^{\text{permute}})$(i) and $(\mathbf{A}^{\text{permute}})$(ii) are both fulfilled.

Construction based on the coordinatewise maximum. Define $K_{\mathfrak{h},\eta} : \mathbb{R}^d \to \mathbb{R}$ by

$$K_{\mathfrak{h},\eta}(\cdot) = K_{\mathfrak{h}\vee\eta}(\cdot), \quad \mathfrak{h} \vee \eta = (\mathfrak{h}_1 \vee \eta_1, \dots, \mathfrak{h}_d \vee \eta_d),$$

and set

$$\hat{G}_{\mathfrak{h},\eta}(\cdot) = n^{-1} \sum_{i=1}^{n} K_{\mathfrak{h},\eta}(X_i - \cdot), \quad \Lambda_{\mathfrak{h},\eta}(f,\cdot) = \mathbb{E}_f^{(n)}\big[\hat{G}_{\mathfrak{h},\eta}(\cdot)\big].$$

Since obviously $K_{\mathfrak{h},\eta} \equiv K_{\eta,\mathfrak{h}}$, we can assert that $\hat{G}_{\mathfrak{h},\eta} \equiv \hat{G}_{\eta,\mathfrak{h}}$ and, therefore, Assumptions $(\mathbf{A}^{\text{permute}})$(i) and $(\mathbf{A}^{\text{permute}})$(ii) are both fulfilled.

Second example. Consider now the estimator collection related to the density estimation under hypothesis of independence presented in Section 2.1.2.

Here, as previously, $\mathcal{D} = \mathbb{R}^d$, $d \geq 2$, and μ is the Lebesgue measure. Recall that $\mathfrak{F} \subseteq \mathfrak{P}(\mathcal{D}, \mu)$, $\mathcal{I}_d$ is the set of all subsets of $\{1, \dots, d\}$, and Π denotes the set of all partitions of $\{1, \dots, d\}$.

Let $\mathbf{K} : \mathbb{R}^1 \to \mathbb{R}^1$ be a univariate kernel, that is, $\mathbf{K} \in \mathbb{L}_1(\mathbb{R}^1)$ and $\int_{\mathbb{R}^1} \mathbf{K} = 1$.

For any $h = (0,1]^d$ and any $I \in \mathcal{I}_d$, set

$$K_{h_I}(u) = V_{h_I}^{-1} \prod_{j \in I} \mathbf{K}(u_j / h_j), \quad V_{h_I} = \prod_{j \in I} h_j.$$

Since the independence hypothesis assumes that there exists a partition $\mathcal{P}$ such that

$$f(x) = \prod_{I \in \mathcal{P}} f_I(x_I), \quad \forall x \in \mathbb{R}^d,$$

the idea is to estimate each marginal density by the kernel method and use the product of these estimators as the final one. Thus, define for any $x \in \mathbb{R}^d$, $\mathfrak{h} \in \mathfrak{H}$, and any $I \in \mathcal{I}_d$,

$$\hat{f}_{h_I}(x_I) = n^{-1} \sum_{i=1}^{n} K_{h_I}(X_{I,i} - x_I)$$

and introduce the following family of estimators:

$$\mathcal{G} = \left\{ \hat{G}_{\mathfrak{h}}(x) = \prod_{I \in \mathcal{P}} \hat{f}_{h_I}(x_I), x \in \mathbb{R}^D, \mathfrak{h} = (h, \mathcal{P}) \in [0,1]^d \times \Pi =: \mathfrak{H} \right\}.$$

Let $*$ denote the convolution operator on $\mathbb{R}$. Set for any $x \in \mathbb{R}^d$, $h, h' \in (0,1]^d$, and any $I \in \mathcal{I}_d$,

$$[K_{h_{\mathbf{I}}} \star K_{h'_I}] = \prod_{j \in I} [\mathbf{K}_{h_j} * \mathbf{K}_{h'_j}]$$

and introduce

$$\hat{f}_{h_I,h'_I}(x_I) = n^{-1}\sum_{i=1}^{n}[K_{h_I} \star K_{h'_I}](X_{I,i} - x_I),$$

Let us endow the set Π with the operation "$\diamond$" putting for any $\mathcal{P}, \mathcal{P}' \in \Pi$,

$$\mathcal{P} \diamond \mathcal{P}' = \left\{I \cap I' \neq \emptyset, I \in \mathcal{P}, I' \in \mathcal{P}'\right\} \in \Pi.$$

Introduce for any $\mathfrak{h}, \eta \in \mathfrak{H}$ the estimator

$$\hat{G}_{\mathfrak{h},\eta}(x) = \prod_{I \in \mathcal{P} \diamond \mathcal{P}'} \hat{f}_{h_I,h'_I}(x_I), \quad x \in \mathbb{R}^d.$$

Obviously, $\hat{G}_{\mathfrak{h},\eta} \equiv \hat{G}_{\eta,\mathfrak{h}}$ and, therefore, Assumption $(\mathbf{A}^{\text{permute}})$(i) is fulfilled. On the other hand, see [30], functionals $\Lambda_{\mathfrak{h}}$ and $\Lambda_{\mathfrak{h},\eta}$ are so complicated that the verification of $(\mathbf{A}^{\text{permute}})$(ii) does not seem possible. We are not even sure that it holds for sufficiently small δ_n.

Third example. Let us now consider the family of estimators which appears in adaptive estimation under the following structural assumption. Let $\mathcal{D} = \mathbb{R}^2$ and μ be the Lebesgue measure. Let $\mathfrak{Q}$ denote the set of all 2×2 rotational matrices and $\mathfrak{P}_1^{\text{sym}}$ denote the set of all symmetric probability densities on $\mathbb{R}^1$. Set

$$\mathcal{A} = \left\{a : \mathbb{R}^2 \to \mathbb{R}^1 : a(\cdot,\cdot) = a_1(\cdot)a_2(\cdot), a_1, a_2 \in \mathfrak{P}_1^{\text{sym}}\right\},$$

and assume that there exist $a \in \mathcal{A}$ and $M \in \mathfrak{Q}$ such that $f(\cdot) = a(M^T\cdot)$. The latter means that

$$X_i = M\xi_i, \quad i = 1, \dots, n,$$

where ξ_i, $i = 1, \dots, n$, are i.i.d. random vectors with a common density a.

If M is known then $\xi_i = M^T X_i, \dots, \xi_n = M^T X_n$ are observable i.i.d. random vectors with *independent coordinates*. Indeed, the density of ξ_1 is $a_1(\cdot)a_2(\cdot)$. Hence the estimation of a is the estimation under hypothesis of independence, which, as it was mentioned above, allows one to improve the accuracy of estimation of the density a, and, therefore, of the density f as well. However, if M is unknown, the sequence $\xi_i = M^T X_i, \dots, \xi_n = M^T X_n$ is not observable anymore and the estimation of f can be viewed as the problem of adaptation to an unknown rotation of the coordinate system.

Let the kernel $\mathbf{K} : \mathbb{R}^1 \to \mathbb{R}^1$ be the same as in the previous example and set $K_h(\cdot) = h^{-1}\mathbf{K}(\cdot/h)$, $h \in (0, 1]$. Later on $Q \in \mathfrak{Q}$ will be presented as

$$Q = (q, q_\perp) = \begin{pmatrix} q_1 & -q_2 \\ q_2 & q_1 \end{pmatrix},$$

where $q, q_\perp \in \mathbb{S}^1$. For any $\mathfrak{h} := (h, Q) \in [0, 1] \times \mathfrak{Q}$ and $x \in \mathbb{R}^2$, set

$$\hat{G}_{\mathfrak{h}}(x) = \left[n^{-1}\sum_{k=1}^{n} K_h\big(q^T(X_k - x)\big)\right]\left[n^{-1}\sum_{k=1}^{n} K_h\big(q_\perp^T(X_k - x)\big)\right],$$

and introduce the following family of estimators:

$$\mathcal{G} = \left\{\hat{G}_{\mathfrak{h}}(x), x \in \mathbb{R}^2, \mathfrak{h} \in \mathfrak{H} \subseteq [0, 1] \times \mathfrak{Q}\right\}.$$

In order to construct estimator $\hat{G}_{\mathfrak{h},\eta}(\cdot)$, $\mathfrak{h}, \eta \in \mathfrak{H}$, we will need the following notation.

For any $Q, D \in \mathfrak{Q}$, define

$$p(D, Q) = q^T d_\perp, \quad \pi(D, Q) = q^T d.$$

Set also $\mathcal{K}_h(t) = K_h(t_1)K_h(t_2)$, $t \in \mathbb{R}^2$, $h \in (0, 1]$, and let

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

Define, see [35], for any $\mathfrak{h} = (h, Q) \in \mathfrak{H}$ and $\eta = (\varkappa, D) \in \mathfrak{H}$,

$$\hat{G}_{\mathfrak{h},\eta}(x) = \frac{1}{n(n-1)} \sum_{k,l=1, k\neq l}^{n} \mathcal{K}_{h\vee\varkappa}\big(p(D, Q)\Omega\Gamma X_k + \pi(D, Q)X_l - \Omega\Gamma QD\Omega x\big)$$

and let

$$\Lambda_{\mathfrak{h},\eta}(f, \cdot) = \mathbb{E}_f^{(n)}\big[\hat{G}_{\mathfrak{h},\eta}(\cdot)\big].$$

Note that for any $D, Q \in \mathfrak{Q}$,

$$p(D, Q) = -p(Q, D), \quad \pi(D, Q) = \pi(Q, D), \quad DQ = QD. \tag{4.3}$$

Obviously, $\hat{G}_{\mathfrak{h},\eta}(\cdot) \neq \hat{G}_{\eta,\mathfrak{h}}(\cdot)$ and, therefore, Assumption $(\mathbf{A}^{\text{permute}})$(i) does not hold.

On the other hand,

$$\Lambda_{\mathfrak{h},\eta}(f, \cdot) = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \mathcal{K}_{h\vee\varkappa}\big(p(D, Q)\Omega\Gamma u + \pi(D, Q)v - \Omega\Gamma QD\Omega x\big) f(u) f(v)\mathrm{d}u\mathrm{d}v.$$

Since $f(\cdot) = a(M^T\cdot)$ and a is symmetric, f is a symmetric function as well, and we have

$$\begin{aligned}\Lambda_{\mathfrak{h},\eta}(f, \cdot) &= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \mathcal{K}_{h\vee\varkappa}\big(-p(D, Q)\Omega\Gamma u + \pi(D, Q)v - \Omega\Gamma QD\Omega x\big) f(u) f(v)\mathrm{d}u\mathrm{d}v \\ &= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \mathcal{K}_{h\vee\varkappa}\big(p(Q, D)\Omega\Gamma u + \pi(Q, D)v - \Omega\Gamma DQ\Omega x\big) f(u) f(v)\mathrm{d}u\mathrm{d}v \\ &= \Lambda_{\eta,\mathfrak{h}}(f, \cdot).\end{aligned}$$

To get the penultimate equality, we have used (4.3). We conclude that Assumption $(\mathbf{A}^{\text{permute}})$(ii) holds with any δ_n whatever the semimetric ℓ is considered.

4.2. Estimator collections in White Gaussian Noise Model

First example. Let $\mathcal{D}$ be a set endowed with the Borel measure μ and $\mu(\mathcal{D}) < \infty$. Recall that the observation $X^{(n)} = \{X_n(g), g \in \mathbb{L}_2(\mathcal{D}, \mu)\}$ is given in (1.2).

Let $\{\psi_m, m \in \mathbb{M}\}$ be an orthonormal basis in $\mathbb{L}_2(\mathcal{D}, \mu)$ and let $\mathfrak{H} = \{\mathfrak{h} = (\mathfrak{h}_m, m \in \mathbb{M})\}$ be a given subset of l_2. Introduce, for any $t, x \in \mathcal{D}$,

$$K_{\mathfrak{h}}(t, x) = \sum_{m\in\mathbb{M}} \mathfrak{h}_m \psi_m(t)\psi_m(x), \quad \mathfrak{h} \in \mathfrak{H},$$

and consider the following estimation collection:

$$\mathcal{G} = \big\{\hat{G}_{\mathfrak{h}}(x) = X_n\big(K(\cdot, x)\big), x \in \mathcal{D}, \mathfrak{h} \in \mathfrak{H}\big\}.$$

The estimator $\hat{G}_{\mathfrak{h}}(\cdot)$ is used in the estimation of unknown f under $\mathbb{L}_2$-loss, that is, $\mathfrak{S} = \mathfrak{F}$, $G(f) = f$, and $\ell(f,g) = \|f-g\|_{2,\mathcal{D}}$, $f, g \in \mathfrak{F} \subset \mathbb{L}_2(\mathcal{D}, \mu)$. Let

$$\Lambda_{\mathfrak{h}}(f,\cdot) = \mathbb{E}_f^{(n)}\big[\hat{G}_{\mathfrak{h}}(\cdot)\big] = \int_{\mathcal{D}} K_{\mathfrak{h}}(t,\cdot) f(t)\mu(\mathrm{d}t) = \sum_{m\in\mathbb{M}} \mathfrak{h}_m \psi_m(\cdot) \int_{\mathcal{D}} \psi_m(t) f(t)\mu(\mathrm{d}t).$$

Denoting the mth Fourier coefficient of f by f_m, we get

$$\Lambda_{\mathfrak{h}}(f,\cdot) = \sum_{m\in\mathbb{M}} \mathfrak{h}_m f_m \psi_m(\cdot).$$

In particular, in view of Parseval's identity,

$$\big\|\Lambda_{\mathfrak{h}}(f) - f\big\|_{2,\mathcal{D}} = \sum_{m\in\mathbb{M}} (\mathfrak{h}_m - 1)^2 f_m^2.$$

For any $\mathfrak{h}, \eta \in \mathfrak{H}$, set

$$K_{\mathfrak{h},\eta}(t,x) = \int_{\mathcal{D}} K_{\mathfrak{h}}(t,y) K_{\eta}(y,x)\mu(\mathrm{d}y), \quad t, x \in \mathcal{D},$$

and put, for any $x \in \mathcal{D}$,

$$\hat{G}_{\mathfrak{h},\eta}(x) = X_n\big(K_{\mathfrak{h},\eta}(\cdot,x)\big).$$

Noting that, for any $t, x \in \mathcal{D}$,

$$K_{\mathfrak{h},\eta}(t,x) = \sum_{m\in\mathbb{M}}\sum_{j\in\mathbb{M}} \mathfrak{h}_m \eta_j \psi_m(t)\psi_j(x) \int_{\mathcal{D}} \psi_m(y)\psi_j(y)\mu(\mathrm{d}t) = \sum_{m\in\mathbb{M}} \mathfrak{h}_m \eta_m \psi_m(t)\psi_m(x),$$

we can assert that $K_{\mathfrak{h},\eta} \equiv K_{\eta,\mathfrak{h}}$. This implies $\hat{G}_{\mathfrak{h},\eta} \equiv \hat{G}_{\eta,\mathfrak{h}}$ and, therefore,

$$\Lambda_{\mathfrak{h},\eta} := \mathbb{E}_f^{(n)}[\hat{G}_{\mathfrak{h},\eta}] \equiv \mathbb{E}_f^{(n)}[\hat{G}_{\eta,\mathfrak{h}}] =: \Lambda_{\eta,\mathfrak{h}}.$$

Hence, Assumptions $(\mathbf{A}^{\mathrm{permute}})$(i) and $(\mathbf{A}^{\mathrm{permute}})$(ii) are both fulfilled.

Second example. Here and later, $D = R^d$, $d \geq 1$, μ is the Lebesgue measure, and $X^{(n)} = \{X_n(g), g \in \mathbb{L}_2(\mathbb{R}^d, \mu)\}$ is given in (1.2).

Let $b > 0$ be given and denote by $\mathfrak{H}(b)$ the set of all Borel functions $\mathfrak{h} : (-b,b)^d \to (0,1]^d$. As before let $K : \mathbb{R}^d \to \mathbb{R}$, $K \in \mathbb{L}_1(\mathbb{R}^d)$ be a function satisfying $\int K = 1$.

With any $\mathfrak{h} \in \mathfrak{H}(b)$, we associate the function

$$K_{\mathfrak{h}(x)}(t,x) = V_{\mathfrak{h}}^{-1}(x) K\left(\frac{t-x}{\mathfrak{h}(x)}\right), \quad t \in \mathbb{R}^d, \quad x \in (-b,b)^d,$$

where $V_{\mathfrak{h}}(x) = \prod_{i=1}^d \mathfrak{h}_i(x)$ and $\mathfrak{h}(\cdot) = (\mathfrak{h}_1(\cdot), \ldots, \mathfrak{h}_d(\cdot))$.

Consider the family of estimators

$$\mathcal{G} = \big\{\hat{G}_{\mathfrak{h}(x)}(x) = X_n\big(K_{\mathfrak{h}(x)}(\cdot,x)\big), \mathfrak{h} \in \mathfrak{H}(b), x \in (-b,b)^d\big\}. \tag{4.4}$$

The estimators from this collection are called *kernel estimators with varying bandwidth*. Let

$$\Lambda_{\mathfrak{h}(\cdot)}(f,\cdot) = \mathbb{E}_f^{(n)}\big[\hat{G}_{\mathfrak{h}(\cdot)}(\cdot)\big] = \int_{\mathbb{R}^d} K_{\mathfrak{h}(\cdot)}(t,\cdot) f(t)\mu(\mathrm{d}t).$$

For any $\mathfrak{h}, \eta \in \mathfrak{H}(b)$, set

$$\hat{G}_{\mathfrak{h}(x)\vee\eta(x)}(x) = X_n\big(K_{\mathfrak{h}(x)\vee\eta(x)}(\cdot,x)\big), \quad x \in (-b,b)^d,$$

where as previously $\mathfrak{h}(\cdot) \vee \eta(\cdot) = (\mathfrak{h}_1(\cdot) \vee \eta_1(\cdot), \dots, \mathfrak{h}_d(\cdot) \vee \eta_d(\cdot))$. Let also

$$\hat{G}_{\mathfrak{h}(\cdot)\vee\eta(\cdot)}(\cdot) = \mathbb{E}_f^{(n)}[\hat{G}_{\mathfrak{h}(\cdot)}] = \int_{\mathbb{R}^d} K_{\mathfrak{h}(\cdot)}(t,\cdot) f(t)\mu(\mathrm{d}t).$$

Since obviously $K_{\mathfrak{h}\vee\eta} \equiv K_{\eta\vee\mathfrak{h}}$ for any $\mathfrak{h}, \eta \in \mathfrak{H}(b)$, we can assert that both Assumptions $(\mathbf{A}^{\text{permute}})$(i) and $(\mathbf{A}^{\text{permute}})$(ii) are fulfilled whatever the semimetric ℓ is considered.

5. ONE EXAMPLE OF ESTIMATOR COLLECTION SATISFYING ASSUMPTION ($\mathbf{A}^{\text{upper}}$)

In this section we continue to consider the estimator family given in (4.4). Our objective here is to find $\mathfrak{H}_n \subset \mathfrak{H}(b)$ and $\{\Psi_n(\mathfrak{h}), \mathfrak{h} \in \mathfrak{H}_n\}$ for which Assumption $(\mathbf{A}^{\text{upper}})$ can be checked in the case where ℓ is the $\mathbb{L}_p$-norm on $(-b,b)^d$, $1 \le p < \infty$.

For any $\mathfrak{h} \in \mathfrak{H}(b)$, define

$$\xi_{\mathfrak{h}(x)}(x) = \int_{\mathbb{R}^d} K_{\mathfrak{h}(x)}(t,x) W(\mathrm{d}t), \quad x \in (-b,b)^d,$$

and note that, in view of (1.2),

$$\ell\big(\hat{G}_{\mathfrak{h}}, \Lambda_{\mathfrak{h}}(f)\big) = n^{-\frac{1}{2}} \|\xi_{\mathfrak{h}}\|_{p,(-b,b)^d}.$$

We remark that $\xi_{\mathfrak{h}(\cdot)}(\cdot)$ is independent of f and n. Hence, Assumption $(\mathbf{A}^{\text{upper}})$ will be checked if we find $\mathfrak{H}_n$ and nonrandom $\{\Psi_n^*(\mathfrak{h}), \mathfrak{h} \in \mathfrak{H}_n\}$ such that

$$\mathbb{E}\Big(\sup_{\mathfrak{h}\in\mathfrak{H}_n} \big[\|\xi_{\mathfrak{h}}\|_{p,(-b,b)^d} - \Psi_n^*(\mathfrak{h})\big]_+^q \Big) \le \varepsilon_n^q n^{\frac{q}{2}}; \tag{5.1}$$

$$\mathbb{E}\Big(\sup_{\mathfrak{h},\eta\in\mathfrak{H}_n} \big[\|\xi_{\mathfrak{h}\vee\eta}\|_{p,(-b,b)^d} - \big\{\Psi_n^*(\mathfrak{h}) \wedge \Psi_n^*(\eta)\big\}\big]_+^q \Big) \le \varepsilon_n^q n^{\frac{q}{2}}. \tag{5.2}$$

Here and later, $\mathbb{E}$ denotes the mathematical expectation with respect to the law of W. Also, furthermore, we will assume that

$$K(x) = \prod_{i=1}^d \mathcal{K}(x_i), \quad \forall x \in \mathbb{R}^d,$$

where $\mathcal{K} : \mathbb{R}^1 \to \mathbb{R}^1$ is such that $\int \mathcal{K} = 1$, $\mathrm{supp}(\mathcal{K}) \subset [-1,1]$, and, for some $M > 0$,

$$\big|\mathcal{K}(s) - \mathcal{K}(t)\big| \le M|s-t|, \quad \forall s,t \in \mathbb{R}.$$

5.1. Functional classes of bandwidths

Let $\alpha_n \to 0$, $n \to \infty$, be a given sequence and let

$$\omega_n = e^{-\sqrt{|\ln(\alpha_n)|}}, \quad \Omega_n = e^{\ln^2(\alpha_n)}.$$

Set $H_n = \{h_s = e^{-s}, s \in \mathbb{N}\} \cap (0, \omega_n]$ and denote by $\mathfrak{H}_{1,n}$ the set of all measurable functions defined on $(-b,b)^d$ and taking values in H_n^d. Obviously, $\mathfrak{H}_{1,n} \subset \mathfrak{H}(b)$. For any $\mathfrak{h} \in \mathfrak{H}_{1,n}$ and any $\mathbf{s} = (s_1, \dots, s_d) \in \mathbb{N}^d$, define

$$\Upsilon_{\mathbf{s}}[\mathfrak{h}] = \bigcap_{j=1}^d \Upsilon_{s_j}[\mathfrak{h}_j], \quad \Upsilon_{s_j}[\mathfrak{h}_j] = \big\{x \in (-b,b)^d : \mathfrak{h}_j(x) = h_{s_j}\big\}.$$

Let $\tau \in (0,1)$ and $L > 0$ be given constants. Define

$$\mathfrak{H}_n(\tau, L) = \Big\{ \mathfrak{h} \in \mathfrak{H}_{1,n} : \sum_{\mathbf{s} \in \mathbb{N}^d} \mu^{\tau}\big(\Upsilon_{\mathbf{s}}[\mathfrak{h}]\big) \le L \Big\}.$$

Set $\mathbb{N}_p = \{\lfloor p \rfloor + 1, \lfloor p \rfloor + 2, \ldots\}$ and introduce

$$\mathfrak{H}_{2,n} = \bigcup_{r \in \mathbb{N}_p} \mathfrak{H}_n(r), \quad \mathfrak{H}_n(r) = \big\{ \mathfrak{h} \in \mathfrak{H}_{1,n} : \big\| V_{\mathfrak{h}}^{-\frac{1}{2}} \big\|_{\frac{rp}{r-p},(-b,b)^d} \le \Omega_n \big\}.$$

We will establish (5.1) and (5.2) with $\mathfrak{H}_n = \mathfrak{H}_n^*(\tau, L) := \mathfrak{H}_{2,n} \cap \mathfrak{H}_n(\tau, L)$.

5.2. Verification of (5.1)

For any $\mathfrak{h} \in \mathfrak{H}_{2,n}$, define

$$\mathbb{N}_{p,n}(\mathfrak{h}) = \mathbb{N}_p \cap [r_n(\mathfrak{h}), \infty), \quad r_n(\mathfrak{h}) = \inf\big\{ r \in \mathbb{N}_p : \mathfrak{h} \in \mathfrak{H}_n(r) \big\}.$$

Obviously, $r_n(\mathfrak{h}) < \infty$ for any $\mathfrak{h} \in \mathfrak{H}_{2,n}$. For any $\mathfrak{h} \in \mathfrak{H}_{2,n}$. define

$$\Psi_n(\mathfrak{h}) = \inf_{r \in \mathbb{N}_{p,n}(\mathfrak{h})} C(r, \tau, L) \big\| V_{\mathfrak{h}}^{-\frac{1}{2}} \big\|_{\frac{rp}{r-p},(-b,b)^d},$$

where $C(r, \tau, L), \tau \in (0,1), L > 0$, can be found in [32, SECTION 3.2.2]. Here we only mention that $C(r, \tau, L)$ is finite for any given r, τ, L but $\lim_{r \to \infty} C(r, \tau, L) = \infty$.

Note also that the condition $\mathfrak{h} \in \mathfrak{H}_{2,n}$ guarantees that $\Psi_n(\mathfrak{h}) < \infty$.

Theorem 5.1 ([32, COROLLARY 1]). *For any $\tau \in (0,1)$ and $q \ge 1$, one can find $n(\tau, q)$ such that for any $n \ge n(\tau, q)$,*

$$\mathbb{E}\Big\{ \sup_{\mathfrak{h} \in \mathfrak{H}_n^*(\tau, L)} \big[\|\xi_{\mathfrak{h}}\|_{p,(-b,b)^d} - \Psi_n(\mathfrak{h}) \big]_+ \Big\}^q \le (c\alpha_n)^q,$$

where c depends on $\mathcal{K}, p, q, b$, and d only.

Choosing $\alpha_n = c^{-1}\varepsilon_n \sqrt{n}$, we can assert that is (5.1) holds for any $\Psi *_n (\cdot) \ge \Psi_n(\cdot)$.

5.3. Verification of (5.2)

The verification of (5.2) is mostly based on two facts.

First, the following result has been proved in [31, LEMMA 1].

Lemma 5.2. *For any $d \ge 1$, $\tau \in (0, 1/d)$, and $L > 0$, there exist $n(\tau, d, L)$ such that for all $n \ge n(\tau, L, d)$,*

$$\mathfrak{h} \vee \eta \in \mathfrak{H}_n\big(d\tau, (2L)^d\big), \quad \forall \mathfrak{h}, \eta \in \mathfrak{H}_n(\tau, L).$$

Hence, setting

$$\Psi_n^*(\mathfrak{h}) = \inf_{r \in \mathbb{N}_{p,n}(\mathfrak{h})} C^*(r, \tau, L) \big\| V_{\mathfrak{h}}^{-\frac{1}{2}} \big\|_{\frac{rp}{r-p},(-b,b)^d},$$

where $C^*(r, \tau, L) = C(r, \tau, L) \vee C(r, d\tau, (2L)^d)$, we can assert that the statement of Theorem 5.1 remains true for $\Psi_n^*(\cdot)$ as well if $\tau > 1/d$. This follows from the fact that $\Psi_n^*(\cdot) \ge \Psi_n(\cdot)$.

Moreover, in view of Theorem 5.1, for all n large enough,

$$\mathbb{E}\Big\{\sup_{\mathfrak{h}\in\mathfrak{S}_n^*(d\,\tau,(2L)^d)}\big[\|\xi_{\mathfrak{h}}\|_{p,(-b,b)^d}-\Psi_n^*(\mathfrak{h})\big]_+\Big\}^q\le(c\alpha_n)^q. \tag{5.3}$$

Since, in view Lemma 5.2, if $\tau>1/d$, we have

$$\sup_{\mathfrak{h},\eta\in\mathfrak{S}_n^*(\tau,L)}\big[\|\xi_{\mathfrak{h}\vee\eta}\|_{p,(-b,b)^d}-\Psi_n^*(\mathfrak{h}\vee\eta)\big]_+\le\sup_{\rho\in\mathfrak{S}_n^*(d\,\tau,(2L)^d)}\big[\|\xi_{\rho}\|_{p,(-b,b)^d}-\Psi_n^*(\rho)\big]_+,$$

we deduce from (5.3) that

$$\mathbb{E}\Big\{\sup_{\mathfrak{h},\eta\in\mathfrak{S}_n^*(\tau,L)}\big[\|\xi_{\mathfrak{h}\vee\eta}\|_{p,(-b,b)^d}-\Psi_n^*(\mathfrak{h}\vee\eta)\big]_+\Big\}^q\le(c\alpha_n)^q. \tag{5.4}$$

It remains to note that for any $1\le t<\infty$ and any $\mathfrak{h}\in\mathfrak{S}$,

$$\big\|V_{\mathfrak{h}\vee\eta}^{-\frac12}\big\|_{t,(-b,b)^d}\le\big\|V_{\mathfrak{h}}^{-\frac12}\big\|_{t,(-b,b)^d}\wedge\big\|V_{\eta}^{-\frac12}\big\|_{t,(-b,b)^d},$$

which implies

$$\Psi_n^*(\mathfrak{h}\vee\eta)\le\Psi_n^*(\mathfrak{h})\wedge\Psi_n^*(\eta),\quad\forall\mathfrak{h},\eta\in\mathfrak{S}. \tag{5.5}$$

Inequality (5.2) follows now from (5.4) and (5.5) if one chooses $\alpha_n=c^{-1}\varepsilon_n\sqrt{n}$.

ACKNOWLEDGMENTS

The author is grateful to A. Goldenshluger who read the manuscript and made useful comments.

FUNDING

This work was partially supported by the Labex Archimède (ANR-11-LABX-0033) and the A*MIDEX project (ANR-11-IDEX-0001-02).

REFERENCES

[1] A. Barron, L. Birgé, and P. Massart, Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113** (1999), 301–413.

[2] L. Birgé and P. Massart, Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** (2001), no. 3, 203–268.

[3] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, Aggregation for Gaussian regression. *Ann. Stat.* **35** (2007), no. 4, 1674–1697.

[4] T. T. Cai, Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Stat.* **27** (1999), no. 3, 898–924.

[5] T. T. Cai and M. G. Low, On adaptive estimation of linear functionals. *Ann. Stat.* **33** (2005), no. 5, 2311–2343.

[6] T. T. Cai and M. G. Low, Optimal adaptive estimation of a quadratic functional. *Ann. Stat.* **34** (2006), no. 5, 2298–2325.

[7] L. Cavalier and G. K. Golubev, Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Stat.* **34** (2006), 1653–1677.

[8] L. Cavalier and A. B. Tsybakov, Penalized blockwise Stein's method, monotone oracle and sharp adaptive estimation. *Math. Methods Stat.* **10** (2001), 247–282.

[9] A. Dalalyan and A. B. Tsybakov, Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** (2008), 39–61.

[10] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, Density estimation by wavelet thresholding. *Ann. Stat.* **24** (1996), 508–539.

[11] S. Yu. Efroimovich, Non-parametric estimation of the density with unknown smoothness. *Theory Probab. Appl.* **30** (1986), 557–568.

[12] S. Yu. Efroimovich, Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Stat.* **36** (2008), 1127–1155.

[13] S. Yu. Efroimovich and M. G. Low, Adaptive estimates of linear functionals. *Probab. Theory Relat. Fields* **98** (1994), 261–275.

[14] S. Yu. Efroimovich and M. S. Pinsker, An adaptive algorithm of nonparametric filtering. *Autom. Remote Control* **45** (1984), 58–65.

[15] A. Goldenshluger, A universal procedure for aggregating estimators. *Ann. Stat.* **37** (2009), no. 1, 542–568.

[16] A. Goldenshluger and O. V. Lepski, Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Relat. Fields* **143** (2009), 41–71.

[17] A. Goldenshluger and O. V. Lepski, Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Stat.* **39** (2011), 1608–1632.

[18] A. Goldenshluger and O. V. Lepski, General selection rule from the family of linear estimators. *Theory Probab. Appl.* **57** (2012), no. 2, 257–277.

[19] A. Goldenshluger and O. V. Lepski, On adaptive minimax density estimation on $\mathbb{R}^d$. *Probab. Theory Relat. Fields* **159** (2014), 479–543.

[20] G. K. Golubev, Non-parametric estimation of smooth probability densities. *Probl. Inf. Transm.* **1** (1992), 52–62.

[21] G. K. Golubev and M. Nussbaum, An adaptive spline estimate in nonparametric regression model. *Theory Probab. Appl.* **37** (1992), no. 3, 553–560.

[22] A. B. Iouditski, O. V. Lepski, and A. B. Tsybakov, Nonparametric estimation of composite functions. *Ann. Stat.* **37** (2009), no. 3, 1360–1440.

[23] A. Juditsky and A. Nemirovski, Functional aggregation for nonparametric regression. *Ann. Stat.* **28** (2000), no. 3, 681–712.

[24] G. Kerkyacharian, O. V. Lepski, and D. Picard, Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Relat. Fields* **121** (2001), 137–170.

[25] N. Klutchnikoff, Pointwise adaptive estimation of a multivariate function. *Math. Methods Statist.* **23** (2014), no. 2, 132–150.

[26] O. V. Lepskii, One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** (1990), 459–470.

[27] O. V. Lepskii, Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** (1991), no. 4, 682–697.

[28] O. V. Lepskii, Asymptotically minimax adaptive estimation. II. Statistical model without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37** (1992), no. 3, 468–481.

[29] O. V. Lepskii, On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, pp. 87–106, Adv. Sov. Math. 12, Amer. Math. Soc., Providence, RI, 1992.

[30] O. V. Lepskii, Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *Ann. Stat.* **41** (2013), no. 2, 1005–1034.

[31] O. V. Lepskii, Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Stat.* **43** (2015), no. 3, 1178–1242.

[32] O. V. Lepskii, Upper functions for $\mathbb{L}_p$-norm of Gaussian random fields. *Bernoulli* **22** (2016), no. 2, 732–773.

[33] O. V. Lepskii, A new approach to estimator selection. *Bernoulli* **24** (2018), 2776–2810.

[34] O. V. Lepskii and B. Ya. Levit, Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.* **7** (1998), no. 2, 123–156.

[35] O. V. Lepskii and G. Rebelles, Structural adaptation in the density model. *Math. Stat. Learn* **3** (2020), 1–42.

[36] G. Leung and A. R. Barron, Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52** (2006), no. 8, 3396–3410.

[37] A. S. Nemirovski, Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour)*, pp. 85–277, Lecture Notes in Math. 1738, Springer, Berlin, 1998.

[38] G. Rebelles, Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli* **21** (2015), no. 4, 1984–2023.

[39] P. Rigollet, Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12** (2006), 351–370.

[40] P. Rigollet and A. B. Tsybakov, Exponential screening and optimal rates of sparse estimation. *Ann. Stat.* **39** (2011), no. 2, 731–771.

[41] A. Tsybakov, Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Stat.* **26** (1998), 2420–2469.

[42] A. Tsybakov, Optimal rate of aggregation. In *Proc. COLT*, pp. 303–313, Lecture Notes in Artificial Intelligence 2777, Springer, Heidelberg, 2003.

[43] M. H. Wegkamp, Model selection in nonparametric regression. *Ann. Stat.* **31** (2003), 252–273.

OLEG V. LEPSKI

Aix-Marseille Université, Institut de Mathématiques de Marseille, 39, rue F. Joiliot Curie, 13453 Marseille, France, oleg.lepski@univ-amu.fr

MEAN ESTIMATION IN HIGH DIMENSION

GÁBOR LUGOSI

ABSTRACT

In this note we discuss the statistical problem of estimating the mean of a random vector based on independent, identically distributed data. This classical problem has recently attracted a lot of attention both in mathematical statistics and in theoretical computer science and numerous intricacies have been revealed. We discuss some of the recent advances, focusing on high-dimensional aspects.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 62G05; Secondary 62G35, 68Q32

KEYWORDS

Mean estimation, robustness, high-dimensional statistics

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5500–5514
DOI 10.4171/ICM2022/75

1. INTRODUCTION

We consider the statistical problem of estimating the mean of a random vector based on independent, identically distributed data. This seemingly innocent classical problem has drawn renewed attention both in mathematical statistics and theoretical computer science.

The problem is formulated as follows: let $X_1, \dots, X_n$ be independent, identically distributed random vectors taking values in $\mathbb{R}^d$ such that their mean $\mu = \mathbb{E}X_1$ exists. Upon observing these random variables, one would like to estimate the vector μ. An estimator $\widehat{\mu}_n = \widehat{\mu}_n(X_1, \dots, X_n)$ is simply a measurable function of the "data" $X_1, \dots, X_n$, taking values in $\mathbb{R}^d$.

Naturally, the standard empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the first estimate that comes to mind. Indeed, the strong law of large numbers guarantees that $\overline{\mu}_n$ converges to μ almost surely without any further conditions on the distribution. However, here we are interested in the finite-sample behavior of mean estimators and for any meaningful statement one needs to make further assumptions on the distribution. Throughout this note, we assume that the covariance matrix $\Sigma = \mathbb{E}(X_1 - \mu)(X_1 - \mu)^T$ exists.

The empirical mean is known to be sensitive to "outliers" that are inevitably present in the data when the distribution may be *heavy-tailed*. This concern gave rise to the area of *robust statistics*. Classical references include Huber [19], Huber and Ronchetti [20], Hampel, Ronchetti, Rousseeuw, and Stahel [14], Tukey [44].

The quality of an estimator may be measured in various ways. While most of the early statistical work focused on expected risk measures such as the *mean-squared error*

$$\mathbb{E}\big[\|\widehat{\mu}_n - \mu\|^2\big]$$

(with $\|\cdot\|$ denoting the Euclidean norm), such risk measures may be misleading. Indeed, if the distance $\|\widehat{\mu}_n - \mu\|$ is not sufficiently concentrated, the expected value does not necessarily reflect the "typical" behavior of the error. For such reasons, estimators $\widehat{\mu}_n$ that are close to μ *with high probability* are desirable.

Thus, our aim is to understand, for any given sample size n and confidence parameter $\delta \in (0,1)$, the smallest possible value $\varepsilon = \varepsilon(n,\delta)$ such that

$$\mathbb{P}\big\{\|\widehat{\mu}_n - \mu\| > \varepsilon\big\} \le \delta.$$

In Section 2 we briefly discuss the one-dimensional case and lay out some of the basic ideas behind the more complex high-dimensional estimators. In Section 3 we present so-called sub-Gaussian estimators that guarantee the optimal order of magnitude for the accuracy $\varepsilon(n,\delta)$. Finally, in Section 4 we discuss the more refined requirement of estimators being close to the mean in each direction.

Bibliographic remark. It is beyond the scope of this note to offer an exhaustive bibliography of the topic. We refer the reader to the recent—though already somewhat outdated—survey of Lugosi and Mendelson [27].

2. BASIC IDEAS: THE ONE-DIMENSIONAL CASE

First consider the case $d = 1$, that is, when the X_i are real-valued random variables. In this case, if σ^2 denotes the variance of X_1, then the central limit theorem guarantees that the empirical mean satisfies

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mu}_n - \mu| > \frac{\sigma\Phi^{-1}(1-\delta/2)}{\sqrt{n}}\right\} = \delta,$$

where $\Phi(x) = \mathbb{P}\{G \le x\}$ is the cumulative distribution function of a standard normal random variable G. This implies the slightly loose asymptotic inequality

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mu}_n - \mu| > \frac{\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right\} \le \delta.$$

Motivated by this property, we introduce a corresponding nonasymptotic notion as follows: for a given sample size n and confidence level δ, we say that a mean estimator $\widehat{\mu}_n$ is *L-sub-Gaussian* if there is a constant $L > 0$, such that, with probability at least $1 - \delta$,

$$|\widehat{\mu}_n - \mu| \le \frac{L\sigma\sqrt{\log(2/\delta)}}{\sqrt{n}}.$$

As it is pointed out in [11], if one considers the class of distributions with finite variance, the best accuracy one can hope for is of the order $\sqrt{\log(1/\delta)/n}$ and in this sense sub-Gaussian estimators are optimal. Perhaps surprisingly, sub-Gaussian estimators exist under the only assumption that the X_i have a finite second moment.

One such estimator is the so-called *median-of-means* estimator. It has been proposed in different forms in various papers, see Nemirovsky and Yudin [41], Jerrum, Valiant, and Vazirani [21], Alon, Matias, and Szegedy [1].

The definition of the median-of-means estimator calls for partitioning the data into k groups of roughly equal size, computing the empirical mean in each group, and taking the median of the obtained values.

Formally, recall that the median of k real numbers $x_1, \dots, x_k \in \mathbb{R}$ is defined as $M(x_1, \dots, x_k) = x_i$ where x_i is such that

$$\left|\{j \in [k] : x_j \le x_i\}\right| \ge \frac{k}{2} \quad \text{and} \quad \left|\{j \in [k] : x_j \ge x_i\}\right| \ge \frac{k}{2}.$$

(If several indices i fit the above description, we take the smallest one.)

Now let $1 \le k \le n$ and partition $[n] = \{1, \dots, n\}$ into k blocks $B_1, \dots, B_k$, each of size $|B_i| \ge \lfloor n/k \rfloor \ge 2$.

Given $X_1, \dots, X_n$, compute the sample mean in each block

$$Z_j = \frac{1}{|B_j|}\sum_{i\in B_j} X_i$$

and define the median-of-means estimator by $\widehat{\mu}_n = M(Z_1, \dots, Z_k)$.

To grasp intuitively why this estimator works, note that for each block, the empirical mean is an unbiased estimator of the mean, with controlled standard deviation $\sigma/\sqrt{n/k}$. Hence, the median of the distribution of the blockwise empirical mean lies within $\sigma/\sqrt{n/k}$ from the expectation. Now the empirical median is a highly concentrated estimator of this

median. Now it is easy to derive the following performance bound. For simplicity, assume that n is divisible by k so that each block has $m = n/k$ elements.

Let $X_1, \dots, X_n$ be independent, identically distributed random variables with mean μ and variance σ^2. For any $\delta \in (0,1)$, if $k = \lceil 8 \log(1/\delta) \rceil$, and $n = mk$, then, with probability at least $1 - \delta$, the median-of-means estimator $\widehat{\mu}_n$ satisfies

$$|\widehat{\mu}_n - \mu| \le \sigma \sqrt{\frac{32 \log(1/\delta)}{n}} \, .$$

In other words, the median-of-means estimator has a sub-Gaussian performance with $L = \sqrt{32}$ for all distributions with a finite variance.

An even more natural mean estimator is based on removing possible outliers using a truncation of X. Indeed, the so-called *trimmed-mean* (or *truncated*-mean) estimator is defined by removing a fraction of the sample, consisting of the εn largest and smallest points for some parameter $\varepsilon \in (0,1)$, and then averaging over the rest. This idea is one of the most classical tools in robust statistics, see, Tukey and McLaughlin [45], Huber and Ronchetti [20], Bickel [3], Stigler [43] for early work on the trimmed-mean estimator. The nonasymptotic sub-Gaussian property of the trimmed mean was established recently by Oliveira and Orenstein [42] who proved that if ε is chosen proportionally to $\log(1/\delta)/n$, then the trimmed-mean estimator has a sub-Gaussian performance for all distributions with a finite variance (see also [27]).

A quite different approach was introduced and analyzed by Catoni [4]. Catoni's idea is based on the fact that the empirical mean $\overline{\mu}_n$ is the solution $y \in \mathbb{R}$ of the equation

$$\sum_{i=1}^{n} (X_i - y) = 0.$$

Catoni proposed to replace the left-hand side of the equation above by another strictly decreasing function of y of the form

$$\sum_{i=1}^{n} \psi\big(\alpha(X_i - y)\big),$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is an antisymmetric increasing function and $\alpha \in \mathbb{R}$ is a parameter. The idea is that if $\psi(x)$ increases much slower than x, then the effect of "outliers" present due to heavy tails is diminished. Catoni offers a whole range of "influence" functions ψ and proves that by an appropriate choice of ψ the estimator has a sub-Gaussian performance.

We close this section by noting that in a recent work Lee and Valiant [23] construct a sub-Gaussian estimator with the (almost) optimal constant $L = \sqrt{2} + o(1)$. Their estimator builds on a clever combination of median of means, trimmed mean, and Catoni's estimator. A different approach was proposed by Minsker and Ndaoud [39]. Just like median of means, their mean estimator also starts by computing empirical averages on disjoint blocks of the data. Then they reweight the block averages in function of their empirical standard deviation. Using nontrivial properties of self-normalized sums, they obtain an estimator that is not only

sub-Gaussian but it is also asymptotically efficient, in the sense that the estimator is asymptotically normal with an asymptotic variance that is as small as possible in the minimax sense.

3. MULTIVARIATE SUB-GAUSSIAN ESTIMATORS

Next we discuss the substantially more complex multivariate problem. Recall that X is a random vector taking values in $\mathbb{R}^d$ with mean $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$. Given n independent, identically distributed samples $X_1, \dots, X_n$ drawn from the distribution of X, one wishes to estimate the mean vector μ.

In order to obtain guidance of what a desirable performance is for a mean estimator, it is instructive to consider the properties of the empirical mean $\overline{\mu}_n$ when X has a multivariate normal distribution. In that case, it is not difficult to see that the Gaussian concentration inequality implies that for $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\overline{\mu}_n - \mu\| \le \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}}\,.$$

where $\mathrm{Tr}(\Sigma)$ and $\lambda_{\max}$ denote the trace and spectral norm of the covariance matrix Σ. Inspired by this, we may generalize the definition of a sub-Gaussian mean estimator to the multivariate setting as follows: we say that for a given confidence level $\delta \in (0,1)$ and sample size n, a mean estimator $\widehat{\mu}_n$ is *sub-Gaussian* if there exists a constant C such that, for all distributions whose covariance matrix exists, with probability at least $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \le C\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}\log(1/\delta)}{n}}\right). \tag{3.1}$$

Naive attempts to generalize the one-dimensional median-of-means estimator do not necessarily achieve the desired sub-Gaussian property. For example, one may define the *geometric median-of-means* estimator defined as follows (see Minsker [37], Hsu and Sabato [18], Lerasle and Oliveira [25]): we start by partitioning $[n] = \{1, \dots, n\}$ into k blocks $B_1, \dots, B_k$, each of size $|B_i| \ge \lfloor n/k \rfloor \ge 2$, where $k \sim \log(1/\delta)$. Just like in the univariate case, we compute the sample mean within each block: for $j = 1, \dots, k$, let

$$Z_j = \frac{1}{m}\sum_{i \in B_j} X_i.$$

The estimator may be defined as the geometric median of the Z_j, defined as

$$\widehat{\mu}_n = \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{j=1}^{k} \|Z_i - m\|.$$

This estimator was proposed by Minsker [37] and independently by Hsu and Sabato [18] (see also Lerasle and Oliveira [25]). Minsker [37] proved that there exists a constant C such that, whenever the covariance matrix exists, with probability at least $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \le C\sqrt{\frac{\mathrm{Tr}(\Sigma)\log(1/\delta)}{n}}.$$

This is quite nice since the inequality does not require any assumption other than the existence of the covariance matrix. However, it is not quite a sub-Gaussian bound as in (3.1). An important advantage of the geometric median-of-means estimator is that it can be computed efficiently by solving a convex optimization problem. See Cohen, Lee, Miller, Pachocki, and Sidford [8] for a recent result and for the history of the computational problem.

3.1. Median-of-means tournaments

The existence of a sub-Gaussian mean estimator was first proved by Lugosi and Mendelson [29]. Their estimator is an instance of *median-of-means tournaments* and may be defined as follows. Let $Z_1, \dots, Z_k$ be the sample means within each block exactly as above. For each $a \in \mathbb{R}^d$, let

$$T_a = \left\{x \in \mathbb{R}^d : \exists J \subset [k] : |J| \geq k/2 \text{ such that for all } j \in J,\ \|Z_j - x\| \leq \|Z_j - a\|\right\} \quad (3.2)$$

and define the mean estimator by

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \operatorname{radius}(T_a),$$

where $\operatorname{radius}(T_a) = \sup_{x \in T_a} \|x - a\|$. Thus, $\widehat{\mu}_n$ is chosen to minimize, over all $a \in \mathbb{R}^d$, the radius of the set T_a defined as the set of points $x \in \mathbb{R}^d$ for which $\|Z_j - x\| \leq \|Z_j - a\|$ for the majority of the blocks. If there are several minimizers, one may pick any one of them.

The set T_a may be seen as the set of points in $\mathbb{R}^d$ that are at least as close to the point cloud $\{Z_1, \dots, Z_k\}$ as the point a. The estimator $\widehat{\mu}_n$ is obtained by minimizing the radius of T_a. The sub-Gaussian performance of this estimator is established in [29]:

> Let $X_1, \dots, X_n$ be independent, identically distributed random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. There exist constants $c, C > 0$ such that for any $\delta \in (0,1)$, if $k = c\lceil \log(1/\delta) \rceil$ and $n = mk$, then, with probability at least $1 - \delta$,
>
> $$\|\widehat{\mu}_n - \mu\| \leq C\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max} \log(1/\delta)}{n}}\right).$$

An equivalent way of defining the median-of-means tournament estimator is

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \sup_{u \in S^{d-1}} \left(\operatorname{Median}\left\{\langle Z_j, u\rangle\right\}_{j \in [k]} - \langle a, u\rangle\right).$$

We may regard this as another notion of multivariate median of the block centers $Z_1, \dots, Z_k$. Unfortunately, unlike the geometric median, computing this median is hard in the sense that computing it (at least in its naive implementation) takes time exponential in the dimension d. However, Hopkins [15] introduced a semidefinite relaxation of the median-of-means tournament estimator that can be computed in time $O(nd + d\log(1/\delta)^c)$ for a dimension-independent constant c and, at the same time, achieves the desired sub-Gaussian guarantee under the only assumption that the covariance matrix exists. Subsequent improvements managed to decrease the running time further. For example, Cherapanamjeri, Flammarion, and

Bartlett [7] combined Hopkins' ideas with gradient-descent optimization to construct an sub-Gaussian mean estimator that is computable in time $O(nd + d\log(1/\delta)^2 + \log(1/\delta)^4)$. Based on ideas of "spectral reweighting" of Cheng, Diakonikolas, and Ge [6], Depersin and Lecué [9], and Lei, Luh, Venkat, and Zhang [24] further improve the running time. Hopkins, Li, and Zhang [17] show how spectral reweighting is essentially equivalent to the median notion introduced above. We refer to these papers for an exhaustive review of the rapidly growing literature of computational aspects of robust mean estimation.

3.2. Multivariate trimmed mean

Here we describe a quite different construction that also results in a sub-Gaussian mean estimator. The estimator, proposed and analyzed by Lugosi and Mendelson [31], is a multivariate version of the trimmed-mean estimator discussed in Section 2. The construction is as follows.

First split the data in two halves. For simplicity of the exposure, suppose we have $2n$ data points $X_1, \dots, X_n, Y_1, \dots, Y_n$. Set $\varepsilon = c\frac{\log(1/\delta)}{n}$ for an appropriate constant $c > 0$. For every $v \in S^{d-1}$, let α_v and β_v be the empirical $\varepsilon/2$ and $1-\varepsilon/2$ quantiles based on the second half of the data $Y_1, \dots, Y_n$. Define

$$\phi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x > \beta, \\ x & \text{if } x \in [\alpha, \beta], \\ \alpha & \text{if } x < \alpha. \end{cases}$$

and for a parameter $Q > 0$, compute the univariate trimmed estimators

$$U_Q(v) = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle X_i, v\rangle\big).$$

Each of these estimators is just the trimmed mean estimator of $\mathbb{E}\langle X, v\rangle = \langle \mu, v\rangle$ for a given direction v. Note that the trimming interval is widened by the global parameter Q whose role is to make sure that the univariate estimators work simultaneously. In order to convert the estimators of the projected means into a single vector, define the "slabs"

$$\Gamma(v, Q) = \big\{x \in \mathbb{R}^d : \big|\langle x, v\rangle - U_Q(v)\big| \le 2\varepsilon Q\big\}$$

and let

$$\Gamma(Q) = \bigcap_{v \in S^{d-1}} \Gamma(v, Q).$$

If $x \in \Gamma(Q)$, then the projection of x to every direction v is close to the trimmed mean estimator of $\langle \mu, v\rangle$. The main technical result of [31] is that, when

$$Q \sim \max\left(\frac{1}{\varepsilon}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, \sqrt{\frac{\lambda_1}{\varepsilon}}\right),$$

the set $\Gamma(Q)$ contains the mean μ, with probability $1-\delta$. Since the diameter of $\Gamma(Q)$ is at most $4\varepsilon Q$, this guarantees the sub-Gaussian property of any element of the set $\Gamma(Q)$. The problem with such an estimator is that its construction requires knowledge of the correct value of Q that depends on the (unknown) covariance matrix Σ. This problem may

be circumvented by a simple adaptive choice of Q: let i^* be the smallest integer such that $\bigcap_{i \geq i^*} \Gamma(2^i) \neq \emptyset$. Then define $\widehat{\mu}_n$ to be any point in the set

$$\bigcap_{i \in \mathbb{Z}: i \geq i^*} \Gamma\left(2^i\right).$$

This choice is sufficient to guarantee the sub-Gaussian property of the estimator.

Remark. In some situations the Euclidean norm is not necessarily the most adequate way of measuring the accuracy of a mean estimator. Hence, it is natural to ask the following: given a norm $\|\cdot\|$, a confidence parameter $\delta \in (0,1)$, and an i.i.d. sample of cardinality n, what is the best possible accuracy ε for which there exists a mean estimator $\widehat{\mu}_n$ for which

$$\|\widehat{\mu}_n - \mu\| \leq \varepsilon \quad \text{with probability at least} 1 - \delta?$$

The optimal order of magnitude of ε is now well understood even in this general setting, see Lugosi and Mendelson [28], Bahmani [2], Depersin and Lecué [10].

4. DIRECTION-DEPENDENT ACCURACY

An equivalent way of formulating the sub-Gaussian inequality (3.1) for a mean estimator $\widehat{\mu}_n$ is as follows: with probability at least $1 - \delta$,

$$\forall u \in S^{d-1} : \langle \widehat{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\lambda_1 \log(1/\delta)}{n}} + \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} \right), \tag{4.1}$$

where $\lambda_1 \geq \cdots \geq \lambda_d$ denote the eigenvalues of the covariance matrix Σ and $\mathrm{Tr}(\Sigma) = \sum_{i=1}^d \lambda_i$. We refer to the two terms on the right-hand side as the *weak* and *strong* terms. The strong term corresponds to a global component, while the weak term controls fluctuations in the worst direction, leading to the weak term which involves λ_1.

If one wanted to estimate the projection $\langle \mu, u \rangle$ in a fixed direction $u \in S^{d-1}$ by an estimator $\widehat{v}_n(u)$, as discussed in Section 2, the best accuracy one could hope for would be

$$\left| \widehat{v}_n(u) - \langle \mu, u \rangle \right| \leq C \sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}},$$

where $\sigma^2(u) = \mathrm{Var}(\langle X, u \rangle)$. Now it is natural to ask whether one can improve the inequality of (4.1) in a direction-sensitive way. In particular, a natural question is if the weak term on the right-hand side of (4.1) can be improved to $\sqrt{\sigma^2(u) \log(1/\delta)/n}$ and if it can, what price one has to pay in the strong term for such an improvement. This problem was studied by Lugosi and Mendelson [30] and in this section we recall the main results of that paper.

Once again, we turn to the canonical case of Gaussian vectors to obtain guidance about what kind of properties one can hope for. One can show (see [30]) that if the X_i are independent Gaussian vectors, then the empirical mean $\overline{\mu}_n$ satisfies that, with probability at least $1 - \delta$,

$$\forall u \in S^{d-1} : \langle \overline{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} \right),$$

where C is a numerical constant. Thus, in the Gaussian case one can indeed obtain a weak term that scales optimally, without giving up anything in the strong term. In fact, the bound can be slightly improved to

$$\forall u \in S^{d-1} : \langle \overline{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\sum_{i>k_1} \lambda_i}{n}} \right)$$

where $k_1 = c \log(1/\delta)$, for some constant c. This bound is, in fact, the best one can hope for in the following sense:

Proposition 1. *Let $\overline{\mu}_n = (1/n) \sum_{i=1}^n X_i$ where the X_i are independent Gaussian vectors with mean μ and covariance matrix Σ. Suppose that there exists a constant C such that, for all δ, n, μ, and Σ, with probability at least $1 - \delta$,*

$$\forall u \in S^{d-1} : \langle \overline{\mu}_n - \mu, u \rangle \leq C \sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + S. \tag{4.2}$$

Then there exists a constant C' depending on C only, such that the "strong term" S has to satisfy

$$S \geq C' \sqrt{\frac{\sum_{i>k_0} \lambda_i}{n}}$$

where $k_0 = 1 + (2C + \sqrt{2})^2 \log(1/\delta)$.

The observation above shows that even in the well-behaved example of a Gaussian distribution, the strong term needs to be at least of the order

$$\sqrt{\frac{\sum_{i>k} \lambda_i}{n}}$$

where k is proportional to $\log(1/\delta)$.

The main result of [30] is that under an additional assumption on the distribution of X, one can construct an estimator that, up to the optimal strong term, preforms in every direction as if it were an optimal estimator of the one-dimensional marginal:

Let $X_1, \ldots, X_n$ be i.i.d. random vectors, taking values in $\mathbb{R}^d$, with mean μ and covariance matrix Σ whose eigenvalues are $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d \geq 0$. Suppose that there exists $q > 2$ and a constant κ such that, for all $u \in S^{d-1}$,

$$\left(\mathbb{E}\left|\langle X - \mu, u \rangle\right|^q\right)^{1/q} \leq \kappa \left(\mathbb{E}\langle X - \mu, u \rangle^2\right)^{1/2}. \tag{4.3}$$

Then for every $\delta \in (0,1)$ there exists a mean estimator $\widehat{\mu}_n$ and constants $0 < c, c', C < \infty$ (depending on κ and q only) such that, if $\delta \geq e^{-c'n}$, then, with probability, at least $1 - \delta$,

$$\forall u \in S^{d-1} : \langle \widehat{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\sum_{i=c \log(1/\delta)}^{d} \lambda_i}{n}} \right). \tag{4.4}$$

Mean estimators with sub-Gaussian performance of the type (3.1) exist without assuming anything more than the existence of the covariance matrix. However, to achieve the improved direction-dependent performance formulated above, we need to assume that moments of order q exist for some $q > 2$. Moreover, we assume that the L_q norm of each one-dimensional marginal is related to the L_2-norm in a uniform manner, as described by (4.3). We call this a *norm-equivalence* condition. This condition is used repeatedly in a crucial way in the construction of the estimator. It is an intriguing question whether such a condition is necessary or if there exists a mean estimator satisfying an inequality of the type (4.4) under the only assumption of finite second moments. The mean estimator and the constants in the performance bound depend on the values κ and q of the norm-equivalence condition.

Next we describe the construction of the mean estimator. It is a quite complex variation of the trimmed mean estimator described in the previous section. In the form defined here, it is hopeless to have an algorithm that computes it efficiently, that is, in time polynomial in the sample size, the dimension, and $\log(1/\delta)$. It is an open question how far computationally efficient mean estimators can reach in terms of their statistical performance. In particular, it would be interesting to understand whether there is a true (i.e., rigorously provable) conflict between statistical accuracy and computational efficiency in the mean estimation problem. We note that in the related problem of robust mean estimation under adversarial contamination, such conflicts indeed seem to exist, see Hopkins and Li [16].

In the first step of the construction of the estimator, we divide the sample $X_1, \dots, X_n$ into n/m blocks of size m and compute, for each block

$$Y_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} X_{m(j-1)+i}.$$

Here m is chosen to be a constant depending on q and κ, the constants appearing in the norm equivalence condition. The purpose of this "smoothing" is to ensure that the Y_j satisfy certain "small-ball" properties.

Next, for each direction $u \in S^{d-1}$, we compute the trimmed-mean estimators

$$\widehat{v}_n(u) = \frac{1}{\sqrt{m}} \frac{1}{n/m - 2\theta n/m} \sum_{j \in [n/m] \setminus J_+(u) \cup J_-(u)} Y_j,$$

where the sets $J_+(u)$ and $J_-(u)$ correspond to the indices of the $\theta n/m$ smallest and $\theta n/m$ largest values of $\langle Y_j, u \rangle$ and $\theta \in (0, 1/2)$ is another constant that depends on q and κ only.

Now one can prove that the directional mean estimates $\widehat{v}_n(u)$ work as desired, simultaneously for all $u \in S^{d-1}$. More precisely, there exist constants $c, C' > 0$ depending on κ and q such that, with probability at least $1 - \delta$, for all $u \in S^{d-1}$,

$$\left|\widehat{v}_n(u) - \langle \mu, u \rangle\right| \le C' \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d} \lambda_i}{n}} \right).$$

Once we have the "directional" mean estimators $\widehat{v}_n(u)$ with the desired property, similarly to the multivariate trimmed-mean estimator discussed in Section 3 above, we need to find a vector $\widehat{\mu}_n$ such that $\langle \widehat{\mu}_n, u \rangle$ is close to $\widehat{v}_n(u)$ for all $u \in S^{d-1}$ (at the appropriate direction-dependent scale).

To this end, similarly to the case of the trimmed-mean estimator, we define "slabs." In order to define slabs of the correct width, we need to estimate the directional variances $\sigma^2(u)$. This is the problem of *covariance estimation* that has received quite a lot of attention, see Catoni [5], Giulini [13], Koltchinskii and Lounici [22], Lounici [26], Mendelson [35], Mendelson and Zhivotovksiy [36], Minsker [38], Minsker and Wei [40] for a sample of the relevant literature.

For our purposes, we only need to accurately estimate the variances $\sigma^2(u)$ in those directions $u \in S^{d-1}$ in which the variance is "not too small," meaning that it is above a certain critical level. Below the critical level, all we need is that the estimator detects that the variance is small. More precisely, we construct an estimator $\psi_n(u)$, such that, on an event of probability at least $1 - e^{-cn}$,

$$\begin{aligned} &\frac{1}{4}\sigma^2(u) \le \psi_n(u) \le 2\sigma^2(u) \quad \forall u \in S^{d-1} \text{ such that } \sigma^2(u) \ge r^2, \\ &\psi_n(u) \le Cr^2 \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

Here c and C are constants depending on κ and q only and

$$r = \sqrt{\frac{c_0}{n}\sum_{i \ge c_0 n} \lambda_i}$$

for another constant $c_0 > 0$ depending on κ and q.

Once such a covariance estimator $\psi_n(u)$ is constructed, for a parameter $\rho > 0$, we may define the slabs

$$E_{u,\rho} = \left\{ v \in \mathbb{R}^d : \left| \widehat{v}_n(u) - \langle v, u \rangle \right| \le \rho + 2C' \sqrt{\frac{\psi_n(u)\log(1/\delta)}{n}} \right\}$$

and let

$$S_\rho = \bigcap_{u \in S^{d-1}} E_{u,\rho}.$$

Since $\rho > 0$, the set S_ρ is compact, and therefore the set

$$S = \bigcap_{\rho > 0 : S_\rho \neq \emptyset} S_\rho$$

is not empty. We may now define the mean estimator as any element $\widehat{\mu}_n \in S$. This estimator satisfies the announced property.

It remains to define the variance estimator $\psi_n(u)$. To this end, first we define

$$\tilde{X}_i = \frac{X_i - X_i'}{2}, \quad i \in [n]$$

(defined on a sample of size $2n$ that is independent of that used to construct the directional mean estimators $\widehat{v}_n(u)$) to obtain a sample of centered vectors with the same covariance as X.

Next we divide this sample into n/m equal blocks, where m is an appropriately chosen constant (depending on κ and q). For each block, we compute

$$Z_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \tilde{X}_{m(j-1)+i}.$$

The purpose of this step is to guarantee a certain "small-ball" property of the distribution, similarly to the definition of $\widehat{\nu}_n(u)$. Once again, $\psi_n(u)$ is a trimmed-mean estimator. More precisely, for every $u \in S^{d-1}$, if we denote by $J_+(u)$ the set of indices of the $\theta n/m$ largest values of $\langle Z_j, u\rangle$, we define

$$\psi_n(u) = \frac{1}{n/m} \sum_{j \in [n/m] \setminus J_+(u)} \langle Z_j, u\rangle^2.$$

The proof of the desired properties of both the directional mean estimator $\widehat{\nu}_n(u)$ and directional variance estimator $\psi_n(u)$ relies on novel bounds for the ratio of empirical and true probabilities that hold uniformly over certain classes of random variables. The main technical machinery that leads to the necessary directional control requires bounds for *ratios* of empirical and true probabilities that hold uniformly in a class of functions. Informally, one needs to control

$$\sup_{\{f \in \mathcal{F}, \|f\|_{L_2} \geq r\}} \sup_{t : \mathbb{P}\{f(X) > t\} \geq \Delta} \left| \frac{n^{-1} \sum_{i=1}^n \mathbb{1}_{f(X_i) > t}}{\mathbb{P}\{f(X) > t\}} - 1 \right|$$

for appropriate values of r and Δ.

In other words, in [30] it is shown that, under minimal assumptions on the class $\mathcal{F}$, the empirical frequencies of level sets of every $f \in \mathcal{F}$ are close, in a multiplicative sense, to their true probabilities, as long as $\|f\|_{L_2} = \sqrt{\mathbb{E} f(X)^2}$ and $\mathbb{P}\{f(X) > t\}$ are large enough. Estimates of this flavor had been derived before, but only in a limited scope. Examples include the classical inequalities of Vapnik–Chervonenkis in VC theory, dealing with small classes of binary-valued functions (see also, Giné and Koltchinskii [12] for some results for real-valued classes). Existing ratio estimates are often based on the restrictive assumption that the collection of level sets, say of the form $\{\{x : f(x) > t\} : f \in \mathcal{F},\ t \geq t_0\}$, is small in the VC sense.

The method developed in [30] is based on a completely different argument that builds on the so-called *small-ball method* pioneered by Mendelson [32–34].

5. CONCLUSION

The problem of estimating the mean of a random vector has received a lot of recent attention both in mathematical statistics and in theoretical computer science. Understanding the possibilities and limitations of general mean estimation is an intriguing problem and the computational aspects enrich the area further with many nontrivial and exciting questions. In spite of the significant progress, many interesting questions remain to be explored. The lessons learnt from this prototypical statistical problem are expected to infuse other areas of statistics and machine learning with valuable ideas.

ACKNOWLEDGMENTS

Shahar Mendelson was my coauthor and main driving force behind a series of papers that form the basis of this article. Thanks Shahar for all the fun!

FUNDING

This work was supported by the Spanish Ministry of Economy and Competitiveness, Grant PGC2018-101643-B-I00 and by "Google Focused Award Algorithms and Learning for AI".

REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy, The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** (2002), 137–147.

[2] S. Bahmani, Nearly optimal robust mean estimation via empirical characteristic function. *Bernoulli* **27** (2021), no. 3, 2139–2158.

[3] P. Bickel, On some robust estimates of location. *Ann. Math. Stat.* **36** (1965), 847–858.

[4] O. Catoni, Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** (2012), no. 4, 1148–1185.

[5] O. Catoni, Pac-Bayesian bounds for the Gram matrix and least squares regression with a random design. 2016, arXiv:1603.05229.

[6] Y. Cheng, I. Diakonikolas, and R. Ge, High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the thirtieth annual ACM–SIAM symposium on discrete algorithms*, pp. 2755–2771, SIAM, 2019.

[7] Y. Cherapanamjeri, N. Flammarion, and P. Bartlett, Fast mean estimation with sub-Gaussian rates. 2019, arXiv:1902.01998.

[8] M. Cohen, Y. Lee, G. Miller, J. Pachocki, and A. Sidford, Geometric median in nearly linear time. In *Proceedings of the 48th annual ACM SIGACT symposium on theory of computing*, pp. 9–21, ACM, 2016.

[9] J. Depersin and G. Lecué, Robust subgaussian estimation of a mean vector in nearly linear time. 2019, arXiv:1906.03058.

[10] J. Depersin and G. Lecué, Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. 2021, arXiv:2102.00995.

[11] L. Devroye, M. Lerasle, G. Lugosi, and R. Oliveira, Sub-Gausssian mean estimators. *Ann. Statist.* **44** (2016), 2695–2725.

[12] E. Giné and V. Koltchinskii, Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** (2006), no. 3, 1143–1216.

[13] I. Giulini, Robust dimension-free Gram operator estimates. *Bernoulli* **24** (2018), 3864–3923.

[14] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust statistics: the approach based on influence functions*. Wiley Ser. Probab. Stat. 196, John Wiley & Sons, 1986.

[15] S. Hopkins, Sub-Gaussian mean estimation in polynomial time. *Ann. Statist.* **48** (2020), no. 2, 1193–1213.

[16] S. B. Hopkins and J. Li, How hard is robust mean estimation? In *Conference on learning theory*, pp. 1649–1682, PMLR, 2019.

[17] S. B. Hopkins, J. Li, and F. Zhang, Robust and heavy-tailed mean estimation made simple, via regret minimization. 2020, arXiv:2007.15839.

[18] D. Hsu and S. Sabato, Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** (2016), 1–40.

[19] P. Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35** (1964), no. 1, 73–101.

[20] P. Huber and E. Ronchetti, *Robust statistics*. Wiley, New York, 2009.

[21] M. Jerrum, L. Valiant, and V. Vazirani, Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** (1986), 186–188.

[22] V. Koltchinskii and K. Lounici, Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** (2017), no. 1, 110–133.

[23] J. C. Lee and P. Valiant, Optimal sub-gaussian mean estimation in R. 2020, arXiv:2011.08384.

[24] Z. Lei, K. Luh, P. Venkat, and F. Zhang, A fast spectral algorithm for mean estimation with sub-Gaussian rates. In *Conference on learning theory*, pp. 2598–2612, PMLR, 2020.

[25] M. Lerasle and R. I. Oliveira, Robust empirical mean estimators. 2012, arXiv:1112.3914.

[26] K. Lounici, High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** (2014), no. 3, 1029–1058.

[27] G. Lugosi and S. Mendelson, Mean estimation and regression under heavy-tailed distributions—a survey. *Found. Comput. Math.* **19** (2019), no. 5, 1145–1190.

[28] G. Lugosi and S. Mendelson, Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields* **175** (2019), 957–973.

[29] G. Lugosi and S. Mendelson, Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** (2019), 783–794.

[30] G. Lugosi and S. Mendelson, Multivariate mean estimation with direction-dependent accuracy. 2020, arXiv:2010.11921.

[31] G. Lugosi and S. Mendelson, Robust multivariate mean estimation: the optimality of trimmed mean. *Ann. Statist.* **49** (2021), 393–410.

[32] S. Mendelson, Learning without concentration. *J. ACM* **62** (2015), 21.

[33] S. Mendelson, An optimal unrestricted learning procedure. 2017, arXiv:1707.05342.

[34] S. Mendelson, Learning without concentration for general loss functions. *Probab. Theory Related Fields* **171** (2018), no. 1–2, 459–502.

[35] S. Mendelson, Approximating the covariance ellipsoid. *Commun. Contemp. Math.* **22** (2020), no. 08, 1950089.

[36] S. Mendelson and N. Zhivotovskiy, Robust covariance estimation under L_4–L_2 norm equivalence. 2018, arXiv:1809.10462.

[37] S. Minsker, Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** (2015), 2308–2335.

[38] S. Minsker, Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** (2018), 2871–2903.

[39] S. Minsker and M. Ndaoud, Robust and efficient mean estimation: approach based on the properties of self-normalized sums. 2020, arXiv:2006.01986.

[40] S. Minsker and X. Wei, Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli* **26** (2020), no. 1, 694–727.

[41] A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[42] R. I. Oliveira and P. Orenstein, *The sub-Gaussian property of trimmed means estimators*. Tech. rep., IMPA, 2019.

[43] S. Stigler, The asymptotic distribution of the trimmed mean. *Ann. Statist.* **1** (1973), 472–477.

[44] J. Tukey, Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, pp. 523–531, 1975.

[45] J. Tukey and D. McLaughlin, Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorization 1. *Sankhya, Ser. A* **25** (1963), 331–352.

GÁBOR LUGOSI

Department of Economics and Business, Pompeu Fabra University, ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain, and Barcelona School of Economics, Barcelona, Spain, gabor.lugosi@gmail.com

ON SOME INFORMATION-THEORETIC ASPECTS OF NON-LINEAR STATISTICAL INVERSE PROBLEMS

RICHARD NICKL AND GABRIEL P. PATERNAIN

ABSTRACT

Results by van der Vaart (1991) from semi-parametric statistics about the existence of a non-zero Fisher information are reviewed in an infinite-dimensional non-linear Gaussian regression setting. Information-theoretically optimal inference on aspects of the unknown parameter is possible if and only if the adjoint of the linearisation of the regression map satisfies a certain range condition. It is shown that this range condition may fail in a commonly studied elliptic inverse problem with a divergence form equation ('Darcy's problem'), and that a large class of smooth linear functionals of the conductivity parameter cannot be estimated efficiently in this case. In particular, Gaussian 'Bernstein von Mises'-type approximations for Bayesian posterior distributions do not hold in this setting.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 62G20; Secondary 35R30

KEYWORDS

Bernstein von Mises theorems, Darcy's problem, adjoint score condition

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5516–5538
DOI 10.4171/ICM2022/18

1. INTRODUCTION

The study of *inverse problems* forms an active scientific field at the interface of the physical, mathematical and statistical sciences and machine learning. A common setting is where one considers a 'forward map' $\mathcal{G}$ between two spaces of functions, and the 'inverse problem' is to recover θ from the 'data' $\mathcal{G}_\theta \equiv \mathcal{G}(\theta)$. In real-world measurement settings, data is observed *discretely*, for instance one is given point evaluations $\mathcal{G}(\theta)(X_i)$ of the function $\mathcal{G}(\theta)$ on a finite discretisation $\{X_i\}_{i=1}^N$ of the domain of $\mathcal{G}_\theta$. Each time a measurement is taken, a statistical error is incurred, and the resulting noisy data can then be described by a statistical regression model $Y_i = \mathcal{G}_\theta(X_i) + \varepsilon_i$, with regression functions $\{\mathcal{G}_\theta : \theta \in \Theta\}$ indexed by the parameter space Θ. Such models have been studied systematically at least since C. F. Gauss [9] and constitute a core part of statistical science ever since.

In a large class of important applications, the family of regression maps $\{\mathcal{G}_\theta : \theta \in \Theta\}$ arises from physical considerations and is described by a *partial differential equation* (PDE). The functional parameter θ is then naturally *infinite- (or after discretisation step, high-) dimensional*, and the map $\theta \mapsto \mathcal{G}_\theta$ is often *non-linear*, which poses challenges for statistical inference. Algorithms for such 'non-convex' problems have been proposed and developed in the last decade since influential work by A. Stuart [28], notably based on ideas from *Bayesian inference*, where the parameter θ is modelled by a Gaussian process (or related) prior Π. The inverse problem is 'solved' by approximately computing the posterior measure $\Pi(\cdot|(Y_i, X_i)_{i=1}^N)$ on Θ by an iterative (e.g. MCMC) method. While the success of this approach has become evident empirically, an objective mathematical framework that allows giving rigorous statistical and computational guarantees for such algorithms in non-linear problems has only emerged more recently. The types of results obtained so far include statistical *consistency and contraction rate* results for posterior distributions and their means, see [1, 13, 19] and also [14, 16, 21–23], as well as *computational guarantees* for MCMC based sampling schemes [3, 15, 25].

Perhaps the scientifically most desirable guarantees are those for 'statistical uncertainty quantification' methods based on posterior distributions, and these are notoriously difficult to obtain. Following a programme originally developed by [4–6, 26] in classical 'direct' regression models, one way to address this issue is by virtue of the so-called *Bernstein–von Mises theorems* which establish asymptotically (as $N \to \infty$) exact Gaussian approximations to posterior distributions. These exploit the precise but delicate machinery from semi-parametric statistics and Le Cam theory (see [31]) and aim at showing that the actions $\langle \psi, \theta \rangle | (Y_i, X_i)_{i=1}^N$ of infinite-dimensional posterior distributions on a well-chosen set of test functions ψ converge – after rescaling by $\sqrt{N}$ (and appropriate re-centering) – to fixed normal $\mathcal{N}(0, \sigma_\theta^2(\psi))$-distributions (with high probability under the data $(Y_i, X_i)_{i=1}^N$). The limiting variance $\sigma_\theta^2(\psi)$ has an information-theoretic interpretation as the *Cramér–Rao lower bound* (inverse Fisher information) of the model (see also Section 2.4). Very few results of this type are currently available in PDE settings. Recent progress in [20] (see also related work in [12, 18, 21, 22]) has revealed that Bernstein–von Mises theorems may hold true if the PDE underlying $\mathcal{G}_\theta$ has certain analytical properties. Specifically, one has to solve

'information equations' that involve the 'information operator' $D\mathscr{G}_\theta^* D\mathscr{G}_\theta$ generated by the linearisation $D\mathscr{G}_\theta$ of $\mathscr{G}_\theta$ (with appropriate adjoint $D\mathscr{G}_\theta^*$). The results in [20, 21] achieve this for a class of PDEs where a base differential operator (such as the Laplacian, or the geodesic vector field) is attenuated by an unknown potential θ, and where ψ can be any smooth test function.

In the present article we study a different class of elliptic PDEs commonly used to model steady state diffusion phenomena, and frequently encountered as a 'fruitfly example' of a non-linear inverse problem in applied mathematics ('Darcy's problem'; see the many references in [13, 28]). While this inverse problem can be solved in a statistically consistent way (with 'nonparametric convergence rates' to the ground truth, see [13, 24]), we show here that, perhaps surprisingly, semi-parametric Bernstein–von Mises phenomena for posterior distributions of a large class of linear functionals of the relevant 'conductivity' parameter *do in fact not hold* for this PDE, not even just locally in a 'smooth' neighbourhood of the standard Laplacian. See Theorems 6 and 7, which imply in particular that the inverse Fisher information $\sigma_\theta^2(\psi)$ does not exist for a large class of smooth ψ's. The results are deduced from a theorem of van der Vaart [30] in general statistical models, combined with a thorough study of the mapping properties of $D\mathscr{G}_\theta$ and its adjoint for the PDE considered. Our negative results should help to appreciate the mathematical subtlety underpinning exact Gaussian approximations to posterior distributions in non-linear inverse problems arising with PDEs.

2. INFORMATION GEOMETRY IN NON-LINEAR REGRESSION MODELS

In this section we review some by now classical material on information-theoretical properties of infinite-dimensional regular statistical models [30, 31], and develop the details for a general vector-valued non-linear regression model relevant in inverse problems settings. Analogous results could be obtained in the idealised Gaussian white noise model (cf. Chapter 6 in [11]) sometimes considered in the inverse problems literature.

2.1. Measurement setup

Let $(\mathcal{X}, \mathcal{A}, \lambda)$ be a probability space and let V be a finite-dimensional vector space of fixed finite dimension $p_V \in \mathbb{N}$ with inner product $\langle \cdot, \cdot \rangle_V$ and norm $|\cdot|_V$. We denote by $L^\infty(\mathcal{X})$ and $L^2(\mathcal{X}) = L^2_\lambda(\mathcal{X}, V)$ the bounded measurable and λ-square integrable V-valued functions defined on $\mathcal{X}$ normed by $\|\cdot\|_\infty$ and $\|\cdot\|_{L^2_\lambda(\mathcal{X})}$, respectively. The inner product on $L^2(\mathcal{X})$ is denoted by $\langle \cdot, \cdot \rangle_{L^2(\mathcal{X})}$. We will also require Hilbert spaces $L^2(P) = L^2(V \times \mathcal{X}, P)$ of real-valued functions defined on $V \times \mathcal{X}$ that are square integrable with respect to a probability measure P on the produce space $V \times \mathcal{X}$, with inner product $\langle \cdot, \cdot \rangle_{L^2(P)}$.

We will consider a parameter space Θ that is subset of a (separable) Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ on which measurable 'forward maps'

$$\theta \mapsto \mathscr{G}(\theta) = \mathscr{G}_\theta, \quad \mathscr{G} : \Theta \to L^2_\lambda(\mathcal{X}, V), \tag{2.1}$$

are defined. Observations then arise in a general random design regression setup where one is given jointly i.i.d. random vectors $(Y_i, X_i)_{i=1}^N$ of the form

$$Y_i = \mathcal{G}_\theta(X_i) + \varepsilon_i, \quad \varepsilon_i \sim^{\text{i.i.d}} \mathcal{N}(0, I_V),\ i = 1, \dots, N, \tag{2.2}$$

where the X_i's are random i.i.d. covariates drawn from law λ on $\mathcal{X}$. We assume that the covariance I_V of each Gaussian noise vector $\varepsilon_i \in V$ is diagonal for the inner product of V. [Most of the content of this section is not specific to Gaussian errors ε_i in (2.2), cf. Example 25.28 in [31] for discussion.]

We consider a 'tangent space' H at any fixed $\theta \in \Theta$ such that H is a linear subspace of $\mathbb{H}$ and such that perturbations of θ in directions $h \in H$ satisfy $\{\theta + sh, h \in H, s \in \mathbb{R}, |s| < \epsilon\} \subset \Theta$ for some ϵ small enough. We denote by $\bar{H}$ the closure of H in $\mathbb{H}$ and will regard $\bar{H}$ itself as a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$. We employ the following assumption in the sequel.

Condition 1. *Suppose $\mathcal{G}$ is uniformly bounded* $\sup_{\theta\in\Theta} \|\mathcal{G}(\theta)\|_\infty \le U_{\mathcal{G}}$. *Moreover, for fixed $\theta \in \Theta$, $x \in \mathcal{X}$, and every $h \in H$, suppose that $\mathcal{G}_\theta(x)$ is Gateaux-differentiable in direction h, that is, for all $x \in \mathcal{X}$,*

$$\left|\mathcal{G}(\theta + sh)(x) - \mathcal{G}(\theta)(x) - s\mathbb{I}_\theta[h](x)\right|_V = o(s) \quad \text{as } s \to 0, \tag{2.3}$$

for some continuous linear operator $\mathbb{I}_\theta : (H, \|\cdot\|_{\mathbb{H}}) \to L^2_\lambda(\mathcal{X}, V)$, and that for some $\epsilon > 0$ small enough and some finite constant $B = B(h, \theta)$,

$$\sup_{|s|<\epsilon} \frac{\|\mathcal{G}(\theta + sh) - \mathcal{G}(\theta)\|_\infty}{|s|} \le B. \tag{2.4}$$

2.2. The DQM property

We will now derive the semi-parametric 'score' and 'information' operators (cf. [30, 31]) in the observational model (2.2). If P_θ is the law of $(Y_1, X_1) = (\mathcal{G}(\theta)(X_1) + \varepsilon_1, X_1)$ on $V \times \mathcal{X}$ then (2.2) is an i.i.d. statistical model of product laws

$$\mathcal{P}_N = \left\{P_\theta^N = \otimes_{i=1}^N P_\theta : \theta \in \Theta\right\}, \quad N \in \mathbb{N}, \tag{2.5}$$

on $(V \times \mathcal{X})^N$, and we can identify all information-theoretic properties in terms of the model $\mathcal{P} = \mathcal{P}_1 = \{P_\theta : \theta \in \Theta\}$ for the coordinate distributions. The model $\mathcal{P}$ is *differentiable in quadratic mean (DQM)* at $\theta \in \Theta$ along the tangent space H with score operator

$$\mathbb{A}_\theta : H \to L^2(V \times \mathcal{X}, P_\theta) \tag{2.6}$$

(cf. (3.2) in [30]) if for each path $\theta_{s,h} = \theta + sh$, $h \in H$, we have as $s \to 0$,

$$\int_{V\times\mathcal{X}} \left[\frac{1}{s}\left(dP_{\theta_{s,h}}^{1/2} - dP_\theta^{1/2}\right) - \frac{1}{2}\mathbb{A}_\theta[h]\, dP_\theta^{1/2}\right]^2 \to 0 \tag{2.7}$$

where

$$dP_\theta^{1/2}(y, x) = (2\pi)^{-p_V/4} e^{-|y-\mathcal{G}(\theta)(x)|_V^2/4}\, dy\, dx, \quad y \in V, x \in \mathcal{X},$$

are the square-root probability densities of P_θ with respect to Lebesgue measure on $V \times \mathcal{X}$.

Theorem 1. *Assuming Condition 1, the model (2.5) is differentiable in quadratic mean (DQM) at $\theta \in \Theta$ along every path $(\theta + sh : |s| < \epsilon, h \in H)$ with ϵ small enough. The 'score' operator $\mathbb{A}_\theta : H \to L^2(V \times \mathcal{X}, P_\theta)$ is given by*

$$\mathbb{A}_\theta[h](y,x) = \big\langle y - \mathscr{G}(\theta)(x), \mathbb{I}_\theta(h)(x)\big\rangle_V, \quad h \in H,\ (y,x) \in V \times \mathcal{X}, \tag{2.8}$$

which extends to a continuous linear operator $\mathbb{A}_\theta : \bar{H} \to L^2(P_\theta)$.

Proof. Fix $h \in H$. Using that the densities dP_θ are strictly positive, the left-hand side in (2.7) equals

$$\begin{aligned}
&\int_{V\times\mathcal{X}} \left[\frac{1}{s}\left(\frac{dP_{\theta_{s,h}}^{1/2}}{dP_\theta^{1/2}} - 1\right) - \frac{1}{2}\mathbb{A}_\theta[h]\right]^2 dP_\theta \\
&\quad = \int_{V\times\mathcal{X}} \left[\frac{1}{s}\Big[e^{\langle \frac{y}{2}, \mathscr{G}(\theta_{s,h})(x) - \mathscr{G}(\theta)(x)\rangle_V - \frac{|\mathscr{G}(\theta_{s,h})(x)|_V^2 - |\mathscr{G}(\theta)(x)|_V^2}{4}} - 1\Big] - \frac{1}{2}\mathbb{A}_\theta[h]\right]^2 dP_\theta(y,x) \\
&\quad = \int_{V\times\mathcal{X}} \left[\frac{1}{s}\left[e^{f(s)} - 1 - \frac{s}{2}\mathbb{A}_\theta[h]\right]\right]^2 dP_\theta
\end{aligned}$$

where, for y, x fixed,

$$f(s) = \left\langle \frac{y}{2}, \mathscr{G}(\theta_{s,h})(x) - \mathscr{G}(\theta)(x)\right\rangle_V - \frac{|\mathscr{G}(\theta_{s,h})(x)|_V^2 - |\mathscr{G}(\theta)(x)|_V^2}{4}.$$

Clearly, $f(0) = 0$ and, by Condition 1 and the chain rule, we have

$$f'(0) = \left\langle \frac{y}{2}, \mathbb{I}_\theta[h](x)\right\rangle_V - \frac{\langle \mathscr{G}(\theta)(x), \mathbb{I}_\theta[h](x)\rangle_V}{2} = \frac{1}{2}\mathbb{A}_\theta[h](y,x),$$

so that the last integrand converges to zero for every $(y,x) \in V \times \mathcal{X}$, as $s \to 0$. By Condition 1 and the Cauchy–Schwarz inequality, we see that $[e^{f(s)} - 1]/s$ is bounded by a constant multiple of $e^{C|y|_V}$, $C = C(B, U_\mathscr{G}) < \infty$, uniformly in $|s| < \epsilon$. Furthermore, again from Condition 1,

$$\big\|\mathbb{A}_\theta[h]\big\|_{L^2(P_\theta)} \lesssim \big[E|Y|_V + U_\mathscr{G}\big]\|\mathbb{I}_\theta[h]\|_{L^2_\lambda} \lesssim \|h\|_\mathbb{H}$$

and

$$E_\theta\big[e^{C|Y|_V} + |\mathbb{A}_\theta[h](Y,X)|\big]^2 < \infty,$$

so the last limit can be P_θ-integrated by the dominated convergence theorem to give that the last displayed integral converges to zero, verifying the DQM property. The first inequality in the last display also implies that $\mathbb{A}_\theta$ extends to a continuous linear map from $\bar{H}$ to $L^2(P_\theta)$. ∎

2.3. The adjoint score and information operator

The bounded linear operator $\mathbb{A}_\theta : (\bar{H}, \langle\cdot,\cdot\rangle_\mathbb{H}) \to L^2(V \times \mathcal{X}, P_\theta)$ has adjoint operator

$$\mathbb{A}_\theta^* : L^2(V \times \mathcal{X}, P_\theta) \to \big(\bar{H}, \langle\cdot,\cdot\rangle_\mathbb{H}\big)$$

which satisfies

$$\langle w, \mathbb{A}_\theta h\rangle_{L^2(P_\theta)} = \big\langle \mathbb{A}_\theta^* w, h\big\rangle_\mathbb{H}, \quad \text{for all } w \in L^2(V \times \mathcal{X}, P_\theta),\ h \in \bar{H}.$$

The *information operator* is then defined as

$$\mathbb{A}_\theta^* \mathbb{A}_\theta : \bar{H} \to \bar{H}. \tag{2.9}$$

Note that the 'complexity' of the statistical model enters via the choice of 'tangent space' H for which the adjoint is computed, but we suppress this in the notation.

In the present model the information operator can be entirely described in terms of the operator $\mathbb{I}_\theta : (H, \langle\cdot,\cdot\rangle_{\mathbb{H}}) \to L^2_\lambda(\mathcal{X}, V)$ from Condition 1, and its adjoint

$$\mathbb{I}_\theta^* : L^2_\lambda(\mathcal{X}, V) \to \big(\bar{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}}\big).$$

Proposition 1. *Assuming Condition 1, we have*

$$\mathbb{A}_\theta^* \mathbb{A}_\theta[h] = \mathbb{I}_\theta^* \mathbb{I}_\theta[h], \quad \forall h \in H. \tag{2.10}$$

Proof. Writing ϕ for the pdf of an $\mathcal{N}(0, I_V)$ distribution, we have from Fubini's theorem, for any $w \in L^2(P_\theta)$,

$$\begin{aligned}
\langle \mathbb{A}_\theta h, w\rangle_{L^2(P_\theta)} &= \int_V \int_{\mathcal{X}} \big\langle y - \mathcal{G}_\theta(x), \mathbb{I}_\theta(h)(x)\big\rangle_V w(y,x) dP_\theta(y,x) \\
&= \int_{\mathcal{X}} \left\langle \mathbb{I}_\theta(h)(x), \int_V \big(y - \mathcal{G}_\theta(x)\big) w(y,x) \phi\big(y - \mathcal{G}_\theta(x)\big) dy \right\rangle_V d\lambda(x) \\
&= \big\langle \mathbb{I}_\theta(h), E_\theta\big[\big(Y - \mathcal{G}_\theta(X)\big) w(Y,X) | X = \cdot\big]\big\rangle_{L^2_\lambda} \\
&= \big\langle h, \mathbb{I}_\theta^*\big[E_\theta\big[\big(Y - \mathcal{G}_\theta(X)\big) w(Y,X) | X = \cdot\big]\big]\big\rangle_{\mathbb{H}},
\end{aligned}$$

that is, the adjoint $\mathbb{A}_\theta^* = \mathbb{I}_\theta^* \circ \mathcal{E}_\theta$ is the composition of the adjoint $\mathbb{I}_\theta^*$ of $\mathbb{I}_\theta$ with the conditional expectation (projection) operator

$$\mathcal{E}_\theta : L^2(P_\theta) \to L^2_\lambda(\mathcal{X}, V), \quad \mathcal{E}_\theta[w](x) = E_\theta\big[\big(Y - \mathcal{G}_\theta(X)\big) w(Y,X) \,\big|\, X = x\big], \quad x \in \mathcal{X}. \tag{2.11}$$

Now for $h \in H$, we see for $\varepsilon \sim \mathcal{N}(0, I_V)$ and λ-a.e. $x \in \mathcal{X}$,

$$\begin{aligned}
\mathcal{E}_\theta\big[\mathbb{A}_\theta[h]\big](x) &= E_\theta\big[\big(Y - \mathcal{G}_\theta(X)\big)\big\langle Y - \mathcal{G}_\theta(X), \mathbb{I}_\theta h(X)\big\rangle_V \,\big|\, X = x\big] \\
&= E\big[\varepsilon\big\langle \varepsilon, \mathbb{I}_\theta h(x)\big\rangle_V\big] = \mathbb{I}_\theta h(x),
\end{aligned}$$

and therefore $\mathbb{A}_\theta^* \mathbb{A}_\theta[h] = \mathbb{I}_\theta^* \mathcal{E}_\theta[\mathbb{A}_\theta[h]] = \mathbb{I}_\theta^* \mathbb{I}_\theta[h]$, completing the proof. ∎

One can think of $\mathcal{E}_\theta$ in the previous proof as a projection onto the 'space of residuals' of the regression equation (2.2), which vanishes in the representation of the information operator (2.10). In particular, the model (2.2) is LAN (locally asymptotically normal) for LAN-norm $\|\cdot\|_{\mathrm{LAN}}$ arising from LAN inner product

$$\langle h_1, h_2\rangle_{\mathrm{LAN}} := \langle \mathbb{I}_\theta h_1, \mathbb{I}_\theta h_2\rangle_{L^2_\lambda} = \langle \mathbb{A}_\theta h_1, \mathbb{A}_\theta h_2\rangle_{L^2(P_\theta)}, \quad h_1, h_2 \in \bar{H}. \tag{2.12}$$

Proposition 2. *Let $D_N \equiv (Y_i, X_i)_{i=1}^N \sim P_\theta^N$ arise from model (2.2) for some $\theta \in \Theta$ and suppose Condition 1 holds. Then the likelihood ratio process satisfies*

$$\log \frac{dP^N_{\theta + h/\sqrt{N}}}{dP_\theta}(D_N) \to^d_{N\to\infty} \mathcal{N}\left(-\frac{1}{2}\|h\|^2_{\mathrm{LAN}}, \|h\|^2_{\mathrm{LAN}}\right), \quad h \in H.$$

The proof follows from Theorem 1 in conjunction with Lemma 25.14 in [31] (and the central limit theorem). This, in particular, justifies the use of the terminology 'information operator' for $\mathbb{I}_\theta^*\mathbb{I}_\theta$ instead of $\mathbb{A}_\theta^*\mathbb{A}_\theta$.

In what is to follow, the range of the adjoint score operator $\mathbb{A}_\theta^*$ will play a crucial role, and we wish to record a few preparatory remarks here. By what precedes, that range equals

$$R\big(\mathbb{A}_\theta^*\big) = \big\{\psi = \mathbb{I}_\theta^*\mathscr{E}_\theta w, \text{ for some } w \in L^2(P_\theta)\big\}, \tag{2.13}$$

where $\mathscr{E}_\theta$ is from (2.11). Since $\mathscr{E}_\theta$ maps $L^2(P_\theta)$ into L^2_λ, a fortiori any $\psi \in R(\mathbb{A}_\theta^*)$ has to satisfy

$$\psi \in R\big(\mathbb{I}_\theta^*\big) = \big\{\psi = \mathbb{I}_\theta^* h, \text{ for some } h \in L^2_\lambda(\mathcal{X}, V)\big\}, \tag{2.14}$$

so $R(\mathbb{A}_\theta^*) \subset R(\mathbb{I}_\theta^*)$. Likewise, taking $w(y,x) = \langle y - \mathscr{G}(\theta)(x), h(x)\rangle_V \in L^2(P_\theta)$, we can realise (arguing as in the proof of the last proposition) any $h \in L^2_\lambda(\mathcal{X})$ as $\mathscr{E}_\theta w = h$ and so if $\psi \in R(\mathbb{I}_\theta^*)$ then $\psi \in R(\mathbb{A}_\theta^*)$, too. We conclude that

$$R\big(\mathbb{I}_\theta^*\big) = R\big(\mathbb{A}_\theta^*\big). \tag{2.15}$$

2.4. Lower bounds for estimation of functionals

Suppose the problem is to estimate a *linear* functional $\Psi : \Theta \to \mathbb{R}$ of the unknown parameter θ. Let

$$\mathscr{P}_H := \big\{w = \mathbb{A}_\theta(h) : h \in H\big\} \subset L^2(V \times \mathcal{X}, P_\theta)$$

denote the tangent space of the model $\mathscr{P}$ induced by H. Suppose further we can find $\tilde\psi_\theta \in L^2(P_\theta)$ (the 'efficient influence function') such that

$$\Psi(h) = \langle \tilde\psi_\theta, \mathbb{A}_\theta h\rangle_{L^2(P_\theta)}, \quad h \in H. \tag{2.16}$$

If such $\tilde\psi_\theta$ exists, we can always take it to belong to the closure $\overline{\mathscr{P}_H}$ of $\mathscr{P}_H$ in $L^2(P_\theta)$ (simply by $L^2(P_\theta)$-projection onto $\overline{\mathscr{P}_H}$, if necessary). A lower bound for the optimal efficient asymptotic variance for $\sqrt{N}$-consistent estimators of $\Psi(\theta)$ over the model $\{\theta + h/\sqrt{N}, h \in H\}$ is then given by

$$\sup_{0 \neq w \in \mathscr{P}_H} \frac{\langle \tilde\psi_\theta, w\rangle^2_{L^2(P_\theta)}}{\langle w, w\rangle_{L^2(P_\theta)}} = \|\tilde\psi_\theta\|^2_{L^2(P_\theta)}, \tag{2.17}$$

with equality holding in view of $\tilde\psi_\theta \in \overline{\mathscr{P}_H}$ and the Cauchy–Schwarz inequality. Specifically, by Theorem 25.21 in [31], one has

$$\liminf_{N\to\infty} \inf_{\tilde\psi_N : (V\times\mathcal{X})^N \to \mathbb{R}} \sup_{h \in H, \|h\|_{\mathbb{H}} \le 1/\sqrt{N}} N E^N_{\theta+h}\big(\tilde\psi_N - \Psi(\theta+h)\big)^2 \ge \|\tilde\psi_\theta\|^2_{L^2(P_\theta)}. \tag{2.18}$$

If the functional is of the form $\Psi(h) = \langle \psi, h\rangle_{\mathbb{H}}$ for some fixed test function ψ, and if $\mathbb{A}_\theta^*$ is the adjoint of $\mathbb{A}_\theta$ from the previous subsection, the requirement (2.16) can be written as

$$\langle \psi, h\rangle_{\mathbb{H}} = \langle \tilde\psi_\theta, \mathbb{A}_\theta h\rangle_{L^2(P_\theta)} = \big\langle \mathbb{A}_\theta^*\tilde\psi_\theta, h\big\rangle_{\mathbb{H}}, \quad h \in H, \tag{2.19}$$

and hence reduces to $\psi = \mathbb{A}_\theta^*\tilde\psi_\theta$ for some $\tilde\psi_\theta \in L^2(P_\theta)$, that is, $\psi \in R(\mathbb{A}_\theta^*)$ from (2.13).

2.5. Non-existence of $\sqrt{N}$-consistent estimators of linear functionals

Arguing along the traditional lines of the proof of the Cramer–Rao inequality, the inverse of

$$i_{\theta,h,\psi} := \frac{\|\mathbb{A}_\theta h\|^2_{L^2(P_\theta)}}{\langle \psi, h\rangle^2_{\mathbb{H}}} \tag{2.20}$$

provides an a priori lower bound for the variance of any estimator $\widehat{\Psi}$ of $\Psi(\theta) = \langle \psi, \theta\rangle_{\mathbb{H}}$ that is unbiased (i.e. satisfies $E_\theta \widehat{\Psi} = \Psi(\theta)$) for all θ in the one-dimensional model $\{\theta + sh : |s| < \epsilon\}$. The *efficient* Fisher information for estimating Ψ optimally for all elements $h \in H$ of the tangent space is then given by

$$i_{\theta,H,\psi} := \inf_{h\in H, \langle \psi,h\rangle_{\mathbb{H}}\neq 0} \frac{\|\mathbb{A}_\theta h\|^2_{L^2(P_\theta)}}{\langle \psi, h\rangle^2_{\mathbb{H}}}. \tag{2.21}$$

Note that when $\psi = \mathbb{A}^*_\theta \tilde{\psi}_\theta$ is in the range of $\mathbb{A}^*_\theta$ then we can rewrite the last number as

$$\inf_{h\in H, \langle \psi,h\rangle_{\mathbb{H}}\neq 0} \frac{\|\mathbb{A}_\theta h\|^2_{L^2(P_\theta)}}{\langle \mathbb{A}^*_\theta \tilde{\psi}_\theta, h\rangle^2_{\mathbb{H}}} = \inf_{h\in H, \langle \psi,h\rangle_{\mathbb{H}}\neq 0} \frac{\|\mathbb{A}_\theta h\|^2_{L^2(P_\theta)}}{\langle \tilde{\psi}_\theta, \mathbb{A}_\theta h\rangle^2_{L^2(P_\theta)}}. \tag{2.22}$$

Since $\psi \in R(\mathbb{A}^*_\theta)$ is orthogonal on $\ker(\mathbb{A}_\theta)$, using also (2.17), we thus arrive at

$$\|\tilde{\psi}_\theta\|^2_{L^2(P_\theta)} = \sup_{h\in H, \mathbb{A}_\theta h\neq 0} \frac{\langle \tilde{\psi}_\theta, \mathbb{A}_\theta h\rangle^2_{L^2(P_\theta)}}{\langle \mathbb{A}_\theta h, \mathbb{A}_\theta h\rangle_{L^2(P_\theta)}} = i^{-1}_{\theta,H,\psi}, \tag{2.23}$$

explaining the relationship to the best asymptotic variance in (2.18).

An important observation of van der Vaart (Theorem 4.1 in [30]) is that a necessary and sufficient condition for the Fisher information for estimating $\Psi(\theta) = \langle \theta, \psi\rangle_{\mathbb{H}}$ to be non-zero is that ψ indeed lies in the range of $\mathbb{A}^*_\theta$.

Theorem 2. *For $\theta \in \Theta$ and tangent space H, let $i_{\theta,H,\psi}$ be the efficient Fisher information (2.21) for estimating the functional $\Psi(\theta) = \langle \theta, \psi\rangle_{\mathbb{H}}$, $\psi \in \bar{H}$. Then $i_{\theta,H,\psi} > 0$ if and only if $\psi \in R(\mathbb{I}^*_\theta)$.*

If $\psi \in R(\mathbb{I}^*_\theta)$ then positivity $i_{\theta,H,\psi} > 0$ follows directly from (2.15), (2.22) and the Cauchy–Schwarz inequality. The converse is slightly more involved – we include a proof in Section 4.2 below for the case most relevant in inverse problems when the information operator $\mathbb{I}^*_\theta \mathbb{I}_\theta$ from (2.10) is *compact* on $\bar{H}$ (see after Proposition 4 below for the example relevant here).

It follows that if $\psi \notin R(\mathbb{I}^*_\theta)$ then $\Psi(\theta)$ cannot be estimated at $\sqrt{N}$-rate.

Theorem 3. *Consider estimating a functional $\Psi(\theta) = \langle \psi, \theta\rangle_{\mathbb{H}}$, $\psi \in \bar{H}$, based on i.i.d. data $(Y_i, X_i)_{i=1}^N$ in the model (2.2) satisfying Condition 1 for some $\theta \in \Theta$ and tangent space H. Suppose $i_{\theta,H,\psi} = 0$. Then*

$$\liminf_{N\to\infty} \inf_{\tilde{\psi}_N : (V\times\mathcal{X})^N \to \mathbb{R}} \sup_{h\in H, \|h\|_{\mathbb{H}} \le 1/\sqrt{N}} N E^N_{\theta+h}\big(\tilde{\psi}_N - \Psi(\theta + h)\big)^2 = \infty. \tag{2.24}$$

The last theorem can be proved following the asymptotic arguments leading to the proof of (2.18) in Theorem 25.21 in [31]. A proof that follows more directly from the preceding developments is as follows: Augment the observation space to include measurements

$(Z_i, Y_i, X_i)_{i=1}^N \sim \bar{P}_\theta^N$ where the $Z_i \sim^{iid} \mathcal{N}(\langle\theta,\psi\rangle_{\mathbb{H}}, \sigma^2)$ are independent of the (Y_i, X_i)'s, and where σ^2 is known but arbitrary. The new model $\bar{\mathscr{P}}_N = \{\bar{P}_\theta^N : \theta \in \Theta\}$ has 'augmented' LAN norm from (2.12) given by

$$\|\bar{\mathbb{A}}_\theta h\|^2_{L^2(\bar{P}_\theta)} = \|\mathbb{A}_\theta h\|^2_{L^2(P_\theta)} + \sigma^{-2}\langle\psi, h\rangle^2_{\mathbb{H}}, \quad h \in \bar{H},$$

as can be seen from a standard tensorisation argument for independent sample spaces and the fact that a $\mathcal{N}(\langle\theta,\psi\rangle_{\mathbb{H}}, \sigma^2)$ model has LAN 'norm' $\sigma^{-2}\langle\psi, h\rangle^2_{\mathbb{H}}$, by a direct calculation with Gaussian densities. In particular, the efficient Fisher information from (2.21) for estimating $\langle\psi,\theta\rangle_{\mathbb{H}}$ from the augmented data is now of the form

$$\bar{i}_{\theta,H,\psi} = \inf_h \frac{\|\mathbb{A}_\theta h\|^2_{L^2(P_\theta)} + \sigma^{-2}\langle\psi, h\rangle^2_{\mathbb{H}}}{\langle\psi, h\rangle^2_{\mathbb{H}}} = i_{\theta,H,\psi} + \sigma^{-2} = \sigma^{-2} > 0.$$

Note next that *mutatis mutandis* (2.17), (2.18) and (2.23) all hold in the augmented model $\bar{\mathscr{P}}_N$ with score operator $\bar{\mathbb{A}}_\theta$ and tangent space H, and that the linear functional $\Psi(\cdot) = \langle\psi,\cdot\rangle_{\mathbb{H}}$ now verifies (2.16) as it is continuous on H for the $\|\bar{\mathbb{A}}_\theta[\cdot]\|_{L^2(\bar{P}_\theta)}$-norm, so that we can invoke the Riesz representation theorem to the effect that

$$\Psi(h) = \langle\bar{\mathbb{A}}\tilde{h}, \bar{\mathbb{A}}h\rangle_{L^2(\bar{P}_\theta)}, \quad h \in H, \text{ and some } \tilde{\psi}_\theta = \bar{\mathbb{A}}\tilde{h} \in \overline{(\bar{\mathscr{P}})_H}.$$

Thus the asymptotic minimax theorem in the augmented model gives

$$\liminf_{N\to\infty} \inf_{\bar{\psi}_N:(\mathbb{R}\times V\times\mathcal{X})^N\to\mathbb{R}} \sup_{h\in H, \|h\|_{\mathbb{H}}\le 1/\sqrt{N}} NE^N_{\theta+h}\big(\bar{\psi}_N - \Psi(\theta+h)\big)^2 \ge \bar{i}^{-1}_{\theta,H,\psi} = \sigma^2 \tag{2.25}$$

for estimators $\bar{\psi}$ based on the more informative data. The asymptotic local minimax risk in (2.24) exceeds the quantity in the last display, and letting $\sigma^2 \to \infty$ implies the result.

3. APPLICATION TO A DIVERGENCE FORM PDE

The results from the previous section describe how in a non-linear regression model (2.2) under Condition 1, the possibility of $\sqrt{N}$-consistent estimation of linear functionals $\Psi(\theta) = \langle\psi,\theta\rangle_{\mathbb{H}}$ essentially depends on whether ψ lies in the range of $\mathbb{I}^*_\theta$. A sufficient condition for this is that ψ lies in the range of the information operator $\mathbb{A}^*_\theta\mathbb{A}_\theta = \mathbb{I}^*_\theta\mathbb{I}_\theta$, and the results in [20] show that the lower bound in (2.18) can be attained by concrete estimators in this situation. The general theory was shown to apply to a class of PDEs of Schrödinger type [20,21] and to non-linear X-ray transforms [18,20], with smooth test functions $\psi \in C^\infty$.

We now exhibit a PDE inverse problem where the range constraint from Theorem 2 fails, fundamentally limiting the possibility of efficient $\sqrt{N}$-consistent estimation of 'nice' linear functionals. In particular, we will show that, unlike for the Schrödinger type equations considered in [20,21], for this PDE the inverse Fisher information $\sigma^2_\theta(\psi)$ does not exist for a large class of functionals $\Psi(\theta) = \langle\theta,\psi\rangle_{L^2}$, including generic examples of *smooth non-negative* $\psi \in C^\infty$. This implies in particular the non-existence of a 'functional' Bernstein–von Mises phenomenon that would establish asymptotic normality of the posterior distribution of the process $\{\langle\theta,\psi\rangle_{L^2} : \psi \in C^\infty\}$ (comparable to those obtained in [4,5,21]).

3.1. Basic setting

Let $\mathcal{O} \subset \mathbb{R}^d$ be a bounded smooth domain with boundary $\partial\mathcal{O}$ and, for convenience, of unit volume $\lambda(\mathcal{O}) = 1$, where λ is Lebesgue measure. Denote by $C^\infty(\mathcal{O})$ the set of all smooth real-valued functions on $\mathcal{O}$ and by $C_0^\infty(\mathcal{O})$ the subspace of such functions of compact support in $\mathcal{O}$. Let $L^2 = L^2_\lambda(\mathcal{O})$ be the usual Hilbert space with inner product $\langle \cdot, \cdot \rangle_{L^2}$. The L^2_λ-Sobolev spaces $H^\beta = H^\beta(\mathcal{O})$ of order $\beta \in \mathbb{N}$ are also defined in the standard way, as are the spaces $C^\beta(\mathcal{O})$ that have all partial derivatives bounded and continuous up to order β.

For a *conductivity* $\theta \in C^\infty(\mathcal{O})$, *source* $f \in C^\infty(\mathcal{O})$ and *boundary temperatures* $g \in C^\infty(\partial\mathcal{O})$, consider solutions $u = u_\theta = u_{\theta,f,g}$ of the PDE

$$\begin{aligned} \nabla \cdot (\theta \nabla u) &= f \quad \text{in } \mathcal{O}, \\ u &= g \quad \text{on } \partial\mathcal{O}. \end{aligned} \tag{3.1}$$

Here $\nabla, \Delta, \nabla\cdot$ denote the gradient, Laplace and divergence operator, respectively. We ensure ellipticity by assuming $\theta \geq \theta_{\min} > 0$ throughout $\mathcal{O}$.

We write $\mathcal{L}_\theta = \nabla \cdot (\theta \nabla(\cdot))$ for the 'divergence form' operator featuring on the left-hand side in (3.1). A unique solution $u_\theta \in C^\infty(\mathcal{O})$ to (3.1) exists (e.g. Theorem 8.3 and Corollary 8.11 in [10]). The operator $\mathcal{L}_\theta$ has an inverse integral operator

$$V_\theta : L^2_\lambda(\mathcal{O}) \to H^2(\mathcal{O}) \cap \{h_{|\partial\mathcal{O}} = 0\} \tag{3.2}$$

for Dirichlet boundary conditions, that is, it satisfies $V_\theta[f] = 0$ at $\partial\mathcal{O}$ and $\mathcal{L}_\theta V_\theta[f] = f$ on $\mathcal{O}$ for all $f \in L^2_\lambda(\mathcal{O})$. Moreover, the operator V_θ is self-adjoint on $L^2_\lambda(\mathcal{O})$. One further shows that whenever $f \in H^2(\mathcal{O})$ satisfies $f_{|\partial\mathcal{O}} = 0$, then $V_\theta \mathcal{L}_\theta[f] = f$. These standard facts for elliptic PDEs can be proved, e.g. as in Section 5.1 in [29] or Chapter 2 in [17].

To define the 'forward map' $\mathcal{G}$ we consider a model Θ of conductivities arising as a H^β-neighbourhood of the standard Laplacian of radius $\eta > 0$, specifically

$$\Theta = \left\{ \theta \in C^\infty(\mathcal{O}), \inf_x \theta(x) > \frac{1}{2}, \theta_{|\partial\mathcal{O}} = 1 : \|\theta - 1\|_{H^\beta(\mathcal{O})} < \eta \right\}, \quad \beta > 1 + d. \tag{3.3}$$

The inverse problem is to recover θ from solutions

$$\mathcal{G} : \Theta \to L^2_\lambda(\mathcal{O}), \quad \mathcal{G}(\theta) \equiv u_\theta \tag{3.4}$$

of (3.1) where we emphasise that f, g, as well as $\theta_{|\partial\mathcal{O}}$, are assumed to be *known* (see also Remark 3). The particular numerical choices $1 = \theta_{|\partial\mathcal{O}}$ and $1/2 = \theta_{\min}$ are made for notational convenience. For independent $\varepsilon_i \sim^{iid} \mathcal{N}(0,1)$, $X_i \sim^{iid} \lambda$, we then observe data

$$(Y_i, X_i)_{i=1}^N \in (\mathbb{R} \times \mathcal{O})^N \sim P_\theta^N, \quad Y_i = u_\theta(X_i) + \varepsilon_i, \tag{3.5}$$

from model (2.2). Note that unlike in statistical 'Calderón problems' [1], we measure u_θ throughout the entire domain $\mathcal{O}$. Before we take a closer look at the *local* information geometry of the map $\mathcal{G}$ arising from the PDE (3.1), let us first give conditions under which the problem of inferring θ from $(Y_i, X_i)_{i=1}^N$ in (3.5) has a consistent solution.

3.2. Global injectivity and model examples

Under suitable constellations of f, g in (3.1), the non-linear map $\theta \mapsto u_\theta$ can be injective, and 'stability' properties of $\mathscr{G}$ are well studied at least since [27], we refer to the recent contributions [2,13,24] and the many references therein. For instance, one can show:

Proposition 3. *Let $\theta_1, \theta_2 \in C^\infty(\mathcal{O})$ be conductivities such that $\|\theta_i\|_{C^1} \leq B$, $\theta_1 = \theta_2$ on $\partial\mathcal{O}$, and denote by u_{θ_i} the corresponding solutions to (3.1). Assume*

$$\inf_{x\in\mathcal{O}}\left[\Delta u_\theta(x) + \mu\left|\nabla u_\theta(x)\right|^2_{\mathbb{R}^d}\right] \geq c_0 > 0 \tag{3.6}$$

holds for $\theta = \theta_1$ and some $\mu > 0$. Then we have for some $C = C(B, \mu, c_0, \mathcal{O}) > 0$,

$$\|\theta_1 - \theta_2\|_{L^2} \leq C\|u_{\theta_1} - u_{\theta_2}\|_{H^2}. \tag{3.7}$$

Based on (3.7), one can show (see [13,24]) that we can recover θ in L^2-loss by some estimator $\hat\theta = \hat\theta((Y_i, X_i)_{i=1}^N)$ at a 'non-parametric rate' $\|\hat\theta - \theta\|_{L^2(\mathcal{O})} = O_{P_\theta^N}(N^{-\gamma})$ for some $0 < \gamma < 1/2$, uniformly in Θ. We wish to study here inference on linear functionals

$$\Psi(\theta) = \langle\psi, \theta\rangle_{L^2(\mathcal{O})}, \quad \psi \in C_0^\infty(\mathcal{O}).$$

As we can bound the 'plug-in' estimation error $|\langle\psi, \theta - \hat\theta\rangle_{L^2}|$ by $\|\hat\theta - \theta\|_{L^2}$, the convergence rate $N^{-\gamma}$ carries over to estimation of Ψ. Nevertheless, we will show that there are fundamental limitations for *efficient* inference on Ψ at the 'semi-parametric' rate ($\gamma = 1/2$). This will be illustrated with two model examples for which the 'injectivity' condition (3.6) can be checked.

Example 1 (No critical points). In (3.1), take

$$f = 2, \quad g = \frac{|\cdot|^2_{\mathbb{R}^d} - 1}{d}. \tag{3.8}$$

Then for the standard Laplacian $\theta = 1$, we have $u_1 = g$ on $\bar{\mathcal{O}}$, $\Delta u_1 = 2$, and hence $\nabla u_1 = 2x/d$, which satisfies $\inf_{x\in\mathcal{O}} |\nabla u_1(x)|_{\mathbb{R}^d} \geq c > 0$ for any domain $\mathcal{O} \subset \mathbb{R}^d$ *separated away from the origin*. This lower bound extends to

$$\inf_{\theta\in\Theta}\inf_{x\in\mathcal{O}} \left|\nabla u_\theta(x)\right|_{\mathbb{R}^d} \geq c_\nabla > 0 \tag{3.9}$$

for η small enough in (3.3), by perturbation: arguing as in (3.16) below and from standard elliptic regularity estimates (Lemma 23 in [24] and as in (3.15)), we have for $b > 1 + d/2$, $\beta > b + d/2$ (such that $H^\beta \subset C^b$),

$$\begin{aligned}\|u_\theta - u_1\|_{C^1} &\lesssim \left\|V_1\left[\nabla\cdot[(\theta-1)\nabla u_\theta]\right]\right\|_{H^b} \lesssim \left\|(\theta-1)\nabla u_\theta\right\|_{H^{b-1}} \\ &\lesssim \|\theta - 1\|_{H^{b-1}}\|u_\theta\|_{C^b} \leq \|\theta-1\|_{H^\beta}\|u_\theta\|_{H^\beta} < C\eta.\end{aligned} \tag{3.10}$$

In view of $\sup_{\theta\in\Theta}\|\Delta u_\theta\|_\infty < \infty$ and (3.9), condition (3.6) is verified for μ large enough and all $\theta \in \Theta$.

The situation in Example 1 where the gradient ∇u_θ never vanishes is somewhat atypical, and one may expect u_θ to possess a *finite* number of isolated critical points x_0

(where $\nabla u_\theta(x_0)$ vanishes); see, e.g. [2] and references therein. The next example encompasses a prototypical such situation with an interior minimum. See also Remark 1 for the case of a saddle point. Further examples with more than one critical point are easily constructed, too.

Example 2 (Interior minimum). Consider the previous example where now $\mathcal{O}$ is the unit disk in $\mathbb{R}^2$ centred at the origin. In other words, in (3.1) we have $f = 2$ and $g_{|\partial\mathcal{O}} = 0$, corresponding to a classical Dirichlet problem with source f. In this case u_1 takes the same form as in the previous example but now has a gradient $\nabla u_1 = x$ that vanishes at the origin $0 \in \mathbb{R}^2$, corresponding to the unique minimum of u_1 on $\mathcal{O}$. The injectivity condition (3.6) is still satisfied for all $\theta \in \Theta$ simply since (3.1) implies

$$0 < 2 = \theta \Delta u_\theta + \nabla\theta \cdot \nabla u_\theta \quad \text{on } \mathcal{O},$$

so that either $\Delta u_\theta \geq 1/(2\|\theta\|_\infty)$ or $|\nabla u_\theta(x)|_{\mathbb{R}^d} \geq 1/(2\|\theta\|_{C^1})$ has to hold on $\mathcal{O}$. In this example, the constraints that η be small enough as well as that $\theta_1 = \theta_2$ on $\partial\mathcal{O}$ in Proposition 3 can in fact be removed, see Lemma 24 in [24].

3.3. The score operator and its adjoint

To connect to Section 2, let us regard Θ from (3.3) as a subset of the Hilbert space $\mathbb{H} = L^2_\lambda(\mathcal{O})$, and take $\mathcal{G}(\theta)$ from (3.4); hence we set $\mathcal{X} = \mathcal{O}$, $V = \mathbb{R}$, $\lambda = dx$ (Lebesgue measure).

As 'tangent space' $H \subset \mathbb{H}$, we take all smooth perturbations of θ of compact support,

$$H = C_0^\infty(\mathcal{O}), \tag{3.11}$$

so that the paths $\theta_{s,h} = \theta + sh$, $\theta \in \Theta$, $h \in H$, lie in Θ for all $s \in \mathbb{R}$ small enough. The closure $\bar{H}$ of H for $\|\cdot\|_{\mathbb{H}}$ equals $\bar{H} = \mathbb{H} = L^2_\lambda(\mathcal{O})$. We now check Condition 1, restricting to $d \leq 3$ to expedite the proof.

Theorem 4. *Assume $d \leq 3$. Let Θ be as in (3.3) and let the tangent space H be as in (3.11). The forward map $\theta \mapsto \mathcal{G}(\theta)$ from (3.4) satisfies Condition 1 for every $\theta \in \Theta$, with uniform bound $U_{\mathcal{G}} = U_{\mathcal{G}}(\|g\|_\infty, \|f\|_\infty)$ and with*

$$\mathbb{I}_\theta(h) \equiv -V_\theta\big[\nabla \cdot (h \nabla u_\theta)\big], \quad h \in H. \tag{3.12}$$

In particular, $\mathbb{I}_\theta$ extends to a bounded linear operator on $\mathbb{H}$.

Proof. We can represent the solutions u_θ of (3.1) by a Feynman–Kac-type formula as

$$u_\theta(x) = \mathbb{E}^x g(X_{\tau_\mathcal{O}}) - \mathbb{E}^x \int_0^{\tau_\mathcal{O}} f(X_s) ds, \quad x \in \mathcal{O}, \tag{3.13}$$

where $(X_s : s \geq 0)$ is a Markov diffusion process started at $x \in \mathcal{O}$ with infinitesimal generator $\mathcal{L}_\theta/2$, law $\mathbb{P}^x = \mathbb{P}^x_\theta$, and exit time $\tau_\mathcal{O}$ from $\mathcal{O}$, see Theorem 2.1 on p. 127 in [8]. As in the proof of Lemma 20 in [24], one bounds $\sup_{x\in\mathcal{O}} \mathbb{E}^x \tau_\mathcal{O}$ by a constant that depends only

on $\mathcal{O}$, $\theta_{\min}$, and we conclude from the last display that therefore

$$\|u_\theta\|_\infty \le \|g\|_\infty + \|f\|_\infty \sup_{x\in\mathcal{O}} \mathbb{E}^x \tau_{\mathcal{O}} < \infty \tag{3.14}$$

so that the bound $U_{\mathscr{G}}$ for $\mathscr{G}$ required in Condition 1 follows.

We will repeatedly use the following elliptic regularity estimates:

$$\big\|V_\theta[h]\big\|_\infty \le c_0 \big\|V_\theta[h]\big\|_{H^2} \le c_1 \|h\|_{L^2}, \quad \|u_\theta\|_{H^2} \le c_2, \tag{3.15}$$

with constants $c_0 = c_0(\mathcal{O}), c_1 = c_1(\theta_{\min}, \mathcal{O}, \beta, \eta), c_2 = c_2(U_{\mathscr{G}}, \|f\|_{L^2}, \|g\|_{H^2}, \theta_{\min}, \mathcal{O}, \beta, \eta)$ that are *uniform* in $\theta \in \Theta$. The first inequality in (3.15) is just the Sobolev imbedding. The second follows from Lemma 21 in [24], noting also that $\sup_{\theta\in\Theta} \|\theta\|_{C^1} \le C(\beta, \eta, \mathcal{O})$ by another Sobolev imbedding $H^\beta \subset C^1$. The final inequality in (3.15) follows from Theorem 8.12 in [10] and (3.14).

To verify (2.4), notice that the difference $u_{\theta+sh} - u_\theta$ solves (3.1) with $g = 0$ and appropriate right-hand side, specifically we can write

$$\mathscr{G}(\theta + sh) - \mathscr{G}(\theta) = -sV_\theta\big[\nabla \cdot (h\nabla u_{\theta+sh})\big], \quad h \in H, \tag{3.16}$$

for $|s|$ small enough. Then (2.4) follows from (3.15) since

$$\begin{aligned}\big\|V_\theta\big[\nabla \cdot (h\nabla u_{\theta+sh})\big]\big\|_\infty &\lesssim \big\|\nabla \cdot (h\nabla u_{\theta+sh})\big\|_{L^2} \lesssim \|h\nabla u_{\theta+sh}\|_{H^1} \\ &\lesssim \|h\|_{C^1} \sup_{\theta\in\Theta} \|u_\theta\|_{H^2} \le B < \infty.\end{aligned}$$

We will verify (2.3) by establishing a stronger '$\|\cdot\|_\infty$-norm' differentiability result: fix $\theta \in \Theta$ and any $h \in H$ such that $\theta + h \in \Theta$. Denote by $D\mathscr{G}_\theta[h]$ the solution $v = v_h$ of the PDE

$$\begin{aligned}\nabla \cdot (\theta\nabla v) &= -\nabla \cdot (h\nabla u_\theta) \quad \text{on } \mathcal{O}, \\ v &= 0 \quad \text{on } \partial\mathcal{O}\end{aligned}$$

where u_θ is the given solution of the original PDE (3.1). Then the function $w_h = u_{\theta+h} - u_\theta - D\mathscr{G}_\theta[h]$ solves the PDE

$$\begin{aligned}\mathscr{L}_{\theta+h} w_h &= -\nabla \cdot \big(h\nabla V_\theta[\nabla \cdot (h\nabla u_\theta)]\big) \quad \text{on } \mathcal{O}, \\ w_h &= 0 \quad \text{on } \partial\mathcal{O}.\end{aligned}$$

As a consequence, applying (3.15) and standard inequalities repeatedly, we have

$$\begin{aligned}\big\|u_{\theta+h} - u_\theta - D\mathscr{G}_\theta[h]\big\|_\infty &= \big\|V_{\theta+h}\big[\nabla \cdot \big(h\nabla V_\theta[\nabla \cdot (h\nabla u_\theta)]\big)\big]\big\|_\infty \\ &\lesssim \big\|\nabla \cdot \big(h\nabla V_\theta[\nabla \cdot (h\nabla u_\theta)]\big)\big\|_{L^2} \\ &\lesssim \|h\|_{C^1} \big\|V_\theta\big[\nabla \cdot (h\nabla u_\theta)\big]\big\|_{H^2} \\ &\lesssim \|h\|_{C^1} \big\|\nabla \cdot (h\nabla u_\theta)\big\|_{L^2} \\ &\lesssim \|h\|_{C^1}^2 \|u_\theta\|_{H^2} = O\big(\|h\|_{C^1}^2\big).\end{aligned} \tag{3.17}$$

In particular $D\mathcal{G}_\theta[sh] = \mathbb{I}_\theta[sh]$ is the linearisation of the forward map $\theta \mapsto \mathcal{G}(\theta) = u_\theta$ along any path $\theta + sh$, $|s| > 0$, $h \in H$. Finally, by duality, self-adjointness of V_θ and the divergence theorem (Proposition 2.3 on p. 143 in [29]), we can bound for every $h \in H$,

$$\begin{aligned}\|\mathbb{I}_\theta h\|_{L^2} &= \sup_{\|\phi\|_{L^2}\le 1} \left| \int_{\mathcal{O}} \phi V_\theta\big[\nabla\cdot(h\nabla u_\theta)\big] \right| = \sup_{\|\phi\|_{L^2}\le 1} \left| \int_{\mathcal{O}} \nabla V_\theta[\phi]\cdot h\nabla u_\theta] \right| \\ &\lesssim \sup_{\|\phi\|_{L^2}\le 1} \big\|V_\theta[\phi]\big\|_{H^1} \|h\|_{L^2}\|u_\theta\|_{C^1} \lesssim \|h\|_{L^2},\end{aligned}$$

using also (3.15) and that $\|u_\theta\|_{C^1} < \infty$ (here for fixed θ) as u_θ is smooth. By continuity and since H is dense in $L^2_\lambda = \mathbb{H}$, we can extend $\mathbb{I}_\theta$ to a bounded linear operator on $\mathbb{H}$, completing the proof. ∎

Theorem 1 gives the score operator $\mathbb{A}_\theta$ mapping H into $L^2(\mathbb{R}\times\mathcal{O}, P_\theta)$ of the form

$$\mathbb{A}_\theta[h](x,y) = \big(y - u_\theta(x)\big)\times \mathbb{I}_\theta(h)(x), \quad y\in\mathbb{R},\ x\in\mathcal{O}. \tag{3.18}$$

For the present tangent space H, we have $\bar{H} = \mathbb{H}$. To apply the general results from Section 2, we now calculate the adjoint $\mathbb{I}_\theta^* : L^2_\lambda(\mathcal{O}) \to \bar{H} = L^2_\lambda(\mathcal{O})$ of $\mathbb{I}_\theta : \bar{H} \to L^2(\mathcal{O})$.

Proposition 4. *The adjoint $\mathbb{I}_\theta^* : L^2_\lambda(\mathcal{O}) \to L^2_\lambda(\mathcal{O})$ of $\mathbb{I}_\theta$ is given by*

$$\mathbb{I}_\theta^*[g] = \nabla u_\theta \cdot \nabla V_\theta[g], \quad g \in L^2_\lambda(\mathcal{O}). \tag{3.19}$$

Proof. Since $\mathbb{I}_\theta$ from (3.12) defines a bounded linear operator on the Hilbert space $L^2_\lambda = \mathbb{H}$, a unique adjoint operator I_θ^* exists by the Riesz representation theorem. Let us first show that

$$\big\langle h, \big(I_\theta^* - \mathbb{I}_\theta^*\big)g\big\rangle_{L^2} = 0, \quad \forall h, g \in C_0^\infty(\mathcal{O}). \tag{3.20}$$

Indeed, since V_θ is self-adjoint for L^2_λ and satisfies $[V_\theta g]_{|\partial\mathcal{O}} = 0$, we can apply the divergence theorem (Proposition 2.3 on p. 143 in [29]) with vector field $X = h\nabla u_\theta$ to deduce

$$\begin{aligned}\big\langle h, I_\theta^* g\big\rangle_{L^2(\mathcal{O})} &= \langle \mathbb{I}_\theta h, g\rangle_{L^2(\mathcal{O})} = -\big\langle V_\theta\big[\nabla\cdot(h\nabla u_\theta)\big], g\big\rangle_{L^2(\mathcal{O})} \\ &= -\int_{\mathcal{O}} \big[\nabla\cdot(h\nabla u_\theta)\big] V_\theta[g] d\lambda \\ &= \int_{\mathcal{O}} h\nabla u_\theta\cdot\nabla V_\theta[g] d\lambda = \big\langle h, \mathbb{I}_\theta^* g\big\rangle_{L^2(\mathcal{O})},\end{aligned}$$

so that (3.20) follows. Since $C_0^\infty(\mathcal{O})$ is dense in $L^2_\lambda(\mathcal{O})$ and since I_θ^*, $\mathbb{I}_\theta^*$ are continuous on $L^2_\lambda(\mathcal{O})$ (by construction in the former case and by (3.15), $u_\theta \in C^\infty(\mathcal{O})$, in the latter case), the identity (3.20) extends to all $g \in L^2_\lambda(\mathcal{O})$ and hence $I_\theta^* = \mathbb{I}_\theta^*$, as desired. ∎

Note further that for $\theta\in\Theta$ fixed, using (3.15), $u_\theta \in C^\infty$ and L^2-continuity of $\mathbb{I}_\theta$, we have $\|\mathbb{I}_\theta^*\mathbb{I}_\theta h\|_{H^1} \lesssim \|\mathbb{I}_\theta h\|_{L^2} \lesssim \|h\|_{L^2}$. The compactness of the embedding $H^1 \subset L^2$ now implies that the information operator $\mathbb{I}_\theta^*\mathbb{I}_\theta$ is a compact and self-adjoint operator on $L^2(\mathcal{O})$.

3.4. Injectivity of $\mathbb{I}_\theta$, $\mathbb{I}_\theta^*\mathbb{I}_\theta$

Following the developments in Section 2, our ultimate goal is to understand the range $R(\mathbb{I}_\theta^*)$ of the adjoint operator $\mathbb{I}_\theta^*$. A standard Hilbert space duality argument implies that

$$R\big(\mathbb{I}_\theta^*\big)^\perp = \ker(\mathbb{I}_\theta), \tag{3.21}$$

that is, the ortho-complement (in $\mathbb{H}$) of the range of $\mathbb{I}_\theta^*$ equals the kernel (null space) of $\mathbb{I}_\theta$ (in $\mathbb{H}$). Thus if ψ is in the kernel of $\mathbb{I}_\theta$ then it cannot lie in the range of the adjoint and the non-existence of the inverse Fisher information in Theorem 2 for such ψ can be attributed simply to the lack of injectivity of $\mathbb{I}_\theta$.

We first show that under the natural 'global identification' condition (3.6), the mapping $\mathbb{I}_\theta$ from (3.12) is injective on the tangent space H (and hence on our parameter space Θ). The proof (which is postponed to Section 4.1) also implies injectivity of the information operator $\mathbb{I}_\theta^*\mathbb{I}_\theta$ on H, and in fact gives an $H^2 - L^2$ Lipschitz stability estimate for $\mathbb{I}_\theta$.

Theorem 5. *In the setting of Theorem 4, suppose also that (3.6) holds true. Then for $\mathbb{I}_\theta$ from (3.12), every $\theta \in \Theta$ and some $c = c(\mu, c_0, \theta, \mathcal{O})$,*

$$\big\|\mathbb{I}_\theta[h]\big\|_{H^2} \geq c\|h\|_{L^2} \quad \forall h \in H. \tag{3.22}$$

In particular, $\mathbb{I}_\theta(h) = 0$ or $\mathbb{I}_\theta^*\mathbb{I}_\theta(h) = 0$ imply $h = 0$ for all $h \in H$.

Using (3.15), one shows further that the operator $\mathbb{I}_\theta$ is continuous from $H^1(\mathcal{O}) \to H^2(\mathcal{O})$ and, by taking limits in (3.22), Theorem 5 then extends to all $h \in H_0^1(\mathcal{O})$ obtained as the completion of H for the $H^1(\mathcal{O})$-Sobolev norm.

Of course, the kernel in (3.21) is calculated on the Hilbert space $\mathbb{H} = L^2(\mathcal{O})$, so the previous theorem does not characterise $R(\mathbb{I}_\theta^*)^\perp$, yet. Whether $\mathbb{I}_\theta$ is injective on all of $L^2(\mathcal{O})$ depends on finer details of the PDE (3.1). Let us illustrate this in the model examples from above.

3.4.1. Example 1 continued; on the kernel in $L^2(\mathcal{O})$

In our first example, $\mathbb{I}_\theta$ starts to have a kernel already when $h_{|\partial\mathcal{O}} \neq 0$. Indeed, from the proof of Theorem 5, a function $h \in C^\infty(\bar{\mathcal{O}})$ is in the kernel of $\mathbb{I}_\theta$ if and only if

$$T_\theta(h) = \nabla \cdot (h\nabla u_\theta) = \nabla h \cdot \nabla u_\theta + h\Delta u_\theta = 0. \tag{3.23}$$

Now fix any $\theta \in \Theta$ with u_θ satisfying (3.9). The integral curves $\gamma(t)$ in $\mathcal{O}$ associated to the smooth vector field $\nabla u_\theta \neq 0$ are given near $x \in \mathcal{O}$ as the unique solutions (e.g. [29, P. 9]) of the vector ODE

$$\frac{d\gamma}{dt} = \nabla u_\theta(\gamma), \quad \gamma(0) = x. \tag{3.24}$$

Since ∇u_θ does not vanish, we obtain through each $x \in \mathcal{O}$ a unique curve $(\gamma(t) : 0 \leq t \leq T_\gamma)$ originating and terminating at the boundary $\partial\mathcal{O}$, with finite 'travel time' $T_\gamma \leq T(\mathcal{O}, c_\nabla) < \infty$. Along this curve, (3.23) becomes the ODE

$$\frac{d}{dt}h\big(\gamma(t)\big) + h\big(\gamma(t)\big)\Delta u_\theta\big(\gamma(t)\big) = 0, \quad 0 < t < T_\gamma.$$

Under the constraint $h|_{\partial\mathcal{O}} = 0$ for $h \in H$, the unique solution of this ODE is $h = 0$, which is in line with Theorem 5. But for other boundary values of h, non-zero solutions exist. One can characterise the elements $h \in C^\infty(\bar{\mathcal{O}})$ in the kernel of $\mathbb{I}_\theta$ as follows. Since the vector field ∇u_θ is non-trapping, there exists (see [7, THEOREM 6.4.1]) $r \in C^\infty(\bar{\mathcal{O}})$ such that $\nabla u_\theta \cdot \nabla r = \Delta u_\theta$. Thus

$$\nabla u_\theta \cdot \nabla(he^r) = e^r T_\theta(h)$$

and it follows that $T_\theta(h) = 0$ iff he^r is a first integral of ∇u_θ. Observe that the set of first integrals of ∇u_θ is rather large: using the flow of ∇u_θ, we can pick coordinates $(x_1, \dots, x_d)$ in $\mathcal{O}$ such that $t \mapsto (t + x_1, x_2, \dots, x_d)$ are the integral curves of ∇u_θ and thus any function that depends only on $x_2, \dots, x_d$ is a first integral.

3.4.2. Example 2 continued; injectivity on $L^2(\mathcal{O})$

We now show that in the context of Example 2, the injectivity part of Theorem 5 *does* extend to all of $L^2(\mathcal{O})$.

Proposition 5. *Let $\mathbb{I}_\theta$ be as in (3.12) where u_θ solves (3.1) with f, g as in (3.8) and $\mathcal{O}$ is the unit disk in $\mathbb{R}^2$ centred at $(0,0)$. Then for $\theta = 1$, the map $\mathbb{I}_1 : L^2(\mathcal{O}) \to L^2(\mathcal{O})$ is injective.*

Proof. Let us write $I = \mathbb{I}_1$ and suppose $I(f) = 0$ for $f \in L^2(\mathcal{O})$. Then for any $h \in C^\infty(\mathcal{O})$ we have by Proposition 4

$$0 = \langle If, h\rangle_{L^2(\mathcal{O})} = \langle f, I^* h\rangle_{L^2(\mathcal{O})} = \big\langle f, XV_1[h]\big\rangle_{L^2(\mathcal{O})} \tag{3.25}$$

with vector field $X = \nabla u_1 \cdot \nabla(\cdot) = x_1\partial x_1 + x_2 \partial x_2$, $(x_1, x_2) \in \mathcal{O}$. Choosing $h = \Delta g$ for any smooth g of compact support, we deduce that

$$\int_{\mathcal{O}} X(g) f \, d\lambda = 0, \quad \forall g \in C_0^\infty(\mathcal{O}), \tag{3.26}$$

and we now show that this implies $f = 0$. A somewhat informal dynamical argument would say that (3.26) asserts that $f d\lambda$ is an invariant density under the flow of X. Since the flow of X in backward time has a sink at the origin, the density can only be supported at $(x_1, x_2) = 0$ and thus $f = 0$.

One can give a distributional argument as follows. Suppose we consider polar coordinates $(r, \vartheta) \in (0,1) \times S^1$ and functions g of the form $\phi(r)\psi(\vartheta)$, where $\phi \in C_0^\infty(0,1)$ and $\psi \in C^\infty(S^1)$. In polar coordinates $X = r\partial_r$, and hence we may write (3.26) as

$$\int_0^1 \left(r^2 \left(\int_0^{2\pi} f(r,\vartheta)\psi(\vartheta)\, d\vartheta \right) \partial_r \phi \right) dr = 0. \tag{3.27}$$

By Fubini's theorem, for each ψ we have an integrable function

$$F_\psi(r) := \int_0^{2\pi} f(r,\vartheta)\psi(\vartheta)\, d\vartheta$$

and thus $r^2 F_\psi$ defines an integrable function on $(0,1)$ whose distributional derivative satisfies $\partial_r(r^2 F_\psi) = 0$ by virtue of (3.27). Thus $r^2 F_\psi = c_\psi$ (using that a distribution on $(0,1)$ with zero derivative must be a constant). Now consider $\psi \in C^\infty(S^1)$ also as a function in $L^2(\mathcal{O})$ and compute the pairing

$$(f, \psi)_{L^2(\mathcal{O})} = \int_0^1 r F_\psi(r)\, dr = c_\psi \int_0^1 r^{-1}\, dr = \pm\infty$$

unless $c_\psi = 0$. Thus $f = 0$. ∎

By perturbation (similar as in (3.10)) and the Morse lemma, we can show that $u_\theta, \theta \in \Theta$, has a gradient u_θ that vanishes only at a single point in a neighbourhood of 0, and so the proof of the previous theorem extends to any $\theta \in \Theta$.

3.5. The range of $\mathbb{I}_\theta^*$ and transport PDEs

From (3.21) we see $\overline{R(\mathbb{I}_\theta^*)} = \ker(\mathbb{I}_\theta)^\perp$, but in our infinite-dimensional setting care needs to be exercised as the last identity holds in the (complete) Hilbert space $\mathbb{H} = L^2(\mathcal{O})$ rather than in our tangent space H (on which the kernel of $\mathbb{I}_\theta$ is trivial). We will now show that the range $R(\mathbb{I}_\theta^*)$ remains strongly constrained. This is also true in Example 2 when $\ker(\mathbb{I}_\theta) = \{0\}$: the range may not be closed $\overline{R(\mathbb{I}_\theta^*)} \neq R(\mathbb{I}_\theta^*)$, and this 'gap' can be essential in the context of Theorems 2 and 3. To understand this, note that from Proposition 4 we have

$$R\big(\mathbb{I}_\theta^*\big) = \big\{\psi = \nabla u_\theta \cdot \nabla V_\theta[g], \text{ for some } g \in L^2_\lambda(\mathcal{O})\big\}. \tag{3.28}$$

The operator V_θ maps L^2_λ into $H^2_0 = \{y \in H^2 : y_{|\partial\mathcal{O}} = 0\}$ and hence if ψ is in the range of $\mathbb{I}_\theta^*$ then the equation

$$\begin{aligned} \nabla u_\theta \cdot \nabla y &= \psi \quad \text{on } \mathcal{O}, \\ y &= 0 \quad \text{on } \partial\mathcal{O} \end{aligned} \tag{3.29}$$

necessarily has a solution $y = y_\psi \in H^2_0$. The existence of solutions to the transport PDE (3.29) depends crucially on the compatibility of ψ with geometric properties of the vector field ∇u_θ, which in turn is determined by the geometry of the forward map $\mathcal{G}$ (via f, g, θ) in the base PDE (3.1). We now illustrate this in our two model Examples 1 and 2.

3.5.1. Example 1 continued; range constraint

Applying the chain rule to $y \in H^2(\mathcal{O})$ and using (3.24), we see

$$\frac{d}{dt} y\big(\gamma(t)\big) = \frac{d\gamma(t)}{dt} \cdot \nabla y\big(\gamma(t)\big) = (\nabla u_\theta \cdot \nabla y)\big(\gamma(t)\big), \quad 0 < t < T_\gamma.$$

Hence along any integral curve γ of the vector field ∇u_θ, the PDE (3.29) reduces to the ODE

$$\frac{dy}{dt} = \psi. \tag{3.30}$$

Now suppose $\psi \in R(\mathbb{I}_\theta^*)$. Then a solution $y \in H^2_0$ to (3.29) satisfying $y_{|\partial\mathcal{O}} = 0$ must exist. Such y then also solves the ODE (3.30) along each curve γ, with initial and terminal values $y(0) = y(T_\gamma) = 0$. By the fundamental theorem of calculus (and uniqueness of solutions), this forces

$$\int_0^{T_\gamma} \psi\big(\gamma(t)\big)\, dt = 0 \tag{3.31}$$

to vanish. In other words, ψ permits a solution y to (3.29) only if ψ integrates to zero along each integral curve (orbit) induced by the vector field ∇u_θ. Now consider any smooth (non-zero) *nonnegative* ψ in the tangent space $H = C^\infty_0(\mathcal{O})$, and take $x \in \mathcal{O}$ such that $\psi \geq c > 0$ near x. For γ the integral curve passing through x, we then cannot have (3.31) as the integrand never takes negative values while it is positive and continuous near x. Conclude by way of contradiction that $\psi \notin R(\mathbb{I}_\theta^*)$. Applying Theorems 2 and 3, we have proved:

Theorem 6. *Consider estimation of the functional $\Psi(\theta) = \langle \theta, \psi \rangle_{L^2(\mathcal{O})}$ from data $(Y_i, X_i)_{i=1}^N$ drawn i.i.d. from P_θ^N in the model (3.5) where f, g in (3.1) are chosen as in (3.8), the domain $\mathcal{O}$ is separated away from the origin, and Θ is as in (3.3) with η small enough and $\beta > 1 + d$,*

$d \leq 3$. *Suppose* $0 \neq \psi \in C_0^\infty(\mathcal{O})$ *satisfies* $\psi \geq 0$ *on* $\mathcal{O}$. *Then for every* $\theta \in \Theta$, *the efficient Fisher information for estimating* $\Psi(\theta)$ *satisfies*

$$\inf_{h \in H, \langle h, \psi \rangle_{L^2} \neq 0} \frac{\|\mathbb{I}_\theta h\|_{L^2_\lambda}^2}{\langle \psi, h \rangle_{L^2_\lambda}^2} = 0. \tag{3.32}$$

In particular, for any $\theta \in \Theta$,

$$\liminf_{N \to \infty} \inf_{\tilde{\psi}_N : (\mathbb{R} \times \mathcal{O})^N \to \mathbb{R}} \sup_{\theta' \in \Theta, \|\theta' - \theta\|_{\mathbb{H}} \leq 1/\sqrt{N}} N E_{\theta'}^N \left(\tilde{\psi}_N - \Psi(\theta') \right)^2 = \infty. \tag{3.33}$$

Let us notice that one can further show that (3.31) is also a *sufficient* condition for ψ to lie in the range of $\mathbb{I}_\theta^*$ (provided ψ is smooth and with compact support in $\mathcal{O}$). As this condition strongly depends on θ via the vector field ∇u_θ, it seems difficult to describe any choices of ψ that lie in $\bigcap_{\theta \in \Theta} R(\mathbb{I}_\theta^*)$.

3.5.2. Example 2 continued; range constraint

We showed in the setting of Example 2 that $\mathbb{I}_\theta$ is injective on all of $L^2(\mathcal{O})$, and hence any $\psi \in L^2(\mathcal{O})$ lies in *closure* of the range of $\mathbb{I}_\theta^*$. Nevertheless, there are many relevant ψ's that are not contained in $R(\mathbb{I}_\theta^*)$. In Example 2, the gradient of u_θ vanishes and the integral curves γ associated to $\nabla u_\theta = (x_1, x_2)$ emanate along straight lines from $(0, 0)$ towards boundary points $(z_1, z_2) \in \partial\mathcal{O}$ where $y((z_1, z_2)) = 0$. If we parameterise them as $\{(z_1 e^t, z_2 e^t) : -\infty < t \leq 0\}$, then as after (3.30) we see that if a solution $y \in H_0^2$ to (3.29) exists then ψ must necessarily satisfy

$$\int_{-\infty}^0 \psi\left(z_1 e^t, z_2 e^t\right) dt = 0 - y(0) = \text{const.} \quad \forall (z_1, z_2) \in \partial\mathcal{O}. \tag{3.34}$$

This again cannot happen, for example, for any non-negative non-zero $\psi \in H$ that vanishes along a given curve γ (for instance if it is zero in any given quadrant of $\mathcal{O}$), as this forces const $= 0$. Theorems 2 and 3 again yield the following for Example 2:

Theorem 7. *Consider the setting of Theorem 6 but where now* $\mathcal{O}$ *is the unit disk centred at* $(0, 0)$, *and where* $0 \leq \psi \in C_0^\infty(\mathcal{O})$, $\psi \neq 0$, *vanishes along some straight ray from* $(0, 0)$ *to the boundary* $\partial\mathcal{O}$. *Then* (3.32) *and* (3.33) *hold at* $\theta = 1$.

Arguing as after Proposition 5, the result can be extended to any $\theta \in \Theta$ by an application of the Morse lemma.

3.6. Concluding remarks

Remark 1 (Interior saddle points of u_θ). To complement Examples 1, 2, suppose we take $\theta = 1$, $f = 0$ in (3.1) so that $u = u_1 = x_1^2 - x_2^2$ if $g = u_{\partial\mathcal{O}}$ (and $\mathcal{O}$ is the unit disk, say). Then $\nabla u = 2(x_1, -x_2)$ and the critical point is a saddle point. In this case we can find integral curves γ_x running through x away from $(0, 0)$ between boundary points in finite time. Then is ψ is nonnegative and supported near x it cannot integrate to zero along γ_x. An analogue of Theorem 6 then follows for this constellation of parameters in (3.1), too. Note that in this example, the kernel of $\mathbb{I}_\theta$ contains at least all constants.

Remark 2 (Local curvature of $\mathscr{G}$). The quantitative nature of (3.22) in Theorem 5 is compatible with 'gradient stability conditions' employed in [3, 25] to establish polynomial time posterior computation time bounds for gradient based Langevin MCMC schemes. Specifically, arguing as in Lemma 4.7 in [25], for a neighbourhood $\mathcal{B}$ of θ_0 one can deduce local average 'curvature'

$$\inf_{\theta\in\mathcal{B}} \lambda_{\min} E_{\theta_0}\big[-\nabla^2\ell(\theta)\big] \geq c_2 D^{-4/d},$$

of the average-log-likelihood function ℓ when the model Θ is discretised in the eigenbasis $E_D \equiv (e_n : n \leq D) \subset H$ arising from the Dirichlet Laplacian. In this sense (using also the results from [13]) one can expect a Bayesian inference method based on data (2.2) and Gaussian process priors to be consistent and computable even in high-dimensional settings. This shows that such local curvature results are not sufficient to establish (and hence distinct from) Gaussian 'Bernstein–von Mises-type' approximations.

Remark 3 (Boundary constraints on θ). As the main flavour of our results is 'negative', the assumption of knowledge of the boundary values of θ in (3.3) strengthens our conclusions – it is also natural as the regression function $u = g$ is already assumed to be known at $\partial\mathcal{O}$. In the definition of the parameter space Θ, we could further have assumed that all outward normal derivatives up to order $\beta - 1$ of θ vanish at $\partial\mathcal{O}$. This would be in line with the parameter spaces from [13, 24]. All results in this section remain valid because our choice of tangent space H in (3.11) is compatible with this more constrained parameter space.

Remark 4 (Ellipticity). The Bernstein–von Mises theorems from [18, 20, 21] exploit *ellipticity* of the information operator $\mathbb{I}_\theta^*\mathbb{I}_\theta$ in their settings, allowing one to solve for y in the equation $\mathbb{I}_\theta^*\mathbb{I}_\theta y = \psi$ so that $R(\mathbb{I}_\theta^*)$ contains at least all smooth compactly supported ψ (and this is so for *any* parameter $\theta \in \Theta$). In contrast, in the present inverse problem arising from (3.1), the information operator does not have this property and solutions y to the critical equation $\mathbb{I}_\theta^* y = \psi$ exist only under stringent geometric conditions on ψ. Moreover, these conditions exhibit a delicate dependence on θ, further constraining the set $\bigcap_{\theta\in\Theta} R(\mathbb{I}_\theta^*)$ relevant for purposes of statistical inference.

4. APPENDIX

For convenience of the reader we include here a few more proofs of some results of this article.

4.1. Proofs of Theorem 5 and Proposition 3

Define the operator

$$T_\theta(h) = \nabla\cdot(h\nabla u_\theta), \quad h \in H,$$

so that (3.12) becomes $\mathbb{I}_\theta = V_\theta \circ T_\theta$. The map $u \mapsto (\mathcal{L}_\theta u, u_{|\partial\mathcal{O}})$ is a topological isomorphism between $H^2(\mathcal{O})$ and $L^2(\mathcal{O}) \times H^{3/2}(\partial\mathcal{O})$ (see [17], Theorem II.5.4), and hence with $u = V_\theta[w]$ we deduce $\|V_\theta[w]\|_{H^2} \gtrsim \|w\|_{L^2}$ for all $w \in C^\infty(\mathcal{O})$. As a consequence, using

also Lemma 1,

$$\|\mathbb{I}_\theta[h]\|_{H^2} \gtrsim \|T_\theta(h)\|_{L^2} \gtrsim \|h\|_{L^2}, \quad h \in H,$$

which proves the inequality in Theorem 5. Next, as $\mathbb{I}_\theta$ is linear, we see that whenever $\mathbb{I}_\theta[h_1] = \mathbb{I}_\theta[h_2]$ for $h_1, h_2 \in H$ we have $\mathbb{I}_\theta[h_1 - h_2] = 0$, and so by the preceding inequality $h = h_1 - h_2 = 0$ in L^2, too. Likewise, if $h_1, h_2 \in H$ are such that $\mathbb{I}_\theta^*\mathbb{I}_\theta h_1 = \mathbb{I}_\theta^*\mathbb{I}_\theta h_2$, then $0 = \langle \mathbb{I}_\theta^*\mathbb{I}_\theta(h_1 - h_2), h_1 - h_2\rangle_{L^2_\lambda} = \|\mathbb{I}_\theta(h_1 - h_2)\|^2_{L^2_\lambda}$ so $\mathbb{I}_\theta h_1 = \mathbb{I}_\theta h_2$ and thus by what precedes $h_1 = h_2$.

Lemma 1. *We have $\|T_\theta(h)\|_{L^2} = \|\nabla \cdot (h\nabla u_\theta)\|_{L^2} \geq c\|h\|_{L^2}$ for all $h \in H$ and some constant $c = c(\mu, B, c_0) > 0$, where $B \geq \|u_\theta\|_\infty$.*

Proof. Applying the Gauss–Green theorem to any $v \in C^1(\mathcal{O})$ vanishing at $\partial\mathcal{O}$ gives

$$\langle \Delta u_\theta, v^2\rangle_{L^2} + \frac{1}{2}\langle \nabla u_\theta, \nabla(v^2)\rangle_{L^2} = \frac{1}{2}\langle \Delta u_\theta, v^2\rangle_{L^2}.$$

For $v = e^{-\mu u_\theta} h$, $h \in H$, with $\mu > 0$ to be chosen, we thus have

$$\frac{1}{2}\int_{\mathcal{O}} \nabla(v^2)\cdot\nabla u_\theta = -\int_{\mathcal{O}} \mu\|\nabla u_\theta\|^2 v^2 + \int_{\mathcal{O}} v e^{-\mu u_\theta}\nabla h\cdot\nabla u_\theta,$$

so that by the Cauchy–Schwarz inequality

$$\begin{aligned}\left|\int_{\mathcal{O}}\left(\frac{1}{2}\Delta u_\theta + \mu\|\nabla u_\theta\|^2\right)v^2\right| &= \left|\left\langle\left(\Delta u_\theta + \mu\|\nabla u_\theta\|^2\right), v^2\right\rangle_{L^2} + \frac{1}{2}\langle\nabla u_\theta, \nabla(v^2)\rangle_{L^2}\right| \\ &= \left|\langle h\Delta u_\theta + \nabla h\cdot\nabla u_\theta, h e^{-2\mu u_\theta}\rangle_{L^2}\right| \\ &\leq \mu\left\|\nabla\cdot(h\nabla u_\theta)\right\|_{L^2}\|h\|_{L^2} \end{aligned} \tag{4.1}$$

for $\bar{\mu} = \exp(2\mu\|u_\theta\|_\infty)$. We next lower bound the multipliers of v^2 in the left-hand side of (4.1). By (3.6),

$$\left|\int_{\mathcal{O}}\left(\frac{1}{2}\Delta u_\theta + \mu\|\nabla u_\theta\|^2\right)v^2\right| \geq c_0\int_{\mathcal{O}} v^2$$

and, combining this with (4.1), we deduce

$$\left\|\nabla\cdot(h\nabla u_\theta)\right\|_{L^2}\|h\|_{L^2} \geq c'\|v\|^2_{L^2(\mathcal{O})} \gtrsim \|h\|^2_{L^2},\ h \in H,$$

which is the desired estimate. ∎

The last lemma also immediately implies Proposition 3. Let us write $h = \theta_1 - \theta_2$ which defines an element of H. Then by (3.1) we have $\nabla\cdot(h\nabla u_{\theta_1}) = \nabla\cdot(\theta_2\nabla(u_{\theta_2} - u_{\theta_1}))$ and hence $\|\nabla\cdot(h\nabla u_{\theta_1})\|_{L^2} \lesssim \|u_{\theta_2} - u_{\theta_1}\|_{H^2}$. By Lemma 1 the left-hand side is lower bounded by a constant multiple of $\|h\|_{L^2} = \|\theta_1 - \theta_2\|_{L^2}$, so that the result follows.

4.2. Proof of Theorem 2 for $\mathbb{I}_\theta^*\mathbb{I}_\theta$ compact

Let us assume $\bar{H} = \mathbb{H}$ without loss of generality, write $I \equiv \mathbb{I}_\theta$, $L^2 = L^2_\lambda(\mathcal{X})$ in this proof, and let $\ker(I^*I) = \{h \in \mathbb{H} : I^*Ih = 0\}$. If I^*I is a compact operator on $\mathbb{H}$ then by the spectral theorem for self-adjoint operators, there exists an orthonormal system of $\mathbb{H}$

of eigenvectors $\{e_k : k \in \mathbb{N}\}$ spanning $\mathbb{H} \ominus \ker(I^*I)$ corresponding to eigenvalues $\lambda_k > 0$ so that

$$I^*Ie_k = \lambda_k e_k, \quad \text{and} \quad I^*Ih = \sum_k \lambda_k \langle h, e_k \rangle_{\mathbb{H}} e_k, \quad h \in \mathbb{H}.$$

We can then define the usual square-root operator $(I^*I)^{1/2}$ by

$$\big(I^*I\big)^{1/2} h = \sum_k \lambda_k^{1/2} \langle h, e_k \rangle_{\mathbb{H}} e_k, \quad h \in \mathbb{H}. \tag{4.2}$$

If we denote by P_0 the $\mathbb{H}$-projection onto $\ker(I^*I)$, then the range of $(I^*I)^{1/2}$ equals

$$R\big(\big(I^*I\big)^{1/2}\big) = \Big\{ g \in \mathbb{H} : P_0(g) = 0, \sum_k \lambda_k^{-1} \langle e_k, g \rangle_{\mathbb{H}}^2 < \infty \Big\}. \tag{4.3}$$

Indeed, using standard Hilbert space arguments, (a) since $P_0(e_k) = 0$ for all k, for any $h \in \mathbb{H}$ the element $g = (I^*I)^{1/2}h$ belongs to the right-hand side in the last display, and conversely (b) if g satisfies $P_0(g) = 0$ and $\sum_k \lambda_k^{-1} \langle e_k, g \rangle_{\mathbb{H}}^2 < \infty$ then $h = \sum_k \lambda_k^{-1/2} \langle e_k, g \rangle e_k$ belongs to $\mathbb{H}$ and $(I^*I)^{1/2}h = g$.

Next, Lemma A.3 in [30] implies that $R(I^*) = R((I^*I)^{1/2})$. Now suppose $\psi \in \mathbb{H}$ is such that $\psi \notin R(I^*)$ and hence $\psi \notin R((I^*I)^{1/2})$. Then from (4.3), either $P_0(\psi) \neq 0$ or $\sum_k \lambda_k^{-1} \langle e_k, \psi \rangle_{\mathbb{H}}^2 = \infty$ (or both). In the first case, let $\bar{h} = P_0(\psi)$, so

$$\|I\bar{h}\|_{L^2} = \big\|I\big(P_0(\psi)\big)\big\|_{L^2} = \big\langle I^*I\big(P_0(\psi)\big), P_0(\psi)\big\rangle_{\mathbb{H}} = 0,$$

but $\langle \psi, \bar{h} \rangle_{\mathbb{H}} = \|P_0\psi\|_{\mathbb{H}}^2 = \delta$ for some $\delta > 0$. Since H is dense in $\mathbb{H}$, for any $\epsilon, 0 < \epsilon < \min(\delta/(2\|\psi\|_{\mathbb{H}}), \delta^2/4)$, we can find $h \in H$ such that $\|h - \bar{h}\|_{\mathbb{H}} < \epsilon$ and by continuity also $\|I(h - \bar{h})\|_{L^2} < \epsilon$. Then

$$\sqrt{i_{\theta,h,\psi}} = \frac{\|Ih\|_{L^2}}{|\langle \psi, h \rangle_{\mathbb{H}}|} \leq 2\frac{\epsilon}{\delta} \leq \sqrt{\epsilon}.$$

Using also (2.12), we conclude that $i_{\theta,H,\psi} < \epsilon$ in (2.21), so that the result follows since ϵ was arbitrary. In the second case we have $\sum_k \lambda_k^{-1} \langle e_k, \psi \rangle_{\mathbb{H}}^2 = \infty$ and define

$$\psi_N = \sum_{k \leq N} \lambda_k^{-1} e_k \langle e_k, \psi \rangle_{\mathbb{H}}, \quad N \in \mathbb{N},$$

which defines an element of $\mathbb{H}$. By density we can choose $h_N \in H$ such that $\|h_N - \psi_N\|_{\mathbb{H}} < 1/\|\psi\|_{\mathbb{H}}$, as well as $\|I(h_N - \psi_N)\|_{L^2} < 1$, for every N fixed. Next observe that

$$\langle \psi, \psi_N \rangle_{\mathbb{H}} = \sum_{k \leq N} \lambda_k^{-1} \langle e_k, \psi \rangle_{\mathbb{H}}^2 \equiv M_N,$$

$$\big\|I(\psi_N)\big\|_{L^2}^2 = \big\langle I^*I(\psi_N), \psi_N \big\rangle_{\mathbb{H}} = \sum_{k \leq N} \lambda_k^{-1} \langle e_k, \psi \rangle_{\mathbb{H}}^2 = M_N,$$

and that $M_N \to \infty$ as $N \to \infty$. Then by our choice of $h_N \in H$ and if $M_N \geq 2$, we have by the triangle inequality,

$$\big|\langle \psi, h_N \rangle_{\mathbb{H}}\big| \geq \big|\langle \psi, \psi_N \rangle_{\mathbb{H}}\big| - \big|\langle \psi, \psi_N - h_N \rangle_{\mathbb{H}}\big| \geq M_N - 1 \geq M_N/2,$$

$$\big\|I(h_N)\big\|_{L^2} \leq \big\|I(\psi_N)\big\|_{L^2} + \big\|I(h_N - \psi_N)\big\|_{L^2} \leq \sqrt{M_N} + 1 \leq 2\sqrt{M_N}.$$

From this and (2.12) we conclude that the inverse of (2.21) satisfies

$$i_{\theta,H,\psi}^{-1} \geq \frac{\langle \psi, h_N \rangle_{\mathbb{H}}^2}{\|I h_N\|_{L^2}^2} \geq \frac{1}{16} \frac{M_N^2}{M_N} \geq M_N/16.$$

As N was arbitrary and $M_N \to_{N\to\infty} \infty$, we must have $i_{\theta,H,\psi} = 0$, as desired.

ACKNOWLEDGMENTS

We are grateful to Jan Bohr and Lauri Oksanen for helpful remarks and discussions.

REFERENCES

[1] K. Abraham and R. Nickl, On statistical Caldéron problems. *Math. Statist. Learn.* **2** (2019), no. 2, 165–216.

[2] G. S. Alberti, G. Bal, and M. Di Cristo, Critical points for elliptic equations with prescribed boundary conditions. *Arch. Ration. Mech. Anal.* **226** (2017), no. 1, 117–141.

[3] J. Bohr and R. Nickl, On log-concave approximations of high-dimensional posterior measures and stability properties in non-linear inverse problems. 2021, arXiv:2105.07835.

[4] I. Castillo and R. Nickl, Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** (2013), no. 4, 1999–2028.

[5] I. Castillo and R. Nickl, On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** (2014), no. 5, 1941–1969.

[6] I. Castillo and J. Rousseau, A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** (2015), no. 6, 2353–2383.

[7] J. J. Duistermaat and L. Hörmander, Fourier integral operators. II. *Acta Math.* **128** (1972), no. 3–4, 183–269.

[8] M. Freidlin, *Functional integration and partial differential equations*. Ann. of Math. Stud. 109, Princeton University Press, Princeton, NJ, 1985.

[9] C.-F. Gauß, *Theoria Motus Corporum Coelestium*. Perthes, Hamburg, 1809.

[10] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*. Springer, Berlin–New York, 1998.

[11] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, New York, 2016.

[12] M. Giordano and H. Kekkonen, Bernstein–von Mises theorems and uncertainty quantification for linear inverse problems. *SIAM/ASA J. Uncertain. Quantificat.* **8** (2020).

[13] M. Giordano and R. Nickl, Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem. *Inverse Probl.* (2020).

[14] M. Giordano and K. Ray, Nonparametric Bayesian inference for reversible multidimensional diffusions. 2018, arXiv:1802.05635.

[15] M. Hairer, A. M. Stuart, and S. J. Vollmer, Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* **24** (2014), no. 6, 2455–2490.

[16] H. Kekkonen, Consistency of Bayesian inference with Gaussian process priors for a parabolic inverse problem. 2021, arXiv:2103.13213.

[17] J.-L. Lions and E. Magenes, *Non-homogeneous boundary value problems and applications. Vol. I*. Springer, New York–Heidelberg, 1972.

[18] F. Monard, R. Nickl, and G. P. Paternain, Efficient nonparametric Bayesian inference for X-ray transforms. *Ann. Statist.* **47** (2019), no. 2, 1113–1147.

[19] F. Monard, R. Nickl, and G. P. Paternain, Consistent inversion of noisy non-abelian X-ray transforms. *Comm. Pure Appl. Math.* **74** (2021), 1045–1099.

[20] F. Monard, R. Nickl, and G. P. Paternain, Statistical guarantees for Bayesian uncertainty quantification in non-linear inverse problems with Gaussian process priors. *Ann. Statist.* **49** (2021), 3255–3298.

[21] R. Nickl, Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc. (JEMS)* **22** (2020), 2697–2750.

[22] R. Nickl and K. Ray, Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.* **48** (2020), no. 3, 1383–1408.

[23] R. Nickl and J. Söhl, Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.* **45** (2017), no. 4, 1664–1693.

[24] R. Nickl, S. van de Geer, and S. Wang, Convergence rates for penalised least squares estimators in PDE-constrained regression problems. *SIAM/ASA J. Uncertain. Quantificat.* **8** (2020).

[25] R. Nickl and S. Wang, On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. 2020, arXiv:2009.05298.

[26] K. Ray, Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** (2017), no. 6, 2511–2536.

[27] G. R. Richter, An inverse problem for the steady state diffusion equation. *SIAM J. Appl. Math.* **41** (1981), no. 2, 210–221.

[28] A. M. Stuart, Inverse problems: a Bayesian perspective. *Acta Numer.* **19** (2010), 451–559.

[29] M. E. Taylor, *Partial differential equations I. Basic theory*. 2nd edn., Appl. Math. Sci. 115, Springer, New York, 2011.

[30] A. W. van der Vaart, On differentiable functionals. *Ann. Statist.* **19** (1991), no. 1, 178–204.

[31] A. W. van der Vaart, *Asymptotic statistics*. Cambridge Univ. Press, 1998.

RICHARD NICKL

University of Cambridge, Faculty of Mathematics, Cambridge CB3 0WA, UK,
nickl@maths.cam.ac.uk

GABRIEL P. PATERNAIN

University of Cambridge, Faculty of Mathematics, Cambridge CB3 0WA, UK,
g.p.paternain@dpmms.cam.ac.uk

FROM STATISTICAL TO CAUSAL LEARNING

BERNHARD SCHÖLKOPF AND
JULIUS VON KÜGELGEN

ABSTRACT

We describe basic ideas underlying research to build and understand artificially intelligent systems: from symbolic approaches via statistical learning to interventional models relying on concepts of causality. Some of the hard open problems of machine learning and AI are intrinsically related to causality, and progress may require advances in our understanding of how to model and infer causality from data.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68T05; Secondary 68Q32, 68T01, 68T10, 68T30, 68T37

KEYWORDS

Causal inference, machine learning

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5540–5593
DOI 10.4171/ICM2022/173

1. INTRODUCTION

In 1958, the New York Times reported on a new machine called the *perceptron*. Frank Rosenblatt, its inventor, demonstrated that the perceptron was able to learn from experience. He predicted that later perceptrons would be able to recognize people, or instantly translate spoken language. Now a reality, this must have sounded like distant science fiction at the time. In hindsight, we may consider it the birth of machine learning, the field fueling most of the current advances in artificial intelligence (AI).

Around the same time, another equally revolutionary development took place: scientists understood that computers could do more than compute numbers: they can process symbols. Although this insight was also motivated by artificial intelligence, in hindsight it was the birth of the field of computer science. There was great optimism that the manipulation of symbols, in programs written by humans, implementing rules designed by humans, should be enough to generate intelligence. Below, we shall refer to this as the *symbol–rule hypothesis*.[1]

There was initially encouraging progress on seemingly hard problems such as automatic theorem proving and computer chess. One of the fathers of the field, Herb Simon, predicted in 1956 that "machines will be capable, within twenty years, of doing any work a man can do." However, problems that appeared simple, such as most things animals could do, turned out to be hard. This came to be known as *Moravec's paradox*. When IBM's *Deep Blue* chess computer beat Garry Kasparov in 1997, Kasparov was physically facing a human during the match: while Deep Blue was capable of analyzing the game's search tree in unprecedented detail, it was unable to recognize and physically move chess pieces, so this task had to be relegated to a human, in an inversion of the famous *mechanical turk*.[2]

In the years to follow, the field of AI entered what came to be known as the *AI winter*. The community got disillusioned with the lack of progress and prospects, and interest greatly declined. However, largely independently of the field of classic AI, *machine learning* eventually started to boom. Like Rosenblatt's early work, it was built on the observation that all existing examples of truly intelligent systems—i.e., animals, including humans—were not built on the symbol–rule hypothesis: both the representations and the rules implemented by natural intelligent systems are acquired from experience, through processes of evolution and learning.

Rather than exploring the well-known dichotomy between rule- and learning-based approaches, we will explore the less known questions of causality and interventions. While the field of causality in computer science was initially strongly linked to classic AI, recent years have witnessed great interest in connecting it to machine learning [111]. Below, we explore some of these connections, drawing from [125, 133]. We will argue that the causal view is relevant when it comes to addressing crucial open problems of machine learning, related to notions of robustness and generalization beyond the training distribution.

---

1 The term should be taken with a grain of salt, since it suggests a separation between representations and computations which is hard to uphold in practice.

2 https://en.wikipedia.org/wiki/Mechanical_Turk.

Overview. In statistical learning, our starting point is a joint distribution $p(\mathbf{X})$ generating the observable data. Here, $\mathbf{X}$ is a random vector, and we are usually given a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_m$ sampled i.i.d. from p. We are often interested in estimating properties of conditionals of some components of $\mathbf{X}$ given others, e.g., a classifier (which may be obtained by thresholding a conditional at 0.5). This is a nontrivial inverse problem, giving rise to statistical learning theory (Section 2).

Causal learning is motivated by shortcomings of statistical learning (Section 3). Its starting point is a structural causal model (SCM) [104] (Section 4). In an SCM, the components $X_1, \ldots, X_n$ of $\mathbf{X}$ are identified with vertices of a directed graph whose arrows represent direct causal influences, and there is a random variable U_i for each vertex, along with a function f_i which computes X_i from its graph parents $\mathbf{PA}_i$ and U_i, i.e.,

$$X_i := f_i(\mathbf{PA}_i, U_i). \tag{1.1}$$

Given a distribution over the U_i, which are assumed independent, this also gives rise to a probabilistic model $p(\mathbf{X})$. However, the model in (1.1) is more structured: the graph connectivity and the functions f_i create particular dependences between the observables. Moreover, it describes how the system behaves under intervention: by replacing functions by constants, we can compute the effect of setting some variables to specific values.

Causal learning builds on assumptions different from standard machine learning (Section 5), and addresses a different level in the modeling hierarchy (Section 6). It also comes with new problems, such as causal discovery, where we seek to infer properties of graph and functions from data (Section 7). In some cases, conditional independences among the X_i contain information about the graph [144]; but novel assumptions let us handle some cases that were previously unsolvable [68]. Those assumptions have nontrivial implications for machine learning tasks such as semisupervised learning, covariate shift adaptation and transfer learning [128] (Section 8). Once provided with a causal model, causal reasoning (Section 9) allows us to identify and estimate certain causal queries of interest from observational data. We conclude with a list of some current and open problems (Section 10), with a particular emphasis on the topic of causal representation learning.

The presentation and notation will be somewhat informal in several respects. We generally assume that all distributions possess densities (with respect to a suitable reference measure). We sometimes write $p(x)$ for the distribution (or density) of a random variable X. Accordingly, the same p can denote another distribution $p(y)$, distinguished by the argument of $p(\cdot)$. We also sometimes use summation for marginalization which supposes discrete variables; the corresponding expressions for continuous quantities would use integrals.

2. STATISTICAL LEARNING THEORY

Suppose we have measured two statistically dependent observables and found the points to lie approximately on a straight line. An empirical scientist might be willing to hypothesize a corresponding law of nature (see Figure 1). However, already Leibniz pointed out that if we scatter spots of ink randomly on a piece of paper by shaking a quill pen, we

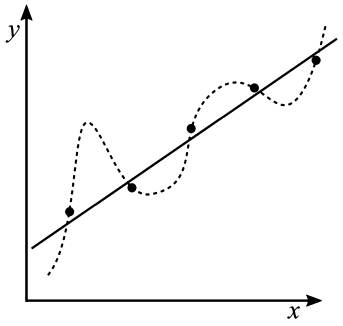


FIGURE 1
Given a small number of observations, how do we find a law underlying them? Leibniz argued that even if we generate a random set of points, we can always find a mathematical equation satisfied by these points.

can also find a mathematical equation satisfied by these points [81]. He argued that we would not call this a law of nature, because no matter how the points are distributed, there always exists such an equation; we would only call it a law of nature only if the equation is simple. This raises the question of what makes an equation simple. The physicist Rutherford took the pragmatic view that if there is a law, it should be directly evident from the data: "if your experiment needs statistics, you ought to have done a better experiment."[3] This view may have been a healthy one when faced with low-dimensional inference problems where regularities are immediately obvious; however, modern AI is facing inference problems that are harder: they are often high-dimensional and nonlinear, yet we may have little prior knowledge about the underlying regularity (e.g., for medical data, we usually do not have a mechanistic model).

Statistical learning theory studies the problem of how to still perform valid inference, provided that we have sufficiently large datasets and the computational means to process them. Let us look at some theoretical results for the simplest learning scenario, drawing from [130]; for details, see [153]. Suppose we are given empirical observations,

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}, \tag{2.1}$$

where $\mathcal{X}$ is some nonempty set from which the *inputs* come, and $\mathcal{Y} = \{\pm 1\}$ is the *output* set, in our case consisting of just two *classes*. This situation is called *pattern recognition*, and our goal is to use the *training data* (2.1) to infer a function $f : \mathcal{X} \to \{\pm 1\}$ (from some function class chosen a priori) which will produce the correct output for a new input x which we may not have seen before. To formalize what we mean by "correct," we make the assumption that all observations (x_i, y_i) have been generated independently by performing a random experiment described by an unknown probability distribution $p(x, y)$—a setting referred to as *i.i.d. (independent and identically distributed) data*. Our goal will be to minimize the

3 Cited after http://www.warwick.ac.uk/statsdept/staff/JEHS/data/jehsquot.pdf.

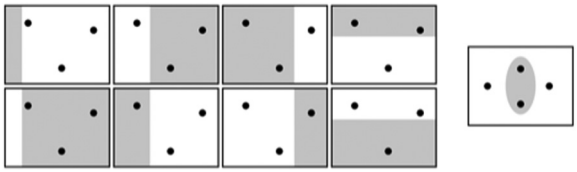


FIGURE 2

Using straight lines, we can separate three points in all possible ways; we cannot do this for four points, no matter how they are placed. The class of linear separations is not "falsifiable" using three points, but it becomes falsifiable once we have four or more points.

expected error (or risk)

$$R[f] = \int_{\mathcal{X}\times\mathcal{Y}} c(y, f(x))dp(x, y), \tag{2.2}$$

where c is a so-called loss function, e.g., the misclassification error $c(y, f(x)) = \frac{1}{2}|f(x) - y|$ taking the value 0 whenever $f(x) = y$ and 1 otherwise.

The difficulty of the task stems from the fact that we are trying to minimize a quantity that we cannot evaluate: since we do not know p, we cannot compute (2.2). We do know, however, the training data (2.1) sampled from p. We can thus try to infer a function f from the training sample whose risk is close to the minimum of (2.2). To this end, we need what is called an *induction principle*.

One way to proceed is to use the training sample to approximate (2.2) by a finite sum, referred to as the *empirical risk*

$$R_{\text{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i)). \tag{2.3}$$

The *empirical risk minimization (ERM) induction principle* recommends that we choose (or "learn") an f that minimizes (2.3). We can then ask whether the ERM principle is statistically *consistent*: in the limit of infinitely many data points, will ERM lead to a solution which will do as well as possible on future data generated by p?

It turns out that if the function class over which we minimize (2.3) is too large, then ERM is not consistent. Hence, we need to suitably restrict the class of possible functions. For instance, ERM is consistent for all probability distributions, provided that the *VC dimension* of the function class is finite. The VC dimension is an example of a *capacity measure*. It is defined as the maximal number of points that can be separated (classified) in all possible ways using functions from the class. For example, using linear classifiers (separating classes by straight lines) on $\mathbb{R}^2$, we can realize all possible classifications for 3 suitably chosen points, but we can no longer do this once we have 4 points, no matter how they are placed (see Figure 2). This means that the VC dimension of this function class is 3. More generally, for linear separations in $\mathbb{R}^d$, the VC dimension is $d + 1$.

Whenever the VC dimension is finite, our class of functions (or explanations) becomes falsifiable in the sense that starting from a certain number of observations, no

longer all possible labelings of the points can be explained (cf. Figure 2). If we can nevertheless explain a sufficiently large set of observed data, we thus have reason to believe that this is a meaningful finding.

Much of machine learning research is concerned with restrictions on classes of functions to make inference possible, be it by imposing prior distributions on function classes, through other constraints, or by designing self-regularizing learning procedures, e.g., gradient descent methods for neural networks [79]. While there is a solid theoretical understanding of supervised machine learning as described above (i.e., function learning from input–output examples), there are still details under investigation, such as the recently observed phenomenon of "double descent" [7].

A popular constraint, implemented in the *Support Vector Machine (SVM)* [130,153], is to consider linear separations with large margin: it turns out that for large margin separations in high-dimensional (or infinite-dimensional) spaces, the capacity can be much smaller than the dimensionality, making learning possible in situations where it would otherwise fail.

For some learning algorithms, including SVMs and nearest neighbor classifiers, there are strong universal consistency results, guaranteeing convergence of the algorithm to the lowest achievable risk, for any problem to be learned [28,130,146,153]. Note, however, that this convergence can be arbitrarily slow.

For a given sample size, it will depend on the problem being learned whether we achieve low expected error. In addition to asymptotic consistency statements, learning theory makes finite sample size statements: one can prove that with probability at least $1-\delta$ (for $\delta > 0$), for all functions f in a class of functions with VC dimension h,

$$R[f] \leq R_{\text{emp}}[f] + \sqrt{\frac{1}{m}\left(h(\log(2m/h)+1) + \log\frac{4}{\delta}\right)}. \tag{2.4}$$

This is an example of a class of results that relate the training error $R_{\text{emp}}[f]$ and the test error $R[f]$ using a confidence interval (the square root term) depending on a capacity measure of a function class (here, its VC dimension h). It says that with high probability, the expected error $R[f]$ on future observations generated by the unknown probability distribution is small, provided the two terms on the right-hand side are small: the training error $R_{\text{emp}}[f]$ (i.e., the error on the examples we have already seen), and the square root term, which will be small whenever the capacity h is small compared to the number of training observations m. If, on the other hand, we try to learn something that may not make sense, such as the mapping from the name of people to their telephone number, we would find that to explain all the training data (i.e., to obtain a small $R_{\text{emp}}[f]$), we need a model whose capacity h is large, and the second term becomes large. In any case, it is crucial for both consistency results and finite sample error bounds such as (2.4) that we have i.i.d. data.

Kernel methods. A symmetric function $k : \mathcal{X}^2 \to \mathbb{R}$, where $\mathcal{X}$ is a nonempty set, is called a positive definite (pd) *kernel* if for arbitrary points $x_1, \ldots, x_m \in \mathcal{X}$ and coefficients $a_1, \ldots, a_m \in \mathbb{R}$:

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0.$$

The kernel is called strictly positive definite if for pairwise distinct points, the implication $\sum_{i,j} a_i a_j k(x_i, x_j) = 0 \Longrightarrow \forall i : a_i = 0$ is valid. Any positive definite kernel induces a mapping

$$\Phi : x \mapsto k(x, \cdot) \tag{2.5}$$

into a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ satisfying

$$\big\langle k(x, \cdot), k(x', \cdot) \big\rangle = k(x, x') \tag{2.6}$$

for all $x, x' \in \mathcal{X}$. Although $\mathcal{H}$ may be infinite-dimensional, we can construct practical classification algorithms in $\mathcal{H}$ provided that all computational steps are carried out in terms of scalar products, since those can be reduced to kernel evaluations (2.6).

In the SVM algorithm, the capacity of the function class is restricted by enforcing a large margin of class separation in $\mathcal{H}$ via a suitable RKHS regularization term. The solution can be shown to take the form

$$f(x) = \operatorname{sgn}\Big(\sum_i \alpha_i k(x_i, x) + b \Big), \tag{2.7}$$

where the learned parameters α_i and b are the solution of a convex quadratic optimization problem. A similar expansion of the solution in terms of kernel functions evaluated at training points holds true for a larger class of kernel algorithms beyond SVMs, regularized by an RKHS norm [126].

In kernel methods, the kernel plays three roles which are crucial for machine learning: it acts as a similarity measure for data points, induces a representation in a linear space[4] via (2.5), and parametrizes the function class within which the solution is sought, cf. (2.7).

Kernel mean embeddings. Consider two sets of points $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$ and $Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$. We define the mean map μ as [130]

$$\mu(X) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot). \tag{2.8}$$

For polynomial kernels $k(x, x') = (\langle x, x' \rangle + 1)^d$, we have $\mu(X) = \mu(Y)$ if all empirical moments up to order d coincide. For strictly pd kernels, the means coincide only if $X = Y$, rendering μ injective [131]. The mean map has some other interesting properties [143], e.g., $\mu(X)$ represents the operation of taking a mean of a function on the sample X:

$$\big\langle \mu(X), f \big\rangle = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i).$$

Moreover, we have

$$\big\| \mu(X) - \mu(Y) \big\| = \sup_{\|f\| \le 1} \big| \big\langle \mu(X) - \mu(Y), f \big\rangle \big| = \sup_{\|f\| \le 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right|.$$

4 Note that the data domain $\mathcal{X}$ need not have any structure other than being a nonempty set.

If $\mathbb{E}_{x,x'\sim p}[k(x,x')],\ \mathbb{E}_{x,x'\sim q}[k(x,x')] < \infty$, then the above statements, including the injectivity of μ, generalize to Borel measures p, q, if we define the mean map as

$$\mu : p \mapsto \mathbb{E}_{x\sim p}\big[k(x,\cdot)\big],$$

and replace the notion of strictly pd kernels by that of characteristic kernels [33]. This means that we do not lose information when representing a probability distribution in the RKHS. This enables us to work with distributions using Hilbert space methods, and construct practical algorithms analyzing distributions using scalar product evaluations.

Note that the mean map μ can be viewed as a generalization of the *moment generating function* M_p of a random variable x with distribution p,

$$M_p(\cdot) = \mathbb{E}_{x\sim p}\big[e^{\langle x,\cdot\rangle}\big].$$

The map μ has applications in a number of tasks, including computing functions of random variables [129] and testing for homogeneity [41] or independence [43]. The latter will be of particular interest to causal inference: we can develop a kernel-based independence test by computing the distance between sample-based embeddings of a joint distribution $p(X,Y)$ and the product of its marginals $p(X)p(Y)$ [42–44,114,165], and generalize it to conditional independence testing [33,100], as required for certain causal discovery methods (see Section 7).

3. FROM STATISTICAL TO CAUSAL MODELS

Methods relying on i.i.d. data. In current successes of machine learning [79], we generally *(i) have large amounts of data*, often from simulations or large-scale human labeling, *(ii) use high capacity machine learning models* (e.g., neural networks with many adjustable parameters), and *(iii) employ high performance computing*. Statistical learning theory offers a partial explanation for recent successes of learning: huge datasets enable training complex models and thus solving increasingly difficult tasks.

However, a crucial aspect that is often ignored is that we *(iv) assume that the data are i.i.d.* This assumption is crucial for good performance in practice, and it underlies theoretical statements such as (2.4). When faced with problems that violate the i.i.d. assumption, all bets are off. Vision systems can be grossly misled if an object that is normally recognized with high accuracy is placed in a context that *in the training set* may be negatively correlated with the presence of the object. For instance, such a system may fail to recognize a cow standing on the beach. In order to successfully generalize in such settings, we would need to construct systems which do not merely rely on statistical dependences, but instead model mechanisms that are robust across certain violations of the i.i.d. assumption. As we will argue, causality provides a natural framework for capturing such stable mechanisms and reasoning about different types of distribution shifts.

Correlation vs. causation. It is a commonplace that *correlation does not imply causation*. Two popular and illustrative examples are the positive correlation between chocolate consumption and Nobel prizes per capita [91], and that between the number of stork breeding

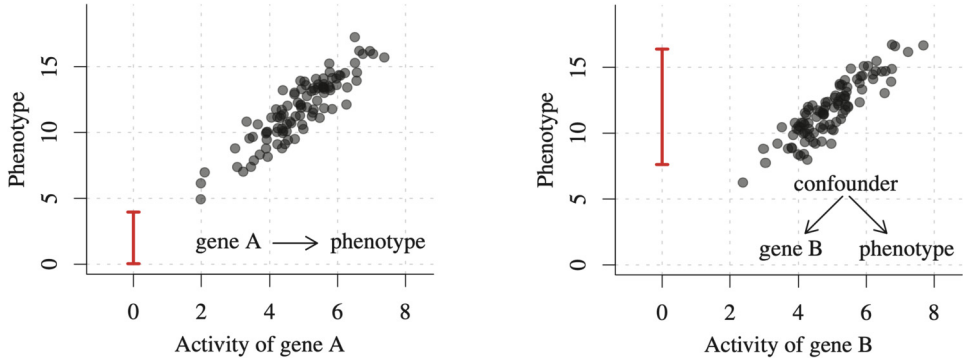


FIGURE 3

Measurements of two genes (x-axis), gene A (left) and gene B (right), show the same strong positive correlation with a phenotype (y-axis). However, this statistical information alone is insufficient to predict the outcome of a knock-out experiment where the activity of a gene is set to zero (vertical lines at $x = 0$). Answering such *interventional* questions requires additional causal knowledge (inset causal graphs): knocking out gene A, which is a direct cause, would lead to a reduction in phenotype, whereas knocking out gene B, which shares a common cause, or confounder, with the phenotype but has no causal effect on it, would leave the phenotype unaffected. This shows that correlation alone is not enough to predict the outcome of perturbations to a system (toy data, figure from [111]).

pairs and human birth rates [89], neither of which admit a sensible interpretation in terms of direct causation. These examples naturally lead to the following questions: What exactly do we mean by "causation"? What is its relationship to correlation? And, if correlation alone is not enough, what is needed to infer causation?

Here, we adopt a notion of causality based on manipulability [159] and intervention [104] which has proven useful in fields such as agriculture [161], econometrics [46, 52], and epidemiology [118].

Definition 3.1 (Causal effect). We say that a random variable X has a causal effect on a random variable Y if there exist $x \neq x'$ such that the distribution of Y after intervening on X and setting it to x differs from the distribution of Y after setting X to x'.

Inherent to the notion of causation, there is a directionality and asymmetry which does not exist for correlation: if X is correlated with Y, then Y is equally correlated with X; but, if X has a causal effect on Y, the converse (in the generic case) does not hold.

We illustrate the intervention-based notion of causation and its difference from correlation (or, more generally, statistical dependence) in Figure 3. Here, knocking out two genes X_A and X_B that are indistinguishable based on their correlation with a phenotype Y would have very different effects. Only intervening on X_A would change the distribution of Y, whereas X_B does not have a causal effect on Y—instead, their correlation arises from a different (confounded) causal structure. Such causal relationships are most commonly represented in the form of *causal graphs* where directed arrows indicate a direct causal effect.

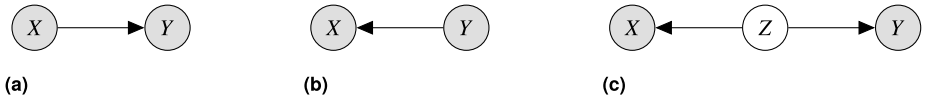


FIGURE 4

Reichenbach's common cause principle **[116]** postulates that statistical dependence between two random variables X and Y has three elementary possible causal explanations shown as causal graphs in (a)–(c). It thus states that association is always induced by an underlying causal process. In (a) the common cause Z coincides with X, and in (b) it coincides with Y. Grey nodes indicate observed and white nodes unobserved variables.

The example in Figure 3 shows that the same correlation can be explained by multiple causal graphs which lead to different experimental outcomes, i.e., *correlation does not imply causation*. However, there is a connection between correlation and causation, expressed by Reichenbach **[116]** as the Common Cause Principle, see Figure 4.

Principle 3.2 (Common cause). *If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a random variable Z which causally influences both of them and which explains all their dependence in the sense of rendering them conditionally independent ($X \perp\!\!\!\perp Y \mid Z$). As a special case, Z may coincide with X or Y.*

According to Principle 3.2, statistical dependence always results from underlying causal relationships by which variables, including potentially unobserved ones, influence each other. Correlation is thus an epiphenomenon, the byproduct of a causal process.

For the example of chocolate consumption (X) and Nobel laureates (Y), common sense suggests that neither of the two variables should have a causal effect on the other, i.e., neither chocolate consumption driving scientific success ($X \to Y$; Figure 4a) nor Nobel laureates increasing chocolate consumption ($Y \to X$; Figure 4b) seem plausible. Principle 3.2 then tells us that the observed correlation must be explained by a common cause Z as in Figure 4c. A plausible candidate for such a confounder could, for example, be economic factors driving both consumer spending and investment in education and science.

Without such background knowledge or additional assumptions, however, we cannot distinguish the three cases in Figure 4 through passive observation, i.e., in a purely data-driven way: the class of observational distributions over X and Y that can be realized by these models is the same in all three cases.

To be clear, this does not mean that correlation cannot be useful, or that causal insight is always required. Both genes in Figure 3 remain useful features for making predictions in a passive, or *observational*, setting in which we measure the activities of certain genes and are asked to predict the phenotype. Similarly, chocolate consumption remains predictive of winning Nobel prizes. However, if we want to answer interventional questions, such as the outcome of a gene-knockout experiment or the effect of a policy enforcing higher chocolate consumption, we need more than correlation: *a causal model*.

4. CAUSAL MODELING FRAMEWORKS

Causal inference has a long history in a variety of disciplines, including statistics, econometrics, epidemiology, and AI. As a result, different frameworks for causal modeling have emerged over the years and coexist today. The first framework described below (CGM) starts from the distribution of the observables, combining it with a directed graph to endow it with causal semantics. The second (SCM) starts from a graph and a set of functional assignments, and generates the observed distribution as the push-forward of an unobserved noise distribution. Finally, we cover a nongraphical approach (PO) popular in statistics.

Causal graphical models (CGMs). The graphical models framework [75, 78] provides a compact way of representing joint probability distributions by encoding the dependence structure between variables in graphical form. Directed graphical models are also known as *Bayesian networks* [101]. While they do not offer a causal interpretation per se—indeed, different graphical models can be compatible with the same distribution (cf. Principle 3.2)—when edges are endowed with the notion of direct causal effect (Definition 3.1), we refer to them as causal graphical models (CGM) [144].

Definition 4.1 (CGM). A CGM $\mathcal{M} = (G, p)$ over n random variables $X_1, \ldots, X_n$ consists of: (i) a directed acyclic graph (DAG) G in which directed edges ($X_j \to X_i$) represent a direct causal effect of X_j on X_i; and (ii) a joint distribution $p(X_1, \ldots, X_n)$ which is Markovian with respect to G:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathbf{PA}_i) \tag{4.1}$$

where $\mathbf{PA}_i = \{X_j : (X_j \to X_i) \in G\}$ denotes the set of parents, or direct causes, of X_i in G.

We will refer to (4.1) as the *causal (or disentangled) factorization*. While many other *entangled factorizations* are possible, e.g.,

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid X_{i+1}, \ldots, X_n), \tag{4.2}$$

only (4.1) decomposes the joint distribution into causal conditionals, or *causal mechanisms*, $p(X_i \mid \mathbf{PA}_i)$, which can have a meaningful physical interpretation, rather than being mere mathematical objects such as the factors on the RHS of (4.2).

It turns out that (4.1) is equivalent to the following condition.

Definition 4.2 (Causal Markov condition). A distribution p satisfies the causal Markov condition with respect to a DAG G if every variable is conditionally independent of its nondescendants in G given its parents in G.

Definition 4.2 can equivalently be expressed in terms of *d-separation*, a graphical criterion for directed graphs [104], by saying that *d-separation in G implies (conditional) independence in p*. The causal Markov condition thus provides a link between properties of p and G.

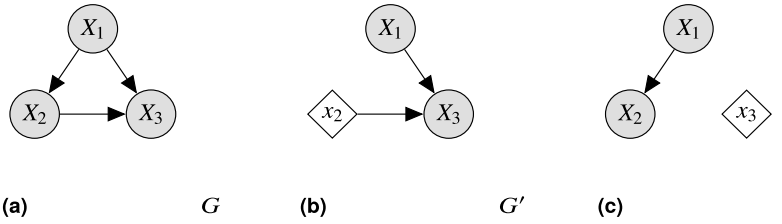


FIGURE 5

(a) A directed acyclic graph (DAG) G over three variables. A causal graphical model (G, p) with causal graph G and observational distribution p can be used to answer interventional queries using the concept of *graph surgery*: when a variable is intervened upon and set to a constant (white diamonds), this removes any influence from other variables, captured graphically by removing all incoming edges. (b) and (c) show postintervention graphs G' and G'' for $\mathrm{do}(X_2 := x_2)$ and $\mathrm{do}(X_3 := x_3)$, respectively. (An intervention on X_1 would leave the graph unaffected.)

What makes CGMs causal is the interpretation of edges as cause–effect relationships which enables reasoning about the outcome of interventions using the *do-operator* [104] and the concept of *graph surgery* [144]. The central idea is that intervening on a variable, say by externally forcing it to take on a particular value, renders it independent of its causes and breaks their causal influence on it, see Figure 5 for an illustration. For example, if a gene is knocked out, it is no longer influenced by other genes that were previously regulating it; instead, its activity is now solely determined by the intervention. This is fundamentally different from conditioning since passively observing the activity of a gene provides information about its driving factors (i.e., its direct causes).

To emphasize this difference between passive observation and active intervention, Pearl [104] introduced the notation $\mathrm{do}(X := x)$ to denote an intervention by which variable X is set to value x. The term *graph surgery* refers to the idea that the effect of such an intervention can be captured in the form of a modification to the original graph by removing all incoming edges to the intervened variable. Interventional queries can then be answered by performing probabilistic inference in the modified postintervention graph which typically implies additional (conditional) independences due to the removed edges.

Example 4.3. The interventional distribution $p(X_3|\,\mathrm{do}(X_2 := x_2))$ for the CGM in Figure 5 is obtained via probabilistic inference with respect to the postintervention graph G' where $X_1 \perp\!\!\!\perp X_2$:

$$p\big(X_3|\,\mathrm{do}(X_2 := x_2)\big) = \sum_{x_1 \in \mathcal{X}_1} p(x_1)p(X_3|x_1, x_2) \tag{4.3}$$

$$\neq \sum_{x_1 \in \mathcal{X}_1} p(x_1|x_2)p(X_3|x_1, x_2) = p(X_3|x_2). \tag{4.4}$$

It differs from the conditional $p(X_3|x_2)$ for which inference in done over G where $X_1 \not\perp\!\!\!\perp X_2$. Note the marginal $p(x_1)$ in (4.3), in contrast to the conditional $p(x_1|x_2)$ in (4.4): this is precisely the link which is broken by the intervention $\mathrm{do}(X_2 := x_2)$, see Figure 5b. The right-

hand side of (4.3) is an example of covariate adjustment: it controls for the confounder X_1 of the causal effect of X_2 on X_3, see Section 9 for more details on adjustment and computing interventions.

CGMs have been widely used in constraint- and score-based approaches to causal discovery [47,144] which we will discuss in Section 7. Due to their conceptual simplicity, they are a useful and intuitive model for reasoning about interventions. However, their capacity as a causal model is limited in that they do not support *counterfactual* reasoning, which is better addressed by the two causal modeling frameworks which we will discuss next.

Structural causal models (SCMs). Structural causal models, also referred to as functional causal models or nonparametric structural equation models, have ties to the graphical approach presented above, but rely on using directed functional parent–child relationships rather than causal conditionals. While conceptually simple in hindsight, this constituted a major step in the understanding of causality, as later expressed by [104, PAGE 104]:

> *"We played around with the possibility of replacing the parents–child relationship* $p(X_i|\mathbf{PA}_i)$ *with its functional counterpart* $X_i = f_i(\mathbf{PA}_i, U_i)$ *and, suddenly, everything began to fall into place: We finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities* $p(X_i|\mathbf{PA}_i)$ *with which we had been working so long in the study of Bayesian networks."*

Definition 4.4 (SCM). An SCM $\mathcal{M} = (\mathbf{F}, p_{\mathbf{U}})$ over a set $\mathbf{X}$ of n random variables $X_1, \ldots, X_n$ consists of (i) a set $\mathbf{F}$ of n assignments (the structural equations),

$$\mathbf{F} = \left\{X_i := f_i(\mathbf{PA}_i, U_i)\right\}_{i=1}^{n} \tag{4.5}$$

where f_i are deterministic functions computing each variable X_i from its causal parents $\mathbf{PA}_i \subseteq \mathbf{X} \setminus \{X_i\}$ and an exogenous noise variable U_i; and (ii) a joint distribution $p_{\mathbf{U}}(U_1, \ldots, U_n)$ over the exogenous noise variables.

The paradigm of SCMs views the processes f_i by which each observable X_i is generated from others as a physical mechanism. All randomness comes from the unobserved (also referred to as *unexplained*) noise terms U_i which capture both possible stochasticity of the process, as well as uncertainty due to unmeasured parts of the system.

Note also the assignment symbol ":=" which is used instead of an equality sign to indicate the asymmetry of the causal relationship: the left-hand side quantity is defined to take on the right-hand side value. For example, we cannot simply rewrite a structural equation $X_2 := f_2(X_1, U_2)$ as $X_1 = g(X_2, U_2)$ for some g, as would be the case for a standard (invertible) equation.

In parametric, linear form (i.e., with linear f_i), SCMs are also known as structural equation models and have a long history in path analysis [161] and economics [46,52].

Each SCM induces a corresponding causal graph via the input variables to the structural equations which is useful as a representation and provides a link to CGMs.

Definition 4.5 (Induced causal graph). The causal graph G induced by an SCM $\mathcal{M}$ is the directed graph with vertex set $\mathbf{X}$ and a directed edge from each vertex in $\mathbf{PA}_i$ to X_i for all i.

Example 4.6. Consider an SCM over $\mathbf{X} = \{X_1, X_2, X_3\}$ with some $p_{\mathbf{U}}(U_1, U_2, U_3)$ and

$$X_1 := f_1(U_1), \quad X_2 := f_2(X_1, U_2), \quad X_3 := f_3(X_1, X_2, U_3).$$

Following Definition 4.5, the induced graph then corresponds to G in Figure 5.

Definition 4.4 allows for a rich class of causal models, including those with cyclic causal relations and ones which do not obey the causal Markov condition (Definition 4.2) due to complex covariance structures between the noise terms. While work exists on such cyclic or confounded SCMs [13], it is common to make the following two assumptions.

Assumption 4.7 (Acyclicity). The induced graph G is a DAG: it does not contain cycles.

Assumption 4.8 (Causal sufficiency/no hidden confounders). The U_i are jointly independent, i.e., their distribution factorizes, $p_{\mathbf{U}}(U_1, \ldots, U_n) = p_{U_1}(U_1) \times \cdots \times p_{U_n}(U_n)$.

Assumption 4.7 implies[5] the existence of a well-defined, unique (observational) distribution over $\mathbf{X}$ from which we can draw via *ancestral sampling*:[6] first, we draw the noise variables from $p_{\mathbf{U}}$, and then we iteratively compute the corresponding X_i's in topological order of the induced DAG (i.e., starting at the root node of the graph), substituting previously computed X_i into the structural equations where necessary. Formally, the (observational) distribution $p(X_1, \ldots, X_n)$ induced by an SCM under Assumption 4.7 is defined as the push-forward of the noise distribution $p_{\mathbf{U}}$ through the structural equations $\mathbf{F}$. Under Assumption 4.8, the causal conditionals are thus given by

$$p(X_i | \mathbf{PA}_i = \mathbf{pa}_i) := p_{U_i}\left(f_{\mathbf{pa}_i}^{-1}(X_i)\right) \quad \text{for } i = 1, \ldots, n, \tag{4.6}$$

where $f_{\mathbf{pa}_i}^{-1}(X_i)$ denotes the preimage of X_i under f_i for fixed $\mathbf{PA}_i = \mathbf{pa}_i$.

Assumption 4.8 rules out the existence of hidden confounders because any unmeasured variables affecting more than one of the X_i simultaneously would constitute a dependence between some of the noise terms (which account for any external, or exogenous, influences not explained by the observed X_i). In combination with Assumption 4.7, Assumption 4.8 (also known as *causal sufficiency*) thus ensures that the distribution induced by an SCM factorizes according to its induced causal graph as in (4.1). In other words, it guarantees that the causal Markov condition is satisfied with respect to the induced causal graph [104]. Below, unless explicitly stated otherwise, we will assume causal sufficiency.

Due to the conceptual similarity between interventions and the assignment character of structural equations, the computation of interventional distributions fits in naturally

5 Acyclicity is a sufficient, but not a necessary condition.

6 Since neither $\mathbf{F}$ nor p are known a priori, ancestral sampling should be seen as a hypothetical sampling procedure; inference and learning are still necessary in general.

into the SCM framework. To model an intervention, we simply replace the corresponding structural equation and consider the resulting entailed distribution.

Definition 4.9 (Interventions in SCMs). An intervention $\mathrm{do}(X_i := x_i)$ in an SCM $\mathcal{M} = (\mathbf{F}, p_U)$ is modeled by replacing the ith structural equation in $\mathbf{F}$ by $X_i := x_i$, yielding the intervened SCM $\mathcal{M}^{\mathrm{do}(X_i:=x_i)} = (\mathbf{F}', p_U)$. The interventional distribution $p(\mathbf{X}_{-i} | \mathrm{do}(X_i := x_i))$, where $\mathbf{X}_{-i} = \mathbf{X} \setminus \{X_i\}$, and intervention graph G' are those induced by $\mathcal{M}^{\mathrm{do}(X_i:=x_i)}$.

This way of handling interventions coincides with that for CGMs, e.g., after performing $\mathrm{do}(X_2 := x_2)$ in Example 4.6, X_1 no longer appears in the structural equation for X_2, and the edge $X_1 \to X_2$ hence disappears in the intervened graph, as is the case for G' in Figure 5.

In contrast to CGMs, SCMs also provide a framework for *counterfactual reasoning*. While (i) observations describe what is passively seen or measured and (ii) interventions describe active external manipulation or experimentation, (iii) counterfactuals are statements about what would or could have been, given that something else was in fact observed. These three modes of reasoning are sometimes referred to as the three rungs of the "ladder of causation" [107]. As an example, consider the following counterfactual query:

> *Given that patient X received treatment A and his/her health got worse, what would have happened if he/she had been given treatment B instead, all else being equal?*

The "all else being equal" part highlights the difference between interventions and counterfactuals: observing the factual outcome (i.e., what actually happened) provides information about the background state of the system (as captured by the noise terms in SCMs) which can be used to reason about alternative, counterfactual, outcomes. This differs from an intervention where such background information is not available. For example, observing that treatment A did not work may tell us that the patient has a rare condition and that treatment B would have therefore worked. However, given that treatment A has been prescribed, the patient's condition may have changed, and B may no longer work in a future intervention.

Note that counterfactuals cannot be observed empirically by their very definition and are therefore unfalsifiable. Some therefore consider them unscientific [115] or at least problematic [26]. On the other hand, humans seem to perform counterfactual reasoning in practice, developing this ability in early childhood [14].

Counterfactuals are computed in SCMs through the following three-step procedure:

1. Update the noise distribution to its posterior given the observed evidence ("abduction").
2. Manipulate the structural equations to capture the hypothetical intervention ("action").
3. Use the modified SCM to infer the quantity of interest ("prediction").

Definition 4.10 (Counterfactuals in SCMs). Given evidence $\mathbf{X} = \mathbf{x}$ observed from an SCM $\mathcal{M} = (\mathbf{F}, p_{\mathbf{U}})$, the counterfactual SCM $\mathcal{M}^{\mathbf{X}=\mathbf{x}}$ is obtained by updating $p_{\mathbf{U}}$ with its posterior: $\mathcal{M}^{\mathbf{X}=\mathbf{x}} = (\mathbf{F}, p_{\mathbf{U}|\mathbf{X}=\mathbf{x}})$. Counterfactuals are then computed by performing interventions in the counterfactual SCM $\mathcal{M}^{\mathbf{X}=\mathbf{x}}$, see Definition 4.9.

Note that while computing interventions only involved manipulating the structural equations, counterfactuals also involve updating the noise distribution, highlighting the conceptual difference between the two. Updating p_U requires knowledge of the interaction between noise and observed variables, i.e., of the structural equations, which explains why additional assumptions are necessary. Note that the updated noise variables no longer need to be independent, even if the original system was causally sufficient (Assumption 4.8).

Example 4.11 (Computing counterfactuals with SCMs). Consider an SCM $\mathcal{M}$ defined by

$$X := U_X, \quad Y := 3X + U_Y, \quad U_X, U_Y \sim \mathcal{N}(0, 1). \tag{4.7}$$

Suppose we observe $X = 2$ and $Y = 6.5$ and want to answer the counterfactual "what would Y have been, had $X = 1$?," i.e., we are interested in $p(Y_{X=1}|X = 2, Y = 6.5)$. Updating the noise using the observed evidence via (4.7), we obtain the counterfactual SCM $\mathcal{M}^{X=2,Y=6.5}$,

$$X := U_X, \quad Y := 3X + U_Y, \quad U_X \sim \delta(2), \quad U_Y \sim \delta(0.5), \tag{4.8}$$

where $\delta(\cdot)$ denotes the Dirac delta measure. Performing the intervention $\mathrm{do}(X := 1)$ in (4.8) then gives the result $p(Y_{X=1}|X = 2, Y = 6.5) = \delta(3.5)$, i.e., "$Y$ would have been 3.5." This differs from the interventional distribution $p(Y\,|\,\mathrm{do}(X = 1)) = \mathcal{N}(3, 1)$, since the factual observation helped determine the background state ($U_X = 2$, $U_Y = 0.5$).

The SCM viewpoint is intuitive and lends itself well to studying restrictions on function classes to enable induction (Section 2). For this reason, we will mostly focus on SCMs in the subsequent sections.

Potential outcomes (PO). The potential outcomes framework was initially proposed by Neyman [98] for randomized studies [31], and later popularized and extended to observational settings by Rubin [124] and others. It is popular within statistics and epidemiology and perhaps best understood in the context of the latter. This is also reflected in its terminology: in the most common setting, we consider a binary treatment variable T, with $T = 1$ and $T = 0$ corresponding to treatment and control, respectively, whose causal effect on an outcome variable Y (often a measure of health) is of interest.

One interpretation of the PO framework consistent with its roots in statistics is to view *causal inference as a missing data problem*. In the PO framework, for each individual (or unit) i and treatment value t there is a PO, or potential response, denoted $Y_i(t)$ capturing what would happen if individual i received treatment t. The POs are considered deterministic quantities in the sense that for a given individual i, $Y_i(1)$ and $Y_i(0)$ are fixed and all randomness in the realized outcome Y_i stems from randomness in the treatment assignment,

$$Y_i = TY_i(1) + (1 - T)Y_i(0). \tag{4.9}$$

TABLE 1

Causal inference as a missing data problem: for each individual i (rows), only the PO $Y_i(T_i)$ corresponding to the assigned treatment T_i is observed; the other PO is a counterfactual. Hence, the unit-level causal effect $\tau_i = Y_i(1) - Y_i(0)$ is unidentifiable.

i	T_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	7	**?**	?
2	0	**?**	8	?
3	1	3	**?**	?
4	1	6	**?**	?
5	0	**?**	4	?
6	0	**?**	1	?

To decide whether patient i should receive treatment, we need to reason about the *individualized treatment effect* (ITE) τ_i as captured by the difference of the two POs.

Definition 4.12 (ITE). The ITE for individual i under a binary treatment is defined as

$$\tau_i = Y_i(1) - Y_i(0). \tag{4.10}$$

The *"fundamental problem of causal inference"* [51] is that only one of the POs is ever observed for each i. The other, unobserved PO becomes a counterfactual,

$$Y_i^{\text{CF}} = (1 - T)Y_i(1) + TY_i(0). \tag{4.11}$$

Consequently, τ_i can never be measured or computed from data, i.e., it is not identifiable (without further assumptions), as illustrated in Table 1.

Implicit in the form of (4.9) and (4.11) are the following two assumptions.

Assumption 4.13 (Stable unit treatment value; SUTVA). The observation on one unit should be unaffected by the particular assignment of treatments to the other units [23].

Assumption 4.14 (Consistency). If individual i receives treatment t, then the observed outcome is $Y_i = Y_i(t)$, i.e., the potential outcome for t.

Assumption 4.13 is usually understood as (i) units do not interfere, and (ii) there is only one treatment level per group (treated or control) leading to well-defined POs [61]. It can be violated, e.g., through (i) population dynamics such as herd immunity from vaccination or (ii) technical errors or varying within-group dosage, respectively. However, for many situations such as controlled studies it can be a reasonable assumption, and we can then view different units as independent samples from a population.

So far, we have considered POs for a given unit as deterministic quantities. However, most times it is impossible to fully characterize a unit, e.g., when dealing with complex subjects such as humans. Such lack of complete information introduces uncertainty, so that

POs are often instead treated as random variables. This parallels the combination of deterministic structural equations with exogenous noise variables in SCMs.[7] Indeed, there is a equivalence between POs and SCMs [104]:

$$Y_i(t) = Y \mid \mathrm{do}(T := t) \quad \text{in an SCM with } \mathbf{U} = \mathbf{u}_i\,,$$

An individual in the PO framework thus corresponds to a particular instantiation of the U_j in an SCM: the outcome is deterministic given $\mathbf{U}$, but since we do not observe $\mathbf{u}_i$ (nor can we characterize a given individual based on observed covariates), the counterfactual outcome is treated as a random variable. In practice, all we observe is a featurized description $\mathbf{x}_i$ of an individual i and have to reason about expected POs, $\mathbb{E}[Y(1), Y(0)|\mathbf{x}_i]$.

Another common assumption is that of no hidden confounders which we have already encountered in form of the causal Markov condition (Definition 4.2) for CGMs and causal sufficiency (Assumption 4.8) for SCMs. In the PO framework this becomes no hidden confounding between treatment and outcome and is referred to as (conditional) ignorability.

Assumption 4.15 (Conditional ignorability). Given a treatment $T \in \{0, 1\}$, potential outcomes $Y(0), Y(1)$, and observed covariates $\mathbf{X}$, we have

$$Y(0) \perp\!\!\!\perp T \mid \mathbf{X} \quad \text{and} \quad Y(1) \perp\!\!\!\perp T \mid \mathbf{X}. \tag{4.12}$$

The PO framework is tailored toward studying the (confounded) effect of a typically binary treatment variable on an outcome and is mostly used for causal reasoning, i.e., estimating individual and population level causal effects (Section 9). In this context, it is sometimes seen as an advantage that an explicit graphical representation is not needed. At the same time, the lack of a causal graph and the need for special treatment and outcome variables make POs rather unsuitable for causal discovery where other frameworks prevail.

5. INDEPENDENT CAUSAL MECHANISMS

We now return to the disentangled factorization (4.1) of the joint distribution $p(X_1, \ldots, X_n)$. This factorization according to the causal graph is always possible when the U_i are independent, but we will now consider an additional notion of independence relating the factors in (4.1) to one another.

Consider a dataset that consists of altitude A and average annual temperature T of weather stations [111]. Variables A and T are correlated, which we believe is due to the fact that the altitude has a causal effect on the temperature. Suppose we had two such datasets, one for Austria and one for Switzerland. The two joint distributions may be rather different, since the marginal distributions $p(A)$ over altitudes will differ. The conditionals $p(T|A)$, however, may be rather similar, since they follow from physical mechanisms generating temperature from altitude. The causal factorization $p(A)\,p(T|A)$ thus contains a component $p(T|A)$ that

7 When all noise variables in an SCM are fixed, the other variables are uniquely determined; without complete background knowledge, on the other hand, they are random.

generalizes across countries, while the entangled factorization $p(T)p(A|T)$ does not. Cum grano salis, the same applies when we consider interventions in a system. For a model to correctly predict the effect of interventions, it needs to be robust with respect to generalizing from an observational distribution to certain *interventional* distributions.

One can express the above insights as follows [111, 128]:

Principle 5.1 (Independent causal mechanisms (ICM)). *The causal generative process of system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

This principle subsumes several notions important to causality, including separate intervenability of causal variables, modularity and autonomy of subsystems, and invariance [104, 111]. If we have only two variables, it reduces to an independence between the cause distribution and the mechanism producing the effect distribution from the cause distribution.

Applied to the causal factorization (4.1), the principle tells us that the factors should be independent in two senses:

(*influence*) changing (or intervening upon) one mechanism $p(X_i|\mathbf{PA}_i)$ does not change the other mechanisms $p(X_j|\mathbf{PA}_j)$ $(i \neq j)$, and

(*information*) knowing some mechanisms $p(X_i|\mathbf{PA}_i)$ $(i \neq j)$ does not give us information about a mechanism $p(X_j|\mathbf{PA}_j)$.

We view any real-world distribution as a product of causal mechanisms. A change in such a distribution (e.g., when moving from one setting/domain to a related one) will always be due to changes in at least one of those mechanisms. Consistent with Principle 5.1, we hypothesize [133]:

Principle 5.2 (Sparse mechanism shift (SMS)). *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization* (4.1)*, i.e., they should usually not affect all factors simultaneously.*

In contrast, if we consider a noncausal factorization, e.g., (4.2), then many terms will be affected simultaneously as we change one of the physical mechanisms responsible for a system's statistical dependences. Such a factorization may thus be called *entangled*. The notion of disentanglement has recently gained popularity in machine learning [9, 50, 83, 147], sometimes loosely identified with statistical independence. The notion of invariant, autonomous, and independent mechanisms has appeared in various guises throughout the history of causality research, see [1, 104, 111, 133].

Measures of dependence of mechanisms. Note that the dependence of two mechanisms $p(X_i|\mathbf{PA}_i)$ and $p(X_j|\mathbf{PA}_j)$ does not coincide with the statistical dependence of the random variables X_i and X_j. Indeed, in a causal graph, many of the random variables will be dependent even if all the mechanisms are independent.

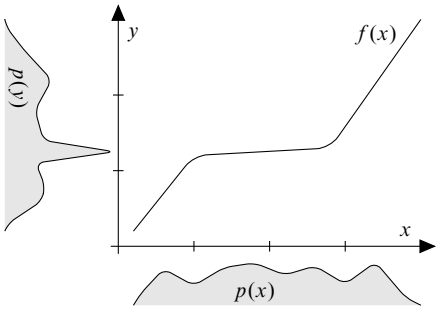


FIGURE 6
If $p(x)$ and f are chosen independently, then peaks of $p(y)$ tend to occur in regions where f has small slope. Hence $p(y)$ contains information about f^{-1} (figure from **[111]**).

Consider two variables and structural assignments $X := U$ and $Y := f(X)$, i.e., the cause X is a noise variable (with density $p(x)$), and the effect Y is a deterministic function of the cause. Let us, moreover, assume that the ranges of X and Y are both $[0, 1]$, and f is strictly monotonically increasing. The ICM principle then reduces to an independence of $p(x)$ and f. Let us consider $p(x)$ and the derivative f' as random variables on the probability space $[0, 1]$ with Lebesgue measure, and use their correlation as a measure of dependence of mechanisms. It can be shown that for $f \neq \mathrm{id}$, independence of $p(x)$ and f' implies dependence between $p(y)$ and $(f^{-1})'$ (see Figure 6). Other measures are possible and admit information-geometric interpretations. Intuitively, under the ICM assumption (Principle 5.1), the "irregularity" of the effect distribution becomes a *sum* of (i) irregularity already present in the input distribution and (ii) irregularity introduced by the mechanism f, i.e., the irregularities of the two mechanisms add up rather than (partly) compensating each other. This would not be the case in the opposite ("anticausal") direction (for details, see **[68]**). Other dependence measures have been proposed for high-dimensional linear settings and time series **[12, 63, 67, 136]**.

Algorithmic independence. So far, we have discussed links between causal and statistical structures. The fundamental of the two is the causal structure, since it captures the physical mechanisms that generate statistical dependences in the first place. The statistical structure is an epiphenomenon that follows if we make the unexplained variables random. It is awkward to talk about the (statistical) information contained in a mechanism, since deterministic functions in the generic case neither generate nor destroy information. This motivated us to devise an algorithmic model of causal structures in terms of Kolmogorov complexity **[65]**. The Kolmogorov complexity (or algorithmic information) of a bit string is essentially the length of its shortest compression on a Turing machine, and thus a measure of its information content. Independence of mechanisms can be defined as vanishing mutual algorithmic information: two conditionals are considered independent if knowing (the shortest compression of) one does not help achieve a shorter compression of the other one.

Algorithmic information theory provides a natural framework for nonstatistical graphical models. Just like statistical CGMs are obtained from SCMs by making the unexplained variables U_i random, we obtain algorithmic CGMs by making the U_i bit strings (jointly independent across nodes) and viewing the node X_i as the output of a fixed Turing machine running program U_i with input $\mathbf{PA}_i$. Similar to the statistical case, one can define a local causal Markov condition, a global one in terms of d-separation, and a decomposition of the joint Kolmogorov complexity in analogy to (4.1), and prove that they are implied by the SCM [65]. This approach shows that causality is not intrinsically bound to statistics: causality is about *mechanisms* governing flow of information which may or may not be statistical.

The assumption of algorithmically independent mechanisms has interesting implications for physics: it implies the second law of thermodynamics (i.e., the arrow of time). Consider a process where an incoming ordered beam of photons (the cause) is scattered by an object (the mechanism). Then the outgoing beam (the effect) contains information about the object. Microscopically, the time evolution is reversible; however, the photons contain information about the object only *after* the scattering. What underlies Loschmidt's paradox [86]?

The asymmetry can be explained by applying the ICM Principle 5.1 to initial state and system dynamics, postulating that the two be algorithmically independent, i.e., knowing one does not allow a shorter description of the other. The Kolmogorov complexity of the system's state can then be shown to be nondecreasing under time evolution [62]. If we view Kolmogorov complexity as a measure of entropy, this means that the entropy of the state can only stay constant or increase, amounting to the second law of thermodynamics.

Note that the resulting state after time evolution is clearly *not* independent of the system dynamic: it is precisely the state that, when fed to the inverse dynamics, would return us to the original (ordered) state.

6. LEVELS OF CAUSAL MODELING

Coupled differential equations are the canonical way of modeling physical phenomena. They allow us to predict the future behavior of a system, to reason about the effect of interventions, and—by suitable averaging procedures—to predict *statistical* dependences that are generated by a coupled time evolution. They also allow us to gain insight into a system, explain its functioning, and, in particular, read off its causal structure.

Consider a coupled set of ordinary differential equations

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \tag{6.1}$$

with initial value $\mathbf{x}(t_0) = \mathbf{x}_0$. We assume that they correctly describe the physical mechanisms of a system.[8] The Picard–Lindelöf theorem states that, at least locally, if f is Lipschitz, there

8 In other words, they do not merely phenomenologically describe its time evolution without capturing the underlying mechanisms (e.g., due to unobserved confounding, or a form of coarse-graining that does not preserve the causal structure [123, 133]).

TABLE 2

A simple taxonomy of models. The most detailed model (top) is a mechanistic or physical one, usually in terms of differential equations. At the other end of the spectrum (bottom), we have a purely statistical model; this can be learned from data and is useful for predictions but often provides little insight beyond modeling associations between epiphenomena. Causal models can be seen as descriptions that lie in between, abstracting away from physical realism while retaining the power to answer certain interventional or counterfactual questions.

Model	Predict in i.i.d. setting	Predict under distribution shift/intervention	Answer counterfactual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

exists a unique solution $\mathbf{x}(t)$. This implies, in particular, that the immediate future of $\mathbf{x}$ is implied by its past values.

In terms of infinitesimal differentials dt and $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$, (6.1) reads

$$\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f\big(\mathbf{x}(t)\big). \tag{6.2}$$

From this, we can ascertain which entries of the vector $\mathbf{x}(t)$ cause the future of others $\mathbf{x}(t + dt)$, i.e., the causal structure.

Compared to a differential equation, a statistical model derived from the joint distribution of a set of (time-independent) random variables is a rather superficial description of a system. It exploits that some of the variables allow the prediction of others as long as the experimental conditions do not change. If we drive a differential equation system with certain types of noise, or if we average over time, statistical dependences between components of $\mathbf{x}$ may emerge, which can be exploited by machine learning. In contrast to the differential equation model, such a model does not allow us to predict the effect of interventions; however, its strength is that it can often be learned from data.

Causal modeling lies in between these two extremes. It aims to provide understanding and predict the effect of interventions. Causal discovery and learning tries to arrive at such models in a data-driven way, using only weak assumptions (see Table 2, from [111,133]).

While we may naively think that causality is always about time, most existing causal models do not (and need not) consider time. For instance, returning to our example of altitude and temperature, there is an underlying dynamical physical process that results in higher places tending to be colder. On the level of microscopic equations of motion for the involved particles, there is a temporal causal structure. However, when we talk about dependence or causality between altitude and temperature, we need not worry about the details of this temporal structure; we are given a dataset where time does not appear, and we can reason about how that dataset would look if we were to intervene on temperature or altitude.

Some work exists trying to build bridges between these different levels of description. One can derive SCMs that describe the interventional behavior of a coupled system that is in an equilibrium state and perturbed in an adiabatic way [96], with generalizations to oscillatory systems [122]. In this work, an SCM arises as a high-level abstraction of an underlying system of differential equations. It can only be derived if suitable high-level variables can be defined [123], which in practice may well be the exception rather than the rule.

7. CAUSAL DISCOVERY

Sometimes, domain knowledge or the temporal ordering of events can help constrain the causal relationships between variables, e.g., we may know that certain attributes like age or sex are not caused by others; treatments influence health outcomes; and events do not causally influence their past. When such domain knowledge is unavailable or incomplete, we need to perform *causal discovery*: infer which variables causally influence which others, i.e., learn the causal structure (e.g., a DAG) from data. Since experiments are often difficult and expensive to perform while observational (i.e., passively collected) data is abundant, causal discovery from observational data is of particular interest.

As discussed in Section 3 in the context of the Common Cause Principle 3.2, the case where we have two variables is already difficult since the same dependence can be explained by multiple different causal structures. One might thus wonder if the case of more observables is completely hopeless. Surprisingly, this is not the case: the problem becomes easier (in a certain sense) because there are nontrivial conditional independence properties [25, 35, 145] implied by a causal structure. We first review two classical approaches to the multivariate setting before returning to the two-variable case.

Constraint-based methods. Constraint-based approaches to causal discovery test which (conditional) independences can be inferred from the data and then try to find a graph which implies them. They are therefore also known as independence-based methods. Such a procedure requires a way of linking properties of the data distribution p to properties of the underlying causal graph G. This link is known as the faithfulness assumption.

Assumption 7.1 (Faithfulness). The only (conditional) independences satisfied by p are those implied by the causal Markov condition (Definition 4.2).

Faithfulness can be seen as the converse of the causal Markov condition. Together, they constitute a one-to-one correspondence between graphical separation in G and conditional independence in p. While the causal Markov condition is satisfied by construction, faithfulness is an assumption which may be violated. A classical example for a violation of faithfulness is when causal effects along different paths cancel.

Example 7.2 (Violation of faithfulness). Consider the SCM from Example 4.6 and let

$$X_1 := U_1, \quad X_2 := \alpha X_1 + U_2, \quad X_3 := \beta X_1 + \gamma X_2 + U_3$$

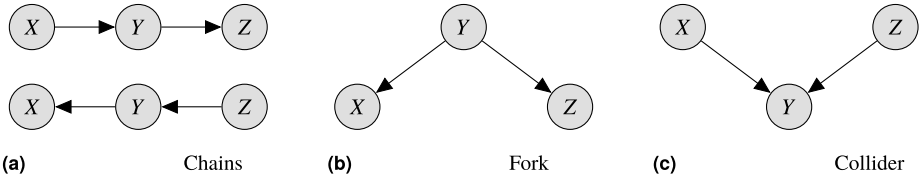


FIGURE 7

Illustration of Markov equivalence using common graph motifs. The chains in (a) and the fork in (b) all imply the relation $X \perp\!\!\!\perp Z \mid Y$ (and no others). They thus form a Markov equivalence class, meaning they cannot be distinguished using conditional independence testing alone. The collider, or v-structure, in (c) implies $X \perp\!\!\!\perp Z$ (but $X \not\perp\!\!\!\perp Z \mid Y$) and forms its own Markov equivalence class, so it can be uniquely identified from observational data. For this reason, v-structures are helpful for causal discovery. It can be shown that two graphs are Markov equivalent if and only if they share the same skeleton and v-structures.

with $U_1, U_2, U_3 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. By substitution, we obtain $X_3 = (\beta + \alpha\gamma)X_1 + \gamma U_2 + U_3$. Hence $X_3 \perp\!\!\!\perp X_1$ whenever $\beta + \alpha\gamma = 0$, even though this independence is not implied by the causal Markov condition over the induced causal graph G, see Figure 5. Here, faithfulness is violated if the direct effect of X_1 on X_3 (β) and the indirect effect via X_2 ($\alpha\gamma$) cancel.

Apart from relying on faithfulness, a fundamental limitation to constraint-based methods is the fact that many different DAGs may encode the same d-separation/independence relations. This is referred to as Markov equivalence and illustrated in Figure 7.

Definition 7.3 (Markov equivalence). Two DAGs are said to be Markov equivalent if they encode the same d-separation statements. The set of all DAGs encoding the same d-separations is called a Markov equivalence class.

Constraint-based algorithms typically first construct an undirected graph, or skeleton, which captures the (conditional) independences found by testing, and then direct as many edges as possible using Meek's orientation rules [90]. The first step carries most of the computational weight and various algorithms have been devised to solve it efficiently.

The simplest procedure is implemented in the IC [109] and SGS [144] algorithms. For each pair of variables (X, Y), these search through all subsets $\mathbf{W}$ of the remaining variables to check whether $X \perp\!\!\!\perp Y \mid \mathbf{W}$. If no such set $\mathbf{W}$ is found, then X and Y are connected with an edge. Since this can be slow due to the large number of subsets, the PC algorithm [144] uses a much more efficient search procedure. It starts from a complete graph and then sequentially test only subsets of the neighbors of X or Y of increasing size, removing an edge when a separating subset is found. This neighbor search is no longer guaranteed to give the right result for causally insufficient systems, i.e., in the presence of hidden confounders. The FCI (short for fast causal inference) algorithm [144] addresses this setting, and produces a partially directed causal graph as output.

Apart from being limited to recovering a Markov equivalence class, constraint-based methods can suffer from statistical issues. In practice, datasets are finite, and conditional

independence testing is a notoriously difficult problem, especially if conditioning sets are continuous and multidimensional. So while, in principle, the conditional independences implied by the causal Markov condition hold true irrespective of the complexity of the functions appearing in an SCM, for finite datasets conditional independence testing is hard without additional assumptions [135]. Recent progress in (conditional) independence testing heavily relies on kernel function classes to represent probability distributions in reproducing kernel Hilbert spaces, see Section 2.

Score-based methods. Score-based approaches to causal discovery assign a score to each graph G from a set of candidate graphs (usually the set of all DAGs). The score S is supposed to reflect how well G explains the observed data $\mathbf{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, and we choose the graph $\hat{G}$ maximizing this score,

$$\hat{G} = \operatorname*{argmax}_G S(G|\mathbf{D}).$$

Various score functions have been proposed, but most methods assume a parametric model which factorizes according to G, parametrized by $\theta \in \Theta$. Two common choices are multinomial models for discrete data [22] and linear Gaussian models for continuous data [34]. For example, a penalized maximum likelihood approach using the BIC [134] as a score yields

$$S_{\mathrm{BIC}}(G|\mathbf{D}) = \log p(\mathbf{D}|G, \hat{\theta}^{\mathrm{MLE}}) - \frac{k}{2} \log m, \tag{7.1}$$

where k is the number of parameters and $\hat{\theta}^{\mathrm{MLE}}$ is the maximum likelihood estimate for θ to D in G. Note that k generally increases with the number of edges in G so that the second term in (7.1) penalizes complex graphs which do not lead to substantial improvements.

Another choice of score function is the marginal likelihood, or evidence, in a Bayesian approach to causal discovery, which requires specifying prior distributions over graphs and parameters, $p(G, \theta) = p(G)p(\theta|G)$. The score for G is then given by

$$S_{\mathrm{BAYES}}(G|\mathbf{D}) = p(\mathbf{D}|G) = \int_\Theta p(\mathbf{D}|G, \theta) p(\theta|G) \mathrm{d}\theta. \tag{7.2}$$

This integral is intractable in general, but can be computed exactly for some models such as a Dirichlet-multinomial under some mild additional assumptions [47, 48].

A major drawback of score-based approaches is the combinatorial size of the search space. The number of DAGs over n random variables grows superexponentially and can be computed recursively (to account for acyclicity constraints) [119]. For example, the number of DAGs for $n = 5$ and $n = 10$ nodes is 29281 and 4175098976430598143, respectively. Finding the best scoring DAG is NP-hard [20]. To overcome this problem, greedy search techniques can be applied, e.g., greedy equivalence search (GES) [21] which optimizes for the BIC.

In recent years, another class of methods has emerged that is based on assuming particular functional forms for the SCM assignments. Those arose from studying the cause-effect inference problem, as discussed below.

Cause–effect inference. In the case of only two variables, the ternary concept of conditional independences collapses and the causal Markov condition (Definition 4.2) thus has no nontrivial implications. However, we have seen in Section 5 that assuming an independence of mechanisms (Principle 5.1) lets us find asymmetries between cause and effect, and thus address the cause–effect inference problem previously considered unsolvable [68]. It turns out that this problem can be also addressed by making additional assumptions on function classes, as not only the graph topology leaves a footprint in the observational distribution, but so do the functions f_i in an SCM. Such assumptions are typical for machine learning, where it is well known that finite-sample generalization without assumptions on function classes is impossible, and where much attention is devoted to properties of function classes (e.g., priors or capacity measures), as discussed in Section 2.

Let us provide an intuition as to why assumptions on the functions in an SCM should help learn about them from data. Consider a toy SCM with only two observables $X \to Y$. In this case, the structural equations (4.5) turn into

$$X := U, \quad Y := f(X, V) \tag{7.3}$$

with noises $U \perp\!\!\!\perp V$. Now think of V acting as a random selector variable choosing from among a set of functions $\mathcal{F} = \{f_v(x) \equiv f(x, v) \mid v \in \mathrm{supp}(V)\}$. If $f(x, v)$ depends on v in a nonsmooth way, it should be hard to glean information about the SCM from a finite dataset, given that V is not observed and it randomly switches between arbitrarily different f_v.[9] This motivates restricting the complexity with which f depends on V. A natural restriction is to assume an *additive noise model*

$$X := U, \quad Y := f(X) + V. \tag{7.4}$$

If f in (7.3) depends smoothly on V, and if V is relatively well concentrated, this can be motivated by a local Taylor expansion argument. Such assumptions drastically reduce the effective size of the function class—without them, the latter could depend exponentially on the cardinality of the support of V.

Restrictions of function classes can break the symmetry between cause and effect in the two-variable case: one can show that given a distribution over X, Y generated by an additive noise model, one cannot fit an additive noise model in the opposite direction (i.e., with the roles of X and Y interchanged) [6,53,76,95,113]. This is subject to certain genericity assumptions, and notable exceptions include the case where U, V are Gaussian and f is linear. It generalizes results of [139] for linear functions, and it can be generalized to include nonlinear rescaling [164], cycles [94], confounders [64], and multivariable causal discovery [112]. There is now a range of methods that can detect causal direction better than chance [97].

9 Suppose X and Y are binary, and U, V are uniform Bernoulli variables, the latter selecting from $\mathcal{F} = \{\mathrm{id}, \mathrm{not}\}$ (i.e., identity and negation). In this case, the entailed distribution for Y is uniform, *independent* of X, even though we have $X \to Y$. We would be unable to discern $X \to Y$ from data. (This would also constitute a violation of faithfulness (Assumption 7.1).)

We have thus gathered some evidence that ideas from machine learning can help tackle causality problems that were previously considered hard. Equally intriguing, however, is the opposite direction: can causality help us improve machine learning?

Nonstationarity-based methods. The last family of causal discovery approaches we mention is based on ideas of nonstationarity and invariance [128]. These approaches do not apply to purely observational data collected in an i.i.d. setting. In contrast, they aim to leverage heterogeneity of data collected from different environments. The main idea is the following: since causal systems are modular in the sense of the ICM Principle 5.1, changing one of the independent mechanisms should leave the other components, or causal conditionals, unaffected (SMS Principle 5.2). A correct factorization of the joint distribution according to the underlying causal structure should thus be able to explain heterogeneity by localized changes in one (or few) of the mechanisms while the others remain invariant.

One of the first works to use this idea [150] analyzed which causal structures can be distinguished given data resulting from a set of mechanism changes. Recent work [55] additionally aims to learn a low-dimensional representation of the mechanism changes. Other works [110,120] have proposed methods for finding the direct causes of a given target variable. Using a recent result on identifiability of nonlinear ICA [59] which also relies on nonstationarity, a method for learning general nonlinear SCMs was proposed [93]. Here the idea is to train a classifier to discriminate between the true value of some nonstationarity variable (such as a time-stamp or environment indicator) and a shuffled version thereof.

8. IMPLICATIONS FOR MACHINE LEARNING

Semisupervised learning. Suppose our underlying causal graph is $X \to Y$, and we are trying to learn a mapping $X \to Y$. The causal factorization (4.1) for this case is

$$p(X, Y) = p(X)p(Y|X). \tag{8.1}$$

The ICM Principle 5.1 posits that the modules in a joint distribution's causal factorization do not inform or influence each other. This means that, in particular, $p(X)$ should contain no information about $p(Y|X)$, which implies that semisupervised learning [17] should be futile, as it is trying to use additional information about $p(X)$ (from unlabeled data) to improve our estimate of $p(Y|X = x)$. What about the opposite direction, is there hope that semisupervised learning should be possible in that case? It turns out the answer is yes, due to the work on cause–effect inference using the ICM Principle 5.1 [24]. It introduced a measure of dependence between the input and the conditional of output given input, and showed that if this dependence is zero in the causal direction, then it is strictly positive in the opposite direction. Independence of cause and mechanism in the causal direction thus implies that in the backward direction (i.e., for anticausal learning), the distribution of the input variable should contain information about the conditional of output given input, i.e., the quantity that machine learning is usually concerned with. This is exactly the kind of information that semisupervised learning requires when trying to improve the estimate of output given input

by using unlabeled inputs. This suggests that *semisupervised learning should be impossible for causal learning problems, but feasible otherwise*, in particular for anticausal ones. A metaanalysis of published semisupervised learning benchmark studies corroborated this prediction [128], and similar results apply for natural language processing [69]. These findings are intriguing since they provide insight into *physical* properties of learning problems, thus going beyond the methods and applications that machine learning studies usually provide.

Subsequent developments include further theoretical analyses [66,111] and a form of conditional semisupervised learning [156]. The view of semisupervised learning as exploiting dependences between a marginal $p(x)$ and a noncausal conditional $p(y|x)$ is consistent with the common assumptions employed to justify semisupervised learning [17,125].

Invariance and robustness. We have discussed the shortcomings of the i.i.d. assumption, which rarely holds true exactly in practice, and the fact that real-world intelligent agents need to be able to generalize not just within a single i.i.d. setting, but across related problems. This notion has been termed *out-of-distribution (o.o.d.) generalization*, attracting significant attention in recent years [133]. While most work so far has been empirical, statistical bounds would be desirable that generalize (2.4), including additional quantities measuring the distance between training and test distribution, incorporating meaningful assumptions [137]. Such assumptions are necessary [8], and could be causal, or related to invariance properties.

The recent phenomenon of "adversarial vulnerability" [148] shows that minuscule targeted violations of the i.i.d. assumption, generated by adding suitably chosen noise to images (imperceptible to humans), can lead to dangerous errors such as confusion of traffic signs. These examples are compelling as they showcase nonrobustnesses of artificial systems which are not shared by human perception. Our own perception thus exhibits invariance or robustness properties that are not easily learned from a single training set.

Early causal work related to domain shift [128] looked at the problem of learning from multiple cause–effect datasets that share a functional mechanism but differ in noise distributions. More generally, given (data from) multiple distributions, one can try to identify components which are robust, and find means to transfer them across problems [4, 36, 54, 163, 166]. According to the ICM Principle 5.1, invariance of conditionals or functions (also referred to as covariate shift in simple settings) should only hold in the causal direction, a reversal of the impossibility described for SSL.

Building on the work of [110,128], the idea of invariance for prediction has also been used for supervised learning [3, 87, 120]. In particular, "invariant risk minimization" (IRM) was proposed as an alternative to ERM, cf. (2.3).

9. CAUSAL REASONING

In contrast to causal discovery (Section 7), which aims to uncover the causal structure underlying a set of variables, *causal reasoning* starts from a known (or postulated) causal graph and answers causal queries of interest. While causal discovery often looks for qualitative relationships, causal reasoning usually aims to quantify them. This requires two steps:

(i) *identifying* the query, i.e., deriving an estimand for it that only involves observed quantities; and (ii) *estimating* this using data. Often, the quantities of interest can be described as treatment effects, i.e., contrasts between two interventions.

Definition 9.1 (Treatment effects). The conditional average treatment effect (CATE), conditioned on (a subset of) features $\mathbf{x}$, is defined as

$$\tau(\mathbf{x}) := \mathbb{E}\big[Y|\mathbf{x}, \mathrm{do}(T=1)\big] - \mathbb{E}\big[Y|\mathbf{x}, \mathrm{do}(T=0)\big] = \mathbb{E}\big[Y(1) - Y(0)|\mathbf{x}\big]. \tag{9.1}$$

The average treatment effect (ATE) is defined as the population average of the CATE,

$$\tau := \mathbb{E}\big[\tau(\mathbf{X})\big] = \mathbb{E}\big[Y \,|\, \mathrm{do}(T=1)\big] - \mathbb{E}\big[Y \,|\, \mathrm{do}(T=0)\big] = \mathbb{E}\big[Y(1) - Y(0)\big]. \tag{9.2}$$

While ITE (Definition 4.12) and CATE (9.1) are sometimes used interchangeably, there is a conceptual difference: ITE refers to the difference of two POs and is thus bound to an individual, while CATE applies to subpopulations, e.g., the CATE for females in their 40s. Since the ITE is fundamentally impossible to observe, it is often estimated by the CATE conditional on an individual's features $\mathbf{x}_i$ using suitable additional assumptions.

As is clear from Definition 9.1, the treatment effects we want to estimate involve interventional expressions. However, we usually only have access to observational data. Causal reasoning can thus be cast as answering interventional queries using observational data and a causal model. This involves dealing with confounders, both observed and unobserved.

Before discussing how to identify and estimate causal effects, we illustrate why causal assumptions are necessary using a well-known statistical phenomenon.

Simpson's paradox and Covid-19. *Simpson's paradox* refers to the observation that aggregating data across subpopulations may yield opposite trends (and thus lead to reversed conclusions) from considering subpopulations separately [142]. We observed a textbook example of this during the Covid-19 pandemic by comparing case fatality rates (CFRs), i.e., the proportion of confirmed Covid-19 cases which end in fatality, across different countries and age groups as illustrated in Figure 8 [154]: for *all* age groups, CFRs in Italy are *lower* than in China, but the *total* CFR in Italy is *higher*.

How can such a pattern be explained? The case demographic (see Figure 8, right) is rather different across the two countries, i.e., there is a statistical association between country and age. In particular, Italy recorded a much larger proportion of cases in older patients who are generally at higher risk of dying from Covid-19 (see Figure 8, left). While this provides a consistent explanation in a *statistical* sense, the phenomenon may still seem puzzling as it defies our *causal* intuition. Humans appear to naturally extrapolate conditional probabilities to read them as causal effects, which can lead to inconsistent conclusions and may leave one wondering: *how can the disease in Italy be less fatal for the young, less fatal for the old, but more fatal for the people overall?* It is for this reason that the reversal of (conditional) probabilities in Figure 8 is perceived as and referred to as a "paradox" [49, 105].

If we consider the country as treatment whose causal effect on fatality is of interest, then causal assumptions (e.g., in the form of a causal graph) are needed to decide how to

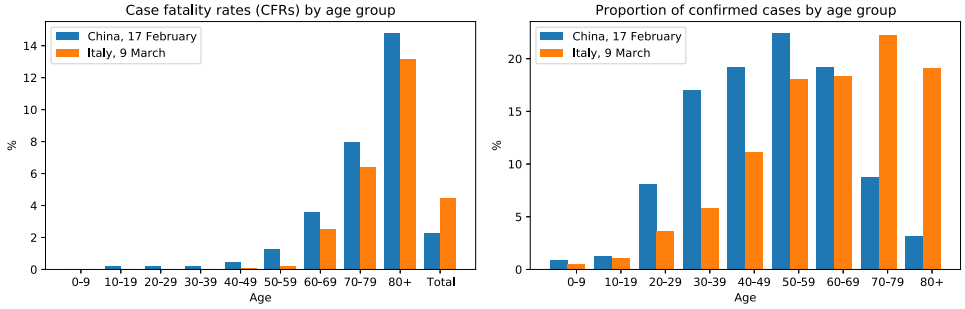


FIGURE 8

(Left) Covid-19 case fatality rates (CFRs) in Italy and China by age and in aggregate ("Total"), including all confirmed cases and fatalities up to the time of reporting in early 2020 (see legend): for all age groups, CFRs in Italy are lower than in China, but the total CFR in Italy is higher, an example of *Simpson's paradox*. (Right) The case demographic differs between countries: in Italy, most cases occurred in the older population (figure from [154]).

handle covariates such as age that are statistically associated with the treatment, e.g., whether to stratify by (i.e., adjust for) age or not. This also explains why randomized controlled trials (RCTs) [31] are the gold standard for causal reasoning: randomizing the assignment breaks any potential links between the treatment variable and other covariates, thus eliminating potential problems of bias. However, RCTs are costly and sometimes unethical to perform, so that causal reasoning often relies on observational data only.[10]

We first consider the simplest setting *without hidden confounders and with overlap*. We start with *identification* of treatment effects on the population level, and then discuss different techniques for *estimating* these from data.

Identification. In absence of unmeasured variables (i.e., without hidden confounding), and provided we know the causal graph, it is straight-forward to compute causal effects by adjusting for covariates. A principled approach to do so for any given graph was proposed by Robins [117] and is known as the *g-computation formula* (where the g stands for general). It is also known as *truncated factorisation* [104] or *manipulation theorem* [144]. It relies on the independence of causal mechanisms (Principle 5.1), i.e., the fact that intervening on a variable leaves the other causal conditionals in (4.1) unaffected:

$$p\big(X_1, \ldots, X_n | \operatorname{do}(X_i := x_i)\big) = \delta(X_i = x_i) \prod_{j \neq i} p(X_j | \mathbf{PA}_j). \tag{9.3}$$

From (9.3) the interventional distribution of interest can then be obtained by marginalization. This is related to the idea of graph surgery (see Figure 5), and leads to a set of three inference rules for manipulating interventional distributions known as *do-calculus* [104] that have been shown to be complete for identifying causal effects [56, 140].

10 For a treatment of more general types of data fusion and transportability of experimental findings across different populations, we refer to [5, 106].

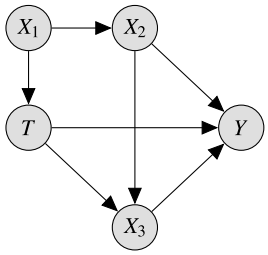


FIGURE 9

Treatment effect estimation with three observed covariates X_1, X_2, X_3: here, the valid adjustment sets for $T \to Y$ (see Proposition 9.3) are $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$. Including X_3 opens the *nondirected path* $T \to X_3 \leftarrow X_2 \to Y$ and lies on the directed path $T \to X_3 \to Y$, both of which can introduce bias.

Note that covariate adjustment may still be needed, even if there are no clear confounders directly influencing both treatment and outcome, as shown by the example in Figure 9.

Example 9.2. Applying the g-computation formula (9.3) to the setting of Figure 9, we obtain

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1) \sum_{x_3} p(x_3|t,x_2)\, p(y|t,x_2,x_3) \tag{9.4}$$

$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1)\, p(y|t,x_2) = \sum_{x_2} p(x_2)\, p(y|t,x_2) \tag{9.5}$$

$$\overset{\text{(a)}}{=} \sum_{x_1,x_2} p(x_1,x_2)\, p(y|t,x_1,x_2) \overset{\text{(b)}}{=} \sum_{x_1} p(x_1)\, p(y|t,x_1), \tag{9.6}$$

where the last line follows by using the following conditional independences implied by the graph: (a) $Y \perp\!\!\!\perp X_1 \mid \{T, X_2\}$, and (b) $X_2 \perp\!\!\!\perp T \mid X_1$.

Note that both the right-hand side in (9.5) and both sides in (9.6) take the form

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_{\mathbf{z}} p(\mathbf{z})\, p(y|t,\mathbf{z}). \tag{9.7}$$

In this case we call $\mathbf{Z}$ a *valid adjustment set* for the effect of T on Y. Here, $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$ are all valid adjustment sets, but it can be shown that, e.g., $\{X_1, X_3\}$ is not (see Figure 9). As computing the g-formula with many covariates can be cumbersome, graphical criteria for which subsets constitute valid adjustment sets are useful in practice, even in the absence of unobserved confounders.

Proposition 9.3 ([141]). *Under causal sufficiency, a set* $\mathbf{Z}$ *is a valid adjustment set for the causal effect of a singleton treatment* T *on an outcome* Y *(in the sense of* (9.7)*) if and only if the following two conditions hold: (i)* $\mathbf{Z}$ *contains no descendant of any node on a directed path from* T *to* Y *(except for descendants of* T *which are not on a directed path from* T *to* Y*); and (ii)* $\mathbf{Z}$ *blocks all non-directed paths from* T *to* Y*.*

Here, a path is called *directed* if all directed edges on it point in the same direction, and *nondirected* otherwise. A path is *blocked* (by a set of vertices $\mathbf{Z}$) if it contains a triple of

consecutive nodes connected in one of the following three ways: $A \to B \to C$ with $B \in \mathbf{Z}$, $A \leftarrow B \to C$ with $B \in \mathbf{Z}$, or $A \to B \leftarrow C$, where neither B nor any descendant of B is in $\mathbf{Z}$.

Two well-known types of adjustment set implied by Proposition 9.3 are *parent adjustment*, where $\mathbf{Z} = \mathbf{Pa}_T$; and the *backdoor criterion*, where $\mathbf{Z}$ is constrained to contain no descendants of T and to block all "back-door paths" from T to Y ($T \leftarrow \ldots Y$).

Note that Proposition 9.3 only holds singleton treatments (i.e., interventions on a single variable). For treatments $\mathbf{T}$ involving multiple variables, a slightly more complicated version of Proposition 9.3 can be given in terms of proper causal paths, and we refer to [102, 108] for details.

Let us briefly return to our earlier example of Simpson's paradox and Covid-19. Considering a plausible causal graph for this setting [154], we find that age A acts as a *mediator* $C \to A \to F$ of the causal effect of country C on fatality F (there is likely also a direct effect $C \to F$, potentially mediated by other, unobserved variables). If we are interested in the (total) causal effect of C on F (i.e., the overall influence of country on fatality), A should not be included for adjustment according to Proposition 9.3, and, subject to causal sufficiency, the total CFRs can be interpreted causally.[11] For another classic example of Simpson's paradox in the context of kidney stone treatment [18], on the other hand, the size of the stone acts as a *confounder* and thus needs to be adjusted for to obtain sound causal conclusions.

Valid covariate adjustment and the g-formula tell us how to compute interventions from the observational distribution when there are no hidden confounders. To actually identify causal effects from data, however, we need to also be able to *estimate* the involved quantities in (9.7). This is a problem if a subgroup of the population never (or always) receives a certain treatment. We thus need the additional assumption of a nonzero probability of receiving each possible treatment, referred to as *overlap*, or common support.

Assumption 9.4 (Overlap/common treatment support). For any treatment t and any configuration of features $\mathbf{x}$, it holds that $0 < p(T = t|\mathbf{X} = \mathbf{x}) < 1$.

The combination of overlap and ignorability (that is, no hidden confounders—see Assumption 4.15) is also referred to as *strong ignorability* and is a sufficient condition for identifying ATE and CATE: the absence of hidden confounders guarantees the existence of a valid adjustment set $\mathbf{Z} \subseteq \mathbf{X}$ for which $p(Y \,|\, \mathrm{do}(T = t), \mathbf{Z}) = p(Y|T = t, \mathbf{Z})$, and overlap guarantees that we can actually estimate the latter term for any $\mathbf{z}$ occurring with nonzero probability.[12]

Regression adjustment. Having identified a valid adjustment set (using Proposition 9.3), *regression adjustment* works by fitting a regression function $\hat{f}$ to $\mathbb{E}[Y|\mathbf{Z} = \mathbf{z}, T = t] = f(\mathbf{z}, t)$ using an observational sample $\{(y_i, t_i, \mathbf{z}_i)\}_{i=1}^m$. We can then use $\hat{f}$ to impute counterfactual outcomes as $\hat{y}_i^{\mathrm{CF}} = \hat{f}(\mathbf{z}_i, 1 - t_i)$ in order to estimate the CATE. The ATE is then

11 Mediation analysis [103] provides tools to tease apart and quantify the direct and indirect effects; the age-specific CFRs in Figure 8 then correspond to *controlled direct effects* [154].

12 The overlap assumption can thus be relaxed to hold for at least one valid adjustment set.

given by the population average and can be estimated as

$$\hat{\tau}_{\text{regression-adj.}} = \frac{1}{m_1} \sum_{i:t_i=1} \big(y_i - \hat{f}(\mathbf{z}_i, 0)\big) + \frac{1}{m_0} \sum_{i:t_i=0} \big(\hat{f}(\mathbf{z}_i, 1) - y_i\big), \tag{9.8}$$

where m_1 and m_0 are the number of observations from the treatment and control groups, respectively. Note the difference to the RCT estimator where no adjustment is necessary,

$$\hat{\tau}_{\text{RCT}} = \frac{1}{m_1} \sum_{i:t_i=1} y_i - \frac{1}{m_0} \sum_{i:t_i=0} y_i. \tag{9.9}$$

Matching and weighting approaches. While regression adjustment indirectly estimates ATE via CATE, matching and weighting approaches aim to estimate ATE directly. The general is idea to emulate the conditions of an RCT as well as possible.

Matching approaches work by splitting the population into subgroups based on feature similarity. This can be done on an individual level (so-called one-to-one or nearest neighbor matching) by matching each individual i with the most similar one, $j(i)$, from the opposite treatment group (i.e., $t_i \neq t_{j(i)}$). The difference of their outcomes, $y_i - y_{j(i)}$, is then considered as a sample of the ATE, and their average taken as an estimate thereof,

$$\hat{\tau}_{\text{NN-matching}} = \frac{1}{m_1} \sum_{i:t_i=1} (y_i - y_{j(i)}) + \frac{1}{m_0} \sum_{i:t_i=0} (y_{j(i)} - y_i). \tag{9.10}$$

Alternatively, the population can be split into larger subgroups with similar features (so-called strata). Each stratum is then treated as an independent RCT. If there are K strata containing $m_1, \ldots, m_K$ observations each, the stratified ATE estimator is

$$\hat{\tau}_{\text{stratified}} = \frac{\sum_{k=1}^{K} m_k \hat{\tau}_{\text{RCT}}^{(k)}}{\sum_{k=1}^{K} m_k}, \tag{9.11}$$

where $\hat{\tau}_{\text{RCT}}^{(k)}$ is the estimator from (9.9) applied to observation in the kth stratum.

Weighting approaches, on the other hand, aim to counteract the confounding bias by reweighting each observation to make the population more representative of an RCT. This means that underrepresented treatment groups are upweighted and overrepresented ones downweighted. An example is the inverse probability weighting (IPW) estimator,

$$\hat{\tau}_{\text{IPW}} = \frac{1}{m_1} \sum_{i:t_i=1} \frac{y_i}{p(T=1|\mathbf{Z}=\mathbf{z}_i)} - \frac{1}{m_0} \sum_{i:t_i=0} \frac{y_i}{p(T=0|\mathbf{Z}=\mathbf{z}_i)}. \tag{9.12}$$

The treatment probability $p(T = 1|\mathbf{Z})$ is also known as *propensity score*. While from a theoretical point of view $\mathbf{Z}$ should be a valid adjustment set, practitioners sometimes use all covariates to construct a propensity score.

Propensity score methods. To overcome the curse of dimensionality and gain statistical efficiency in high-dimensional, low-data regimes, propensity scores can be a useful tool, because covariates and treatment are rendered conditionally independent, $T \perp\!\!\!\perp \mathbf{Z} \mid s(\mathbf{z})$, by the propensity score $s(\mathbf{z}) := p(T = 1|\mathbf{Z} = \mathbf{z})$ [121]. Instead of adjusting for large feature sets or performing matching in high-dimensional spaces, the scalar propensity score can be used instead. Applying this idea to the above methods gives rise to *propensity score adjustment*

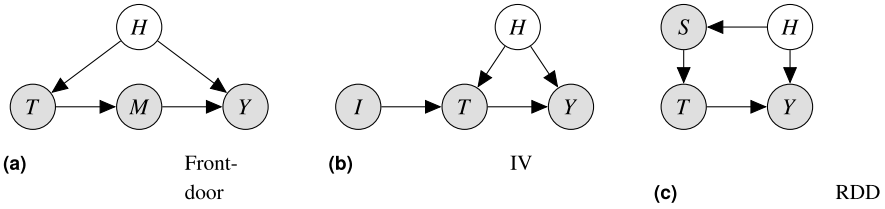


FIGURE 10

Overview of special settings which allow estimating causal effects of treatment T on outcome Y when the strong ignorability assumption (no hidden confounding or overlap) does not hold. In (a) the hidden confounder H is dealt with by means of an observed mediator M, while (b) relies on an instrumental variable (IV) which is independent of H. (c) In a regression discontinuity design (RDD), treatment assignment is a threshold function of some observed decision score S so that there is no overlap between treatment groups.

and *propensity score matching*. For the latter, the difference in propensity scores is used as similarity between instances to find nearest neighbors or to define strata.

While simplifying in one respect, the propensity score needs to be estimated from data which is an additional source of error. The standard approach for this is to estimate $s(\mathbf{z})$ by logistic regression, but more sophisticated methods are also possible. However, propensity score methods still rely on having identified a valid adjustment set $\mathbf{Z}$ to give unbiased results. Using all covariates to estimate s, without checking for validity as an adjustment set, can thus lead to wrong results.

Next, we consider the case of causal reasoning with *unobserved confounders*. While it is not possible to identify causal effects in the general case, we will discuss two particular situations in which ATE can still be estimated. These are shown in Figures 10a and 10b.

Front-door adjustment. The first situation in which identification is possible even though a hidden variable H confounds the effect between treatment and outcome is known as *front-door adjustment*. The corresponding causal graph is shown in Figure 10a. Front-door adjustment relies on the existence of an observed variable M which blocks all directed paths from T to Y, so that T only causally influences Y through M. For this reason M is also called a *mediator*. The other important assumption is that the hidden confounder does not influence the mediator other than through the treatment T, i.e., $M \perp\!\!\!\perp H \mid T$. In this case, and provided $p(t, m) > 0$ for all t and m, the causal effect of T on Y is identifiable and is given by the following.

Proposition 9.5 (Front-door adjustment). *For the causal graph in Figure 10a it holds that*

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_m p(m|t) \sum_{t'} p(t')\,p(y|m,t'). \tag{9.13}$$

We give a sketch of the derivation, and refer to [104] for a proof using the rules of do-calculus. Since M mediates the causal effect of T on Y, we have that

$$p\big(y|\,\mathrm{do}(t)\big) = \sum_m p\big(m|\,\mathrm{do}(t)\big)\,p\big(y|\,\mathrm{do}(m)\big). \tag{9.14}$$

Since there are no back-door paths from T to M, we have $p(m|\,\mathrm{do}(t)) = p(m|t)$.

Moreover, $\{T\}$ is a valid adjustment set for the effect of M on Y by Proposition 9.3, so

$$p\big(y|\,\mathrm{do}(m)\big) = \sum_{t'} p(t')\,p(y|m,t'). \tag{9.15}$$

Substituting into (9.14) then yields expression (9.13).

We point out that the setting presented here is only the simplest form of front-door adjustment which is sufficient to convey the main idea. It can be amended to include observed covariates $\mathbf{X}$ as well, as long as the conditions on the mediator remain satisfied.

Instrumental variables (IVs). The second setting for causal reasoning with hidden confounders is based on the idea of instrumental variables [2, 29, 160], see Figure 10b. The IV approach relies on the existence of a special observed variable I, called instrument.

Definition 9.6 (IV). A variable I is a valid instrument for estimating the effect of treatment T on outcome Y confounded by a hidden variable H if all of the following three conditions hold: (i) $I \perp\!\!\!\perp H$; (ii) $I \not\perp\!\!\!\perp T$; and (iii) $I \perp\!\!\!\perp Y \mid T$.

Condition (i) states that the instrument is independent of any hidden confounders H. Since this assumption cannot be tested, background knowledge is necessary to justify the use of a variable as IV in practice. Conditions (ii) and (iii) state that the instrument is correlated with treatment, and only affects the outcome through T, and are referred to as relevance and exclusion restriction, respectively.

Given a valid IV, we apply a two-stage procedure: first obtain an estimate $\hat{T}$ of the treatment variable T that is independent of H by predicting T from I. Having thus created an unconfounded version of the treatment, a regression of Y on $\hat{T}$ then reveals the correct causal effect. We demonstrate this idea for a simple linear model with continuous treatment variable where the causal effect can be obtained by two-stage least squares (2SLS).

Example 9.7 (Linear IV with 2SLS). Consider the linear SCM defined by

$$T := aI + bH + U_T, \quad Y := cH + dT + U_Y,$$

with U_T, U_Y independent noise terms. Then, since $I \perp\!\!\!\perp H$, linear regression of T on I recovers the coefficient a via $\hat{T} = aI$. Substituting for T in the structural equation for Y gives

$$Y := daI + (c + bd)H + U_Y + dU_T.$$

A second linear regression of Y on $\hat{T} = aI$ recovers the causal effect d because $(I \perp\!\!\!\perp H) \implies (\hat{T} \perp\!\!\!\perp H)$, whereas a naive regression of Y on T would give a different result, as $T \not\perp\!\!\!\perp H$.

Instrumental variables have been studied extensively and more sophisticated versions than the simple example above exist, allowing for nonlinear interactions and observed covariates.

Having discussed some special settings to deal with hidden confounding, we briefly present a technique to deal with violations of the overlap assumption.

Regression discontinuity design. In a *regression discontinuity design* (RDD), the treatment assignment mechanism behaves like a threshold function, i.e., the propensity score is discontinuous [60]. In the simplest setting, the assignment of treatment or control is determined by whether an *observed score* S is above a threshold s_0, $T := \mathbb{I}\{S \geq s_0\}$. This score in turn depends on other covariates which may or may not be observed. For example, patients may be assigned a risk score, and treatment is only prescribed if this score surpasses a given threshold. Since the score may be assigned by another institution, not all relevant covariates H are usually observed. However, it is assumed that the treatment decision only depends on the score, e.g., because doctors comply with the official rules. The causal graph for such a simple RDD setting is shown in Figure 10c. While the score S constitutes a valid adjustment set in principle, the problem with RDDs is the lack of overlap: patients with low scores are always assigned $T = 0$ and patients with high scores are always assigned $T = 1$. Because of this, covariate adjustment, matching, or weighting approaches do not apply. The general idea of an RDD is to overcome this challenge by comparing observations with score in a small neighborhood of the decision cut-off value s_0, motivated by the consideration that patients with close scores but on opposite sides of s_0 differ only in whether they received the treatment or not. For example, if the treatment cut-off value is 0.5 for a score in [0,1], then patients with scores of 0.49 and 0.51 are comparable and can be treated as samples from an RCT. An RDD (in its simplest form) thus focuses on differences in the regression function $\mathbb{E}[Y|S = s, T = t(s)] = f(s)$ for $s \in [s_0 - \varepsilon, s_0 + \varepsilon]$, where $\varepsilon > 0$ is small.

Half-sibling regression and exoplanet detection. We conclude this section with a real-world application performing causal reasoning in a confounded additive noise model. Launched in 2009, NASA's Kepler space telescope initially observed 150000 stars over four years, in search of exoplanet transits. These are events where a planet partially occludes its host star, causing a slight decrease in brightness, often orders of magnitude smaller than the influence of telescope errors. When looking at stellar light curves, we noticed that the noise structure was often shared across stars that were light years apart. Since that made direct interaction of the stars impossible, it was clear that the shared information was due to the telescope acting as a confounder. We thus devised a method that (a) regresses a given star of interest on a large set of other stars chosen such that their measurements contain no information about the star's astrophysical signal, and (b) removes that regression in order to cancel the telescope's influence.[13] The method is called "half-sibling" regression since target and predictors share a parent, namely the telescope. The method recovers the random variable representing the astrophysical signal almost surely (up to a constant offset), for an additive noise model (specifically, the observed light curve is a sum of the unknown astrophysical signal and an unknown function of the telescope noise), subject to the assumption that the telescope's effect on the star is in principle predictable from the other stars [132].

13 For events that are localized in time (such as exoplanet transits), we further argued that the same applies for suitably chosen past and future values of the star itself, which can thus also be used as predictors.

In 2013, the Kepler spacecraft suffered a technical failure, which left it with only two functioning reaction wheels, insufficient for the precise spatial orientation required by the original Kepler mission. NASA decided to use the remaining fuel to make further observations, however, the systematic error was significantly larger than before—a godsend for our method designed to remove exactly these errors. We augmented it with models of exoplanet transits and an efficient way to search light curves, leading to the discovery of 36 planet candidates [32], of which 21 were subsequently validated as bona fide exoplanets [92]. Four years later, astronomers found traces of water in the atmosphere of the exoplanet K2-18b—the first such discovery for an exoplanet in the habitable zone, i.e., allowing for liquid water [10, 151]. The planet turned out to be one that had been first detected in our work [32] (exoplanet candidate EPIC 201912552).

10. CURRENT RESEARCH AND OPEN PROBLEMS

Conservation of information. We have previously argued that the mechanization of information processing currently plays a similar role to the mechanization of energy processing in earlier industrial revolutions [125]. Our present understanding of information is rather incomplete, as was the understanding of energy during the course of the first two industrial revolutions. The profound modern understanding of energy came with Emmy Noether and the insight that energy conservation is due to a symmetry (or covariance) of the fundamental laws of physics: they look the same no matter how we shift time. One might argue that information, suitably conceptualized, should also be a conserved quantity, and that this might also be a consequence of symmetries. The notions of invariance/independence discussed above may be able to play a role in this respect.

Mass seemingly played two fundamentally different roles (inertia and gravitation) until Einstein furnished a deeper connection in general relativity. It is noteworthy that causality introduces a layer of complexity underlying the symmetric notion of statistical mutual information. Discussing source coding and channel coding, Shannon [138] remarked: *This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.*

What is an object? Following the i.i.d. pattern recognition paradigm, machine learning learns objects by extracting patterns from many observations. A complementary view may consider objects as modules that can be separately manipulated or intervened upon [149]. The idea that objects are defined by their behavior under transformation has been influential in fields ranging from psychology to mathematics [74, 88].

Causal representation learning. In hindsight, it appears somewhat naive that first attempts to build AI tried to realize intelligence by programs written by humans, since existing examples of intelligent systems appear much too complex for that. However, there is a second problem, which is just as significant: classic AI assumed that the symbols which were the basis of algorithms were provided a priori by humans. When building a chess program, it is

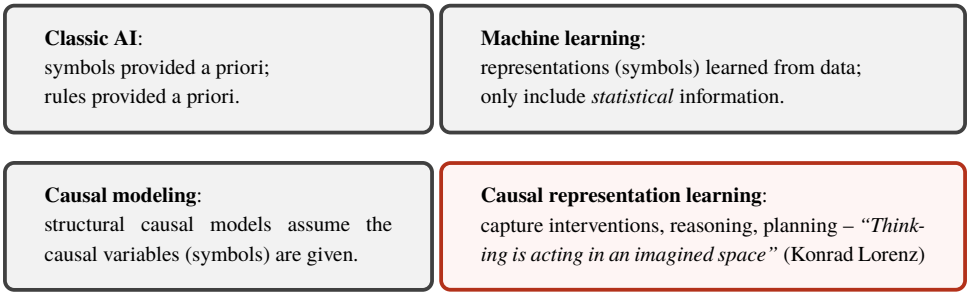


FIGURE 11
Causal representation learning aims to automatically learn representations that contain not just statistical information, but support interventions, reasoning, and planning. The long-term goal of this field is to learn causal world models supporting AI, or causal digital twins of complex systems.

clear that the algorithms operate on chess board positions and chess pieces; however, if we want to solve a real-world problem in an unstructured environment (e.g., recognize spoken language), it is not clear what constitutes the basic symbols to be processed.

Traditional causal discovery and reasoning assumed that the elementary units are random variables connected by a causal graph. Real-world observations, however, are usually not structured into such units to begin with. For instance, objects in images that permit causal reasoning first need to be discovered [84, 85, 149, 157]. The emerging field of *causal representation learning* strives to learn these variables from data, much like machine learning went beyond symbolic AI in not requiring that the symbols that algorithms manipulate be given a priori (see Figure 11).

Defining objects or variables, and structural models connecting them, can sometimes be achieved by coarse-graining of microscopic models, including microscopic SCMs [123], ordinary differential equations [122], and temporally aggregated time series [37]. While most causal models in economics, medicine, or psychology use variables that are abstractions of more elementary concepts, it is challenging to state general conditions under which coarse-grained variables admit causal models with well-defined interventions [15, 16, 123]. The task of identifying suitable units that admit causal models aligns with the general goal of modern machine learning to learn meaningful representations for data, where meaningful can mean *robust*, *transferable*, *interpretable*, *explainable*, or *fair* [70–72, 77, 155, 162]. To combine structural causal modeling (Definition 4.4) and representation learning, we may try to devise machine learning models whose inputs may be high-dimensional and unstructured, but whose inner workings are (partly) governed by an SCM.

Suppose that our high-dimensional, low-level observations $\mathbf{X} = (X_1, \ldots, X_d)$ are explained by a small number of unobserved, or *latent*, variables $\mathbf{S} = (S_1, \ldots, S_n)$ where $n \ll d$, in that $\mathbf{X}$ is generated by applying an injective map $g : \mathbb{R}^n \to \mathbb{R}^d$ to $\mathbf{S}$ (see Figure 12c),

$$\mathbf{X} = g(\mathbf{S}). \tag{10.1}$$

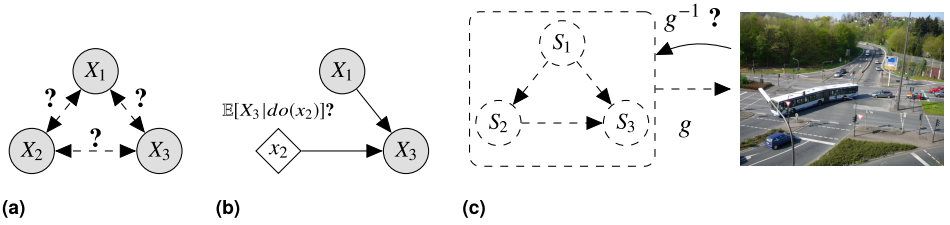


FIGURE 12

Overview of different causal learning tasks: (a) *causal discovery* (Section 7) aims to learn the causal graph (or SCM) connecting a set of *observed* variables; (b) *causal reasoning* (Section 9) aims to answer interventional or counterfactual queries based on a (partial) causal model over observed variables X_i; (c) *causal representation learning* (Section 10) aims to infer a causal model consisting of a small number of high-level, abstract causal variables S_i and their relations from potentially high-dimensional, low-level observations $\mathbf{X} = g(\mathbf{S})$.

A common assumption regarding (10.1) is that the latent S_i are jointly independent, e.g., for independent component analysis (ICA) [57] (where g is referred to as a *mixing*) or disentangled representation learning [9] (where g is called a *decoder*). Presently, however, we instead want think of the latent S_i as *causal variables* that support interventions and reasoning.

The S_i may thus well be dependent, and possess a causal factorization (4.1),

$$p(S_1, \ldots, S_n) = \prod_{i=1}^{n} p(S_i|\mathbf{PA}_i), \tag{10.2}$$

induced by an underlying (acyclic) SCM $\mathcal{M} = (\mathbf{F}, p_{\mathbf{U}})$ with jointly independent U_i and

$$\mathbf{F} = \{S_i := f_i(\mathbf{PA}_i, U_i)\}_{i=1}^{n}. \tag{10.3}$$

Our goal is to learn a latent causal model consisting of (i) the causal representation $\mathbf{S} = g^{-1}(\mathbf{X})$, along with (ii) the corresponding causal graph and (iii) the mechanisms $p(S_i|\mathbf{PA}_i)$ or f_i. This is a challenging task, since none of them are directly observed or known a priori; instead we typically only have access to observations of $\mathbf{X}$. In fact, there is no hope in an i.i.d. setting since already the simpler case with independent S_i (and $n = d$) is not identifiable in general (i.e., for arbitrary nonlinear g in (10.1)): even independence does not sufficiently constrain the problem to uniquely recover, or identify, the true S_i's up to any simple class of ambiguities such as permutations and elementwise invertible transformations of the S_i [58].

To link causal representation learning to the well-studied ICA setting with independent latents in (10.1), we can consider the so-called *reduced form* of an (acyclic) SCM: by recursive substitution of the structural assignments (10.3) in topological order of the causal graph, we can write the latent causal variables $\mathbf{S}$ as function of the noise variables only

$$\mathbf{S} = f_{\mathrm{RF}}(\mathbf{U}). \tag{10.4}$$

Due to acyclicity, this mapping $f_{\mathrm{RF}} : \mathbb{R}^n \to \mathbb{R}^n$ has a lower triangular Jacobian (possibly after reordering the S_i without loss of generality). However, (10.4) is strictly less informative

than (10.3): while they entail the same distribution (10.2), the former no longer naturally supports interventions on the S_i but only changes to the noise distribution $p_{\mathbf{U}}$ (an example of a so-called *soft* intervention [30]). At the same time, the reduced form (10.4) allows us to rewrite (10.1) as

$$\mathbf{X} = g \circ f_{\text{RF}}(\mathbf{U}). \tag{10.5}$$

Through this lens, the task of learning the reduced form (10.4) could be seen as structured form of nonlinear ICA (i.e., (10.1) with independent latents) where we additionally want to learn an intermediate representation through f_{RF}. However, as discussed, we cannot even solve the problem with independent latents (i.e., identify $g \circ f_{\text{RF}}$ in (10.5)) [58], let alone separate the SCM and mixing functions to recover the intermediate causal representation.

It is not surprising that is is not possible to solve the strictly harder causal representation learning problem in an i.i.d. setting and that additional causal learning signals are needed. This gives rise to the following questions: How can we devise causal training algorithms to learn the S_i? And, what types of additional data, assumptions, and constraints do they require beyond the i.i.d. setting? Two general ideas are to (i) build on the ICM Principle 5.1 and enforce some form of (algorithmic) independence between the learned causal mechanisms $p(S_i|\mathbf{PA}_i)$ or f_i, and (ii) use heterogeneous (*non-i.i.d.*) data, e.g., from multiple views or different environments, arising from interventions in the underlying latent SCM (10.3). We briefly discuss some more concrete ideas based on recent work.

Generative approach: Causal autoencoders. One approach is to try to learn the generative causal model (10.1) and (10.3), or its reduced form (10.4), using an *autoencoder* approach [73]. An autoencoder consists of an *encoder* function $q : \mathbb{R}^d \to \mathbb{R}^n$ which maps $\mathbf{X}$ to a latent "bottleneck" representation (e.g., comprising the unexplained noise variables $\mathbf{U}$), and a *decoder* function $\hat{g} : \mathbb{R}^n \to \mathbb{R}^d$ mapping back to the observations. For example, the decoder may directly implement the composition $\hat{g} = g \circ f_{\text{RF}}$ from (10.4). Alternatively, it could consist of multiple modules, implementing (10.1) and (10.3) separately. A standard procedure to train such an autoencoder architecture is to minimize the reconstruction error, i.e., to satisfy $\hat{g} \circ q \approx \text{id}$ on a training set of observations of $\mathbf{X}$. As discussed, this alone is insufficient, so to make it causal we can impose additional constraints on the structure of the decoder [80] and try to make the causal mechanisms independent by ensuring that they are invariant across problems and can be independently intervened upon. For example, if we intervene on the causal variables S_i or noise distribution $p_{\mathbf{U}}$ in our model of (10.3) or (10.4), respectively, this should still produce "valid" observations, as assessed, e.g., by the discriminator of a generative adversarial network [38]. While we ideally want to manipulate the causal variables, another way to intervene is to replace noise variables with the corresponding values computed from other input images, a procedure that has been referred to as hybridization [11]. Alternatively, if we have access to multiple environments, i.e., datasets collected under different conditions, we could rely on the Sparse Mechanism Shift Principle 5.2 by requiring that changes can be explained by shifts in only a few of the $p(S_i|\mathbf{PA}_i)$.

Discriminative approach: Self-supervised causal representation learning. A different machine learning approach for unsupervised representation learning, that is not based on

generative modeling but is discriminative in nature, is *self-supervised learning with data augmentation*. Here, the main idea is to apply some hand-crafted transformations to the observation to generate augmented views that are thought to share the main semantic characteristics with the original observation (e.g., random crops or blurs for images). One then directly learns a representation by maximizing the similarity across encodings of views related to each other by augmentations, while enforcing diversity across those of unrelated views. In a recent work [158], we set out to better understand this approach theoretically, as well as to investigate its potential for learning causal representations. Starting from (10.1), we postulate a latent causal model of the form $\mathbf{S}_c \to \mathbf{S}_s$, where $\mathbf{S}_c$ is a (potentially multivariate) *content* variable, defined as the high-level semantic part of the representation $\mathbf{S} = (\mathbf{S}_c, \mathbf{S}_s)$ that is assumed invariant across views; and $\mathbf{S}_s$ is a (potentially multivariate) *style* variable, defined as the remaining part of the representation that may change. Within this setting, data augmentations have a natural interpretation as counterfactuals under a hypothetical intervention on the style variables, given the original view. It can be shown that in this case, subject to some technical assumptions, common contrastive self-supervised learning algorithms [19, 45, 152] as well as appropriately constrained generative models isolate, or recover, the true content variables $\mathbf{S}_c$ up to an invertible transformation. By extending this approach to use multiple augmented views of the same observation, and linking these to different counterfactuals in the underlying latent SCM, it may be possible to recover a more-fine-grained causal representation.

Independent mechanism analysis. We also explored [40] to what extent the ICM Principle 5.1 may be useful for unsupervised representation learning tasks such as (10.1), particularly for imposing additional constraints on the mixing function g. It turns out that independence between $p(\mathbf{S})$ and the mixing g—measured, e.g., as discussed in Section 5 in the context of Figure 6 and [68]—does not impose nontrivial constraints when $\mathbf{S}$ is not observed, even when the S_i are assumed independent as in ICA. However, by thinking of each S_i as independently *influencing* the observed distribution, we postulate another type of independence between the partial derivatives $\frac{\partial g}{\partial S_i}$ of the mixing g which has a geometric interpretation as an orthogonality condition on the columns of the Jacobian of g. The resulting *independent mechanism analysis* (IMA) approach rules out some of the common examples of nonidentifiability of nonlinear ICA [58, 83] mentioned above. Since IMA does not require independent sources, it may also be a useful constraint for causal representation learning algorithms.

Learning transferable mechanisms and multitask learning. Machine learning excels in i.i.d. settings, and through the use of high capacity learning algorithms we can achieve outstanding performance on many problems, provided we have i.i.d. data for each individual problem (Section 2). However, natural intelligence excels at generalizing across tasks and settings. Suppose we want to build a system that can solve multiple tasks in multiple environments. If we view learning as data compression, it would make sense for that system to utilize components that apply across tasks and environments, and thus need to be stored only once [125].

Indeed, an artificial or natural agent in a complex world is faced with limited resources. This concerns training data, i.e., we only have limited data for each individual task/domain, and thus need to find ways of pooling/reusing data, in stark contrast to the current industry practice of large-scale labeling work done by humans. It also concerns computational resources: animals have constraints on the resources (e.g., space, energy) used by their brains, and evolutionary neuroscience knows examples where brain regions get repurposed. Similar constraints apply as machine learning systems get embedded in physical devices that may be small and battery-powered. Versatile AI models that robustly solve a range of problems in the real world will thus likely need to reuse components, which requires that the components are robust across tasks and environments [127,133]. This calls for a structure whose modules are maximally reusable. An elegant way to do this would be to employ a modular structure that mirrors modularity that exists in the world. In other words, if the are mechanisms at play in the world play similar roles across a range of environments, tasks, and settings, then it would be prudent for a model to employ corresponding computational modules [39]. For instance, if variations of natural lighting (the position of the sun, clouds, etc.) imply that the visual environment can appear in brightness conditions spanning several orders of magnitude, then visual processing algorithms in our nervous system should employ methods that can factor out these variations, rather than building separate sets of object recognizers for every lighting condition. If our brain were to model the lighting changes by a gain control mechanism, say, then this mechanism in itself need not have anything to do with the physical mechanisms bringing about brightness differences. It would, however, play a role in a modular structure that corresponds to the role the physical mechanisms play in the world's modular structure—in other words, it would *represent* the physical mechanism. Searching for the most versatile, yet compact, models would then automatically produce a bias towards models that exhibit certain forms of structural isomorphy to a world that we cannot directly recognize.

A sensible inductive bias to learn such models is to look for independent causal mechanisms [82], and competitive training can play a role in this: for a pattern recognition task, learning causal models that contain independent mechanisms helps in transferring modules across substantially different domains [99].

Interventional world models, surrogate models, digital twins, and reasoning. Modern representation learning excels at learning representations of data that preserve relevant statistical properties [9,79]. It does so, however, without taking into account causal properties of the variables, i.e., it does not care about the interventional properties of the variables it analyzes or reconstructs. Going forward, causality will play a major role in taking representation learning to the next level, moving beyond the representation of statistical dependence structures towards models that support intervention, planning, and reasoning. This would realize Konrad Lorenz' notion of *thinking* as *acting in an imagined space*. It would also provide a means to learn causal *digital twins* that go beyond reproducing statistical dependences captured by *surrogate models* trained using machine learning.

The idea of surrogate modeling is that we may have a complex phenomenon for which we have access to computationally expensive simulation data. If the mappings involved (e.g., from parameter settings to target quantities) can be fitted from data, we can employ machine learning, which will often speed them up by orders of magnitude. Such a speed-up can qualitatively change the usability of a model: for instance, we have recently built a system to map gravitational wave measurements to a probability distribution of physical parameters of a black hole merger event, including sky position [27]. The fact that this model only requires seconds to evaluate makes it possible to immediately start electromagnetic follow-up observations using telescopes as soon as a gravitational wave event has been detected, enabling analysis of transient events.

Going forward, we anticipate that surrogate modeling will benefit from respecting the causal factorization (4.1) decomposing the overall dependence structure into mechanisms (i.e., causal conditionals). We can then build an overall model of a system by modeling the mechanisms independently, each of them using the optimal method. Some of the conditionals we may know analytically, some we may be able to transfer from related problems, if they are invariant. For some, we may have access to real data to estimate them, and for others, we may need to resort to simulations, possibly fitted using surrogate models.

If the model is required to fully capture the effects of all possible interventions, then all components should be fitted as described in the causal directions (i.e., we fit the causal mechanisms). Such a model then allows employing all the causal reasoning machinery described in Sections 4 and 9 (e.g., computing interventional and, in the case of SCMs, counterfactual distributions). If, on the other hand, a model only needs to capture *some* of the possible interventions, and is used in a purely predictive/observational mode for other variables, then we can get away with also using and fitting some noncausal modules, i.e., using a decomposition which lies in-between (4.1) and (4.2).

We believe that this overall framework will be a principled and powerful approach to build such (causal) digital twins or causal surrogate models by combining a range of methods and bringing them to bear according to their strengths.

Concluding remarks. Most of the discussed fields are still in their infancy, and the above account is biased by personal taste and knowledge. With the current hype around machine learning, there is much to say in favor of some humility towards what machine learning can do, and thus towards the current state of AI—the hard problems have not been solved yet, making basic research in this field all the more exciting.

ACKNOWLEDGMENTS

Many thanks to all past and present members of the Tübingen causality team, and to Cian Eastwood and Elias Bareinboim for feedback on the manuscript.

REFERENCES

[1] J. Aldrich, Autonomy. *Oxf. Econ. Pap.* **41** (1989), 15–34.

[2] J. D. Angrist, G. W. Imbens, and D. B. Rubin, Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** (1996), no. 434, 444–455.

[3] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, Invariant risk minimization. 2019, arXiv:1907.02893.

[4] E. Bareinboim and J. Pearl, Transportability from multiple environments with limited experiments: completeness results. In *Advances in neural information processing systems 27*, pp. 280–288, Curran Associates, Inc., 2014.

[5] E. Bareinboim and J. Pearl, Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* **113** (2016), no. 27, 7345–7352.

[6] S. Bauer, B. Schölkopf, and J. Peters, The arrow of time in multivariate time series. In *Proceedings of the 33nd international conference on machine learning 48*, pp. 2043–2051, PMLR, 2016.

[7] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine learning practice and the bias-variance trade-off. 2018, arXiv:1812.11118.

[8] S. Ben-David, T. Lu, T. Luu, and D. Pál, Impossibility theorems for domain adaptation. In *Proceedings of the international conference on artificial intelligence and statistics 13 (AISTATS)*, pp. 129–136, PMLR, 2010.

[9] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** (2013), no. 8, 1798–1828.

[10] B. Benneke, I. Wong, C. Piaulet, H. A. Knutson, I. J. M. Crossfield, J. Lothringer, C. V. Morley, P. Gao, T. P. Greene, C. Dressing, D. Dragomir, A. W. Howard, P. R. McCullough, E. M. R. K. J. J. Fortney, and J. Fraine, Water vapor on the habitable-zone exoplanet K2-18b. 2019, arXiv:1909.04642.

[11] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, Counterfactuals uncover the modular structure of deep generative models. In *International conference on learning representations*, OpenReview.net, 2020.

[12] M. Besserve, N. Shajarisales, B. Schölkopf, and D. Janzing, Group invariance principles for causal generative models. In *Proceedings of the 21st international conference on artificial intelligence and statistics (AISTATS)*, pp. 557–565, PMLR, 2018.

[13] S. Bongers, P. Forré, J. Peters, and J. M. Mooij, Foundations of structural causal models with cycles and latent variables. *Ann. Statist.* **49** (2021), no. 5, 2885–2915.

[14] D. Buchsbaum, S. Bridgers, D. Skolnick Weisberg, and A. Gopnik, The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philos. Trans. R. Soc. B, Biol. Sci.* **367** (2012), no. 1599, 2202–2212.

[15] K. Chalupka, F. Eberhardt, and P. Perona, Multi-level cause-effect systems. In *Artificial intelligence and statistics*, pp. 361–369, PMLR, 2016.

[16] K. Chalupka, F. Eberhardt, and P. Perona, Causal feature learning: an overview. *Behaviormetrika* **44** (2017), no. 1, 137–164.

[17] O. Chapelle, B. Schölkopf, and A. Zien (eds.), *Semi-supervised learning*. MIT Press, Cambridge, MA, USA, 2006.

[18] C. R. Charig, D. R. Webb, S. R. Payne, and J. E. Wickham, Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br. Med. J. (Clin. Res. Ed.)* **292** (1986), no. 6524, 879–882.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations. 2020, arXiv:2002.05709.

[20] D. M. Chickering, Learning Bayesian networks is NP-complete. In *Learning from data*, pp. 121–130, Springer, 1996.

[21] D. M. Chickering, Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** (2002), 507–554.

[22] G. F. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9** (1992), no. 4, 309–347.

[23] D. R. Cox, *Planning of experiments*, Wiley, 1958.

[24] P. Daniušis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf, Inferring deterministic causal relations. In *Proceedings of the 26th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 143–150, AUAI Press, 2010.

[25] A. P. Dawid, Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B.* **41** (1979), no. 1, 1–31.

[26] A. P. Dawid, Causal inference without counterfactuals. *J. Amer. Statist. Assoc.* **95** (2000), no. 450, 407–424.

[27] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Real-time gravitational-wave science with neural posterior estimation. *Phys. Rev. Lett.* **127** (2021), no. 24.

[28] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Appl. Math. 31, Springer, New York, NY, 1996.

[29] V. Didelez, S. Meng, and N. A. Sheehan, Assumptions of IV methods for observational epidemiology. *Statist. Sci.* **25** (2010), 22–40.

[30] F. Eberhardt and R. Scheines, Interventions and causal inference. *Philos. Sci.* **74** (2007), no. 5, 981–995.

[31] R. A. Fisher, *The design of experiments 2*. Oliver & Boyd, Edinburgh & London, 1937.

[32] D. Foreman-Mackey, B. T. Montet, D. W. Hogg, T. D. Morton, D. Wang, and B. Schölkopf, A systematic search for transiting planets in the K2 data. *Astrophys. J.* **806** (2015), no. 2.

[33] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pp. 489–496, Curran Associates, Inc., 2008.

[34] D. Geiger and D. Heckerman, Learning Gaussian networks. In *Proceedings of the tenth international conference on uncertainty in artificial intelligence*, pp. 235–243, AUAI Press, 1994.

[35] D. Geiger and J. Pearl, Logical and algorithmic properties of independence and their application to Bayesian networks. *Ann. Math. Artif. Intell.* **2** (1990), 165–178.

[36] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, Domain adaptation with conditional transferable components. In *Proceedings of the 33nd international conference on machine learning*, pp. 2839–2848, PMLR, 2016.

[37] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao, Causal discovery from temporally aggregated time series. In *Proceedings of the thirty-third conference on uncertainty in artificial intelligence*, pp. 1066–1075, AUAI Press, 2017.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets. In *Advances in neural information processing systems 27*, pp. 2672–2680, Curran Associates, Inc., 2014.

[39] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, Recurrent independent mechanisms. In *International conference on learning representations*, OpenReview.net, 2021.

[40] L. Gresele, J. von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve, Independent mechanism analysis, a new concept? In *Advances in neural information processing systems 34*, Curran Associates, Inc., 2021.

[41] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, A kernel method for the two-sample-problem. In *Advances in neural information processing systems 19*, pp. 513–520, Curran Associates, Inc., 2007.

[42] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic learning theory*, pp. 63–78, Springer, 2005.

[43] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, A kernel statistical test of independence. In *Advances in neural information processing systems 20*, pp. 585–592, Curran Associates, Inc., 2008.

[44] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** (2005), 2075–2129.

[45] U. M. Gutmann and A. Hyvärinen, Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In *International conference on artificial intelligence and statistics*, pp. 297–304, PMLR, 2010.

[46] T. Haavelmo, The probability approach in econometrics. *Econometrica* (1944), iii–115.

[47] D. Heckerman, D. Geiger, and D. M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* **20** (1995), no. 3, 197–243.

[48] D. Heckerman, C. Meek, and G. Cooper, A Bayesian approach to causal discovery. In *Innovations in machine learning*, pp. 1–28, Springer, 2006.

[49] M. A. Hernán, D. Clayton, and N. Keiding, The Simpson's paradox unraveled. *Int. J. Epidemiol.* **40** (2011), no. 3, 780–785.

[50] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, Beta-VAE: learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, OpenReview.net, 2017.

[51] P. W. Holland, Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** (1986), no. 396, 945–960.

[52] K. D. Hoover, *Causality in macroeconomics*. Cambridge University Press, 2001.

[53] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems 21*, pp. 689–696, Curran Associates, Inc., 2009.

[54] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, Behind distribution shift: mining driving forces of changes and causal arrows. In *IEEE 17th international conference on data mining (ICDM 2017)*, pp. 913–918, IEEE, 2017.

[55] B. Huang, K. Zhang, J. Zhang, J. D. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.* **21** (2020), no. 89, 1–53.

[56] Y. Huang and M. Valtorta, Pearl's calculus of intervention is complete. In *Proceedings of the twenty-second conference on uncertainty in artificial intelligence*, pp. 217–224, AUAI Press, 2006.

[57] A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications. *Neural Netw.* **13** (2000), no. 4–5, 411–430.

[58] A. Hyvärinen and P. Pajunen, Nonlinear independent component analysis: existence and uniqueness results. *Neural Netw.* **12** (1999), no. 3, 429–439.

[59] A. Hyvarinen, H. Sasaki, and R. Turner, Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868, PMLR, 2019.

[60] G. W. Imbens and T. Lemieux, Regression discontinuity designs: a guide to practice. *J. Econometrics* **142** (2008), no. 2, 615–635.

[61] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[62] D. Janzing, R. Chaves, and B. Schölkopf, Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New J. Phys.* **18** (2016), no. 093052, 1–13.

[63] D. Janzing, P. Hoyer, and B. Schölkopf, Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th international conference on machine learning*, edited by J. Fürnkranz and T. Joachims, pp. 479–486, PMLR, 2010.

[64] D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf, Identifying confounders using additive noise models. In *Proceedings of the 25th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 249–257, AUAI Press, 2009.

[65] D. Janzing and B. Schölkopf, Causal inference using the algorithmic Markov condition. *IEEE Trans. Inf. Theory* **56** (2010), no. 10, 5168–5194.

[66] D. Janzing and B. Schölkopf, Semi-supervised interpolation in an anticausal learning scenario. *J. Mach. Learn. Res.* **16** (2015), 1923–1948.

[67] D. Janzing and B. Schölkopf, Detecting non-causal artifacts in multivariate linear regression models. In *Proceedings of the 35th international conference on machine learning (ICML)*, pp. 2250–2258, PMLR, 2018.

[68] D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, Information-geometric approach to inferring causal directions. *Artificial Intelligence* **182–183** (2012), 1–31.

[69] Z. Jin, J. von Kügelgen, J. Ni, T. Vaidhya, A. Kaushal, M. Sachan, and B. Schölkopf, Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP)*, pp. 9499–9513, Association for Computational Linguistics, 2021.

[70] A.-H. Karimi, B. Schölkopf, and I. Valera, Algorithmic recourse: from counterfactual explanations to interventions. In *Conference on fairness, accountability, and transparency*, pp. 353–362, ACM, 2021.

[71] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera, Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in neural information processing systems 33*, pp. 265–277, Curran Associates, Inc., 2020.

[72] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems 30*, pp. 656–666, Curran Associates, Inc., 2017.

[73] D. P. Kingma and M. Welling, Auto-encoding variational Bayes. 2013, arXiv:1312.6114.

[74] F. Klein, *Vergleichende Betrachtungen über neuere geometrische Forschungen*. Verlag von Andreas Deichert, Erlangen, 1872.

[75] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[76] S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf, Consistency of causal inference under the additive noise model. In *Proceedings of the 31th international conference on machine learning*, pp. 478–486, PMLR, 2014.

[77] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, Counterfactual fairness. In *Advances in neural information processing systems 30*, pp. 4066–4076, Curran Associates, Inc., 2017.

[78] S. L. Lauritzen, *Graphical models. 17*. Clarendon Press, 1996.

[79] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning. *Nature* **521** (2015), no. 7553, 436–444.

[80] F. Leeb, Y. Annadani, S. Bauer, and B. Schölkopf, Structural autoencoders improve representations for generation and transfer. 2020, arXiv:2006.07796.

[81] G. W. Leibniz, *Discours de métaphysique*, 1686 (cited after Chaitin, 2010).

[82] F. Locatello, D. Vincent, I. Tolstikhin, G. Rätsch, S. Gelly, and B. Schölkopf, Competitive training of mixtures of independent deep generative models. 2018, arXiv:1804.11130.

[83] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th international conference on machine learning*, PMLR, 2019.

[84] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, Object-centric learning with slot attention. In *Advances in neural information processing systems*, pp. 11525–11538, Curran Associates, Inc., 2020.

[85] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou, Discovering causal signals in images. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 58–66, IEEE Computer Society, 2017.

[86] J. Loschmidt, Über den Zustand des Wärmegleichgewichtes eines Systems von Körpern mit Rücksicht auf die Schwerkraft. *Sitzungsber. Akad. Wiss. Wien, Math.-Naturwiss. Kl.* **73** (1876), 128–142.

[87] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf, Nonlinear invariant risk minimization: a causal approach. 2021, arXiv:2102.12353.

[88] S. MacLane, *Categories for the working mathematician*. Springer, New York, 1971.

[89] R. Matthews, Storks deliver babies ($p = 0.008$). *Teach. Stat.* **22** (2000), no. 2, 36–38.

[90] C. Meek, Causal inference and causal explanation with background knowledge. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pp. 403–410, Morgan Kaufmann Publishers Inc., 1995.

[91] F. H. Messerli, Chocolate consumption, cognitive function, and Nobel laureates. *N. Engl. J. Med.* **367** (2012), no. 16, 1562–1564.

[92] B. T. Montet, T. D. Morton, D. Foreman-Mackey, J. A. Johnson, D. W. Hogg, B. P. Bowler, D. W. Latham, A. Bieryla, and A. W. Mann, Stellar and planetary properties of K2 campaign 1 candidates and validation of 17 planets, including a planet receiving earth-like insolation. *Astrophys. J.* **809** (2015), no. 1, 25.

[93] R. P. Monti, K. Zhang, and A. Hyvärinen, Causal discovery with general non-linear relationships using non-linear ICA. In *Uncertainty in artificial intelligence*, pp. 186–195, PMLR, 2020.

[94] J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf, On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems 24 (NIPS)*, pp. 639–647, Curran Associates, Inc., 2011.

[95] J. M. Mooij, D. Janzing, J. Peters, and B. Schölkopf, Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th international conference on machine learning (ICML)*, pp. 745–752, PMLR, 2009.

[96] J. M. Mooij, D. Janzing, and B. Schölkopf, From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the 29th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 440–448, AUAI Press, 2013.

[97] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks. *J. Mach. Learn. Res.* **17** (2016), no. 32, 1–102.

[98] J. S. Neyman, On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Ann. Agric. Sci.* **10** (1923), 1–51. (Translated and edited by D. M. Dabrowska and T. P. Speed, *Statist. Sci.* **5** (1990), 465–480).

[99] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf, Learning independent causal mechanisms. In *Proceedings of the 35th international conference on machine learning (PMLR) 80*, pp. 4036–4044, PMLR, 2018.

[100] J. Park and K. Muandet, A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in neural information processing systems 33 (NEURIPS 2020)*, pp. 21247–21259, Curran Associates, Inc., 2020.

[101] J. Pearl, Bayesian networks: a model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the cognitive science society*, pp. 329–334, University of California (Los Angeles), Computer Science Department, 1985.

[102] J. Pearl, Causal diagrams for empirical research. *Biometrika* 82 (1995), no. 4, 669–-688.

[103] J. Pearl, Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 411–420, AUAI Press, 2001.

[104] J. Pearl, *Causality: models, reasoning, and inference*. 2nd edn., Cambridge University Press, New York, NY, 2009.

[105] J. Pearl, Comment: understanding Simpson's paradox. *Amer. Statist.* **68** (2014), no. 1, 8–13.

[106] J. Pearl and E. Bareinboim, External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** (2014), no. 4, 579–595.

[107] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018.

[108] J. Pearl and A. Paz, Confounding equivalence in causal inference. *J. Causal Inference* 2 (2014), no. 1, 75-–93.

[109] J. Pearl and T. Verma, A theory of inferred causation. In *Principles of knowledge representation and reasoning: proceedings of the second international conference 2*, p. 441, Morgan Kaufmann, 1991.

[110] J. Peters, P. Bühlmann, and N. Meinshausen, Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** (2016), no. 5, 947–1012.

[111] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference – foundations and learning algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[112] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, Identifiability of causal graphs using functional models. In *Proceedings of the 27th annual conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 589–598, AUIA Press, 2011.

[113] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15** (2014), 2009–2053.

[114] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** (2018), no. 1, 5–31.

[115] K. Popper, *The logic of scientific discovery*. 1959.

[116] H. Reichenbach, *The direction of time*. University of California Press, Berkeley, CA, 1956.

[117] J. Robins, A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** (1986), no. 9–12, 1393–1512.

[118] J. M. Robins, M. A. Hernan, and B. Brumback, Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** (2000), no. 5, 550–560.

[119] R. W. Robinson, Counting labeled acyclic digraphs, new directions in the theory of graphs. In *Proc. third Ann Arbor conf., Univ. Michigan, Ann Arbor, MI, 1971*, pp. 239–273, Academic Press, 1973.

[120] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** (2018), no. 36, 1–34.

[121] P. R. Rosenbaum and D. B. Rubin, The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** (1983), no. 1, 41–55.

[122] P. K. Rubenstein, S. Bongers, B. Schölkopf, and J. M. Mooij, From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th conference on uncertainty in artificial intelligence (UAI)*, pp. 114–123, AUAI Press, 2018.

[123] P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf, Causal consistency of structural equation models. In *Proceedings of the thirty-third conference on uncertainty in artificial intelligence*, pp. 808–817, AUAI Press, 2017.

[124] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** (1974), no. 5, 688.

[125] B. Schölkopf, Causality for machine learning. 2019, arXiv:1911.10500. To appear in: R. Dechter, J. Halpern, and H. Geffner, *Probabilistic and causal inference: the works of Judea Pearl*. ACM books, 2019.

[126] B. Schölkopf, R. Herbrich, and A. J. Smola, A generalized representer theorem. In *Annual conference on computational learning theory*, edited by D. Helmbold and R. Williamson, pp. 416–426, Lecture Notes in Comput. Sci. 2111, Springer, Berlin, 2001.

[127] B. Schölkopf, D. Janzing, and D. Lopez-Paz, Causal and statistical learning. *Oberwolfach Rep.* **13** (2016), no. 3, 1896–1899.

[128] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij, On causal and anticausal learning. In *Proceedings of the 29th international conference on machine learning (ICML)*, pp. 1255–1262, PMLR, 2012.

[129] B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters, Computing functions of random variables via reproducing kernel Hilbert space representations. *Stat. Comput.* **25** (2015), no. 4, 755–766.

[130] B. Schölkopf and A. J. Smola, *Learning with kernels*. MIT Press, Cambridge, MA, 2002.

[131] B. Schölkopf, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu, RKHS representation of measures applied to homogeneity, independence, and Fourier optics. *Oberwolfach Rep.* **30** (2008), 42–44.

[132] B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters, Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci.* **113** (2016), no. 27, 7391–7398.

[133] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, Toward causal representation learning. *Proc. IEEE* **109** (2021), no. 5, 612–634.

[134] G. Schwarz, et al., Estimating the dimension of a model. *Ann. Statist.* **6** (1978), no. 2, 461–464.

[135] R. D. Shah and J. Peters, The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** (2020), no. 3, 1514–1538.

[136] N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve, Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd international conference on machine learning (ICML)*, pp. 285–294, PMLR, 2015.

[137] U. Shalit, F. D. Johansson, and D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085, PMLR, 2017.

[138] C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion. In *IRE international convention records 7*, pp. 142–163, Wiley-IEEE Press, 1959.

[139] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen, A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** (2006), 2003–2030.

[140] I. Shpitser and J. Pearl, Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st national conference on artificial intelligence*, pp. 1219–1226, AAAI Press, 2006.

[141] I. Shpitser, T. VanderWeele, and J. M. Robins, On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence*, pp. 527–536, AUAI Press, 2010.

[142] E. H. Simpson, The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** (1951), no. 2, 238–241.

[143] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, A Hilbert space embedding for distributions. In *Algorithmic learning theory: 18th international conference*, pp. 13–31, Springer, 2007.

[144] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*. 2nd edn., MIT Press, Cambridge, MA, 2000.

[145] W. Spohn, *Grundlagen der Entscheidungstheorie*. Scriptor, 1978.

[146] I. Steinwart and A. Christmann, *Support vector machines*. Springer, New York, NY, 2008.

[147] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, Robustly disentangled causal mechanisms: validating deep representations for interventional robustness. In *International conference on machine learning*, pp. 6056–6065, PMLR, 2019.

[148] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks. 2013, arXiv:1312.6199.

[149] M. Tangemann, S. Schneider, J. von Kügelgen, F. Locatello, P. Gehler, T. Brox, M. Kümmerer, M. Bethge, and B. Schölkopf, Unsupervised object learning via common fate. 2021, arXiv:2110.06562.

[150] J. Tian and J. Pearl, Causal discovery from changes. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 512–521, Morgan Kaufmann Publishers Inc., 2001.

[151] A. Tsiaras, I. Waldmann, G. Tinetti, J. Tennyson, and S. Yurchenko, Water vapour in the atmosphere of the habitable-zone eight-earth-mass planet K2-18b. *Nat. Astron.* **3** (2019), 1086–1091.

[152] A. van den Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding. 2018, arXiv:1807.03748.

[153] V. N. Vapnik, *Statistical learning theory*. Wiley, New York, NY, 1998.

[154] J. von Kügelgen, L. Gresele, and B. Schölkopf, Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *IEEE Trans. Artif. Intell.* **2** (2021), no. 1, 18–27.

[155] J. von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf, On the fairness of causal algorithmic recourse. In *36th AAAI conference on artificial intelligence*, AAAI Press, 2022.

[156] J. von Kügelgen, A. Mey, M. Loog, and B. Schölkopf, Semi-supervised learning, causality and the conditional cluster assumption. In *Conference on uncertainty in artificial intelligence*, pp. 1–10, AUAI Press, 2020.

[157] J. von Kügelgen, I. Ustyuzhaninov, P. Gehler, M. Bethge, and B. Schölkopf, Towards causal generative scene models via competition of experts. In *ICLR 2020 workshop on causal learning for decision making*, OpenReview.net, 2020.

[158] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, Self-supervised learning with data augmentations provably isolates content from style. In *Advances in neural information processing systems 34*, Curran Associates, Inc., 2021.

[159] J. Woodward. *Causation and manipulability*, Stanford Encyclopedia of Philosophy, 2001.

[160] P. G. Wright, *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

[161] S. Wright, Correlation and causation. *J. Agric. Res.* **20** (1921), 557–580.

[162] J. Zhang and E. Bareinboim, Fairness in decision-making – the causal explanation formula. In *Proceedings of the thirty-second AAAI conference on artificial intelligence*, pp. 2037–2045, AAAI Press, 2018.

[163] K. Zhang, M. Gong, and B. Schölkopf, Multi-source domain adaptation: a causal view. In *Proceedings of the 29th AAAI conference on artificial intelligence*, pp. 3150–3157, AAAI Press, 2015.

[164] K. Zhang and A. Hyvärinen, On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 647–655, AUAI Press, 2009.

[165] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th annual conference on uncertainty in artificial intelligence (UAI)*, pp. 804–813, AUAI Press, 2011.

[166] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, Domain adaptation under target and conditional shift. In *Proceedings of the 30th international conference on machine learning*, pp. 819–827, PMLR, 2013.

BERNHARD SCHÖLKOPF

Max Planck Institute for Intelligent Systems, Tübingen, Germany, bs@tuebingen.mpg.de

JULIUS VON KÜGELGEN

Max Planck Institute for Intelligent Systems, Tübingen, Germany; and University of Cambridge, Cambridge, United Kingdom, jvk@tuebingen.mpg.de

SECOND- AND HIGHER-ORDER GAUSSIAN ANTICONCENTRATION INEQUALITIES AND ERROR BOUNDS IN SLEPIAN'S COMPARISON THEOREM

CUN-HUI ZHANG

ABSTRACT

This paper presents some second- and higher-order Gaussian anticoncentration inequalities in high dimension and error bounds in Slepian's comparison theorem for the distribution functions of the maxima of two Gaussian vectors. The anticoncentration theorems are presented as upper bounds for the sum of the absolute values of the partial derivatives of a certain order for the joint distribution function of a Gaussian vector or weighted sums of such absolute values. Compared with the existing results where the covariance matrix of the entire Gaussian vector is required to be invertible, the bounds for the mth derivatives developed in this paper require only the invertibility of the covariance matrices of all subsets of m random variables. The second-order anticoncentration inequality is used to develop comparison theorems for the joint distribution functions of Gaussian vectors or, equivalently, the univariate distribution functions of their maxima via Slepian's interpolation. The third- and higher-order anticoncentration inequalities are motivated by recent advances in the central limit theorem and consistency of bootstrap for the maximum component of a sum of independent random vectors in high dimension and related applications in statistical inference and machine learning.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 60E15; Secondary 60G15, 62E20, 62E17

KEYWORDS

Anti-concentration, comparison of distributions, Gaussian process, density of maximum

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5594–5615
DOI 10.4171/ICM2022/186

1. INTRODUCTION

Let $X = (X_1, \ldots, X_d)^\top$ and $Y = (Y_1, \ldots, Y_d)^\top$ be two Gaussian vectors. Slepian's [31] inequality asserts that when X_i and Y_i have the same mean and variance, and $\mathrm{Var}(X_i - X_j) \le \mathrm{Var}(Y_i - Y_j)$ for all $1 \le i < j \le d$, the maximum of Y_i is stochastically larger than the maximum of X_i,

$$\mathbb{P}\Big\{\max_{1\le i\le d} X_i > t\Big\} \le \mathbb{P}\Big\{\max_{1\le i\le d} Y_i > t\Big\}, \quad \forall t \in \mathbb{R}. \tag{1.1}$$

Variations and extensions of Slepian's inequality have been developed to relax the conditions on the mean and variance of the individual components and pairwise contrasts, and to compare more general functions of the Gaussian vectors. Among such results, the Sudakov–Fernique inequality [15, 32, 33] asserts that

$$\mathbb{E}\Big[\max_{i\le d} X_i\Big] \le \mathbb{E}\Big[\max_{i\le d} Y_i\Big] \tag{1.2}$$

when $\mathbb{E}[X] = \mathbb{E}[Y]$ and $\mathrm{Var}(X_i - X_j) \le \mathrm{Var}(Y_i - Y_j)$ for all $1 \le i < j \le d$. Gordon's [16] inequalities extend (1.1) and (1.2) to the minimax function of Gaussian matrices. Chatterjee [5] provided an error bound in the Sudakov–Fernique inequality

$$\Big|\mathbb{E}\Big[\max_{i\le d} Y_i\Big] - \mathbb{E}\Big[\max_{i\le d} X_i\Big]\Big| \le \sqrt{\Delta \log d} \tag{1.3}$$

under the condition $\mathbb{E}[X] = \mathbb{E}[Y]$, where $\Delta = \max_{1\le i<j\le d} |\mathrm{Var}(Y_i - Y_j) - \mathrm{Var}(X_i - X_j)|$.

Comparison theorems such as the above and related anticoncentration inequalities are used in statistical inference, machine learning, reliability, signal processing, extreme value theory, random matrix theory, empirical processes, and more. See, for example, [1, 18, 21–23, 27, 29, 30, 34] and references therein. Anticoncentration inequalities in Slepian's comparison theorem provide upper bounds for the modulus of continuity of the distribution function of the maximum or the corresponding density function. This paper is motivated by the recent developments in the central limit theorem and bootstrap theory for the maximum component of a sum of independent random vectors in high dimension, specifically a crucial role of the Gaussian anticoncentration theory in these developments [6, 8, 10, 12, 13, 19, 25].

We present in this paper second- and higher-order anticoncentration inequalities for the Gaussian maxima and some of their implications in the comparison of Gaussian distribution functions. These anticoncentration inequalities provide upper bounds for the sum of the absolute values of the derivatives of a given order for the Gaussian joint distribution function and thus upper bounds for the derivatives of the distribution function of the Gaussian maxima. While the second-order anticoncentration inequalities are used in the development of our error bounds in Slepian's comparison theorem, the third- and higher-order anticoncentration inequalities can be used in studies of the central limit theorem and bootstrap in high dimension depending on the order of expansion in the related Slepian's [31] or Lindeberg's [24] interpolations in such applications.

We present below some error bounds in Slepian's comparison theorem as consequences of our results in Sections 2 and 3.

Theorem 1. *Let $X = (X_1, \ldots, X_d)^\top$ and $Y = (Y_1, \ldots, Y_d)^\top$ be two Gaussian vectors with $\mathbb{E}[X] = \mathbb{E}[Y] = \mu$. Let $\sigma_i = \sqrt{\mathrm{Var}(X_i) \wedge \mathrm{Var}(Y_i)}$, $\Delta_{i,j} = \{\mathrm{Cov}(Y_i, Y_j) - \mathrm{Cov}(X_i, X_j)\}/(\sigma_i \sigma_j)$, $\Delta_+^{\mathrm{cross}} = \max_{1 \le i < j \le d} (\Delta_{i,j})_+$, and $\Delta^{\mathrm{diag}} = \max_{1 \le i \le d} |\Delta_{i,i}|$. Then, for $d \ge 2$,*

$$\mathbb{P}\Big\{\max_{1 \le k \le d} Y_k \le t\Big\} - \mathbb{P}\Big\{\max_{1 \le k \le d} X_k \le t\Big\} \le \sqrt{\Delta^*}(4 \log d), \tag{1.4}$$

where $\Delta^* = (\Delta_+^{\mathrm{cross}} + \Delta^{\mathrm{diag}})/2 + \max_{1 \le i \le j \le d} |\Delta_{i,j}|/(2 \log d)$. Moreover, for $d \ge 2$,

$$\begin{aligned} &\mathbb{P}\Big\{\max_{1 \le k \le d} Y_k \le t\Big\} - \mathbb{P}\Big\{\max_{1 \le k \le d} X_k \le t\Big\} \\ &\quad \le \left(2\frac{\Delta_+^{\mathrm{cross}} \vee \Delta^{\mathrm{diag}}}{1 - \rho^*} + \frac{\Delta^{\mathrm{diag}}}{2}\right) \min\big\{2 \log d, \big(v_*(t) + 1\big)^2\big\}, \end{aligned} \tag{1.5}$$

where $\rho^* = \max_{i < j \le d} |\mathrm{Corr}(X_i, X_j)| \vee |\mathrm{Corr}(Y_i, Y_j)|$ and $v_*(t) = 1 \vee \max_{i \le d} |t - \mu_i|/\sigma_i$.

Theorem 1 is proved in Section 3. In Theorem 1, (1.4) is a sharper and more explicit version of Corollary 5.1 of [9]. Under the conditions for (1.1), $\Delta_+^{\mathrm{cross}}(s) = \Delta^{\mathrm{diag}}(s) = 0$ in (1.5), so Theorem 1 contains Slepian's inequality as a special case. Inequality (1.5) improves upon (1.4) when $\sqrt{\Delta^*}/(1 - \rho^*)$ is small. Although quantities of different order of smoothness are concerned, the error bounds in Theorem 1 are of a similar form to that of (1.3).

The rest of the paper is organized as follows. We present second-order anticoncentration inequalities in Section 2, comparison theorems for the Gaussian joint distribution functions in Section 3, and higher-order anticoncentration inequalities in Section 4.

We use the following notation to shorten mathematical expressions in the rest of the paper. For positive integers $m < d$, $[d] = \{1, \ldots, d\}$, $i_{1:m} = (i_1, \ldots, i_m)$, $[d]^m = \{i_{1:m} : i_j \in [d]\ \forall\ j \in [m]\}$, $[d]^m_{\ne} = \{i_{1:m} \in [d]^m : i_j \ne i_k\ \forall\ j \ne k\}$, $[d]^m_< = \{i_{1:m} \in [d]^m : i_1 < \cdots < i_m\}$, $[d]_{i_{1:m}} = \{k \in [d] : k \ne i_j\ \forall j \in [m]\}$, and $[d]^2_{i_{1:m},\ne} = \{(j,k) \in [d]^2_{\ne} : j \in [d]_{i_{1:m}}, k \in [d]_{i_{1:m}}\}$. As usual, we denote by $\varphi(t)$ and $\Phi(t)$, respectively, the $N(0,1)$ density and distribution functions, $\|\cdot\|_2$ the Euclidean norm, $\|f\|_{L_\infty} = \sup_{x \in \mathbb{R}^d} |f(x)|$, $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$, and $x_+ = \max(x, 0)$.

2. ANTICONCENTRATION INEQUALITIES FOR GAUSSIAN MAXIMA

Let $X = (X_1, \ldots, X_d)^\top$ be a multivariate Gaussian vector with a joint distribution function

$$G(x) = G(x_1, \ldots, x_d) = \mathbb{P}\big\{X_i \le x_i\ \forall i \in [d]\big\}. \tag{2.1}$$

Let $X_{\max} = \max_{i \in [d]} X_i$ and denote the distribution function of the maximum by

$$G_{\max}(t) = \mathbb{P}\{X_{\max} \le t\} = G(t, \ldots, t). \tag{2.2}$$

While concentration inequalities provide upper bounds for the deviation of $X_{\max}$ from its center, e.g., the median, anticoncentration inequalities bound

$$\mathbb{P}\{a < X_{\max} \le a + \varepsilon\} = G_{\max}(a + \varepsilon) - G_{\max}(a)$$

or the density of $X_{\max}$ from the above.

Among existing results on the anticoncentration of $X_{\max}$, Nazarov's [26] inequality,

$$G_{\max}(a+\varepsilon) - G_{\max}(a) \le \varepsilon \frac{2+\sqrt{2\log d}}{\min_{i\in[d]}\sqrt{\operatorname{Var}(X_i)}}, \quad \forall \varepsilon > 0, \tag{2.3}$$

has found important applications in statistics and machine learning, including bootstrap and central limit theorem [6,8,12,13,19,35]. In terms of the joint distribution G, Nazarov's inequality can be written as an ℓ_1-bound for the gradient of G,

$$\frac{d}{dt}G_{\max}(t) = \sum_{i=1}^{d} \frac{\partial}{\partial x_i} G(x)|_{x_i=t,\forall i\in[d]} \le \frac{2+\sqrt{2\log d}}{\min_{i\in[d]}\sqrt{\operatorname{Var}(X_i)}}. \tag{2.4}$$

In our development of comparison theorems for Gaussian maxima, the second derivative

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\mathbb{P}\Big\{\max_{k\in[d]}(X_k - x_k) \le t\Big\}$$

is involved in Slepian's interpolation. As t can be absorbed into x_k, what we need is a proper upper bound for the second derivative of the distribution function G,

$$G_{i,j}(x) = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\mathbb{P}\big\{X_i \le x_i \ \forall i \in [d]\big\}.$$

In fact, a weighted ℓ_1-norm of $G_{i,j}(x)$ is used in our analysis. Such bounds for the Hessian of $G(x)$ can be viewed as second-order anticoncentration inequalities.

For a standard Gaussian vector $Z = (Z_1, \dots, Z_d)^\top$ with $\mathbb{E}[Z] = 0$ and $\mathbb{E}[ZZ^\top] = I_d$, [3] proved the following anticoncentration inequality of general order:

$$\sup_x \sum_{(i_1,\dots,i_m)\in[d]^m} \left| \frac{\partial^m \mathbb{P}\{Z_k \le x_k \ \forall k \in [d]\}}{\partial x_{i_1}\cdots\partial x_{i_m}} \right| \le C_m (\log d)^{m/2}$$

for some constant C_m depending on m only. Further development of such results and their applications can be found in [2,4,10,11,14,17,20,25,28,36,37] among others. In particular, for applications to Gaussian and bootstrap approximation of the maxima of sums of independent random vectors, the Gaussian vector X was assumed to have a nonsingular covariance matrix Σ and the transformation $Z = \Sigma^{-1/2}X$ was taken to study the anticoncentration of $X_{\max}$ [2,10,14]. The resulting anticoncentration inequality can be written as

$$\sup_x \sum_{(i_1,\dots,i_m)\in[d]^m} \left| \frac{\partial^m G(x)}{\partial x_{i_1}\cdots\partial x_{i_m}} \right| \le \frac{C_m(\log d)^{m/2}}{\lambda_{\min}^{m/2}(\Sigma)}, \tag{2.5}$$

where $\lambda_{\min}$ stands for the smallest eigenvalue. However, the dependence of (2.5) on the smallest eigenvalue is restrictive. We provide below second-order and in Section 4 higher-order anticoncentration inequalities which replace $\lambda_{\min}(\Sigma)$ in (2.5) by the minimum of the eigenvalues of diagonal blocks of dimension m. Such results can be viewed as extensions of Nazarov's inequality (2.4) to higher order.

Before we present the second-order anticoncentration inequality, we give a variation of Nazarov's inequality to explain our approach and write a short proof of it as a road map of the proof in higher order.

Theorem 2. *Let $G(x)$ be the joint distribution function (2.1) of a Gaussian vector $(X_1, \dots, X_d)^\top$ with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $G_i(x) = (\partial/\partial x_i)G(x)$. Let $h(t)$ be a function and $t_0 \ge 0$ such that $h(t)\varphi(t) \le h(t_0)\varphi(t_0)$ for $t \le t_0$, $h(t) \wedge h'(t) \ge 0$ for $t \ge t_0$, and $th(t) - h'(t)$ is nonnegative and increasing in $[t_0, \infty)$. Let $a_1^* = t_0 \vee \max_{i\in[d]}(x_i - \mu_i)/\sigma_i$. Then,*

$$\sum_{i=1}^d \sigma_i G_i(x) h\big((x_i - \mu_i)/\sigma_i\big) \le \min\big\{h\big(a_1^*\big)\big(a_1^* + 1 \wedge \big(1/a_1^*\big)\big), h(t_0 \vee \sqrt{2\log d})\sqrt{2\log d}\big\}, \quad d \ge 2. \tag{2.6}$$

In particular, for $\beta \ge 0$, (2.6) holds for $h(t) = |t|^\beta$ with $t_0 = \sqrt{\beta}$.

For $h(t) = 1$ and $d \ge 2$, (2.6) slightly improves Nazarov's inequality (2.4). Inequality (2.6) with $h(t) = |t|^{m-1}$ is useful in bounding the mth order derivative of $G(x)$. The following corollary demonstrates another way of utilizing the choice $h(\cdot)$ in Theorem 2.

Corollary 1. *Let $d \ge 2$. If $\sigma_i \ge \underline{\sigma} > 0$, then*

$$\frac{d}{dt}\mathbb{P}\Big\{\max_{i\in[d]} X_i \le t\Big\} \le \frac{\sqrt{2\log d}}{\underline{\sigma}}.$$

If $|t - \mu_i| \ge a$ with a certain $a > 0$, then

$$\frac{d}{dt}\mathbb{P}\Big\{\max_{i\in[d]} X_i \le t\Big\} \le (2/a)\log d.$$

Compared with existing literature, Corollary 1 provides an alternative bound to deal with high heteroskedasticity. The second bound in the corollary follows from (2.6) with $h(t) = |t|/a$ as $G_i(x) \le |(x_i - \mu_i)/\sigma_i|\sigma_i G_i(x)/a$. Typically, for $\mathbb{E}[X_i] = 0$, the magnitude of $\mathbb{E}[\max_{i\in[d]} X_i]$ is of the order $\overline{\sigma}\sqrt{\log d}$ for some $\overline{\sigma}$ representing the average of σ_i or larger and the probability outside a small neighborhood of $\mathbb{E}[\max_{i\in[d]} X_i]$ is very small due to Gaussian concentration, so that the most interesting application of (2.4) is for $t \ge a$ with $a \asymp \overline{\sigma}\sqrt{\log d}$. In such applications, Corollary 1 replaces $\min_{i\in[d]} \sigma_i$ in (2.4) with a quantity of the order $\overline{\sigma}$. This is related to a variation of (2.3) in [7] where the $\sqrt{\log d}$ rate is replaced by $O(\mathbb{E}[\max_{k\le d} X_k/\sigma_k] + \sqrt{1 \vee \log(\min_i \sigma_i/\varepsilon)})$ for centered X. Another variation of (2.4) can be found in [13] where the upper bound is

$$1/\sigma_{(1)} + \max_{1\le j\le d}(1 + \sqrt{2\log j})/\sigma_{(j)}$$

where $\sigma_{(1)} \le \cdots \le \sigma_{(d)}$ are the ordered values of $\sigma_1, \dots, \sigma_d$. This variation of Nazarov's inequality is extended to higher orders in Section 4.

Proof of Theorem 2. Let $\phi_i(t) = \varphi((t - \mu_i)/\sigma_i)/\sigma_i$ be the density of X_i, $X_i' = (X_i - x_i)/\sigma_i$ and $\rho_{i,j} = \mathrm{Corr}(X_i, X_j)$. Because X_i is independent of $X_k' - \rho_{i,k}X_i'$,

$$\begin{aligned} G_i(x) &= \mathbb{P}\big\{X_k \le x_k, \forall k \in [d]_i | X_i = x_i\big\}\phi_i(x_i) \\ &= \mathbb{P}\big\{X_k' \le \rho_{i,k}X_i', \forall k \in [d]_i\big\}\phi_i(x_i). \end{aligned} \tag{2.7}$$

Let $\pi_i = \mathbb{P}\{X_i' > 0\}G_i(x)/\phi_i(x_i)$, $\nu_i(x) = (x_i - \mu_i)/\sigma_i$ and $R_i = \sum_{k\in[d]} I\{X_k' \geq X_i'\}$ be the rank of X_i'. Due to the independence of X_i' and $\{X_k' - \rho_{i,k}X_i', k \in [d]_i\}$ and the fact that $\{X_k' \leq \rho_{i,k}X_i', k \in [d]_i, X_i' > 0\} \subseteq \{R_i = 1\}$,

$$\pi_i = \mathbb{P}\big\{X_k' \leq \rho_{i,k}X_i', \forall k \in [d]_i\big\}\Phi\big(-\nu_i(x)\big) \leq \mathbb{P}\{R_i = 1\}, \tag{2.8}$$

so that $\sum_{i\in[d]} \pi_i \leq 1$. Let $\psi_1(t) = \Phi(-t)/\varphi(t)$. We have $h(\nu_i(x))/\psi_1(\nu_i(x)) \leq h(a_1^*)/\psi_1(a_1^*)$ because $a_1^* \geq t_0$, $h(t)/\psi_1(t) \leq h(t_0)/\psi_1(t_0)$ for $t \leq t_0$ and $h(t)/\psi_1(t)$ is nondecreasing for $t \geq t_0$. Because $\sigma_i G_i(x) = \pi_i/\psi_1(\nu_i(x))$ by (2.7) and (2.8), the first upper bound in (2.6) follows from $1/\psi_1(t) \leq t + \psi_1(t)$ and $h(t) \wedge h'(t) \geq 0$ for $t \geq t_0$. For the second upper bound in (2.6), (2.7) and (2.8) yield

$$\begin{aligned}
&\sum_{i=1}^d \sigma_i G_i(x)h\big(\nu_i(x)\big) \\
&\quad\leq \max_{\nu_i,\pi_i,i\in[d]} \left\{ \sum_{\nu_i\leq t_0} \frac{h(t_0)\varphi(t_0)\pi_i}{\Phi(-\nu_i)} + \sum_{\nu_1>t_0} \frac{h(\nu_i)\pi_i}{\psi_1(\nu_i)} : \sum_{i=1}^d \pi_i \leq 1, 0 \leq \pi_i \leq \Phi(-\nu_i) \right\}. \\
&\quad= \max_{\nu_i\geq t_0,\pi_i,i\in[d]} \left\{ \sum_{i=1}^d h(\nu_i)\varphi(\nu_i) : \sum_{i=1}^d \Phi(-\nu_i) \leq 1 \right\}
\end{aligned} \tag{2.9}$$

due to $h(t)\varphi(t) \leq h(t_0)\varphi(t_0)$ in $(\infty, t_0]$ and the monotonicity of $\Phi(t)$ in $\mathbb{R}$ and $h(t)/\psi_1(t)$ in $[t_0, \infty)$. The global maximum on the right-hand side of (2.9) must be attained when $\nu_i h(\nu_i) - h'(\nu_i) = \lambda$ for all i with a Lagrange multiplier λ. As $th(t) - h'(t) = \lambda$ has one solution in $[t_0, \infty)$, the global maximum is attained at $\nu_i = t$ for all $i \in [d]$ and some $t \geq t_0$. As $(d/dt)\{h(t)\varphi(t)\} \leq 0$ for $t \geq t_0$, the maximum is attained at $t = t_0 \vee t_1$ and given by $dh(t_0 \vee t_1)\varphi(t_0 \vee t_1)$, where t_1 is the solution of $\Phi(-t_1) = 1/d$. This gives the second upper bound in (2.6) because $t_1 \leq \sqrt{2\log d}$ and $d\varphi(t_1) = 1/\psi_1(t_1) \leq \sqrt{2\log d}$ for $d \geq 2$. ■

To extend Theorem 2 to the second order, we need to define certain quantities $\alpha_{i,j}$ as an extension of the weights σ_i. Let $\Sigma^{i,j}$ and $\phi_{i,j}(\cdot)$ be respectively the covariance matrix and joint density of $(X_i, X_j)^\top$. As $1/\sigma_i = \max_t \sqrt{2\pi}\phi_i(t)$, $\alpha_{i,j}$ is expected to involve $|\det(\Sigma^{i,j})|^{1/2}$ as the Jacobian in the denominator of $\phi_{i,j}(\cdot)$. However, $\alpha_{i,j}$ also involves a certain threshold level t_i for a two-dimensional extension of (2.8). Let $\rho_{i,j} = \mathrm{Corr}(X_i, X_j)$. The threshold level t_i is defined as

$$t_i = \min\Big\{\sqrt{(1-\rho_{i,j})/(1+\rho_{i,j})} : j \in [d]_i\Big\}, \tag{2.10}$$

which can be viewed as the tangent of the minimum half-angle between standardized X_i and X_j in $L_2(\mathbb{P})$. Let $Y_i = (X_i - \mu_i)/\sigma_i$ be the standardized X_i, $\theta_{i,j} = \arccos(\rho_{i,j}) \in [0, \pi]$ be the $L_2(\mathbb{P})$ angle between Y_i and Y_j, and $\theta_{i,\min} = \min\{\theta_{i,j} : j \in [d]_i\}$ be the angle between Y_i and its nearest neighbor. The threshold level in (2.10) can be written as

$$t_i = \tan(\theta_{i,\min}/2).$$

The quantity $\alpha_{i,j}$ is then defined as

$$\begin{aligned}\alpha_{i,j} &= 2\tan(\theta_{i,\min}/4)\big|\det\big(\Sigma^{i,j}\big)\big|^{1/2}\\ &= 2\tan\big(\arctan(t_i)/2\big)\sigma_i\sigma_j\big(1-\rho_{i,j}^2\big)^{1/2}\end{aligned} \tag{2.11}$$

with $\tan(\theta_{i,\min}/4) \in [0,1]$ and t_i as in (2.10). We note that $2\tan(\theta_{i,\min}/4) \approx t_i$ when t_i is small. We also consider quantities

$$\tilde{\nu}_{i,j} = \tilde{\nu}_{i,j}(x) = \big(\nu_i^2 + \nu_{j|i}^2\big)^{1/2} \wedge \big(\nu_i + t_i(\nu_{j|i})_+\big)_+ \tag{2.12}$$

as signed versions of

$$\nu_{i,j} = \nu_{i,j}(x) = \big(\nu_i^2 + \nu_{j|i}^2\big)^{1/2}, \tag{2.13}$$

where $\nu_i = \nu_i(x) = (x_i - \mu_i)/\sigma_i$ and $\nu_{j|i} = \nu_{j|i}(x) = (\nu_j - \rho_{i,j}\nu_i)/(1-\rho_{i,j}^2)^{1/2}$. We are now ready to state a second order Gaussian anticoncentration theorem.

Theorem 3. *Let $d \geq 2$ and $X = (X_1, \ldots, X_d)^\top$ be a Gaussian vector with a joint distribution function $G(x)$. Let $G_{i,j}(x) = (\partial/\partial x_i)(\partial/\partial x_j)G(x)$, $\alpha_{i,j}$ as in (2.11), $\rho_{i,j} = \mathrm{Corr}(X_i, X_j)$, and $a_2^* = a_2^*(x) = \sqrt{2} \vee \max_{i,j} \tilde{\nu}_{i,j}(x)$ with $\tilde{\nu}_{i,j}(x)$ as in (2.12). Then,*

$$\sum_{(i,j)\in[d]^2_{\neq}} \alpha_{i,j}G_{i,j}(x) \leq \min\big\{(1/\pi) \vee \big(2\log\big(d(d-1)/2\big)\big), \big(a_2^* + \sqrt{2}\big)^2\big\}. \tag{2.14}$$

Moreover, with $a_1^* = 1 \vee \max_{i\in[d]}(x_i - \mu_i)/\sigma_i$,

$$\sum_{i=1}^d \Big|\sigma_i^2 G_{i,i}(x) + \sum_{j\in[d]_i} \rho_{i,j}\sigma_i\sigma_j G_{i,j}(x)\Big| \leq \min\big\{2\log d, \big(a_1^*\big)^2 + 1\big\}. \tag{2.15}$$

Before we move ahead to proving Theorem 3, we state in the following corollary a scaled ℓ_1-bound for the Hessian of the joint distribution function $G(x)$ as a direct consequence of the theorem using $\tan(\theta_{i,\min}/4) \geq \sqrt{(1 - \max_{k\neq i}\rho_{i,k})/8}$ in (2.11).

Corollary 2. *With $\sigma_i = \mathrm{Var}^{1/2}(X_i)$ and $\rho_{i,j} = \mathrm{Corr}(X_i, X_j)$,*

$$\sum_{i=1}^d\sum_{j=1}^d \sigma_i\sigma_j\big|G_{i,j}(x)\big| \leq \max_{(i,j,k)\in[d]^3_{\neq}} \frac{8\log d}{\sqrt{(1-|\rho_{i,j}|)(1-\rho_{j,k})}} + 2\log d, \quad d \geq 2.$$

Proof of Theorem 3. To prove (2.14), we define

$$X_i' = \frac{X_i - x_i}{\sigma_i}, \quad X_{j|i}' = \frac{X_j' - \rho_{i,j}X_i'}{(1-\rho_{i,j}^2)^{1/2}}, \quad \rho_{(j,k)|i} = \mathrm{Corr}\big(X_j', X_k'|X_i'\big). \tag{2.16}$$

Let $\phi_{i,j}(\cdot)$ be the joint density of $(X_i, X_j)^\top$. As in (2.7), it holds for all $(i,j) \in [d]^2_{\neq}$ that

$$\begin{aligned}G_{i,j}(x) &= \mathbb{P}\big\{X_k' < 0,\ \forall k \in [d]_{i,j}|X_i' = X_j' = 0\big\}\phi_{i,j}(x_i, x_j)\\ &= \mathbb{P}\big\{X_{k|i}' - \rho_{(j,k)|i}X_{j|i}' < 0,\ \forall k \in [d]_{i,j}\big\}\phi_{i,j}(x_i, x_j).\end{aligned} \tag{2.17}$$

For the second step above, we note that $X'_{k|i} - \rho_{(j,k)|i} X'_{j|i}$ is independent of $(X'_i, X'_j)^\top$. Similar to the proof leading to (2.9), we set $\mathscr{C}_{i,j} = \{0 < X'_{j|i} < t_i X'_i\}$ with the threshold level t_i in (2.10), and define

$$\pi_{i,j} = \mathbb{P}\{\mathscr{C}_{i,j}\} G_{i,j}(x)/\phi_{i,j}(x_i, x_j). \tag{2.18}$$

Let $R_i = \sum_{j=1}^d I\{X'_j \geq X'_i\}$ and $R_{j|i} = \sum_{k\in[d]_i} I\{X'_{k|i} \geq X'_{j|i}\}$ be respectively the marginal and conditional ranks of X'_i and $X'_{j|i}$ in (2.16). By the definition of t_i, $t_i \leq \sqrt{(1-\rho_{i,k})/(1+\rho_{i,k})}$, so that $\rho_{i,k} + (1-\rho_{i,k}^2)^{1/2} t_i \leq 1$ for all $k \in [d]_i$. It follows that

$$\begin{aligned}
\pi_{i,j} &= \mathbb{P}\big\{X'_{k|i} - \rho_{(j,k)|i} X'_{j|i} < 0 \ \forall k \in [d]_{i,j}\big\} \mathbb{P}\{\mathscr{C}_{i,j}\} \\
&= \mathbb{P}\big\{X'_{k|i} < \rho_{(j,k)|i} X'_{j|i}, \forall k \in [d]_{i,j}, 0 \leq X'_{j|i} < t_i X'_i, X'_i > 0\big\} \\
&\leq \mathbb{P}\big\{R_{j|i} = 1,\ X'_{k|i} < t_i X'_i,\ \forall k \in [d]_i, X'_i > 0\big\} \\
&= \mathbb{P}\big\{R_{j|i} = 1,\ X'_k < \big(\rho_{i,k} + \big(1-\rho_{i,k}^2\big)^{1/2} t_i\big) X'_i \leq X'_i,\ \forall k \in [d]_i\big\} \\
&\leq \mathbb{P}\{R_{j|i} = 1, R_i = 1\}.
\end{aligned}$$

Consequently,

$$\sum_{(i,j)\in[d]^2_{\neq}} \pi_{i,j} \leq 1. \tag{2.19}$$

We still need a lower bound for $\mathbb{P}\{\mathscr{C}_{i,j}\}$ to use (2.19). To this end, we prove

$$\begin{aligned}
\mathbb{P}\{\mathscr{C}_{i,j}\} &= \pi_{i,j} \phi_{i,j}(x_i, x_j)/G_{i,j}(x) \\
&\geq 2\tan(\theta_{i,\min}/4) \varphi(\nu_{i,j}) \varphi(0) \psi_2(\tilde{\nu}_{i,j}) \\
&= \alpha_{i,j} \phi_{i,j}(x_i, x_j) \psi_2(\tilde{\nu}_{i,j}),
\end{aligned} \tag{2.20}$$

where $\nu_{i,j} = \nu_{i,j}(x)$ are as in (2.13), $\tilde{\nu}_{i,j} = \tilde{\nu}_{i,j}(x) = \min\{\nu_{i,j}, (\nu_i + t_i(\nu_{j|i})_+)_+\}$ are as in (2.12), $\alpha_{i,j}$ and $\theta_{i,\min}$ are as in (2.11), and

$$\psi_2(t) = \int_0^\infty \int_0^{y_1} e^{-y_2^2/2 - t y_1 - y_1^2/2} dy_2 dy_1. \tag{2.21}$$

Moreover, with $\psi_1(t) = \Phi(-t)/\varphi(t)$ as in (2.9), we prove that for all $t \geq 0$,

$$1/\psi_2(t) \leq 1/\psi_1^2(t) + 1 + 2/\big(1 + \psi_1^2(t)\big). \tag{2.22}$$

The first equality in (2.20) is from the definition of $\pi_{i,j}$ in (2.18), and the last follows from $\varphi(\nu_i)\varphi(\nu_{j|i}) = |\det(\Sigma^{i,j})|^{1/2} \phi_{i,j}(x_i, x_j)$ and the definition of $\alpha_{i,j}$ in (2.11). We note that $\nu_{i,j} = (\nu_i^2 + \nu_{j|i}^2)^{1/2}$ and the variables $X'_i \sim N(-\nu_i, 1)$ and $X'_{j|i} \sim N(-\nu_{j|i}, 1)$ are independent by (2.16). It follows that

$$\mathbb{P}\{\mathscr{C}_{i,j}\} = \int_0^\infty \int_0^{t_i y_1} \varphi(y_1 + \nu_i) \varphi(y_2 + \nu_{j|i}) dy_2 dy_1 \tag{2.23}$$

with $t_i = \tan(\theta_{i,\min}/2)$. Given $\nu_{i,j} = (\nu_i^2 + \nu_{j|i}^2)^{1/2}$, the above integral is minimized when $\nu_i \wedge \nu_{j|i} \geq 0$ and $\nu_{j|i}/\nu_i = \tan(\theta_{i,\min}/4)$. Thus, after proper rotation

$$\mathbb{P}\{\mathscr{C}_{i,j}\} \geq \int_0^\infty \int_{|y_2| \leq \tan(\theta_{i,\min}/4) y_1} \varphi(y_1 - \nu_{i,j}) \varphi(y_2) dy_2 dy_1,$$

which implies the inequality in (2.20) for $\nu_{i,j} = \tilde{\nu}_{i,j}$. For $\nu_{i,j} > \tilde{\nu}_{i,j}$ and $0 < y_2 \le t_i y_1$, $\nu_i y_1 + \nu_{j|i} y_2 \le \nu_i y_1 + t_i (\nu_{j|i})_+ y_1 \le \tilde{\nu}_{i,j} y_1$, so that by (2.23)

$$\mathbb{P}\{\mathscr{C}_{i,j}\} \ge \frac{\varphi(\nu_{i,j})}{\sqrt{2\pi}} \int_0^\infty \int_0^{t_i y_1} e^{-y_1^2/2 - y_2^2/2 - \tilde{\nu}_{i,j} y_1} dy_2 dy_1,$$

which again implies the inequality in (2.20). For (2.22), we note that by (2.21)

$$\psi_2(t) = \int_0^\infty \{\psi_1((t + y_1)/\sqrt{2})/\sqrt{2}\} e^{-t y_1 - y_1^2/2} dy_1.$$

As in Lemma 9 of [13], $1/(t + \psi_1(t)) < \psi_1(t) < 1/t$, so that

$$\int_0^\infty y_1 e^{-t y_1 - y_1^2/2} dy_1 = 1 - t\psi_1(t) \le \psi_1^2(t).$$

As $\psi_1(\cdot)$ is convex and decreasing in $[0, \infty)$, an application of Jensen's inequality yields $\psi_2(t) \ge \{\psi_1(t)/\sqrt{2}\}\psi_1((t + \psi_1(t))/\sqrt{2})$. Thus, as $(1/t)/(1 + 1/t^2) < \psi_1(t) < 1/t$,

$$\psi_2(t) \ge \frac{\psi_1(t)}{\sqrt{2}} \psi_1\left(\frac{1 + \psi_1^2(t)}{\sqrt{2}\psi_1(t)}\right) \ge \frac{\psi_1^2(t)/(1 + \psi_1^2(t))}{1 + 2\psi_1^2(t)/(1 + \psi_1^2(t))^2},$$

which gives (2.22).

Let $\pi'_{i,j} = \alpha_{i,j} G_{i,j}(x) \psi_2(\tilde{\nu}_{i,j})$. It follows from (2.20) that $\pi'_{i,j} \le \pi_{i,j}$. By (2.11) and (2.17),

$$\pi'_{i,j} \le 2\big|\det\big(\Sigma^{i,j}\big)\big|^{1/2} \phi_{i,j}(x_i, x_j) \psi_2(\tilde{\nu}_{i,j}) = \sqrt{2/\pi}\varphi(\nu_{i,j})\psi_2(\tilde{\nu}_{i,j}). \tag{2.24}$$

This gives (2.14) for $d = 2$ as $\alpha_{1,2} G_{1,2}(x) \le 1/\pi$. By (2.19) and (2.22),

$$\sum_{(i,j)\in[d]^2_{\ne}} \alpha_{i,j} G_{i,j}(x) = \sum_{(i,j)\in[d]^2_{\ne}} \frac{\pi'_{i,j}}{\psi_2(\tilde{\nu}_{i,j})} \le \big(a_2^* + \sqrt{2}\big)^2$$

due to $a_2^* = \sqrt{2} \vee \max_{(i,j)\in[d]^2_{\ne}} \tilde{\nu}_{i,j}$ and $1/\psi_1(t) \le t + 1/t$. In general, (2.19) and (2.24) yield

$$\begin{aligned}
&\sum_{(i,j)\in[d]^2_{\ne}} \alpha_{i,j} G_{i,j}(x) \\
&\le \max_{\nu_{i,j}\ge 0, \pi'_{i,j}} \left\{ \sum_{(i,j)\in[d]^2_{\ne}} \frac{\pi'_{i,j}}{\psi_2(\nu_{i,j})} : \sum_{(i,j)\in[d]^2_{\ne}} \pi'_{i,j} \le 1, \pi'_{i,j} \le \sqrt{2/\pi}\varphi(\nu_{i,j})\psi_2(\nu_{i,j}) \right\} \\
&= \max_{\nu_{i,j}\ge 0} \left\{ \sum_{(i,j)\in[d]^2_{\ne}} \sqrt{2/\pi}\varphi(\nu_{i,j}) : \sum_{(i,j)\in[d]^2_{\ne}} \sqrt{2/\pi}\varphi(\nu_{i,j})\psi_2(\nu_{i,j}) \le 1 \right\}
\end{aligned} \tag{2.25}$$

because $\psi_2(t)$ and $\varphi(t)$ are both decreasing in $[0, \infty)$. Let $d_2 = d(d-1)/2$. By (2.21), $\psi_2(t) - \psi_2'(t)/t$ is decreasing in t in $[0, \infty)$, so that the optimization problem is solved by $\nu_{i,j} = t_2$ with a Lagrange multiplier, where t_2 is the solution of $\sqrt{2/\pi}\varphi(t_2)\psi_2(t_2) = 1/(2d_2)$. For $d \ge 3$ and $t = \sqrt{(2\log(2d_2/(2\pi \log d_2)))_+}$, we have $1/\psi_2(t) \le 2 \log d_2$ via (2.22). Thus, the right-hand side of (2.25) is no greater than $2 \log d_2$.

Finally, it follows from (2.16) and (2.7) that

$$G_{i,i}(x) = -G_i(x)\nu_i(x)/\sigma_i - \sum_{j\in[d]_i} G_{i,j}(x)\rho_{i,j}\sigma_j/\sigma_i \tag{2.26}$$

with $\nu_i(x) = (x_i - \mu_i)/\sigma_i$, so that (2.15) follows from Theorem 2 with $h(t) = |t|$. ∎

3. COMPARISON OF GAUSSIAN DISTRIBUTION FUNCTIONS

The Gaussian anticoncentration theorem in Section 2 yields the following error bounds in the comparison of Gaussian joint distribution functions.

Let $X = (X_1, \ldots, X_d)^\top$ and $Y = (Y_1, \ldots, Y_d)^\top$ be two Gaussian vectors with common mean μ and respective covariance matrices Σ^X and Σ^Y and joint distribution functions

$$G^X(x) = \mathbb{P}\{X_k \le x_k \forall k \in [d]\}, \quad G^Y(y) = \mathbb{P}\{Y_k \le y_k \forall k \in [d]\}. \tag{3.1}$$

For $0 \le s \le 1$, let $\Sigma_{i,j}(s)$ be the elements of $\Sigma(s) = (1-s)\Sigma^X + s\Sigma^Y$,

$$\nu_i(x;s) = (x_i - \mu_i)/\sqrt{\Sigma_{i,i}(s)}, \tag{3.2}$$

$$\Delta_{i,j}(s) = \frac{\Sigma^Y_{i,j} - \Sigma^X_{i,j}}{\sqrt{\Sigma_{i,i}(s)\Sigma_{j,j}(s)}}, \quad \rho_{i,j}(s) = \frac{\Sigma_{i,j}(s)}{\sqrt{\Sigma_{i,i}(s)\Sigma_{j,j}(s)}}, \tag{3.3}$$

and

$$\Delta_{i,j,\pm}(s) = \max_{k\ne i,\ell\ne j} \frac{(2\Delta_{i,j}(s))_\pm \vee |\Delta_{i,i}(s) + \Delta_{j,j}(s)|}{\sqrt{(1-|\rho_{i,j}(s)|)}(\sqrt{1-\rho_{i,k}(s)} + \sqrt{1-\rho_{j,\ell}(s)})}. \tag{3.4}$$

Theorem 4. *Let $G^X(x)$ and $G^Y(y)$ be as in (3.1), $\nu_*(x;s) = 1 \vee \max_{i\in[d]} |\nu_i(x;s)|$, $\Delta^*_+(s) = \max_{(i,j)\in[d]^2_{\ne}} \Delta_{i,j,+}(s)$ and $\Delta^{\mathrm{diag}}(s) = \max_{i\in[d]} |\Delta_{i,i}(s)|$, where $\nu_i(x;s)$, $\Delta_{i,i}(s)$, and $\Delta_{i,j,\pm}(s)$ are as in (3.2), (3.3), and (3.4), respectively. Then, for $d \ge 2$,*

$$G^Y(x) - G^X(x) \le \int_0^1 \big(2\Delta^*_+(s) + \Delta^{\mathrm{diag}}(s)/2\big) \min\big\{2\log d, \big(\nu_*(x,s)+1\big)^2\big\} ds. \tag{3.5}$$

Because $G^X_{\max}(t) = G^X(t,\ldots,t)$ and $G^Y_{\max}(t) = G^Y(t,\ldots,t)$, Theorem 1 is an immediate consequence of Theorem 4. Conversely, as t can be absorbed into the mean, Theorem 1 is a simplified version of Theorem 4.

Assume, without loss of generality, that X and Y are independent as the theorem does not involve the joint distribution of X and Y. With $\mu = \mathbb{E}[X]$, write $X(s) = \sqrt{1-s}(X-\mu) + \sqrt{s}(Y-\mu) + \mu$, $s \in [0,1]$, as Slepian's interpolation and

$$\phi(x;s) = (2\pi)^{-d} \int_{\mathbb{R}^d} \exp\big[\sqrt{-1}(\mu - x)^\top u - u^\top\big((1-s)\Sigma^X + s\Sigma^Y\big)u/2\big]du$$

as the joint density of $X(s)$. Slepian's inequality was proved by passing the differentiation of $\mathbb{E}[f(X(s))]$ to twice differentiation of f through the above formula,

$$\frac{d}{ds}\mathbb{E}\big[f\big(X(s)\big)\big] = \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d\big(\Sigma^Y_{i,j} - \Sigma^X_{i,j}\big)\int\left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)\phi(x;s)dx, \tag{3.6}$$

provided the twice differentiability of $f(x)$. However, for comparison of distribution functions, this is not feasible as f is an indicator function. Instead, with $y = (y_1,\ldots,y_d)^\top$, we may exchange the differentiation and integration in (3.6) and write

$$\frac{d}{ds}\mathbb{E}\big[f\big(X(s)\big)\big] = \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d\big(\Sigma^Y_{i,j} - \Sigma^X_{i,j}\big)\frac{\partial^2 F(y;s)}{\partial y_i \partial y_j}\bigg|_{y=0} \tag{3.7}$$

with $F(y;s) = \int f(x)\phi(x+y;s)dx = \mathbb{E}[f(X(s)-y)]$. In the proof of Theorem 4, we directly apply the weighted anticoncentration inequality in Theorem 3 to (3.7).

Proof of Theorem 4. Let $\sigma_i(s) = \Sigma_{i,i}^{1/2}(s)$ and

$$G(x;s) = \mathbb{P}\{X_k(s) \le x_k, k \in [d]\},$$

so that (3.7) becomes

$$(\partial/\partial s)G(x;s) = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\big(\Sigma_{i,j}^{Y} - \Sigma_{i,j}^{X}\big)G_{i,j}(x;s), \tag{3.8}$$

where $G_{i,j}(x;s) = (\partial/\partial x_i)(\partial/\partial x_j)G(x;s)$. As in (2.11), let

$$\alpha_{i,j}(s) = 2\sigma_i(s)\sigma_j(s)\big(1-\rho_{i,j}^2(s)\big)^{1/2}\tan\big(\theta_{i,\min}(s)/4\big),$$

with $\theta_{i,\min}(s) = \min\{\arccos(\rho_{i,k})(s), k \in [d]_i\} \in [0,\pi]$. We have

$$\frac{\sigma_i(s)\sigma_j(s)(1+|\rho_{i,j}(s)|)}{\alpha_{i,j}(s)+\alpha_{j,i}(s)} \le \max_{k\ne i,\ell\ne j}\frac{\sqrt{2(1+|\rho_{i,j}(s)|)}}{\sqrt{1-|\rho_{i,j}(s)|}(\sqrt{(1-\rho_{i,k})}+\sqrt{(1-\rho_{j,\ell})})}$$

due to $\tan(\theta_{i,\min}/4) \ge \sqrt{(1-\max_{k\ne i}\rho_{i,k})/8}$. Let $v_i(x;s) = (x_i-\mu_i)/\sigma_i(s)$. We use $\sigma_i(s)$ to scale (3.8) and apply (2.26) and Theorem 3 as follows:

$$\begin{aligned}
&(\partial/\partial s)G(x;s)\\
&\quad= \frac{1}{2}\sum_{(i,j)\in[d]^2_{\ne}}\Delta_{i,j}(s)\sigma_i(s)\sigma_j(s)G_{i,j}(x;s) - \frac{1}{2}\sum_{i=1}^{d}\Delta_{i,i}(s)\sigma_i(s)G_i(x;s)v_i(x;s)\\
&\qquad-\frac{1}{4}\sum_{(i,j)\in[d]^2_{\ne}}\big(\Delta_{i,i}(s)+\Delta_{j,j}(s)\big)\rho_{i,j}(s)\sigma_i(s)\sigma_j(s)G_{i,j}(x;s)\\
&\quad\le \sum_{(i,j)\in[d]^2_{\ne}}\Delta_{i,j,+}(s)\left(\frac{\alpha_{i,j}(s)+\alpha_{j,i}(s)}{2}\right)G_{i,j}(x;s)\\
&\qquad+\frac{1}{2}\sum_{i=1}^{d}\big|\Delta_{i,i}(s)v_i(x;s)\big|\sigma_i(s)G_i(x;s)\\
&\quad\le \max_{(i,j)\in[d]^2_{\ne}}\Delta_{i,j,+}(s)\{(\sqrt{2}v_*+\sqrt{2})^2\wedge(4\log d)\}\\
&\qquad+\max_{i\in[d]}\big|\Delta_{i,i}(s)/2\big|\{(v_*+1)^2\wedge(2\log d)\}.
\end{aligned}$$

This gives (3.5) by integrating over $s \in [0,1]$. ∎

In the rest of this section we prove Theorem 1.

Proof of Theorem 1. Let $\mathrm{Err}_t = \mathbb{P}\{\max_{1\le i\le d}Y_i \le t\} - \mathbb{P}\{\max_{1\le i\le d}X_i \le t\}$ and write

$$\mathrm{Err}_t = \mathbb{P}\Big\{\max_{1\le i\le d}Y_i' \le 0\Big\} - \mathbb{P}\Big\{\max_{1\le i\le d}X_i' \le 0\Big\},$$

with $X_i' = (X_i-t)/\sigma_i$ and $Y_i' = (Y_i-t)/\sigma_i$. Let $\varepsilon \ge \varepsilon' > 0$, $\beta = (\log d)/(2\varepsilon')$, $g(x) = \beta^{-1}\log(\sum_{i=1}^{d}e^{\beta x_i})$ for $x = (x_1,\dots,x_d)^\top$, and $f_\varepsilon(t)$ be the nonincreasing function

with $f_\varepsilon(\varepsilon) = 0$ and derivative $f_\varepsilon'(t) = -\varepsilon^{-1}(1 - |t|/\varepsilon)_+$. Let $x_{\max} = \max_{1\le i\le d} x_i$. Similar to [5], we approximate $I\{x_{\max} \le 0\}$ by $f_\varepsilon(g(x) - \varepsilon')$. Because $x_{\max} \le g(x) \le x_{\max} + 2\varepsilon'$,

$$\begin{aligned} &I\{y_{\max} \le 0\} - I\{x_{\max} \le 0\} - f_\varepsilon\big(g(y) - \varepsilon'\big) + f_\varepsilon\big(g(x) - \varepsilon'\big) \\ &\quad \le I\{y_{\max} \le 0\}\big\{1 - f_\varepsilon\big(y_{\max} + \varepsilon'\big)\big\} + I\{x_{\max} > 0\} f_\varepsilon\big(x_{\max} - \varepsilon'\big), \end{aligned}$$

for any x and $y = (y_1, \dots, y_d)^\top$, where $y_{\max} = \max_{1\le i\le d} y_i$. Set $\varepsilon'/\varepsilon = 3/10$. As $\operatorname{Var}(X_i') \wedge \operatorname{Var}(Y_i') \ge 1$ and $f(t) + f(-t) = 1$, Corollary 1 provides

$$\begin{aligned} &\operatorname{Err}_t - \mathbb{E}\big[f\big(g(Y)\big)\big] + \mathbb{E}\big[f\big(g(X)\big)\big] \\ &\quad \le \sqrt{2\log d}\left(\int_{-\varepsilon-\varepsilon'}^{0} \big(1 - f_\varepsilon\big(t + \varepsilon'\big)\big)dt + \int_0^{\varepsilon+\varepsilon'} f_\varepsilon\big(t - \varepsilon'\big)dt\right) \\ &\quad = \sqrt{2\log d}\big\{2\varepsilon' + \varepsilon\big(1 - \varepsilon'/\varepsilon\big)^3/3\big\} \\ &\quad \le (3\varepsilon/4)\sqrt{2\log d}. \end{aligned} \tag{3.9}$$

The approximation allows us to apply (3.6) to X' and Y'. Let $p_i = p_i(x) = e^{\beta x_i}/\sum_{j=1}^d e^{\beta x_j}$. We have $\partial g(x)/\partial x_i = p_i$ and $\partial p_i/\partial x_j = \beta I_{\{i=j\}} p_i - \beta p_i p_i$. It follows that

$$\begin{aligned} |\operatorname{Err}_t| &\le (3\varepsilon/4)\sqrt{2\log d} + \frac{1}{2}\int_{\mathbb{R}^d} \beta f_\varepsilon'\big(g(x)\big) \sum_{i=1}^d p_i \Delta_{i,i}\phi(x;s)dx \\ &\quad + \frac{1}{2}\int_{\mathbb{R}^d} \big\{f_\varepsilon''\big(g(x)\big) - \beta f_\varepsilon'\big(g(x)\big)\big\} \sum_{i=1}^d \sum_{j=1}^d p_i p_j \Delta_{i,j}\phi(x;s)dx \\ &\le (3\varepsilon/4)\sqrt{2\log d} + \frac{\Delta}{2}\int_{\mathbb{R}^d} \big|f_\varepsilon''\big(g(x)\big)\big|\phi(x;s)dx \\ &\quad + \frac{(\Delta^{\text{diag}} + \Delta_+^{\text{cross}})\log d}{4\varepsilon'}\int_{\mathbb{R}^d} \big|f_\varepsilon'\big(g(x)\big)\big|\phi(x;s)dx, \end{aligned} \tag{3.10}$$

due to $f_\varepsilon'(t) \le 0$, where $\Delta = \max_{1\le i\le j\le d} |\Delta_{i,j}|$. Similar to (3.9), we have

$$\frac{1}{4\varepsilon'\sqrt{2\log d}}\int_{\mathbb{R}^d} \big|f_\varepsilon'\big(g(x)\big)\big|\phi(x;s)dx \le \frac{1 + 2\varepsilon'/\varepsilon}{4\varepsilon'} = \frac{4}{3\varepsilon}$$

and $\int_{\mathbb{R}^d} |f_\varepsilon''(g(x))|\phi(x;s)dx/(2\sqrt{2\log d}) \le (\varepsilon'/\varepsilon + 1)/\varepsilon \le 4/(3\varepsilon)$. Inserting the above bounds for the integrals into (3.10), we find that

$$\frac{|\operatorname{Err}_t|}{\sqrt{2\log d}} \le \frac{3\varepsilon}{4} + \frac{4}{3\varepsilon}\big\{\big(\Delta^{\text{diag}} + \Delta_+^{\text{cross}}\big)\log d + \Delta\big\}.$$

This gives (1.4) with ε minimizing the right-hand side. Theorem 4 implies (1.5) due to $|\rho_{i,j}(s)| \le \rho^*\{(1-s)\sqrt{\sigma_i(0)\sigma_j(0)} + s\sqrt{\sigma_i(1)\sigma_j(1)}\}/\{\sigma_i(s)\sigma_j(s)\} \le \rho^*$ in (3.4). ∎

4. HIGHER-ORDER ANTICONCENTRATION

In this section we extend Theorem 3 to higher order by developing upper bounds for weighted sums of the absolute values of the derivatives

$$G_{i_1,\dots,i_m}(x) = \frac{\partial^m \mathbb{P}\{X_k \le x_k \ \forall k \in [d]\}}{\partial x_{i_1}\cdots\partial x_{i_m}} \tag{4.1}$$

for Gaussian vectors $X = (X_1, \ldots, X_d)^\top$, where $x = (x_1, \ldots, x_d)^\top$. We shall defer proofs to the end of the section after the statement and discussion of these extensions.

We first present an mth-order anticoncentration inequality in terms of partial correlations between the components of the Gaussian vector X. For $i_{1:m} = (i_1, \ldots, i_m) \in [d]^m_{\neq}$ and $(j,k) \in [d]^2_{i_{1:m},\neq}$, the partial correlation of X_j and X_k given $X_{i_{1:m}} = (X_{i_1}, \ldots, X_{i_m})^\top$ is

$$\rho_{j,k|i_{1:m}} = \mathrm{Corr}(X_j, X_k | X_{i_{1:m}}) \tag{4.2}$$

with the convention $\rho_{(j,k)|i_{1:0}} = \rho_{j,k} = \mathrm{Corr}(X_j, X_k)$. Define threshold levels

$$t_{i_{1:j}} = 1 \wedge \min\left\{ \frac{\sqrt{1 - \rho_{i_j,k|i_{1:(j-1)}}}}{\sqrt{1 + \rho_{i_j,k|i_{1:(j-1)}}}}, k \in [d]_{i_{1:j}} \right\}, \quad i_{1:j} \in [d]^j_{\neq}, \tag{4.3}$$

and an extension of a simplification of (2.11) as

$$\alpha'_{i_{1:m}} = \left|\det\left(\Sigma^{i_{1:m}}\right)\right|^{1/2} \prod_{j=1}^{m-1} t_{i_{1:j}}, \quad i_{1:m} \in [d]^m_{\neq}, \tag{4.4}$$

with $\alpha'_i = \sigma_i = \Sigma^{1/2}_{i,i}$, where $\Sigma^{i_{1:m}}$ is the $m \times m$ covariance matrix of $X_{i_{1:m}}$. Compared with (2.11) where $\cos(\theta_{i,\min}) = \rho_{i,\max} = \max\{\rho_{i,k} : k \in [d]_i\}$, $t_i = 1 \wedge \tan(\theta_{i,\min}/2)$ for $i_1 = i$ in (4.3), so that $\alpha'_{i,j}$ in (4.4) and $\alpha_{i,j}$ in (2.11) are within a factor of 2 of each other.

Theorem 5. *For any positive integer $m < d$, there exists a finite numerical constant C_m depending on m only such that for any set of positive constants $\{b_{i_{1:m}} : i_{1:m} \in [d]^m_<\}$ with ordered values $b_{(1)} \le b_{(2)} \le \cdots$, the mth-order derivatives in (4.1) are bounded by*

$$\sup_x \sum_{i_{1:m} \in [d]^m_<} \frac{\alpha'_{i_{1:m}}}{b_{i_{1:m}}} \left| G_{i_{1:m}}(x) \right| \le C_m \max_{1 \le k \le d} \frac{(1 + \sqrt{2 \log k})^m}{b_{(k)}}, \tag{4.5}$$

where $\alpha'_{i_{1:m}}$ are as in (4.4) for $i_{1:m} \in [d]^m_<$.

As mentioned in the discussion of (2.5), the upper bound in our anticoncentration inequality can be expressed in terms of the minimum eigenvalue of the correlation matrix of no more than m components of X. For $i_{1:m} = (i_1, \ldots, i_m) \in [d]^m_{\neq}$, let $\rho^{i_{1:m}}$ be the $m \times m$ correlation matrix of $X_{i_{1:m}} = (X_{i_1}, \ldots, X_{i_m})^\top$ and define the corresponding minimum eigenvalue as

$$\lambda^{i_{1:m}}_{\min} = \min\left\{ u^\top \rho^{i_{1:m}} u : u \in \mathbb{R}^m, \|u\|_2 = 1 \right\}. \tag{4.6}$$

The following theorem asserts that the quantity $\alpha'_{i_{1:m}}$ in Theorem 5 can be replaced by

$$\alpha''_{i_{1:m}} = (\sigma_{i_1} \cdots \sigma_{i_m}) \left(\lambda^{i_{1:m}}_{\min} \prod_{j=1}^{m-1} \min\left\{ \lambda^{i_1,\ldots,i_j,k}_{\min} : k \in [d]_{i_1,\ldots,i_j} \right\} \right)^{1/2}. \tag{4.7}$$

For $m = 2$, $\min\{\lambda^{i,k}_{\min} : k \in [d]_i\} = 1 - \max\{|\rho_{i,k}| : k \in [d]_i\}$ in (4.7) while the sharper one-sided $t_i = 1 \wedge \tan(\theta_{i,\min}/2)$ and $2\tan(\theta_{i,\min}/4)$ are respectively used in (4.4) and (2.11), where $\cos(\theta_{i,\min}) = \max\{\rho_{i,k} : k \in [d]_i\}$.

Theorem 6. *For any positive integer $m < d$, there exists a finite numerical constant C_m depending on m only such that (4.5) holds with the quantity $\alpha'_{i_{1:m}}$ replaced by the quantity $\alpha''_{i_{1:m}}$ in (4.7). In particular,*

$$\sup_x \sum_{i_{1:m}\in[d]^m_<} \left|G_{i_{1:m}}(x)\right| \le \frac{C_m(1+\sqrt{2\log d})^m}{\min\{\alpha''_{i_{1:m}} : i_{1:m} \in [d]^m_<\}}, \tag{4.8}$$

and in terms of the sparse eigenvalue $\lambda_{\min,j} = \min\{\lambda^{i_{1:j}}_{\min} : i_{1:j} \in [d]^j_{\neq}\}$ with the $\lambda^{i_{1:j}}_{\min}$ in (4.6)

$$\sup_x \sum_{i_{1:m}\in[d]^m_<} \left(\prod_{j=1}^m \sigma_{i_j}\right)\left|G_{i_{1:m}}(x)\right| \le \frac{C_m(1+\sqrt{2\log d})^m}{\lambda_{\min,m}\sqrt{\lambda_{\min,m-1}\cdots\lambda_{\min,2}}}. \tag{4.9}$$

While the quantity $\alpha''_{i_{1:m}}$ in (4.7) is expressed in terms of the more familiar minimum eigenvalues, it is bounded from the above by the quantity $\alpha'_{i_{1:m}}$ in (4.4) up to a constant factor. Moreover, compared with $\alpha''_{i_{1:m}}$, the quantity $\alpha'_{i_{1:m}}$ is potentially of larger order as it involves one-sided threshold levels $t_{i_{1:j}}$ in (4.3). Thus, Theorem 5 is slightly sharper than Theorem 6. We present next an upper bound of the ratio $\alpha''_{i_{1:m}}/\alpha'_{i_{1:m}}$ through a Cholesky decomposition of correlation matrices, and thus the validity of Theorem 6 as a corollary of Theorem 5.

Because the quantity $\alpha'_{i_{1:m}}$ involves partial correlations in (4.3), we construct the Cholesky decomposition through a Gram–Schmidt orthogonalization process. Let $Y_i = (X_i - \mu_i)/\sigma_i$. In the Gram–Schmidt orthogonalization process, we write

$$Y_{k|i_{1:j}} = \frac{Y_{k|i_{1:(j-1)}} - \rho_{i_j,k|i_{1:(j-1)}} Y_{i_j|i_{1:(j-1)}}}{(1-\rho^2_{i_j,k|i_{1:(j-1)}})^{1/2}}, \quad k \in [d]_{i_{1:j}},\ j = 0,\dots,m-1, \tag{4.10}$$

with the convention $Y_{k|i_{1:0}} = Y_k$. Let $A^{i_{1:m}}$ be the matrix satisfying

$$\begin{pmatrix} Y_{i_1} \\ Y_{i_2|i_1} \\ \vdots \\ Y_{i_m|i_{1:(m-1)}} \end{pmatrix} = A^{i_{1:m}} \begin{pmatrix} Y_{i_1} \\ Y_{i_2} \\ \vdots \\ Y_{i_m} \end{pmatrix}. \tag{4.11}$$

Because $\{Y_{k|i_{1:j}}, k \in [d]_{i_{1:j}}\}$ and $Y_{i_{1:j}}$ are independent, $Y_{i_1}, Y_{i_2|i_1}, \dots, Y_{i_m|i_{1:(m-1)}}$ are iid $N(0,1)$ variables, so that $A^{i_{1:m}}$ gives a Cholesky decomposition of $\rho^{i_{1:m}}$ in the sense of

$$I_{m\times m} = A^{i_{1:m}} \rho^{i_{1:m}} \left(A^{i_{1:m}}\right)^\top. \tag{4.12}$$

As the spectrum norm of $A^{i_{1:m}}$ is bounded by $(\lambda^{i_{1:m}}_{\min})^{-1/2}$ and the elements of $A^{i_{1:m}}$ are expressed in terms of partial correlations, (4.10), (4.11), and (4.12) lead to the following lemma.

Lemma 1. *For $i_{1:m} \in [d]^m_{\neq}$, let $\rho^{i_{1:m}}$ be the $m \times m$ correlation matrix of the Gaussian vector $X_{i_{1:m}} = (X_{i_1}, \dots, X_{i_m})^\top$. For $j \in [d]_{i_{1:m}}$, let $\rho_{(i_m,j)|i_{1:(m-1)}}$ be the partial correlation as defined in (4.2). Then, the determinant of $\rho^{i_{1:m}}$ is given by*

$$\det\left(\rho^{i_{1:m}}\right) = \prod_{k=2}^m \prod_{j=1}^{k-1} \left(1-\rho^2_{i_j,i_k|i_{1:(j-1)}}\right) \tag{4.13}$$

with $\rho_{i_k,i_1|i_{1:0}} = \rho_{i_k,i_1}$. *Consequently, with* $\lambda_{\min}^{i_{1:m}}$ *being the smallest eigenvalue of* $\rho^{i_{1:m}}$,

$$\det\bigl(\rho^{i_{1:m}}\bigr)\prod_{k=1}^{m-1}\min\biggl(1,\frac{1-\rho_{\ell_{k+1},i_k|i_{1:(k-1)}}}{1+\rho_{\ell_{k+1},i_k|i_{1:(k-1)}}}\biggr)\geq\lambda_{\min}^{i_{1:m}}\prod_{k=1}^{m-1}\bigl(\lambda_{\min}^{i_{1:k},\ell_{k+1}}/5\bigr) \tag{4.14}$$

for all $\ell_{k+1}\in[d]_{i_{1:k}}$, $k=1,\dots,m-1$.

It follows from Lemma 1 that $\alpha''_{i_{1:m}}\leq 5^{(m-1)/2}\alpha'_{i_{1:m}}$ for the quantities in (4.7) and (4.4), respectively, so that Theorem 6 is a consequence of Theorem 5.

We still need to consider the case where the differentiation is taken multiple times in some of the directions. As a general study of such results is beyond the scope of this paper, we present here an upper bound for the third derivative and discuss the main difficulties in the higher-order cases.

Theorem 7. *Let* $G_{i,j,k}(x)$ *be as in* (4.1) *for a Gaussian vector* $X_{1:d}$ *with marginal distributions* $X_i\sim N(\mu_i,\sigma_i^2)$. *Let* $\lambda_{\min,j}$ *be the lower sparse eigenvalue as in Theorem* 6 *for the correlation matrices of* j*-components of* $X_{1:d}$. *Then, for some numeric constant* C_3,

$$\sup_x\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d\sigma_i\sigma_j\sigma_k\bigl|G_{i,j,k}(x)\bigr|\leq\frac{C_3(1+\sqrt{2\log d})^3}{\lambda_{\min,3}\sqrt{\lambda_{\min,2}}}. \tag{4.15}$$

In our approach, the proof of Theorem 7 and the analysis in higher-order cases involve factors which can be expressed as regression coefficients. Let $Y_i=(X_i-\mu_i)/\sigma_i$ as in (4.10). Given $i_{1:m}\in[d]_{\neq}^m$ and $k\in[d]_{i_{1:m}}$, the linear regression of Y_k against $Y_{i_{1:m}}$ is given by

$$\mathbb{E}[Y_k|Y_{i_{1:m}}]=\sum_{j=1}^m\beta_{i_j}^{k|i_{1:m}}Y_{i_j}. \tag{4.16}$$

These regression coefficients $\beta_{i_j}^{k|i_{1:m}}$ appear in the derivatives (4.1) as follows. Let

$$P_{i_{1:m}}(x)=\mathbb{P}\bigl\{X_k\leq x_k\ \forall\,k\in[d]_{i_{1:m}}|X_{i_{1:m}}=x_{i_{1:m}}\bigr\} \tag{4.17}$$

and $\phi_{i_{1:m}}(x)$ be the joint density of $X_{i_{1:m}}$. As in (2.7), we have

$$G_{i_{1:m}}(x)=P_{i_{1:m}}(x)\phi_{i_{1:m}}(x). \tag{4.18}$$

As $Y_k-\mathbb{E}[Y_k|Y_{i_{1:m}}]$ is independent of $Y_{i_{1:m}}$ and $Y_{i_{1:m}}$ is linear in $X_{i_{1:m}}$, the conditional probability in (4.17) can be written as

$$P_{i_{1:m}}(x)=\mathbb{P}\biggl\{Y_k-\mathbb{E}[Y_k|Y_{i_{1:m}}]\leq\frac{x_k-\mu_k}{\sigma_k}-\sum_{j=1}^m\beta_{i_j}^{k|i_{1:m}}\frac{x_{i_j}-\mu_{i_j}}{\sigma_{i_j}},\ \forall k\in[d]_{i_{1:m}}\biggr\}.$$

Thus, for $a\in[m]$,

$$\begin{aligned}(\partial/\partial x_{i_a})G_{i_{1:m}}(x)&=P_{i_{1:m}}(x)(\partial/\partial x_{i_a})\phi_{i_{1:m}}(x)+\phi_{i_{1:m}}(x)(\partial/\partial x_{i_a})P_{i_{1:m}}(x)\\&=G_{i_{1:m}}(x)(\partial/\partial x_{i_a})\log\phi_{i_{1:m}}(x)\\&\quad-\sum_{i_{m+1}\in[d]_{i_{1:m}}}G_{i_{1:m+1}}\beta_{i_a}^{i_{m+1}|i_{1:m}}\sigma_{i_{m+1}}/\sigma_{i_a}.\end{aligned} \tag{4.19}$$

In general, the scaled mth partial derivative $(\sigma_{i_1}\partial/\partial x_{i_1})^{j_1}\cdots(\sigma_{i_k}\partial/\partial x_{i_k})^{j_k}G(x)$ with $j_1+\cdots+j_k=m$ would involve a term of the form

$$(-1)^{m-k}\sigma_{i_1}\cdots\sigma_{i_m}G_{i_1,\ldots,i_m}\beta_{\ell_{k+1}}^{i_{k+1}|i_{1:k}}\cdots\beta_{\ell_m}^{i_m|i_{1:(m-1)}}$$

such that i_a appears j_a-1 times in $\ell_{k+1},\ldots,\ell_m$, optionally in the order of $a=1,\ldots,k$. While $\beta_{\ell_{k+1}}^{i_{k+1}|i_{1:k}}\lesssim 1/\lambda_{\min}^{i_{1:k}}$, a difficulty is to find sharper bounds for

$$\sigma_{i_1}\cdots\sigma_{i_m}\beta_{\ell_{k+1}}^{i_{k+1}|i_{1:k}}\cdots\beta_{\ell_m}^{i_m|i_{1:(m-1)}}/\alpha'_{i_{1:m}}$$

to extend Theorem 7 to higher order in the same form as that of (4.15) and (4.9).

Proof of Theorem 5. Consider a fixed sequence of integers $i_{1:m}\in[d]_{\neq}^m$. Define

$$X'_{j|i_{1:m}}=\frac{X'_{j|i_{1:(m-1)}}-\rho_{(i_m,j)|i_{1:(m-1)}}X'_{i_m|i_{1:(m-1)}}}{(1-\rho^2_{(i_m,j)|i_{1:(m-1)}})^{1/2}} \tag{4.20}$$

as in (2.16) with the partial correlation $\rho_{(i_m,j)|i_{1:(m-1)}}$ in (4.2) and initialization $X'_{j|i_{1:0}}=X'_j=(X_j-x_j)/\sigma_j$. This is the same Gram–Schmidt orthogonalization process as in (4.10) but the X'_j are not centered to have mean zero at the initialization. Still the covariance structure of $X'_{j|i_{1:m}}$ is the same as that of $Y_{j|i_{1:m}}$. Because $X'_{k|i_{1:m}},k\in[d]_{i_{1:m}}$ are independent of $X_{i_{1:m}}$,

$$\begin{aligned}G_{i_{1:m}}(x)&=\int_{y_k\le x_k,\forall k\in[d]_{i_{1:m}}}\phi_{[d]}(y)\prod_{k\in[d]_{i_{1:m}}}dy_k|_{y_{i_{1:m}}=x_{i_{1:m}}}\\&=\mathbb{P}\{X'_k<0\ \forall k\in[d]_{i_{1:m}}|X'_{i_1}=\cdots=X'_{i_m}=0\}\phi_{i_{1:m}}(x_{i_{1:m}})\\&=\mathbb{P}\{X'_{k|i_{1:m}}<0\ \forall k\in[d]_{i_{1:m}}\}\phi_{i_{1:m}}(x_{i_{1:m}})\end{aligned} \tag{4.21}$$

as in (2.17) and (4.18). To bound the probability $\mathbb{P}\{X'_{k|i_{1:m}}<0\ \forall k\in[d]_{i_{1:m}}\}$, we define

$$\pi_{i_{1:m}}=\mathbb{P}\{\mathscr{C}_{i_{1:m}}\}G_{i_{1:m}}(x)/\phi_{i_{1:m}}(x_{i_{1:m}}), \tag{4.22}$$

where $\mathscr{C}_{i_{1:m}}$ is defined with the threshold levels $t_{i_{1:j}}$ in (4.3) as

$$\mathscr{C}_{i_{1:m}}=\{0<X'_{i_{j+1}|i_{1:j}}\le t_{i_{1:j}}X'_{i_j|i_{1:(j-1)}},1\le j<m,X_{i_1}>0\}.$$

Given integers $j\ge 0$ and $i_{1:j}$, define the rank of $X'_{k|i_{1:j}}$ as

$$R_{k|i_{1:j}}=\sum_{\ell\in[d]_{i_{1:j}}}I\{X'_{\ell|i_{1:j}}\ge X'_{k|i_{1:j}}\},\quad k\in[d]_{i_{1:j}}.$$

Here $R_{k|i_{1:0}}=\sum_{\ell\in[d]}I\{X'_\ell\ge X'_k\}$ is the marginal rank of X'_k as $X'_{\ell|i_{1:0}}=X'_\ell$ in (4.20). In the event $\{X'_{k|i_{1:m}}<0\ \forall k\in[d]_{i_{1:m}}\}\cap\mathscr{C}_{i_{1:m}}$, we have $R_{i_m|i_1,\ldots,i_{m-1}}=1$ due to

$$X'_{k|i_{1:(m-1)}}\le\rho_{i_m,k|i_{1:(m-1)}}X'_{i_m|i_{1:(m-1)}}\le X'_{i_m|i_{1:(m-1)}},$$

and by induction $R_{i_j|i_{1:(j-1)}}=1$ given $R_{i_{j+1}|i_{1:j}}=1$ for $j=m-1,\ldots,1$ due to

$$\begin{aligned}X'_{k|i_{1:(j-1)}}&=\rho_{i_j,k|i_{1:(j-1)}}X'_{i_j|i_{1:(j-1)}}+\{1-\rho^2_{i_j,k|i_{1:(j-1)}}\}^{1/2}X'_{k|i_{1:j}}\\&\le\rho_{i_j,k|i_{1:(j-1)}}X'_{i_j|i_{1:(j-1)}}+\{1-\rho^2_{i_j,k|i_{1:(j-1)}}\}^{1/2}X'_{i_{j+1}|i_{1:j}}\\&\le\{\rho_{i_j,k|i_{1:(j-1)}}+(1-\rho^2_{i_j,k|i_{1:(j-1)}})^{1/2}t_{i_{1:j}}\}X'_{i_j|i_{1:(j-1)}}\\&\le X'_{i_j|i_{1:(j-1)}},\end{aligned}$$

by the choice of $t_{i_{1:j}}$ in (4.3). Thus, due to the independence between the event $\mathscr{C}_{i_{1:m}}$ and the set of random variables $\{X'_{k|i_{1:m}}, k \in [d]_{i_{1:m}}\}$,

$$\pi_{i_{1:m}} = \mathbb{P}\{X'_{k|i_{1:m}} < 0 \ \forall k \in [d]_{i_{1:m}}, \mathscr{C}_{i_{1:m}}\} \le \mathbb{P}\{R_{i_j|i_{1:(j-1)}}, 1 \le j \le m\}.$$

Consequently,

$$\sum_{i_{1:m} \in [d]^m_{\neq}} \pi_{i_{1:m}} \le 1. \tag{4.23}$$

We still need to find a suitable lower bound for $\mathbb{P}\{\mathscr{C}_{i_{1:m}}\}$ to use (4.23). Let $\nu_i = \mathbb{E}[X'_i]$,

$$\nu_{i_{1:m}} = \{(\nu_{i_1}, \dots, \nu_{i_m})\big(\Sigma^{i_{1:m}}\big)^{-1}(\nu_{i_1}, \dots, \nu_{i_m})^\top\}^{1/2}, \quad \nu_{i_j|i_{1:(j-1)}} = \mathbb{E}\big[X'_{i_j|i_{1:(j-1)}}\big],$$

and $\varphi_{i_j|i_{1:(j-1)}}$ be the $N(\nu_{i_j|i_{1:(j-1)}}, 1)$ density. We shall prove that

$$\mathbb{P}\{\mathscr{C}_{i_{1:m}}\} \ge \alpha'_{i_{1:m}} \phi_{i_{1:m}}(x) C'_m J_m(\nu_{i_{1:m}})/\nu^m_{i_{1:m}}, \tag{4.24}$$

with $J_m(t) = \int_0^\infty y^{m-1} e^{-y-y^2/(2t^2)} dy$ and $C'_m = 2\pi^{m/2}/\{2^m \Gamma(m/2) m!\}$, and that

$$\big|\det\big(\Sigma^{i_{1:m}}\big)\big|^{1/2} \phi_{i_{1:m}}(x) = (2\pi)^{-m/2} \exp\big(-\nu^2_{i_{1:m}}/2\big). \tag{4.25}$$

Because $X'_{i_j|i_{1:(j-1)}}$ are defined by the Gram–Schmidt process, they are independent $N(\nu_{i_j|i_{1:(j-1)}}, 1)$ variables. Thus, as the Jacobian of a linear transformation of $X'_{i_1}, \dots, X'_{i_m}$ is a constant, $|\det(\Sigma^{i_{1:m}})|^{1/2} \phi_{i_{1:m}}(x) = \prod_{j=1}^m \varphi_{i_j|i_{1:(j-1)}}(0)$ and $\sum_{j=1}^m \nu^2_{i_j|i_{1:(j-1)}} = \nu^2_{i_{1:m}}$. This gives (4.25). Because $t_{i_{1:j}} \le 1$ for all j, it follows that

$$\begin{aligned}
\mathbb{P}\{\mathscr{C}_{i_{1:m}}\} &= \mathbb{P}\big\{0 < X'_{i_{j+1}|i_{1:j}} \le t_{i_{1:j}} X'_{i_j|i_{1:(j-1)}}, 1 \le j < m, X_{i_1} > 0\big\} \\
&= \int_0^\infty \int_0^{t_{i_1} x_1} \cdots \int_0^{t_{i_{1:(m-1)}} x_{m-1}} \prod_{j=1}^m \varphi(x_j - \nu_{i_j|i_1,\dots,i_{j-1}}) dx_j \\
&\ge \int_0^\infty \int_0^{t_{i_1} x_1} \cdots \int_0^{t_{i_{1:(m-1)}} x_{m-1}} \frac{\exp(-\nu^2_{i_{1:m}}/2 - \nu_{i_{1:m}} \|x\|_2 - \|x\|_2^2/2)}{(2\pi)^{m/2}} dx \\
&\ge \left(\prod_{j=1}^{m-1} t_{i_{1:j}}\right) \int_0^\infty \int_0^{x_1} \cdots \int_0^{x_{m-1}} \frac{\exp(-\nu^2_{i_{1:m}}/2 - \nu_{i_{1:m}} \|x\|_2 - \|x\|_2^2/2)}{(2\pi)^{m/2}} dx \\
&= \left(\prod_{j=1}^{m-1} t_{i_{1:j}}\right) \frac{2}{m! 2^m \Gamma(m/2) 2^{m/2}} \int_0^\infty y^{m-1} e^{-\nu^2_{i_{i:m}}/2 - \nu_{i_{i:m}} y - y^2/2} dy \\
&= \left(\prod_{j=1}^{m-1} t_{i_{1:j}}\right) \frac{2\pi^{m/2} |\det(\Sigma^{i_{1:m}})|^{1/2} \phi_{i_{1:m}}(x)}{2^m \Gamma(m/2) m!} \frac{J_m(\nu_{i_{1:m}})}{\nu^m_{i_{1:m}}} \\
&= \alpha'_{i_{1:m}} \phi_{i_{1:m}}(x) C'_m J_m(\nu_{i_{1:m}})/\nu^m_{i_{1:m}}.
\end{aligned}$$

Putting together (4.4), (4.21), (4.22), (4.23), (4.24), and (4.25), we find that

$$\begin{aligned}\sum_{i_{1:m}\in[d]^m_<}\frac{\alpha'_{i_{1:m}}}{b_{i_{1:m}}}G_{i_{1:m}}(x) &= \sum_{i_{1:m}\in[d]^m_<}\min\left\{\frac{e^{-v^2_{i_{1:m}}/2}}{(2\pi)^{m/2}b_{i_{1:m}}},\frac{\alpha'_{i_{1:m}}}{b_{i_{1:m}}}\frac{\pi_{i_{1:m}}\phi_{i_{1:m}}(x_{i_{1:m}})}{\mathbb{P}\{\mathscr{C}_{i_{1:m}}\}}\right\}\\ &\le \sum_{i_{1:m}\in[d]^m_<}\min\left\{\frac{e^{-v^2_{i_{1:m}}/2}}{(2\pi)^{m/2}b_{i_{1:m}}},\frac{\pi_{i_{1:m}}v^m_{i_{1:m}}}{b_{i_{1:m}}C'_mJ_m(v_{i_{1:m}})}\right\}\\ &\le \sum_{k\notin K}\frac{e^{-L_k^2/2}}{(2\pi)^{m/2}b_{(1)}}+\max_{k\in K}\frac{m!L_k^m}{b_{(k)}C'_mJ_m(L_k)}\\ &\le C_m\max_{k\in[d]}\frac{(1+\sqrt{2\log k})^m}{b_{(k)}},\end{aligned}\tag{4.26}$$

with $L_k = 1+\sqrt{2\log k}$ and $K=\{k : v_{(k)}\le L_k\}$ due to the monotonicity $J_m(t)/t^m\uparrow$ in $(0,\infty)$ and $J_m(L_k)\ge J_m(L_1)=J_m(1)$. This completes the proof of Theorem 5. ■

Proof of Theorem 6. In view of the definitions of $\alpha'_{i_{1:m}}$ and $\alpha''_{i_{1:m}}$ in (4.4) and (4.7), respectively, Theorem 6 follows directly from Theorem 5 and Lemma 1. ■

Proof of Lemma 1. Let $i_{1:m}=1:m$, as a permutation of labels does not change the conclusions. It follows from (4.12) that

$$\det\bigl(\rho^{1:m}\bigr)\bigl(\det\bigl(A^{1:m}\bigr)\bigr)^2=1.$$

Because $A^{1:m}$ is a lower-triangular matrix with diagonal elements $A^{1:m}_{1,1}=1$ and $A^{1:m}_{k,k}=\prod_{j=1}^{k-1}(1-\rho^2_{k,j|1:(j-1)})^{-1/2}$ for $2\le k\le m$,

$$\det\bigl(\rho^{1:m}\bigr)=\frac{1}{\det^2(A^{1:m})}=\prod_{k=2}^m\frac{1}{(A^{1:m}_{k,k})^2}=\prod_{k=2}^m\prod_{j=1}^{k-1}\bigl(1-\rho^2_{k,j|1:(j-1)}\bigr).$$

This gives (4.13). Now we write for $k<\ell_{k+1}\le d$

$$\begin{aligned}&\det\bigl(\rho^{1:m}\bigr)\prod_{k=1}^{m-1}\min\left(1,\frac{1-\rho_{\ell_{k+1},k|1:(k-1)}}{1+\rho_{\ell_{k+1},k|1:(k-1)}}\right)\\ &=\left(\prod_{j=1}^{m-1}\bigl(1-\rho^2_{m,j|1:(j-1)}\bigr)\right)\prod_{k=1}^{m-1}\left\{\min\left(1,\frac{1-\rho_{\ell_{k+1},k|1:(k-1)}}{1+\rho_{\ell_{k+1},k|1:(k-1)}}\right)\prod_{j=1}^{k-1}\bigl(1-\rho^2_{k,j|1:(j-1)}\bigr)\right\},\end{aligned}\tag{4.27}$$

with the convention $\prod_{j=1}^{k-1}(1-\rho^2_{k,j|1:(j-1)})=1$ for $k=1$. By (4.12),

$$\left(\prod_{j=1}^{m-1}\bigl(1-\rho^2_{m,j|1:(j-1)}\bigr)\right)^{-1}=\bigl(A^{1:m}_{m,m}\bigr)^2\le\frac{1}{\lambda_{\min}(\rho^{1:m})},$$

as the spectral norm of $A^{1:m}$ is no greater than $1/\lambda_{\min}^{1/2}(\rho^{1:m})$. For $1 \le k \le m-1$,

$$\begin{aligned}
&\max\left(1, \frac{\sqrt{1+\rho_{k+1,k|1:(k-1)}}}{\sqrt{1-\rho_{k+1,k|1:(k-1)}}}\right)\prod_{j=1}^{k-1}\bigl(1-\rho_{k,j|1:(j-1)}^2\bigr)^{-1/2}\\
&\quad\le \left(1+2\frac{|\rho_{k+1,k|1:(k-1)}|}{\sqrt{1-\rho_{k+1,k|1:(k-1)}^2}}\right)\prod_{j=1}^{k-1}\bigl(1-\rho_{k,j|1:(j-1)}^2\bigr)^{-1/2}\\
&\quad= A_{k.k}^{1:(k+1)} + 2\bigl|A_{k+1,k}^{1:(k+1)}\bigr|\\
&\quad\le \sqrt{5/\lambda_{\min}\bigl(\rho^{1:(k+1)}\bigr)},
\end{aligned}$$

due to $\sqrt{(1+t)/(1-t)} \le 1+2|t|/\sqrt{1-t^2}$ or, equivalently, $1+t \le \sqrt{1-t^2}+2|t|$ for $|t|<1$. This and (4.27) give (4.14) because labels do not matter. ∎

Proof of Theorem 7. Let $\phi_{i,j}(x)$ be the joint density of $(X_i, X_j)^\top$, $v_i(x) = (x_i-\mu_i)/\sigma_i$, and $v_{i,j}(x)$ be given by $-v_{i,j}^2(x)/2 = \log(2\pi\det^{1/2}(\Sigma^{i,j})\phi_{i,j}(x))$ as in (4.25). As in (4.19) and similar to (2.26), for $i \ne j$,

$$G_{i,j,j}(x) = G_{i,j}(x)(\partial/\partial x_j)\log\phi_{i,j}(x) - \sum_{k\in[d]_{i,j}} G_{i,j,k}(x)\beta_j^{k|i,j}\sigma_k/\sigma_j,$$

with $|(\partial/\partial x_j)\log\phi_{i,j}(x)| = |e_j^\top(\rho^{i,j})^{-1}(v_i(x), v_j(x))^\top|/\sigma_j \le (\lambda_{\min}^{i,j})^{-1/2}v_{i,j}(x)/\sigma_j$ and

$$\beta_j^{k|i,j} = \frac{(1-\rho_{k,j|i}^2)^{-1/2}\rho_{k,j|i}(1-\rho_{j,i}^2)^{-1/2}}{(1-\rho_{k,j|i}^2)^{-1/2}(1-\rho_{k,i}^2)^{-1/2}} = \frac{\rho_{k,j|i}(1-\rho_{k,i}^2)^{1/2}}{(1-\rho_{i,j}^2)^{1/2}}.$$

The formula for the regression coefficient is obtained by noticing that $\beta_j^{k|i,j} = -A_{k,j}^{i,j,k}/A_{k,k}^{i,j,k}$ in the Cholesky decomposition (4.11) with the matrix elements $A_{k,j}^{i,j,k}$ and $A_{k,k}^{i,j,k}$ determined by the Gram–Schmidt formula (4.10). By (4.13),

$$\det\bigl(\rho^{i,j,k}\bigr) = \bigl(1-\rho_{k,j|i}^2\bigr)\bigl(1-\rho_{k,i}^2\bigr)\bigl(1-\rho_{i,j}^2\bigr).$$

As in the proof of Lemma 1, we have, by (4.4) and (4.3),

$$\begin{aligned}
\left(\frac{\sigma_i\sigma_j\sigma_k\beta_j^{k|i,j}}{\alpha'_{i,j,k}}\right)^2 &= \frac{\rho_{k,j|i}^2(1-\rho_{k,i}^2)(1-\rho_{i,j}^2)^{-1}}{t_{i,j}^2t_i^2(1-\rho_{k,j|i}^2)(1-\rho_{k,i}^2)(1-\rho_{i,j}^2)}\\
&\le \frac{1}{t_i^2}\times\frac{1}{t_{i,j}^2(1-\rho_{i,j}^2)}\times\frac{1}{(1-\rho_{j,k|i}^2)(1-\rho_{i,j}^2)}\\
&\le \frac{5}{\lambda_{\min}^{i,\ell_j}}\times\frac{5}{\lambda_{\min}^{i,j,\ell_k}}\times\frac{1}{\lambda_{\min}^{i,j,k}},
\end{aligned}$$

for some $\ell_j \in [d]_i$ and $\ell_k \in [d]_{i,j}$. It follows that

$$\begin{aligned}
\sum_{(i,j)\in[d]_{\ne}^2}\sigma_i\sigma_j^2\bigl|G_{i,j,j}(x)\bigr| \le{}& \sum_{(i,j)\in[d]_{\ne}^2}\sigma_i\sigma_j G_{i,j}(x)\bigl(\lambda_{\min}^{i,j}\bigr)^{-1/2}v_{i,j}(x)\\
&+\sum_{(i,j,k)\in\in[d]_{\ne}^3}\frac{5\alpha'_{i,k,j}|G_{i,k,j}(x)|}{\lambda_{\min,3}\sqrt{\lambda_{\min,2}}}.
\end{aligned}\tag{4.28}$$

Similar to (4.26) in the proof of Theorem 5, the first term on the right-hand side above is bounded by

$$\sum_{(i,j)\in[d]^2_{\neq}} \frac{\sigma_i\sigma_j G_{i,j}(x)\nu_{i,j}(x)}{\sqrt{\lambda_{\min,2}}} \le \frac{C_3'(1+\sqrt{2\log d})^3}{\lambda_{\min,2}}.$$

By Theorem 5, $\sum_{(i,j,k)\in\in[d]^3_{\neq}} \alpha'_{i,j,k}|G_{i,j,k}(x)| \le 6C_3(1+\sqrt{2\log d})^3$. Thus, the right-hand side of (4.28) is bounded by $C_3''(1+\sqrt{2\log d})^3/(\lambda_{\min,3}\sqrt{\lambda_{\min,2}})$. Similarly,

$$\sum_{i=1}^{2} \sigma_i^3 \big|G_{i,i,i}(x)\big| \le \frac{C_3''(1+\sqrt{2\log d})^3}{\lambda_{\min,3}\sqrt{\lambda_{\min,2}}}$$

by differentiating the identity

$$G_{i,i}(x) = G_i(x)\nu_i(x)/\sigma_i - \sum_{j:i\neq j\in[d]} G_{i,j}(x)\rho_{i,j}\sigma_j/\sigma_i$$

in (2.26). The conclusion follows as the sum over $[d]^3_{\neq}$ is bounded in Theorem 6. ∎

FUNDING

This work was partially supported by NSF grants IIS-1741390, CCF-1934924, and DMS-2052949.

REFERENCES

[1] R. J. Adler, An introduction to continuity, extrema, and related topics for general gaussian processes. *IMS* (1990).

[2] N. H. Anderson, P. Hall, and D. Titterington, Edgeworth expansions in very-high-dimensional problems. *J. Statist. Plann. Inference* **70** (1998), no. 1, 1–18.

[3] V. Bentkus, Smooth approximations of the norm and differentiable functions with bounded support in Banach space l_k^∞. *Lith. Math. J.* **30** (1990), no. 3, 223–230.

[4] V. Bentkus, On the dependence of the Berry–Esseen bound on dimension. *J. Statist. Plann. Inference* **113** (2003), no. 2, 385–402.

[5] S. Chatterjee, An error bound in the Sudakov–Fernique inequality. 2005, arXiv:math/0510424.

[6] V. Chernozhukov, D. Chetverikov, and K. Kato, Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** (2013), no. 6, 2786–2819.

[7] V. Chernozhukov, D. Chetverikov, and K. Kato, Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probab. Theory Related Fields* **162** (2015), no. 1, 47–70.

[8] V. Chernozhukov, D. Chetverikov, and K. Kato, Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** (2017), no. 4, 2309–2352.

[9] V. Chernozhukov, D. Chetverikov, K. Kato, and Y. Koike, Improved central limit theorem and bootstrap approximations in high dimensions. 2019, arXiv:1912.10529.

[10] V. Chernozhukov, D. Chetverikov, and Y. Koike, Nearly optimal central limit theorem and bootstrap approximations in high dimensions. 2020, arXiv:2012.09513.

[11] D. Das and S. Lahiri, Central limit theorem in high dimensions: The optimal bound on dimension growth rate. *Trans. Amer. Math. Soc.* **374** (2021), no. 10, 6991–7009.

[12] H. Deng, Slightly conservative bootstrap for maxima of sums. 2020, arXiv:2007.15877.

[13] H. Deng and C.-H. Zhang, Beyond gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *Ann. Statist.* **48** (2020), no. 6, 3643–3671.

[14] X. Fang and Y. Koike, High-dimensional central limit theorems by Stein's method. *Ann. Appl. Probab.* **31** (2021), no. 4, 1660–1686.

[15] X. Fernique, Des resultats nouveaux sur les processus gaussiens. *C. R. Acad. Sci., Sér. A–B* **278** (1974), A363–A365.

[16] Y. Gordon, Some inequalities for Gaussian processes and applications. *Israel J. Math.* **50** (1985), no. 4, 265–289.

[17] F. Gotze, On the rate of convergence in the multivariate CLT. *Ann. Probab.* (1991), 724–739.

[18] F. Götze, A. Naumov, V. Spokoiny, and V. Ulyanov, Gaussian comparison and anti-concentration inequalities for norms of Gaussian random elements. 2017, arXiv:1708.08663.

[19] Y. Koike, Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *Jpn. J. Stat. Data Sci.* (2020), 1–41.

[20] A. K. Kuchibhotla and A. Rinaldo, High-dimensional CLT for sums of non-degenerate random vectors: $n^{-1/2}$-rate. 2020, arXiv:2009.13673.

[21] M. Ledoux and M. Talagrand, *Probability in Banach spaces: isoperimetry and processes* 23. Springer, 1991.

[22] W. V. Li and Q.-M. Shao, Gaussian processes: inequalities, small ball probabilities and applications. *Handb. Statist.* **19** (2001), 533–597.

[23] W. V. Li and Q.-M. Shao, A normal comparison inequality and its applications. *Probab. Theory Related Fields* **122** (2002), no. 4, 494–508.

[24] J. W. Lindeberg, Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* **15** (1922), no. 1, 211–225.

[25] M. E. Lopes, Central limit theorem and bootstrap approximation in high dimensions with near $1/\sqrt{n}$ rates. 2020, arXiv:2009.06004.

[26] F. Nazarov, On the maximal perimeter of a convex set in r^n with respect to a Gaussian measure. In *Geometric aspects of functional analysis*, pp. 169–187, Springer, 2003.

[27] I. Nourdin and F. Viens, Density formula and concentration inequalities with Malliavin calculus. *Electron. J. Probab.* **14** (2009), 2287–2309.

[28] R. O'Donnell, R. A. Servedio, and L.-Y. Tan, Fooling polytopes. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, pp. 614–625, ACM, 2019.

[29] M. Rudelson and R. Vershynin, The Littlewood–Offord problem and invertibility of random matrices. *Adv. Math.* **218** (2008), no. 2, 600–633.

[30] M. Rudelson and R. Vershynin, Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.* **62** (2009), no. 12, 1707–1739.

[31] D. Slepian, The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech. J.* **41** (1962), no. 2, 463–501.

[32] V. N. Sudakov, Gaussian random processes and measures of solid angles in Hilbert space. *Dokl. Akad. Nauk* **197** (1971), 43–45.

[33] V. N. Sudakov, Geometric problems of the theory of infinite-dimensional probability distributions. *Tr. Mat. Inst. Steklova* **141** (1976), 3–191.

[34] M. Talagrand, *Spin glasses: a challenge for mathematicians: cavity and mean field models* 46, Springer, 2003.

[35] D. Zhang and W. B. Wu, Gaussian approximation for high dimensional time series. *Ann. Statist.* **45** (2017), no. 5, 1895–1919.

[36] X. Zhang and G. Cheng, Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli* **24** (2018), 2640–2675.

[37] M. Zhilova, Nonclassical Berry–Esseen inequalities and accuracy of the bootstrap. *Ann. Statist.* **48** (2020), no. 4, 1922–1939.

CUN-HUI ZHANG

Department of Statistics, Hill Center, Busch Campus, Rutgers University, Piscataway, NJ 08854, USA, czhang@stat.rutgers.edu

18. STOCHASTIC AND DIFFERENTIAL MODELLING

LOWER BOUNDS ON THE LYAPUNOV EXPONENTS OF STOCHASTIC DIFFERENTIAL EQUATIONS

JACOB BEDROSSIAN, ALEX BLUMENTHAL, AND SAM PUNSHON-SMITH

ABSTRACT

In this article, we review our recently introduced methods for obtaining strictly positive lower bounds on the top Lyapunov exponent of high-dimensional, stochastic differential equations such as the weakly-damped Lorenz-96 (L96) model or Galerkin truncations of the 2D Navier–Stokes equations. This hallmark of chaos has long been observed in these models, however, no mathematical proof had been provided for either deterministic or stochastic forcing.

The method we proposed combines (A) a new identity connecting the Lyapunov exponents to a Fisher information of the stationary measure of the Markov process tracking tangent directions (the so-called "projective process"); and (B) an L^1-based hypoelliptic regularity estimate to show that this (degenerate) Fisher information is an upper bound on some fractional regularity. For L96 and GNSE, we then further reduce the lower bound of the top Lyapunov exponent to proving that the projective process satisfies Hörmander's condition. We review the recent contributions of the first and third authors on the verification of this condition for the 2D Galerkin–Navier–Stokes equations in a rectangular, periodic box of any aspect ratio. Finally, we briefly contrast this work with our earlier work on Lagrangian chaos in the stochastic Navier–Stokes equations. We end the review with a discussion of some open problems.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 37H15; Secondary 37D25, 58J65, 35B65

KEYWORDS

Lyapunov exponents, fluid dynamics, random dynamical systems

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5618–5654
DOI 10.4171/ICM2022/44

1. LYAPUNOV EXPONENTS FOR STOCHASTIC DIFFERENTIAL EQUATIONS

Understanding the "generic" long-term dynamics of high (or infinite-) dimensional nonlinear systems far from equilibrium remains a daunting task. In physical applications of interest, many such systems are both subject to unpredictable external forcing and observed to be chaotic in the sense of being very sensitive to the initial condition and forcing. Hence, for all practical purposes, the exact dynamics of any specific trajectory cannot be predicted far in advance and any controlled experiments will not be exactly repeatable. Instead of reckoning such systems one trajectory at a time, a common practice is to view initial conditions as *random*, i.e., distributed according to some probabilistic law, and to attempt to understand how this law evolves as it is transported by the dynamics. In this context, the relevant "time-invariant" objects are *equilibrium probabilistic laws* on the phase space of the system, often referred to as *invariant measures* or *stationary measures*.

There is a well-developed abstract theory (smooth ergodic theory) for understanding the invariant measures of chaotic systems, their geometric properties, and how these relate to the asymptotic regimes of trajectories initiated from "typical" initial conditions. On the other hand, it is quite hard to verify mathematically that this abstract program applies to systems of practical interest. There are already extremely challenging open problems for vastly simplified 2D toy models of the kinds of chaotic behavior seen in fluid dynamics, e.g., the Chirikov standard map discussed below in Section 1.1.

It turns out that verifying and understanding chaotic properties is far more tractable for systems subjected to *random* noise. The kinds of systems we have in mind are, for example, hydrodynamical settings such as with wind over a sail, a weather or climate system, or nonlinear wave systems. In these settings it has long been suggested to study the random dynamical system generated by the PDE or ODE subjected to random external forcing, and this is often done in applied mathematics (see, e.g., [26,69] and the references therein). Even with the simplifications coming from the random forcing, and despite considerable efforts, a thorough, mathematically rigorous understanding of these random systems is still in its infancy, with many basic open questions remaining.

In this article we will review existing work and our recent contributions [17,20] in proving that a given system of interest modeled by a stochastic differential equation is chaotic, i.e., it is highly sensitive to initial conditions for trajectories initiated at Lebesgue-typical points in the phase space. The specific systems we apply our methods to are the Lorenz-96 system [67] and Galerkin truncations of the 2D Navier–Stokes equations in a rectangular, periodic box (of any aspect ratio), provided they are subjected to sufficiently strong stochastic forcing[1] (equivalently, sufficiently weak damping) and are sufficiently high dimensional. These are the first results of this type for such models, despite overwhelming numerical evidence (see, e.g., [26,53,69,74]). Specifically we prove for these models that if the damping parameter is ε, then the top Lyapunov exponent (see Sections 1.1 and 1.2 for definition)

1 The deterministic case remains very far out of reach.

satisfies

$$\lim_{\varepsilon \to 0} \frac{\lambda_1^\varepsilon}{\varepsilon} = \infty$$

as $\varepsilon \to 0$, and in particular, $\exists \varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$, $\lambda_1^\varepsilon > 0$.

Outline

In Section 1 we give a background on Lyapunov exponents for stochastic differential equations (SDEs). Section 2 concerns formulae of Lyapunov exponents through the stationary statistics of tangent directions and contains both classical and our recent results from [17] which connect Lyapunov exponents to a certain Fisher information-type quantity. We discuss in Section 3 how to connect the Fisher information to regularity using ideas from hypoellipticity theory (also original work from [17]), and in Section 4 we discuss applications to a class of weakly-driven, weakly-dissipated SDE with bilinear nonlinear drift term (original work in [17] for Lorenz-96 and for Galerkin Navier–Stokes in [20]). In Section 5 we briefly discuss our earlier related work on Lagrangian chaos in the (infinite-dimensional) stochastic Navier–Stokes equations [14]. Finally, in Section 6 we discuss some open problems and potential directions for research.

1.1. Lyapunov exponents and their challenges

Let $\Phi^t : \mathbb{R}^n \to \mathbb{R}^n$, $t \in \mathbb{R}_{\geq 0}$ be a flow (autonomous or not) with differentiable dependence on initial conditions. The *Lyapunov exponent* at $x \in \mathbb{R}^n$, when it exists, is the limit

$$\lambda(x) = \lim_{t \to \infty} \frac{1}{t} \log \left| D_x \Phi^t \right|,$$

where $D_x \Phi^t$ is the Jacobian of Φ^t at x, i.e., the derivative with respect to the initial condition. Hence, $\lambda(x)$ gives the asymptotic exponential growth rate of the Jacobian as $t \to \infty$.

The exponent $\lambda(x)$ contains information about the divergence of trajectories: heuristically at least, if $d(x, y)$ is small then

$$d\left(\Phi^t(x), \Phi^t(y)\right) \approx e^{\lambda(x)t} d(x, y)$$

and hence $\lambda(x) > 0$ implies *exponential sensitivity with respect to initial conditions*, commonly popularized as the "butterfly effect." Morally, a positive Lyapunov exponent at a "large" proportion of initial conditions $x \in \mathbb{R}^n$ is a hallmark of chaos, the tendency of a dynamical system to exhibit disordered, unpredictable behavior. In this note we refer to a system such that $\lambda(x) > 0$ for Lebesgue a.e. x as *chaotic*.[2]

The existence of Lyapunov exponents is usually justified using tools from ergodic theory, and forms a starting point for obtaining more refined dynamical features, such as stable/unstable manifolds in the moving frame along "typical" trajectories. These ideas form

2 We caution the reader that there is no single mathematical definition of "chaos." Some definitions refer to the *existence* of a subset of the phase space exhibiting chaotic behavior, e.g., Li–Yorke chaos or the presence of a hyperbolic horseshoe. The results discussed in this note pertain to the long-time behavior of Lebesgue-typical initial conditions.

the fundamentals of *smooth ergodic theory*, which aims to study *statistical* properties of chaotic systems, such as decay of correlations, i.e., how $\Phi^t(x)$, $t \gg 1$ can "forget" the initial $x \in \mathbb{R}^n$, and probabilistic laws such as a strong law of large numbers or central limit theorem for $g \circ \Phi^t(x)$, where $g : \mathbb{R}^n \to \mathbb{R}$ is a suitable *observable* of the system; see, e.g., discussions in [7,10,66,83,85].

A discrete-time example

Unfortunately, estimating $\lambda(x)$ or proving $\lambda(x) > 0$ for specific systems turns out to be extremely challenging. A simple, classical model which shows the challenges is the Chirikov standard map family [29], written here as

$$F_L : \mathbb{T}^2 \circlearrowleft, \quad F_L(x, y) = \big(2x + L \sin(2\pi x) - y, x\big),$$

where $\mathbb{T}^2$ is parametrized as $[0, 1)^2$ and both coordinates in $F := F_L$ are taken modulo 1. Here, $L \geq 0$ is a fixed parameter which for purposes of the discussion here will be taken large. The diffeomorphism F is smooth and volume-preserving, and ergodic theory affirms that the Lyapunov exponent $\lambda(x, y) = \lim_n \frac{1}{n} \log |D_{(x,y)} F|$ exists for Lebesgue a.e. x and satisfies $\lambda(x, y) \geq 0$ where it exists. The Chirikov standard map itself is frequently used as a toy model of more complicated chaotic systems, e.g., the Navier–Stokes equations in transition from laminar flow to turbulence [68].

Observe that when $L \gg 1$ and away from an $O(L^{-1})$ neighborhood of $\{\cos(2\pi x) = 0\}$, the Jacobian $D_{(x,y)} F$ exhibits strong expansion along tangent directions roughly parallel to the x-axis (matched by strong contraction roughly parallel to the y-axis). In view of this, it is widely conjectured that $\{\lambda(x) > 0\}$ has positive Lebesgue measure. Nevertheless, this *standard map conjecture* remains wide open [32,75]. A key obstruction is "cone twisting": on long timescales, vectors roughly parallel to the x-axis are strongly expanded until the first visit to the "critical strip" near $\{\cos(2\pi x) = 0\}$, where DF is approximately a rotation by 90 degrees. At this point, vectors roughly parallel to the x axis are rotated to be roughly parallel to the y axis, where strong contraction occurs and previously accumulated expansion can be negated. Indeed, an estimate on a Lyapunov exponent requires understanding the asymptotic cancelations in the Jacobian as $t \to \infty$. One manifestation of the subtlety is the wildly tangled coexistence of hyperbolic trajectories [45] and elliptic islands [35].

The problem of estimating Lyapunov exponents for the standard map is far more tractable in the presence of noise/stochastic driving. Let us consider the standard map subjected to small noise: let $\omega_1, \omega_2, \ldots$ be i.i.d. random variables uniformly distributed in $[-\varepsilon, \varepsilon]$ for some $\varepsilon > 0$, and consider the random compositions

$$F^n = F_{\omega_n} \circ \cdots \circ F_{\omega_1}, \quad F_{\omega_i}(x, y) = F(x + \omega_i, y).$$

One can show show that $\forall \varepsilon > 0$, the corresponding Lyapunov exponent $\lambda = \lambda(x, y)$ is *deterministic* (independent of the random samples almost surely) and constant (independent of (x, y)) with probability 1. It is a folklore theorem that $\lambda > 0$ $\forall \varepsilon > 0$, while for $L \gg 1$ and $\varepsilon \gtrsim e^{-L}$, one can show $\lambda \geq \frac{1}{2} \log L$, commensurate with exponential expansion in the x-direction over the bulk of phase space [23]; in a related vein, see also [22,24,25,64,80].

1.2. Lyapunov exponents for SDE

The topic of this note is to discuss developments in the context of the random dynamical systems generated by stochastic differential equations (SDE), i.e., ODE subjected to Brownian motion driving terms. In this continuous-time framework, numerous additional tools not present in the discrete-time setting become available, e.g., infinitesimal generators, which as we show below, connect the estimation of Lyapunov exponents to regularity estimates (e.g., Sobolev regularity) of solutions to certain (degenerate) elliptic PDE. A highlight of this approach is our application to the Lyapunov exponents of a class of weakly-driven, weakly-forced SDEs, including famous models such as Lorenz 96 and Galerkin truncations of the Navier–Stokes equations.

For simplicity, in this note we restrict our attention to SDE on $\mathbb{R}^n$, however, our more general results apply to SDEs posed on orientable, geodesically complete, smooth manifolds; see [17]. Let $X_0, X_1, \dots, X_r : \mathbb{R}^n \to \mathbb{R}^n$ be smooth vector fields on $\mathbb{R}^n$, and let $W_t^1, \dots, W_t^r$ be a collection of independent, real-valued Brownian motions, with Ω denoting the corresponding canonical space with probability $\mathbf{P}$ and $(\mathcal{F}_t)_{t\geq 0}$ denoting the increasing filtration generated by $\{W_s^k, s \leq t\}_{k=1}^r$. We consider continuous-time processes (x_t) on $\mathbb{R}^n$ solving the SDE

$$\mathrm{d}x_t = X_0(x_t)\,\mathrm{d}t + \sum_{k=1}^{r} X_k(x_t) \circ \mathrm{d}W_t^k, \tag{1.1}$$

for fixed initial data $x_0 \in \mathbb{R}^n$.

Under mild conditions on the vector fields $X_0, \dots, X_r$ (for example, regularity and the existence of a suitable Lyapunov function to rule out finite time blow-up), global-in-time solutions (x_t) to (1.1) exist, are unique, and have differentiable dependence of x_t on x_0; in particular, for $\mathbf{P}$-a.e. $\omega \in \Omega$ and all $t \geq 0$, there exists a stochastic flow of diffeomorphisms Φ_ω^t such that $\forall x_0 \in \mathbb{R}^n$, the law of the process $(x_t)_{t\geq 0}$ solving (1.1) is the same as that of the process $(\Phi_\omega^t(x_0))_{t\geq 0}$; see, e.g., [60] for the details and general theory of SDEs and stochastic flows.

This *stochastic flow of diffeomorphisms* Φ_ω^t is the analogue of the flow Φ^t corresponding to solutions of the initial value problem of an ODE. However, the external stochastic forcing implies a time-inhomogeneity which must be accounted for. One can show that there exists a $\mathbf{P}$-measure preserving semiflow $\theta^t : \Omega \circlearrowleft, t \geq 0$ corresponding to time-shifts on the Brownian paths, i.e., shifting the path $(W_s)_{s\geq 0}$ to $(W_{t+s} - W_t)_{s\geq 0}$. Equipped with this time shift, one has the following with probability 1 and for all $s, t \geq 0$:

$$\Phi_\omega^{s+t} = \Phi_{\theta^s\omega}^t \circ \Phi_\omega^s. \tag{1.2}$$

We now set about summarizing the ergodic theory tools used to study such stochastic flows. First, we note that the trajectories $x_t = \Phi_\omega^t(x_0)$ for fixed initial $x_0 \in \mathbb{R}^n$ form a Markov process adapted to the filtration $(\mathcal{F}_t)$. Moreover, Φ_ω^t has *independent increments*: $\forall s, t \geq 0$, Φ_ω^s and $\Phi_{\theta^s\omega}^t$ are independent.

1.2.1. Stationary measures and long-term statistics

Markov semigroups. We write $P_t(x, A) = \mathbf{P}(\Phi_\omega^t(x) \in A)$ for the time-t transition kernel of (x_t). Let $\mathcal{P}_t$ denote the *Markov semigroup* associated to (x_t), defined for bounded, measurable *observables* $h : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{P}_t h(x) = \mathbf{E}\big[h(x_t) \mid x_0 = x\big] = \int_{\mathbb{R}^n} h(y) P_t(x, \mathrm{d}y).$$

This semigroup gives the expected value of a given observable given a fixed initial condition. Via the pairing of functions and measures, we derive the (formal) dual $\mathcal{P}_t^*$, which gives the evolution of the law of the solution (x_t) given a distribution for the initial condition: for a probability measure $\mu_0 \in \mathcal{P}(\mathbb{R}^n)$ and Borel set $A \subseteq \mathbb{R}^n$,

$$\mathcal{P}_t^* \mu_0(A) = \int_{\mathbb{R}^n} P_t(x, A) \mathrm{d}\mu_0(x).$$

That is, $\mathcal{P}_t^* \mu_0$ is the law of x_t assuming μ_0 is the law of x_0.

Taking a time derivative ∂_t, we (formally) obtain the *backward Kolmogorov equation*

$$\partial_t \mathcal{P}_t h(x) = \mathcal{L} \mathcal{P}_t h(x), \quad \text{where} \quad \mathcal{L} = X_0 + \frac{1}{2} \sum_{i=1}^{r} X_i^2, \tag{1.3}$$

where, for a given vector field X and $f \in C^\infty$, Xf denotes the derivative of f in the direction X. The differential operator $\mathcal{L}$ is called the *(infinitesimal) generator*. Assuming that the law of x_t has a density p_t with respect to Lebesgue, the formal dual of (1.3) is the *Fokker–Plank equation* (or *Forward Kolmogorov equation*) given by the following PDE

$$\partial_t p_t = \mathcal{L}^* p_t, \tag{1.4}$$

where $\mathcal{L}^*$ denotes the formal L^2 adjoint of $\mathcal{L}$. See, e.g., [60] for mathematical details.

Stationary measures. We say a measure μ is *stationary* if $\mathcal{P}_t^* \mu = \mu$. That is, if x_0 is distributed with law μ, then x_t is distributed with law[3] μ for all $t > 0$. We say that a set $A \subset \mathbb{R}^n$ is *invariant* if $P_t(x, A) = 1$ for all $x \in A$ and $t \geq 0$, and we say that a stationary measure μ is *ergodic* if all invariant sets have μ-measure 0 or 1. By the pointwise ergodic theorem, ergodic stationary measures determine the long-term statistics of a.e. initial datum in their support [33]: if μ is an ergodic stationary measure, then for any bounded, measurable $\varphi : \mathbb{R}^n \to \mathbb{R}$ and $\mu \times \mathbf{P}$-a.e. (x, ω) we have that

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \varphi\big(\Phi_\omega^t(x)\big) \, \mathrm{d}t = \int_{\mathbb{R}^n} \varphi(x) \mathrm{d}\mu(x).$$

Unlike for deterministic systems, stationary measures are usually much easier to characterize for SDEs. In particular, it is often possible to show that there exists a unique stationary measure and that it has a smooth density with respect to Lebesgue. In such a case, Lebesgue-generic initial conditions all have the same long-term statistics, a property often observed in nature and experiments for the physical systems we are interested in.

---

3 It is important to note that (x_t) itself is *not* constant in t; consider, e.g., water flowing past a stone in a river.

Existence of stationary measures. If the domain of the Markov process were compact (e.g., $\mathbb{T}^n$ instead of $\mathbb{R}^n$) then the existence of stationary measures would follow from a standard Krylov–Bogoliubov argument: given an initial probability measure $\mu_0 \in \mathcal{P}(\mathbb{R}^n)$, one considers the time-averaged measures

$$\bar{\mu}_t := \frac{1}{t}\int_0^t \mathcal{P}_s^* \mu_0 \, \mathrm{d}s.$$

The weak-$*$ compactness of probability measures on a compact space ensures that the sequence $\{\bar{\mu}_t\}_{t\geq 0}$ has a weak-$*$ limit point μ which by construction must be stationary (assuming some mild well-posedness properties for the original SDE). On a noncompact domain, one must show the tightness of the measures $\{\bar{\mu}_t\}_{t\geq 0}$ (this is essentially saying that solutions do not wander off to infinity too often) and use Prokorov's theorem to pass to the limit in the narrow topology. This is often achieved using the method of Lyapunov functions[4]/drift conditions [71], or by using a special structure and the damping in the system (such as the case for, e.g., the Navier–Stokes equations [59]).

Uniqueness of stationary measures. The Doob–Khasminskii theorem [33] implies that uniqueness is connected to (A) irreducibility and (B) regularization of the Markov semigroups[5] and, in particular, one can deduce that any stationary measure is unique if these properties hold in a sufficiently strong sense.

Let us first discuss irreducibility. For a Markov process (x_t) on $\mathbb{R}^n$, we say that (x_t) is *topologically irreducible* if for all open $U \subset \mathbb{R}^n$, $\exists t = t(U, x) \geq 0$ such that

$$P_t(x, U) > 0.$$

That is, every initial condition has a positive probability of being in U. This is stronger than necessary to deduce uniqueness, but is sufficient for our discussions.

Regularity is a little more subtle. A sufficient condition is the requirement of being *strong Feller*:

$$\forall \varphi : \mathbb{R}^n \to \mathbb{R} \quad \text{bounded, measurable,} \quad \mathcal{P}_t\varphi \in C\left(\mathbb{R}^n; \mathbb{R}\right), \quad t > 0.$$

For finite-dimensional SDEs, it is reasonably common and there exists a machinery to characterize this.[6] When $\mathrm{Span}\{X_i(x), 1 \leq i \leq r\} = \mathbb{R}^n$ for all $x \in \mathbb{R}^n$, $\mathcal{L}$ is elliptic and hence being strong Feller follows from classical parabolic regularity theory [65] applied to (1.3) (assuming suitable regularity conditions on the $\{X_j\}$). When this direct spanning is absent (e.g., when $r < n$), $\mathcal{L}$ is only degenerate elliptic. However, nearly sharp sufficient conditions for the regularization of $\mathcal{L}$ were derived by Hörmander [50], who obtained a condition (now called *Hörmander's condition*), in terms of the Lie algebra generated by the vector fields $\{X_i, 0 \leq i \leq r\}$. We will return to this important topic of *hypoellipticity* in Section 3.1.

---

4 These are the probabilistic analogues of Lyapunov's "first method" for ODEs, used to ensure convergence to compact attractors. This is not to be confused with Lyapunov exponents, which refer to Lyapunov's "second method."

5 In essence, this is equivalent to how $x \mapsto P_t(x, \cdot)$ behaves, i.e., whether trajectories with nearby initial conditions have similar statistics.

6 In infinite dimensions it is much more rare; luckily, it is stronger than what is required just to prove uniqueness (see, e.g., [47, 58]).

1.2.2. Lyapunov exponents

We saw that the long-term behavior of scalar observables is determined by stationary measures, which is due to the ergodic theorem. A more sophisticated ergodic theorem connects stationary measures to Lyapunov exponents. Given $x \in \mathbb{R}^n$, $v \in \mathbb{R}^n \setminus \{0\}$ (with v being considered a direction here) and a random sample $\omega \in \Omega$, the Lyapunov exponent at (ω, x, v) is defined as the limit (if it exists)

$$\lambda(\omega, x, v) = \lim_{t\to\infty} \frac{1}{t} \log\left|D_x \Phi_\omega^t v\right|.$$

The following (truncated) version of Oseledets' *Multiplicative Ergodic Theorem* (MET) addresses the existence of the limit [55,73,77].

Theorem 1.1 (Oseledets' multiplicative ergodic theorem [73]). *Let μ be an ergodic stationary measure, and assume a mild integrability condition (see, e.g., [55,73]) then, there exist (deterministic) constants $\lambda_1 > \lambda_2 > \cdots > \lambda_\ell \geq -\infty$ such that for $\mathbf{P} \times \mu$-almost all $(\omega, x) \in \Omega \times \mathbb{R}^n$ and for* all *$v \in \mathbb{R}^n \setminus \{0\}$, the limit defining $\lambda(\omega, x, v)$ exists and takes one of the values $\lambda_i, 1 \leq i \leq \ell$.*

Moreover, there exists a $\mathbf{P} \times \mu$-measurably-varying flag of strictly increasing subspaces

$$\emptyset =: F_{\ell+1}(\omega, x) \subset F_\ell(\omega, x) \subset \cdots \subset F_1(\omega, x) := \mathbb{R}^n$$

such that for $\mathbf{P} \times \mu$-a.e. (ω, x) and $\forall v \in F_j \setminus F_{j+1}$,

$$\lambda_j = \lim_{t\to\infty} \frac{1}{t} \log\left|D_x \Phi_\omega^t v\right| = \lambda(\omega, x, v).$$

In particular, the top Lyapunov exponent λ_1 is realized at $\mathbf{P} \times \mu$-a.e. (ω, x) and all $v \in \mathbb{R}^n$ outside a positive-codimension subspace $F_2(\omega, x) \subset \mathbb{R}^n$.

We note that under very mild conditions, if the stationary measure μ is unique, it is automatically ergodic; otherwise, each distinct ergodic stationary measure admits its own set of Lyapunov exponents.

The sign of the largest Lyapunov exponent λ_1 is the most relevant to the stability analysis of typical trajectories, in view of the fact that $\lambda(\omega, x, v) = \lambda_1$ for v in an open and dense set. For this reason we frequently refer to λ_1 as "the" Lyapunov exponent. The *sum Lyapunov exponent* also turns out to be crucial:

$$\lambda_\Sigma = \sum_{j=1}^{\ell} m_j \lambda_j = \lim_{t\to\infty} \frac{1}{t} \log\left|\det D_x \Phi_\omega^t\right|,$$

which gives the asymptotic exponential expansion/compression of Lebesgue volume under the flow. Here, $m_j = \dim F_j - \dim F_{j+1}$ is the *multiplicity* of the jth Lyapunov exponent.

2. FORMULAE FOR THE LYAPUNOV EXPONENTS

Throughout this section, we assume that Φ_ω^t is the stochastic flow of diffeomorphisms corresponding to the SDE (1.1) with associated Markov process $x_t = \Phi_\omega^t(x), x \in \mathbb{R}^n$.

2.1. The projective process

As we have seen, Lyapunov exponents are naturally viewed as depending on the tangent direction $v \in \mathbb{R}^n$ at which the derivative $D_x\Phi_\omega^t$ is evaluated. For this reason, to estimate Lyapunov exponents, it is natural to consider an auxiliary process on *tangent directions* themselves. To this end, let $\mathbb{SR}^n = \mathbb{R}^n \times \mathbb{S}^{n-1}$ denote the unit tangent bundle of $\mathbb{R}^n$, where $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. Given a fixed initial $(x, v) \in \mathbb{SR}^n$, we define the process (v_t) on $\mathbb{S}^{n-1}$ by

$$v_t = \frac{D_x\Phi_\omega^t(v)}{|D_x\Phi_\omega^t(v)|}.$$

The full process $z_t = (x_t, v_t)$ on $\mathbb{SR}^n$ is Markovian, and in fact solves an SDE

$$dz_t = \tilde{X}_0(z_t)\,\mathrm{d}t + \sum_{i=1}^{r} \tilde{X}_i(z_t) \circ \mathrm{d}W_t^{(i)},$$

where the "lifted" fields $\tilde{X}_i$ are defined as

$$\tilde{X}_i(x, v) := \big(X_i(x), (I - \Pi_v)\nabla X_i(x)v\big).$$

Here, we have written $\Pi_v = v \otimes v$ for the orthogonal projection onto the span of $v \in \mathbb{S}^{n-1}$. Below, we denote the corresponding generator by

$$\tilde{\mathscr{L}} := \tilde{X}_0 + \frac{1}{2}\sum_{i=1}^{r} \tilde{X}_i^2.$$

Lyapunov exponents and stationary measures. Let (x_t, v_t) be a trajectory of the projective process with fixed initial $(x, v) \in \mathbb{SR}^n$, and observe that at integer times $t \in \mathbb{Z}_{>0}$, we have by (1.2)

$$\frac{1}{t}\log\big|D_x\Phi_\omega^t(v)\big| = \frac{1}{t}\sum_{i=0}^{t-1} \log\big|D_{x_i}\Phi_{\theta^i\omega}^1 v_i\big|.$$

Hence, $\log|D_x\Phi_\omega^t|$ is an *additive observable* of (x_t, v_t), i.e., a sum iterated over the trajectory (x_t, v_t). Therefore, the strong law of large numbers for a Markov chain implies the following formula for the Lyapunov exponent:

Proposition 2.1 (See, e.g., [55]). *Let ν be an ergodic stationary measure for (x_t, v_t). Assuming the integral is finite, for ν-a.e. initial $(x, v) \in \mathbb{SR}^{n-1}$ and $t \geq 0$, we have*

$$t\lambda(\omega, x, v) = \mathbf{E}\int \log\big|D_x\Phi_\omega^t v\big|\,\mathrm{d}\nu(x, v).$$

with probability 1 *(*$\mathbf{E}$ *denotes integration with respect to* $d\mathbf{P}(\omega)$*).*

Moreover, if ν is the unique stationary measure for the (x_t, v_t) process, then for μ-a.e. x, and all *$v \in \mathbb{R}^n$, we have $\lambda_1 = \lambda(\omega, x, v)$ with probability* 1 *and*

$$t\lambda_1 = \mathbf{E}\int \log\big|D_x\Phi_\omega^t v\big|\,\mathrm{d}\nu(x, v). \tag{2.1}$$

Remark 2.2. The latter statement can be interpreted as saying that the existence of a unique stationary measure for the projective process gives a kind of nondegeneracy of the Oseledets' subspace $F_2(\omega, x)$ with respect to ω [55].

A time-infinitesimal version: the Furstenberg–Khasminskii formula. One of the key benefits of the SDE framework is the ability to take time derivatives, which turns dynamical questions (e.g., estimates of Lyapunov exponents, identification of stationary densities) into functional-analytic ones (e.g., solutions of degenerate elliptic or parabolic equations) for which many tools are available. Taking the time derivative of (2.1) gives what is known as the *Furstenberg–Khasminskii formula* (see, e.g., [7,54]):

Proposition 2.3. *Assume (x_t, v_t) admits a unique stationary measure ν on $\mathbb{SR}^n$ projecting to a stationary measure μ on $\mathbb{R}^n$ for (x_t). For $(x, v) \in \mathbb{SR}^n$, define*

$$Q(x) = \operatorname{div} X_0(x) + \frac{1}{2}\sum_{i=1}^{r} X_i \operatorname{div} X_i(x),$$

$$\tilde{Q}(x, v) = \operatorname{div} \tilde{X}_0(x, v) + \frac{1}{2}\sum_{i=1}^{r} \tilde{X}_i \operatorname{div} \tilde{X}_i(x, v).$$

Then, provided $Q \in L^1(\mathrm{d}\mu)$ and $\tilde{Q} \in L^1(\mathrm{d}\nu)$, one has

$$\lambda_\Sigma = \int Q \,\mathrm{d}\mu \quad \textit{and}$$

$$n\lambda_1 - \lambda_\Sigma = \int_{\mathbb{R}^n} Q \,\mathrm{d}\mu - \int_{\mathbb{SR}^n} \tilde{Q} \,\mathrm{d}\nu.$$

The first formula expresses $Q(x)$ as the time-infinitesimal rate at which $D_x\Phi_\omega^t$ compresses or expands Lebesgue measure, which in this formula is directly related to the asymptotic exponential volume growth or contraction rate λ_Σ. Similarly, $\tilde{Q}(x, v)$ is the time-infinitesimal rate at which $D_x\Phi_\omega^t$ compresses or expands volume on *the sphere bundle* $\mathbb{SR}^n = \mathbb{R}^n \times \mathbb{S}^{n-1}$. Roughly speaking, contraction of volumes along the $\mathbb{S}^{n-1}$ coordinate is associated with expansion in the Jacobian, while expansion of $\mathbb{S}^{n-1}$-volume is related to contraction in the Jacobian; this reversal is the reason for the minus sign in front of $\tilde{Q}$. For some intuition, observe that $(1, 0)$ is a sink and $(0, 1)$ is a source for the discrete-time system $v_n = A^n v/|A^n v|$ on S^1, where $A = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$.

2.2. Sign-definite formulas for Lyapunov exponents

The Furstenberg–Khasminskii formula is highly remarkable in that it reduces the problem of estimating Lyapunov exponents to computing the ensemble average of a single *deterministic* observable, $\tilde{Q}$, with respect to the stationary measure of (x_t, v_t). On the other hand, the formula itself is sign-indefinite, as $\tilde{Q}(x, v)$ takes on both positive and negative values as (x, v) is varied. This is reflective of the cancelation problem mentioned earlier in the estimation of Lyapunov exponents: previously accumulated tangent growth can be "canceled out" by rotation into contracting directions later on in the trajectory. Hence, without a very precise characterization of ν, it would be very challenging to obtain any useful quantitative estimates on λ_1 from this formula.

Given the above, it makes sense to seek a *sign-definite* formula for the Lyapunov exponent. Below, given measures $\lambda, \eta, \eta \ll \lambda$ on a measurable space X, the *relative entropy*

$H(\eta|\lambda)$ of η given λ is defined by

$$H(\eta \mid \lambda) = \int_X \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\lambda}\right)\mathrm{d}\eta.$$

Observe that $H(\eta \mid \lambda) \geq 0$, while by strict convexity of log and Jensen's inequality, we have $H(\eta|\lambda) = 0$ iff $\eta = \lambda$. We also write $\widehat{\Phi}^t_\omega : \mathbb{SR}^d \circlearrowleft$ for the stochastic flow associated to full lifted process (x_t, v_t) on $\mathbb{SR}^n$, that is, $\widehat{\Phi}^t_\omega(x_0, v_0) = (x_t, v_t)$. Lastly, given a diffeomorphism Φ of a Riemannian manifold M and a density g on M, we define $\Phi_* g$ to be the density

$$\Phi_* g(x) = g \circ \Phi^{-1}(x)\left|\det D_x \Phi^{-1}\right|,$$

noting that if x is distributed like $g\,\mathrm{dVol}_M$, then $\Phi(x)$ is distributed like $\Phi_* g\,\mathrm{dVol}_M$.

The following deep formula has its roots in Furstenberg's seminal paper [40] and ideas à la Furstenberg have been developed by a variety of authors (e.g., [12, 28, 63, 79, 82]), and can be stated as follows: if $\nu \in \mathcal{P}(\mathbb{SR}^n)$ is a stationary probability measure for the projective process (x_t, v_t) and $\mathrm{d}\nu(x, v) = \mathrm{d}\nu_x(v)\mathrm{d}\mu(x)$ is the disintegration of ν, then for all $t > 0$, the following identity (often an inequality in more general settings) holds.

Proposition 2.4 (See, e.g., [12]). *Assume (x_t, v_t) admits a unique stationary measure ν with density $f = \frac{d\nu}{dq}$, where $\mathrm{d}q = \mathrm{dVol}_{\mathbb{SR}^n}$ is the Riemannian volume measure on $\mathbb{SR}^n = \mathbb{R}^n \times \mathbb{S}^{n-1}$. Let μ be the corresponding stationary measure for (x_t) with density $\rho = \frac{\mathrm{d}\mu}{\mathrm{d}x}$. Writing*

$$f_t := \left(\widehat{\Phi}^t_\omega\right)_* f, \quad \rho_t := \left(\Phi^t_\omega\right)_* \rho,$$

we have (under the same integrability condition as Theorem 1.1)

$$\mathbf{E}H(\rho_t|\rho) = -t\lambda_\Sigma \quad \textit{and} \quad \mathbf{E}H(f_t|f) = t(n\lambda_1 - 2\lambda_\Sigma).$$

At least in simple settings, such as for SDEs with a unique stationary measure for the projective process, the formula follows from a slightly more subtle analysis of volume compression/expansion on $\mathbb{SR}^n$ suitably combined with ergodic theory. Furstenberg [40] was the first to relate relative entropy to Lyapunov exponents; at the generality above, the proof is due to Baxendale [12].

To explore the consequences of Proposition 2.4, let us rewrite it in a more suggestive form. Let $f_x(v) = f(x, v)/\rho(x)$, $f_{t,x}(v) = f_t(x, v)/\rho_t(x)$ denote the *conditional densities* of f and f_t along the fiber $\mathbb{S}_x\mathbb{R}^n \simeq \mathbb{S}^{n-1}$. One can then combine the above formulae into the identity

$$\mathbf{E}H(f_t|f) - \mathbf{E}H(\rho_t|\rho) = \mathbf{E}\int_{\mathbb{R}^n} H(f_{t,x}|f_x)\mathrm{d}\mu(x) = t(n\lambda_1 - \lambda_\Sigma). \tag{2.2}$$

The left-hand side of this identity is the expectation of a positive quantity, while the right-hand side is nonnegative due to the general inequality $n\lambda_1 \geq \lambda_\Sigma$. By the strict convexity, we have

$$n\lambda_1 = \lambda_\Sigma \quad \Longleftrightarrow \quad f_{t,x} \equiv f_x \quad \text{with probability 1 for all } t \geq 0 \text{ and } \mu \text{ almost every } x.$$

Unraveling the definitions, $f_{t,x} \equiv f_x$ means that

$$\left(D_x \Phi^t_\omega\right)_* f_x = f_{\Phi^t_\omega(x)},$$

i.e., the matrices $D_x\Phi_\omega^t$, viewed as acting on $\mathbb{S}^{n-1}$ embedded in $\mathbb{R}^n$, transform the conditional density f_x into the density $f_{\Phi_\omega^t(x)}$ of tangent directions at $\Phi_\omega^t(x)$. This is a very rigid condition in view of the fact that given any two (absolutely continuous) densities h, h' on $\mathbb{S}^{n-1}$,

$$\left\{A \in \mathrm{GL}_n(\mathbb{R}) : A_* h = h'\right\}$$

has empty interior in the space of $n \times n$ matrices. One can obtain the following beautiful dichotomy by a more detailed analysis of the rigidity in a group of matrices in SL_n that preserve a given probability measure; see, e.g., [12,40,63].

Theorem 2.5 (Furstenberg criterion). *Suppose the same setting as Proposition 2.4. If $n\lambda_1 = \lambda_\Sigma$, then one of the following holds:*

(a) *There is a continuously-varying family of inner products $x \mapsto \langle\cdot,\cdot\rangle_x$ with the property that $D_x\Phi_\omega^t$ is an isometry from $\langle\cdot,\cdot\rangle_x$ to $\langle\cdot,\cdot\rangle_{\Phi_\omega^t(x)}$ with probability 1 for all $t \geq 0$.*

(b) *There is a (locally) continuously-varying family of proper subspaces $x \mapsto L_x^i \subset \mathbb{R}^d$ with the property that $D_x\Phi_\omega^t(\bigcup_i L_x^i) = \bigcup_i L_{\Phi_\omega^t(x)}^i$ with probability 1 for all $t \geq 0$.*

Remark 2.6. Note that in the above, the inner products and the L^i are *deterministic*, which is highly rigid for many random systems. Note that they are also *continuously-varying*.

However, if one is interested in deducing $\lambda_1 > 0$, this criterion is really only useful if $\lambda_\Sigma = 0$, i.e., the system is volume preserving, otherwise one only obtains the nondegeneracy $n\lambda_1 > \lambda_\Sigma$. Moreover, Theorem 2.5 lacks any quantitative information, and so it cannot be used to obtain concrete estimates with respect to parameters. Hence, it generally cannot be applied to dissipative systems, even weakly dissipative.

In the volume preserving case, however, criteria à la Furstenberg can be a very powerful tool. In our previous work [14], we used a suitable (partially) infinite-dimensional extension of Theorem 2.5 to show that the Lagrangian flow map (i.e., the trajectories of particles in a fluid) is chaotic when the fluid evolves by the stochastically forced 2D Navier–Stokes equations (called *Lagrangian chaos* in the fluid mechanics literature). See Section 5 for more information.

2.3. The best of both worlds: sign-definite and time-infinitesimal

Proposition 2.4 is, on its face, a quantitative and sign-definite formula for Lyapunov exponents, and this leads to a strong and relatively easy-to-rule-out dichotomy for the degenerate scenario $n\lambda_1 = \lambda_\Sigma$. On the other hand, the formula itself is not straightforward to work with, requiring both the stationary density f for (x_t, v_t) as well as the time-t flow Φ_ω^t and its derivative $D_x\Phi_\omega^t$ as ω varies. In particular, it is unclear how to glean *quantitative* information beyond the "soft" inequality $n\lambda_1 > \lambda_\Sigma$, as would be relevant for a damped system (i.e., $\lambda_\Sigma < 0$).

In view of the sign-indefinite formula (2.1) and its time-infinitesimal version, the Furstenberg–Khasminskii formula, it is reasonable to hope that a time-infinitesimal version of Proposition 2.4 might exist. The authors establish such a formula in our recent work [17].

Proposition 2.7 (Theorem A in [17]). *Assume (x_t, v_t) has a unique stationary measure ν with density $f = \frac{d\nu}{dq}$ on $\mathbb{SR}^n$. Let μ denote the corresponding stationary measure for (x_t) on $\mathbb{R}^n$ with density $\rho = \frac{d\mu}{dx}$. Define the modified* Fisher information

$$\mathrm{FI}(f) = \frac{1}{2}\sum_{i=1}^{r}\int_{\mathbb{SR}^n}\frac{|\tilde{X}_i^* f|^2}{f}\,dq, \quad \mathrm{FI}(\rho) = \frac{1}{2}\sum_{i=1}^{r}\int_{\mathbb{R}^n}\frac{|X_i^*\rho|^2}{\rho}\,dx.$$

Under a mild moment criterion (see [17]), we have

$$\mathrm{FI}(\rho) = -\lambda_\Sigma \quad \textit{and} \quad \mathrm{FI}(f) = n\lambda_1 - 2\lambda_\Sigma.$$

Recall that $\tilde{X}_i^*$ denotes the adjoint of $\tilde{X}_i$ viewed as an operator on $L^2(dq)$.

Remark 2.8. One can show that $\mathrm{FI}(f) - \mathrm{FI}(\rho)$ corresponds to an analogous Fisher information on the conditional densities $\hat{f}_x(v)$, providing the exact time-infinitesimal analogue of (2.2) (see [17]).

These *Fisher-information*-type formulas for Lyapunov exponents enjoy many of the best qualities of the previous formulas: (A) they are sign-definite, like those in Proposition 2.4, and (B) are also time-infinitesimal like those in Proposition 2.3, and so are inherently simpler, requiring only the stationary density f for (x_t, v_t) and how it is acted on by the first-order differential operators $\tilde{X}_i^*$.

A key feature of Proposition 2.7 is that a lower bound on $\mathrm{FI}(f)$ implies a lower bound on $n\lambda_1 - 2\lambda_\Sigma$. The $\mathrm{FI}(f)$ itself has the connotation of a *partial regularity* of f along the forcing directions $\tilde{X}_i$. This is reminiscent of techniques in Hörmander's theory of hypoelliptic operators, where partial regularity along forcing directions implies regularity in *all* directions under an appropriate Lie algebra spanning condition involving the drift X_0. This connection is explored in the next section.

3. QUANTITATIVE LOWER BOUNDS BY THE FISHER INFORMATION

Let us now set about obtaining quantitative estimates on Lyapunov exponents using the Fisher information as in Proposition 2.7. For this, it will be most useful to consider the weakly-forced system

$$dx_t = X_0^\varepsilon(x_t)\,dt + \sqrt{\varepsilon}\sum_{k=1}^{r} X_k^\varepsilon(x_t)\circ dW_t^k, \tag{3.1}$$

where we have also allowed ε dependence in the vector fields X_j^ε. In this case, Proposition 2.7 gives the following Fisher information formula on the stationary density f^ε of the projective process associated to (3.1)

$$\frac{\varepsilon}{2}\sum_{j=1}^{r}\int\frac{|\tilde{X}_j^* f^\varepsilon|^2}{f^\varepsilon}dq = n\lambda_1 - 2\lambda_\Sigma.$$

If $\tilde{X}_j$ has a bounded divergence,[7] by Cauchy–Schwarz inequality, $\exists C > 0$ such that

$$\sum_{j=1}^{r} \left\| \tilde{X}_j f^\varepsilon \right\|_{L^1}^2 \leq C + \mathrm{FI}\left(f^\varepsilon\right) = \left(C + \frac{n\lambda_1 - 2\lambda_\Sigma}{\varepsilon} \right).$$

Hence, we have related L^1-type directional regularity in the forcing directions to the Lyapunov exponents. *If* the lifted forcing directions $\{\tilde{X}_j\}_{j=1}^r$ spanned the entire tangent space $T_w \mathbb{SR}^n$ everywhere, then we would obtain a lower bound of the Lyapunov exponents of the type

$$\left\| f^\varepsilon \right\|_{\dot{W}^{1,1}}^2 \lesssim \left(1 + \frac{n\lambda_1 - 2\lambda_\Sigma}{\varepsilon} \right), \tag{3.2}$$

and so we would find a straightforward lower bound on $n\lambda_1 - 2\lambda_\Sigma$ in terms of the regularity of f^ε. This kind of lower bound is clearly most useful if λ_Σ is small, especially $O(\varepsilon)$, but crucially, it does not have to be *exactly* zero. In this manner, we can treat systems which are close to being volume preserving, but not necessarily exactly volume preserving. This is at the crux of why we can treat systems like Lorenz-96 and Galerkin–Navier–Stokes whereas traditional à la Furstenberg methods based on, e.g., Theorem 2.5 cannot.

3.1. Hypoellipticity

It is not usually the case that $\{\tilde{X}_j\}_{j=1}^r$ spans $T_w \mathbb{SR}^n$ and so the lower bound (3.2) is generally false. For example, for additive noise, the lifts satisfy $\tilde{X}_j = (X_j, 0)$ and so clearly this fails to span $T_w \mathbb{SR}^n$, regardless of whether or not $\{X_j\}_{j=1}^r$ spans $T_x \mathbb{R}^n$. Hence, in general, the Fisher information connects regularity in the lifted forcing directions to the Lyapunov exponents, but a priori, not any other directions in $T_w \mathbb{SR}^n$. For this, we need a concept known as *hypoellipticity*, by which solutions to Kolmogorov equations such as (1.3) or (1.4) can be smooth even when $\mathscr{L}$ is degenerate, i.e., even when the forcing directions do not span the tangent space. This effect was studied first by Kolmogorov [57] in 1934, however, clarity on the effect was not fully obtained until Hörmander's 1967 work [50].

Let us discuss Hörmander's main insights from [50]. It will make sense to quantify fractional regularity along a vector field X using the group e^{tX} and the L^p Hölder-type seminorm (brushing aside minor technical details)

$$|h|_{X,s} := \sup_{t \in (-1,1)} |t|^{-s} \left\| e^{tX} h - h \right\|_{L^p}.$$

Hörmander's original work was based in L^2; our work will be based in L^1. For now, we set $p = 2$.

There are two key ideas in [50]. The first, and simpler idea, comes from the Campbell–Baker–Hausdorff formula, which implies for any two vector fields X, Y that (essentially, the Zassenhaus formula):

$$e^{-tX} e^{-tY} e^{tX} e^{tY} = e^{t^2[X,Y] + O(t^3)},$$

7 This is not the case for our examples, but this will not be important as we will eventually work only locally.

where here $[X, Y]$ is the Lie bracket, i.e., the commutator (see [49] and [50]). In particular, marching forward and then backward by two vector fields X, Y does not quite get us back to where we started (unless X, Y commute). Therefore we have (using that the e^{tX} are bounded on L^p),

$$\left\| e^{t^2[X,Y]+O(t^3)} - I \right\|_{L^p} \lesssim \left\| e^{tX} - I \right\|_{L^p} + \left\| e^{tY} - I \right\|_{L^p} + \left\| e^{-tX} - I \right\|_{L^p} + \left\| e^{-tY} - I \right\|_{L^p},$$

which suggests the remarkable property that any fractional regularity of a function h in directions X, Y, i.e., $|h|_{X,s} + |h|_{Y,s} < \infty$, implies that h also has (a little less) fractional regularity in the commutator direction $[X, Y]$. Another version of Campbell–Baker-0Hausdorff (see [50]) gives

$$e^{t(X+Y)} = e^{tX} e^{tY} e^{t^2[X,Y]} \cdots,$$

where the "$\cdots$" corresponds to a formal product expansion of higher commutators of tX and tY (and thus higher powers in t). Combined with the previous formal discussion, this suggests that regularity in directions X, Y should also supply regularity in the direction $X + Y$ (and indeed, any linear combination). By iterating these heuristics, we get the suggestion that a priori regularity along any set of vector fields $\{Z_0, \dots, Z_r\}$ should imply that there should also be some regularity in *any* direction $Z \in \mathrm{Lie}(Z_0, \dots, Z_r)$, where the *Lie algebra* is given by the span of all possible combinations of commutators

$$\mathrm{Lie}(Z_0, \dots, Z_r) := \mathrm{span}\big\{\mathrm{ad}(Y_m) \dots \mathrm{ad}(Y_1) Y_0 : Y_j \in \{Z_0, Z_1, \dots, Z_r\} m \geq 0\big\},$$

and where $\mathrm{ad}(X)Y := [X, Y]$. In [50], these heuristics are made rigorous with the following functional inequality: Suppose that $\forall z \in \mathbb{R}^n$, $\mathrm{Lie}_z(Z_0, \dots, Z_r) = \{Z(z) : Z \in \mathrm{Lie}(Z_0, \dots, Z_r)\} = T_z\mathbb{R}^n$. Then $\forall s_j \in (0, 1)$, $\exists s_\star$ such that for all $0 < s < s_\star$, $\forall R > 0$, and $\forall h \in C_c^\infty(B(0, R))$, one has

$$\|h\|_{H^s} \lesssim_R \|h\|_{L^2} + \sum_{j=0}^{r} |h|_{Z_j, s_j}. \tag{3.3}$$

In particular, this inequality holds a priori for any $h \in C_c^\infty(B_R(0))$ and it has nothing to do directly with solutions to any PDE. Making this rigorous requires dealing with the errors in the CBH formulas used above. At any step of the argument, these errors are of lower regularity but in new directions, and so dealing with them requires a little finesse and interpolations to close the argument.

Inequality (3.3) is already an interesting observation that can expand the directions of regularity. In particular, one can use an L^1-analogue of (3.3) to provide a lower bound on the Fisher information based on regularity in any direction contained in the Lie algebra of the forcing directions $\{\tilde{X}_1, \dots, \tilde{X}_r\}$. However, Hörmander was *far* from done. Indeed, this is clearly unsatisfying to some degree as this will not even depend on the underlying deterministic dynamical system under consideration, encoded in the drift vector field $\tilde{X}_0$. Moreover, for additive forcing, (3.3) fails to add anything at all. For Hörmander's second

main insight, consider the backward Kolmogorov equation

$$\mathcal{L}g = Z_0 g + \frac{1}{2}\sum_{j=1}^{r} Z_j^2 g = F. \tag{3.4}$$

Assuming $\{Z_j\}_{j=0}^r$ have bounded divergence,[8] one obtains the standard L^2-"energy" estimate:

$$\sum_{j=1}^{r} \|Z_j g\|_{L^2}^2 \lesssim \|g\|_{L^2}^2 + \|F\|_{L^2}^2.$$

After applying a smooth cutoff $\chi_R(x) = \chi(x/R)$ where $\chi \in C_c^\infty(B_2(0))$, $0 \le \chi \le 1$, and $\chi(x) = 1$ for $|x| \le 1$, and dealing with the commutators as in a Caccioppoli estimate, the functional inequality (3.3) combined with this estimate implies that *if* $\mathrm{Lie}_z(Z_1, \dots, Z_r) = T_z\mathbb{R}^n$ at all z, then we would obtain an estimate like

$$\|\chi_R g\|_{H^s} \lesssim_R \|g\|_{L^2(B_{2R}(0))} + \|F\|_{L^2(B_{2R}(0))}.$$

However, as discussed above, this condition on the vector fields is often too strong to be useful for us here.

However, another natural a priori estimate on g is available from (3.4). Indeed, pairing (3.4) with a test function φ, we obtain

$$\left|\int \varphi Z_0 f \,\mathrm{d}q\right| \le \frac{1}{2}\sum_{j=1}^{r} \left\|Z_j^*\varphi\right\|_{L^2} \|Z_j g\|_{L^2} \lesssim \frac{1}{2}\sum_{j=1}^{r} \left(\|\varphi\|_{L^2} + \|Z_j\varphi\|_{L^2}\right)\left(\|g\|_{L^2} + \|F\|_{L^2}\right).$$

This simple observation shows that for solutions of $\mathcal{L}g = F$, H^1-type regularity in the forcing directions automatically provides a corresponding dual H^{-1}-type regularity on $Z_0 g$. The cornerstone of [50] is the following functional inequality (i.e., again, not directly related to solutions of any PDEs): if one has $\mathrm{Lie}_z(Z_0, Z_1, \dots, Z_r) = \mathbb{R}^n$ everywhere, then $\exists s \in (0,1)$ such that if $R > 0$ and $h \in C_c^\infty(B_R(0))$, then

$$\|h\|_{H^s} \lesssim \|h\|_{L^2} + \sup_{\varphi:\|\varphi\|_{L^2}+\sum_{j=1}^r \|Z_j\varphi\|_{L^2}\le 1} \left|\int \varphi Z_0 h \,\mathrm{d}q\right| + \sum_{j=1}^{r} \|Z_j h\|_{L^2} =: \|h\|_{H^1_{\mathrm{hyp}}}. \tag{3.5}$$

The key heuristic behind this functional inequality is the following observation:

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left\|e^{tZ_0}h - h\right\|_{L^2}^2 &= \left\langle e^{tZ_0}h - h, Z_0 e^{tZ_0}h\right\rangle \\ &\le \left(\|h\|_{L^2} + \sum_{j=1}^{r}\left\|Z_j e^{tZ_0^*}\left(e^{tZ_0}h - h\right)\right\|_{L^2}\right)\|g\|_{H^1_{\mathrm{hyp}}}.\end{aligned}$$

Therefore, *if* we had something like

$$\sum_{j=1}^{r}\left\|Z_j e^{tZ_0^*}\left(e^{tZ_0}h - h\right)\right\|_{L^2} \lesssim \sum_{j=1}^{r}\|Z_j h\|_{L^2}, \tag{3.6}$$

8 Alternatively, one can consider the estimates suitably localized.

then we could combine the L^2-estimate on $\{Z_j\}_{j=1}^r$ with the corresponding dual negative regularity in the Z_0 direction to obtain some positive fractional regularity in the Z_0 direction, specifically we would have $1/2$ regularity from

$$\left\| e^{tZ_0} h - h \right\|_{L^2}^2 \lesssim t \|h\|_{H^1_{\text{hyp}}}^2 .$$

Unfortunately (3.6) does not generally hold,[9] and Hörmander uses a rather ingenious regularization argument to turn this heuristic into reality. We shall henceforth call functional inequalities of the type (3.5) *Hörmander inequalities*.

The gain in regularity from (3.5) combines with the Kolmogorov equation to get the estimate

$$\|g\chi_R\|_{H^s} \lesssim \|g\|_{L^2(B_{2R}(0))} + \|F\|_{L^2(B_{2R}(0))},$$

and so provides an analogue of the gain of regularity when studying elliptic equations (though only fractional regularity). As in that theory, this regularity gain can be iterated to imply that any L^2-solution of $\mathcal{L}g = F$ is C^∞ if $F \in C^\infty$ [50].

3.2. Uniform hypoellipticity

Next, we want to make the arguments which are quantitative with respect to parameters, and hence we will introduce the notion of *uniform* hypoellipticity. Let us formalize the definition of Hörmander's condition for elliptic- and parabolic-type equations. For a manifold M, we denote by $\mathfrak{X}(M)$ the set of smooth vector fields on M.

Definition 3.1 (Hörmander's condition). Given a manifold $\mathcal{M}$ and a collection of vector fields

$$\{Z_0, Z_1, \dots, Z_r\} \subset \mathfrak{X}(\mathcal{M}),$$

we define collections of vector fields $\mathscr{X}_0 \subseteq \mathscr{X}_1 \subseteq \cdots$ recursively by

$$\begin{aligned} \mathscr{X}_0 &= \{Z_j : j \geq 1\}, \\ \mathscr{X}_{k+1} &= \mathscr{X}_k \cup \big\{[Z_j, Z] : Z \in \mathscr{X}_k,\ j \geq 0\big\}. \end{aligned}$$

We say that $\{Z_i\}_{i=0}^r$ satisfies the *parabolic Hörmander condition* if there exists k such that for all $w \in \mathcal{M}$,

$$\operatorname{span}\big\{Z(w) : Z \in \mathscr{X}_k\big\} = T_w\mathcal{M}.$$

We say that $\{Z_i\}_{i=0}^r$ satisfies the *(elliptic) Hörmander condition* if this holds with $\mathscr{X}_0 = \{Z_j : j \geq 0\}$.

Note that the parabolic Hörmander condition is slightly stronger than the elliptic Hörmander condition.

---

9 As in the easier inequalities above, the heuristic (3.6) neglects the creation of higher-order commutators; in fact, one requires regularity in many other directions in $\operatorname{Lie}(Z_0, \dots, Z_r)$ as a result.

Definition 3.2 (Uniform Hörmander's condition). Let $\mathcal{M}$ be a manifold, and let $\{Z_0^\varepsilon, Z_1^\varepsilon, \dots, Z_r^\varepsilon\} \subset \mathfrak{X}(\mathcal{M})$ be a set of vector fields parameterized by $\varepsilon \in (0,1]$. With $\mathscr{X}_k$ defined as in Definition 3.1 in the parabolic case (resp. elliptic), we say $\{Z_0^\varepsilon, Z_1^\varepsilon, \dots, Z_r^\varepsilon\}$ satisfies the uniform parabolic (resp. elliptic) Hörmander condition on $\mathcal{M}$ if $\exists k \in \mathbb{N}$ such that for any open, bounded set $U \subseteq \mathcal{M}$ there exist constants $\{K_n\}_{n=0}^\infty$ such that for all $\varepsilon \in (0,1]$ and all $x \in U$, there is a finite subset $V(x) \subset \mathscr{X}_k$ such that $\forall \xi \in T_x\mathcal{M}$,

$$|\xi| \le K_0 \sum_{Z \in V(x)} \big|Z(x) \cdot \xi\big|, \qquad \sum_{Z \in V(x)} \|Z\|_{C^n} \le K_n.$$

This definition stipulates that any ε dependence is *locally* (on the manifold) uniform in terms of both regularity and spanning. Now we are ready to state the uniform L^1-type Hörmander inequality suitable for use with the Fisher information, proved in [17]. There are many works extending Hörmander's theory in various ways see, e.g., [1,4,19,44,56,62,72] and the references therein. However, as far as the authors are aware, there are no works in the L^1–L^∞ framework. We also need to consider the forward Kolmogorov equation $\tilde{\mathcal{L}}^* f = 0$, as opposed to the case of the backward Kolmogorov equation considered by Hörmander [50]; this changes some details but little of significant consequence is different.

Theorem 3.3 (L^1-type uniform Hörmander inequality, [17, THEOREM 4.2]). *Let $\{X_0^\varepsilon, X_1^\varepsilon, \dots, X_r^\varepsilon\}$ be a collection of vector fields on $\mathbb{SR}^n$ satisfying the uniform elliptic Hörmander condition as in Definition 3.2. Then, $\exists s_\star \in (0,1)$ such that if $B_R(x_0) \subset \mathbb{R}^n$ is an open ball and $h \in C_c^\infty(B_R(x_0) \times \mathbb{S}^{n-1})$, then for all $0 < s < s_\star$, $\exists C = C(R, x_0, s)$ such that $\forall \varepsilon \in (0,1)$ the following fractional regularity[10] estimate holds* uniformly in ε:

$$\|h\|_{W^{s,1}} \le C\left(\|h\|_{L^1} + \sup_{\varphi : \|\varphi\|_{L^\infty} + \sum_{j=1}^r \|X_j^\varepsilon \varphi\|_{L^\infty} \le 1} \left|\int \varphi\big(X_0^\varepsilon\big)^* h \, \mathrm{d}q\right| + \sum_{j=1}^r \big\|\big(X_j^\varepsilon\big)^* h\big\|_{L^1}\right).$$

In particular, applying a smooth cutoff $\chi_R := \chi(x/R)$ for some $\chi \in C_c^\infty(B_2(0))$ with $0 \le \chi \le 1$ and $\chi \equiv 1$ if $|x| \le 1$ to the Kolmogorov equation $\tilde{\mathcal{L}}^* f^\varepsilon = 0$ (assuming also $\|f^\varepsilon\|_{L^1} = 1$) and suitably estimating the commutators, we obtain

$$\big\|\chi_R f^\varepsilon\big\|_{W^{s,1}}^2 \lesssim_R 1 + \mathrm{FI}(f^\varepsilon). \tag{3.7}$$

Remark 3.4. Hypoellipticity plays a classical role in the theory of SDEs. In particular, the parabolic Hörmander condition of Definition 3.1 is exactly the condition most often used to deduce that the Markov semigroup $\mathcal{P}_t$ is strong Feller (the exposition of [46] is especially intuitive). The parabolic Hörmander condition also often plays a role in proving irreducibility via geometric control theory (see discussions in [42,48,52] and specifically in [17] in regards

---

10 For $s \in (0,1)$, we may define $W^{s,1}$ on a geodesically complete, n-dimensional Riemannian manifold with bounded geometry $\mathcal{M}$ as

$$\|w\|_{W^{s,1}} = \|w\|_{L^1} + \left(\int_{\mathcal{M}} \int_{h \in T_x\mathcal{M} : |h| < \delta_0} \frac{|w(\exp_x h) - w(x)|}{|h|^{s+n}} \mathrm{d}h \mathrm{d}q(x)\right),$$

where $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ is the exponential map on $\mathcal{M}$ and $\mathrm{d}q$ is the Riemannian volume measure. See, e.g., [81] for more details.

to the projective process). For many applications, it is likely that the parabolic Hörmander's condition will be used to prove that there exists a unique stationary measure ν for the projective process (via Doob–Khasminskii [33]), as required to apply Proposition 2.7. Hence the condition of uniformity-in-ε in Definition 3.2 will usually be the only additional information required to apply Theorem 3.3.

Remark 3.5. Quantitative arguments based on L^2 Hörmander inequalities can be found in [2, 19] (completed concurrently with or after [17]). Thinking about hypoellipticity in terms of functional inequalities, rather than qualitative statements about regularity of solutions to PDEs, has other important advantages as well, for example, it is easier to adapt classical elliptic and parabolic PDE methods, such as De Giorgi or Moser iterations, into hypoelliptic equations [19, 44, 72].

Obtaining the above Theorem 3.3 follows an argument generally based on Hörmander's original paper [50], however, the L^1–L^∞ framework, as opposed to the self-dual L^2 framework in [50], necessitates a more complicated regularization argument than that used [50] (which was already quite delicate!). Moreover, as we are always interested in sphere bundles here, one cannot avoid working on smooth manifolds, which at least under the assumption of geodesic completeness, only adds some technical complexity rather than fundamental difficulties.

Let us briefly see, heuristically, how one would approach the proof of Theorem 3.3. Motivated by the above discussion regarding [50], the main challenge is to obtain $1/2$ of a derivative of L^1 Hölder-type regularity in the $\tilde{X}_0^*$ "direction." By a bootstrap-type argument, we may assume that we have corresponding regularity along all of the other vector fields in $\mathrm{Lie}_z(\tilde{X}_0, \tilde{X}_1, \dots, \tilde{X}_r)$ (see [17] for details). Let S_t be a (carefully designed) regularization operator $S_t : L^p \to L^p$. We obtain for any $w \in C_c^\infty(B_R(0) \times \mathbb{S}^{n-1})$,

$$\left\| e^{t\tilde{X}_0^*} w - w \right\|_{L^1} \le \left\| e^{t\tilde{X}_0^*} \left(S_\tau^* w - w \right) \right\|_{L^1} + \left\| S_\tau^* w - w \right\|_{L^1} + \left\| e^{t\tilde{X}_0^*} S_\tau^* w - S_\tau^* w \right\|_{L^1}.$$

We eventually set $\tau \sim \sqrt{t}$ and the regularization operator will be designed so that the first two terms are $O(\tau)$, thus we need mainly to work on the latter term, which by duality is estimated by

$$\left\| e^{t\tilde{X}_0^*} S_\tau^* w - S_\tau^* w \right\|_{L^1} \le \sup_{\|v\|_{L^\infty} \le 1} \left| \int_0^t \int_{\mathbb{SR}^n} (e^{s\tilde{X}_0} v) X_0^* S_\tau^* w \, \mathrm{d}q \mathrm{d}s \right|,$$

and, for any fixed $v \in L^\infty$, we have

$$\begin{aligned} \left| \int_{\mathbb{SR}^n} (e^{s\tilde{X}_0} v) X_0^* S_\tau^* w \, \mathrm{d}q \right| &\le \left| \int_{\mathbb{SR}^n} (e^{s\tilde{X}_0} v) [\tilde{X}_0, S_\tau]^* w \, \mathrm{d}q \right| + \left| \int_{\mathbb{SR}^n} (S_\tau e^{s\tilde{X}_0} v) \tilde{X}_0^* w \, \mathrm{d}q \right| \\ &\le \left\| e^{sX_0} v \right\|_{L^\infty} \left\| [\tilde{X}_0, S_\tau]^* w \right\|_{L^1} \\ &\quad + \left(\left\| S_\tau e^{s\tilde{X}_0} v \right\|_\infty + \sum_{j=1}^r \left\| X_j S_\tau e^{s\tilde{X}_0} v \right\|_{L^\infty} \right) \mathfrak{D}(w), \end{aligned}$$

where

$$\mathfrak{D}(h) := \sup_{\varphi : \|\varphi\|_{L^\infty} + \sum_{j=1}^r \|X_j^\varepsilon \varphi\|_{L^\infty} \le 1} \left| \int_{\mathbb{SR}^n} \varphi \left(\tilde{X}_0^\varepsilon \right)^* h \, \mathrm{d}q \right|.$$

Hence, the challenge is designing a regularizer such that the commutator $[\tilde{X}_0, S_\tau]^*$ loses only $O(\tau^{-1})$ using no a priori regularity in the $\tilde{X}_0$ direction, and similarly that S_τ regularizes the forcing fields $\tilde{X}_j$ like $O(\tau^{-1})$. To do this, we let S_τ be a modified version of Hörmander's regularizer, which averages the function along directions in $\mathrm{Lie}_z(\tilde{X}_0, \dots, \tilde{X}_r)$ a corresponding amount (higher commutators corresponding to less regularization) in a carefully ordered way. Specifically, because these "directional mollifiers" do not commute, the order in which they are applied is very important. Hörmander regularized with S_τ, whereas we are fundamentally regularizing with its adjoint S_τ^*, which reverses the delicate ordering. Despite the added difficulty, this turns out to be an important choice for our framework.

4. CHAOS FOR 2D GALERKIN–NAVIER–STOKES AND RELATED MODELS

In this section, we outline how to apply the above ideas to prove a positive Lyapunov exponent for Galerkin truncations of the stochastic 2D Navier–Stokes. A general class of models with similar bilinear drift term, which we call *Euler-like* systems, are given by the following SDE:

$$\mathrm{d}x_t^\varepsilon = \big(B(x_t^\varepsilon, x_t^\varepsilon) - \varepsilon A x_t^\varepsilon\big)\mathrm{d}t + \sum_{k=1}^{r} X_k \mathrm{d}W_t^k. \tag{4.1}$$

Here, $\{X_k\}_{k=1}^r$ is a collection of *constant* (x-independent) forcing vector fields (i.e., additive forcing) while $B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a nontrivial (not identically zero) bilinear drift that satisfies

$$\operatorname{div} B = 0, \quad x \cdot B(x, x) = 0,$$

so in particular the unforced $\varepsilon = 0$ dynamics preserve the norm,[11] given by $\frac{1}{2}\|x\|^2$, and volume in $\mathbb{R}^n$ (i.e., the Liouville property). The term $-\varepsilon A$ provides weak linear damping, where A is assumed to be a symmetric, positive-definite $n \times n$ matrix. Stochastically forced versions of the Lorenz 96 model (L96) [67], Galerkin truncations of 2D and 3D Navier–Stokes on a torus (of arbitrary aspect ratio) [20, 36, 78] and truncations of commonly used shell models for turbulence [34,43,61,84] can be cast in this form. The 2D stochastic Galerkin–Navier–Stokes equations will be described in more detail in Section 4.3 below.

The bilinearity of B implies that solutions can be naturally rescaled into a weakly-damped, weakly-driven system, and the two scalings are equivalent as far as Lyapunov exponents are concerned. Indeed, while the scaling (4.1) is common among models of complex real-world systems, the stationary measure μ has characteristic energy $\int |x|^2 \mathrm{d}\mu(x) \approx \varepsilon^{-1}$. Since we are concerned with the regime $\varepsilon \ll 1$, it is natural to rescale and consider a weakly-damped, weakly-driven system. Hence, it is more natural to rescale so that the long-time behavior remains bounded and nonvanishing as $\varepsilon \to 0$. By rescaling $x_t^\varepsilon \mapsto \sqrt{\varepsilon} x^\varepsilon_{\sqrt{\varepsilon}t}$, replacing $\varepsilon \mapsto \varepsilon^{3/2}$, and using the self-similarity of Brownian motion, we get an equivalent in law,

11 In the case of the vorticity form of the 2D Navier–Stokes equations that we will be studying below, this quantity is the *enstrophy*.

weakly-driven, weakly damped form

$$\mathrm{d}x_t^\varepsilon = \big(B(x_t^\varepsilon, x_t^\varepsilon) - \varepsilon A x_t^\varepsilon\big)\mathrm{d}t + \sqrt{\varepsilon}\sum_{k=1}^{r} X_k \mathrm{d}W_t^k. \tag{4.2}$$

Most importantly, this rescaling does not affect our results on Lyapunov exponents, since upon setting $\hat{\varepsilon} = \varepsilon^{3/2}$, the Lyapunov exponent $\hat{\lambda}_1^{\hat{\varepsilon}}$ of (4.2) with parameter $\hat{\varepsilon}$ is related to the Lyapunov exponent λ_1^ε of (4.1) by the identity $\frac{\hat{\lambda}_1^{\hat{\varepsilon}}}{\hat{\varepsilon}} = \frac{\lambda_1^\varepsilon}{\varepsilon}$. This kind of scaling is sometimes called *fluctuation–dissipation* due to the balance between the forcing and the dissipation.

For this class of systems (4.1), our result below gives a sufficient condition for a positive Lyapunov exponent in terms of projective hypoellipticity, i.e., if the lifted vector fields $\{\tilde{X}_0^\varepsilon, \tilde{X}_1, \dots, \tilde{X}_r\}$ corresponding to the projective process $(x_t^\varepsilon, v_t^\varepsilon)$ (denoting $X_0^\varepsilon(x) = B(x,x) - \varepsilon A x$) satisfy Hörmander's condition on $\mathbb{SR}^n$.

Theorem 4.1 ([17, THEOREM C]). *Assume that*

(i) $\{\tilde{X}_0^\varepsilon, \tilde{X}_1, \dots, \tilde{X}_r\}$ *satisfy the elliptic Hörmander's condition uniformly in* $\varepsilon \in (0,1)$ *as in Definition* 3.2*;*

(ii) *the bilinear term* B *is nontrivial, i.e.,* $B(x,x) \neq 0$ *for some* $x \in \mathbb{R}^n$*; and*

(iii) *the process* $(x_t^\varepsilon, v_t^\varepsilon)$ *admits a unique stationary density* f^ε.

Then, the limit defining the Lyapunov exponent λ_1^ε *of* (4.1) *exists, and satisfies*

$$\lim_{\varepsilon \to 0} \frac{\lambda_1^\varepsilon}{\varepsilon} = \infty.$$

In particular, $\exists \varepsilon_0 > 0$ *such that for all* $\varepsilon \in (0, \varepsilon_0)$*, one has* $\lambda_1^\varepsilon > 0$.

A sketch of the proof of Theorem 4.1 is given in Section 4.1 below. The most difficult part of applying this result to a concrete system, e.g., Galerkin–Navier–Stokes, is to prove the parabolic Hörmander condition for the projective process: general comments on this problem are given in Section 4.2, while the issue of affirming this for Galerkin–Navier–Stokes is taken up in Section 4.3.

Given parabolic Hörmander's condition, unique existence of f^ε follows, via the Doob–Khasminskii theorem, from topological irreducibility of $(x_t^\varepsilon, v_t^\varepsilon)$, i.e., the ability to approximately control random trajectories by controlling noise paths. For Euler-like models such as (4.1), this follows from geometric control theory arguments and the following well-known cancelation condition on $B(x,x)$ (known to hold for many models such as Galerkin–Navier–Stokes, cf. [42, 48]): there exists a collection of vectors $\{e_1, \dots, e_s\} \subset \mathbb{R}^n$ with

$$\operatorname{span}\{e_1, \dots, e_s\} = \operatorname{span}\{X_1, \dots, X_r\}$$

such that for each $1 \le k \le s$, $B(e_k, e_k) = 0$. For more details, see Section 5.3 of [17].

Remark 4.2. The inverse Lyapunov exponent $(\lambda_1^\varepsilon)^{-1}$ is sometimes called the Lyapunov time, and is the "typical" length of time one must wait for tangent vectors to grow by a factor of e. Thus, the estimate $\lambda_1^\varepsilon \gg \varepsilon$ implies that the Lyapunov time is $\ll \varepsilon^{-1}$. On the other hand,

ε^{-1} is the typical amount of time it takes for the Brownian motion $\sqrt{\varepsilon}W_t$ to reach an $O(1)$ magnitude; for this reason, it is reasonable to refer to ε^{-1} as a kind of "diffusion timescale." So, stated differently, our results indicate that as $\varepsilon \to 0$, arbitrarily many Lyapunov times elapse before a single "diffusion time" has elapsed, indicating a remarkable sensitivity of the Lyapunov exponent to the presence of noise.

Based on these ideas, one would like to assert that the scaling $\lambda_1^\varepsilon \gg \varepsilon$ implies that the deterministic dynamics are "close" to positive Lyapunov exponent dynamics, agnostic as to whether the zero-noise system has a positive exponent on a positive area set. However, this assertion does not follow from the scaling $\lambda_1^\varepsilon \gg \varepsilon$ alone: even if the Brownian motion itself is small, there could already be a substantial difference between random and corresponding deterministic (zero-noise) trajectories well before time ε^{-1}, e.g., if there is already strong vector growth in the deterministic dynamics. For more on this, see the open problems in Section 6.

4.1. Zero-noise limit and rigidity: proof sketch of Theorem 4.1

Applying the Fisher information identity (Proposition 2.7) to the Euler-like system (4.2) and using that $\lambda_\Sigma^\varepsilon = -\varepsilon \operatorname{tr} A$, we obtain

$$\mathrm{FI}(f^\varepsilon) = \frac{n\lambda_1^\varepsilon}{\varepsilon} + 2\operatorname{tr} A.$$

By the regularity lower bound (3.7), this implies that, for each open ball $B_R(0)$, we have the lower bound

$$\left\| \chi_R f^\varepsilon \right\|_{W^{s,1}}^2 \lesssim_R 1 + \frac{\lambda_1^\varepsilon}{\varepsilon},$$

where the regularity $s \in (0,1)$ and the implicit constant $C = C_R$ are independent of ε.

From this, we see that if $\liminf_\varepsilon \varepsilon^{-1}\lambda_1^\varepsilon$ were to remain bounded, then f^ε would be bounded in $W^{s,1}_{\mathrm{loc}}$ uniformly in ε. As $W^{s,1}$ is locally compactly embedded in L^1 and f^ε naturally satisfies certain uniform-in-ε moment bounds, one can deduce, by sending $\varepsilon \to 0$, that at least one of the following must hold true (see Proposition 6.1, [17] for details):

(a) either $\lim_{\varepsilon\to 0} \frac{\lambda_1^\varepsilon}{\varepsilon} = \infty$; or

(b) the zero-noise flow (x_t^0, v_t^0) admits a stationary density $f^0 \in L^1(\mathbb{SR}^n)$.

Let us consider alternative (b). While it is natural and common for the projective processes of SDE to admit stationary densities, the existence of an absolutely continuous invariant measure $f^0 \mathrm{d}q$ for the projective process of the $\varepsilon = 0$ problem

$$\dot{x}_t = B(x_t, x_t), \tag{4.3}$$

is quite rigid. Indeed, in view of the fact that vector growth implies concentration of Lebesgue measure in projective space (cf. the discussion in Section 2.1 after Proposition 2.3), the existence of an invariant density essentially rules out *any* vector growth for the $\varepsilon = 0$ projective process (x_t^0, v_t^0). Precisely, a generalization of Theorem 2.32 in [8] (see [17] for details)

implies that there is a measurably varying Riemannian metric $x \mapsto g_x$ such that Φ^t is an *isometry* with respect to g_x, namely

$$g_x\big(D_x\Phi^t v, D_x\Phi^t w\big) = g_{\Phi^t(x)}(v,w), \quad v,w \in T_x\mathbb{R}^n,$$

where $\Phi^t : \mathbb{R}^n \to \mathbb{R}^n$ is the flow associate to the $\varepsilon = 0$ dynamics (4.3). So, we see that if $\liminf_{\varepsilon\to 0} \varepsilon^{-1}\lambda_1 < \infty$, then the deterministic, measure-preserving $\varepsilon = 0$ dynamics must be in a situation analogous to possibility (a) in Theorem 2.5.

In our setting, we show that there is necessarily *some* norm growth as $t \to \infty$ for the $\varepsilon = 0$ dynamics due to *shearing between conserved energy shells* $\{x \in \mathbb{R}^n : |x|^2 = E\}$. This is straightforward to check: due to the scaling symmetry $\Phi^t(\alpha x) = \alpha\Phi^{\alpha t}(x)$, $\alpha > 0$, we have the following orthogonal decomposition of the linearization $D_x\Phi^t$ in the direction $x \in \mathbb{R}^n$:

$$D_x\Phi^t x = \Phi^t(x) + tB\big(\Phi^t(x), \Phi^t(x)\big),$$

noting that $y \cdot B(y,y) \equiv 0$ for all $y \in \mathbb{R}^n$. Hence, one obtains the lower bound

$$\big|D_x\Phi^t\big| \geq t\frac{|B(\Phi^t(x),\Phi^t(x))|}{|x|}$$

for each $x \in \mathbb{R}^n \setminus \{0\}$ and each $t > 0$. This contradicts the existence of the Riemannian metric g_x via a Poincaré recurrence argument and the fact that the set of stationary points $\{x \in \mathbb{R}^n : B(x,x) = 0, |x|^2 \leq R\}$ is a zero volume set. This is summarized in the following proposition (a proof of which is given in [17]).

Proposition 4.3 ([17, PROPOSITION 6.2]). *Assume that the bilinear mapping B is not identically 0. Let ν be any invariant probability measure for $\widehat{\Phi}^t$ (the flow corresponding to the (deterministic) $\varepsilon = 0$ projective process) with the property that $\nu(A \times \mathbb{S}^{n-1}) = \mu(A)$, where $\mu \ll \mathrm{Leb}_{\mathbb{R}^n}$. Then, ν is singular with respect to volume measure $\mathrm{d}q$ on $\mathbb{SR}^n$.*

4.2. Verifying projective hypoellipticity: a sufficient condition

We address here the challenge of verifying the parabolic Hörmander condition on the sphere bundle $\mathbb{SR}^n$. Recall that given a smooth vector field X on $\mathbb{R}^n$ we define its lift $\tilde{X}$ to the sphere bundle $\mathbb{SR}^n$ by

$$\tilde{X}(x,v) = \big(X(x), \nabla X(x)v - v\big\langle v, \nabla X(x)v\big\rangle\big),$$

where $\nabla X(x)$ denotes the (covariant) derivative of X at x and is viewed as a linear endomorphism on $T_x\mathbb{R}^n$. Many of the following general observations about the lifted fields were made in [12]; see also [17] for detailed discussions.

An important property is that the lifting operation can be seen to be a Lie algebra isomorphism onto its range with respect to the Lie bracket, i.e., $[\tilde{X},\tilde{Y}] = [X,Y]\tilde{}$. Using this observation, the parabolic Hörmander condition (see Definition 3.1) on $\mathbb{SR}^n$ for the lifts of a collection of vector fields

$$\{X_0, X_1, \ldots, X_r\} \subset \mathfrak{X}(\mathbb{R}^n)$$

can be related to nondegeneracy properties of the Lie subalgebra $\mathfrak{m}_x(X_0; X_1, \dots, X_r)$ of $\mathfrak{sl}(T_x\mathbb{R}^n)$ defined by

$$\mathfrak{m}_x(X_0; X_1, \dots, X_r) := \left\{ \nabla X(x) - \frac{1}{n} \operatorname{div} X(x) \operatorname{Id} : X \in \operatorname{Lie}(X_0; X_1, \dots, X_r),\ X(x) = 0 \right\},$$

where

$$\operatorname{Lie}(X_0; X_1, \dots, X_r) := \operatorname{Lie}\big(X_1, \dots, X_r, [X_0, X_1], \dots, [X_0, X_r]\big),$$

is the *zero-time ideal* generated by $\{X_0, X_1, \dots, X_r\}$, with X_0 a distinguished "drift" vector field (recall that $\mathfrak{sl}_n(T_x\mathbb{R}^n)$ is the Lie algebra of traceless linear endomorphisms of $T_x\mathbb{R}^n$).

Particularly, if for each $x \in \mathbb{R}^n$, $\mathfrak{m}_x(X_0; X_1, \dots, X_r)$ acts *transitively* on $\mathbb{S}^{n-1}$ in the sense that, for each $(x, v) \in \mathbb{SR}^n$, one has

$$\big\{ Av - v\langle v, Av\rangle : A \in \mathfrak{m}_x(X_0; X_1, \dots, X_r) \big\} = T_v\mathbb{S}^{n-1}, \tag{4.4}$$

then the parabolic Hörmander condition for $\{X_0, X_1, \dots, X_r\}$ on $\mathbb{R}^n$ is equivalent to the parabolic Hörmander condition for the lifts $\{\tilde{X}_0, \tilde{X}_1, \dots, \tilde{X}_r\}$ on $\mathbb{SR}^n$. Moreover, the uniform parabolic Hörmander condition is satisfied on $\mathbb{SR}^n$ if and only if it is satisfied on $\mathbb{R}^n$ and (4.4) holds uniformly in the same sense as in Definition 3.2. Since $\mathfrak{sl}(\mathbb{R}^n)$ acts transitively on $\mathbb{R}^n \setminus \{0\}$ (see, for instance, [27]), a sufficient condition for transitivity on $\mathbb{S}^{n-1}$ is

$$\mathfrak{m}_x(X_0; X_1, \dots, X_r) = \mathfrak{sl}\big(T_x\mathbb{R}^n\big).$$

In the specific case of Euler-like models (4.2) with $X_0^\varepsilon(x) = B(x, x) - \varepsilon A$ and $\{X_k\}_{k=1}^r$ as in (4.2), the situation can be simplified if $\operatorname{Lie}(X_0; X_1, \dots, X_r)$ contains the constant vector fields $\{\partial_{x_k}\}_{k=1}^n$. In this case, the family of x and ε-independent endomorphisms

$$H_k := \nabla\big[\partial_{x_k}, X_0^\varepsilon\big] = \nabla[\partial_{x_k}, B], \quad k = 1, \dots, n,$$

generate the Lie algebra $\mathfrak{m}_x(X_0^\varepsilon; X_1, \dots, X_r)$ at all $x \in \mathbb{R}^n$. This argument implies the following sufficient condition for projective spanning.

Corollary 4.4 (See [17]). *Consider the bilinear Euler-like models* (4.2). *If* $\operatorname{Lie}(X_0; X_1, \dots, X_r)$ *contains* $\{\partial_{x_k}\}_{k=1}^n$, *then* $\{\tilde{X}_0^\varepsilon, \tilde{X}_1, \dots, \tilde{X}_r\}$ *satisfy the uniform parabolic Hörmander condition (in the sense of Definition* 3.2*) on* $\mathbb{SR}^n$ *if*

$$\operatorname{Lie}(H^1, \dots, H^n) = \mathfrak{sl}(\mathbb{R}^n).$$

This criterion is highly useful, having reduced projective spanning to a question about a single Lie algebra of trace-free matrices.

In [17], we verified this condition directly for the Lorenz 96 system [67], which is defined for n unknowns in a periodic array by the nonlinearity B given by

$$B_\ell(x, x) = x_{\ell+1}x_{\ell-1} - x_{\ell-2}x_{\ell-1}. \tag{4.5}$$

The traditional case is $n = 40$, but it can be considered in any finite dimension. In particular, we proved the following.

Corollary 4.5 ([17, COROLLARY D]). *Consider the L96 system given by (4.2) with the nonlinearity (4.5) and $X_k = q_k e_k$ for $k \in \{1, \dots, r\}$, $q_k \in \mathbb{R}$, and e_k the canonical unit vectors. If $q_1, q_2 \neq 0$ and $n \geq 7$, then*

$$\lim_{\varepsilon \to 0} \frac{\lambda_1^\varepsilon}{\varepsilon} = \infty.$$

In particular $\exists \varepsilon_0 > 0$ such that $\lambda_1^\varepsilon > 0$ if $\varepsilon \in (0, \varepsilon_0)$.

4.3. Projective hypoellipticity for 2D Galerkin–Navier–Stokes

Let us now see how we can go about verifying the projective hypoellipticity condition for a high-dimensional model of physical importance, namely Galerkin truncations of the 2D stochastic Navier–Stokes equations on the torus of arbitrary side-length ratio $\mathbb{T}_r^2 = [0, 2\pi) \times [0, \frac{2\pi}{r})$ (periodized) for $r > 0$. Recall that the Navier–Stokes equations on $\mathbb{T}_r^2$ in vorticity form are given by

$$\partial_t w + u \cdot \nabla w - \varepsilon \Delta w = \sqrt{\varepsilon} \dot{W}_t,$$

where w is the vorticity and u is the divergence-free velocity field coming from the Biot–Savart law $u = \nabla^\perp (-\Delta)^{-1} w$ and $\dot{W}_t$ is a white-in time, colored-in-space Gaussian forcing which we will take to be diagonalizable with respect to the Fourier basis with Fourier transform supported on a small number of modes.

In the work [20] by the first and last authors of this note, we consider a Galerkin truncation of the 2D stochastic Navier–Stokes equations at an arbitrary frequency $N \geq 1$ in Fourier space by projecting onto the Fourier modes in the truncated lattice

$$\mathbb{Z}_{0,N}^2 := \big\{(k_1, k_2) \in \mathbb{Z}^2 \setminus \{0\} : \max\big\{|k_1|, |k_2|\big\} \leq N\big\} \subseteq \mathbb{Z}^2,$$

giving rise to an $n = |\mathbb{Z}_{0,N}^2| = (2N+1)^2 - 1$ dimensional stochastic differential equation with the reality constraint $w_{-k} = \overline{w}_k$ for $w = (w_k) \in \mathbb{C}^{\mathbb{Z}_{0,N}^2}$ (that is, the vector is indexed over $\mathbb{Z}_{0,N}^2$) governed by

$$\mathrm{d}w_k = \big(B_k(w, w) - \varepsilon |k|_r^2 w_k\big)\mathrm{d}t + \sqrt{\varepsilon}\mathrm{d}W^k, \tag{4.6}$$

where $|k|_r^2 := k_1^2 + r^2 k_2^2$, and $W_t^k = \alpha_k W_t^{a,k} + i\beta_k W_t^{b,k}$ are independent complex Wiener processes satisfying $W_t^k = \overline{W}_t^{-k}$ ($W_t^{a,k}, W_t^{b,k}$ are standard i.i.d. Wiener processes) with α_k, β_k arbitrary such that $\alpha_k = 0 \Leftrightarrow \beta_k = 0$. The symmetrized nonlinearity $B_k(w, w)$ is given by

$$B_k(w, w) := \frac{1}{2} \sum_{j+\ell=k} c_{j,\ell} w_j w_\ell, \quad c_{j,\ell} := \langle j^\perp, \ell \rangle_r \left(\frac{1}{|\ell|_r^2} - \frac{1}{|j|_r^2} \right)$$

where the sum runs over all $j, \ell \in Z_{0,N}^2$ such that $j + \ell = k$ and we are using the notation $\langle j^\perp, \ell \rangle_r := r(j_2 \ell_1 - j_1 \ell_2)$. In what follows, the coefficient $c_{j,\ell}$ always depends on r, but we suppress the dependence for notational simplicity.

We will regard the configuration space $\mathbb{C}^{\mathbb{Z}_{0,N}^2}$ as a complex manifold with complexified tangent space spanned by the complex basis vectors $\{\partial_{w_k} : k \in \mathbb{Z}_{0,N}^2\}$ (Wirtinger derivatives) satisfying $\partial_{w_{-k}} = \bar{\partial}_{w_k}$. See [51] for the notion of complexified tangent space

and [20] for discussion on how to use this complex framework for checking Hörmander's condition. In this basis, we can formulate the SDE (4.6) in the canonical form

$$\mathrm{d}w_t = X_0^\varepsilon(w_t) + \sum_{k\in\mathcal{Z}^0} \sqrt{\varepsilon}\partial_{w_k}\mathrm{d}W_t^k,$$

where the drift vector field X_0^ε is given by $X_0^\varepsilon(w) := \sum_{k\in\mathbb{Z}^2_{0,N}}(B_k(w,w) - \varepsilon|k|_r^2 w_k)\partial_{w_k}$ and the set of driving modes $\mathcal{Z}^0$ is given by $\mathcal{Z}^0 := \{k \in \mathbb{Z}^2_{0,N} : \alpha_k, \beta_k \neq 0\}$.

As in the setting of [36, 47], we consider very degenerate forcing and study how it spreads throughout the system via the nonlinearity $B_\ell(w,w)$. Specifically, define the sets

$$\mathcal{Z}^n = \big\{\ell \in \mathbb{Z}^2_{0,N} : \ell = j + k,\ j \in \mathcal{Z}^0,\ k \in \mathcal{Z}^{n-1}, c_{j,k} \neq 0\big\}, \quad n \geq 0$$

and assume that the driving modes $\mathcal{Z}^0$ satisfy $\bigcup_{n\geq 0}\mathcal{Z}^n = \mathbb{Z}^2_{0,N}$. Under this assumption on $\mathcal{Z}^0$, it can be shown (see [17] Proposition 3.6 or [36, 47]) that the complexified Lie algebra $\mathrm{Lie}(X_0^\varepsilon; \{\partial_{w_k} : k \in \mathcal{Z}^0\})$ contains the constant vector fields $\{\partial_{w_k} : k \in \mathbb{Z}^2_{0,N}\}$ and therefore satisfies the uniform parabolic Hörmander condition on $\mathbb{C}^{\mathbb{Z}^2_{0,N}}$.

4.3.1. A distinctness condition on a diagonal subalgebra

As discussed in Section 4.2, in order to verify projective hypoellipticity for the vector fields $X_0^\varepsilon; \{\partial_{w_k} : k \in \mathcal{Z}^0\}$, it suffices to study the generating properties of a suitable matrix Lie algebra. In [20], we show this can be reformulated to a condition on the constant, *real valued* matrices $H^k = \nabla[\partial_{w_k}, B]$, $k \in \mathbb{Z}^2_{0,N}$, represented in $\{\partial_{w_k}\}$ coordinates by $(H^k)_{\ell,j} = \partial_{w_j}\partial_{w_k}B_\ell(w,w) = c_{j,k}\delta_{\ell=j+k}$. After obtaining this reformulation, the main result of [20] is the following nondegeneracy property of the matrices $\{H^k\}$.

Theorem 4.6 ([20, THEOREM 2.13], see also Proposition 3.11). *Consider the 2D stochastic Galerkin–Navier–Stokes equations with frequency truncation N on $\mathbb{T}^2_r$ and suppose that $N \geq 392$. Then, the following holds:*

$$\mathrm{Lie}\big(\{H^k : k \in \mathbb{Z}^2_{0,N}\}\big) = \mathfrak{sl}_{\mathbb{Z}^2_{0,N}}(\mathbb{R}), \tag{4.7}$$

where $\mathfrak{sl}_{\mathbb{Z}^2_{0,N}}(\mathbb{R})$ denotes the Lie algebra of real-valued traceless matrices indexed by the truncated lattice $\mathbb{Z}^2_{0,N}$. Therefore projective hypoellipticity holds for (4.2) and, by Theorem 4.1, the top Lyapunov exponent satisfies $\lim_{\varepsilon\to 0}\varepsilon^{-1}\lambda_1^\varepsilon = \infty$.

Remark 4.7. Verifying the Lie algebra generating condition (4.7) is quite challenging due to the fact that there there are $n = |\mathbb{Z}^2_{0,N}|$ matrices and $n^2 - 1$ degrees of freedom to span. The matrices are also banded in the sense that for each k, $(H^k)_{\ell,j}$ couples most of the lattice values ℓ, j along the band $k = \ell - j$ and therefore it is a major challenge to isolate elementary matrices (matrices with only one nonzero entry) as one can do rather easily in "local in frequency" models like L96 (4.5) (see [17]). Moreover, brute force computational approaches that successively generate Lie bracket generations and count the rank by Gaussian elimination (such as the Lie tree algorithm in [37]) are only available for fixed $r \in \mathbb{R}_+$ and $N \in \mathbb{Z}_+$, and can be subject to numerical error (for instance, if r is chosen irrational) which destroy the validity of the proof.

In order to show that (4.7) holds, in [20] we take an approach inspired by the root-space decomposition of semisimple Lie algebras and study genericity properties of the following diagonal subalgebra of $\mathrm{Lie}(\{H^k\})$

$$\mathfrak{h} := \mathrm{span}\{\mathbb{D}^k : k \in \mathbb{Z}^2_{0,N}\},$$

where $\mathbb{D}^k = [H^k, H^{-k}]$ are a family of diagonal matrices with diagonal elements $\mathbb{D}^k_i = (\mathbb{D}^k)_{ii}$ given by

$$\mathbb{D}^k_i = c_{i,k}c_{i+k,k}\mathbb{1}_{\mathbb{Z}^2_{0,N}}(i+k) - c_{i,k}c_{i-k,k}\mathbb{1}_{\mathbb{Z}^2_{0,N}}(i-k).$$

Using that, for a given diagonal matrix $\mathbb{D} \in \mathfrak{sl}_{\mathbb{Z}^2_{0,N}}(\mathbb{R})$, the adjoint action $\mathrm{ad}(\mathbb{D}) : \mathfrak{sl}_{\mathbb{Z}^2_{0,N}}(\mathbb{R}) \to \mathfrak{sl}_{\mathbb{Z}^2_{0,N}}(\mathbb{R})$, where $\mathrm{ad}(\mathbb{D})H = [\mathbb{D}, H]$, has eigenvectors given by the elementary matrices $E^{i,j}$ (i.e., a matrix with 1 in the ith row and jth column and 0 elsewhere), $\mathrm{ad}(\mathbb{D})E^{i,j} = (\mathbb{D}_i - \mathbb{D}_j)E^{i,j}$ means that $\mathrm{ad}(\mathbb{D})$ has a simple spectrum if the diagonal entries of $\mathbb{D}$ have *distinct differences*, $\mathbb{D}_i - \mathbb{D}_j \neq \mathbb{D}_{i'} - \mathbb{D}_{j'}$, $(i,j) \neq (i',j')$. This implies that if H is a matrix with nonzero nondiagonal entries and $\mathbb{D}$ has distinct differences, then for $M = n^2 - n$, the Krylov subspace

$$\mathrm{span}\{H, \mathrm{ad}(\mathbb{D})H, \mathrm{ad}(\mathbb{D})^2 H, \ldots, \mathrm{ad}(\mathbb{D})^{M-1}H\}$$

contains the set $\{E^{i,j} : i,j \in \mathbb{Z}^2_{0,N},\ i \neq j\}$, which is easily seen to generate $\mathfrak{sl}_{\mathbb{Z}^2_{0,N}}(\mathbb{R})$.

However, in our setting the diagonal matrices $\mathbb{D}^k$ have an inversion symmetry $\mathbb{D}^k_{-i} = -\mathbb{D}^k_i$ and therefore there *cannot* be a matrix in $\mathfrak{h}$ with all differences distinct. Moreover, we do not have a matrix with *all* off diagonal entries nonzero due to the degeneracies present in $c_{j,k}$ and the presence of the Galerkin cut-off. Nevertheless, in [20] we are able to deduce the following sufficient condition on the family $\{\mathbb{D}^k\}$, ensuring that (4.7) holds:

Proposition 4.8 ([20, COROLLARY 4.9 AND LEMMA 5.2]). *Let $N \geq 8$. If for each $(i,j,\ell,m) \in (\mathbb{Z}^2_{0,N})^4$ satisfying $i+j+\ell+m = 0$ and $(i+j,\ell+m) \neq 0$, $(i+\ell, j+m) \neq 0$, $(i+m, j+\ell) \neq 0$, there exists a $k \in \mathbb{Z}^2_{0,N}$ such that*

$$\mathbb{D}^k_i + \mathbb{D}^k_j + \mathbb{D}^k_\ell + \mathbb{D}^k_m \neq 0, \tag{4.8}$$

then (4.7) holds.

The proof of Proposition 4.8 is not straightforward. However, its proof uses some similar ideas as the proof of (4.8) but is otherwise significantly easier, so we only discuss the latter.

4.3.2. Verifying the distinctness condition using computational algebraic geometry

The distinctness condition (4.8) is not a simple one to verify. Indeed, ignoring the Galerkin cut-off N for now, $\mathbb{D}^k_i$ are rational algebraic expressions in the variables (i,k,r) (by comprising products and sums of the coefficients $c_{j,k}$), and therefore proving (4.8) amounts

to showing that the family of *Diophantine equations*[12]

$$\mathbb{D}_i^k + \mathbb{D}_j^k + \mathbb{D}_\ell^k + \mathbb{D}_m^k = 0, \quad \text{for each } k \in \mathbb{Z}_{0,N}^2 \tag{4.9}$$

have *no solutions* (i, j, ℓ, m, r) satisfying the constraints of Proposition 4.8. Due to the complexity of the expression for $\mathbb{D}_i^k$, there is little hope to verify such a result by hand (the resulting polynomials are of degree 16 in 9 variables). But, if one extends each of the 9 variables $(i, j, \ell, m, r) = (i_1, i_2, j_1, j_2, \ell_1, \ell_2, m_1, m_2, r)$ to the algebraically closed field $\mathbb{C}$, then (4.9) along with $i + j + \ell + m = 0$ defines a polynomial ideal I with an associated algebraic variety $\mathbb{V}(I)$ in $\mathbb{C}^9$. Such a high dimensional variety is rather complicated due to the inherent symmetries of the rational equation in (4.9); however, its analysis is nonetheless amenable to techniques from algebraic geometry, particularly the strong Nullstellensatz and computer algorithms for computing Gröbner bases (see **[30]** for a review of the algebraic geometry concepts). Indeed, without the Galerkin cut-off (the formal infinite-dimensional limit), in **[20]** we proved, by computing Gröbner bases in rational arithmetic using the F4 algorithm **[38]** implemented in the computer algebra system Maple **[70]**, that the identity $\mathbb{V}(I) = \mathbb{V}(g)$ holds, where g is the following "saturating" polynomial

$$\begin{aligned} g(i, j, \ell, m, r) = r^2 |i|_r^2 |j|_r^2 |\ell|_r^2 |m|_r^2 \big(|i + j|_r^2 + |\ell + m|_r^2\big)\big(|i + \ell|_r^2 + |j + m|_r^2\big) \\ \times \big(|i + m|_r^2 + |j + \ell|_r^2\big) \end{aligned}$$

whose nonvanishing encodes the constraints in Proposition 4.8, thereby showing that (4.8) holds.

Dealing with the Galerkin truncation adds significant difficulties to the proof as the associated rational system (4.9) is instead piecewise defined (depending on k and N) and therefore does not easily reduce to a problem about polynomial inconsistency. Nonetheless, by considering 34 different polynomial ideals associated to different possible algebraic forms, in **[20]** we were able to show that if N is taken large enough (bigger than 392 to be precise) then (4.8) still holds with the Galerkin truncation present and therefore Theorem 4.6 holds.

Finally, it is worth remarking that even without the Galerkin cut-off, the system of rational equations (4.9) is complex enough to become computationally intractable (even for modern computer algebra algorithms) without some carefully chosen simplifications, variable orderings, choice of saturating polynomial g, and sheer luck; see **[20]** for more details.

5. LAGRANGIAN CHAOS IN STOCHASTIC NAVIER–STOKES

At present, the results above based on Proposition 2.7 are restricted to finite-dimensional problems. Indeed, even while the Fisher information can potentially be extended to infinite dimensions under certain conditions,[13] for any parabolic SPDE problem, we will

12 At least considering $r = 1$ or another fixed, rational number.

13 If $X^*\nu \ll \nu$ and we define $\beta_{X^*}^\nu := \frac{\mathrm{d}\tilde{X}^*\nu}{\mathrm{d}\nu}$, then $\mathrm{FI}(f) = \frac{1}{2}\sum_k \|\beta_{\tilde{X}^*}^\nu\|_{L^2(\nu)}^2$, and there is no explicit dependence on any reference measure or Riemannian metric; see **[17]** for more details.

always have $\lambda_\Sigma = -\infty$. The existence of positive Lyapunov exponents for the infinite-dimensional, stochastic Navier–Stokes equations remains open as of the writing of this note.

However, there is another important problem in fluid mechanics where we have been able to make progress. Consider the (infinite-dimensional) 2D Navier–Stokes equations[14] in $\mathbb{T}^2$,

$$\partial_t u_t + (u_t \cdot \nabla u_t + \nabla p - \nu \Delta u_t) = \sum_k q_k e_k \dot{W}_t^k, \quad \operatorname{div} u_t = 0, \tag{5.1}$$

where the $q_k \in \mathbb{R}$ and e_k are eigenfunctions of the Stokes operator. The *Lagrangian flow map* $\varphi_{\omega,u}^t : \mathbb{T}^2 \mapsto \mathbb{T}^2$ is defined by the trajectories of particles moving with the fluid

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_{\omega,u}^t(x) = u_t\big(\varphi_{\omega,u}^t(x)\big), \quad \varphi_{\omega,u}^0(x) = x,$$

where note that the diffeomorphism $\varphi_{\omega,u}^t$ depends on the initial velocity u and the noise path ω and is therefore a cocycle over the skew product $\Theta_t : \Omega \times H^s \circlearrowleft$, where $\Theta_t(\omega, u) = (\theta_t\omega, \Psi_\omega^t(u))$ and $\Psi_\omega^t : H^s \circlearrowleft$ is the 2D Navier–Stokes flow on H^s associated with (5.1). One can naturally ask whether or not (u_t) is chaotic, as we have done in previous sections, or if the motion of particles immersed in the fluid is chaotic, e.g., if the Lagrangian Lyapunov exponent is strictly positive. The latter is known as *Lagrangian chaos* [3,5,9,26,31,41,86] (to distinguish it from chaos of (u_t) itself, which is sometimes called *Eulerian chaos*). While both are expected to be observed in turbulent flows, Lagrangian chaos is not incompatible with Eulerian "order," i.e., a negative exponent for the (u_t) process.

In [14] we proved, under the condition that $|q_k| \approx |k|^{-\alpha}$ for some $\alpha > 10$, that $\exists \lambda_1 > 0$ deterministic and independent of initial x and initial velocity u such that the following limit holds almost surely:

$$\lim_{t\to\infty} \frac{1}{t} \log\big|D_x \varphi_{\omega,u}^t\big| = \lambda_1 > 0. \tag{5.2}$$

This Lagrangian chaos was later upgraded in [15,18] to the much stronger property of uniform-in-diffusivity, almost sure exponential mixing. To formulate this notion, we consider (g_t) a passive scalar solving the (random) advection–diffusion equation

$$\partial_t g_t + u_t \cdot \nabla g_t = \kappa \Delta g_t, \quad g_0 = g,$$

for $\kappa \in [0, 1]$ and a fixed, mean-zero scalar $g \in L^2(\mathbb{T}^2)$. In [15,18], we proved that there exists a (deterministic) constant $\mu > 0$ such that for all $\kappa \in [0, 1]$ and initial divergence free $u \in H^s$ (for some sufficiently large s), there exists a random constant $D = D(\omega, \kappa, u)$ such that for all $g \in H^1$ (mean-zero)

$$\|g_t\|_{H^{-1}} \le D e^{-\mu t} \|g\|_{H^1}$$

where D is almost surely finite and satisfies the *uniform-in-κ* moment bound (for some fixed constant q and for any $\eta > 0$),

$$\mathbf{E} D^2 \lesssim_\eta \big(1 + \|u\|_{H^s}\big)^q e^{\eta \|u\|_{H^1}^2}.$$

14 The 3D Navier–Stokes equations can be treated provided the $-\nu\Delta u_t$ is replaced with the hyperviscous damping $\nu\Delta^2 u_t$.

One can show that this result is essentially optimal up to getting sharper quantitative estimates on μ and D, at least if $\kappa = 0$ [15,18]. This uniform, exponential mixing plays the key role in obtaining a proof of Batchelor's power spectrum [11] of passive scalar turbulence in some regimes [16].

Let us simply comment on the Lagrangian chaos statement (5.2), as it is most closely related to the rest of this note. The main step is to deduce an analogue of Theorem 2.5 for the Lagrangian flow map, using that while the Lagrangian flow map depends on an infinite-dimensional Markov process, the Jacobian $D_x\varphi^t_{\omega,u}$ itself is finite dimensional. This is done in our work [14] by extending Furstenberg's criterion to handle general linear cocycles over infinite-dimensional processes in the same way that $D_x\varphi^t_{\omega,u}$ depends on the sample paths (u_t).

The Lagrangian flow is divergence-free, and thus the Lagrangian Lyapunov exponents satisfy $\lambda_\Sigma = 0$ and $\lambda_1 \geq 0$, so ruling out the degenerate situations in Theorem 2.5 would immediately imply $\lambda_1 > 0$. A key difficulty in this infinite-dimensional context is to ensure that the rigid invariant structures (now functions of the fluid velocity field u and the Lagrangian tracer position x) in our analogue of Theorem 2.5 vary *continuously* as functions of u and x. It is at this step that we require the nondegeneracy type condition on the noise $|q_k| \gtrsim |k|^{-\alpha}$, which is used to ensure that the Markov process $(u_t, \varphi^t(x))$ is strong Feller.

At the time of writing, it remains an interesting open problem to extend our works [14,15,18] to degenerate noise such as that used in [47] or [58]. It bears remarking that the methods of [47] apply to the one-point process $(u_t, \varphi^t(x))$ (this is used in our work [15]), however, it is nevertheless unclear how to prove Lagrangian chaos without a sufficiently strong analogue of Theorem 2.5, and it is unclear how to obtain such a theorem without the use of the strong Feller property.

6. LOOKING FORWARD

The work we reviewed here raises a number of potential research directions.

Tighter hypoelliptic regularity estimates. The scaling $\lambda_1^\varepsilon \gg \varepsilon$ that naturally follows from our above analysis is surely suboptimal – even if the deterministic problem were to be completely integrable, the scaling would likely be $O(\varepsilon^\gamma)$ for some $\gamma < 1$ depending on dimension (see, e.g., [13,76]). To begin with, one may attempt to strengthen the hypoelliptic regularity estimate by refining the ε scaling to something like

$$\|f^\varepsilon\|^2_{W^{s,1}} \lesssim 1 + \frac{n\lambda_1^\varepsilon - 2\lambda_\Sigma^\varepsilon}{\varepsilon^\gamma},$$

for some constant $0 < \gamma < 1$. If such an estimate were true, the same compactness-rigidity argument of Theorem 4.1 would imply a scaling like $\lambda_1^\varepsilon \gtrsim \varepsilon^\gamma$. An improvement of this type seems plausible given the proof of Theorem 3.3. It might be necessary, in general, to use a more specialized norm on the left-hand side, but local weak L^1 compactness, i.e., equiintegrability, is all that is really required for the compactness-rigidity argument to apply.

Beyond compactness-rigidity. Compactness-rigidity arguments may remain limited in their ability to yield optimal or nearly optimal scalings for λ_1, regardless of the ways one can improve Theorem 3.3. Another approach is to find some way to work more directly on $\varepsilon > 0$. This was essentially the approach of works [13,76], however, the method of these papers only applies if one has a nearly-complete understanding of the pathwise random dynamics. We are unlikely to ever obtain such a complete understanding of the dynamics of models such as L96 or Galerkin–Navier–Stokes, but there may be hope that partial information, such as the isolation of robust, finite-time exponential growth mechanisms, could be used to obtain better lower bounds on $\|f^\varepsilon\|_{W^{s*,1}}$. An approach with a vaguely related flavor for random perturbations of discrete-time systems, including the Chirikov standard map, was carried out in the previous work [23].

Finer dynamical information: moment Lyapunov exponents. Lyapunov exponents provide asymptotic exponential growth rates of the Jacobian, but they provide no quantitative information on how long it takes for this growth to be realized with high probability. One tool to analyze this is the study of large deviations of the convergence of the sequences $\frac{1}{t}\log|D_x\Phi_\omega^t v|$. The associated rate function is the Legendre transform of the *moment Lyapunov exponent function* $p \mapsto \Lambda(p) := \lim_{t\to\infty}\frac{1}{t}\log \mathbf{E}|D_x\Phi_\omega^t v|^p$ (the limit defining $\Lambda(p)$ exists and is independent of (x, v, ω) under fairly general conditions [6]). It would be highly interesting to see if the quantitative estimates obtained by, e.g., Theorem 4.1 extend also to quantitative estimates on the moment Lyapunov exponents. We remark that the moment Lagrangian Lyapunov exponents play a key role in our works [15,18].

Lyapunov times of small-noise perturbations of completely integrable systems. The phase space of a completely integrable Hamiltonian flow is foliated by invariant torii along which the dynamics is a translation flow—such systems are highly ordered and nonchaotic. On the other hand, small perturbations of the Hamiltonian are known to break the most "resonant" of these torii, while torii with sufficiently "nonresonant" frequencies persist due to KAM theory. It is an interesting and highly challenging open problem to prove that this "breakage" results in the formation of a positive-volume set admitting a positive Lyapunov exponent. For the most part such problems are wide open, and related to the standard map conjecture discussed in Section 1.1. The recent work of Berger and Turaev [21] established a renormalization technique for proving the *existence* of smooth perturbations resulting in a positive Lyapunov exponent, but it remains open to affirm how "generic" such perturbations actually are.

The following is a closely related *stochastic* dynamics problem: starting from a completely integrable system and adding a small amount of noise, how many Lyapunov times elapse for the random dynamics before the "stochastic divergence" timescale when the deterministic flow and the stochastic flow differ by $O(1)$? Estimating the stochastic divergence timescale is essentially a large deviations problem, and has already been carried out for small random perturbations of completely integrable systems; see, e.g., [39]. On the other hand, estimating Lyapunov times beyond the crude $(\lambda_1^\varepsilon)^{-1}$ estimate is a large deviations estimate for the convergence of finite-time Lyapunov exponents to their asymptotic value

λ_1^ε. The associated rate function in this case is the Legendre transform of the moment Lyapunov exponent $\Lambda(p)$ mentioned earlier; a positive result for the program described above would require quantitative-in-ε estimates on $\Lambda(p)$.

More general noise models. One simple potential extension is Theorem 4.1 to different types of multiplicative noise. Another important extension would be to noise models which are not white-in-time, for example, noise of the type used in [58], which is challenging because our work is deeply tied to the elliptic nature of the generator $\tilde{\mathcal{L}}^*$. A simpler example of nonwhite forcing can be constructed from "towers" of coupled Ornstein–Uhlenbeck processes, which can be built to be C^k in time for any $k \geq 0$ (see, e.g., [14,15] for details).

Lagrangian chaos. There are several directions of research to extend our results in [14, 15, 18], such as studying degenerate noise as in [47, 58], extending to more realistic physical settings such as bounded domains with stochastic boundary driving, and extending Proposition 2.7 to the Lagrangian flow map in a variety of settings, which would help to facilitate quantitative estimates (note one will have to use the conditional density version so that one does not see the effect of the λ_Σ associated to the Navier–Stokes equations themselves).

FUNDING

J.B. was supported by National Science Foundation CAREER grant DMS-1552826 and National Science Foundation RNMS #1107444 (Ki-Net). A.B. was supported by National Science Foundation grant DMS-2009431. This material was based upon work supported by the National Science Foundation under Award No. DMS-1803481.

REFERENCES

[1] F. Abedin and G. Tralli, Harnack inequality for a class of Kolmogorov–Fokker–Planck equations in non-divergence form. *Arch. Ration. Mech. Anal.* **233** (2019), no. 2, 867–900.

[2] D. Albritton, R. Beekie, and M. Novack, Enhanced dissipation and Hörmander's hypoellipticity. 2021, arXiv:2105.12308.

[3] C. H. Amon, A. M. Guzmán, and B. Morel, Lagrangian chaos, Eulerian chaos, and mixing enhancement in converging–diverging channel flows. *Phys. Fluids* **8** (1996), no. 5, 1192–1206.

[4] F. Anceschi, S. Polidoro, and M. A. Ragusa, Moser's estimates for degenerate Kolmogorov equations with non-negative divergence lower order coefficients. *Nonlinear Anal.* **189** (2019), 111568.

[5] T. M. Antonsen Jr., Z. Fan, E. Ott, and E. Garcia-Lopez, The role of chaotic orbits in the determination of power spectra of passive scalars. *Phys. Fluids* **8** (1996), no. 11, 3094–3104.

[6] L. Arnold, A formula connecting sample and moment stability of linear stochastic systems. *SIAM J. Appl. Math.* **44** (1984), no. 4, 793–802.

[7] L. Arnold, Random dynamical systems. In *Dynamical systems*, pp. 1–43, Springer, 1995.

[8] L. Arnold, D. C. Nguyen, and V. Oseledets, Jordan normal form for linear cocycles. *Random Oper. Stoch. Equ.* **7** (1999), no. 4, 303–358.

[9] E. Balkovsky and A. Fouxon, Universal long-time properties of Lagrangian statistics in the Batchelor regime and their application to the passive scalar problem. *Phys. Rev. E* **60** (1999), no. 4, 4164.

[10] L. Barreira and Y. B. Pesin, *Lyapunov exponents and smooth ergodic theory*. Univ. Lecture Ser. 23, American Mathematical Soc., 2002.

[11] G. K. Batchelor, Small-scale variation of convected quantities like temperature in turbulent fluid part 1. general discussion and the case of small conductivity. *J. Fluid Mech.* **5** (1959), no. 1, 113–133.

[12] P. H. Baxendale, Lyapunov exponents and relative entropy for a stochastic flow of diffeomorphisms. *Probab. Theory Related Fields* **81** (1989), no. 4, 521–554.

[13] P. H. Baxendale and L. Goukasian, Lyapunov exponents for small random perturbations of Hamiltonian systems. *Ann. Probab.* **30** (2002), 101–134.

[14] J. Bedrossian, A. Blumenthal, and S. Punshon-Smith, Lagrangian chaos and scalar advection in stochastic fluid mechanics. *J. Euro. Math. Soc.* (to appear), arXiv:1809.06484.

[15] J. Bedrossian, A. Blumenthal, and S. Punshon-Smith, Almost-sure exponential mixing of passive scalars by the stochastic Navier-Stokes equations. *Ann. Probab.* (to appear), arXiv:1905.03869.

[16] J. Bedrossian, A. Blumenthal, and S. Punshon-Smith, The Batchelor spectrum of passive scalar turbulence in stochastic fluid mechanics at fixed reynolds number. *Comm. Pure Appl. Math.* (to appear), arXiv:1911.11014.

[17] J. Bedrossian, A. Blumenthal, and S. Punshon-Smith, A regularity method for lower bounds on the Lyapunov exponent for stochastic differential equations. *Invent. Math.* (to appear), arXiv:2007.15827.

[18] J. Bedrossian, A. Blumenthal, and S. Punshon-Smith, Almost-sure enhanced dissipation and uniform-in-diffusivity exponential mixing for advection-diffusion by stochastic Navier-Stokes. *Probab. Theory Related Fields* **179** (2021), no. 3, 777–834.

[19] J. Bedrossian and K. Liss, Quantitative spectral gaps and uniform lower bounds in the small noise limit for Markov semigroups generated by hypoelliptic stochastic differential equations. *Probab. Math. Phys.* **2** (2020), no. 3, 477–532.

[20] J. Bedrossian and S. Punshon-Smith, Chaos in stochastic 2d Galerkin–Navier–Stokes. 2021, arXiv:2106.13748.

[21] P. Berger and D. Turaev, On Herman's positive entropy conjecture. *Adv. Math.* **349** (2019), 1234–1288.

[22] A. Blumenthal, J. Xue, and Y. Yang, Lyapunov exponents for random perturbations of coupled standard maps. 2020, arXiv:2004.10626.

[23] A. Blumenthal, J. Xue, and L.-S. Young, Lyapunov exponents for random perturbations of some area-preserving maps including the standard map. *Ann. of Math.* **185** (2017), 285–310.

[24] A. Blumenthal, J. Xue, and L.-S. Young, Lyapunov exponents and correlation decay for random perturbations of some prototypical 2d maps. *Comm. Math. Phys.* **359** (2018), no. 1, 347–373.

[25] A. Blumenthal and Y. Yang, Positive Lyapunov exponent for random perturbations of predominantly expanding multimodal circle maps. 2018, arXiv:1805.09219.

[26] T. Bohr, M. H. Jensen, G. Paladin, and A. Vulpiani, *Dynamical systems approach to turbulence*. Cambridge University Press, 2005.

[27] W. M. Boothby and E. N. Wilson, Determination of the transitivity of bilinear systems. *SIAM J. Control Optim.* **17** (1979), no. 2, 212–221.

[28] A. Carverhill, Furstenberg's theorem for nonlinear stochastic systems. *Probab. Theory Related Fields* **74** (1987), no. 4, 529–534.

[29] B. V. Chirikov, et al., A universal instability of many-dimensional oscillator systems. *Phys. Rep.* **52** (1979), no. 5, 263–379.

[30] D. Cox, J. Little, and D. OShea, *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 2013.

[31] A. Crisanti, M. Falcioni, A. Vulpiani, and G. Paladin, Lagrangian chaos: transport, mixing and diffusion in fluids. *La Rivista del Nuovo Cimento* **14** (1991), no. 12, 1–80.

[32] S. Crovisier and S. Senti, A Problem for the 21st/22nd Century. *EMS Newsl.* **114** (2019), 8–-13.

[33] G. Da Prato and J. Zabczyk, *Ergodicity for infinite dimensional systems*. London Math. Soc. Lecture Note Ser. 229. Cambridge University Press, 1996.

[34] P. D. Ditlevsen, *Turbulence and shell models*. Cambridge University Press, 2010.

[35] P. Duarte, Plenty of elliptic islands for the standard family of area preserving maps. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **11** (1994), 359–409, Elsevier.

[36] W. E and J. C. Mattingly, Ergodicity for the Navier-Stokes equation with degenerate random forcing: Finite-dimensional approximation. *Comm. Pure Appl. Math.* **54** (2001), no. 11, 1386–1402.

[37] D. Elliott, *Bilinear control systems: Matrices in action*. Springer, Dordrecht, 2009.

[38] J.-C. Faugere, A new efficient algorithm for computing gröbner bases (f4). *J. Pure Appl. Algebra* **139** (1999), no. 1–3, 61–88.

[39] M. I. Freidlin and A. D. Wentzell, *Random perturbations of Hamiltonian systems*. Mem. Amer. Math. Soc. 523, American Mathematical Soc., 1994.

[40] H. Furstenberg, Noncommuting random products. *Trans. Amer. Math. Soc.* **108** (1963), no. 3, 377–428.

[41] S. Galluccio and A. Vulpiani, Stretching of material lines and surfaces in systems with lagrangian chaos. *Phys. A* **212** (1994), no. 1–2, 75–98.

[42] N. E. Glatt-Holtz, D. P. Herzog, and J. C. Mattingly, Scaling and saturation in infinite-dimensional control problems with applications to stochastic partial differential equations. *Ann. PDE* **4** (2018), no. 2.

[43] E. Gledzer, Hydrodynamic-type system admitting two quadratic integrals of motion. *Dokl. Akad. Nauk SSSR* **209** (1973), 1046–1048.

[44] F. Golse, C. Imbert, C. Mouhot, and A. Vasseur, Harnack inequality for kinetic fokker-planck equations with rough coefficients and application to the Landau equation. *Ann. Sc. Norm. Super. Pisa Cl. Sci.* **19** (2016), no. 1, 253–295.

[45] A. Gorodetski, On the stochastic sea of the standard map. *Comm. Math. Phys.* **309** (2012), no. 1, 155–192.

[46] M. Hairer, On Malliavin's proof of Hörmander's theorem. *Bull. Sci. Math.* **135** (2011), no. 6–7, 650–666.

[47] M. Hairer and J. C. Mattingly, Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing. *Ann. of Math.* **164** (2006), no. 3, 993–1032.

[48] D. P. Herzog and J. C. Mattingly, A practical criterion for positivity of transition densities. *Nonlinearity* **28** (2015), no. 8, 2823.

[49] G. Hochschild, *The structure of lie groups*. Holden-day, 1965.

[50] L. Hörmander, Hypoelliptic second order differential equations. *Acta Math.* **119** (1967), no. 1, 147–171.

[51] D. Huybrechts, *Complex geometry: an introduction*. Springer, 2005.

[52] V. Jurdjevic, *Geometric control theory*. Cambridge university press, 1997.

[53] A. Karimi and M. R. Paul, Extensive chaos in the Lorenz-96 model. *Chaos* **20** (2010), no. 4, 043105.

[54] R. Khasminskii, *Stochastic stability of differential equations*. Stoch. Model. Appl. Probab. 66, Springer, 2011.

[55] Y. Kifer, *Ergodic theory of random transformations*. Progr. Probab. 10, Springer, 2012.

[56] A. E. Kogoj and S. Polidoro, Harnack inequality for hypoelliptic second order partial differential operators. *Potential Anal.* **45** (2016), no. 14, 545–555.

[57] A. Kolmogorov, Zufällige Bewegungen (zur Theorie der Brownschen Bewegung). *Ann. of Math. (2)* **35** (1934), no. 1, 116–117.

[58] S. Kuksin, V. Nersesyan, and A. Shirikyan, Exponential mixing for a class of dissipative pdes with bounded degenerate noise. *Geom. Funct. Anal.* **30** (2020), 1–62.

[59] S. Kuksin and A. Shirikyan, *Mathematics of two-dimensional turbulence*. Cambridge Tracts in Math. 194, Cambridge University Press, 2012.

[60] H. Kunita, *Stochastic flows and stochastic differential equations*. Cambridge Stud. Adv. Math. 24, Cambridge university press, 1997.

[61] V. S. L'vov, E. Podivilov, A. Pomyalov, I. Procaccia, and D. Vandembroucq, An optimal shell model of turbulence. *Phys. Rev. E* **58** (1998), no. 2, 1811.

[62] A. Lanconelli, A. Pascucci, and S. Polidoro, Gaussian lower bounds for non-homogeneous kolmogorov equations with measurable coefficients. *J. Evol. Equ.* **20** (2020), 1–19.

[63] F. Ledrappier, Positivity of the exponent for stationary sequences of matrices. In *Lyapunov exponents*, pp. 56–73, Springer, 1986.

[64] Z. Lian and M. Stenlund, Positive Lyapunov exponent by a random perturbation. *Dyn. Syst.* **27** (2012), no. 2, 239–252.

[65] G. M. Lieberman, *Second order parabolic differential equations*. World scientific, 1996.

[66] P.-D. Liu and M. Qian, *Smooth ergodic theory of random dynamical systems*. Springer, 2006.

[67] E. N. Lorenz, Predictability: A problem partly solved. In *Predictability of weather and climate*, pp. 40–58, Cambridge University Press, 2006.

[68] R. MacKay, An appraisal of the ruelle-takens route to turbulence. In *The global geometry of turbulence*, pp. 233–246, Springer, 1991.

[69] A. J. Majda, *Introduction to turbulent dynamical systems in complex systems*. Springer, 2016.

[70] Maple, *maplesoft, a division of waterloo maple inc*, waterloo, ontario, 2020.

[71] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer, 2012.

[72] C. Mouhot, De Giorgi–Nash–Moser and Hörmander theories: new interplays. In *Proceedings of the international congress of mathematicians rio de janeiro*, pp. 2467–2493, World Scientific, 2018.

[73] V. I. Oseledets, A multiplicative ergodic theorem. characteristic Ljapunov exponents of dynamical systems. *Tr. Mosk. Mat. Obs.* **19** (1968), 179–210.

[74] E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. A. Yorke, A local ensemble Kalman filter for atmospheric data assimilation. *Tellus, Ser. A Dyn. Meteorol. Oceanogr.* **56** (2004), no. 5, 415–428.

[75] Y. Pesin and V. Climenhaga, Open problems in the theory of non-uniform hyperbolicity. *Discrete Contin. Dyn. Syst.* **27** (2010), no. 2, 589–607.

[76] M. A. Pinsky and V. Wihstutz, Lyapunov exponents of nilpotent Itô systems. *Stochastics* **25** (1988), no. 1, 43–57.

[77] M. S. Raghunathan, A proof of Oseledec's multiplicative ergodic theorem. *Israel J. Math.* **32** (1979), no. 4, 356–362.

[78] M. Romito and L. Xu, Ergodicity of the 3D stochastic Navier–Stokes equations driven by mildly degenerate noise. *Stochastic Process. Appl.* **121** (2011), no. 4, 673–700.

[79] G. Royer, Croissance exponentielle de produits Markoviens de matrices aléatoires. *Ann. Inst. Henri Poincaré Probab. Stat.* **16** (1980), 49–62.

[80] M. Shamis and T. Spencer, Bounds on the lyapunov exponent via crude estimates on the density of states. *Comm. Math. Phys.* **338** (2015), no. 2, 705–720.

[81] H. Triebel, *Theory of function spaces II*. Birkhauser, 1992.

[82] A. Virtser, On products of random matrices and operators. *Theory Probab. Appl.* **24** (1980), no. 2, 367–377.

[83] A. Wilkinson, What are Lyapunov exponents, and why are they interesting? *Bull. Amer. Math. Soc.* **54** (2017), no. 1, 79–105.

[84] M. Yamada and K. Ohkitani, Lyapunov spectrum of a chaotic model of three-dimensional turbulence. *J. Phys. Soc. Jpn.* **56** (1987), no. 12, 4210–4213.

[85] L.-S. Young, Mathematical theory of Lyapunov exponents. *J. Phys. A* **46** (2013), no. 25, 254001.

[86] G.-C. Yuan, K. Nam, T. M. Antonsen Jr., E. Ott, and P. N. Guzdar, Power spectrum of passive scalars in two dimensional chaotic flows. *Chaos* **10** (2000), no. 1, 39–49.

JACOB BEDROSSIAN

Department of Mathematics, University of Maryland, College Park, MD 20742, USA, jacob@math.umd.edu

ALEX BLUMENTHAL

School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA, ablumenthal6@gatech.edu

SAM PUNSHON-SMITH

School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, USA, spsmith@math.ias.edu

Department of Mathematics, Tulane University, New Orleans LA 70118, USA, punshs@brown.edu

MULTISCALE ECO-EVOLUTIONARY MODELS: FROM INDIVIDUALS TO POPULATIONS

NICOLAS CHAMPAGNAT, SYLVIE MÉLÉARD, AND VIET CHI TRAN

ABSTRACT

Motivated by recent biological experiments, we emphasize the effects of small and random populations in various biological/medical contexts related to evolution such as invasion of mutant cells or emergence of antibiotic resistances. Our main mathematical challenge is to quantify such effects on macroscopic approximations. The individual behaviors are described by the mean of stochastic multiscale models. The latter, in the limit of large population and according to the assumptions on mutation size and frequency, converge to different macroscopic equations. Sufficiently rare mutations yield a timescale separation between competition and mutation. In that case, the stochastic measure-valued process at the mutation timescale converges to a jump process which describes the successive invasions of successful mutants. The gene transfer can drastically affect the evolutionary outcomes. For faster mutation timescales, numerical simulations indicate that these models exhibit as cyclic behaviors. Mathematically, population sizes and times are considered on a log-scale to keep track of small subpopulations that have negligible sizes compared with the size of the resident population. Explicit criteria on the model parameters are given to characterize the possible evolutionary outcomes. The impact of these time and size scales on macroscopic approximations is also investigated, leading to a new class of Hamilton–Jacobi equations with state constraint boundary conditions.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 92D25; Secondary 92D15, 60K35, 60F99

KEYWORDS

Multiscale models, stochastic individual-based models, measure-valued Markov processes, long time behaviour, large population approximations, branching processes, partial differential equations

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5656–5678
DOI 10.4171/ICM2022/24

1. INTRODUCTION AND PRESENTATION OF THE INDIVIDUAL-BASED MODEL

Since Darwin's revolutionary work on evolution and natural selection [18], many mathematicians have worked on modeling his theories. Different schools of thought have developed, involving different classes of mathematical objects. Ecological models of structured population dynamics usually rely on deterministic models in large populations, such as dynamical systems (as the famous Lotka–Volterra system) and partial differential equations. Population genetics are more interested in random fluctuations of gene frequencies in small populations (like in the Wright–Fisher model) and therefore make extensive use of probabilistic tools. A few decades ago, eco-evolutionary models emerged, seeking to link these two approaches. Our work is placed in this framework. Our point of view consists in focusing on stochastic individual behaviors, taking into account demographic parameters (birth and death rates), evolutionary parameters (mutations, gene transfer), and ecological parameters (interactions between individuals), all these parameters depending on the genetic or phenotypic characteristics of the individual. This point of view is strongly reinforced by the ability of biologists to obtain more and more individual data, for example, for bacteria, thanks to single cell microscopes or microfluidic techniques. The notion of individual variability took a long time to emerge, especially for the biology of microorganisms, and it was not until the 2000s that biologists began to take it into account [24, 39].

There are three main sources of randomness in eco-evolutionary mechanisms which happen at different time and size scales: at the molecular level (errors in DNA replication or genetic information exchanges), at the individual level (division time, life span, contacts, access to resources), and at a macroscopic level (environmental variations). Mathematically, it is very exciting that all the parameters we have mentioned have their own scales, which can be different according to the species considered and also can vary according to the environment. Depending on these scales, the mathematical models and the associated mathematical questions can be of different nature and challenging, and open new fields of investigation.

We consider bacteria or cell populations. The ability of an individual to survive or divide depends on phenotypic or genetic parameters whose quantitative expression (real or vectorial) is called a trait. The evolution of the trait distribution results from different main mechanisms. The heredity is the vertical transmission of the ancestral trait to offspring, except when a mutation occurs. Mutations generate trait variability in the population. The selection process takes place at two levels. The variability in traits allows an individual with a higher probability of survival or a better ability to reproduce to create a subpopulation of offspring that will invade the population (genetic selection). In addition, selection also favors those individuals best able to survive in competition with others (ecological selection). Although their reproduction is asexual, bacteria or cells can also horizontally exchange genetic information during their life. Horizontal gene transfer is obtained by direct contact between cells, either by the transfer of small parts of chromosomal DNA or by the transfer of plasmids, small circular double-stranded DNA structures which can be very costly

for the cell in terms of energy used. Gene transfer plays an essential role in the evolution, maintenance, transmission of virulence and antibiotic resistance.

Our goal in this paper is to show the richness of models, mathematical questions and theorems that can emerge from these eco-evolutionary dynamics and from the understanding of their long-term evolution. One is faced with the fundamental question: how to describe and quantify the successive invasions of favorable mutants? All our constructions will be based on the stochastic behavior of the individuals from which we will derive different macroscopic approximations depending on the parameter assumptions.

The seminal papers concerning eco-evolutionary modeling are based on game theory and dynamical systems, see Hofbauer–Sigmund [28], Marrow–Law–Cannings [33], Metz et al. [35, 36]. Then more general models for structured populations have been introduced based either on partial differential equations, see, for example, the founding papers of Diekmann [21], Diekmann–Jabin–Mischler–Perthame [22], Barles–Mirrahimi–Perthame [3], Desvillettes–Jabin–Mischler–Raoul [19], or on stochastic individual-based models as in the theoretical biological papers by Dieckmann–Law [20], Bolker–Paccala [9], or in the rigorous mathematical papers by Fournier–Méléard [25], Champagnat–Ferrière–Méléard [13], Champagnat [11], Champagnat–Méléard [15]. Models including horizontal transfer have been proposed in the literature based on the seminal contribution of Anderson and May on host-pathogen deterministic population dynamics [1] (see also Levin et al. [30, 40]) or on a population genetics framework without ecological concern (see [4, 38, 41]).

The basis of our approach is a stochastic individual-based model: it is a pure jump point measure-valued process in continuous time, weighted by the carrying capacity K of the system (order of magnitude of the population size), whose jump events are births with or without mutation, transfers, and deaths. The jump rates depend on the trait value of each individual, on the total population and for some of them on K. From this basic process, one can derive different approximations following the main biological assumptions of the adaptive biology. The population size is assumed to be large ($K \to \infty$), but we will also need to keep track of small populations. Mutations are rare (p_K tends to 0), but not necessarily from the population standpoint, depending on whether Kp_K tends to 0 or not. Mutation steps in the trait space may be considered small or not. The population process will be considered on different time scales: of order 1, of order $\frac{1}{Kp_K}$, or of order $\log K$.

After introducing in Section 2 the individual-based model scaled by the carrying capacity K, we will study in Section 3 large population limits on finite time intervals when K tends to infinity, using ideas developed in [25]. The stochastic process is shown to converge to the unique solution of a nonlinear integro-differential equation (see also Billiard et al. [5, 6] for models with horizontal transfer). In the case where the trait support is composed of two values, the equation reduces to a nonstandard two-dimensional dynamical system whose long-time behavior is studied. In Section 4, we analyze the invasion probability and time to fixation of an initially rare mutant population. In this case, the stochastic behavior of the mutant population is fundamental and needs to be combined with the deterministic approximation of the resident population size. In Section 5 we assume that mutations are rare at the population scale to imply a separation between the competition and mutation time

scales, following ideas of [11,13,15]. Under an "invasion implies fixation" assumption, a pure jump (single support) measure-valued process is derived from the population process at the mutation time scale. When the mutation steps tend to 0, a limiting differential equation for the support dynamics is also derived in a longer time scale. These results are illustrated by simulations of a simple model in Section 6. Depending on the transfer rate, we obtain dramatically different behaviors, ranging from expected evolution toward the optimal trait, to extinction (evolutionary suicide). When the individual mutation rate is small, but not from the population standpoint, intermediary values of transfer rates lead to surprising cyclic behaviors related to reemergence of traits. To capture these phenomena, we consider in Section 7 the small populations of order K^{β_K} for $0 < \beta_K \leq 1$ that can be observed in the long time scale $\log K$. We study the asymptotic dynamics of the exponents $(\beta_K(t), t \geq 0)$ and analyze the first reemergence of the optimal trait. In Section 8, under the additional assumption that the individual mutations are small, we establish in a simple framework that the stochastic discrete exponent process converges to the viscosity solution of a Hamilton–Jacobi equation with state constraint boundary conditions, allowing us to fill the gap between the stochastic [11,15] and deterministic [3,22] approaches of Dirac concentration in adaptive dynamics. In the coming years, we hope to generalize this result in a much more general framework.

Notation. The set E being a Polish space, the Skorohod space $\mathbb{D}([0,T],E)$ is the functional space of right-continuous and left-limited functions from $[0,T]$ to E. It is endowed by the Skorohod topology (cf. Billingsley [7]) which makes it a Polish space.

2. A GENERAL STOCHASTIC INDIVIDUAL-BASED MODEL FOR VERTICAL AND HORIZONTAL TRAIT TRANSMISSION

2.1. The model

The population dynamics is described by a stochastic system of interacting individuals (cf. [12,13,25]). The individuals are characterized by a quantitative parameter x, called trait, belonging to a compact subset $\mathcal{X}$ of $\mathbb{R}^d$, which summarizes the phenotypic or genotypic information of each individual. The trait determines the demographic rates. It is inherited from parent to offspring, except when a mutation occurs, in which case the trait of the offspring takes a new value. It can also be transmitted by horizontal transfer from an individual to another one. The demographic and ecological rates are scaled by the *carrying capacity* K which is taken as a measure of the "system size" (resource limitation, living area, initial number of individuals). We will derive macroscopic behaviors for the population by letting K tend to infinity with the appropriate scaling $\frac{1}{K}$ for individuals' weight.

At each time t, the population state at time t is described by the point measure

$$\nu_t^K(dx) = \frac{1}{K}\sum_{i=1}^{N_t^K} \delta_{X_i(t)}(dx), \quad N_t^K = K\int \nu_t^K(dx),$$

where $X_i(t)$ is the trait of the ith individual living at t, individuals being ranked according to the lexicographic order of their trait values. Recall that notation δ_x means the Dirac mea-

sure at x. Later we will denote indifferently, for a measurable bounded function f on $\mathbb{R}^d$, $\langle \nu_t^K, f\rangle = \int_{\mathbb{R}^d} f(x)\nu_t^K(dx) = \sum_{i=1}^{N_t^K} f(X_i(t))/K$.

The right-continuous and left-limited measure-valued process $(\nu_t^K, t \geq 0)$ is a Markov process whose transitions are described as follows. An individual with trait x gives birth to a new individual with rate $b(x)$. With probability $1 - p_K$, the new individual carries the trait x and with probability p_K, there is a mutation on the trait. The trait z of the new individual is chosen according to the probability distribution $m(x, dz)$. An individual with trait x dies with intrinsic death rate $d(x)$ and from the competition with any other individual alive at the same time. If the competitor has the trait y, the competition death rate is $\frac{C(x,y)}{K}$, leading for a population $\nu = \frac{1}{K}\sum_{i=1}^n \delta_{x_i}$ to a total individual death rate $d(x) + \frac{1}{K}\sum_{i=1}^n C(x, x_i) = d(x) + C * \nu(x)$. Horizontal transfers can occur from individuals x to y, or vice versa. In a population ν, an individual with trait x chooses a partner with trait y at rate $\frac{1}{K}\frac{\tau(x,y)}{\langle \nu, 1\rangle}$. After transfer, (x, y) becomes (x, x).

2.2. Generator

We denote by $\mathcal{M}_K$ the set of point measures on $\mathcal{X}$ weighted by $1/K$ and by $\mathcal{M}_F$ the set of finite measures on $\mathcal{X}$. The generator of the process $(\nu_t^K)_{t\geq 0}$ is given for measurable bounded functions F on $\mathcal{M}_K$ and $\nu = \frac{1}{K}\sum_{i=1}^n \delta_{x_i}$ by

$$\begin{aligned}
&\sum_{i=1}^n b(x_i)\Big((1-p_K)\Big(F\Big(\nu + \frac{1}{K}\delta_{x_i}\Big) - F(\nu)\Big)\\
&\quad + p_K\int_{\mathcal{X}}\Big(F\Big(\nu + \frac{1}{K}\delta_z\Big) - F(\nu)\Big)m(x_i, dz)\Big)\\
&\quad + \sum_{i=1}^n \big(d(x_i) + C * \nu(x_i)\big)\Big(F\Big(\nu - \frac{1}{K}\delta_{x_i}\Big) - F(\nu)\Big)\\
&\quad + \sum_{i,j=1}^n \frac{\tau(x_i, x_j)}{K\langle \nu, 1\rangle}\Big(F\Big(\nu + \frac{1}{K}\delta_{x_i} - \frac{1}{K}\delta_{x_j}\Big) - F(\nu)\Big).
\end{aligned}$$

It is standard to construct the measure-valued process ν^K as the solution of a stochastic differential equation driven by Poisson point measures and to derive the following moment and martingale properties (see, for example, [25] or Bansaye–Méléard [2]).

Theorem 2.1. *Under the previous assumptions and assuming also that for some $p \geq 2$, $\mathbb{E}(\langle \nu_0^K, 1\rangle^p) < \infty$, the following properties hold. For a bounded measurable function f on $\mathcal{X}$,*

$$\begin{aligned}
\int f(x)\nu_t^K(dx) = \int f(x)\nu_0^K(dx) + M_t^{K,f}\\
+ \int_0^t\int_{\mathcal{X}}\Big\{\big((1-p_K)b(x) - d(x) - C * \nu_s^K(x)\big)f(x)\\
+ p_K b(x)\int_{\mathcal{X}} f(z)m(x, dz)\\
+ \int_{\mathcal{X}}\frac{\tau(x,y)}{\langle \nu_s^K, 1\rangle}\big(f(x) - f(y)\big)\nu_s^K(dy)\Big\}\nu_s^K(dx)ds,
\end{aligned}$$

where $M^{K,f}$ *is a right-continuous and left-limited square-integrable martingale starting from* 0 *with quadratic variation*

$$\langle M^{K,f}\rangle_t = \frac{1}{K}\int_0^t \int_{\mathcal{X}} \Big\{ \big((1-p_K)b(x) + d(x) + C * \nu_s^K(x)\big) f^2(x) + p_K b(x) \int_{\mathcal{X}} f^2(z) m(x,dz) + \int_{\mathcal{X}} \frac{\tau(x,y)}{\langle \nu_s^K, 1\rangle} \big(f(x)-f(y)\big)^2 \nu_s^K(dy) \Big\} \nu_s^K(dx)ds.$$

3. LARGE POPULATION LIMIT AND RARE MUTATION IN THE ECOLOGICAL TIME-SCALE

3.1. A deterministic approximation

Assuming that p_K converges to p when K tends to infinity, we derive a macroscopic approximation of the population process on any finite time interval.

Assumptions (H)**.** (i) When $K \to +\infty$, the stochastic initial point measures ν_0^K converge in probability (and for the weak topology) to the deterministic measure $u_0 \in \mathcal{M}_F(\mathcal{X})$ and $\sup_K \mathbb{E}(\langle \nu_0^K, 1\rangle^3) < +\infty$.

(ii) The functions b, d, C, and τ are continuous. The intrinsic growth rate of the subpopulation of trait x is denoted by $r(x) = b(x) - d(x)$. For any $x, y \in \mathcal{X}$, we also assume $r(x) > 0, C(x,y) > 0$. It means that, in absence of competition, the subpopulation with trait x has a tendency to grow and the regulation of the population size comes from the competition pressure.

Proposition 3.1. *Assume* (H) *and that* $p_K \to p$ *when* K *tends to infinity. Then, for* $T > 0$ *and when* $K \to \infty$*, the sequence* $(\nu^K)_{K\geq 1}$ *converges in probability in* $\mathbb{D}([0,T], \mathcal{M}_F(\mathcal{X}))$ *to the deterministic function* $u \in \mathcal{C}([0,T], \mathcal{M}_F(\mathcal{X}))$*, the unique weak measure-solution of*

$$\partial_t u(t,x) = \big(r(x) - C * u(t,x)\big)u(t,x) + p\int_{\mathcal{X}} b(y)m(y,x)u(t,y)dy + \frac{u(t,x)}{\|u(t,\cdot)\|_1}\int_{\mathcal{X}} \alpha(x,y)u(t,y)dy, \tag{3.1}$$

with $C * u(t,x) = \int C(x,y)u(t,y)dy$ *and* $\alpha(x,y) = \tau(x,y) - \tau(y,x)$.

The proof is standard and consists of a tightness and uniqueness argument, see [2,25] or [6] for details. Let us note that the horizontal transfer acts on the dynamics (3.1) through the "horizontal flux" rate α which quantifies the asymmetry between transfers and can be positive as well as negative (or zero in the case of perfectly symmetrical transfer). Nevertheless, the fully stochastic population process depends not only on α but also on τ itself. Let us mention that, to the best of our knowledge, the long-time behavior of a solution of (3.1) is unknown, except in the case without transfer studied by Desvillettes et al. [19]. The existence of steady-states for some similar equations has been studied in Hinow et al. [27] and Magal–Raoul [32].

3.2. Particular cases when $p = 0$

Standard biological observations lead us to assume small individual mutation rate,

$$\lim_{K\to\infty} p_K = 0. \tag{3.2}$$

Under this assumption, the mutational term in (3.1) disappears, meaning that mutation events are too rare to be observed at the demographic/ecological timescale (of births, deaths, and interaction). In the particular case when the support of the initial measure u_0 is a single point x, i.e., $u_0 = n_x(0)\delta_x$, $n_x(0) \in \mathbb{R}_+$, the support of the measure u_t is $\{x\}$ for all $t > 0$ and $u_t = n_x(t)\delta_x$. From (3.1), we deduce that $n_x(t)$ is the solution of the logistic equation

$$n_x'(t) = n_x(t)\big(r(x) - C(x,x)n_x(t)\big).$$

This equation has a unique stable equilibrium

$$\bar{n}_x = \frac{r(x)}{C(x,x)}. \tag{3.3}$$

Similarly, in the case when the support of u_0 is composed of two points x and y, i.e., $u_0 = n_x(0)\delta_x + n_y(0)\delta_y$, $n_x(0), n_y(0) \in \mathbb{R}_+$, the support of the measure u_t is $\{x, y\}$ for all $t > 0$ and $u_t = n_x(t)\delta_x + n_y(t)\delta_y$, and $(n_x(t), n_y(t))$ is the solution of the dynamical system

$$\begin{aligned}
\frac{dn_x}{dt} &= \left(r(x) - C(x,x)n_x - C(x,y)n_y + \frac{\alpha(x,y)}{(n_x+n_y)}n_y\right)n_x,\\
\frac{dn_y}{dt} &= \left(r(y) - C(y,x)n_x - C(y,y)n_y - \frac{\alpha(x,y)}{(n_x+n_y)}n_x\right)n_y.
\end{aligned} \tag{3.4}$$

This system can be seen as a perturbation of a competitive Lotka–Volterra system, but presents more possible limit behaviors (but no cycles, see [5] for a detailed study). It is easy to see that trait y will invade a resident population of trait x and get fixed if and only if

$$r(y) - r(x) + \alpha(y,x) > 0. \tag{3.5}$$

In particular, the horizontal transfer can revert the outcome of the dynamical system without transfer, provided that $|\alpha(y,x)| > |r(y) - r(x)|$ and $\operatorname{sign}(\alpha(y,x)) = -\operatorname{sign}(r(y) - r(x))$, where $\operatorname{sign}(x) = 1$ if $x > 0$; 0 if $x = 0$; -1 if $x < 0$.

The situation is even simpler if the function C is constant. The system becomes

$$\begin{aligned}
\frac{dn}{dt} &= n\big(qr(x) + (1-q)r(y) - Cn\big),\\
\frac{dq}{dt} &= q(1-q)\big(r(y) - r(x) + \alpha(y,x)\big),
\end{aligned}$$

where $n = n_x + n_y$ and $q = n_x/(n_x + n_y)$. There are only two equilibria for the second equation, $q = 0$ and $q = 1$, corresponding to the equilibria $(\frac{r(x)}{C}, 1)$ and $(\frac{r(y)}{C}, 0)$, respectively. This illustrates an important assumption, called the 'invasion implies fixation" principle (IIF).

Assumption (IIF). Given any $x \in \mathcal{X}$ and Lebesgue almost any $y \in \mathcal{X}$, either $(\bar{n}_x, 0)$ is a stable steady state of (3.4), or $(\bar{n}_x, 0)$ and $(0, \bar{n}_y)$ are respectively unstable and stable steady states, and any solution of (3.4) with an initial state in $(\mathbb{R}_+^*)^2$ converges to $(0, \bar{n}_y)$ when $t \to \infty$.

Biologically speaking, this means that the ecological coefficients impede the coexistence of two traits (which is biologically accepted when there is only one type of resource, see [14]).

4. RARE MUTATIONS – FIXATION PROBABILITY

For this section, we refer to [11, 13, 15] for rigorous proofs.

Let us now assume (3.2) and that the resident population is uniquely composed of individuals with trait x and near its size equilibrium, i.e., when K is large, the population size $N^{x,K}$ is then close to the equilibrium $K\bar{n}^x$. Let us now investigate the fate of a newly mutated individual with trait y in this resident population, as observed in Figure 1. When the mutant appears, it begins to develop (by heredity) a small population with trait y whose size is initially negligible. During this first phase, the number $N^{y,K}$ of individuals with trait y is very small with respect to $N^{x,K}$. Its dynamics can be approximated by a linear birth and death stochastic process, at least until it reaches the threshold ηK, for a given small $\eta > 0$. The transfer $x \to y$ acts as a birth term and the transfer $y \to x$ as a death term. Therefore, the growth rate of an individual with trait y for this first phase is approximately given by

$$S(y;x) = r(y) - C(y,x)\bar{n}_x + \alpha(y,x) = r(y) - C(y,x)\frac{r(x)}{C(x,x)} + \alpha(y,x). \qquad (4.1)$$

The quantity $S(y;x)$ is called invasion fitness of trait y in the resident population of trait x. Note that S is not symmetric and null on the diagonal; for C constant, it is given by (3.5). When K tends to infinity, the probability for the process $N^{y,K}$ to reach ηK (for some $\eta > 0$) is approximately the survival probability of the underlying linear birth and death process, i.e., the positive part of the growth rate $S(y;x)$ divided by the birth rate $b(y) + \tau(y,x)$,

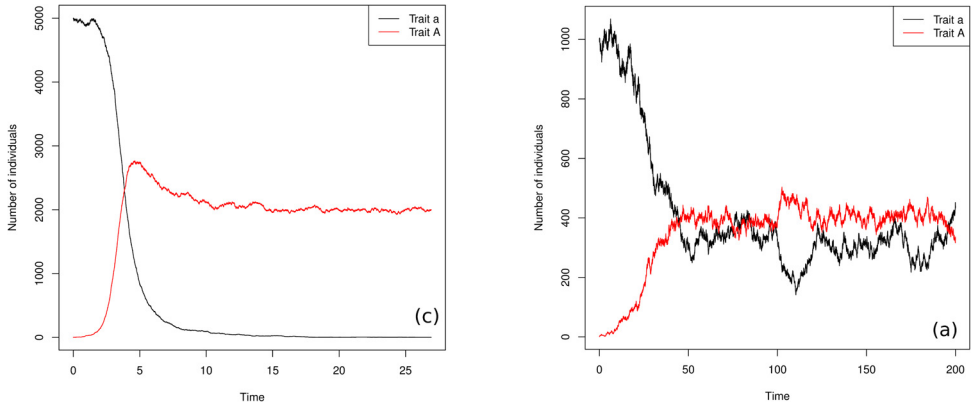


FIGURE 1

Invasion and fixation or polymorphic persistence of a deleterious mutation for unilateral transfer rate: (left) $C \equiv 1$, $b(y) = 0.5$, $b(x) = 1$, $d(x) = d(y) = 0$, $K = 5\,000$, $\alpha(y,x) = \tau(y,x) = 0.7$; (right) $C(y,x) = C(x,x) = 2$, $C(y,y) = 4$, $C(x,y) = 1$, $b(y) = 0.8$, $b(x) = 1$, $d(x) = d(y) = 0$, $K = 1\,000$, $\alpha(y,x) = \tau(y,x) = 0.5$.

namely

$$P(y;x) = \frac{[r(y) - C(y,x)\bar{n}^x + \alpha(y,x)]_+}{b(y) + \tau(y,x)}. \tag{4.2}$$

In particular, invasion is impossible if $S(y;x) \leq 0$.

Let us assume that $S(y;x) > 0$. Then, the duration of the first phase (growth of the y-population from 1 to ηK individuals) is of order $\log K/S(y;x)$. It can be proved rigorously but, to be convinced of this, it is enough to notice that if t is the time elapsed from the appearance of the single mutant individual with trait y to threshold ηK, then $\mathbb{E}(N_t^{y,K}) \approx e^{S(y;x)t} = \eta K$, and $t = \log K/S(y;x)$. Then the second phase begins, where the processes $(N^{x,K}, N^{y,K})$ stay close to the dynamical system (3.4) with nonnegligible initial data η. Under Assumption (IIF), the trait y invades the population and the x-population size decreases to $N_t^{x,K} < \eta K$ in a duration of order of magnitude 1. Should the latter happen, the third phase begins and $N^{x,K}$ can be approximated by a subcritical linear birth and death process, until y is fixed and x is lost. In this case, the transfer $y \to x$ acts as a birth term and the transfer $x \to y$ as a death term. The duration of this third phase behaves as $\log K/(d-b)$ when $K \to \infty$ (see [34, SECTION 5.5.3, P. 190] for precise computation) where $b = b(x) + \tau(x,y)$, $d = d(x) + \frac{C(x,y)r(y)}{C(y,y)} + \tau(y,x)$. Summing up, the fixation time of an initially rare trait y going to fixation is of order

$$T_{\text{fix}} = \log K \left(\frac{1}{S(y;x)} + \frac{1}{|S(x;y)|} \right) + o(\log K). \tag{4.3}$$

5. VERY RARE MUTATIONS IN AN EVOLUTIONARY TIME SCALE

We wish to rigorously define and quantify the evolutionary process describing the successive invasions of successful mutants under hypothesis (3.2). In Section 3, mutations are not seen in the limit. *To observe the dynamical impact of mutations, we have to wait for a longer time than* $O(1)$. Depending on the rate of convergence of p_K to 0, different timescales will be considered in the next sections.

We assume here that not only $p_K \to 0$ but also $K\, p_K \to 0$, meaning that both individual and population mutation rates are small. We will consider the behavior of the population process at the very long time scale $\frac{1}{K\, p_K}$. Moreover, we will assume that

$$\forall\, V > 0, \quad \log K \ll \frac{1}{K\, p_K} \ll e^{VK}. \tag{5.1}$$

This assumption leads to a separation of time scales between competition phases and mutation arrivals. Indeed, by (4.3), mutations are rare enough so that the selection has time to eliminate deleterious traits or to fix advantageous traits before the arrival of a new mutant.

5.1. Trait substitution sequence

Let us study the convergence of the process $(\nu^K_{./(Kp_K)})_{K\geq 1}$ when K tends to infinity, under the assumption (5.1). By simplicity we assume the *invasion implies fixation* (IIF) principle. This implies that, for a monomorphic ancestral population, the dynamics at the

time scale $t/(Kp_K)$ can be approximated by a jump process over singleton measures on $\mathcal{X}$ whose mass at any time is at equilibrium. More precisely, we have

Theorem 5.1. *Assume* (*H*), (5.1), *and* (IIF). *Suppose the initial conditions are* $\nu_0^K(dx) = N_0^K \delta_{x_0}(dx)$ *with* $x_0 \in \mathcal{X}$, $\lim_{K\to\infty} N_0^K = \bar{n}_{x_0}$ *in probability, and* $\sup_{K\in\mathbb{N}^*} \mathbb{E}((N_0^K)^3) < +\infty$.

Then, the sequence of processes $(\nu^K_{\cdot/(Kp_K)})_{K\geq 1}$ *converges in law (for finite-dimensional distributions) to the* $\mathcal{M}_F(\mathcal{X})$*-valued process* $(V_t(dx) = \bar{n}_{Y_t}\delta_{Y_t}(dx), t \geq 0)$ *where* $(Y_t)_{t\geq 0}$ *is a pure jump process on* $\mathcal{X}$*, started at* x_0*, with the jump measure from* x *to* y *being*

$$b(x)\bar{n}_x P(y;x)m(x,dy) \tag{5.2}$$

and $P(y;x)$ *being defined in* (4.2).

The jump process $(Y_t, t \geq 0)$ (with $Y_0 = x_0$) describes the support of $(V_t, t \geq 0)$. It has been heuristically introduced in [35] and rigorously studied in [11], in the case without transfer. It is often called the trait substitution sequence (TSS). Theorem 5.1 can be generalized when the assumption (IIF) is not satisfied, see [15].

Main ideas for the proof of Theorem 5.1. The proof is a direct adaptation of [11]. The birth and death rates of the resident x and mutant y are

$$b(x) + \frac{\tau(x,y)N^{y,K}}{N^K}, \quad d(x) + C(x,x)N^{x,K} + C(x,y)N^{y,K} + \frac{\tau(y,x)N^{y,K}}{N^K},$$
$$b(y) + \frac{\tau(y,x)N^{x,K}}{N^K}, \quad d(y) + C(y,x)N^{x,K} + C(y,y)N^{y,K} + \frac{\tau(x,y)N^{x,K}}{N^K}.$$

The proof consists in combining (5.1), the results in Section 4, and the Markov property. Let us fix $\eta > 0$. At $t = 0$, the population is monomorphic with trait x_0 and satisfies the assumptions of Theorem 5.1. As long as no mutation occurs, the population stays monomorphic with trait x_0 and, for t and K large enough, the density process $\langle \nu_t^K, \mathbf{1}_{x_0}\rangle$ belongs to the η-neighborhood of $\bar{n}_{x_0}$ with large probability (cf. Proposition 3.1). From the large deviations principle (see Freidlin–Wentzell [26]), one deduces that the time taken by the density process in absence of mutations to leave the η-neighborhood of $\bar{n}_{x_0}$ is larger than $\exp(VK)$, for some $V > 0$, with high probability. Hence assumption (5.1) ensures that the approximation of the population process by $\bar{n}^{x_0}\delta_{x_0}$ stays valid until the first mutation occurrence.

The invasion dynamics of a mutant with trait y in the resident population has been studied in Section 4. If $S(y;x_0) > 0$, the process $N^{y,K}$ is supercritical, and therefore, for large K, the probability for the mutant population's density to attain η is close to the probability $P(y;x_0)$. After this threshold and thanks to Assumption (IIF), the density process $(\langle \nu^K_{\frac{t}{Kp_K}}, \mathbf{1}_{x_0}\rangle, \langle \nu^K_{\frac{t}{Kp_K}}, \mathbf{1}_y\rangle)$ will attain, when K tends to infinity, an η-neighborhood of the unique stable equilibrium $(0, \bar{n}_y)$ of (3.4) and will stabilize around this equilibrium. We have shown in Section 4 that the time elapsed between the occurrence of the mutant and the final stabilization is given by (4.3). Hence, if $\log K \ll \frac{1}{Kp_K}$, with a large probability this phase of competition–stabilization will be complete before the occurrence of the next mutation. Using Markovian arguments, we reiterate the reasoning after each mutation event. Therefore, the population process on the time-scale t/Kp_K only keeps in the limit the successive stationary

states corresponding to successive advantageous mutations. If the process belongs to an η-neighborhood of $\bar{n}_x$, the mutation rate from an individual with trait x is close to $Kp_K b(x)\bar{n}_x$. At the time scale $\frac{t}{Kp_K}$, it becomes $b(x)\bar{n}_x$. The limiting process is a pure jump process $(V_t, t \geq 0)$ whose jump measure from a state $\bar{n}_x\delta_x$ is $b(x)\bar{n}_x P(y;x)m(x,dy)$. ■

Example 5.2. Let us consider a simple model with trait $x \in [0,4]$, C being constant, and $b(x) = 4 - x$, $d \equiv 1$, $\tau(x,y) = \tau e^{x-y}$. Then $S(x+h;x) = -h + \tau(e^h - e^{-h})$ and, for $\tau > 1/2$, it is positive if and only if $h > 0$. Thus the evolution with transfer is directed towards larger and larger traits, decreasing the growth rate until possible extinction. For τ small enough, $S(x+h;x) < 0$ for $h > 0$ so that a mutant of trait $x+h$ with $h > 0$ would disappear at the TSS scale. In this case, evolution drives the population to smaller and smaller traits until trait 0. The evolution for intermediary τ's is an open challenging question.

5.2. Canonical equation of the adaptive dynamics

Let us now assume that the mutation effects are very small: the mutation distribution m_σ depends on a parameter $\sigma > 0$ as follows:

$$\int g(z)m_\sigma(x,dz) = \int g(x+\sigma h)m_1(x,dh),$$

where m_1 is a reference symmetric measure with finite variance. Then the generator of the TSS Y^σ (which now depends on the parameter σ) is given by

$$L^\sigma g(x) = \int \big(g(x+\sigma h) - g(x)\big)b(x)\bar{n}_x \frac{[S(x+\sigma h;x)]_+}{b(x+\sigma h) + \tau(x+\sigma h,x)\bar{n}_x} m_1(x,dh).$$

For smooth S and since $S(x;x) = 0$, we have when σ tends to 0,

$$L^\sigma g(x) \sim \sigma^2 \frac{1}{2} g'(x)\bar{n}_x \partial_1 S(x;x) \int h^2 m_1(x,dh).$$

Let us observe that $\sigma \to 0$ makes the dynamics stop at this time scale. To observe a nontrivial behavior, we have to wait a longer time of order of magnitude $1/\sigma^2$.

Standard tightness and identification arguments allow showing the convergence in probability in $\mathbb{D}([0,T],\mathcal{X})$ of the process $(Y^\sigma_{t/\sigma^2}, t \in [0,T])$ to the deterministic function $(x(t), t \in [0,T])$, solving the equation

$$x'(t) = \frac{1}{2}\bar{n}_{x(t)}\partial_1 S\big(x(t);x(t)\big) \int h^2 m_1\big(x(t),dh\big), \tag{5.3}$$

the so-called *canonical equation of adaptive dynamics* introduced in [20] (cf. [15] for a rigorous proof). Note also that there is another candidate for the canonical equation obtained from partial differential equation arguments related to Hamilton–Jacobi equations [22,31,37].

Let us come back to Example 5.2 introduced previously. We assume that $m_1(x,dh)$ is a symmetric measure keeping the trait in $[0,4]$, i.e., with support in $[-x, 4-x]$. In this case, $\bar{n}_x = \frac{3-x}{C}$ and the canonical equation is given by

$$x'(t) = \frac{3-x(t)}{C}(2\tau - 1) \int h^2 m_1\big(x(t),dh\big),$$

since $r'(x) = -1$ and $\partial_1\tau(x,x) = -\partial_2\tau(x,x) = \tau$. Then for $\tau > 1/2$, the trait support is an increasing function, the population size $\bar{n}_{x(t)}$ is decreasing to 0, and therefore evolution

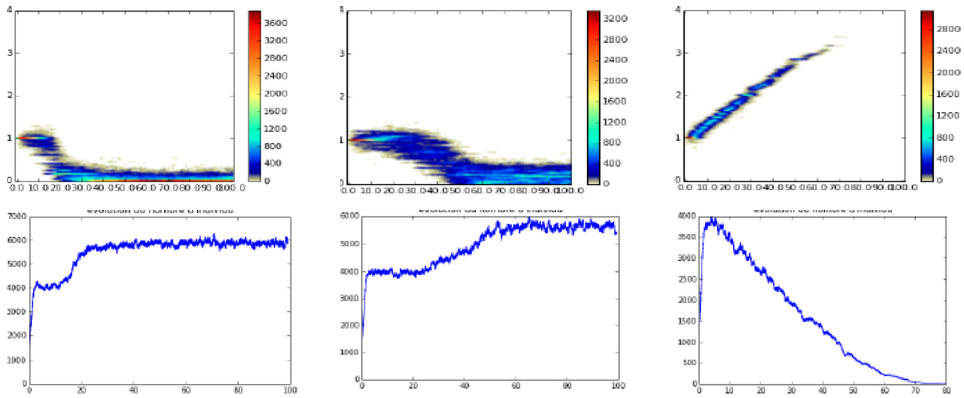
drives the population to an evolutionary suicide. Conversely, for $\tau < 1/2$, evolution leads to the optimal null trait (which maximizes the growth rate).

6. SIMULATIONS — CASE OF FREQUENCY-DEPENDENCE

(Simulations due to the Master students Lucie Desfontaines and Stéphane Krystal)

We focus on the special case of unilateral transfer, that is, $\tau(x, y) = \tau \mathbf{1}_{x>y}$, which is relevant for plasmids transfer. The next simulations are concerned with Example 5.2, with $C \equiv 0.5$, $p = 0.03$, and $m_\sigma(x,h)dh = \mathcal{N}(0,\sigma^2)$, conditioned on $x + h \in [0,4]$, with $\sigma = 0.1$. The initial state is composed of $K = 1\,000$ individuals with trait 1. Thus the corresponding population size at equilibrium is $1\,000 \times \frac{b(1)-d(1)}{C} = 4\,000$ individuals.

The constant τ will be the varying parameter. Figure 2(a) shows the evolution dynamics when $\tau = 0$. The evolution drives the population to its optimal trait 0 corresponding to a size at equilibrium equal to $1\,000 \times \frac{b(0)-d(0)}{C} = 6\,000$ individuals. The case $\tau = 0.2$ in Figure 2(b) shows a scenario similar to the case $\tau = 0$, although the evolution to optimal trait 0 takes a longer time. Conversely, when $\tau = 1$ (Figure 2(c)), the transfer drives the traits to larger and larger values, corresponding to lower and lower population sizes until extinction (evolutionary suicide). These simulations correspond to the theoretical study of the previous section. Let us now consider the intermediary value $\tau = 0.7$ (Figure 3). The evolution exhibits different patterns. In the first picture, high transfer converts at first individuals to larger traits and at the same time the population decreases. At some point, the population size is so small that the transfer does not play a role anymore leading to the brutal resurgence of a quasiinvisible strain, issued from a few individuals with small traits (and then with larger growth rate). We observe cyclic resurgences driving the mean trait towards



(a) (b) (c)

FIGURE 2

(a) $\tau = 0$; (b) $\tau = 0.2$ – almost no modification; (c) $\tau = 1$ – evolutionary suicide. Time in abscissa. First line, trait evolution; second line, size evolution.

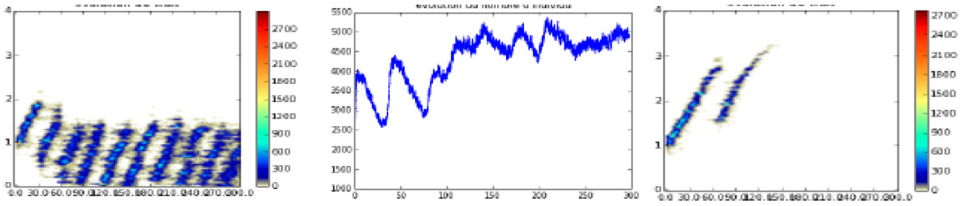


FIGURE 3

$\tau = 0.7$ – stepwise evolution with the trait evolution (left), and population size (center). Another pattern with extinction (right).

the optimal trait 0. In the last picture, we observe extinction of the population: the remaining individuals with smaller traits allow for a single resurgence of a new strain, but the traits of the individuals alive are too large to allow for survival.

7. STOCHASTIC ANALYSIS OF EMERGENCE OF EVOLUTIONARY CYCLIC BEHAVIOR – A SIMPLE MODEL

From now on, we are interested in the mathematical understanding of the previous simulations. In the latter, the chosen mutation probability p was small, but not the population mutation rate Kp, so (5.1) was not satisfied. We have to consider different time and size scales than the previous ones to capture the surprising resurgence behaviors. This part is largely inspired from Champagnat–Méléard–Tran [17].

7.1. A trait-discretized model

From now on, we consider a model inspired by Example 5.2 with a discrete trait space of mesh $\delta > 0$: $\mathcal{X} = [0, 4] \cap \delta\mathbb{N} = \{0, \delta, \dots, L\delta\}$ where $L = \lfloor 4/\delta \rfloor$. We choose $b(x) = 4 - x$, $\tau(x, y) = \tau \mathbf{1}_{x>y}$, $d(\cdot) \equiv 1$ and $C(\cdot, \cdot) \equiv C$. Therefore, $\bar{n}_x = \frac{3-x}{C}$ and the invasion fitness of a mutant individual of trait y in the population of resident trait x and size $K\bar{n}_x$ is

$$S(y; x) = x - y + \tau \mathbf{1}_{x<y} - \tau \mathbf{1}_{x>y} = x - y + \tau \operatorname{sign}(y - x). \tag{7.1}$$

We also define the fitness of an individual of trait y in a negligible population (of size $o(K)$) with dominant trait x to be

$$\hat{S}(y; x) = 3 - y + \tau \operatorname{sign}(y - x). \tag{7.2}$$

Indeed, the competition part is negligible in that case and vanishes at the limit when $K \to \infty$.

We assume that

$$p_K = K^{-\alpha} \quad \text{with } \alpha \in (0, 1), \tag{7.3}$$

and when a mutation occurs from an individual with trait $\ell\delta$, the new offspring carries the mutant trait $(\ell + 1)\delta$ (the mutations are directed to the right). The total mutation rate in a

population with size of order K is thus equal to $K^{1-\alpha}$ and then goes to infinity with K. We are very far from the situation described in [6, 11, 15] where (5.1) was satisfied. Here, small populations of size order K^β, $\beta < 1$ can have a nonnegligible contribution to evolution by mutational events, and we need to take into account all subpopulations with size of order K^β.

The population is described by the vector $(N_0^K(t), \dots, N_\ell^K(t), \dots, N_L^K(t))$, where $N_\ell^K(t)$ is the number of individuals of trait $x = \ell\delta$ at time t. The total population size N_t^K is now $N_t^K = \sum_{\ell=0}^L N_\ell^K(t)$. Our study of the (evolutionary) long-time dynamics of the process is based on a fine analysis of the size order, as power of K, of each subpopulation. These powers of K evolve on the timescale $\log K$, as can be easily seen in the case of branching processes (see Lemma 7.1). We thus define $\beta_\ell^K(t)$ for $0 \le \ell \le L$ such that

$$N_\ell^K(t \log K) = K^{\beta_\ell^K(t)} - 1, \quad \text{i.e., } \beta_\ell^K(t) = \frac{\log(1 + N_\ell^K(t \log K))}{\log K}. \tag{7.4}$$

We assume that $N^K(0) = (\lfloor \frac{3K}{C} \rfloor, \lfloor K^{1-\alpha} \rfloor, \dots, \lfloor K^{1-\ell\alpha} \rfloor, \dots, \lfloor K^{1-\lfloor 1/\alpha \rfloor \alpha} \rfloor, 0, \dots, 0)$. Then trait $x = 0$ is initially resident, with density $3/C$. With this initial condition, we have

$$\beta_\ell^K(0) \xrightarrow[K \to +\infty]{} (1 - \ell\alpha)\mathbf{1}_{0 \le \ell < \frac{1}{\alpha}}. \tag{7.5}$$

The main result of this section will give the asymptotic dynamics of $\beta^K(t) = (\beta_0^K(t), \dots, \beta_L^K(t))$ for $t \ge 0$ when $K \to +\infty$. We show that the limit is a piecewise affine continuous function, which can be described along successive phases determined by their resident or dominant traits. When the latter trait changes, the fitnesses governing the slopes are modified. Moreover, inside each phase, other changes of slopes are possible due to a delicate balance between mutations, transfer, and growth of subpopulations. We will deduce from the asymptotic dynamics of $\beta^K(t)$ explicit criteria for some of the evolutionary outcomes observed in Section 6 (Theorem 7.5).

Such an approach based on the behavior of the exponents β_K at the time scale $\log K$ has also been used in Durrett–Mayberry [23] for constant population size or pure birth process, with directional mutations and increasing fitness parameter, in Bovier et al. [10] for a density-dependent model where the evolution crosses the fitness valley constituted of unfit traits, in Blath et al. [8] for models with dormancy. In a deterministic setting with similar scales, we also refer to Kraut–Bovier [29]. In our case, the dynamics is far more complex due to the trade-off between larger birth rates for small trait values and transfer to higher traits, leading to diverse evolutionary outcomes. As a consequence, we need to consider cases where the dynamics of a given trait is completely driven by immigrations (see Lemma 7.2). This complexifies a lot the analysis.

7.2. Some enlightening lemmas

Before stating the main result (Theorem 7.3) which can be difficult to read and understand, we state two lemmas whose proof can be found in the Appendix of [17]. These lemmas are interesting by themselves.

(i) Assume first that a mutant with trait y appears in a resident population with trait x such that $y < x$. Then the dynamics of the initial (small) y-subpopulation size behaves

as a linear birth and death process with birth rate approximated by $4 - y$ and death rate by $1 + \frac{CN^{x,K}(t)}{K} + \tau$. We are thus led to study the following process.

Lemma 7.1. *Let us consider a linear birth and death process $(Z_t^K, t \geq 0)$, i.e., a binary branching process, with individual birth rate $b \geq 0$, individual death rate $d \geq 0$, and initial value $Z_0^K = K^\beta$ with $\beta > 0$.*

The process $(\frac{\log(1+Z^K_{s\log K})}{\log K}, s \in [0,T])$ converges in probability in $L^\infty([0,T])$ for all $T > 0$ to $((\beta + rs) \vee 0, s \in [0,T])$ when K tends to infinity, with $r = b - d$.

In addition, if $b < d$, for all $s > \beta/r$, then $\lim_{K\to+\infty} \mathbb{P}(Z^K_{s\log K} = 0) = 1$.

The limit can be understood from $\mathbb{E}(Z^K(t)) = K^\beta e^{rt}$. The proof of Lemma 7.1 uses the martingale property of $(e^{-rt} Z_t^K)_{t\geq 0}$. The proof is easy for $r \geq 0$ and more technical in the case $r < 0$, necessitating to control the extinction events after a certain time.

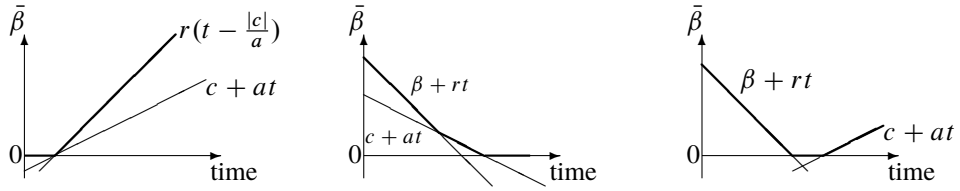
(ii) Assume now that a mutant with trait $y = x + \delta$ appears in a resident population with trait x. Then the dynamics of the initial (small) y-subpopulation size behaves as a linear birth and death process with birth rate approximated by $4 - y + \tau$ and death rate by $1 + \frac{CN^{x,K}(t)}{K}$. But in addition, trait y may receive a contribution from x at time t due to mutations at total rate $N^{x,K}(t)K^{-\alpha}$. By Lemma 7.1, we know that $N^{x,K}(s \log K) \approx K^{c+as}$ for constant $a, c \in \mathbb{R}$. This justifies the following lemma.

Lemma 7.2. *Let us consider a linear birth and death process with immigration $(Z_t^K, t \geq 0)$, with individual birth rate $b \geq 0$, individual death rate $d \geq 0$, initial value $Z_0^K = K^\beta$ with $\beta > 0$, and immigration rate at time t given by $K^c e^{at}$, with $a, c \in \mathbb{R}$.*

The process $(\frac{\log(1+Z^K_{s\log K})}{\log K}, s \in [0,T])$ converges when K tends to infinity in probability in $L^\infty([0,T])$ for all $T > 0$ to a continuous deterministic function $\bar\beta(s)$.

For $c \leq \beta$ and $\beta > 0$, $\bar\beta(s) = (\beta + rs) \vee (c + as) \vee 0$. For $\beta = 0$, $c < 0$ and $a > 0$, $\bar\beta(s) = ((r \vee a)(s - |c|/a)) \vee 0$. For $\beta = 0$, $c < 0$, and $a \leq 0$, $\bar\beta(s) = 0$. The other cases are immediate (see [17]).

This convergence is illustrated in Figure 4.



(a) $c < \beta = 0, 0 < a < r$ (b) $0 < c < \beta, r < a < 0$ (c) $c < 0 < \beta, r < 0 < a$

FIGURE 4

(a) Initially, $\bar\beta = 0$, but thanks to immigration, the population is revived. Once this happens, the growth rate r being larger than a, immigration has a negligible effect after time $|c|/a$. (b) After time $(\beta - c)/(a - r)$, the dynamics is driven by mutation before getting extinct. (c) We observe a local extinction before the population is revived thanks to incoming mutations.

7.3. Dynamics of the exponents

Let us come back to the asymptotic dynamics of $\beta^K(t) = (\beta_0^K(t), \dots, \beta_L^K(t))$ for $t \geq 0$ when $K \to +\infty$, which are characterized in the next result by a succession of deterministic time intervals $[s_{k-1}, s_k], k \geq 1$, called phases and delimited by changes of resident or dominant traits. The latter are unique except at times s_k and are denoted by $\ell_k^* \delta, k \geq 1$. This asymptotic result holds until a time T_0, which guarantees that there is ambiguity neither on these traits nor on the extinct subpopulations at the phase transitions. We will not give the exact (and technical) definition of T_0 and refer to [17].

Theorem 7.3. *Assume* (7.3) *with* $\alpha \in (0,1)$, $\delta \in (0,4)$, *and* (7.5).

(i) *For* $0 < T \leq T_0$, *the sequence* $(\beta^K(t), t \in [0,T])$ *converges in probability in* $\mathbb{D}([0,T],[0,1]^{L+1})$ *to a deterministic piecewise affine continuous function* $(\beta(t) = (\beta_0(t), \dots, \beta_L(t)), t \in [0,T])$, *such that* $\beta_\ell(0) = (1 - \ell\alpha)\mathbf{1}_{0 \leq \ell < \frac{1}{\alpha}}$. *The functions* β *are parameterized by* α, δ, *and* τ *defined as follows.*

(ii) *There exist an increasing nonnegative sequence* $(s_k)_{k \geq 0}$ *and a sequence* $(\ell_k^*)_{k \geq 1}$ *in* $\{0, \dots, L\}$ *defined inductively:* $s_0 = 0$, $\ell_1^* = 0$, *and, for all* $k \geq 1$, *assuming that* ℓ_k^* *have been constructed, we can construct* $s_k > s_{k-1}$ *as follows:*

$$s_k = \inf\{t > s_{k-1} : \exists \ell \neq \ell_k^*,\ \beta_\ell(t) = \beta_{\ell_k^*}(t)\}. \tag{7.6}$$

If $\beta_{\ell_k^*}(s_k) > 0$, *we set*

$$\ell_{k+1}^* = \arg\max_{\ell \neq \ell_k^*} \beta_\ell(s_k), \tag{7.7}$$

if the argmax is unique. In the other cases, we stop the induction.

(iii) *The functions* β_ℓ *are defined, for all* $t \in [s_{k-1}, s_k]$ *and* $\ell \in \{0, \dots, L\}$, *by*

$$\beta_\ell(t) = \begin{cases} [\mathbb{1}_{\beta_0(s_{k-1})>0}(\beta_0(s_{k-1}) + \int_{s_{k-1}}^t \tilde{S}_{s,k}(0; \ell_k^*\delta)\, ds)] \vee 0, & \text{if } \ell = 0, \\ (\beta_\ell(s_{k-1}) + \int_{t_{\ell-1,k} \wedge t}^t \tilde{S}_{s,k}(\ell\delta; \ell_k^*\delta)\, ds) \\ \qquad \vee (\beta_{\ell-1}(t) - \alpha) \vee 0, & \text{otherwise,} \end{cases} \tag{7.8}$$

where, for all traits x, y, $\tilde{S}_{t,k}(y;x) = \mathbb{1}_{\beta_{\ell_k^*}(t)=1} S(y;x) + \mathbb{1}_{\beta_{\ell_k^*}(t)<1} \hat{S}(y;x)$ *and where*

$$t_{\ell-1,k} = \begin{cases} \inf\{t \geq s_{k-1},\ \beta_{\ell-1}(t) = \alpha\}, & \text{if } \beta_\ell(s_{k-1}) = 0, \\ s_{k-1}, & \text{otherwise.} \end{cases} \tag{7.9}$$

In addition, for all ℓ *and all* $a < b$ *such that the time interval* $[a,b]$ *is included in the interior of the zero-set of* β_ℓ, *the event* $\{N_\ell^K(t \log K) = 0, \forall t \in [a,b]\}$ *has a probability converging to* 1 *as* K *tends to infinity.*

Simulations are shown in Figure 5 for various parameter values.

Roughly speaking, slope changes of the exponents $((\beta_0(t), \dots, \beta_L(t)), t \in [0,T])$ can take place at the times when a new exponent reaches 1 and there is a change of the resident

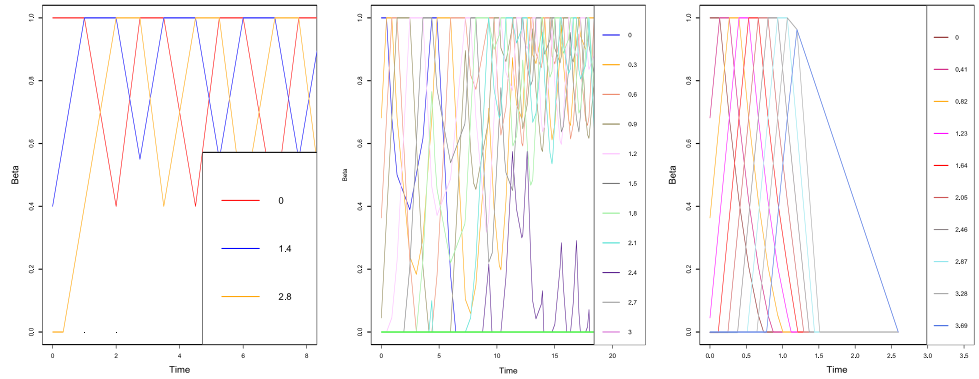


FIGURE 5

Exponents $\beta_\ell(t)$ as functions of time: (left) $\delta = 1.4$, $\alpha = 0.6$, $\tau = 2$. We see a periodic behavior showing reemergence of the fittest traits; (center) $\delta = 0.3$, $\alpha = 1/\pi$, $\tau = 1$. A cyclic but aperiodic behavior is observed; (right) $\delta = 0.41$, $\alpha = 1/\pi$, $\tau = 2.8$. The population is directly driven to evolutionary suicide.

trait, when a new exponent reaches 0 and there is extinction of the trait, and when the slope of an exponent formerly directed by its fitness becomes directed by incoming mutations.

Remark 7.4. (i) By the definition of s_k and ℓ^*_{k+1}, $\max_\ell \beta_\ell(t) = \beta_{\ell^*_k}(t)$ for $t \in [s_{k-1}, s_k)$.

(ii) The previous result keeps track of populations of size K^β for $0 < \beta \leq 1$, but not of populations of smaller order, which go fast to extinction on the time scale $\log K$.

Main ideas of the proof. We need to consider in the sequel two different situations: either there is a single trait x with population size of order K, called *resident* trait, or the total population size is $o(K)$. We explain the proof for simplicity assuming that there is always a resident trait. Theorem 7.3 is obtained by a fine comparison of the size of each subpopulation defined by a given trait value with carefully chosen branching processes with immigration. The stochastic dynamics consists in a succession of steps, composed of long phases $[\sigma_k^K \log K, \theta_k^K \log K]$ for $k \geq 1$ (with $\sigma_1^K = 0$) followed by short intermediate phases $[\theta_k^K \log K, \sigma_{k+1}^K \log K]$, where the stopping time θ_k^K is defined as the first time when the resident population size exits a neighborhood of its equilibrium density, or when the other subpopulations stop to be negligible with respect to the resident population. In each long phase, there is a single resident trait. Short intermediate phases correspond to the replacement of the resident trait, where two subpopulations are of maximal order. We prove that θ_k^K converges in probability to s_k, $k \geq 1$. In the limit, intermediate steps vanish on the time scale $\log K$. The proof proceeds by induction on k until some step k_0 where one of the three following events occurs: the exponents of three traits become maximal simultaneously, extinction, or the exponent of some trait vanishes at the same time as a change of resident population.

We then stop the induction and set $T_0 = s_{k_0}$ in the first and third cases, or $T_0 = +\infty$ in the second case.

To control the exponents $\beta_\ell^K(t)$, we proceed by a double induction, first on the steps, and second, inside each step, on the traits $\ell\delta$, for $\ell = 0$ to $\ell = L$. The exponents are approximately piecewise affine. Changes of slopes may happen when a new trait emerges, when a trait dies or when the dynamics of a trait becomes driven by incoming mutations. We use Lemma 7.2. During intermediate phases, we use comparisons with dynamical systems, described in Section 3. ■

7.4. Reemergence of trait 0

Recall that we work with birth, death and transfer rates presented in Section 7.1. In Figure 5, we have exhibited different evolutionary dynamics (reemergence of a trait, cyclic behavior, local extinction, evolutionary suicide). By reemergence of a trait $\ell\delta$, we mean that $\beta_\ell(s) = 1$ on some nonempty time interval $[t_1, t_2]$, then $\beta_\ell(s) < 1$ on some nonempty interval (t_2, t_3), and then $\beta_\ell(s) = 1$ again on some nonempty interval $[t_3, t_4]$. We would like to predict the evolutionary outcome as a function of parameters α, δ, τ. There are so many situations that we are not able to fully characterize the outcomes (see [17] for a detailed study in the case of three traits). Therefore, we focus on the beginning of the dynamics until either global extinction or reemergence of one trait occurs.

The resurgence of trait 0 is a prerequisite for a cyclic dynamics as those observed in Figure 5. We assume here that $\delta < 4/3$ (so that the cardinal of $\mathcal{X}$ is $L + 1 \geq 4$) and only consider the case $\delta < \tau < 3$. Computing the fitness functions, one can observe that for the first phases, $s_k = \frac{k\alpha}{\tau - \delta}$, and the trait $k\delta$ is resident on $[s_k, s_{k+1})$ ($\beta_k(s) = 1$) and for all $s \in [s_k, s_{k+1})$,

$$\beta_0(s) = 1 - \frac{\alpha(k-1)}{\tau - \delta}\left(\tau - \frac{k}{2}\delta\right) - (\tau - k\delta)(s - s_k).$$

This formula stays valid until either $\beta_0(s) = 0$ (loss of 0), or $\beta_0(s) = 1$ for some $s > s_1$ (reemergence of 0), or when the population size becomes $o(K)$. The slope of the function $\beta_0(s)$ becomes positive at time $s_{\tilde{k}}$, where $\tilde{k} := \lceil \frac{\tau}{\delta} \rceil$. Hence its minimal value is equal to

$$m_0 = \beta_0(s_{\tilde{k}}) = 1 - \frac{\alpha(\tilde{k}-1)}{\tau - \delta}\left(\tau - \frac{\tilde{k}}{2}\delta\right). \tag{7.10}$$

If the latter is positive, β_0 reaches 1 again in phase $[s_{\bar{k}}, s_{\bar{k}+1})$, where $\bar{k} = \lfloor 2\frac{\tau}{\delta} \rfloor$, at time

$$\bar{s} := s_{\bar{k}} + \frac{\alpha(\bar{k}-1)}{\tau - \delta}\frac{\tau - \frac{\bar{k}}{2}\delta}{\bar{k}\delta - \tau} = s_{\lfloor 2\frac{\tau}{\delta} \rfloor} + \frac{\alpha(\lfloor 2\frac{\tau}{\delta} \rfloor - 1)}{\tau - \delta}\frac{\tau - \frac{\lfloor 2\frac{\tau}{\delta} \rfloor}{2}\delta}{\lfloor 2\frac{\tau}{\delta} \rfloor\delta - \tau}. \tag{7.11}$$

The previous calculations give the intuition for the following theorem (see the proof in [17]).

Theorem 7.5. *Assuming $\delta < \tau < 3$, $\delta < 4/3$ and, under the assumptions of Theorem 7.3,*

(a) *If $m_0 > 0$ and $\bar{k}\delta < 3$, then the first reemerging trait is 0 and the maximal exponent is always 1 until this reemergence time;*

(b) *If* $m_0 < 0$*, the trait* 0 *gets lost before its reemergence and there is global extinction of the population before the reemergence of any trait;*

(c) *If* $m_0 > 0$ *and* $\bar{k}\delta > 3$*, there is reemergence of some trait* $\ell\delta < 3$ *and, for some time* t *before the time of first reemergence,* $\max_{1\le\ell\le L} \beta_\ell(t) < 1$.

Biologically, case (b) corresponds to evolutionary suicide. In cases (a) and (c), very few individuals with small traits remain, which are able to reinitiate a population of size of order K (reemergence) after the resident trait becomes too large. In these cases, one can expect successive reemergences. However, we do not know if there exists a limit cycle for the dynamics. Case (c) means that the total population is $o(K)$ on some time interval, before reemergence occurs after populations with too large traits become small enough.

It seems very difficult to go further with probabilistic tools. Another approach could consist in obtaining a macroscopic approximation of the exponents β^K in a trait continuum in terms of Hamilton–Jacobi equations and then using the tools of analysis.

8. MACROSCOPIC HAMILTON–JACOBI APPROXIMATION OF THE EXPONENTS

This part is a collaboration in progress with S. Mirrahimi [16]. We will give the ideas of our ongoing results, in particular a partial result concerning the simple case of stochastic supercritical birth–death–mutation process without transfer and competition. We assume that trait x belongs to the continuum $[0, 1]$. Starting from a finite population, our goal is to recover, by a direct scaling, the Hamilton–Jacobi equation that has been introduced in [3,22]. For this, we consider a discretization of the trait space $[0, 1]$ with step $\delta_K \to 0$, scale the mutation steps by a factor $1/\log K$ (small mutation steps), and assume that the initial population sizes are of the order of K^{β_0} for an exponent β_0 that can depend on the trait. More precisely, the population is composed of individuals with traits belonging to the discrete space $\mathcal{X}_K := \{i\delta_K : i \in \{0, 1, \dots, [\frac{1}{\delta_K}]\}\}$. The number of individuals with trait $i\delta_K$ is described by the stochastic process $(N_i^K(t), t \ge 0)$. As in the previous sections, an individual with trait $x \in \mathcal{X}_K$ gives birth to a new individual with same trait x at rate $b(x)$, dies at rate $d(x)$, but we assume that, for all $y \in \mathcal{X}_K$, it gives birth to a mutant individual with trait y at rate

$$p(x)\delta_K \log K m\big(\log K(x - y)\big).$$

Assumption 8.1. (i) The functions b, d, and p are nonnegative C^1-functions defined on $[0, 1]$ such that, for all $x \in [0, 1]$, $b(x) > d(x)$.

(ii) The function m is nonnegative, continuous, defined on $\mathbb{R}$, satisfies $\int_{\mathbb{R}} m(y)\,dy = 1$. It has exponential moments of any order and behaves as the Gaussian kernel $m(h) = \frac{1}{\sqrt{2\pi}\sigma} e^{-h^2/2\sigma^2}$ at infinity.

(iii) There exists $a > 0$ such that, for all $K \in \mathbb{N}$ and all $i \in \{0, 1, \dots, [\frac{1}{\delta_K}]\}$, $N_i^K(0) \ge K^a$.

(iv) There exists $a_2 < a$ such that $K^{-a_2/4} << \delta_K << 1/\log(K)$. Then, for $h_K := \delta_K \log K$, we have $\lim_{K\to+\infty} h_K = 0$.

Note that points 1 and 3 of Assumption 8.1 impede the subpopulations to be extinct. Note also that, for all $x \in (0,1)$, the total mutation rate from an individual with trait $x_K = i_K\delta_K$ with $i_K = [x/\delta_K]$, converges as $K \to +\infty$ to

$$\lim_{K\to+\infty} p(x_K) \sum_{j=0}^{[\frac{1}{\delta_K}]} h_K m\big(h_K(i_k - j)\big) = p(x) \int_{\mathbb{R}} m(y)\,dy = p(x).$$

Defining the exponents $\beta_i^K(t)$ as in (7.4), we introduce their interpolations: for all $x \in [0,1]$ and $K \geq 1$, let i be such that $x \in [i\delta_K, (i+1)\delta_K)$ and define

$$\tilde\beta^K(t,x) = \beta_i^K(t)\left(1 - \frac{x}{\delta_K} + i\right) + \beta_{i+1}^K(t)\left(\frac{x}{\delta_K} - i\right).$$

The sequence of processes $(\tilde\beta^K)_{K\geq 1}$ belongs to $\mathbb{D}([0,T], \mathcal{C}([0,1],\mathbb{R}))$, where $\mathcal{C}([0,1],\mathbb{R})$ is endowed with the topology of uniform convergence.

Theorem 8.2. *We assume that Assumptions 8.1 hold, and that the sequence $(\tilde\beta^K(0,\cdot))$ converges in probability on $\mathcal{C}([0,1],\mathbb{R})$ to a deterministic function $\beta_0(\cdot)$ and that there exists a constant A such that*

$$\lim_{K\to+\infty} \mathbb{P}\big(L_0^K > A\big) = 0, \quad \text{where } L_0^K := \sup_{i\neq j} \frac{|\beta_i^K(0) - \beta_j^K(0)|}{\delta_K |i-j|}.$$

Then $\tilde\beta^K$ converges in probability in $\mathbb{D}([0,T], \mathcal{C}([0,1],\mathbb{R}))$ to the unique viscosity solution β of the Hamilton–Jacobi equation with state constraint boundary conditions

$$\begin{cases} \dfrac{\partial}{\partial t}\beta(t,x) = b(x) - d(x) + p(x)\displaystyle\int_{\mathbb{R}} m(h) e^{h\partial_x \beta(t,x)} dh, & (t,x) \in \mathbb{R}_+ \times (0,1), \\ \beta(0,x) = \beta_0(x), & x \in [0,1]. \end{cases} \tag{8.1}$$

More precisely, β is a viscosity supersolution of (8.1) in $(0,+\infty)\times(0,1)$ and a viscosity subsolution in $(0,+\infty)\times[0,1]$.

Usually, the analytical proof of such concentration results is based on the maximum principle (see [3]) which does not hold in this stochastic framework. To prove the tightness of the sequence $\tilde\beta^K$, a technical and delicate point consists in showing that the increments $(\beta_{i+1}^K(t) - \beta_i^K(t))/\delta_K$ are bounded uniformly in time for K large enough. These increments are semimartingales, and we easily obtain their Doob–Meyer decomposition. The martingale part is proved to be small for large K. The maximum principle is used to control the finite variation part, with an ω-by-ω argument. Once the tightness is obtained, we have to identify the limiting values of $\tilde\beta^K$, which only charge deterministic and continuous trajectories. We identify the limiting paths as viscosity solutions of the Hamilton–Jacobi equation (8.1).

FUNDING

This work was partially supported by the Chair "Modélisation Mathématique et Biodiversité" of Veolia Environnement-Ecole Polytechnique-Museum National d'Histoire Naturelle-Fondation X and by Labex Bézout (ANR-10-LABX-58).

REFERENCES

[1] R. Anderson and R. May, Population biology of infectious diseases: Part I. *Nature* **280** (1979), 361–367.

[2] V. Bansaye and S. Méléard, *Stochastic models for structured populations. Scaling limits and long time behavior*. MBI Lecture Ser. 1.4, Springer, 2015.

[3] G. Barles, S. Mirrahimi, and B. Perthame, Concentration in Lotka–Volterra parabolic equations: a general convergence result. *Methods Appl. Anal.* **16** (2009), 321–340.

[4] F. Baumdicker and P. Pfaffelhuber, The infinitely many genes model with horizontal gene transfer. *Electron. J. Probab.* **19** (2014), 1–27.

[5] S. Billiard, P. Collet, R. Ferrière, S. Méléard, and V. Tran, The effect of competition and horizontal trait inheritance on invasion, fixation and polymorphism. *J. Theoret. Biol.* **411** (2016), 48–58.

[6] S. Billiard, P. Collet, R. Ferrière, S. Méléard, and V. Tran, Stochastic dynamics for adaptation and evolution of microorganisms. In *Proceedings of 7th European Congress of Mathematics*, edited by V. Mehrmann and M. Skutella, pp. 527–552, European Mathematical Society, 2018.

[7] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, New York, 1968.

[8] J. Blath, T. Paul, and A. Tobias, A Stochastic Adaptive Dynamics Model for Bacterial Populations with Mutation, Dormancy and Transfer. 2021, arXiv:2105.09228.

[9] B. Bolker and S. Pacala, Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. *Theor. Popul. Biol.* **52** (1997), 179–197.

[10] A. Bovier, L. Coquille, and C. Smadi, Crossing a fitness valley as a metastable transition in a stochastic population model. *Ann. Appl. Probab.* **29** (2019), no. 6, 3541–3589.

[11] N. Champagnat, A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic Process. Appl.* **116** (2006), 1127–1160.

[12] N. Champagnat, R. Ferrière, and S. Méléard, Individual-based probabilistic models of adaptive evolution and various scaling approximations. In *Proceedings of the 5th seminar on stochastic analysis, random fields and applications*, Probab. Prog. Ser., Birkhäuser, Ascona, Suisse, 2006.

[13] N. Champagnat, R. Ferrière, and S. Méléard, Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models via timescale separation. *Theor. Popul. Biol.* **69** (2006), 297–321.

[14] N. Champagnat, P.-E. Jabin, and S. Méléard, Adaptive dynamics in a stochastic multi-resources chemostat model. *J. Math. Pures Appl.* **101** (2014), no. 6, 755–788.

[15] N. Champagnat and S. Méléard, Polymorphic evolution sequence and evolutionary branching. *Probab. Theory Related Fields* **151** (2011), no. 1–2, 45–94.

[16] N. Champagnat, S. Méléard, S. Mirrahimi, and V. C. Tran, Filling the gap between individual-based evolutionary models and Hamilton–Jacobi equations. 2022, in preparation.

[17] N. Champagnat, S. Méléard, and V. C. Tran, Stochastic analysis of emergence of evolutionary cyclic behaviour in population dynamics with transfer. *Ann. Appl. Probab.* **31** (2021), no. 4, 1820–1867.

[18] C. Darwin, *On the origin of species: A facsimile of the first edition*. Harvard University Press, 1964.

[19] L. Desvillettes, P. E. Jabin, S. Mischler, and G. Raoul, On selection dynamics for continuous structured populations. *Commun. Math. Sci.* **6** (2008), no. 3, 729–747.

[20] U. Dieckmann and R. Law, The dynamical theory of coevolution: a derivation from stochastic ecological processes. *J. Math. Biol.* **34** (1996), 579–612.

[21] O. Diekmann, A beginner's guide to adaptive dynamics. *Banach Center Publ.* **63** (2003), 47–86.

[22] O. Diekmann, P.-E. Jabin, S. Mischler, and B. Perthame, The dynamics of adaptation: an illuminating example and a Hamilton–Jacobi approach. *Theor. Popul. Biol.* **67** (2005), 257–271.

[23] R. Durrett and J. Mayberry, Travelling waves of selective sweeps. *Ann. Appl. Probab.* **21** (2011), no. 2, 699–744.

[24] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Stochastic Gene Expression in a Single Cell. *Science* **297** (2002), no. 5584, 1183–1186.

[25] N. Fournier and S. Méléard, A microscopic probabilistic description of a locally regulated population and macroscopic approximations. *Ann. Appl. Probab.* **14** (2004), no. 4, 1880–1919.

[26] M. I. Freidlin and A. Ventzell, *Random Perturbations of Dynamical Systems*. Springer, Berlin, 1984.

[27] P. Hinow, F. Le Foll, P. Magal, and G. Webb, Analysis of a model for transfer phenomena in biological populations. *SIAM J. Appl. Math.* **70** (2009), 40–62.

[28] J. Hofbauer and R. Sigmund, Adaptive dynamics and evolutionary stability. *Appl. Math. Lett.* **3** (1990), 75–79.

[29] A. Kraut and A. Bovier, From adaptive dynamics to adaptive walks. *J. Math. Biol.* **79** (2019), 1699–1747.

[30] B. Levin, F. Stewart, and V. Rice, Kinetics of conjugative plasmid transmission: fit of a simple mass action model. *Plasmid* **2** (1979), 247–260.

[31] A. Lorz, S. Mirrahimi, and B. Perthame, Dirac mass dynamics in multidimensional nonlocal parabolic equations. *Comm. Partial Differential Equations* **36** (2011), no. 6, 1071–1098.

[32] P. Magal and G. Raoul, Dynamics of a kinetic model describing protein exchanges in a cell population. 2015, arXiv:1511.02665.

[33] P. Marrow, R. Law, and C. Cannings, The coevolution of predator–prey interactions – ESSs and Red Queen dynamics. *Proc. R. Soc. Lond., B Biol. Sci.* **250** (1992), 133–141.

[34] S. Méléard, *Modèles aléatoires en Ecologie et Evolution*. Springer, 2016.

[35] J. Metz, S. Geritz, G. Meszéna, F. Jacobs, and J. V. Heerwaarden, Adaptative dynamics, a geometrical study of the consequences of nearly faithful reproduction. In *Stochastic and Spatial Structures of Dynamical Systems*, edited by S. J. Van Strien and S. M. Verduyn Lunel, Konink. Nederl. Akad. Wetensch. Verh. Afd. Natuurk. Eerste Reeks 45, North-Holland, Amsterdam, pp. 183–231, 1996.

[36] J. Metz, R. Nisbet, and S. Geritz, How should we define 'fitness' for general ecological scenarios? *Trends Ecol. Evol.* **7** (1992), 198–202.

[37] S. Mirrahimi, B. Perthame, and J. Y. Wakano, Evolution of species trait through resource competition. *J. Math. Biol.* **64** (2011), no. 7, 1189–1223.

[38] A. Novozhilov, G. Karev, and E. Koonin, Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.* **22** (2005), 1721–1732.

[39] A. Raj and A. van Oudenaarden, Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135** (2008), no. 2, 216–226.

[40] F. Stewart and B. Levin, The population biology of bacterial plasmids: A priori conditions for the existence of conjugationally transmitted factors. *Genetics* **87** (1977), 209–228.

[41] S. J. Tazzyman and S. Bonhoeffer, Fixation probability of mobile elements such as plasmids. *Theor. Popul. Biol.* **90** (2013), 49–55.

NICOLAS CHAMPAGNAT

Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France, nicolas.champagnat@inria.fr

SYLVIE MÉLÉARD

Institut Universitaire de France and Ecole polytechnique, CNRS, Institut polytechnique de Paris, route de Saclay, 91128 Palaiseau Cedex-France, sylvie.meleard@polytechnique.edu

VIET CHI TRAN

LAMA, Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, F-77454 Marne-la-Vallée, France, chi.tran@univ-eiffel.fr

QUANTITATIVE ANALYSIS OF FIELD CONCENTRATION IN PRESENCE OF CLOSELY LOCATED INCLUSIONS OF HIGH CONTRAST

HYEONBAE KANG

ABSTRACT

In composites consisting of inclusions and a matrix of different materials, some inclusions are located closely to each other. If the material properties of inclusions are of high contrast with that of the matrix, field concentration occurs in the narrow region between closely located inclusions. Understanding the field concentration quantitatively is important in the theory of composites and imaging since it represents stress or field enhancement. The last 30 years or so have witnessed significant progress in analyzing this phenomena of field concentration: optimal estimates and asymptotic characterization capturing the field concentration have been derived in the contexts of the conductivity equation (or antiplane elasticity), the Lamé system of linear elasticity, and the Stokes system. The purpose of this paper is to review some of them in a coherent manner.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 35J15; Secondary 35B65

KEYWORDS

Transmission problem, conductivity equation, Lamé system, Stokes system, high contrast, stress, field concentration, optimal estimates, spectrum, Neumann–Poincaré operator

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5680–5699
DOI 10.4171/ICM2022/83

1. INTRODUCTION

Typical composites consist of inclusions imbedded in the matrix (the background medium), where the inclusions have material properties different from that of the matrix. In some composites, two inclusions are located closely to each other, and if their material properties are in high contrast with that of the matrix, then a strong concentration of the field or stress may occur in the narrow region between two inclusions. It is important to quantitatively analyze the field concentration or the stress since it may cause material failure (see, for example, [5]).

Composites may have multiple inclusions. But, since the region of interest is the local narrow area in-between two inclusions which are closely located, all other inclusions except for the considered two are ignored and the mathematical problem is formulated with just two inclusions, that is, the problem is formulated in terms of disjoint bounded domains D_1 and D_2 in $\mathbb{R}^d$ $(d = 2, 3)$ representing the two inclusions. They are assumed to have Lipschitz continuous boundaries, and the interface conditions along ∂D_j $(j = 1, 2)$ are given by the perfectly bonding conditions, namely, continuity of the flux and the potential (see (2.5) and below). With these interface conditions, we consider the homogeneous and inhomogeneous transmission problems of various equations such as the equation of conductivity or antiplane elasticity, the Lamé system for linear elasticity, and the Stokes system for fluid flow. The inclusions represent conductors or insulators for conductivity equations, elastic inclusions for antiplane elasticity equations or Lamé systems, and suspensions for Stokes systems.

Throughout this paper, ε denotes the distance between two inclusions, namely,

$$\varepsilon := \operatorname{dist}(D_1, D_2). \tag{1.1}$$

The characteristic feature of the configuration for the problem is that ε is arbitrarily small. The mathematical problem here is to capture in a quantitative way the behavior of the field (the gradient of the solution) and its derivatives in the narrow region between D_1 and D_2 in terms of ε and, if possible, the contrast of material parameters. As mentioned before, this problem arises from the stress analysis in composites. It also arises from the effective medium theory [12, 25] (see also [19]): in order to compute the effective properties of composites with the periodic array of densely packed inclusions, it is necessary to capture the asymptotic behavior of the field in-between inclusions. Sometimes the two inclusions are designed to create the field concentration to achieve a desired enhancement of the field.

During the last three decades or so, significant development on the problem has been made: optimal estimates for the gradient and its derivatives have been obtained and asymptotic characterizations of the field concentration have been derived. The purpose of this paper is to review them. Despite all this progress, some outstanding and challenging problems remain unsolved. We discuss them as well.

The rest of this paper consists of three sections, reviewing the conductivity equation, the Lamé system, and the Stokes systems in turn. A short discussion is added at the end of the paper.

2. THE CONDUCTIVITY EQUATION

Let D_1 and D_2 be disjoint bounded domains in $\mathbb{R}^d$ $(d = 2, 3)$ whose boundaries are assumed to be Lipschitz continuous. Let k_j be the conductivity of D_j for $j = 1, 2$, while that of $\mathbb{R}^d \setminus (D_1 \cup D_2)$ is assumed to be 1. So the conductivity distribution is given by

$$\sigma = \chi_{\mathbb{R}^d \setminus (D_1 \cup D_2)} + k_1 \chi_{D_1} + k_2 \chi_{D_2}, \tag{2.1}$$

where χ denotes the characteristic function on the respective set. We assume that $0 < k_j \neq 1 < \infty$ $(j = 1, 2)$.

We consider the inhomogeneous transmission problem: for a given function f,

$$\begin{cases} \nabla \cdot \sigma \nabla u = f & \text{in } \mathbb{R}^d, \\ u(x) = c \ln |x| + O\big(|x|^{-1}\big) & \text{as } |x| \to \infty, \end{cases} \tag{2.2}$$

for some constant c. The constant c can be nonzero if $d = 2$, and it is zero if $d = 3$. We also consider the homogeneous transmission problem

$$\begin{cases} \nabla \cdot \sigma \nabla u = 0 & \text{in } \mathbb{R}^d, \\ u(x) - H(x) = O\big(|x|^{-d+1}\big) & \text{as } |x| \to \infty, \end{cases} \tag{2.3}$$

where H is a given function harmonic in $\mathbb{R}^d$. Instead of the free-space problems (2.2) and (2.3), one may consider the corresponding boundary value problems, which are equivalent to the above problems. However, the free-space problems seem more natural since the problems arise from the composite theory, and all but the two closely located inclusions are ignored.

When the conductivities k_1 and k_2 simultaneously tend ∞ or 0, it is expected for the ∇u of the solution u to become arbitrarily large as the distance ε between the two inclusions tends to 0. The problem is to derive estimates for ∇u in terms of ε (and k_1, k_2, if possible) as ε tends to 0. The conductivity being ∞ means that the inclusion is perfectly conducting, while 0 means insulating. The two-dimensional equation may represent the antiplane elasticity, and in such a case it means that the inclusion is either stiff or void. When $k_1 = 0$ and $k_2 = \infty$, or the other way around, a quite different singular behavior (blow-up) occurs as ε tends to 0 as we will see later.

When the distance ε tends to 0, the numerical computation of u becomes quite difficult since the blow-up of ∇u forces us to use a refined mesh. In this respect, an asymptotic characterization of the singularity of ∇u has an important role. By an asymptotic characterization, as ε tends to 0, we mean a decomposition of the form

$$u = s + r, \tag{2.4}$$

where s is the singular part, namely, ∇s carries the full information about the singularity of ∇u, while r is the regular part, namely, ∇r is bounded. To be used effectively for numerical computations, the singular part s needs to be the solution of the conductivity equation, and explicit.

The problem (2.2) can be expressed as

$$\begin{cases} \Delta u = f & \text{in } \mathbb{R}^d \setminus \overline{D}, \\ \Delta u = k_j^{-1} f & \text{in } D_j,\ j = 1,2, \\ u|_+ - u|_- = 0 & \text{on } \partial D_j,\ j = 1,2, \\ \partial_\nu u|_+ - k_j \partial_\nu u|_- = 0 & \text{on } \partial D_j,\ j = 1,2, \\ u(x) = c \ln |x| + O\big(|x|^{-1}\big) & \text{as } |x| \to \infty. \end{cases} \tag{2.5}$$

Here and throughout this paper, ∂_ν denotes the outward normal derivative on ∂D_j and the subscripts $\pm$ denote the limits from outside and inside of D_j, respectively. The third and fourth lines in (2.5) represent the perfect-bonding conditions along ∂D, namely the continuity of the potential and flux, respectively.

Let F be the (weighted) Newtonian potential of f, namely,

$$F(x) = \int_{\mathbb{R}^d \setminus D} \Gamma(x-y) f(y)\, dy + \sum_{j=1}^{2} \frac{1}{k_j} \int_{D_j} \Gamma(x-y) f(y)\, dy, \quad x \in \mathbb{R}^d, \tag{2.6}$$

where $\Gamma(x)$ is the fundamental solution to the Laplacian, i.e.,

$$\Gamma(x) = \begin{cases} \dfrac{1}{2\pi} \ln |x|, & d = 2, \\ -\dfrac{1}{4\pi} |x|^{-1}, & d = 3. \end{cases} \tag{2.7}$$

Since $\Delta F = f$ in $\mathbb{R}^d \setminus \overline{D}$ and $\Delta F = k_j^{-1} f$ in D_j, $v := u - F$ (u is the solution to (2.2)) is the solution to

$$\begin{cases} \Delta v = 0 & \text{in } D \cup (\mathbb{R}^d \setminus \overline{D}), \\ v|_+ - v|_- = 0 & \text{on } \partial D_j,\ j = 1,2, \\ \partial_\nu v|_+ - k_j \partial_\nu v|_- = (k_j - 1)\eta_j & \text{on } \partial D_j,\ j = 1,2, \\ v(x) = c \ln |x| + O\big(|x|^{-1}\big) & \text{as } |x| \to \infty, \end{cases} \tag{2.8}$$

with $\eta_j = \partial_\nu F|_{\partial D_j}$ ($j = 1,2$). That is, the inhomogeneous problem (2.2) is reduced to (2.8). By putting $v := u - H$, we see that the homogeneous (2.3) is reduced to (2.8) with $\eta_j = \partial_\nu H|_{\partial D_j}$.

The solution to (2.8) can be represented in terms of the single-layer potentials, and if it is done so, the problem is reduced to a system integral equations for the Neumann–Poincaré operator on $\partial D_1 \times \partial D_2$. In a recent paper [14], explicit solutions to (2.2) and (2.3) have been constructed when inclusions are circular using the complete knowledge of the spectrum for the Neumann–Poincaré operator on two circles. In Section 2.1, we review them and optimal estimates of the derivatives of the solution as consequences. We then review in Section 2.2 important generalizations to inclusions, of more general shape in two and three dimensions, of results for circular inclusions. These are actually results established earlier than the circular case of [14]; the review of this section is in reverse historical order. The merit in doing so is that the fine results for the case of circular inclusions may serve as milestones of which problems have been solved and which still need to be solved.

In the last subsection, we review the results on the asymptotic characterizations of the singular behavior of the gradient of the solution.

2.1. Estimates for circular inclusions

2.1.1. Explicit representation of the solution

Suppose that D_1 and D_2 are disks of radii r_1 and r_2, respectively. Explicit solutions are constructed in [14] by transforming circles ∂D_1 and ∂D_2 to two concentric circles. In order for the transformation to take a simple form, we make some necessary translations and rotations so that after them centers of D_1 and D_2 are located at $(c_1, 0)$ and $(c_2, 0)$, where

$$c_1 = \frac{r_2^2 - r_1^2 - (r_1 + r_2 + \varepsilon)^2}{2(r_1 + r_2 + \varepsilon)} - \frac{\beta}{2}, \quad c_2 = c_1 + r_1 + r_2 + \varepsilon, \tag{2.9}$$

with

$$\beta = \frac{\sqrt{\varepsilon}\sqrt{(2r_1 + \varepsilon)(2r_2 + \varepsilon)(2r_1 + 2r_2 + \varepsilon)}}{r_1 + r_2 + \varepsilon}. \tag{2.10}$$

Then, ∂D_1 and ∂D_2 are mapped onto two concentric circles by the transformation

$$z^* = Tz := \frac{\beta}{z} + 1, \tag{2.11}$$

namely, $T(\partial D_j)$ $(j = 1, 2)$ is the circle of the radius R_j centered at 0, where R_j is given by

$$R_1^2 = 1 + \frac{\beta}{c_1}, \quad R_2^2 = 1 + \frac{\beta}{c_2}. \tag{2.12}$$

Let

$$D_1^* := T(D_1) = \{|\zeta| < R_1\}, \quad D_2^* := T(D_2) = \{|\zeta| > R_2\}. \tag{2.13}$$

Let $H^{-1/2}(\partial D_j)$ denote the Sobolev space of order $-1/2$ on ∂D_j and $H_0^{-1/2}(\partial D_j)$ be the subspace of $H^{-1/2}(\partial D_j)$ whose element f satisfies $\int_{\partial D_j} f = 0$. Suppose that the function η_j appearing in (2.8) belongs to $H_0^{-1/2}(\partial D_j)$ and let H_j be the unique solution to the following Neumann boundary value problem:

$$\begin{cases} \Delta H_j = 0 & \text{in } D_j, \\ \partial_\nu H_j = \eta_j & \text{on } \partial D_j. \end{cases} \tag{2.14}$$

Let h_j be the analytic function in D_j^* such that $h_1(0) = 0$, $\lim_{|\zeta|\to\infty} h_2(\zeta) = 0$, and

$$H_j(z) = \Re(h_j \circ T)(z) + C_j, \quad z \in D_j, \tag{2.15}$$

for some constant C_j. Here and afterwards, $\Re$ indicates the real part. Let

$$\rho := \frac{R_1}{R_2} \quad \text{and} \quad \lambda_j := \frac{k_j + 1}{2(k_j - 1)}, \quad j = 1, 2, \tag{2.16}$$

and define functions w_j by

$$w_1(\zeta) = \sum_{l=0}^{\infty} \frac{h_1(\rho^{2l}\zeta)}{(4\lambda_1\lambda_2)^{l+1}}, \quad |\zeta| < R_1, \tag{2.17}$$

and

$$w_2(\zeta) = \sum_{l=0}^{\infty} \frac{h_2(\rho^{-2l}\zeta)}{(4\lambda_1\lambda_2)^{l+1}}, \quad |\zeta| > R_2. \tag{2.18}$$

Using functions w_1 and w_2, we define

$$A_1(\zeta) := \begin{cases} (\lambda_1+\lambda_2)w_1(\zeta) & \\ \quad + (\lambda_1-\lambda_2)w_1(\rho\zeta) - w_1(\rho^2\zeta), & |\zeta| \le R_1, \\ (\lambda_1+\lambda_2)w_1(R_1^2\bar{\zeta}^{-1}) & \\ \quad + (\lambda_1-\lambda_2)w_1(\rho\zeta) - w_1(\rho^2\zeta), & R_1 < |\zeta| \le R_2, \\ (\lambda_1+\lambda_2)w_1(R_1^2\bar{\zeta}^{-1}) & \\ \quad + (\lambda_1-\lambda_2)w_1(R_1R_2\bar{\zeta}^{-1}) - w_1(R_1^2\bar{\zeta}^{-1}), & R_2 < |\zeta|, \end{cases} \tag{2.19}$$

and

$$A_2(\zeta) := \begin{cases} (\lambda_1+\lambda_2)w_2(R_2^2\zeta^{-1}) & \\ \quad - (\lambda_1-\lambda_2)w_2(R_1R_2\zeta^{-1}) - w_2(R_2^2\zeta^{-1}), & |\zeta| \le R_1, \\ (\lambda_1+\lambda_2)w_2(R_2^2\zeta^{-1}) & \\ \quad - (\lambda_1-\lambda_2)w_2(\rho^{-1}\bar{\zeta}) - w_2(\rho^{-2}\bar{\zeta}), & R_1 < |\zeta| \le R_2, \\ (\lambda_1+\lambda_2)w_2(\bar{\zeta}) & \\ \quad - (\lambda_1-\lambda_2)w_2(\rho^{-1}\bar{\zeta}) - w_2(\rho^{-2}\bar{\zeta}), & R_2 < |\zeta|. \end{cases} \tag{2.20}$$

We have the following representation formula for the solution to (2.8).

Proposition 2.1. *Suppose $\eta_j \in H_0^{-1/2}(\partial D_j)$ $(j = 1, 2)$. The solution v to (2.8) is given by*

$$v(z) = \Re\big(A_1\big(T(z)\big) + A_2\big(T(z)\big)\big), \quad z \in \mathbb{R}^2. \tag{2.21}$$

For the inhomogeneous problem (2.2), $\eta_j = \partial_\nu F|_{\partial D_j}$, and hence the condition that η_j belongs to $H_0^{-1/2}(\partial D_j)$ $(j = 1, 2)$ amounts to

$$\int_{D_1} f = \int_{D_2} f = 0. \tag{2.22}$$

Thus we have the following corollary for (2.2).

Corollary 2.2. *Suppose that f satisfies (2.22). The solution u to (2.2) is represented as*

$$u(z) = F(z) + \Re\big(A_1\big(T(z)\big) + A_2\big(T(z)\big)\big) + \textit{const.} \tag{2.23}$$

For the general case when f does not necessarily satisfy condition (2.22), we can (explicitly) construct functions V_1 and V_2 such that the function f_0, defined by

$$f_0 = f - \left(\int_{D_1} f\right)\nabla\cdot\sigma\nabla V_1 - \left(\int_{D_2} f\right)\nabla\cdot\sigma\nabla V_2,$$

satisfies (2.22), and hence the solution u to (2.2) takes the form

$$u = \left(\int_{D_1} f\right)V_1 + \left(\int_{D_2} f\right)V_2 + u_0, \tag{2.24}$$

where u_0 is the solution to (2.2) of the form (2.21). The construction of functions V_1 and V_2 in [14] heavily uses the fact that D_1 and D_2 are disks.

For the homogeneous problem (2.3), $\eta_j = \partial_\nu H|_{\partial D_j}$ and hence $H_j = H$. Thus, we have the following corollary:

Corollary 2.3. *The solution u to (2.2) is represented as*

$$u(z) = H(z) + \Re\big(A_1\big(T(z)\big) + A_2\big(T(z)\big)\big). \tag{2.25}$$

2.1.2. Optimal estimates for the solution

We now present estimates for the solutions and their derivatives. These estimates are optimal and derived from the explicit representations of the solution presented in the previous subsection. The derivation is far from trivial.

We first introduce some norms for regularity of functions. A function g defined on $\mathbb{R}^2$ (with inclusions D_1 and D_2) is said to be piecewise $C^{n,\alpha}$ for some nonnegative integer n and $0 < \alpha < 1$ if g is $C^{n,\alpha}$ on $\overline{D_1}$, $\overline{D_2}$ and $\mathbb{R}^2 \setminus D$ ($D = D_1 \cup D_2$) separately. For piecewise $C^{n,\alpha}$ functions g, the norm is defined by

$$\|g\|_{n,\alpha} := \|g\|_{C^{n,\alpha}(\overline{D_1})} + \|g\|_{C^{n,\alpha}(\overline{D_2})} + \|g\|_{C^{n,\alpha}(\mathbb{R}^2\setminus D)}. \tag{2.26}$$

When $\alpha = 0$, we denote it by $\|g\|_{n,0}$. We also use the following norm:

$$\|g\|^*_{n,\alpha} := \frac{1}{k_1}\|g\|_{C^{n,\alpha}(\overline{D_1})} + \frac{1}{k_2}\|g\|_{C^{n,\alpha}(\overline{D_2})} + \|g\|_{C^{n,\alpha}(\mathbb{R}^2\setminus D)}. \tag{2.27}$$

When $(k_1 - 1)(k_2 - 1) > 0$ which includes the case when $k_1 = k_2 = \infty$ or $k_1 = k_2 = 0$ in limits, we obtain the following theorems for the inhomogeneous and homogeneous transmission problems. Here and throughout this paper, we put

$$r_* := \sqrt{\frac{2(r_1 + r_2)}{r_1 r_2}}. \tag{2.28}$$

We assume that the inhomogeneity f is given by $f = \nabla \cdot g$ for some g. It is assumed that g is compactly supported in $\mathbb{R}^2$ for the sake of simplicity.

Theorem 2.4. *Suppose $(k_1 - 1)(k_2 - 1) > 0$ and $f = \nabla \cdot g$ for some piecewise $C^{n-1,\alpha}$ function g with the compact support (n is a positive integer and $0 < \alpha < 1$). There is a constant $C > 0$ independent of k_1, k_2, ε, and g such that the solution u to (2.2) satisfies*

$$\|u\|_{n,0} \le C\|g\|^*_{n-1,\alpha}(4\lambda_1\lambda_2 - 1 + r_*\sqrt{\varepsilon})^{-n}. \tag{2.29}$$

This estimate is optimal in the sense that there is g such that the reverse inequality (with a different constant C) holds when $n = 1$.

Theorem 2.5. *Let Ω be a bounded set containing $\overline{D_1 \cup D_2}$. Let u be the solution to (2.3). If $(k_1 - 1)(k_2 - 1) > 0$, then there is a constant $C > 0$ independent of k_1, k_2, ε, and the function H such that*

$$\|u\|_{n,\Omega} \le C\|H\|_{C^n(\Omega)}(4\lambda_1\lambda_2 - 1 + r_*\sqrt{\varepsilon})^{-n}. \tag{2.30}$$

This estimate is optimal in the sense that there is a harmonic function H such that the reverse inequality (with a different constant C) holds for the case $n = 1$. Here, $\|u\|_{n,\Omega}$ denotes the piecewise C^n norm on Ω, namely,

$$\|u\|_{n,\Omega} := \|u\|_{C^n(\overline{D_1})} + \|u\|_{C^n(\overline{D_2})} + \|u\|_{C^n(\Omega\setminus D)}. \tag{2.31}$$

The estimates (2.29) and (2.30) are not new. The estimate (2.29) (for the inhomogeneous problem with circular inclusions) was obtained in [11]. The estimate (2.30) for the gradient for the homogeneous problem (with circular inclusions), namely, for $n = 1$, is obtained in [3, 4], while that for higher n in [11].

Since

$$4\lambda_1\lambda_2 - 1 = \frac{2(k_1 + k_2)}{(k_1 - 1)(k_2 - 1)},$$

the estimate (2.29) shows that if either k_1 or k_2 is finite (away from 0 and ∞), then $\|u\|_{n,0}$ is bounded regardless of the distance ε, while if both k_1 and k_2 tend to ∞, then the right-hand side of (2.29) is of order $\varepsilon^{-n/2}$. As explained at the end of this subsection, ∇u may actually blow up at the order of $\varepsilon^{-1/2}$. If k_1 and k_2 tend to 0, then the right-hand side of (2.29) is also of order $\varepsilon^{-n/2}$ provided that $\|g\|^*_{n-1,\alpha}$ is bounded, in particular, if there is no source in $D_1 \cup D_2$, namely, $g = 0$ in $D_1 \cup D_2$. The estimate (2.30) yields the same findings.

If $(k_1 - 1)(k_2 - 1) < 0$ which includes the case when $k_1 = 0$ and $k_2 = \infty$ (or the other way around) in limits, then $4\lambda_1\lambda_2 < 0$. Thus the right-hand sides of (2.29) and (2.30) are bounded and cannot be the right estimates for this case. Instead, we obtain the following theorems.

Theorem 2.6. *Suppose $(k_1 - 1)(k_2 - 1) < 0$ and $f = \nabla \cdot g$ for some piecewise $C^{n,\alpha}$ function g with compact support (n is a positive integer and $0 < \alpha < 1$). There is a constant $C > 0$ independent of k_1, k_2, ε, and g such that the solution u to (2.2) satisfies*

$$\|u\|_{n,0} \le C\|g\|^*_{n,\alpha}\left(4|\lambda_1\lambda_2| - 1 + r_*\sqrt{\varepsilon}\right)^{-n+1}. \tag{2.32}$$

This estimate is optimal in the sense that there is f such that the reverse inequality (with a different constant C) holds for $n = 2$.

Theorem 2.7. *Let Ω be a bounded set containing $\overline{D_1 \cup D_2}$. Let u be the solution to (2.3). If $(k_1 - 1)(k_2 - 1) < 0$, then there is a constant $C > 0$ independent of k_1, k_2, ε, and the function H such that*

$$\|u\|_{n,\Omega} \le C\|H\|_{C^{n+1}(\Omega)}\left(4|\lambda_1\lambda_2| - 1 + r_*\sqrt{\varepsilon}\right)^{-n+1}. \tag{2.33}$$

This estimate is optimal in the sense that there is a harmonic function H such that the reverse inequality (with a different constant C) holds for $n = 2$.

Estimates (2.32) and (2.33) show that if $(k_1 - 1)(k_2 - 1) < 0$, then ∇u is bounded regardless of the k_1, k_2, and ε. But, the nth ($n \ge 2$) order derivative may blow up at the rate of $\varepsilon^{-(n-1)/2}$ if, for example, $k_1 = 0$ and $k_2 = \infty$. The second derivative of u actually blows up at the rate of $\varepsilon^{-1/2}$ in some cases as explained in the next subsection. These results are new and waiting to be generalized to inclusions of general shape and to higher dimensions.

2.1.3. Optimality of the estimates

Let F be a smooth function in $\mathbb{R}^2$ with a compact support such that $F(z) = x_1$ in a neighborhood of $\overline{D_1 \cup D_2}$. Let $f := \Delta F$. Then the following hold [14]:

(i) Let $k_1 = k_2 = \infty$. The solution u to (2.2) satisfies

$$\left|\nabla u(z)\right| \gtrsim \varepsilon^{-1/2} \tag{2.34}$$

for some $z \in \mathbb{R}^2 \setminus \overline{D}$.

(ii) For the case when $(k_1 - 1)(k_2 - 1) < 0$, we take either $k_1 = 0$, $k_2 = \infty$ or $k_1 = \infty$, $k_2 = 0$. The solution u to (2.2) satisfies

$$\left|\nabla^2 u(z)\right| \gtrsim \varepsilon^{-1/2} \tag{2.35}$$

for some $z \in \mathbb{R}^2 \setminus \overline{D}$, while ∇u is bounded.

Similar estimates hold for the solution to the homogeneous problem (2.3) with $H(x) = x_1$ (the optimality of the gradient estimate is also shown [3]).

2.2. Estimates for inclusions of general shape

The estimate (2.30) shows that if k_1, k_2 are finite, namely, $0 < C_1 \le k_1$, $k_2 \le C_2 < \infty$ for some constants C_1, C_2, then ∇u is bounded regardless of ε. This fact is known to be true in a more general setting where there are several inclusions of arbitrary shape [29] (see [10] for the case of circular inclusions).

If $k_1 = k_2 = \infty$ (the perfectly conducting case), then we see from (2.30) that

$$\left|\nabla u(z)\right| \lesssim \varepsilon^{-1/2}. \tag{2.36}$$

This estimate and its optimality for the case of strictly convex inclusions (more generally, if they are strictly convex near the points of the shortest distance) in two dimensions has been proved in [34]. In three dimensions, the optimal estimate for ∇u has been obtained in [6] as

$$\left|\nabla u(z)\right| \lesssim \frac{1}{\varepsilon|\ln \varepsilon|}. \tag{2.37}$$

(See [26, 31] for the case of spherical inclusions.) In [21], a bow-tie structure, where two vertices are points of the shortest distance, is considered. It is proved that two kinds of singularities appear, one due to the corners and the other due to the interaction between the two inclusions.

If $k_1 = k_2 = 0$ (the insulating case), the same estimate for $|\nabla u|$ as the perfectly conducting case holds in two dimensions. This is due to the existence of harmonic conjugates and does not extend to three dimensions. In fact, the three-dimensional case is completely different. It is proved in [7] that if $k_1 = k_2 = 0$, the estimate

$$\left|\nabla u(z)\right| \lesssim \varepsilon^{-s} \tag{2.38}$$

holds with $s = 1/2$ when inclusions are strictly convex inclusions in three dimensions. It is then proved in [35] that the surprising estimate with $s = \frac{2-\sqrt{2}}{2}$ holds on the shortest line

segment between two spherical inclusions of the same radii. Recently in [30] the estimate with $s = 1/2 - \gamma$ for some $\gamma > 0$ was derived on strictly convex inclusions and for dimensions $d \geq 3$. An upper bound of γ for $d \geq 4$ has been derived in [33].

It is likely that in the three-dimensional insulating case the behavior of the gradient depends heavily on the geometry of inclusions, and it is not clear at all what the best possible s is in (2.38). It is not even clear if such a number exists; it may depend on the position x of the estimate. Clarifying this is now an outstanding open problem to be solved.

For the inhomogeneous problem, estimates on conducting inclusions of circular and bow-tie shapes in two dimensions and of spherical shape in three dimensions when the source function is an emitter, namely, $f = a \cdot \delta_z$ for some z outside inclusions, have been obtained [22–24]. Here, δ_z denotes the Dirac-delta function. Such a problem is considered in relation to the patched antenna where the field excited by an emitter of the dipole-type is enhanced by closely located antenna (see, for example, [32]).

Theorems 2.6 and 2.7 for the case $(k_1 - 1)(k_2 - 1) < 0$ are new and unexpected, and their extension to inclusions of general shape and to higher dimensions is wide open. Particular interest lies in the high contrast case, namely, $k_1 = 0$ and $k_2 = \infty$; we do not know whether the gradient is bounded and the higher order derivatives blow up, if so at what rate. The case of spherical inclusions seems already quite challenging.

2.3. Asymptotic characterizations of the gradient blow-up

The problem (2.3) in the limit $k_1 \to \infty$ and $k_2 \to \infty$ can be rewritten as

$$\begin{cases} \Delta u = 0 & \text{in } D^e, \\ u = \lambda_j \text{ (constant)} & \text{on } \partial D_j, \ j = 1, 2, \\ u(x) - H(x) = O\big(|x|^{1-d}\big) & \text{as } |x| \to \infty, \end{cases} \tag{2.39}$$

where $D^e := \mathbb{R}^d \setminus \overline{(D_1 \cup D_2)}$. The problem (2.39) is not an exterior Dirichlet problem since the constants λ_j are not prescribed. Rather, they are determined by the conditions

$$\int_{\partial D_j} \partial u|_+ \, dS = 0, \quad j = 1, 2. \tag{2.40}$$

The constants λ_1 and λ_2 may or may not be the same depending on the given H (and the configuration of inclusions). When they are different, a sharp gradient occurs if the distance between D_1 and D_2 is short.

The singular behavior of ∇u where u is the solution to (2.39) can be characterized by the singular function $q = q_D$ which is the solution to

$$\begin{cases} \Delta q = 0 & \text{in } D^e, \\ q = \text{constant} & \text{on } \partial D_j, \ j = 1, 2, \\ \displaystyle\int_{\partial D_j} \partial q|_+ dS = -(-1)^j, & j = 1, 2, \\ q(x) = O\big(|x|^{1-d}\big) & \text{as } |x| \to \infty. \end{cases} \tag{2.41}$$

For general inclusions D_1 and D_2, there is a unique solution to (2.41) (see [1]).

Using the singular function q_D, the solution u to (2.39) can be decomposed as

$$u = \alpha q_D + r, \tag{2.42}$$

where

$$\alpha = \frac{u|_{\partial D_2} - u|_{\partial D_1}}{q_D|_{\partial D_2} - q_D|_{\partial D_1}}. \tag{2.43}$$

Here the constant α and functions q_D, r depend on ε. Observe that r attains constant values on ∂D_1 and ∂D_2, and $r|_{\partial D_1} = r|_{\partial D_2}$, so that ∇r is bounded on D^e (see [16]). Thus the term $\alpha \nabla q_D$ characterizes the blow-up of ∇u as $\varepsilon \to 0$. In particular, since ∇q_D is of order $\varepsilon^{-1/2}$, α represents the magnitude of the blow-up, and hence is called the stress concentration factor.

If $D_1 = B_1$ and $D_2 = B_2$ are two disjoint disks, the solution q (we denote it by q_B in this case) can be found explicitly. Let R_j be the inversion with respect to ∂B_j ($j = 1, 2$), and let δ_1 and δ_2 be the unique fixed points of the combined inversions $R_1 \circ R_2$ and $R_2 \circ R_1$, respectively. Let

$$q_B(x) = \frac{1}{2\pi}\big(\ln|x - \delta_1| - \ln|x - \delta_2|\big). \tag{2.44}$$

The function q_B is the solution to (2.41). In particular, q_B is constant on ∂B_j because ∂B_1 and ∂B_2 are circles of Apollonius of points δ_1 and δ_2. The function q_B appears in the bipolar coordinate system for ∂B_1 and ∂B_2 and was used for analysis of the field concentration for the first time in [34]. Using the explicit form of the function q_B, it is proved that

$$\|\nabla q_B\|_{L^\infty(\mathbb{R}^2 \setminus (B_1 \cup B_2))} \sim \varepsilon^{-1/2}. \tag{2.45}$$

Results on asymptotic characterizations of the gradient blow-up in two dimensions may be summarized as follows:

(i) If $D_1 = B_1$ and $D_2 = B_2$ are disks, then

$$\alpha = \frac{4\pi r_1 r_2}{r_1 + r_2} \frac{(z_2 - z_1) \cdot \nabla H(\frac{z_1 + z_2}{2})}{|z_2 - z_1|} + O(\sqrt{\varepsilon}) \quad \text{as } \varepsilon \to 0, \tag{2.46}$$

where r_j is the radius of D_j, $j = 1, 2$ [16].

(ii) Suppose that ∂D_j is $\mathcal{C}^{2,\gamma}$ for some $\gamma \in (0, 1)$. We further suppose that there are unique points $z_1 \in \partial D_1$ and $z_2 \in \partial D_2$ such that $|z_1 - z_2| = \operatorname{dist}(D_1, D_2)$ and there is a common neighborhood U of z_1 and z_2 such that $D_j \cap U$ is strictly convex for $j = 1, 2$. Let B_j be the disk osculating to D_j at z_j ($j = 1, 2$). Then,

$$\nabla q_D = \nabla q_B\big(1 + O\big(\varepsilon^{\gamma/2}\big)\big) + O(1), \tag{2.47}$$

and

$$\alpha = \frac{\sqrt{2}\pi}{\sqrt{\kappa_1 + \kappa_2}} \frac{1}{\sqrt{\varepsilon}} \int_{\partial D_1 \cup \partial D_2} H \partial_\nu q_D \, d\sigma \big(1 + O\big(\varepsilon^{\gamma/2}\big)\big). \tag{2.48}$$

In particular, α is bounded regardless of ε [1].

(iii) Let D_1^0 and D_2^0 be the touching inclusions obtained as the limit of D_1 and D_2 as $\varepsilon \to 0$ (D_1 and D_2 are still assumed to satisfy assumptions of (ii)), and let u_0 be the solution for the touching case, namely,

$$\begin{cases} \Delta u_0 = 0 & \text{in } D_0^e, \\ u_0 = \lambda_0 & \text{on } \partial D_0^e, \\ u_0(x) - H(x) = O(|x|^{-1}) & \text{as } |x| \to \infty, \end{cases} \tag{2.49}$$

where $D_0^e := \mathbb{R}^2 \setminus \overline{(D_1^0 \cup D_2^0)}$ and λ_0 is a constant determined by the additional condition

$$\int_\Omega |\nabla(u_0 - H)|^2 dA < \infty. \tag{2.50}$$

Then,

$$\alpha = \int_{\partial D_1^0} \partial_\nu u_0 + O\big(\varepsilon |\log \varepsilon|\big) \tag{2.51}$$

as $\varepsilon \to 0$ [15].

The decomposition formula (2.42) (together with (2.47) and (2.51)) has some important consequences. Since ∇q_D is bounded from below and above by $\varepsilon^{-1/2}$ (up to constant multiples), the blow-up estimates for ∇u can be obtained from the formula. It can be used to compute u numerically. Since the formula extracts the leading singular term in an explicit way, it suffices to compute the residual term b for which only regular meshes are required. This idea appeared and was exploited in [16] in the special case when D_j are disks.

The formula (2.42) has another very interesting implication. The quantity $\nabla u \cdot n$ represents the charge density on $\partial D_1 \cup \partial D_2$ induced by the field $-\nabla H$, and $\nabla u_0 \cdot n$ does that on $\partial D_1^0 \cup \partial D_2^0$. Note that the charge densities on the separated inclusions have a singular part $\alpha \nabla q_D \cdot n$ and a regular part $\nabla r \cdot n$. It is proved in [15] that $\nabla r \cdot n$ converges to $\nabla u_0 \cdot n$ as $\varepsilon \to 0$, that is, as the separated inclusions approach the touching ones. So the singular part suddenly disappears when the two inclusions become touching. It is reminiscent of the electrical spark occurring between two separated conductors which suddenly disappears when the conductors are touching.

The decomposition formula of the kind (2.42) when D_1 and D_2 are three-dimensional balls of the same radii has been derived in [17] (see [27] for the case of different radii). In this case the singular function is given as an infinite superposition of point charges.

3. LAMÉ SYSTEM

In this section we review results on the field concentration for the Lamé system of linear elasticity. If Lamé parameters are finite so that inclusions are of low contrast with the matrix, then the gradient of the solution is bounded regardless of the distance between inclusions. This is the well-known result of Li–Nirenberg [28]. The only known results for the high contrast case are when inclusions are hard and strictly convex. We review them here. Hard inclusions for the elasticity correspond to the perfect conductors for the electricity and are characterized by the boundary condition as explained blow.

As before, let D_1 and D_2 be bounded domains in $\mathbb{R}^2$. Let (λ, μ) be the pair of Lamé constants of $D^e = \mathbb{R}^2 \setminus \overline{(D_1 \cup D_2)}$ which satisfies the strong ellipticity conditions, $\mu > 0$ and $\lambda + \mu > 0$ (we only consider the two-dimensional case). The Lamé operator is given by

$$\mathcal{L}_{\lambda,\mu} u := \mu \Delta u + (\lambda + \mu) \nabla \nabla \cdot u, \tag{3.1}$$

where $u = (u_1, u_2)^T$ (T for transpose) is a vector-valued function. Let

$$\Psi_1(x) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \Psi_2(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \Psi_3(x) = \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}, \tag{3.2}$$

which are the displacement fields of the rigid motions.

The problem for the Lamé system is given as follows:

$$\begin{cases} \mathcal{L}_{\lambda,\mu} u = 0 & \text{in } D^e, \\ u = \displaystyle\sum_{j=1}^{3} c_{ij} \Psi_j & \text{on } \partial D_i, i = 1, 2, \\ u(x) - H(x) = O(|x|^{-1}) & \text{as } |x| \to \infty, \end{cases} \tag{3.3}$$

where $H = (h_1, h_2)^T$ is a given function satisfying $\mathcal{L}_{\lambda,\mu} H = 0$ in $\mathbb{R}^2$. The boundary conditions to be satisfied by the displacement u on ∂D_j (the second line in (3.3)) indicate that D_1 and D_2 are hard inclusions. The constants c_{ij} there are not given but determined by the condition similar to (2.40), that is,

$$\int_{\partial D_i} \Psi_j \cdot \sigma[u] n \, ds = 0, \quad i = 1, 2, \ j = 1, 2, 3. \tag{3.4}$$

Here, $\sigma[u]$ denotes the stress tensor corresponding to the displacement vector u, defined by

$$\sigma[u] := \lambda (\nabla \cdot u) + 2\mu (\widehat{\nabla} u),$$

where $\widehat{\nabla} u = \frac{1}{2}(\nabla u + \nabla u^T)$.

An asymptotic characterization of the solution u to (3.3), which captures the singular behavior of ∇u, is obtained in [18]. It is given in terms of singular functions which are constructed by the singular function q_B for the conductivity problem given in (2.44). To describe them, let z_1, z_2, B_1, B_2 be as before (right before (2.47)), namely $z_1 \in \partial D_1$ and $z_2 \in \partial D_2$ are unique points such that $|z_1 - z_2| = \operatorname{dist}(D_1, D_2)$, there is a common neighborhood U of z_1 and z_2 such that $D_i \cap U$ is strictly convex for $i = 1, 2$, and B_i is the disk osculating to D_i at z_i ($i = 1, 2$). Let δ_1 and δ_2 be the points appearing in the definition (2.44) of q_B, namely the fixed points of the combined inversions. After a translation and a rotation if necessary, we may assume that $\delta_1 = (-a, 0)$ and $\delta_2 = (a, 0)$. This number a is actually satisfies $a = 2\beta$, where β is given in (2.10). If we denote the centers of B_i by $(c_i, 0)$ ($i = 1, 2$), then c_i satisfies the relation

$$c_i = (-1)^i \sqrt{r_i^2 + a^2}, \quad i = 1, 2. \tag{3.5}$$

Let $q = q_B$ and let

$$\alpha_1 = \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{\lambda + 2\mu}\right) \quad \text{and} \quad \alpha_2 = \frac{1}{2}\left(\frac{1}{\mu} - \frac{1}{\lambda + 2\mu}\right). \tag{3.6}$$

Singular functions Q_1 and Q_2 for the elasticity problem of this section are defined by

$$Q_1 = \alpha_1 \begin{bmatrix} q \\ 0 \end{bmatrix} - \alpha_2 x_1 \nabla q \tag{3.7}$$

and

$$Q_2 = \alpha_1 \begin{bmatrix} 0 \\ q \end{bmatrix} + \alpha_2 x_1 (\nabla q)^{\perp}, \tag{3.8}$$

where $(a, b)^{\perp} = (-b, a)$. Actually, these functions were found in [18] as linear combinations of point-source functions in linear elasticity called nuclei of strain. It turns out that they can be expressed in simple forms using the function q (see also [20]).

One can easily see that Q_j are solutions to the Lamé system, namely

$$\mathcal{L}_{\lambda,\mu} Q_j = 0 \quad \text{in } \mathbb{R}^2 \setminus \{\delta_1, \delta_2\}. \tag{3.9}$$

It is shown in [18] that Q_j takes "almost" constant values Ψ_j on the osculating circles ∂B_i $(i = 1, 2)$. In fact, there are constants k_{ji} and l_{ji} such that for $i = 1, 2$,

$$Q_1(x) = k_{1i} \Psi_1(x) + l_{1i} x, \quad x \in \partial B_i, \tag{3.10}$$

and

$$Q_2(x) = k_{2i} \Psi_2(x) + l_{2i} x^{\perp}, \quad x \in \partial B_i. \tag{3.11}$$

Actually, the constants k_{ji} and l_{ji} can be easily derived using the simple forms Q_j. Using the fact that q is constant on ∂B_i, one can show that

$$\nabla q(x) = -\frac{a}{2\pi r_i} \frac{1}{x_1} (x_1 - c_i, x_2), \quad x \in \partial B_i,\ i = 1, 2.$$

It thus follows that for $i = 1, 2$,

$$k_{1i} = \alpha_1 q|_{\partial B_i} - \frac{\alpha_2 a c_i}{2\pi r_i}, \quad l_{1i} = \frac{\alpha_2 a}{2\pi r_i}, \tag{3.12}$$

and

$$k_{2i} = \alpha_1 q|_{\partial B_i} + \frac{\alpha_2 a c_i}{2\pi r_i}, \quad l_{2i} = -\frac{\alpha_2 a}{2\pi r_i} \tag{3.13}$$

Another function related with the boundary value Ψ_3 on ∂B_1 and ∂B_2 is constructed in the same paper. But this function has nothing to do with the singular behavior of the field, so we omit it here. It is worth mentioning that the singular functions Q_1 and Q_2 are effectively utilized to prove the Flaherty–Keller formula [12] describing the effective property of densely packed elastic composites [19].

Using the singular functions Q_1 and Q_2, it is proved that the solution u to (3.3) admits the following decomposition:

$$u = C_1 Q_1 + C_2 Q_2 + b, \tag{3.14}$$

where C_1 and C_2 are constants depending on ε, but bounded independently of ε, and b is a function whose gradient is bounded on any bounded subset of D^e. The following estimate is obtained as an immediate consequence of the decomposition formula:

$$\|\nabla u\|_{L^\infty(D^e)} \lesssim \varepsilon^{-1/2}. \tag{3.15}$$

This estimate is also proved in [8]. This estimate is optimal in the sense that the reverse inequality holds in some cases. An extension to three dimensions has been achieved in [9].

We emphasize that the constants C_1 and C_2 appearing in formula (3.14) are not explicit. Thus further investigation on how to determine them (or compute them numerically) is desired.

4. STOKES SYSTEM

In this section we review the result in [2], that is, an asymptotic characterization of the stress concentration for the Stokes flow modeled by $\mu\Delta u = \nabla p$ and $\nabla \cdot u = 0$. Here, μ represents the constant viscosity of the fluid. Even if the result is only for the two-dimensional inclusions of circular shape, the result may serve as a milestone for further development.

Let D_1 and D_2 be disks and let $D^e = \mathbb{R}^2 \setminus \overline{D_1 \cup D_2}$ as before. Let (U, P) is a given background solution to the homogeneous Stokes system in $\mathbb{R}^2$, namely, $\mu\Delta U = \nabla P$ and $\nabla \cdot U = 0$ in $\mathbb{R}^2$. We consider the following problem of the Stokes system:

$$\begin{cases} \mu\Delta u = \nabla p & \text{in } D^e, \\ \nabla \cdot u = 0 & \text{in } D^e, \\ u = \displaystyle\sum_{j=1}^{3} d_{ij}\Psi_j & \text{on } \partial D_i,\ i = 1, 2, \end{cases} \tag{4.1}$$

with the conditions

$$(u - U)(x) = O\big(|x|^{-1}\big), \quad \nabla(u - U)(x) = O\big(|x|^{-2}\big), \quad (p - P)(x) = O\big(|x|^{-2}\big)$$

as $|x| \to \infty$. Here, Ψ_j are the functions given in (3.2), and d_{ij} are constants to be determined from the equilibrium conditions

$$\int_{\partial D_i} \Psi_j \cdot \sigma[u, p]n \, d\sigma = 0, \quad i = 1, 2,\ j = 1, 2, 3. \tag{4.2}$$

Here, $\sigma[u, p]$ is the stress field induced by the velocity-pressure pair (u, p), namely

$$\sigma[u, p] = -pI + 2\mu\widehat{\nabla}u, \tag{4.3}$$

where I is the identity matrix.

As the distance between D_1 and D_2 tends to 0, the solution to (4.1) exhibits singular behavior in its gradient which can be captured in terms of singular functions. The singular functions for (4.1) form the solution (V_j, p_j) $(j = 1, 2)$ to the following problem:

$$\begin{cases} \mu\Delta V_j = \nabla p_j & \text{in } \mathbb{R}^2 \setminus \{\delta_1, \delta_2\}, \\ \nabla \cdot V_j = 0 & \text{in } \mathbb{R}^2 \setminus \{\delta_1, \delta_2\}, \\ V_j = \dfrac{(-1)^i}{2}\Psi_j & \partial B_i,\ i = 1, 2, \end{cases} \tag{4.4}$$

with the conditions

$$V_j(x) = C_j + O\big(|x|^{-1}\big), \quad \nabla V_j(x) = O\big(|x|^{-2}\big), \quad p_j(x) = O\big(|x|^{-2}\big)$$

for some constant C_j as $|x| \to \infty$. Here δ_j is the point appearing in (2.44).

In [2], singular functions (V_j, p_j) are constructed using the stream function formulation for which the bipolar coordinate system is used. We assume $\delta_1 = (-a, 0)$ and $\delta_2 = (a, 0)$ as before. Then, the bipolar coordinates (ζ, θ) are defined by

$$\zeta = 2\pi q_D, \quad \theta = \arg(x - a, y) - \arg(x + a, y). \tag{4.5}$$

Let

$$e_\zeta = \frac{\nabla\zeta}{|\nabla\zeta|}, \quad e_\theta = \frac{\nabla\theta}{|\nabla\theta|}.$$

Suppose that D_1 and D_2 have the same radius, say R, and let

$$s = \sinh^{-1}(a/R).$$

Define two constants A_1 and B_1 by

$$A_1 := \frac{1}{2s - \tanh 2s}, \quad B_1 := -\frac{1}{2\cosh 2s} A_1. \tag{4.6}$$

Then, the velocity V_1 is given by $V_1 = v_{1\zeta} e_\zeta + v_{1\theta} e_\theta$ where

$$v_{1\zeta} = (A_1\zeta + B_1 \sinh 2\zeta) \frac{1 - \cosh\zeta \cos\theta}{\cosh\zeta - \cos\theta}, \tag{4.7}$$

$$v_{1\theta} = \sin\theta \left(A_1 + 2B_1 \cosh 2\zeta - \frac{\sinh\zeta (A_1\zeta + B_1 \sinh 2\zeta)}{\cosh\zeta - \cos\theta} \right), \tag{4.8}$$

and the pressure p_1 is given by

$$p_1 = \frac{2\mu}{a} \big((A_1 - 2B_1) \cosh\zeta \cos\theta + B_1 \cosh 2\zeta \cos 2\theta \big) - \frac{2\mu}{a} (A_1 - B_1). \tag{4.9}$$

The formulas for (V_2, p_2) are quite involved. But it is proved in [2] that

$$V_2 = -A_2 \begin{bmatrix} 0 \\ \zeta \end{bmatrix} + A_2 x (\nabla\zeta)^\perp + V_{2o} \tag{4.10}$$

and

$$p_2 = -\frac{2\mu}{a} A_2 \sinh\zeta \sin\theta + p_{2o}, \tag{4.11}$$

where (V_{2o}, p_{2o}) is a solution to the Stokes system whose gradient is bounded regardless of ε, and A_2 is the constant defined by

$$A_2 = -\frac{1}{2s + \sinh 2s}. \tag{4.12}$$

The function V_2 is similar to the function Q_2 for the Lamé system given in (3.8).

It is proved in the same paper that if the background velocity field U is given by

$$U(x_1, x_2) = \begin{bmatrix} \alpha & \beta \\ \gamma & -\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \big(\alpha^2 + (\beta + \gamma)^2 \neq 0 \big) \tag{4.13}$$

for some constants α, β, and γ, the background pressure $P = 0$, and if D_1 and D_2 are disks of the same radius R, then the solution (u, p) admits a decomposition of the following form:

$$(u, p) = \alpha \frac{2}{\sqrt{R}} \varepsilon^{3/2} (V_1, p_1) + \frac{\beta + \gamma}{2} \sqrt{R\varepsilon} (V_2, p_2) + (u_0, p_0), \tag{4.14}$$

where (u_0, p_0) is a solution to the Stokes problem whose stress tensor is bounded. Thus we have

$$\sigma[u, p] = \alpha \frac{2}{\sqrt{R}} \varepsilon^{3/2} \sigma[V_1, p_1] + \frac{\beta + \gamma}{2} \sqrt{R\varepsilon}\, \sigma[V_2, p_2] + \sigma[u_0, p_0]. \tag{4.15}$$

Since $\|\sigma[V_1, p_1]\|_{L^\infty(D^e)} \approx \varepsilon^{-2}$ and $\|\sigma[V_2, p_2]\|_{L^\infty(D^e)} \approx \varepsilon^{-1}$ as proved in [2], we have

$$\|\sigma[u, p]\|_{L^\infty(D^e)} \approx \varepsilon^{-1/2}, \tag{4.16}$$

which says that the stress always blows up at the rate of $\varepsilon^{-1/2}$ provided that U is linear as given in (4.13) and inclusions are circular. It is quite interesting and challenging to extend this result to the noncircular case.

5. CONCLUSIONS

In this paper we review significant results on optimal estimates of the derivatives and asymptotic characterizations of the solution in the presence of two inclusions when the distance between them tends to zero. A special emphasis is laid on the case of high contrast. We review results on the conductivity equation, the Lamé system, and the Stokes system. Apart from these equations, the stress concentration factor for the p-Laplacian has been derived in [13].

As mentioned in the text, many challenging problems remain unsolved. Among them, the problem for the three-dimensional insulating case is outstanding. The case when the conductivities k_1 and k_2 satisfy the condition $(k_1 - 1)(k_2 - 1) < 0$ is also quite interesting. It goes without saying that the studies of problems for the Lamé and Stokes systems are in their early stage. Extensions to general shape and higher dimensions are quite challenging.

FUNDING

This work was partially supported by NRF (of S. Korea) grant No. 2019R1A2B5B0106-9967.

REFERENCES

[1] H. Ammari, G. Ciraolo, H. Kang, H. Lee, and K. Yun, Spectral analysis of the Neumann–Poincaré operator and characterization of the stress concentration in anti-plane elasticity. *Arch. Ration. Mech. Anal.* **208** (2013), 275–304.

[2] H. Ammari, H. Kang, D. W. Kim, and S. Yu, Quantitative estimates for stress concentration of the Stokes flow between adjacent circular cylinders. 2020, arXiv:2003.06578.

[3] H. Ammari, H. Kang, H. Lee, J. Lee, and M. Lim, Optimal Estimates for the Electrical Field in Two Dimensions. *J. Math. Pures Appl.* **88** (2007), 307–324.

[4] H. Ammari, H. Kang, and M. Lim, Gradient estimates for solutions to the conductivity problem. *Math. Ann.* **332** (2005), 277–286.

[5] I. Babuška, B. Andersson, P. Smith, and K. Levin, Damage analysis of fiber composites. I. Statistical analysis on fiber scale. *Comput. Methods Appl. Mech. Engrg.* **172** (1999), 27–77.

[6] E. Bao, Y. Y. Li, and B. Yin, Gradient estimates for the perfect conductivity problem. *Arch. Ration. Mech. Anal.* **193** (2009), 195–226.

[7] E. Bao, Y. Y. Li, and B. Yin, Gradient estimates for the perfect and insulated conductivity problems with multiple inclusions. *Comm. Partial Differential Equations* **35** (2010), 1982–2006.

[8] J. Bao, H. Li, and Y. Li, Gradient estimates for solutions of the Lamé system with partially infinite coefficients. *Arch. Ration. Mech. Anal.* **215** (2015), 307–351.

[9] J. Bao, H. Li, and Y. Li, Gradient estimates for solutions of the Lamé system with partially infinite coefficients in dimensions greater than two. *Adv. Math.* **305** (2017), 298–338.

[10] E. Bonnetier and M. Vogelius, An elliptic regularity result for a composite medium with "touching" fibers of circular cross-section. *SIAM J. Math. Anal.* **31** (2000), 651–677.

[11] H. Dong and H. Li, Optimal estimates for the conductivity problem by Green's function method. *Arch. Ration. Mech. Anal.* **231** (2019), 1427–1453.

[12] J. E. Flaherty and J. B. Keller, Elastic behavior of composite media. *Comm. Pure Appl. Math.* **26** (1973), 565–580.

[13] Y. Gorb and A. Novikov, Blow-up of solutions to a p-Laplace equation. *Multiscale Model. Simul.* **10** (2012), 727–743.

[14] Y. Ji and H. Kang, Spectrum of the Neumann–Poincaré operator and optimal estimates for transmission problems in presence of two circular inclusions. 2021, arXiv:2105.06093.

[15] H. Kang, H. Lee, and K. Yun, Optimal estimates and asymptotics for the stress concentration between closely located stiff inclusions. *Math. Ann.* **363** (2015), 1281–1306.

[16] H. Kang, M. Lim, and K. Yun, Asymptotics and computation of the solution to the conductivity equation in the presence of adjacent inclusions with extreme conductivities. *J. Math. Pures Appl.* **99** (2013), 234–249.

[17] H. Kang, M. Lim, and K. Yun, Characterization of the electric field concentration between two adjacent spherical perfect conductors. *SIAM J. Appl. Math.* **74** (2014), 125–146.

[18] H. Kang and S. Yu, Quantitative characterization of stress concentration in the presence of closely spaced hard inclusions in two-dimensional linear elasticity. *Arch. Ration. Mech. Anal.* **232** (2019), 121–196.

[19] H. Kang and S. Yu, A proof of the Flaherty–Keller formula on the effective property of densely packed elastic composites. *Calc. Var. Partial Differential Equations* **59** (2020), 22.

[20] H. Kang and S. Yu, Singular functions and characterizations of field concentrations: a survey. *Anal. Theory Appl.* **37** (2021), 102–113.

[21] H. Kang and K. Yun, Optimal estimates of the field enhancement in presence of a bow-tie structure of perfectly conducting inclusions in two dimensions. *J. Differential Equations* **266** (2019), 5064–5094.

[22] H. Kang and K. Yun, Precise estimates of the field excited by an emitter in presence of closely located inclusions of a bow-tie shape. *J. Math. Anal. Appl.* **479** (2019), no. 2, 1670–1707.

[23] H. Kang and K. Yun, Quantitative estimates of the field excited by an emitter in a narrow region between two circular inclusions. *Quart. Appl. Math.* **LXXVII** (2019), no. 4, 861–873.

[24] H. Kang and K. Yun, Quantitative estimates for enhancement of the field excited by an emitter due to presence of two closely located spherical inclusions. *J. Differential Equations* **269** (2020), 2977–3002.

[25] J. B. Keller, Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders. *J. Appl. Phys.* **34** (1963), 991–993.

[26] J. Lekner, Electrostatics of two charged conducting spheres. *Proc. R. Soc. A* **468** (2012), 2829–2848.

[27] H. Li, F. Wang, and L. Xu, Characterization of electric fields between two spherical perfect conductors with general radii in 3D. *J. Differential Equations* **267** (2019), 6644–6690.

[28] Y. Y. Li and L. Nirenberg, Estimates for elliptic system from composite material. *Comm. Pure Appl. Math.* **LVI** (2003), 892–925.

[29] Y. Y. Li and M. Vogelius, Gradient estimates for solution to divergence form elliptic equation with discontinuous coefficients. *Arch. Ration. Mech. Anal.* **153** (2000), 91–151.

[30] Y. Y. Li and Z. Yang, Gradient estimates of solutions to the insulated conductivity problem in dimension greater than two. 2020, arXiv:2012.14056.

[31] M. Lim and K. Yun, Blow-up of electric fields between closely spaced spherical perfect conductors. *Comm. Partial Differential Equations* **34** (2009), 1287–1315.

[32] V. Pacheco-Penã, M. Beruete, A. I. Fernández-Domínquez, Y. Luo, and M. Navarro-Cía, Description of bow-tie nanoantennas excited by localized emitters using conformal transformation. *ACS Photonics* **3** (2016), 1223–1232.

[33] B. Weinkove, The insulated conductivity problem, effective gradient estimates and the maximum principle. 2021, arXiv:2103.14143.

[34] K. Yun, Estimates for electric fields blown up between closely adjacent conductors with arbitrary shape. *SIAM J. Appl. Math.* **67** (2007), 714–730.

[35] K. Yun, An optimal estimate for electric fields on the shortest line segment between two spherical insulators in three dimensions. *J. Differential Equations* **261** (2016), 148–188.

HYEONBAE KANG
Department of Mathematics and Institute of Applied Mathematics, Inha University, Incheon 22212, S. Korea, hbkang@inha.ac.kr

19. MATHEMATICAL EDUCATION AND POPULARIZATION OF MATHEMATICS

THE HUG OF THE SCUTOID

CLARA I. GRIMA

ABSTRACT

This paper is a personal account of my work in the popularization of mathematics. How I started doing math popularization, why I think it is important to do this kind of tasks, and, finally, how this work can lead to some fruitful results in pure research that, initially, seems not to be related with that work of popularization.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 00A35; Secondary 00A08, 00A06, 97A20, 97B60

KEYWORDS

Mathematic popularization, mathematic dissemination

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5702–5714
DOI 10.4171/ICM2022/38

1. INTRODUCTION

Day after day, the popularization and dissemination of the results obtained by the researchers is not only considered more important, but it is compulsory in order to fulfill the conditions of many research grants. Many reasons have been given to justify why science popularization and dissemination is crucial. In fact, nowadays, as it is said in [4], "Dissemination and communication of research should be considered as an integral part of any research project. Both help in increasing the visibility of research outputs, public engagement in science and innovation, and confidence of society in research. Effective dissemination and communication are vital to ensure that the conducted research has a social, political, or economical impact. They draw [the] attention of governments and stakeholders to research results and conclusions, enhancing their visibility, comprehension, and implementation." But this paper tries to be just a personal account of my experience in the field of math popularization, and one thing is why universities, governments, and any other institution must encourage dissemination and popularization of science in general, or mathematics in particular, and a very different thing is why any particular individual, myself in this occasion, is doing this kind of tasks. Of course, in order to understand a particular case, we have to keep in mind a more general scope, so, we shall briefly try to answer the typical questions of why, what, where, and how, both from a general and from my very personal point of view. In fact, those topics, to a greater or lesser extent, have been considered previously in the six works on this subject (just six) in other ICM's editions. In this way, Ian Stewart's work [9] considers where, and he analyzes the many possible types of media which can be used for popularization. He focuses on magazines, newspapers, books, radio, and television, but the internet is missing, so, eight years later, Etienne Ghys [1] focused precisely on the role of the internet. A general perspective was discussed in a panel directed by Günter Ziegler [11] and seeking the same goal is the purpose of the first work on this subject presented at ICM [8]. Finally, the other two articles are mainly concerned with one of those big questions, and so [3] tries to give clues about "what" and [6] is focused on "who."

In this work, I am going to try to answer some of those big questions, but mainly treating the "why" one. So, firstly I will tell why I started to work on math popularization and why I think it is important, adding a couple or reasons to those more commonly given by general institutions. For instance, a committee of the British government on strategies for education (see http://nationalstrategies.standards.dcsf.gov.uk/node/16073) addresses the importance of mathematics in society:

> *"Mathematical thinking is important for all members of a modern society as a habit of mind for its use in the workplace, business and finance, and for personal decision-making. Mathematics is fundamental to national prosperity in providing tools for understanding science, engineering, technology and economics. It is essential in public decision-making and for participation in the knowledge economy."*

But those arguments apply only to institutions, not to individuals. Hardly anyone works in any subject, even less in one as math popularization with a difficult rentabilization from the point of view of the academic career, with those high targets in mind. But, something interesting I have learned from my personal perspective is that we can add some additional arguments to those usually given.

The structure of this paper is as follows. Firstly, in the next section, I will tell how and why I started doing math popularization and will show my personal point of view of the other general questions we have mentioned before. I think it is important to point out that one of my main concerns is how to bring mathematics closer to women and girls, and the other way around: how to bring women and girls closer to mathematics, and why I think this is important. Section 3 reports an unexpected benefit of my work in math popularization and provides an additional reason to those usually given by institutions, since we have found how that task is important in order to obtain new results in basic science, even more important, in a very multidisciplinary work. We will finish with some conclusions and plans for future works.

2. WHO, WHAT, WHERE, WHEN, WHY, AND HOW (MY PERSONAL VIEW)

Several reasons have been given to encourage science popularization in general, and mathematics in particular. The rapid development and extensive application of science and technology since 1870 not only promoted a profound transformation in economy and society, but also deeply impacted the mode of production, people's lifestyle, and the basic relationship between science and the public. But it is difficult for the general public to understand the role of mathematics in that development, so almost all institutions try to impel the spreading of a certain knowledge of that role in people's lifestyle. But, I have to confess that I did not get up one morning thinking "mm, I must spread my math knowledge because that is important for society." Not at all, In my case, after obtaining my PhD degree (with a thesis in Computational Geometry), I was lucky enough that a well-known publisher contacted me and so I, and my advisor, embarked into transforming the thesis into a book for that publisher. In the meantime, I continued with my research, trying to publish some papers and sending works to conferences. This is to say, I did not care about math popularization at all; at least, as an actor. So, what was the starting point of my career in this field? Well, I have to confess that after the book two kids came, and with them a lot of questions (after a while, of course). Many of those questions referred to mathematics, simply because their dad and mum are mathematicians. I tried to answer those questions, and their dad encouraged me to write those answers in a personal blog I kept at that time. Nobody read that blog before that entry, but, suddenly those explanations became quite popular and I got thousands of clicks. In one way or the other, that post was read by an illustrator (Raquel García Ulldemolins) and, simultaneously, a very popular blog in that time asked me to collaborate with them. Someone suggested that Raquel and I could make a tandem, so, I wrote the text and she painted some illustrations. In this way, "Mati y sus mateaventuras" (Mati and her mathadventures) was born.

The structure of all the posts in that blog is similar: two brothers (Sal and Ven) are arguing about any subject and then, they meet their friend Mati and she shows then that the

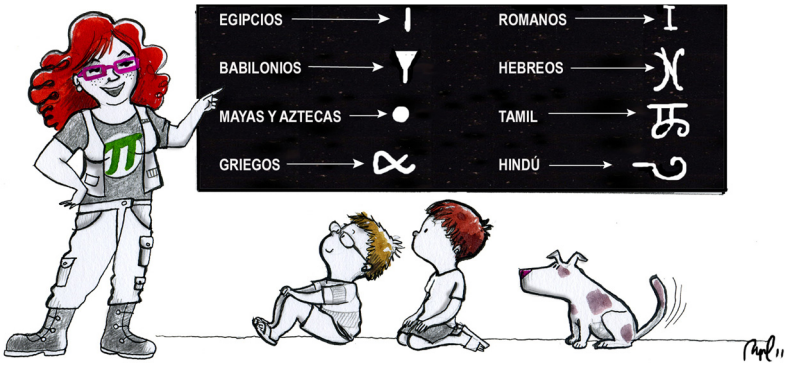


FIGURE 1
One of the illustrations in the first post of "Mati y sus mateaventuras".

subject is full of mathematical facts; and with them their dog (Gauss is its name) is always given a humorous counterpoint (see Figures 1 and 2).

I have to say that this blog had quite an impact and obtained several awards. Initially, we had in mind young people as the target of our stories, but we soon realized that a lot of parents and, especially, teachers started to use the ideas we were showing there as an inspiration or as a pedagogical material for their families or in their classrooms. So, right after the blog, many visits to schools came and, after that, talks in many different places (universities, secondary schools, science museums, even bars, and cafeterias), radio, books, television, etc.

Thus, after all this activity, I can say that my main professional activity in the last ten years has been math popularization. I think in all this process I have learned some things about this discipline. Firstly, if we assume that math popularization is important, it is crucial the support of the institution but, at the same time, we have to try to carry our message to traditional media because a good portion of our possible and desirable target does not access yet to the new (or not so new) places where an important part of the popularization activity is done (around internet, mainly). On the other hand, the institutions must create the adequate climate to foster the work of those who decide to dedicate a part of their time to try to convince the society that mathematics is important (as all the other branches of knowledge, of course). I am not an expert on that side of the equation, but three things can be done: firstly, universities must have their own units in order to organize activities open to the public, with a clear plan ahead and with measurable milestones; additionally, those units must help the members of the university doing actions in this field by giving technical support. Secondly (and in some way, related to point one), it is important to dedicate funds to these tasks. On many occasions, the people doing math popularization do it at their own expense. Lastly, the elements needed to value this work must be established. Not a long time ago, we did math popularization without telling our departments we were doing such a task after work, fearing that other members of those departments could think we were wasting our time.

But now, I think we can focus on the "what" side of popularization. Again, at least from my point of view. Of course, it is impossible to talk about "what" without taking into account the audience ("Who" is your public). It is not the same to give a talk in a bar to adult people as in a school to six-year-old students or a twenty minute slot in a radio show. Nevertheless, there exists a rule of the thumb for any activity: we have to keep in mind that, with very few exceptions, our audience is not very fond of mathematics and, in many cases, they think that the discipline is useful, but they cannot give examples of its usefulness other than elementary arithmetic calculations. So, our central thread must be an application of mathematics, if possible a very unexpected one. Or, at least, to show a puzzle challenging enough to engage and inspire your audience. Then, with the excuse of solving the application we have presented (or the puzzle), we can (we must) the beautiness of the involved mathematics. Of course, the application does not need to be a crucial one for humanity. For instance, in one of the posts, I wrote about one of the few mathematical articles that appeared in Nature [5]. In that work, the author tries to find, under different criteria, the best way to lace the shoes. Or in another one, I reproduce the simple computations needed to solve a very important problem: how to leave the toilet seat after using it (males mainly). And an obvious answer is "clean," of course, following this, more academic work: https://www.scq.ubc.ca/a-game-theoretic-approach-to-the-toilet-seat-problem/. Well, probably those two examples are, indeed, crucial for humanity.

In any case, it is clear, I think, that "What," "Who," and "How" are closely interrelated (and even "Where"), and so every time we try to communicate something about mathematics, we have to keep in mind those three factors and to adapt our message under those conditions.

A side note about "How." In my case, I always use the same style, with the Raquel García Ulldemolins's illustrations (with the exception of the radio, of course) and a certain naive touch.

2.1. Why mathematics?

Briefly, I would like to point out three reasons (among hundreds) to try to answer the question "Why is it important to make math popular?". At least, those are the factors I have in mind when preparing any action on this subject, especially giving a talk or arranging any other activity in schools.

Firstly, to fight against some myths ("mathematics is only for few people (and they are nerds)", "common people only need to know elementary arithmetic operations," "I am not fit for mathematics", etc.). The main problem here is that in an elementary school (and even in a secondary school) in many countries, mathematics is taught just as a tedious repetition of some computations without putting them in context. So, it is important here to show how mathematics is present in many facts of everyday life, and why having a certain knowledge and understanding of mathematics can help when we make some decisions.

Secondly, for the students, the analytic mind that we can train with the proper problems in mathematics can help in almost all the other subjects. And finally, I have noticed the positive value of a talk in those gifted students (and in many cases, they do not know about

FIGURE 2
After some time, the characters in "Mati y sus mateaventuras" were changing a little bit, but preserving the same flavor.

their capacities). There are some studies in this area and, for example, in [10], the authors show, as they pointed out in the abstract, that within the same society, adolescent students who specifically lack mathematical education exhibited reduced brain inhibition levels in a key brain area involved in reasoning and cognitive learning. Importantly, these brain inhibition levels predicted mathematical attainment 19 months later, suggesting they play a role in neuroplasticity. And their study provides a biological understanding of the impact of the lack of mathematical education on the developing brain and the mutual interplay between biology and education. We will see something about the mutual interplay between biology and mathematics later.

2.2. Where are the women?

The second half of the last century has seen an increase of women with mathematical degrees, until reaching the 50% level in many countries and even more than that, but we have seen that this process has been reversed in this century. As Figure 3 shows, the turning point was around the beginning of this century and it coincides (causality or not, we as mathematicians know pretty well that correlation does not imply causality) with the moment when data, information, algorithms, and many other subjects related with mathematics entered our common life with a high demand for mathematicians in many companies. In other words, this happened when teaching became not the first option after obtaining the degree.

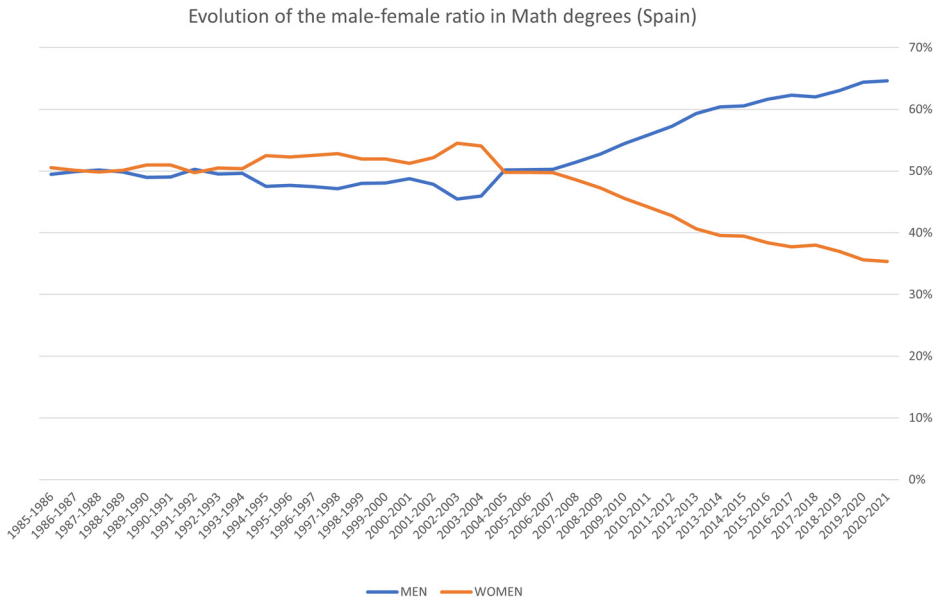


FIGURE 3
One can observe the effect of "scisors" in the evolution of the male–female ratio in math degrees in Spain. The line representing women grows until the year 2000, and then it decreases.

I think it is worthy to copy what is said[1] about this problem and its possible solutions:

> "Male scientists outnumber females two to one, [...]. According to the National Girls Collaborative Project, 'Women make up half of the total U. S. college-educated workforce, but only 29% of the science and engineering workforce.'
>
> Girls are interested in math and science when they're young, but they're often diverted before high school and eventually declare their college major in another field. An American Association of University Women study of 1,226 female science professionals found that girls actually demonstrate interest in science at a young age, but are discouraged due to antagonistic, critical behavior in many math and science departments. Nearly 40 percent of respondents indicated experiencing such behavior.
>
> Maybe the problem isn't gender-based but in the way children's skills are fostered. Math Professor Mary Beth Ruskai argues that both boys and girls need more interactions with scientists to become interested in science. Schools should also identify and encourage students' talents, regardless of academic field. Educational reform efforts often yield increased retention rates for both males and females, simultaneously combating two problems.

1 https://www.learningliftoff.com/encourage-girls-math-and-science-courses-and-careers/.

Here are a few ways teachers and parents can keep that spark of interest in science going for young girls and encourage more women scientist in the future:

1. Create Projects Based on Interest

Instead of letting girls' math and science interests lie dormant or go ignored, we should present them with science projects based on their interest. Sometimes, all it takes is one successful project to give a girl the encouragement she needs to find her passion in math or science. Even some toys for young children can aid in kindling an interest in science.

2. Introduce Female Math and Science Role Models

The U. S. Department of Commerce reports that women hold only 24 percent of STEM occupations, and those with a STEM (Science, Technology, Engineering, and Math) degree typically work in education or healthcare. While the numbers seem bleak, it presents an opportunity for change. Females working in or holding degrees in math or science should serve as role models for girls seeking a career in their field. Introducing a positive role model of the same gender to young girls can keep them interested and have a lifelong impact on their career paths.

3. Emphasize the Positives

Parents and educators should encourage girls to defy the stereotypes that math and science are only for boys. Like any subject, if girls are struggling in math and science, teachers should help them work through their struggles. This can mean playing an active role in helping them better understand these subjects. Just because a female student finds the subjects difficult is not a reason to move away from the field entirely. Working through challenges is part of the learning experience. Confidence plays a large role in a girl's success in science and math, and it's important to help her maintain a high level of confidence.

4. Explore Career Options Early

Often, kids in elementary and high school are unaware of the myriad career options that will be available to them as adults. Many of these careers will require a broader background in subjects they may not have considered or cultivated an interest in. But if they are able to explore career interests early, they can better prepare for them by taking classes they might have avoided otherwise. A student may want to be a doctor or a veterinarian, for example, but not be aware of the important role that science will take for such careers. And some job fields, such as computer coding and programming, encourage students to begin training in high school and even elementary school to be truly competitive. If career education courses are not offered in your child's school, consider an alternative school choice such as an online career academy. Destinations Career Academies and Programs combine traditional high school academics with career education.

In short, science and math are not gender-specific fields, yet girls seem to tune out natural tendencies toward these subjects. We can change their atti-

tudes toward math and science by offering them encouragement, role models, and opportunities to learn, tapping into their innate scientific and math skills."

I think that we, from an external point of view, can relate to those four points, and it is a good guide to follow.

3. AND, SUDDENLY, FLIES

In this section, I am going to try to show an additional benefit of math popularization, based on a personal experience. This benefit is not for the general public, for society, but mainly for other researchers. I will try to show how the popularization of mathematics can help in some interdisciplinary fields, building bridges between other sciences and mathematics, and fostering new and important results.

3.1. Voronoi diagram and 2-dimensional epithelial tissues

As "The New Yorker" says in an article,[2] "One of the many mysteries of living cells is how they manage to blossom into coherent many-celled units. A person or a platypus begins as a single cell, which divides into more cells, which also divide and subdivide. Some of these, the epithelial cells, are destined to become tissues and organs. The cells collect into layers, which bend and fold into greater-than sums: ovaries, kidneys, a heart. In part, it's a packing challenge, a geometry problem; as the layers twist and curve, the individual cells change shape in accordance with the whole, and they do so as efficiently as possible."

So, the research group of Luisma Escudero was focused on the problem of describing the shape of the epithelial cells. In fact, they had made some progress in the 2-dimensional case (when the cells are very flat). And their result was based on the Voronoi diagram. According to Wikipedia, "A Voronoi diagram is a partition of a plane into regions close to each of a given set of objects. In the simplest case, these objects are just finitely many points in the plane (called seeds, sites, or generators). For each seed, there is a corresponding region, called a Voronoi cell, consisting of all points of the plane closer to that seed than to any other." In this way, Luisma Escudero and his group modeled the shape of those 2-dimensional cells using the following method.

Firstly, they generate a set of random points in a plane region, then they compute the Voronoi diagram of that set of points, this can be considered the iteration 0. For next iterations, they compute the centroid of each Voronoi region obtained in the previous iteration and compute, again, the Voronoi diagram of that set of centroids. In this way, they conclude that a good model for this kind of tissues is obtained after five iterations. This means that by comparing some parameters (number cells with a given number of neighbors, quantity of some structures, etc.) of a sample with their model tissue, they can check if the sample corresponds to a healthy individual or if it presents some problem [7] (see Figure 4).

2 https://www.newyorker.com/science/lab-notes/we-are-all-scutoids-a-brand-new-shape-explained.

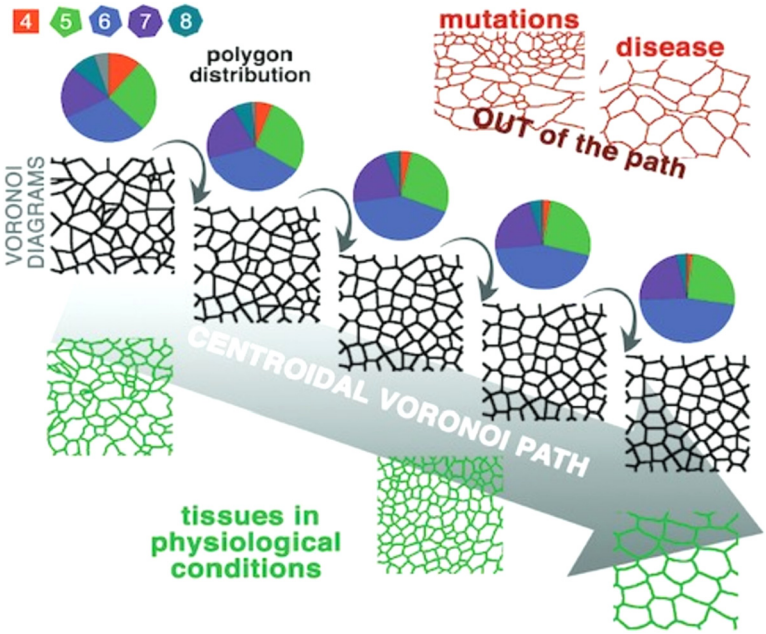


FIGURE 4
A synopsis of the results obtained in [7].

But, as animals develop, tissue bending contributes to shaping the organs into complex three-dimensional structures. However, the architecture and packing of curved epithelia remains largely unknown, and it was known that the results for the 2-dimensional model are not valid anymore. Thus, the transition from planar epithelial sheets to cylindrical, ellipsoidal, or spherical forms involves a fundamental reorganization of the cells.

It must be known that an epithelial tissue must be thought as a sheet and all the cells in that tissue appear on both sides of the tissue (called the apical and basal surfaces). Thus, in all the books of Cellular Biology, those cells are represented by prisms or truncated pyramids. But a close examination under microscope shows that the neighboring cells are not the same on the apical and basal surfaces. So, another model for those cells was needed.

3.2. Scutoids and 3-dimensional epithelial tissues

Of course, the first idea is to consider some variation of Voronoi diagrams, but the problem is that the model must predict what really happens in epithelial tissues, and Escudero's group was stuck. Then, he read a couple of articles I had written for one of the most important platform for science popularization in Spain (Naukas) about Voronoi Diagrams (https://naukas.com/2011/12/23/cada-uno-en-su-region-y-voronoi-en-la-de-todos/ and https://naukas.com/2012/01/28/esta-voronoi-que-se-ponga/) and he decided to contact me. That was the beginning of a beautiful (and fruitful) collaboration.

After some failures, we finally modeled the scutoids, following the following steps:

Regarding the space, we start with a given surface S, then for each point $X(u.v) \in S$, we consider the normal vector to S at $X(u, v)$, $N(u, v)$. Thus, for each $\lambda \in [0, 1]$, it is possible to define a new surface S_λ parallel to S in such a way that any point of S_λ is $X_\lambda(u, v) = X(u, v) + \lambda N(u, v)$ (a point in one of the surfaces has an equivalent point in each one of the other parallel surfaces). The metric in each surface previously defined is just the distance of the shortest geodesic on that surface joining two points. As it is well known, in the case of the cylinder, the geodesics are the helices in the cylinder.

We define every seed starting in a point on the apical surface. That point defines a segment between the basal and apical surfaces by means of its normal (given the point $X(u, v)$, the segment is $X(u, v) + \lambda N(u, v)$, $\lambda \in [0, 1]$). The intersection of these line segments with a given surface determines a seed. Thus, in order to generate all the seeds, in a first step we had chosen n points on the apical surface, then the n segments that were generated by them, and, finally, the intersection of those segments with every surface S_λ defined the seeds for that surface.

The next step is to compute the Voronoi diagrams of the seeds obtained in each one of the parallel surfaces. We linked the Voronoi regions corresponding to the seeds on the same segment, obtaining a three-dimensional figure, which we called a *scutoid* (see Figure 5).

3.3. Almost famous

After the publication of our results in Nature Communications, we made an important work in the dissemination of its results and scutoids appeared everywhere. Of course, in the most important media such as "New York Times," "The New Yorker," "The Times," as well as in the news of most of the TV companies (BBC, CBS, Fox, etc.), and in some shows as "The Late Show with Stephen Colbert." This led artists, designers, architects, and engineers to produce some works in their fields based on the shape and properties of the scutoids. In order to achieve this success, we sent press releases, called press conferences, and, in general, we talked about scutoids everywhere and everytime.

4. THE SHOW MUST GO ON. CONCLUSIONS AND FUTURE WORK

Although the beginning of my work in math popularization started in a unplanned way, it has reached such a volume that, I think, I must concentrate on the things I believe are more productive: firstly, encourage girls to study mathematics or any other science, or obtain a technical degree, and secondly, try to spread the usefulness of mathematics to a very broad public by using some media, such as the radio, that are not so popular in math popularization (it is quite a challenge to describe the shape of a scutoid if we have no images).

But, just as at the beginning, I really do not know what I am going to do in five years from now. Let us see.

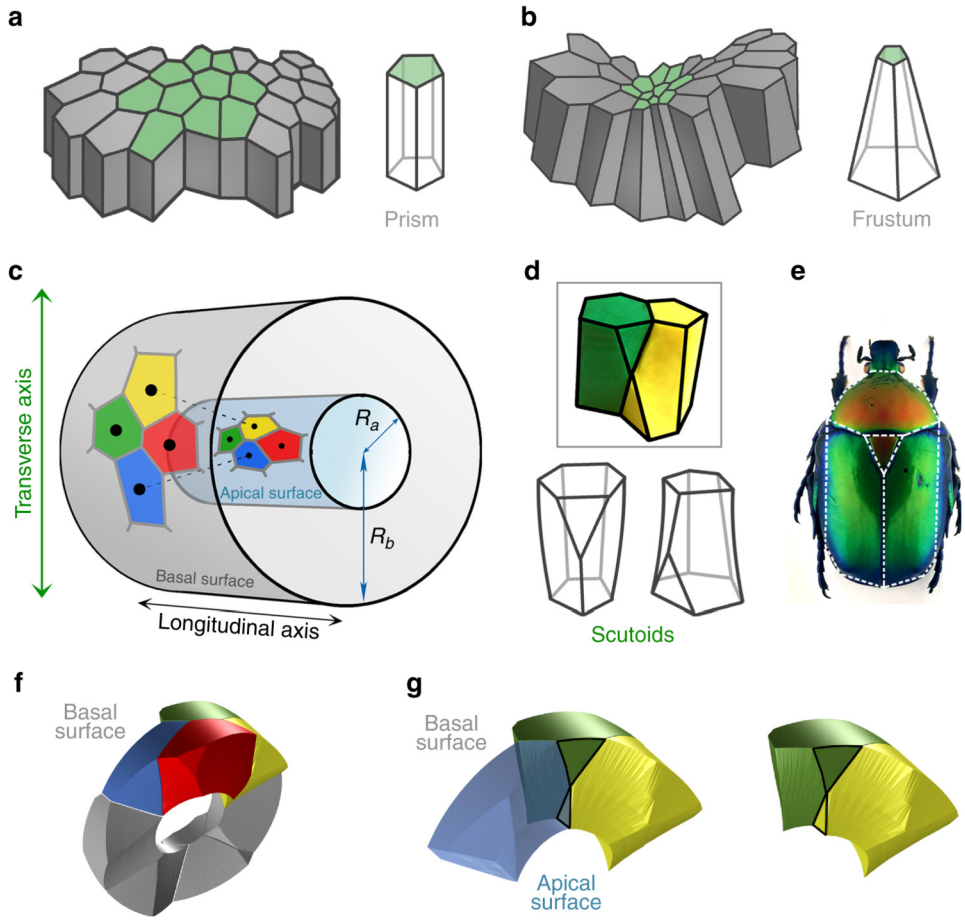


FIGURE 5
A graphical synopsis of the results obtained in [2].

ACKNOWLEDGMENTS

I thank Alberto Márquez for his help in this work, for his help in everything, for teaching me almost everything I know about math, for being an important piece of my life, for caring about me.

FUNDING

This work was partially supported by Secretariado de divulgación científica y cultural (Universidad de Sevilla) and projects PID2019-103900GB-I00 and P18-FR-631.

REFERENCES

[1] E. Ghys, The internet and the popularization of mathematics. In *Proceedings of the ICM, Seoul*, pp. 1187–1202, 2014.

[2] P. Gómez-Gálvez, P. Vicente-Munuera, A. Tagua, C. Forja, A. M. Castro, M. Letrán, A. Valencia-Expósito, C. Grima, M. Bermúdez-Gallardo, Ó. Serrano-Pérez-Higueras, F. Cavodeassi, S. Sotillos, M. D. Martín-Bermudo, A. Márquez, J. Buceta, and L. M. Escudero, Scutoids are a geometrical solution to three-dimensional packing of epithelia. *Nat. Commun.* **9** (2018), no. 1, 2960.

[3] V. Hansen, Popularizing mathematics: From eight to infinity. In *Proceedings of the ICM, Beijing*, pp. 885–896, 2002.

[4] E. Marín-González, D. Malmusi, L. Camprubí, and C. Borrell, The role of dissemination as a fundamental part of a research project: Lessons learned from sophie. *Int. J. Health Serv.* **47** (2017), no. 2, 258–276.

[5] B. Polster, What is the best way to lace your shoes? *Nature* **420** (2002), no. 6915, 476–476.

[6] C. Rousseau, The role of mathematicians in the popularization of mathematics. In *Proceedings of the ICM, Hyderabad*, pp. 723–738, 2010.

[7] D. Sánchez-Gutiérrez, M. Tozluoglu, J. D. Barry, A. Pascual, Y. Mao, and L. M. Escudero, Fundamental physical cellular constraints drive self-organization of tissues. *EMBO J.* **35** (2016), no. 1, 77–88.

[8] J. Schneider, Issues for the popularization of mathematics. In *Proceedings of the ICM, Zurich*, pp. 1551–1558, 1994.

[9] I. Stewart, Mathematics, the media, and the public. In *Proceedings of the ICM, Madrid*, pp. 1631–1644, 2006.

[10] G. Zacharopoulos, F. Sella, and R. Cohen Kadosh, The impact of a lack of mathematical education on brain development and future attainment. *Proc. Natl. Acad. Sci.* **118** (2021), no. 24.

[11] G. Ziegler, Communicating mathematics to society at large. In *Proceedings of the ICM, Hyderabad*, pp. 706–722, 2010.

CLARA I. GRIMA

Departamento Matemática Aplicada I, Universidad de Sevilla, Sevilla, Spain, grima@us.es

THE LONG WAY FROM MATHEMATICS TO MATHEMATICS EDUCATION: HOW EDUCATIONAL RESEARCH MAY CHANGE ONE'S VISION OF MATHEMATICS AND OF ITS LEARNING AND TEACHING

ANNA SFARD

ABSTRACT

Mathematicians and mathematics educators are united by their deep care for mathematics. This said, they are sometimes like parents who have differing ideas about what is good for the child. To improve communication between these two communities, I am telling the story of my own transformation from mathematics to mathematics education. In this account, I explain why I was compelled to revise my vision of mathematics and how I eventually arrived at the "commognitive" conceptualization, according to which mathematics is an activity of telling stories that produce their own objects. This change of vision brought many insights about learning mathematics and about factors that may slow students' progress. I illustrate some of the gains that come with commognitive conceptualization by showing how this approach allowed my colleagues and me to come to grips with some learning-related phenomena that have long been puzzling mathematicians and educators.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 97; Secondary 97C20, 97C30, 97C50, 97C60, 97C70

KEYWORDS

Learning theories, cognition, discourse, development, identity

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5716–5745
DOI 10.4171/ICM2022/90

Let me begin with introducing myself: I am a mathematics education researcher, particularly interested in how people learn mathematics. If I am here, at the convention of mathematicians, it is because our two communities, that of mathematicians and that of mathematics educators, have something centrally important in common: for all of us, 'mathematics' is the keyword around which our professional activities evolve. True, mathematicians spend their time *within* mathematical universe investigating its objects, whereas my colleagues and I sit in school and university classrooms observing those who try to enter that universe. Yet, understanding mathematics is the basic requirement for both of us. Our two communities are also united in their deep care for mathematics. This said, mathematicians and mathematics educators are sometimes like parents who have differing ideas about what is good for the child. The main sources of our occasional disagreements, it seems, are our dissimilar perspectives. Mathematicians never take their eyes off mathematical objects, whereas educational researchers constantly vacillate between this abstract universe and the outside world, populated by human beings. When invited to this conference, I felt this may be a good occasion to take a closer look at similarities and differences of these two outlooks. Getting acquainted with your interlocutor's thinking, even if it is unlikely to turn into your own, is the necessary first step in bridging potential communicational gaps. The best way to do this, I thought, would be by reflecting on what changed in my and my colleagues' journey from mathematics to mathematics education.

For me, this trip began years ago. As I traveled, the view before my eyes evolved all the time, changing from time to time almost beyond recognition. Today, I consider myself as a mathematical insider-turned-outsider, or a participant-turned-observer. I believe that my first-hand experiences as a member of both research communities makes me well equipped for the job of explaining and justifying my current perspective. Retracing the events that transformed me and my colleagues from people-who-think-like-mathematicians into those-who-think-like-educators may do the job best. This is, indeed, what I intend to do in this paper. Mine will be a story of an evolving vision of mathematics and of the deepening understanding of how children and young people turn into mathematical thinkers – or fail to do so. As my narrative unfolds, please keep in mind that if I occasionally speak in the first person singular, it is not because I consider my own history as in any way special or unique. On the contrary, it is because of its being rather common that I find it worth telling. My perspective may not be the only one with some traction within the community of mathematics education, but it can be considered as generic, in that it reflects concerns and sensitivities common to most of those who teach mathematics. The resulting story, therefore, which is not unlike those many of my colleagues could tell, should not be read as an autobiographical exercise but rather as a general reflection on how answers to the questions of what mathematics is and how people learn may change with the change of the storyteller's perspective. As you go through the following pages, please remember that whatever I found in this journey was generated in a collective effort of numerous people.[1]

---

1 I cannot list here the countless encounters with colleagues and texts that contributed to the ideas to be presented in this paper, but I wish to mention the *Haifa Discourse Group*, whose role in this project has been central.

1. CONUNDRUMS THAT TRIGGERED THE TRANSFORMATION

The proper way to begin this "travelogue" is to mention those special events that got me started on the journey and then kept me going. I will present here just three out of the many formative occurrences that raised questions and made me think.

1.1. Why doesn't logic suffice to understand mathematics?

I was still in the middle school when, as I was reading Henri Poincaré's[2] seminal book *Science and method*, I came across a paragraph that gave me pause:

> *One ... fact must astonish us, or rather would astonish us if we were not too much accustomed to it: How does it happen that there are people who do not understand mathematics? If the science invokes only the rules of logic, those accepted by all well-formed minds, how does it happen that there are so many people who are entirely impervious to it?* [14, P. 47]

Poincare's words resonated with what I had been wondering about myself. My classmates seemed split into two groups: some students could clearly grasp mathematical ideas in no time, at a glance; others complained incessantly about their inability to make sense of what was going on in the classroom. The higher the grade, the sharper the split appeared. Those from the first camp, the fluent speakers of the language of mathematical symbols, wondered with Poincaré about the other students' imperviousness to the logic of this language; those from the group of nonspeakers could not understand how this language could ever be mastered.

I agreed with Poincaré about the puzzling nature of this difference and, like him, wondered how this split could be explained. Saying that mathematics, unlike other school subjects, is uniformly abstract did not satisfy me as an explanation. The word "abstract" has been offered as if it was clear what it meant, but was it? For most people, the term signals the intangibility of the mathematical universe, its being inaccessible to senses. But saying what abstract thing *is not* hardly solves Poincare's puzzle. Indeed, the question remains of why and how some people manage to get into this abstract universe despite of its intangibility; and what it is that keeps the rest of humanity behind its closed doors.

1.2. What is so complex about complex numbers?

The formative event to be presented now sharpened this latter question. It took place when I was already a graduate student in mathematics and served as a teaching assistant to a well-known mathematician specializing in mathematical logic. One day, I was briefing the professor about my recent classroom experiences: "The students could recite the definition of complex numbers, but they constantly complained about 'not understanding anything'

---

2 The French thinker Henri Poincaré is known mainly as a mathematician, but he was a polymath who made important contributions also to theoretical physics, engineering, and philosophy of science.

and not being able to cope with the tasks I gave them". And, indeed, these students' minds seemed to be going blank even in the face of problems that would have yielded to just properly applied definition. The professor seemed puzzled. And then, suddenly, he said: "Well, this may be merely a matter of the teaching method. If I was their tutor, I would just discuss the definition and show that it is free from contradiction and consistent with the axioms of a number field. This, I am sure, would have opened their eyes."

I knew intuitively that this simple solution had little chance to work. Just as verbal instructions for juggling would not suffice to make a person able to juggle balls, clubs, and rings, repeating the definition of operations on complex numbers would also be insufficient to make the learners able to juggle a complex number. One mathematician whom I interviewed years ago told me that he could act with only those mathematical objects that appeared to him as having a clear "physiognomy" [17]. This metaphor brought back the issue of abstraction, but this time, it made me zero-in on the idea of a *mathematical object*: whereas it was clear how one develops an image of a person, how does one accord a distinct physiognomy to a new mathematical object, such as a complex number?

All this seemed to constitute at least a partial response to Poincare's question: Only those seem to be doing well in mathematics who have their ways to work out for themselves a good sense of mathematical objects. It is the ability to "see" these objects as they are being juggled by the teacher that allows one to make sense of the teacher's movements; and this is the inability to imagine them that turns these movements into incomprehensible. This was an important insight, and yet, it left me with new questions. Above all, I was now wondering about what mathematical object is, where it comes from, and how it can be turned into "one's own."

1.3. Why cannot children see as the same what grownups cannot see as different?

I was already a beginning researcher in mathematics education when an encounter with two four-year old girls put me and my colleagues on the path toward an all-new vision of mathematical universe. The search began when one of my Masters' student got interested in young children's numerical thinking, which she decided to investigate by watching her four-year old daughter Roni and Roni's 7-month older friend, Einat, performing some numerical comparisons. The girls were presented with pairs of boxes with marbles and then asked "In which box are there more marbles?" It soon became clear that the children could count properly. With a little prodding, they also managed, in most cases, to produce proper answers. And yet, even their successful solutions were accompanied by actions and utterances that we found strange and difficult to account for [21]. The greatest surprise came when the girls faced the pair of boxes with two marbles each. Upon seeing the two pairs of little balls, Roni smiled and said: "In none." Visibly pleased with the girl's answer, the interviewer closed the conversation: "There are more marbles in none of the boxes? Right." And yet, Roni's father,

who watched the scene from behind the camera, was not yet fully satisfied. He asked for explanation, and the following conversation between him and his daughter took place:

1.	Father:	Why? Why do you say this?
2.	Roni:	Because there is [are] 2 in one, and in [this] one there is [are] another 2.
3.	Father:	So, this is why there is more in none of them? So, in both of them there is... what?
4.	Roni:	Two.
5.	Father:	And this is... more or less?
6.	Roni:	Less
7.	Father:	Less than what?
8.	Roni:	Than... than... than big numbers.
9.	Father:	Than big numbers? That means... If there is [are] 2 in one box and 2 also in the other, then what is there in the two boxes?
10.	Roni:	4.
11.	Father:	Aha. Together, there is [are] 4?
12.	Roni:	Yes.
13.	Father:	And in each box there is the sa...
14.	Roni:	Because it is between...
15.	Father:	I see. And there is the same [thing] in each box?
16.	Roni:	.....
17.	Father:	How many in each box?
18.	Roni:	2.
19.	Father:	Oh well...

At the first sight, what happened here, while quite amusing, could have been dismissed as too commonplace to merit a serious investigation: The little girl was unable to guess her father's intentions and did her best to satisfy his expectations by offering any guess she could muster. Anybody who has ever taught mathematics seems familiar with situations such as this. Yet, we were wondering about the futility of the father's multiple attempts to make his daughter use the expression "the same" (as, for instance, in "There is the same number of marbles in these two boxes"[3]). Why were they ineffective, in spite of their versatility? Why did even his "there is the sa..." (see turn 13 in the transcript), which left only one syllable to Roni's discretion,[4] fail to do the trick? And finally, why did his explicit formulation of the desired

---

3 The conversation was in Hebrew, where "the same" translates into an idiomatic expression "oto davar," verbally equivalent to "the same thing" ("the same" cannot be stated without being followed by "thing" or any other noun, such as "number"). Note, therefore, that to fulfill the father's expectation, Roni could use the generic "the same thing" rather than the more specific "the same number."

4 Father said "oto da...", which had to be completed to "oto davar." This single syllable would have also completed Roni's answer because it would have produced a more or less full sentence.

answer (15) leave the girl visibly bewildered (16)? Our own bafflement was not any lesser: Why was this simple expression inaccessible to this obviously intelligent girl in this task, even though, as had already been repeatedly demonstrated, she was perfectly able to use it in other contexts?

After much deliberation, we concluded that our 4-year old participants could not think about any two objects to which the words "the same" could be applied. Evidently, Roni's father wanted this expression to be referred to *numbers*, or *amounts* of marbles in the two boxes. But these two italicized nouns, both of them used by the adult as signifiers of mathematical objects, were nothing of the kind in the eyes of the children. This event sharpened our interest in the nature and origins of mathematical objects. Whereas the previous story was about learners who have not yet developed a sense of a new mathematical object, this one was about students who did not even suspect the existence of such an object. The question now was how to bring this object to their awareness. If the query regarded concrete material objects, the response would have been clear. Objects such as those investigated in physics, biology, or astronomy are pretty straightforward and can be experienced by a person through his/her senses, either directly or indirectly, even before her being able to say anything about them. But the case of mathematical objects is quite different. Numbers, functions, and derivatives, unlike stones, stars, and living creatures, do not wait for the learner out there to be first detected, and investigated only later. So, how to even start talking about such an object?

Let me summarize. This last event, as well as the previous two, although brief and seemingly unremarkable, can be called formative: all three of them made us realize that to teach mathematics we can no longer ignore the question of the nature and origins of mathematical objects. We now needed to confront foundational queries head-on. After the iterative process of proposing tentative answers, which we would then critically examine, put to empirical tests and reject or modify, a far-reaching change in our vision of things eventually occurred. In the rest of this paper, I tell the story of transformations that led us to our current conceptualization. For reasons to be explained latter, we call this framework *discursive* or *commognitive*. The commognitive way of thinking has been working well for us for some time now. It made us able to formulate an answer to Poincare's query, to explain what the learners needed in order to reconcile themselves with complex numbers, and to account for the fact that four-year old children do not consider the expression "the same" as applicable within the context of numerical comparisons. These answers, while probably not the only possible, helped us make sense of what we saw and gave rise to pedagogical decisions that subsequently proved themselves in practice. We thus hold to the commognitive vision, at least for now, fully aware that it may be replaced one day with another, potentially more powerful way of thinking about mathematics, its objects, and its learning.

2. WHAT CHANGED ON THE WAY FROM MATHEMATICS TO MATHEMATICS EDUCATION

In this part, I explain what commognition is, while also telling the story of how this framework came into being. In the beginning, our thinking about mathematics was shaped exclusively by our own first-hand mathematical experience. It then evolved in a series of decisive steps, the first of which was the recognition of the very need to engage with the onto-epistemology of mathematics. Next came a series of small conceptual earthquakes, some of which have been presented above. One after another, these events effectively shook and transformed our foundational approach. I will now present each of these transformations in some detail.

2.1. Recognition of the need to elucidate onto-epistemological foundations

Although deliberations on the ontology and epistemology of mathematics have a long history, meta-mathematical questions usually fail to attract those who actually investigate numbers, functions, and abstract algebraic or geometric constructs. Preoccupied with the study of mathematical universe, they have little patience for conundrums labeled as "philosophical." This unwillingness to engage with foundational issues may be accounted for in a couple of ways.

In some cases, the lack of openness toward a serious conversation on foundational issues comes in a form of a quiet certainty about the mind-independent nature of mathematics. According to thinkers known as Platonists, mathematical objects, although inaccessible to our senses, are as much a part of the mind-independent reality as are stars, trees, and computers. Questioning the origins of mathematical universe would thus be an idle game. Since the times of the eponymous Plato, this view has been voiced over and over again, and most recently was reiterated by some of the most distinguished mathematicians of our times. Thus, for instance, the logician Kurt Gödel stated that "Mathematics describes a non-sensual reality, which exists independently both of the acts and [of] the dispositions of the human mind" [7, P. 311]. René Thom, the founder of catastrophe theory, sounded even more categorical when he stated that "mathematicians should have courage of their most profound convictions and thus affirm that mathematical forms indeed have an existence that is independent of the mind considering them" [22, P. 695].

Another reason that has been keeping mathematicians from engaging in serious foundational debates has been the view, shared by many, that onto-epistemological questions are irrevocably ill-defined and thus cannot lead to verifiable, useful answers. To save yourself embarrassment, it is better to remain silent on these issues, and thus agnosticism may be the safest option. This, indeed, is the spirit of Bertrand Russell's famous description of mathematics "as a subject in which we never know what we are talking about, nor whether what we are saying is true" [16, P. 84].

But this widespread disdain for foundational issues may also be explained in another way. If mathematicians may allow themselves the luxury of ignoring onto-epistemological infrastructure of their research, it is because no foundational resolutions seem necessary to

investigate mathematical reality. Reuben Hersh and Philip Davis, two mathematicians turned philosophers of mathematics, speak explicitly about mathematicians' unwillingness to make a serious ontological commitment while stating, tongue in cheek, that "the typical working mathematician is a Platonist on weekdays and a formalist[5] on Sundays" [2, P. 321]. In short, theories on the nature and origins of mathematical universe seem as irrelevant to those who juggle mathematical objects as the theory of big-bang is to those who juggle balls, rings and clubs.

Well, some may doubt if it is really so. After all, the disbelief with which new mathematical objects have usually been greeted throughout history could usually be traced to uncertainties about the ontological status, and thus legitimacy, of these entities. On the face of it, this kind of problem should have prodded foundational reflections. Historical facts, however, undermine this claim. As explained by the British logician and historian of mathematics, Philip Jourdain, whenever "logically-minded men" objected to such "absurd" notions as a negative number and imaginary numbers, the struggle for the recognition was eventually settled not by rational argument but simply by mathematicians' stubborn application of the problematic entity and their eventual "getting used" to its presence. To put it in Jourdain's own words, "mathematicians simply ignored [the objectors] and said 'Go on; faith will come to you' ... So [the new objects] were used with faith that ... was justified much later" [10, PP. 29–30].

These days, the mathematicians' indifference toward the question of the origins and nature of mathematical objects spreads to education, and the foundational issues remain an elephant also in mathematics classroom. As long as I was involved in mathematical research myself, I was accepting this situation uncritically. My position changed, however, when I started introducing others to the world of mathematics. As explained above, I soon realized that without coming to grips with the sticky foundational questions I would not be able to address properly any of the conundrums I encountered while teaching. Taking exception with the agnostic attitude was the necessary first step on my way toward the kind of understandings that are indispensable for well-reasoned pedagogical decisions. Upon this realization, my colleagues and I began talking about things that, so far, went without saying. In the rest of this section, I present the insights gradually gained on these occasions, especially those of them that withstood empirical tests and have been deemed helpful enough to be retained as a part of our theory of learning mathematics.

2.2. Mathematical object as a *mode de parler* rather than a part of mind-independent reality

The story of our journey toward the commognitive conceptualization of mathematics will now be told as a series of three transformations that resulted from our foundational

5 Formalism, yet another school in the philosophy of mathematics, has been embraced, among others, by Gottlob Frege and David Hilbert. According to formalists, mathematics is, basically, a symbolic game – the art of manipulating "empty" symbols according to well-defined rules.

deliberations. The first of these changes was due to the doubt about the signifier–signified dichotomy. We decided that rather than treating mathematical objects as self-sustained entities, ontologically different from the discursive constructs used to "describe" them, it might be more useful to see them as mere fictitious interpretation of certain communicational forms.

What are mathematical objects? The Platonic stance implies that mathematical objects are entities in their own right, not to be confused with mathematical words, symbols, diagrams, and graphs, all of which play an only the auxiliary role of these objects' "representations" – the mere communicational means. Or, as stated by the French mathematician Alain Connes, "Conceptual tools [signs, representations] aren't to be confused with the mathematical reality itself" **[1, P. 182]**.

You do not need to be the declared Platonist, however, to live in the world of this signifier–signified dichotomy. The idea that words and symbols are mere avatars of the "real things" is entrenched in the way we speak. For instance, we make statements such as

The symbols 13*, XIII, and* $5 + 6$ *represent the same number*.

The expression x^2 *and the basic parabola represent the same function*.

The word *represent* appearing in both these sentences implies that there are two categories of things, one of which comprises the entities that constitute the proper object of mathematical conversation (in this case, these are the number called "thirteen" and the function called "quadratic", respectively), and the other one composed of signifiers – the communicational counterparts of the former (in this case, these are the symbols 13, XIII, $5 + 6$ and x^2 and the words "number" or "function"[6]). The message about the independent existence of numbers or functions is implied by the fact that, as indicated by these last two utterances, a single mathematical object can have many sharply differing representations.

Being inscribed in the expressions we use, and thus in the ways we think, the signifier–signified dualism is difficult to argue with. It is unlikely to become an explicit topic of conversation in the first place. If the issue ever caught my attention, it was because of questions I began asking myself when, as a novice teacher, I was charged with the task of ushering other people to the world of mathematics. Before I could start introducing my students to the concept of negative number, for instance, I had to resolve the problem: How to talk with the class about entities that cannot be shown, while also claiming that these entities constitute products of operations that the young learners considered so far as "impossible"? The textbooks I was using suggested extending the number line to the left of the zero with the help of a symmetric half-line, whose integer points would now be given the names $-1, -2, -3, \ldots$ I was skeptical. Will the students believe me when I try to convince them that calling a point on a line with a new name suffices to conjure an all-new mathematical object? Will I be persuasive while claiming that by this simple act of baptism I had brought into being something

6 Here, I the quotation marks in the expressions "number" and "function" signal that I am speaking about *the words*, not about what is signified by these words.

that these young people had always considered as nonexistent and even "not allowed"? And if I put the new symbol -3 to the right of the equality sign in the expression "$5 - 8 = \ldots$," saying "Now the operation $5 - 8$ can be performed and it gives a result," wouldn't they protest, asking what had been added in this act of arbitrary signification? While wondering about what is the point of all this, they will surely question our human power to conjure something out of nothing. Years later, when I got acquainted with a bunch of classroom studies on children learning about negative numbers, and especially when I also co-conducted one such study myself [18], I found out that all these fears were definitely justified.

The decisive turnaround in my implicitly Platonist vision of mathematics began taking place as a result of my encounter with the 4-year old Roni and Eynat. As I was deliberating on the children's inability to use the expression "the same" in the context of boxes with marbles, I realized: when I introduce a new mathematical object, such as number, to the conversation about what I see – in this case about boxes with marbles, – I add nothing. Rather, I am just changing the way I speak. Of course, there are reasons for this shift in my discourse, and in the longer run, this transformation is going to prove itself very useful. But the change in the way I talk is all that "introduction of a new object" may mean at this point to the uninitiated – to those who, like Roni and Eynat, are hearing the term "number" within this context for the first time. This change in the form of speech is bound to confuse a young person who cannot yet appreciate the prospective benefits of this move. This bafflement will be experienced not just during the introduction of negative numbers, but also when other types of numbers – fractional, irrational, "imaginary" or even the most basic one, the natural – enter the scene for the first time.

To explain, let me engage you in a thought exercise. Please, take a look at the four pictures in Figure 1. Although these images are very different from one another, we may still claim that they present the same person. What is it that justifies this last statement? The answer seems simple: the claim is true because a single person, Sigmunt Freud, served as the model for all of four them. If the four pictures seem dissimilar, this is because they were drawn at different times in his life.

FIGURE 1
What makes us say that these four pictures "present the same person"?.

FIGURE 2
What is it that is "the same" in these six pictures?.

Now, consider the six pictures in Figure 2. Here, too, we may speak in terms of a single thing represented in different ways: all these pictures represent the *number five*. But where is this common element, number five? The truth is that the only feature shared by the six figures is that whenever we count their elements, we end with the word "five." Thus, what makes these figures into "the same" is the five-word long *process of counting*, and not any common *object*, as was the case with Freud's pictures. Yes, only a shared procedure may become the basis for claiming a "sameness" of dissimilar figures. If this fact escapes our attention, it is because also in the case such as that in Figure 2, we use the form of speech that was applied, so far, only for stating the presence of a common object. Saming through common procedure rather than a common object brings results because of which the French mathematician Henri Poincaré defined mathematics as "the art of giving the same name to different things" (quoted in [24, P. 154]).

All this makes us aware of the fact that we are using number-word only *as if* they were names of some independently existing objects, in a metaphorical way. But metaphors have their entailments, and in this case, one of the metaphorical entailments is that such object as "number five" is "represented" in all these very different images here the way Freud was represented in the four photos. Now it became obvious why young children must be able to count long before they can speak about numbers as anything else than the sounds used in counting. Indeed, counting is probably where the very idea of the abstract object called "number" has its roots. Following this insight, we decided to investigate the processes of objectifying the operation of counting, with the term *objectification* to be understood as *a discursive transformation that makes us use mathematical words and symbols as if they signified discourse-independent objects*.

We soon realized that the change in the way of talking called objectification is a combination of two lexico-grammatical transformations. First, there is *nominalization* – the act of replacing lengthy portions of text with a single noun. This is what you do, for instance, when you replace the talk about counting with number-words used as nouns. This is also what happens when you transit from the proposition

(A) *If I extract a square root from x and raise the result to the third power, I get the same result as when I raise x to the 3rd power and extract square root from it.*

to the equivalent objectified sentence

(B) *The third power of square root equals square root of the third power.*

(Note that both propositions can be expressed symbolically as $\sqrt{x}^3 = \sqrt{x^3}$). The verb clauses from (A), "I extract square root from ..." and "I raise ... to the third power" have been replaced in (B) by the noun phrases "square root of ..." and "third power of ...," respectively.

The second component of objectification is *alienation*, that is, the removal of the human subject. Thus, in the example just given, the grammatical subject of (A) is "I," which implies that it is a human being who performs the operation given by the subsequent verb phrase "extract square root." In (B), it is the noun phrase "The third power of square root" that plays the role of grammatical subject. In result, (B) sounds as if it was speaking about a self-sustained entity that does its own thing, without an involvement of any human agent. Only when we adopt this impersonal form of speech, we also begin saying that the nouns or symbols "represent" the object.

It soon became clear to us that objectification is a common phenomenon, to be found almost everywhere, not just in mathematics. You build on the metaphor of object also when you use words such as "velocity," "energy," "identity," "class," "justice" or human "ego." And while the subsequent research taught us that the transition to this objectified form of talk is never straightforward or easy, it also made us aware of the reasons why so many people, in so many domains, are prepared to invest the necessary effort.

Why do we need MOs? So, why do we objectify, in the first place? What do we gain when making transition from talking about actions and operations to talking about objects? The theoretical and empirical scrutiny of what happens in this transition brought to our attention two beneficial consequences of objectification: first, it improves the effectiveness of communication by allowing us to say more with less; second, it widens the range of things we can do, and in particular, of practical tasks we can perform.

To make my first point, let me, once again, compare propositions (A) and (B), the first of them expressed as a story of a series of actions (extracting square root and raising a number to the third power), and the other as a description of properties of mathematical objects (of the square root, of the third power). One difference between the two is readily visible: the objectified statement (B) is much shorted, more concise, than its unobjecti-

fied equivalent, (A). Thus, this example clearly corroborates my first claim: objectification allowed us to express ourselves more briefly, whatever it was we wished to say.

To illustrate the compressing power of objectification in an even more dramatic way, I will engage you in the following *thought exercise*:

> *Suppose you cannot use number words "one," "two," "three,"... except in counting. How would you then present in words the general truths expressed in this equality:* $3 + 4 = 7$*?*

Let me explain: in your response, you are allowed to use the number words, but only as "empty" signifiers, that is, as just strings of letters or of phonemes. Thus, you can say: "I counted the marbles in this box and got 'five' as the last number word", but you cannot say "There are five marbles in this box." I suggest that you give some thought to possible answers before you read my own response below.

And here is my answer. Not allowed to say things like "There are four marbles in the box" or "4 plus 3 equals 7," I would translate the symbolic equality $3 + 4 = 7$ into the following statement:

> *If I have a set so that whenever I count its elements I stop at the word "three,"*
> *and I have yet another set such that whenever I count its elements I stop at the word "four"*
> *and if I put these two sets together,*
> *then,*
> *if I count the elements of the new set, I will always stop at the word "seven."*

This is a very long sentence. Without condensing it and similar ones into objectified expressions such as "$3 + 4 = 7$," or even just, in words, *three plus four equals seven*, how would we be able to develop mathematics at large, and its numerical algorithms in particular? This example shows with particular force how the discursive device called objectification impacts the efficiency of mathematical communication by compressing lengthy expressions into very short ones.

And now, let me substantiate the second claim, according to which objectification extends the range of things we can do. I will help myself with an example that may appear so familiar, commonplace, and simple that you may wonder why I even chose to deal with it. But this is exactly the point. The analysis of this seemingly trivial event will let you see things, of the existence of which you might have been always aware, but which you never scrutinized to see how and why they work. What we notice here can be extrapolated to even most complex cases.

The example is taken from one of our empirical studies, in which we observed young people performing tasks related to numbers. Consider the following conversation between the interviewer and the 18-year-old girl by the name Mira, who was asked to pay for an

imaginary purchase with real coins[7] that have been given to her beforehand:

1.	Interviewer:	You bought 3 cookies from me; each one costs 75 agoras. Now you have to pay me.
2.	Mira:	Three times 75 . . .
3.		150 plus 3 times 25. . . 75. . .
4.		150 plus 75. . . 225.
5.		Here you are: 2 shekels and 25 agoras [*while saying this, Mira passes to the interviewer two coins of 1 shekel, two of 10 agoras, and one of 5 agoras*].

Let us take a close look at what Mira did. While saying "Three times 75" (utterance 2), she translated the required operation on coins into the numerical operation, multiplication of number 75 by 3. She did it by mapping the concrete objects (specific coins) onto mathematical objects (corresponding numbers) and by matching physical operations on the former with arithmetical operations on the latter. Then, in steps (3), (4), and (5), Mira implemented the operations on the mathematical objects, obtaining the number 225.[8] It is only then that she returned to the coins and composed the actual payment. Thus, the conversation that began as one about concrete objects (cookies and coins) has become one about mathematical objects (numbers), and then went back to concrete objects (coins). To sum up, the monetary transaction was a brief drama in three acts, with the middle one, the act of *planning* the action of paying, resulting in the *mediating story* about numbers, "three times 75 equals 225".

It is noteworthy that in a simple case such as that presented above, the task could have been performed also in an unmediated way. Such unmediated action is exemplified in another episode from our study:

1.	Interviewer:	Now you have to pay me. You bought 3 cookies from me; each one costs 75 agoras. Please, pay me.
2.	Talli:	Each one is 75 agoras. . . [*while saying this, hands a coin of 50 agoras (1/2 shekel), two of 10 agoras, and one of 5 agoras to the interviewer*].
3.	Interviewer:	What did you give me?
4.	Talli:	75.
5.	Interviewer:	Yes, you mean half and?
6.	Talli:	20 agoras and 5. Ok. And a shekel [*passes a coin of 1 shekel*]. One shekel and 75. Inside the shekel there is a 75, so there is 25 more. So, here is half a shekel more [*passes the coin of 50 agoras*]. And that's it.

7 The coins are in *shekels* and *agoras*, Israeli monetary units corresponding to dollars and cents, or pounds and pennies. Note that in the last sentence of the conversation (see line 5), the number names 2 and 25 are but labels for coins: the coin of one shekel and the set of coins including 2 coins of 10 agoras and one of five, respectively.

8 She took 50 out of the three 75s and added them together (3), and then, in (4) she first multiplied by 3 the remaining 25s and then added the products, 75, to the 150 obtained in (3).

Here, the required payment was performed directly on coins: Talli simply passed three sets of 75 agoras one by one. No mediating story has been told here and the payer ended up without necessarily knowing the total price of the purchase.

Considering this last example, the question may be asked why we should ever bother about mediating actions involving mathematical objects. Well, whereas this kind of action may appear just optional in simple tasks with which one is closely familiar, other tasks may be unfeasible without it. When the payment is made in the direct, immediate way, one relies on her memory of specific sets of coins that compose different basic values, such as that of 75 agoras. Sometimes, one's repertoire of memorized sums may not suffice to compose the required payment. Even more importantly, unmediated way of acting is applicable only in familiar situations, in which the performer can be guided by her previous experience. In contrast, mediating story used skillfully in one situation, may be appropriate also for a less familiar situation, involving concrete objects of a different kind. Thanks to their universality, therefore, mathematical objects make a person able to act in situations that are new to them, that is, involve objects – concrete or abstract – upon which she has never operated before. Indeed, mediating mechanisms of the kind of those exemplified here are at work even when you perform most complex and sophisticated practical tasks, such as building bridges or computers, flying to the moon, or designing vaccine for corona. One story about a single mathematical object allows us to deal with multiple situation that, so far, have not been considered as having anything in common. To sum up, mathematical objects are powerful tools, which not only make communication effective, but also allow us to deal with ever-new situations and to engaging in ever more complex forms of activity.

How are MOs discursively constructed? The interesting feature of these tools, and more specifically of mathematical objects, is that rather than being applied readymade, they are being constructed as we go. To put it differently, we conjure mathematical object by talking about them. This may sound as paradoxical as saying that a hammer is being put together during, and thanks to, the process of hammering. Yet, this is how it is. I will now take a closer look at the way in which the on-the-run object constructions take place.

After defining objectification as a discursive transformation that makes us use mathematical words and symbols as if they signified discourse-independent objects, I pointed out to two discursive operations that produce the objectifying effect: nominalization and alienation. Alienation has been briefly explained above, and I will now focus on nominalization, the process of replacing portions of text with a noun. Let us take a look at the different ways in which nominalization can be attained.

One of these ways has already been exemplified: I have shown how processes of counting turn into mathematical objects called numbers. Brief utterances with words such as *two* or *five* used as nouns may now replace long statements about human actions, such as "when I count the sides of pentagon, I arrive at the word 'five'." This move of replacing stories of processes with stories of objects is called *reification*. Reifying is also what I do when instead of speaking about my own action of multiplying, as in the narrative "When I multiply odd number by itself, I get odd number," I tell a story of an object: "The square

of odd number is odd." And it is what I did above in transition from the proposition (A) to (B), when I disposed of *verb* phrases "extract square root" and "raise to the third power" appearing in proposition (A) and replaced them in (B) with *noun* phrases, "square root" and "third power." It should be stressed that reification is not restricted to mathematics. We apply it everywhere, even in everyday talk. I reify, for instance, when I replace the story employing the verb "move," as in "The antelope moves fast" with the one that uses the noun "movement," as in "The antelope's movement is fast."

Another nominalizing operation that may lead to the emergence of a new object takes place when we endow several different objects with the same name. As such, it may be called *saming*. Saming is what we do when we refer to things as different as, say, dog and cat with the same word, "domestic animal." We are saming in mathematics when we refer to both the expression x^2 and the curve known as parabola with the same name, "the basic quadratic function."

Finally, there is the operation of *encapsulating*, of replacing the plural form with the singular. This is what we do when instead of saying "The post-office workers *are* efficient," we declare "The post-office staff *is* efficient." Here, the word "staff," in singular, encapsulates "the workers," in plural. And in mathematics, we are encapsulating when, for instance, we replace the claim "The cubes of numbers *are* increasing" with "The function x^3 *is* increasing."

The following example, featuring the object called "the basic quadratic function," shows how these three operations, saming, reifying, and encapsulating, can be iteratively combined in the process of constructing a mathematical object. It is reasonable to conjecture that the idea of the quadratic function emerged when people realized that some stories about x^2 may be translated into narratives about the curve called "parabola" and also into those about a certain table – the one displaying a set of ordered number pairs, in each of which the second element is the square of the first. For instance, the claim that zero is the smallest possible value of x^2 can be translated into the story of the smallest second element of the pair and into one on the lowest point of the parabola.

The benefit of replacing all three signifiers, the algebraic expression, the parabola, and the table with the single term "basic quadratic function" is immediately obvious: this replacement allows us to make all these statements simultaneously, in the single sentence: "Zero is the smallest value of the basic quadratic function." Here, we used the new noun "function" to perform saming of the three original signifiers. Clearly, such saming makes our propositions incomparably more general, and thus more powerful, and it adds to the thriftiness of mathematical communication.

What we call "basic quadratic function" became a combination of three signifiers, which from now will be called *realizations* of the signifier "basic quadratic function." But the process of realizing signifiers with the help of other signifiers is recursive, and the three realizations of the basic quadratic function may themselves be realized by other signifiers. Thus, x^2 may be realized as a square of any specific number. It is obtained from these specific squares by saming. These square numbers, in turn, are reifications of the operation of multiplying numbers by themselves. Similarly, both the table and the parabola can be realized as

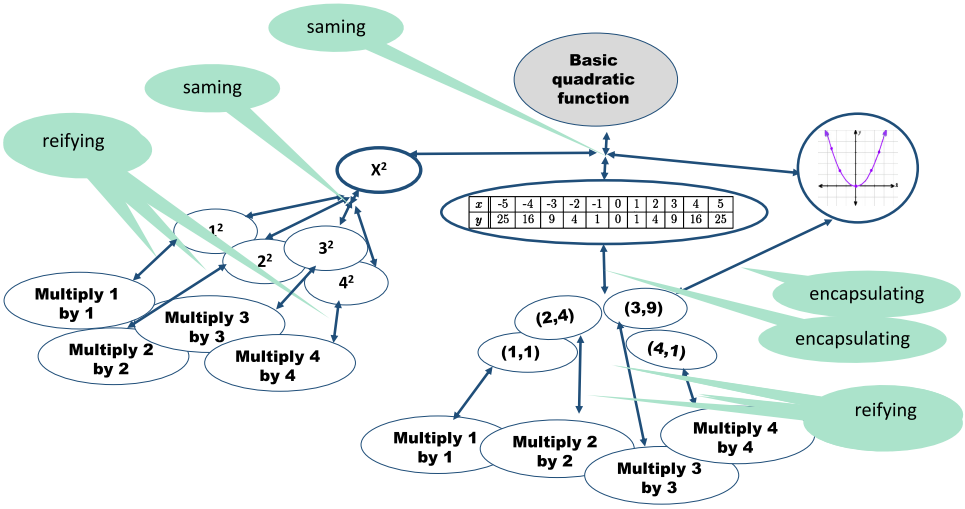


FIGURE 3
The realization tree of the signifier "Basic quadratic function" (adapted from **[20]**).

the set of ordered pairs of numbers and their squares, with encapsulation as the corresponding transformation. As before, these squares are reifications of the operation of squaring. The resulting diagram in Figure 3 presents the object called "basic quadratic function," composed of its name (the signifier) and its realization network (the signified).

Let me now summarize some of the insights, so far, about mathematical objects. First, the last statement of the previous paragraph can be generalized, and the *mathematical object* can now be defined as a *signifier*, such as word, written symbol or icon, together with its *realization network*. Here, the phrase "R is a *realization* of signifier S" is to be understood as saying that for a set of true proposition about S, there is an isomorphic set of true proposition about R (to avoid getting into technicalities, I will skip the definition of the relation of *isomorphism between sets of propositions*, but I hope the term is self-explanatory). Thus, the expression "x^2" is a realization of the signifier "basic quadratic function" and 3^2 is a realization of x^2.

Second, the set of all objects can be split into a pair of categories, and this can be done in two different ways. First, there is the distinction between *primary* and *discursive* objects – between those that exist in the world, independently from the human mind, and those that exist also, or only, in discourse. Second, there is the *concrete–abstract* dichotomy. These distinctions may be explained with the help of Figure 4, which shows in a schematic way how generations of new objects are being built, one after another, from those that precede them. The chain begins with concrete material objects, that is, those material things whose existence does not depend on whether somebody thinks or talks about them. These are the objects that are called *primary*. When a primary object is given a name or denoted with a symbol, we can start communicating about it. In this way, an atomic or elementary

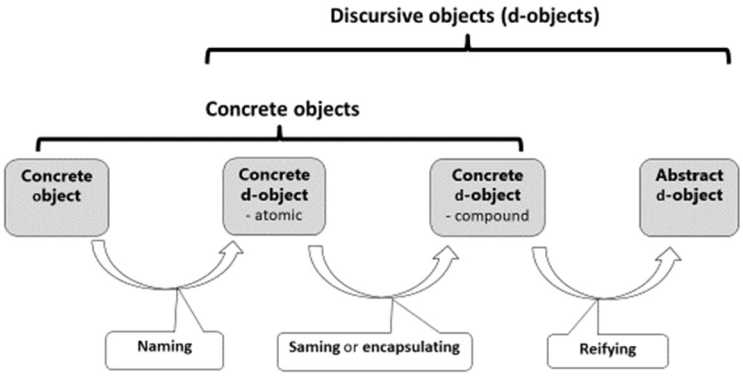


FIGURE 4
Types of objects (adapted from [20]).

discursive object, or *atomic d-object* for short, is created. In the third stage, some atomic d-objects are combined by saming or encapsulating into *compound discursive objects*. Finally, or perhaps in parallel to the stages of saming and encapsulating, additional compound d-objects are obtained by reifying processes that involve previously constructed objects. It is this last category that consists of objects called *abstract*. The other three, reification-free categories, contain objects considered as *concrete*.

Several conclusions follow from what was said so far. First, according to the above definitions, all *mathematical objects are abstract* because their construction involves reification, the operation that appears to be the act of adding a whole new entity but, in fact, introduces just a new figure of speech. Second, there is no ontological distinction anymore between signifier and signified. Any of the material means used for communicating – written or spoken words, visual devises, or touchable things – may serve as signifiers, and these are also the materials of which the signified, this dynamically expanding, never complete network of realizations is made. This means that all objects, whether primary or discursive, whether concrete or abstract, are basically material and accessible to senses, and the only difference between concrete and abstract entities is that in this latter case, the signified may be unbounded: it is always ready to accommodate new elements and is never perceivable in its entirety. Finally, discursive objects are personal constructs that develop gradually as a person learns mathematics. In this process, the realization network of signifiers such as "the basic quadratic function" or "rational number" is constantly expanding, sometimes deviating from the canonic version, accepted by the community of mathematicians.

The main idea to be taken from all that has been said so far is that mathematics is an *autopoietic* communicational system that creates its own objects while telling stories about them. Recognizing this inherently paradoxical nature of object construction is critical to our understanding of how mathematics is learned. Before I turn to new insights about learning that came to us with this recognition, let me add a few words about how this new vision of mathematical objects revolutionized our ideas about mathematics as an activity.

2.3. Mathematics: the activity of telling useful stories about reality rather than a search for the universal truth about it

It seems that the discursive nature of at least some mathematical objects has been already intimated by one of the greatest mathematicians of all times, Johann Carl Friedrich Gauss, who famously stated that "Infinity is merely a *façon de parler"* ([6, P. 216], quoted in [12, P. 337]). From referring to a particular mathematical object, this claim has now been extended to all of the abstract entities. An inescapable conclusion of this nondualist vision is that *mathematics is a form of communication, or discourse, that we adopt when constructing mathematical objects and telling potentially useful stories about them*. I will now unpack this assertion explaining what is meant here by the terms "story" and "useful."

Within this present context, the use of the colloquial word "story" may make some of you feel uneasy. I claim, however, that the word is in place in describing not only mathematics, but also all other domains of research, such as physics, biology, or history. Thus, for instance, a story about living organisms, such as "Plants convert light energy into chemical energy in the process of photosynthesis," is a typical output of research in biology, whereas the formula "$S = 1/2gt^2$" is among stories about bodies in motion told by physicists. Yes, also this last string of symbols, as unlikely as it may seem at the first glance as an example of a story, does turn into a narrative once we decode it and write it in words rather than symbols: "The distance S traveled by a free falling object is equal to half of the gravitational acceleration, g, multiplied by the square of the time of the travel, t." Similarly, mathematical equality $(x^2)' = 2x$ can be seen as a narrative about a function and its derivative. Of course, the three propositions I brought here as examples of scientific or mathematical stories present these stories in a highly condensed form. For elaboration, one needs to consult academic literature.

Let me complete my explanations by clarifying how the term "story" is to be understood in the present context. From now on, I will be using the expression *story about X*, where X is a noun, as referring to *a coherent sequence of utterances* (*propositions*) *that, when taken together, can be said to be "about X"* (or "on X" or "of X"). The "aboutness" means that X is the grammatical object or subject of some of the utterances in the sequence, and the sequence in its entirety is consistent and cohesive. The term "consistent" says that the sequence does not logically imply both a proposition and its negation. This term "cohesive" indicates the presence of lexico-grammatical links that hold the sequence together, that is, connect its successive utterances thematically. The connection may be chronological, as is the case when the successive utterances are linked with words such as "before," "after," or "next"; it can be logical, attained by the use of connectors such as "therefore," "it follows," "and," "or"; and it can be causal, expressing itself in the presence of words such as "because." Most current uses of the term "story" or "narrative" imply chronological interconnection of the different parts, and thus our present definition leads to a wider than the common application of the term.

The last question that needs to be answered in the attempt to complete the definition of mathematics as "the activity of telling potentially useful stories about mathematical objects" regards the term "useful." What does this adjective mean and why should it be pre-

ferred to "true," which is mathematicians' favorite? In the preceding section, I have already stressed that mathematical objects are useful in their roles of "compressors" of mathematical prose and of action-mediating devices. With their help, we are able to perform tasks that would not be workable otherwise. For the operations on mathematical objects to be truly helpful, we have to draw on what we have learned about their properties. In other words, stories about mathematical objects are those that guide our decisions about how to use these entities in problem-solving and in practical action.

Some of you may shrug at my mention of usefulness as the required feature of mathematical stories. Some mathematicians may share D. H. Hardy's conviction that their activities have nothing to do with "anything useful" [8, P. 150]. Yet, the majority of mathematicians seem to be of one mind with Andrew Forsyth [4, P. 35], who famously claimed that almost any mathematical story would eventually turn useful beyond mathematics itself, provided we have the patience to wait for a "real-life" problem that can be solved with its help. The message about potential practical usefulness of even most abstract mathematical ideas can also be heard in Alfred North Whitehead's disclaimer: "It is no paradox to say that in our most theoretical moods we may be nearest to our most practical applications" [25, P. 100].

Of course, not all stories come equal and not all of them can serve as reliable mediators of practical actions. Only those mathematical narratives are endorsed as reliable and potentially useful that have been constructed and shown to be endorsable with the help of well-defined communicational tools, that is, within a special discourse. This latter word, *discourse*, may be defined as referring to a communicational game that determines a community. Its game-like nature expresses itself in its being rules-regulated activity, similar in this respect to, say, the game of chess. It determines a community in that, like chess, it splits the humanity into those who are able to participate in this activity and those who are not (of course, the split is never clear-cut, but the idea of the "community of discourse" is useful nevertheless). It is important to remember that discourse may be in words, but more often than not, it is multimodal. Sometimes, mathematical conversation may take place just in sounds other than words, in body movement, gestures, facial expressions, pictures – any of these or all of them together. Mathematical discourse, as any other, can be practiced with partners or with oneself. In this latter case, the discursive activity it is called "thinking."

Different discourses are created for different types of mathematical objects, and they differ among them along four dimensions. The first and most obvious of the distinctive features is the set of *keywords* pertaining to the discourse's characteristic objects, such as the words "number," "one," "eleven," "sum," "product" in arithmetic, "figure" and "triangle" in geometry, and "function" in mathematical analysis. Most of these keywords come with explicit rules for use, known as definitions. Second, there are the characteristic *visual mediators*, that is, visual means with which one makes clear what it is she or he is talking about. Thus, in mathematical analysis we use algebraic expressions and curves known as graphs, and in geometric discourse we help ourselves with drawings of different shapes. Third, each of mathematical discourses has a well-defined set of communicational *routines*, the patterned, recurrent ways of doing things. Some of these routines are common to all mathematical discourses, whereas some others are discourse-specific. Among them, there

may be routines for reading mathematical notations and for operating on symbols, those to be applied in constructing stories about mathematical objects, and some others, to be performed in testing stories already created or in showing whether they can be endorsed. The routine used in this latter task is known as "proof." Finally, the discourse on X comes with a small set of endorsed narratives on X, known as axioms, on the basis of which other endorsed narratives on X will gradually be constructed. Together, all these endorsed narratives will constitute the *theory of X*. In natural sciences, a collection of narratives, to count as a theory, must be unambiguous, consistent with experience, general rather than specific, and this is only the beginning of the long list of requirements. In mathematics, on the other hand, at least in principle, consistency and cohesiveness are all that is necessary to ensure that a story be seen as a part of theory. Mathematicians strive to make their theories as complete as possible, hoping that for every proposition about X, either this sentence or its negation will turn out to be a part of the theory of X.

Viewing research, at large, and mathematics in particular, as communicational activities has an implication that goes against one widespread belief about mathematics, engender by its Platonic version: it is now clear that many seemingly competing theories, not just one, may be developed about the same X.[9] The phenomenon is well known from science – think, for instance, about Aristotelian, Newtonian, and Einsteinian theories of motion. To see that it occurs also in mathematics, one may consider the Euclidean and non-Euclidean geometries, each of which tells its own story of the construct called "space."[10] The different stories may sometimes appear to be contradicting each other, as is the case for the Euclidean, Bolyai–Lobachevskian (hyperbolic), and Riemannian (spherical) narratives about the sum of angles in a triangle. Here, the apparent contradiction stems from the fact that, in each of the discourses, the use of the basic keywords is defined with the help of a slightly different set of axioms. Some other examples that could be given here are much less obvious, simply because mathematicians agreed to opt for just one version that became canonic, with the others forgotten. This is what happened when integers were extended to rational numbers, and then when unsigned numbers were broadened to signed, or from real to complex. Within the nondualist approach to mathematics, therefore, unlike in the world of Platonic ideas, *the decision to label a narrative as "true" becomes relative to the discourse in which this narrative is told*. It is for this reason that the adjective "useful" may be a more appropriate descriptor for the basic criterion for endorsability than is the word "true" which, whether we want it or not, brings the connotation of universality. To forestall possible protests, let me immediately add that what has been said in this paragraph does not imply that mathematical

---

9 Keep in mind that X is a *noun* that points us to a certain phenomenon, rather than the phenomenon as such. The different discourses on X are likely to use this noun differently, and this entails differing narratives about X.

10 The fact that these three theories can be subsumed under a common metadiscourse may give rise to the assertion that they are parts of a single higher-level theory; this, however, does not contradict the claim that when taken separately, they constitute different theories of the same X, and that these different theories pertain to, or are useful for, different interpretations of the X.

"truth" (or endorsability) is arbitrary. Whereas we are free to opt for any properly constructed mathematical discourse,[11] once we make our choice, we lose our freedom to decide what can count as true. Within the boundaries of the chosen discourse, the veracity of narratives we are going to create will be uniquely determined by the rules and routines of this discourse.

Before concluding this brief introduction to commognition, it is important to stress that this approach, and more generally, our conversion from covert Platonists to overt nondualists did not come out of nowhere. It was inspired by many recent developments in several seemingly unrelated domains, with philosophy of science and learning sciences among them. On the one hand, we followed in the footsteps of leading thinkers of the 20th century who turned to communication as the key to understanding human uniqueness. The word "knowledge," signifying one of the hallmarks of humanity, has been interpreted by Rorty as referring to the "conversation of mankind" [15, P. 389]. In a similar vein, Foucault claimed that discourses are "things said. . . those familiar yet enigmatic groups of statements that are known as medicine, political economy, and biology" (see the blurb on the cover of [5]; mathematics can now be added to this list). This nondualist position with regard to knowledge, as observed at the level of humanity as a whole, paralleled the work of psychologists whose observations on individual human beings and on their cognitive activities was inspired by the ideas of the Austrian–British philosopher Ludwig Wittgenstein and of the Russian thinker Lev Vygotsky. In tune with Vygotsky's claims on the inseparability of word and its meaning, the writers who called themselves "discursive psychologists" started questioning the ontological split between thinking and communication [3, 9]. We have been encouraged by all these thinkers when we decided to view mathematical thinking as a self-dialogue involving the discourse known as mathematics. The unity of these hitherto separate ontological categories, cognition and communicating, is reflected in the portmanteau *commognition* [19].

3. HOW COMMOGNITIVE INSIGHTS ABOUT LEARNING HELPED TO SOLVE THE INITIAL CONUNDRUMS

Having introduced the nondualist way of thinking about mathematics and its object, I now have to convince you that the result was worth the effort. More specifically, I need to show that commognition is a powerful tool for making sense of what people do in their encounters with mathematics, and that it is more successful in this role than any dualist approach so far. I will do this by showing how the discursive conceptualization of mathematics helps us resolve the three conundrums that initiated us on our way toward commognition. I will now attend these conundrums in the order reverse to that in which they are presented above. On my way, I will discuss some of the more general changes brought by commognition to our understanding of what people do when they learn mathematics, what obstacles

11 We choose discourse according to the criterion of prospective usefulness, as it is measured by either its practical applications or by its power to generate a rich mathematical theory or, preferably, according to both these considerations.

they need to tackle on their way, and what may help or obstruct their efforts to overcome the hurdles.

3.1. Seeing as the same what so far appeared as different: the paradoxical conditions for objectification

Just to remind, the heroin of this formative event was the 4-year old Roni who, when faced with two boxes with a pair of marbles each, was unable to say what her father desperately wanted to hear: that there was "the same" number of marbles in the two boxes. This was puzzling because while opening each box, Roni could be heard saying the word "two" and then claiming that there is more "in none." Already when introducing this conundrum, I have raised an explanatory conjecture: for the 4-year old, there was nothing in the two boxes that could be called "the same." Now I can say that at this point, the young child evidently did not yet create for herself any abstract objects, mathematical or otherwise, that could be seen as being present in both boxes with two marbles and described as "the same."

This brief story gives rise to a much more general, and some may say quite unorthodox conclusion about sources of numerical thinking. According to cognitivist theories, produced in the mainstream psychological research, this kind of thinking is an inborn property of humans, with the first signs of "number sense" detectable already in newborns. Commognitive researchers do agree that some special human abilities, rarely found in other species, are necessary to make numerical thinking possible. As a good example, let me mention one ability that may well appear already at birth – the ability to distinguish between small sets of different cardinalities. Yet, once mathematical thinking is conceptualized as a *discursive* activity, the mere recognition of quantitative difference does not yet count as a case of mathematical thinking. According to commognition, mathematical thinking, *by definition*, does not exist before the child developed some uniquely human communicational skills. Note that this disagreement between the dualist and non-dualist visions of mathematical thinking is not just a matter of semantics. Indeed, the difference of opinion on the ontology of numbers has far-reaching consequences for our understanding of how this thinking emerges and how it develops later. Eventually, it is bound to affect our ideas about the ways in which children may be helped – or hindered – on their way toward numeracy.

To give just one example, let me consider yet another conundrum, one that has been challenging cognitive psychologists ever since the seminal studies by Jean Piaget. To put it in their own words, these psychologists have been puzzling over the fact that "children who know how to count may not use counting to compare sets with respect to number" [13, P. 35]. In this sentence, the authors summarized the phenomenon that has been observed time and time again: When presented with two sets of, say, marbles and asked "In which of them are there more marbles?", 4- or 5-year old children would not count even if they could. This, indeed, may seem puzzling to a person who considers numbers as self-sustained things which, like spoons or bicycles, can be experienced by children long before they are able to act with these objects themselves. And the puzzle may go, more or less, like this: The fact the children can count indicates that they are already familiar with the entities called numbers. Of course, they need some time to develop the routine of comparing-by-counting. But even

when they are already adept in this latter routine, why do they stay away from it when asked such question as "Where are there more marbles"? In our long conversations with Roni and Eynat, we observed this phenomenon many times [11, 21]. It was puzzling indeed, but only as long as it was described in this cognitivist language, which we too used at that time. The effect of puzzle disappeared when we began seeing number as but a reification of the discursive action of counting. The commognitive vision reversed the order of learning: the routine of comparing-by-counting, with counting understood at this stage as but an incantation (reciting number words in a constant order) comes first, and the idea of number as an abstract object emerges form it much later. Thus, as long as the child cannot actually do things with number words, there is simply no such thing as number. And even when she gains some mastery over the discursive operations of counting and comparing-by-counting, it must still take time until she reifies counting and stops seeing it as merely the favorite game of the grownups. All this seemed to solve, or rather resolve, the cognitivist conundrum: as long as the process of counting has not been reified, which seems to be the common state of affairs in 4- or 5-year olds, saying that children are trying to "compare sets with respect to numbers" makes no sense – and the puzzling disappears.

All that has been said here evokes also one important metacommognitive reflection. Our studies taught us quite a lesson about ourselves as observers of others. Events such as the latter one opened our eyes to the fact that one's own view of mathematics serves as a highly selective lens for seeing and understanding other people. We realized that unless we take precautions, we tend, as teachers or researchers, to attribute our own numerical way of thinking to those whom we observe, while also assuming that in the learner this thinking may be not as well developed as in an expert. This tendency comes to the fore when the dualistically-minded observer takes for granted that the questions she asked has been interpreted by the young participants according to her intention ("children compared sets *with regard to number*"). In result, when the child's performance does not meet her expectations, the observer tends to put the blame on procedural insufficiencies. She says to herself, "The child did try to do this, but she erred in the procedure." While stressing what is missing in children's actions, the cognitivist observer remains blind to what is actually there. In research, she does not even record the "strange" things children are actually doing in the attempt to cope. This oversight leaves her ignorant of the fact that children could be trying to perform a task quite different from that she had in mind. This is how the observer who thinks in dualist terms is misled by her own language and misses the opportunity to get a deeper insight into the meandering route the children travel before they become skillful participants of the canonic mathematical discourse.

3.2. The complexity of complex numbers: the need to reconcile yourself with the incommensurability between the old and the new discourses of numbers

Another puzzle left us with the question about difficulties students experience while learning about complex numbers. Why, we asked, in order to turn the learner into a skillful, competent participant of the discourse on complex numbers, does it not suffice to provide the definition of these numbers and then ask the learners to practice the well-defined operations?

Within the commognitive approach, one possible answer offers itself immediately: as in the previous case, we are talking here about the introduction of a new mathematical object, and as already stated, processes of objectification take time. Yet, although this statement sounds like answering our question, it leaves us with a new one: Why is the process of objectifying so demanding in the case of complex numbers? And more generally, what obstacles must the learner overcome on his/her way toward a new mathematical object?

Admittedly, not everybody experiences the task of objectifying as an uphill struggle. In some cases, the birth of a mathematical objects is recalled as an exhilarating event, an epiphany. Here is, for example, the story told by the topologist William Thurston:

> *I remember as a child, in fifth grade, coming to the amazing (to me) realization that the answer to 134 divided by 29 is 134 over 29... What a tremendous labor-saving device! To me, "134 divided by 29" meant a certain tedious chore, while 134 over 29 was an object with no implicit work.* **[23, P. 4]**

And Thurston continues: "I went excitedly to my father to explain my discovery. He told me that of course this is so, '*a over b*' and '*a* divided by b' are just synonyms. To him, it was just a small variation in notation" (ibid). Yet, as demonstrated in our examples, not every mathematics learner is as fortunate as Thurston. A closer look shows that the learners' difficulties may have several sources.

First, there is a certain circularity of requirements. If mathematical objects, such as numbers, whether natural or complex, are discursive constructions, then in order to build such an object one needs to talk about it. But to talk about it, the person must have already brought this object into being. And there is also another, slightly different circularity: the learner is unlikely to make the necessary effort without understanding its prospective gains. Indeed, she needs to be aware of the usefulness of the object she is trying to construct. But how can she comprehend its usefulness before she actually uses it?

Another objectification-hindering circumstance is the fact that what happens in the process of reifying may appear counterintuitive. Indeed, when you reify a mathematical process, such as that of extracting a square root from a number, and you write $\sqrt{-1} = i$, you claim that there is a product to the operation that has been considered so far as giving no result and was described as "forbidden." And now, who can say where and why this new number came from? It appeared with the introduction of the new signifier, "i." This new signifier reified the process of subtracting, but it did not add anything. This unlikely act of conjuring something out of nothing seems as counterintuitive (and difficult to digest!) as would be reifying a recipe for a cake and claiming that it constitutes the cake itself.

Objectification may have yet another counterintuitive aspect. To reify, a revolution in the rules of the game is sometimes required. This dramatic change may express itself in adopting a new way of building and endorsing new narratives, in changing how we think about familiar objects, and in disqualifying some of hitherto unquestioned truths. Thus, when complex numbers are to be introduced, some defining features of the object known as "number" may have to be abandoned. So far, numbers have been understood as what

answers such questions as "How many?" or "How much?" Each of them had a magnitude, and for any two of them it was clear which is "bigger." Not any longer. Also some previously endorsed stories must now be compromised. For instance, in the transition from the discourse of real numbers to that of complex ones, the narrative "Some polynomial equations have no solutions" is not true anymore. In spite of the apparent contradiction, the old truth and the new one are not mutually exclusive. They just belong to different discourses, because each one of them is using the word "number" in different way. Such two narratives are called "incommensurable" (as opposed to incompatible), and so do the discourses that produced them. Summing up, objectification projects back onto familiar discourses and transforms them, sometimes beyond recognition.

In the view of all this, it is not surprising that students may struggle to construct mathematical objects for themselves, and that they may take time to succeed. As long as the success refuses to come, they may have a considerable difficulty benefitting from what their teacher does or says. Obviously, the question now cries to be asked of how we can support the learners in their coping with all these hurdles. How to help them overcome the circularity and counterintuitiveness of objectification? A partial answer will be given below, when I show how commognition helped us tackle Poincaré's query. For now, let me just say that those who teach, having long forgotten their own past struggles, are mostly unaware of incommensurability between their own discourse and that of the learners. This was certainly so in the case of the mathematician with whom I discussed students' difficulties with complex numbers. The very awareness of the nature of the problem may take the teacher half way toward a solution.

3.3. The insufficiency of logic for understand mathematics? Some mathematical developments are a matter of choice, not of deductive reasoning

If mathematics "invokes only the rules of logic, those accepted by all well-formed minds, how does it happen that there are so many people who are entirely impervious to it?", wondered Poincaré while pondering on his own abilities as mathematician. As can already be seen from the former examples, commognition dissolves this puzzle by showing the falsity of its premise. Yes, according to commognition, the assumption that mathematics is the exclusive province of logic is untrue. Whereas logic wields the absolute power *inside* every mathematical discourse, the choice of the discourse is not a purely deductive act.

Let me elaborate. One of the implications of the commognitive vision of mathematics and its objects is that the growth of mathematics, whether historical or ontogenetic (in learning), involves two types of developments: adding ever-new stories about already existing objects and, from time to time, adding new objects and reforming the discourse. The first of these changes happens inside an existing discourse, whereas the other is metadiscursive: it is a transformation of the discourses themselves. We can thus speak about two types of learning that can be described, respectively, as object-level and meta-level. I will now argue that only the former kind of learning can be considered as just a matter of logic. Indeed, although mathematics is often described as a purely analytic discipline, that is, one whose narratives are constructed and endorsed exclusively on the basis of deduction, this feature

holds only within the boundaries of a well-defined discourse. Once a discourse is chosen, its rules, combined with those of deduction, uniquely determine how new endorsed narratives are to be derived from those that have been endorsed before, axioms and definitions included. Thus, as long as a skillful participant stays within the confines of a particular mathematical discourse, he can, at least in principle, produce new narratives and test their endorsablity independently, without being helped by others. In school, this is the situation for the learner who is already well acquainted with, say, the discourse on functions and is now supposed to explore properties of different families of functions.

The situation changes, however, when the student faces the need for meta-level learning. Here, in order to proceed, he will have to make the transition to a discourse incommensurable with the one he is coming from. Historically, this kind of transition is an outcome of mathematicians' personal choices – of their assessment of how useful or beautiful would be the results of following in one direction or another. To develop new mathematical discourse, they often needed to revise their shared beliefs on what should count as useful, aesthetic, and as "mathematically permissible." Clearly, these choices were not *dictated* by logic – they were a matter of contingency and of personal preferences rather than of necessity. Making such decisions required the ability to see mathematics as a whole and to foresee the long-term effects of these decisions. Incapable of this kind of considerations, novice participants of mathematical discourse are unlikely to replicate these historical choices on their own, and must thus be ushered into the new incommensurable discourse by others.

The need for meta-level learning appears many times along the school and university curricula, with this need being often invisible even to the teachers. How can meta-level learning happen? It is unlikely to begin in any other way than with the learner's exposure to the new discourse, as practiced by experts. Such exposure is likely to create a communicational conflict between the learner and the teacher: coming from different discourses, the interlocutors will be using the same words in different ways, possibly remaining unaware of this latter difference. If the learner is to enter the new discourse, she needs to recognize the need for a change and must be willing to make it even if she does not yet have any independent rationale for doing this. She must, however, be confident that those who introduced the new discourse had good reasons for doing so, and that once she is better acquainted with how the new discourse works, these reasons will become clear to her. This means she has to start acting according to the rules of the new discourse before she can say what they are good for. Thus, the first stage in learning involves participating in the discourse by imitation. While performing what must appear at this time as a mere ritual, the learner has to engage in the sustained effort to figure out the rationale for implementing these unfamiliar discursive routines. In most cases, the student's persistence may be trusted to pay. In the end, the new discourse and its stories will combine into a sensible, logical whole, and what appeared so far as mere rituals will turn into the activity of genuine mathematical explorations. In short, meta-level learning begins with *emulation of expert activities*, accompanied by a constant attempt at *rationalization*. We call this procedure *reflective imitation*. The gradual objectification is a part and parcel of the process and it is the one that turns the learner from memorizer and rule-follower into an explorer of mathematical universe.

The relevant point in this story of meta-level learning is that rather than being dependent exclusively on the learner' logical thinking, the necessary meta-level developments are predominantly a matter of persistence. They also require suspense of old beliefs and preferences. Exactly as stated by Jourdain, the learner must be able to say to himself "Go on; faith will come to you." This principle, even if recognized by the student, is difficult to implement. Not everybody's confidence in her ability to eventually "see the light" would suffice to persist indefinitely in practices that may sometimes be quite frustrating. The "many people" whose evidently insufficient understanding of mathematics puzzled Poincaré are probably those individuals who, for one reason or another, gave up at a certain point – or perhaps did not ever begin this unending sense-making struggle in the first place.

4. POSTSCRIPT: MY PERSONAL TAKEAWAYS FROM THE JOURNEY

So, what is it that we achieved in our travel from thinking-as-mathematicians to thinking-as-mathematics-educators? To begin with, our vision of mathematics underwent an ontological upheaval. From the task of describing the independently existing world of ideal mathematical objects, it reincarnated into the activity of telling stories whose protagonists are being constructed on the go. As a result, also our vision of mathematics learning changed considerably. From the straightforward, even if at times challenging, activity of *cumulating* "mathematical knowledge" the learning of mathematics was converted into an obstacle-racing, with the obstacles imposing periodic changes of direction. In each resulting transition, a new discourse subsumed an old one, retroactively changing some of the old discourse's metarules and certain uses of its keywords.

Our own transition from crypto-Platonism to commognition was the case of meta-level learning. Indeed, this was a change in our stories about mathematics and in the ways they are told – and it was highly consequential. On the new onto-epistemological foundations, we started developing teaching practices that could now be theoretically justified and rigorously tested. This passage brought also some understandings about ourselves. We realized that because of deep-seated convictions about learning we inherited from our own teachers we were sometimes, unwittingly, teaching mathematics in ways that contributed to students' life-long failure. As researchers, we learned that our own well-developed mathematical discourse, which we once saw as developing by a mere accrual, could be blinding us to what is happening when people learn mathematics. We now know that what one sees from where her long mathematical journey takes her may be quite different from what she experienced in the point of departure. Moreover, we are also aware that by the time a person reaches a certain point in the development of her mathematical discourse, she has already forgotten the initial landscape, and does not even remember that it was once quite different! All this taught us that, as teachers and researchers, we have to be always mindful of this simple caveat: When you see people doing something that does not make any sense to you, do not assume that it is senseless for the actors. The odds are that they are just not doing what you think they are. And if you are aware of the abyss between the learners' present discourse and the discourse you wish them to reach, you no longer expect them to make it to your place

in a leap, simply by hopping over the abyss. Instead, you join them in building a bridge that would take the novices safely to the other side of the dangerous gap. This technique, drawing heavily on insights earned in mathematics education research, can be trusted to save many mathematical lives.

REFERENCES

[1] J.-P. Changeaux and A. Connes, *Conversations on mind, matter, and mathematics. (M. B. DeBevoise, Ed. and Trans.)*. Princeton University Press, Princeton, NJ, 1995.

[2] P. J. Davis and R. Hersh, *The mathematical experience*. Penguin Books, London, 1981.

[3] D. Edwards, Discursive psychology. In *Handbook of language and social interaction*, edited by R. E. Sanders and K. L. Fitch, pp. 257–273, Routledge, London, 2005.

[4] A. R. Forsyth, *Perry's teaching of mathematics*. London Math Society, 1902.

[5] M. Foucault, *The archaeology of knowledge*. Pantheon Books, New York, 1972.

[6] C. F. Gauss, *Werke*. Springer, Berlin, Heidelberg, 1877.

[7] K. Gödel, Some basic theorems on the foundations of mathematics and their implications. In *Collected works, Vol. 3*, edited by S. Feferman, J. J. Dawson, W. Goldfarb, C. Parsons, and R. Solovey, pp. 304–324, Clarendon Press, Oxford, 1951.

[8] G. H. Hardy, *A mathematical apology*. Cambridge University Press, Cambridge, England, 1940/1967.

[9] R. Harré and G. Gillett, *The discursive mind*. Sage Publications, Thousand Oaks, CA, 1995.

[10] P. E. B. Jourdain, The nature of mathematics. In *The world of mathematics*, edited by J. P. Newman, Simon and Schuster, New York, 1956.

[11] I. Lavie and A. Sfard, How children individualize numerical routines: Elements of a discursive theory in making. *J. Learn. Sci.* **28** (2019), no. 4–5, 419–461. DOI 10.1080/10508406.2019.1646650

[12] R. E. Moritz, *Memorabilia mathematica: The philomath's quotation book*. The Mathematical Association of America, Spectrum, New York, 1914/1942.

[13] T. Nunes and P. Bryant, *Children doing mathematics*. Blackwell, Oxford, England, 1996.

[14] H. Poincaré, *Science and method*. Dover Publications, New York, 1952.

[15] R. Rorty, *Philosophy and the mirror of nature*. Princeton University Press, Princeton, NJ, 1979.

[16] B. Russell, Recent works on the principles of mathematics. *Internat. Monthly* **4** (1904), 84.

[17] A. Sfard, Reification as a birth of a metaphor. *Learn. Math.* **14** (1994), no. 1, 44–55.

[18] A. Sfard, When the rules of discourse change, but nobody tells you: Making sense of mathematics learning from a commognitive standpoint. *J. Learn. Sci.* **16** (2007), no. 4, 567–615.

[19] A. Sfard, *Thinking as communicating: Human development, the growth of discourses, and mathematizing*. Cambridge University Press, Cambridge, UK, 2008.

[20] A. Sfard, Taming fantastic beasts of mathematics: Struggling with incommensurability. *Int. J. Res. Undergrad. Math. Educ.*, in press.

[21] A. Sfard and I. Lavie, Why cannot children see as the same what grown-ups cannot see as different? – Early numerical thinking revisited. *Cogn. Instr.* **23** (2005), no. 2, 237–309.

[22] R. Thom, Modern mathematics: An educational and philosophical error? *Amer. Sci.* **59** (1971), 695–699.

[23] W. P. Thurston, Mathematical education. *Not. Amer. Math. Soc.* **37** (1990), no. 7, 844–850.

[24] F. Verhulst, Mathematics is the art of giving the same name to different things. An interview with Henri Poincaré. *Nieuw Arch. Wiskd. (5)* **13** (2012), no. 3, 154–158.

[25] A. N. Whitehead, *An introduction to mathematics*. Thornton Butterworth, London, 1911.

ANNA SFARD

The University of Haifa, Haifa, Israel, sfard@netvision.net.il

20. HISTORY OF MATHEMATICS

GEORGE BIRKHOFF'S FORGOTTEN MANUSCRIPT AND HIS PROGRAMME FOR DYNAMICS

JUNE BARROW-GREEN

ABSTRACT

In 1912 George Birkhoff created a sensation with his proof of Poincaré's so-called "last geometric theorem." He followed it with prize-winning papers on "The restricted problem of three bodies" (1915) and "Dynamical systems with two degrees of freedom" (1917). Many of the essential ideas from these papers can be found in his book *Dynamical Systems* (1927). At the end of the 1920s, Birkhoff began to draw up a programme of research on unsolved problems in dynamics, and in 1941 presented his ideas at the 50th anniversary celebration of the University of Chicago. Soon afterwards a summary of his lecture was published. At the time of his death in 1944, he left unfinished a manuscript of a revised and extended version of his lecture. In this paper I describe Birkhoff's work leading up to this manuscript before describing the contents of the manuscript itself.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 01A60; Secondary 3703, 37B99

KEYWORDS

George Birkhoff, history of topological dynamics

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5748–5770
DOI 10.4171/ICM2022/119

1. BIRKHOFF'S WORK IN DYNAMICS

In 1924 the Russian mathematician Nikolai Krylov described George Birkhoff as "the Poincaré of America."[1] It was an apt description. As a student in Chicago, Birkhoff had been introduced to Poincaré's work by the mathematical astronomer Forest Ray Moulton and he had immersed himself in it, especially Poincaré's great treatise on celestial mechanics—the three volume *Les Méthodes Nouvelles de la Mécanique Céleste*—which had appeared in the last decade of the 19th century. In fact, so closely did Birkhoff's name become linked with that of Poincaré that when Birkhoff died Poincaré's name featured often in the obituaries, an extreme example being the short notice written by Jacques Hadamard in which Poincaré's name appears more often than Birkhoff's **[28]**.[2] Although Birkhoff made significant advances in other fields of mathematics, such as the theory of difference equations and the four-color problem, it is his work in dynamics, notably his proof of Poincaré's "last geometric theorem" and his individual ergodic theorem, on which his fame principally rests.

Indeed, Birkhoff maintained an interest in dynamics throughout his career. His *Collected Mathematical Papers* list 32 papers under the heading, the first published in 1912, when he was aged 28, and the last, posthumously, in 1945. The second was his proof of Poincaré's last geometric theorem which he presented to the American Mathematical Society in October 1912 and which appeared in print in January 1913, with a French translation the following year **[8]**. Poincaré had published the theorem in 1912 shortly before his death, having been working on it for two years previously **[39]**. Despite (correctly) believing it to be true, Poincaré had been unable to prove it except for a few special cases.[3] Birkhoff was not the only mathematician to rise to the challenge but no-one was better prepared—his proof came only a few months after Poincaré's death.[4] Remarkably for an American mathematician at the time, Birkhoff had never been to Europe—he had learnt all his mathematics in the United States. As Norbert Wiener later wrote, "Before 1912 it had been considered indispensable for any young American mathematician of promise to complete his training abroad. Birkhoff marks the beginning of the autonomous maturity of American mathematics" **[42, P. 177]**.

Birkhoff gave Poincaré's theorem in the following form:

> *Let us suppose that a continuous one-to-one transformation T takes the ring [annulus] R formed by concentric circles C_a and C_b of radii a and b, respec-*

---

1 On 9 August 1924, Raymond Archibald, who had just met Krylov at the International Congress of Mathematicians in Toronto, wrote to Birkhoff to tell him that Krylov (whom he described as "a magnificent man") wanted especially to meet him. HUG 4213.2, Birkhoff Papers, Harvard University Archives.

2 Hadamard and Birkhoff were friends for over 30 years, and Hadamard translated some of Birkhoff's work into French. Birkhoff was a popular speaker at the famous Séminaire Hadamard in Paris, and he was one of the mathematicians interviewed by Hadamard for his famous *Psychology of Invention in the Mathematical Field* (1945).

3 In 1992 Golé and Hall would show that Poincaré had been closer to success than he had realized **[25]**.

4 Among those who made a determined but unsuccessful attempt was L. E. J. Brouwer **[40, PP. 147–148]**.

> *tively ($a > b > 0$), into itself in such a way as to advance the points of C_a in a positive sense, and the points of C_b in the negative sense, and at the same time to preserve areas. Then there are at least two invariant points* **[8, P. 14]**.

Birkhoff's proof of the theorem would soon come to be considered as "one of the most exciting mathematical events of the era and widely acclaimed" **[20, P. IV]**, although at the time, as Oswald Veblen wrote to Birkhoff from Germany in December 1913, the reaction in Göttingen was only that Birkhoff was someone who "probably [had] to be reckoned with"! **[7, P. 42]**.

There is a close connection between Poincaré's theorem and what is known as "the restricted three-body problem." This is a particular case of the three-body problem in which two large bodies, with masses μ and $1 - \mu$, respectively, rotate about their center of mass in circular orbits under their mutual gravitational attraction, and a third body of neglible mass, which is attracted by the other two bodies but does not influence their motion, moves in the plane defined by the two revolving bodies. The problem is then to ascertain the motion of the third body. The problem has one integral, which was first obtained by Carl Jacobi in 1836 and hence is known as the Jacobian integral or constant. Although the problem may appear contrived, it turns out to be a reasonable approximation to the Sun–Earth–Moon system. It was first explored by Leonhard Euler in connection with his lunar theory of 1772, but it was Poincaré who brought the problem to prominence in his celebrated memoir of 1890 **[37]**, and who later gave it its name.[5] Poincaré knew that if his theorem could be shown to be true, then it would confirm the existence of an infinite number of periodic motions for the problem for all values of the mass parameter μ. Poincaré also believed that the theorem would eventually be instrumental in establishing whether or not the periodic motions are densely distributed amongst all possible motions. As Aurel Wintner later observed, much of the dynamical work of Birkhoff was either directed towards or influenced by the restricted three-body problem **[44, P. 349]**.

In 1925 Birkhoff extended Poincaré's theorem to a nonmetric form by removing the condition that the outer boundaries of the ring and the transformed ring must coincide, and replacing it instead with the alternative condition that the outer boundary and the transformed outer boundary are met only once by a certain radial line **[11]**. He proved that the revised form held for annular regions with arbitrary boundary curves, and, correcting an earlier omission—he had not taken into account that the first invariant point might have index zero which meant that the existence of a second invariant point does not follow automatically—proved that there are always two distinct invariant points. Since the extension does not involve an invariant area integral it is essentially a topological result. Its importance lies in the fact that it can be used to establish the existence of infinitely many periodic motions near a stable periodic motion in a dynamical system with two degrees of freedom, from which the existence of quasiperiodic motions—that is motions which are not periodic

---

5 Poincaré's work on the three-body problem is discussed in detail in my book **[4]**.

themselves but which are limits of periodic motions—follows.[6] Three years later Birkhoff explored the relationship between the dynamical system and the area-preserving transformation used in the theorem [14]. Having shown that corresponding to such a dynamical problem there exists an area-preserving transformation in which the important properties of the system for motions near periodic motions correspond to properties of the transformation, he showed that a converse form of this correspondence also exists. In other words, given a particular type of area preserving transformation there exists a corresponding dynamical system. In 1931 he generalized the theorem to higher dimensions [22].

Birkhoff published three papers on the restricted three-body problem itself. The first [21], which appeared in 1915 and for which he won the Quirini Stampalia prize of the Royal Venice Institute of Science, provided the first major qualitative attack on the problem since Poincaré. Unlike Poincaré, Birkhoff, in his treatment of the problem, made little concession to analysis, and his investigation was founded almost entirely on topological ideas. By considering the representation from a topological point of view, he was able to illustrate the problem's dependence on the value of the Jacobian constant. He established a transformation of the variables which enabled him to derive a new form of the equations in which the equations are regular, providing the third body is not rejected to infinity. From this he created a geometric representation in which the manifolds of states of motion are represented by the stream-lines of a three-dimensional flow and are without singularity unless the Jacobian constant takes one of five exceptional values. Having excluded these five values, the totality of the states of motion could then be represented by the stream lines of a three-dimensional flow occupying a nonsingular manifold in a four-dimensional space. But, as Poincaré had shown, providing the mass of the one of the two main bodies is sufficiently small, the representation of the problem as a three-dimensional flow can be reduced to a representation which depends on the transformation of a two-dimensional ring into itself [38, PP. 372–381]. Birkhoff showed that Poincaré's transformation could be considered as the product of two involutory transformations, a result he subsequently used to prove the existence of an infinite number of symmetric periodic motions, as well as results concerning their characteristic properties and distribution.

Twenty years elapsed before Birkhoff next published on the problem. In the interim he had worked extensively on general dynamical systems, the crowning result of which was another prize memoir which appeared in 1935 [16], the prize having been awarded by the Pontifical Academy of Sciences. In two later papers on the restricted problem which derived from lectures given at the Scuola Normale Superiore di Pisa, he combined ideas from the prize memoir of 1915 together with some general results from the one of 1935, notably his development of Poincaré's idea of a surface of section (now often called a Poincaré section).[7] In the first of these two later papers [17], he focused on the analytic properties of the surface

6 A modern and slightly modified account of Birkhoff's proof is given by Brown and Neumann [23].

7 Given an n-dimensional phase space, a surface of section is an $(n-1)$-dimensional space embedded in the original space and transversal to the flow of the system.

of section and the transformation he had used in 1935, while in the second [18] he used qualitative methods to explore the results from the first in order to obtain further information about the different types of motion and the relationships existing between them.

In 1923 Birkhoff was awarded the Bôcher Memorial Prize of the American Mathematical Society for a paper in which he provided a general treatment of dynamical systems with two degrees of freedom [9]. Such systems comprise the simplest type of nonintegrable dynamical problems, and, as exemplified in the work of Poincaré, they form the natural starting point for qualitative explorations into questions of dynamics. According to Marston Morse, Birkhoff stated that he thought the Bôcher prize paper was as good a piece of research as he would be likely to do [35, P. 380].

Birkhoff began with the equations of motion in standard Lagrangian form:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial x'} - \frac{\partial L}{\partial x}\right) = 0, \quad \frac{d}{dt}\left(\frac{\partial L}{\partial y'} - \frac{\partial L}{\partial y}\right) = 0,$$

where the function L, which is quadratic in the velocities, involves six arbitrary functions of x and y. By making an appropriate transformation of variables, he reduced the equations to a normal form which involved only two arbitrary functions of x and y. In the reversible case, that is, when the equations of motion remain unchanged when t is replaced by $-t$, the transformation was already well known. In this case the equations of motion can be interpreted as those of a particle constrained to move on a smooth surface and the orbits of the particle interpreted as geodesics on the surface. But in the irreversible case, as, for example, in restricted three-body problem, Birkhoff's transformation was new and he gave a dynamical interpretation in which the motions can be regarded as the orbits of a particle constrained to move on a smooth surface which rotates about a fixed axis with uniform angular velocity and carries with it a conservative force field. The central part of the paper concerned various methods by which the existence of periodic motions could be established. These include his "minimum method," and his "minimax method," the latter later providing a starting point for the work of Morse on calculus of variations in the large. Birkhoff also considered Poincaré's method of analytic continuation which is applicable to both reversible and irreversible periodic motions. One of the problems with the method was that it was only valid for a small variation in the value of the parameter. The restriction was due to the possibility that the period of the motion under consideration might become infinite. Thus to increase the interval of the variation it is necessary to show that this possibility cannot arise and Birkhoff did precisely that for a wide range of periodic motions.

It was in the Bôcher prize paper that Birkhoff first began to generalize Poincaré's idea of a surface of section and formally develop a theory attached to it. Poincaré had used the idea specifically to reduce the restricted three-body problem to the transformation of a ring to itself, but if the method was to have a general validity it was important to establish under what circumstances surfaces of section exist. Birkhoff was able to show that not only do they exist in a wide variety of cases but also that they can be of varying genus and have different numbers of boundaries.

In his "Surface transformations and their dynamical applications" of 1920 [10], Birkhoff elaborated and extended some of the ideas he had broached at the end of the Bôcher

prize paper. By reducing the dynamical problem to a transformation problem and studying certain transformations and their fixed points, which he did at length, he was able to classify certain different types of motion. For example, whether a periodic motion, which is represented by a fixed point, is stable or unstable can be determined by examining the behavior of a point sufficiently close to the fixed point under repeated iterations of the transformation. Later Birkhoff considered the question of stability in more detail **[13]**.

Invited by Gösta Mittag-Leffler in 1926 to contribute to the 50th volume of *Acta Mathematica*—the journal which Mittag-Leffler had edited since its inception in 1882—Birkhoff chose to tackle Poincaré's conjecture concerning the denseness of periodic motions. It was a particularly fitting choice of subject, given Poincaré's early and consistent support of *Acta*.[8] A feature of Birkhoff's paper **[12]** is his introduction of the billiard ball problem—that is, to determine the motion of a billiard ball on a convex table—which he used to show how Poincaré's last geometric theorem could be applied to dynamical systems with two degrees of freedom.[9] Having considered certain types of periodic motion, he was able to conclude that if a dynamical system admits one stable periodic motion of nonexceptional type—the exception being when the period of the perturbed motion is independent of the constants of integration—then it admits an infinite number of stable periodic motions within its immediate vicinity, and the totality of these stable periodic motions form a dense set. Although this does not resolve Poincaré's conjecture, it does show that it cannot be true unconditionally. He was able to prove the conjecture in the case of a transitive system—that is a system in which "motions can be found passing from nearly one assigned state to nearly any other arbitrarily assigned state" **[12, P. 379]**—showing that the periodic motions together with those asymptotic to them are densely distributed.

Birkhoff's influential book, *Dynamical Systems*, which derived from the American Mathematical Society Colloquium Lectures he delivered in Chicago in 1920, was published in 1927, with a new edition appearing in 1966. A Russian translation, which also contained translations of several of Birkhoff's papers including **[15]**, was published in 1941 and reprinted in 1999. Although representing "essentially a continuation of Poincaré's profound and extensive work on Celestial Mechanics" **[20, P. III]**, Birkhoff's book opened a new era in the study of dynamics by detaching the subject from its origins in celestial mechanics and making use of topology **[3]**. It provides a summary of Birkhoff's research in dynamics during the preceding 15 years, with the final three chapters—on the general theory of dynamical systems, the case of two degrees of freedom, and the three-body problem—bringing together the main strands of his work. As Bernard Koopman, one of Birkhoff's former students, remarked, *Dynamical Systems* is better described as a theory than as a book **[31, P. 165]**. Birkhoff's goal was clear: "The final aim of the theory of the motions of a dynamical system must be directed

---

8 Poincaré's contributions to *Acta Mathematica* are discussed in my article **[5, PP. 148–150]**.

9 It is indicative of the paper's status that it was selected by Robert MacKay and James Meiss for reproduction in their book of the most significant writings on Hamiltonian dynamics published since the First World War **[33]**.

toward the qualitative determination of all possible types of motions and of the interrelations of these motions." [20, P. 189].

He started with a general class of dynamical systems, that is systems defined by the differential equations,

$$\frac{dx_i}{X_i} = dt \ldots \quad (i = 1, \ldots, n),$$

where the X_i are n real analytic functions, and a state of motion can be represented by a point in a closed n-dimensional manifold. A motion can then be represented by a trajectory in the manifold, and its domain is its closed set of limit points. The trajectories composed entirely of limit points are those Birkhoff called "recurrent motions." More generally, recurrent motions are those which trace out with uniform closeness, in any sufficiently large period of their entire history, all their states. Since, by definition, every point on the trajectory of a recurrent motion is a limit point, the motion must approach every point on the trajectory infinitely often and arbitrarily closely. Thus the simplest types of recurrent motions are the stationary motions and the periodic motions. As Birkhoff showed, the idea of recurrent motion is a particularly useful one with regard to the general problem of determining all possible motions in a particular dynamical system. For example, he proved that the set of limit motions of any motion contains at least one recurrent motion; and that any point either generates a recurrent motion or generates a motion which approaches with uniform frequency arbitrarily close to a set of recurrent motions. Furthermore, the concept of recurrent motion can be used to derive definite results about the motion in an arbitrary dynamical system; a significant feature of the theory being that it is valid for systems with any degree of freedom. This is in contrast to Poincaré's theory of periodic motion which is known to be valid only for systems with two degrees of freedom.

The theory developed in Birkhoff's papers and further expounded in *Dynamical Systems* formed the bedrock on which Birkhoff's Chicago lecture and its related manuscript were built, and it is to these we now turn.

2. BIRKHOFF'S FORGOTTEN MANUSCRIPT

In September 1941 the University of Chicago celebrated its 50th anniversary. It was a celebration that had been two years in the planning. Honorary degrees were awarded and a symposium was held in conjunction with the American Association for the Advancement of Science. According to an account in the university magazine, the celebration was sufficiently "significant that, in a world at war, it attracted national and even world wide attention" [30, P. 6].

As one of the leading figures in American mathematics and a former student of the university, Birkhoff was a natural choice for an honorary degree and symposium speaker, the citation describing him as the "leading contributor to the fundamentals of dynamics." The only other mathematician amongst the 34 others on the rostrum was Birkhoff's close friend and long-standing colleague Oswald Veblen, also a Chicago protégé.

For the subject of his lecture, Birkhoff chose "Some unsolved problems of theoretical dynamics," a topic well in keeping with the anniversary theme of "New Frontiers in Education and Research." The symposium was well advertised prior to the celebrations and before Birkhoff delivered his lecture he was asked by *Nature* if he could provide the journal with a summary. However, the summary did not appear in *Nature* but in *Science* and it appeared some three months after the lecture had been delivered [19]. In fact, the lecture was ready only about a week before it was due to be delivered, as Birkhoff admitted to Eric O'Connor, one of his former doctoral students:

> *During the last few weeks I have been extremely occupied with the address which I have to give next week at Chicago. In it I take a look at Classical Dynamics from the abstract point of view and suggest about a dozen problems, many of them new, which seem to be most directly in the line of further advance. In one or two instances I indicate a partial answer to these. It now looks as though the paper will be in good shape for the 24th September, when I have to deliver it, but it has been a very close squeak!*[10]

The idea of presenting a programme for research in dynamics was not new for Birkhoff. Some 13 years earlier, in 1928, he had given a series of lectures at the University of Berlin on "Some Problems of Dynamics" and the lectures were published in German in a condensed form [15]. In these lectures, having emphasized the importance of qualitative dynamical ideas for the exact sciences, he discussed various examples including the billiard ball problem, the motion of a particle on a smooth convex surface and on a smooth closed surface of negative curvature, and the three-body problem. On that occasion, he listed six problems:

I To construct a dynamical system on a three-dimensional closed phase space, in which the ordinal r of central motion is > 3.

II To prove that in the case of the Hamiltonian problem with two degrees of freedom, with closed phase space and with at least one stable periodic motion, the periodic motions are everywhere dense.

III To prove that in the case of all Hamiltonian problems with closed phase space the recurrent motions are everywhere dense.

IV To prove, for a given conservative transformation T, the existence of corresponding Hamiltonian systems in particular of geodesic type.

V If T is any conservative transformation with a fixed point P of stable type, then determine the necessary conditions so that there are infinitely many points P_n existing in the neighborhood of P which are fixed points of T^m.

VI To prove, in the case with two degrees of freedom, the existence of a dynamical system that has a periodic motion of stable type, which is not truly stable.

---

10 Letter from Birkhoff to O'Connor, 18 September 1941. HUG 4213.2.2, Birkhoff Papers, Harvard University Archives.

Of these, only the first three relate to problems Birkhoff discussed in Chicago. The first was solved in 1946 by A. G. Maier [34]. In 1941 the problems were republished in Russian to accompany the Russian edition of *Dynamical Systems*, where they are described as "important, unsolved problems." Further work remains to be done to establish the extent of interest generated by these problems subsequent to both the German and the Russian publications.

Nine years after his lectures in Berlin, Birkhoff returned to the same theme but this time in Paris. In 1937 he gave a lecture at the Institut Henri Poincaré entitled "Quelques problèmes de la Dynamique théorique." Birkhoff referred to this lecture in a footnote of the manuscript where he said that in Paris he had made reference to "one or two of the problems listed in the present paper" but without identifying which ones, and no further information on this lecture has so far come to light.

Birkhoff received the Chicago invitation in November 1940, and in April 1941 he was invited by Otto Schmidt and Anisim Bermant to contribute to the celebratory 50th volume of *Matematicheskii Sbornik*, the prestigious Russian mathematical journal founded in 1866.[11] For some time Russian mathematicians had been closely following Birkhoff's work, especially in dynamics, as is evident from Krylov's remark of 1924 given above. Also in the 1920s, a group in Pavel Aleksandrov's topology seminar in Moscow had specialized in studying Birkhoff's publications;[12] and in 1936 Birkhoff had been invited by A. A. Markov to speak on the ergodic theorem and related topics at an international conference due to take place in Leningrad in 1937, although in the event the conference was canceled.[13] Birkhoff cannot have taken long to decide that an article laying out his programme for dynamics would make a fitting contribution to the journal, knowing that the Chicago meeting would provide him with an excellent opportunity to test out his ideas before committing them to print.

In May 1943, Birkhoff wrote to his Russian colleagues to let them know that he had "written out an extensive article not wholly completed as yet on 'Some Unsolved Problems of Theoretical Dynamics'," mentioning that he had spoken on the subject "in a preliminary way" in Chicago (Figure 1), but that he had decided to delay sending the article to Russia until after the cessation of hostilities.[14] But it was not to be. On 12 November 1944, Birkhoff, aged only 60, died unexpectedly.[15] Thus the manuscript, which runs to some 40 pages, was never submitted. It remains as a hand-annotated typescript, with additional handwritten leaves, among Birkhoff's papers in the Harvard University Archives.[16] In a footnote appended to

---

11 Letter from Schmidt and Bermant to Birkhoff, 2 April 1941. HUG 4213.2.2, Birkhoff Papers, Harvard University Archives.

12 Letter from Aleksandrov to Birkhoff, 19 October 1926. HUG 4213.2, Birkhoff Papers, Harvard University Archives.

13 Letter from Birkhoff to Markov, 26 February 1936; letter from Markov to Birkhoff, 7 May 1936. HUG 4213.2, Birkhoff Papers, Harvard University Archives.

14 Letter from Birkhoff to Schmidt and Bermant, 18 May 1943. HUG 4213.2.2, Birkhoff Papers, Harvard University Archives.

15 As described by his Harvard colleague, Edwin B. Wilson, Birkhoff had some time in hand before a lunchtime visit to his son, Garrett, and had taken the occasion to rest but when his wife went to find him he had passed away [43, P. 578].

16 HUG 4213.52, Birkhoff Papers, Harvard University Archives.

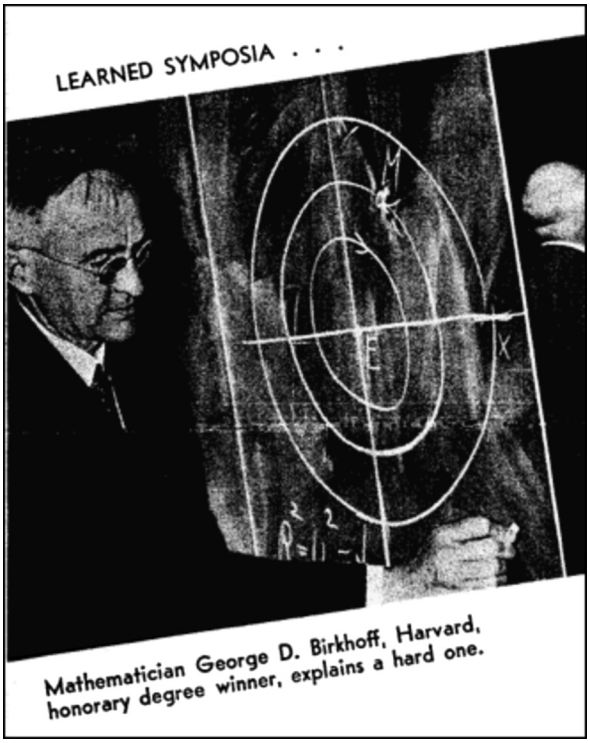


FIGURE 1
Birkhoff delivering his symposium lecture **[30, P. 6]**. Courtesy of the University of Chicago Library.

the title page of the manuscript, Birkhoff stated that he had written the paper with the dual purpose of reading it in Chicago and publishing it in the anniversary volume of the Russian journal. Comparing the manuscript with the summary, and taking into account the delay of the publication of the latter, it seems likely that Birkhoff, having lectured from the manuscript then used it to prepare the summary and in the course of the latter's preparation further annotated the manuscript.

The manuscript opens as follows:

> *It scarcely seems too much to say that all the basic problems of point-set theory, topology, and the theory of functions of real variables present themselves naturally in purely dynamical contexts. Some of these dynamical problems are best formulated and solved in terms of an underlying abstract space, as important recent Russian and American work has shown. Others are inherently of more special character.*
> *In the present paper I venture to set forth certain unsolved problems of this type which seem to me worthy of further study. The problems are arranged as much as possible in order of decreasing abstractness. They are formulated in terms of*

positive conjectures in the belief that this procedure is most likely to stimulate further research. In each case indications of the underlying reasons for these conjectures are made. Some new definitions are given, as for instance that of a "dynamical" flow in an abstract metric space; and some partial results are deduced, as for instance the brief proof in the concluding section that the non-existence of other periodic lunar orbits beside the fundamental variational orbit and the allied retrograde orbit of G. W. Hill's theory of the motion of the Moon about the Earth would imply that all possible lunar orbits with the same constant of Jacobi have the same mean angular advance of perigee per synodic revolution.

The summary opens rather differently. There Birkhoff gives a pathway for the development of his ideas—he traces them from Poincaré, who first realized that the study of dynamical systems led directly to problems in topology, on through the abstract ideas of E. H. Moore—describing how these ideas fed into his own work.[17] Although Moore deserved a high billing, it was also a diplomatic move on Birkhoff's part to be explicit about the contribution of Moore, his former thesis advisor and first head of the University's mathematics department, who had died in 1932. In the manuscript, the reference to Moore, although laudatory, is considerably abbreviated and consigned to a footnote.

Altogether there are 17 problems, the first ten are formulated in terms of abstract spaces, the 11th is concerned with extensions of results of Karl Sundman on the three-body problem to the motion of a gas. And the last six, which are concerned with n-dimensional spaces, are of a topological nature. The paper is also divided up into sections which imposes a useful classification on the problems. The manuscript also includes a "provocative form of conclusion." In the prelecture press release, Birkhoff referred to only ten problems without listing them, so it is possible that he had originally intended to present only ten problems and it was expanding the paper that led to the "close squeak" referred to in the letter to O'Connor mentioned above.[18] In what follows, the section headings and the problems themselves are taken directly from the manuscript. Other material from the manuscript will be given in quotation marks followed by a page number.

The first problem, a conjecture about the interrelationship between continuous and discrete flows in an abstract space R, is precursored by three sections on continuous and discrete flows, including an explanation of geodesic flow. As Birkhoff observed in a footnote, the idea of using "this kind of abstract setting for a dynamical problem" did not originate with him but in an article of 1933 by Hassler Whitney, one of Birkhoff's research students [41].

---

17 Birkhoff felt especially grateful to Moore for impressing him "with the importance of the abstract domain and for stimulating [him] on the abstract side." Letter from Birkhoff to Raymond Archibald, 5 April 1938. HUG 4213.4.5, Birkhoff Papers, Harvard University Archives.

18 Another difference between the press release and the summary is that in the former Hassler Whitney and Norbert Wiener are identified as American authors of recent work on abstract dynamics while the latter refers simply to "American mathematicians." University of Chicago Development Campaigns and Anniversaries Records, Box 12, Folder 11.

The idea is that in R, which is a compact metric space, a type of reduction of a continuous flow to a discrete one may be effected by showing that there exists a surface of section in R on which the flow can be studied. Each point in R represents a state of motion and as time passes there is a steady flow of R into itself, with each point tracing out a "curve of motion," each curve representing a complete motion of the dynamical system [19, p. 598]. As Birkhoff noted, he had already shown "in the n-dimensional case, and very recently Ambrose and Kakutani had established in the abstract case, a kind of converse reduction of a continuous flow to a discrete flow may be made, providing one is content to introduce discontinuous flows" (p. 5). His hope was for a more complete result, and he felt certain that the conjecture would be shown to hold. Since he did not give a citation for the Ambrose and Kakutani paper which had been submitted for publication in 1941 and appeared in 1942 [2], it would appear that he did not return to the manuscript in the years following the lecture apart from reporting on its existence to Schmidt and Bermant.

Problem 1. Any (continuous) flow without equilibrium points in a compact metric space R admits of a complete open surface of section Σ in R, on which the flow defines an extensibly-discrete flow $Q = \phi(P)$ obtained by following any point P of ϕ to the first subsequent point Q of Σ on the same stream line. Conversely, given any metric space Σ on which an extensibly-discrete flow, $Q = \phi(P)$, is defined, then it is possible to imbed Σ in an isometric compact metric space R and to define a continuous flow in R, so that Σ forms a complete open surface of section for this flow, for which $Q = \phi(P)$ in the related extensibly-discrete flow.

Birkhoff next discussed recurrent motions and central motions, central motions being those which recur infinitely often close to any particular state of the motion, or at least have such motions in the infinitesimal vicinity of any state. Having observed that "all the motions of a dynamical system will be central if and only if every molecule of the system overlaps itself as time increases or decreases" (p. 9), he noted that in the classical case there are many examples in which all the motions are central. And it was this that led him to ask the analogous question of recurrent motions, i.e., "what are the circumstances such that *all* the motions of a dynamical system will be recurrent?" (p. 10).

This last question provides the basis for Problems 2 and 3 in which Birkhoff conjectured that all the motions of a continuous flow would be recurrent if and only if the flow may be decomposed into a set of irreducible constituent flows which are "homogeneous," i.e., such that the stream lines are topologically indistinguishable from one another. As an example, he cited the two-body problem—two particles interacting gravitationally with no other forces acting—as being of this type, providing the value of the energy constant is sufficiently small, with the irreducible constituents being the individual periodic motions.

Problem 2. All the motions of a regionally transitive (discrete or continuous) flow in a compact metric space R will be recurrent if and only if the flow is "homogeneous," in the sense that an automorphism of the flow exists (with possible modification of the definition of the "time") which takes an arbitrary point P into a second arbitrary point Q.

Problem 3. All the motions of regionally transitive flows in a compact metric space R will be recurrent if and only if the closest set of motions formed by any motion M and its limit motions is homogeneous in the subspace R_M of these motions [and] every minimal closed component of the flow is homogeneous.

As contrasting illustrations, Birkhoff gave as examples the two-body problem in which all the motions are periodic and the billiard ball problem which, "although 'integrable', has a family of non-recurrent motions, namely those which pass infinitely often through the two foci and are doubly asymptotic (homoclinic in the sense of Poincaré) to the major axis" (p. 13).

Conservative flows. In ordinary dynamical systems, conservative flows are those with an invariant volume integral, e.g., the flow of an incompressible liquid. Here Birkhoff considered the extension of conservative flows to the abstract case. By 1941 this had become an active area of research and in the summary he named several Russian mathematicians (Beboutov, Bogolyubov, Krylov, Stepanov) and American mathematicians (Halmos, Oxtoby, Ulam, von Neumann, Wiener, Wintner) who had made important studies of such flows **[19, p. 599]**, although rather curiously he did not mention them in the manuscript.

In the fourth problem, which was preceded by a four-page introduction, Birkhoff conjectured that if the abstract flow is so regular as to be "geodesic" then it will be conservative if all the motions are central, while in the fifth he conjectured that the recurrent motions are necessarily everywhere densely distributed in the abstract space of a geodesic conservative flow. As he pointed out, Poincaré's recurrence theorem makes the latter conjecture a very natural one.[19] However, he did not "expect the periodic motions to be always everywhere dense in the conservative case or even in the case of a dynamical flow." (p. 17).

Problem 4. A geodesic flow all of whose motions are central always admits an invariant positive volume integral.

Problem 5. The recurrent motions are everywhere dense in any conservative flow, at least if it be geodesic.

Ergodic theory and conservative flow. Birkhoff opened this section with a short discussion relating to his own "individual ergodic theorem," observing that the theorem implies "that for conservative systems almost all motions have definite habits of recurrence with regard to any measurable type of behaviour." (p. 18). He also noted the priority of von Neumann's "mean ergodic theorem."[20] In the summary, he avoided any mention of ergodic theory but instead used features of the billiard ball problem, such as the fact that in the long run the ball will be

---

19 Roughly speaking, Poincaré's recurrence theorem says that if the flow is volume-preserving then, at some point in the future, the system will return arbitrarily close to its initial state. For a discussion of the theorem, see **[4, pp. 86–87]**.

20 An account of the relationship between Birkhoff's individual ergodic theorem and von Neumann's mean ergodic theorem, which also explains the confusing chronology of publication, is given by J. D. Zund **[45]**.

on any designated part of the table a definite proportion of the time, in order to demonstrate the significance of conservative flow. Problem 6 proposes a topological characterization of conservative flows based on this fact of recurrence.

Problem 6. If a continuous flow in a compact metric space has the property that for any open region of R, the exceptional sets for which a positive mean sojourn time τ (the same in both senses of the time) fails to exist are always of measure 0 with respect to some measure $\int dP$, then there exists necessarily at least one invariant integral $\int \mu dP$.

Birkhoff also noted that "an important paper" by Oxtoby and Ulam containing questions "closely related" to Problem 6 was in the pipeline [36]. This paper, which Koopman summarised as a "thorough and detailed study of the group of measure-preserving and measurability preserving automorphisms (homeomorphisms into itself) in polyhedra, their metrical transitivity, equivalence, and the whole bearing of such questions on ergodic theory,"[21] appeared in 1941, the absence of its publication details providing a further indication that Birkhoff did not edit the manuscript in the years after it was written.

Discontinuous conservative flows. This section and its accompanying problem are on two handwritten pages. These pages open with the words "Very recently" (p. 19′) and a footnote gives a full citation for a paper published by Ambrose in July 1941 [1], showing that these pages were written either shortly before, or possibly soon after, the lecture was given. The flows now considered are "measure-preserving flows which are 1–1 except over sets of measure 0 and carry measurable sets into measurable sets (in particular, sets of measure 0 into sets of measure 0) and conserve a positive volume integral" (p. 19′), and such flows are, as Birkhoff noted, of particular interest from the point of view of probability, and in this context he mentioned that they had recently been studied by von Neumann, Kakutani, Ambrose, and Halmos. In the summary Wiener and Wintner are exchanged for Ambrose and Kakutani, and there is no mention of probability.

Having established that the underlying space can be taken as a line segment of unit length, and relaxed the condition of continuity on a conservative flow, Birkhoff proposed a characterization of the invariants of the flow based on what he termed "packing coefficients." He explained the latter as follows: "Make the total μ-measure 1 by choosing the total measure as a unit. Select any $n \geq 1$ and consider all ways of decomposing a minimal metrically transitive constituent into a measurable set Σ and its first $n-1$ images under [a discrete flow] T, say $S^1, \ldots, S^{(n-1)}$ in such a way that these sets are disjoint. To each such decomposition there will be a measure of the complementary point set. We will call the lower bound of these quantities the 'nth packing fraction' and denoted it by σ_n, and it is easy to prove that the inequality $\sigma_n \leq 1/n$ always holds." (p. 19″).

Problem 7. Any such discontinuous conservative transformation T is completely characterized by its "spectrum," determining the nature of the metrically transitive constituents, and by the packing coefficients $\sigma_1, \sigma_2, \ldots$, for every such constituent. These packing coefficients may be taken arbitrarily except for the fact that $n\sigma_n$ forms a decreasing sequence.

21 *Mathematical Reviews* M0005803.

Dynamical flows. The next three problems, Problems 8–10, derive from Birkhoff's attempt to define abstractly a "dynamical flow" where he takes as his model Pfaffian systems,[22] rather than Hamiltonian systems of classical dynamics. However, this part of the manuscript is a little tricky to follow as there are six handwritten pages inserted between two typescript pages (pp. 21–22/23). Unlike the other handwritten pages, there is nothing to show exactly where the text from these pages should be inserted. It is evident that wherever they are inserted the typed text would need to be adjusted for the narrative to flow. The first five of these pages provide the justification for a result he had deduced from the properties of his abstract definition of a line integral, a result he needed for his definition of a continuous dynamical flow which involves the existence of a line integral (in an abstract sense which he made precise). The final handwritten page contains only Problem 8 (after which all subsequent problems in the typescript were renumbered). It is notable that in the summary he remarked, that "the crucial part of the characterization of a dynamical flow lay in the suitable definition of a line integral in any abstract 'geodesic space' R", and a few lines later observed that "the question of an adequate characterization of a dynamical flow beyond the obvious properties of conservativeness and continuity has been especially baffling" **[19, P. 599]**, which suggests that he returned to this part of the manuscript after he gave the lecture.

Problem 8. Any dynamical flow is necessarily conservative with reference to a completely additive measure with positive measure on any open set.

In Problems 9 and 10, Birkhoff returned to the question of the denseness of periodic motions, the question he had addressed in his *Acta Mathematica* paper of 1925. Now he reformulated the question in an abstract setting with the added condition of stability. He defined a periodic motion to be stable (topologically) "if there are other complete motions in its ε-neighborhood", adding that a similar definition can be made for 'stable' recurrent motions', providing neighboring recurrent motions of the same minimal set are excluded from consideration (p. 25). He defined a completely unstable flow as one in which there are no stable periodic or recurrent motions, for example, geodesics on a closed surface of negative curvature, and here he cited the well-known work of Hadamard (1898) and Morse (1921, 1924).

Problem 9. In any regionally-transitive nonhomogeneous flow of dynamical type the periodic motions are everywhere dense.

Problem 10.

(a) In a regionally transitive dynamical flow not of completely unstable type, the stable periodic motions are everywhere dense, and the set of such motions is dense on itself (i.e., in the infinitesimal neighborhood of any stable periodic motion there exist infinitely many other stable periodic motions). Furthermore,

22 Birkhoff had first defined Pfaffian systems in his Colloquium Lectures of 1920, and later considered them in **[13]** and in *Dynamical Systems*. They were brought to further prominence by Lucien Feraud in an explanatory paper of 1930 **[24]**.

in the neighborhood of any stable recurrent motion there are similarly infinitely many other stable periodic motions.

(b) In any regionally transitive dynamical flow of completely unstable type, which is furthermore not homogeneous, the unstable periodic motions are everywhere dense.

On a possible extension of some work of Sundman. Next Birkhoff asked for the generalization to a gas of certain remarkable results on the three-body problem produced in the early years of the 20th century by the Finnish mathematical astronomer Karl Sundman. Birkhoff was a strong advocate for Sundman's theoretical "solution" to the three-body problem which, due to its practical limitations, had met with a mixed reception.[23] He had even gone as far as to say that "the recent work of Sundman is one of the most remarkable contributions to the problem of three bodies which has ever been made" **[13, P. 260]**. Of particular relevance here are Sundman's results concerning triple and binary collisions, namely that a triple collision can occur only if all three integrals of angular momentum are simultaneously zero, and that the singularity at binary collision is of removable type. Birkhoff had already shown in *Dynamical Systems* how the essence of Sundman's argument can be used under other laws of force and for a system of more than three bodies to establish that, with similar initial conditions, a simultaneous near approach of the bodies cannot occur, hence the generalization to a gas was a natural next step. The problem was formulated rather vaguely—indeed, in the summary he admitted it was incomplete **[19, P. 599]**—but he chose to include it because it provided "an interesting illustration of a dynamical flow in a kind of Euclidean space R of infinitely many dimensions, intermediate in type between the flows in abstract metric space and in n-dimensional Euclidean space" (p. 27).

Problem 11. To determine equations of state and initial conditions of a free bounded gas such that the diameter of the gas can never be less than a specifiable $d > 0$ despite the fact that such configurations are compatible with the known integrals.

Following on, Birkhoff now turned to problems relating to motions in n-dimensional space.

A problem concerning central motions in *n*-dimensional space. Problem 12 is essentially the first problem he presented in Berlin, and which was solved by Maier in 1946, now extended to the n-dimensional case. On this occasion, Birkhoff used the notion of "wandering motions" W_0 of a space R, a notion he had introduced in *Dynamical Systems*, and which here he described (none too clearly) as "those which can be embedded in a molecule which never overlaps itself as time increases or decreases" (p. 27). When the wandering motions are removed from R, there remains a closed subspace, $M_1 = R - W_0$, of lower dimension, which can then be considered from the same point of view. Using this idea, Birkhoff formed "a well-ordered set $M = M_0, M_1, \ldots$, which is enumerable and terminates in the set of cen-

23 A detailed discussion of Sundman's work on the three-body problem and its reception is given in my article **[6]**.

tral motions M_c." Emphasizing the fact that in the known cases the series contain at most n terms, he proposed the following form of the problem:

Problem 12. To construct a continuous flow in a closed manifold of $n > 2$ dimensions for which the well-ordered series $M = M_0, M_1, M_2, \ldots$ leading to the central motions M_c contains more than n and if possible an infinite number of terms.

A problem in the 3-dimensional case. Problem 13 was suggested by recent work of the Hungarian mathematician Béla Kerékjártó on "regular" or nearly regular transformations of 2-dimensional closed surfaces of arbitrary genus.[24] Here Birkhoff conjectured that 3-dimensional flows that are "regular" have one of only three different forms.

Problem 13. For an ordinary 3-dimensional manifold R_3, to show that the only regular discrete flows are topologically equivalent to one of the following:

(1) R_3, a 3-dimensional torus with a transformation

$$\theta_1 = \theta_1 + \alpha_1, \quad \theta_2 = \theta_2 + \alpha_2, \quad \theta_3 = \theta_3 + \alpha_3,$$

with $\theta_1, \theta_2, \theta_3$ being angular coordinates for the torus.

(2) R_3, the product of a surface of sphere and circle, and the transformation of each of these a pure rotation.

(3) R_3, a 3-dimensional hypersphere and the transformation of a rigid rotation of this sphere.

A problem in the 2-dimensional case. Birkhoff now moved to problems connected with analytic transformations, the ideas emerging from the first of his papers on the restricted three-body problem [21]. In the first of these problems, he conjectured that a particular transformation of the surface of a sphere into itself with two fixed points, which is such that all iterations of the transformation produce no other fixed points, is a pure rotation when considered topologically.

Problem 14. A 1–1 direct analytic, conservative transformation T of the surface of a sphere into itself with two and only two fixpoints P, Q for T and all its iterations is topologically equivalent to a pure rotation of the sphere about an axis through an angle incommensurable with 2π.

From this he was led to propose the following analogous problem for a plane circular ring:

Problem 15. A 1–1 direct analytic conservative transformation of a circular ring into itself, in which two boundaries are invariant and which possess no periodic points is topologically equivalent to a rotation of the ring through an angle α incommensurable with 2π.

---

24 Birkhoff was well acquainted with Kerékjártó. In 1925 he had supported his promotion in Szeged, and in 1928 he had visited Szeged to lecture on Poincaré's last geometric theorem.

A conjectural supplement to Poincaré's last geometric theorem. In the final two problems Birkhoff returned again to Poincaré's last geometric theorem. Unsurprisingly, he thought these two problems, since they express conjectures which in a sense represent a complement to the theorem, were the ones likely to generate the most interest.

In Problem 16, Birkhoff conjectured that the theorem would hold in the case when the points on the two concentric circles C_a and C_b, are advanced by the same angular distance (in contrast to the original theorem where the points are advanced by distinct distances), that is, their rotation numbers α are equal, provided that some nearby points of the ring become separated widely in an angular sense when the transformation T is repeated sufficiently often, as happens when the rotation numbers are unequal.

Problem 16. Let T be a 1–1 continuous discrete conservative transformation T of a circular ring into itself which leaves the two circular boundaries individually invariant, with equal rotation numbers α along these boundaries. Then if nearby pairs of points exist which separate indefinitely in an angular sense under indefinite iteration of T, there will necessarily exist periodic points.

This was followed by a conjecture on the partial converse, the case when the common rotation number α is not a rational multiple of 2π.

Problem 17. Under the same hypotheses concerning T as in the first part of Problem 16, let us further require only that for no preliminary deformation of the ring in itself can the angular deviation of all pairs of points less than 2π apart in angular sense be made to remain less than $2\pi + \varepsilon$ under all iterations of T (ε arbitrary). There will then exist periodic points on the ring. Furthermore, if $\underline{\alpha}$ and $\overline{\alpha}$ denote the lower and upper bounds of the rates of angular advance for such periodic point groups then we have $\underline{\alpha} < \overline{\alpha}$ and $\underline{\alpha} \leq \alpha \leq \overline{\alpha}$ and, for any relatively prime integers m and $n > 0$ such that

$$\underline{\alpha} < 2\frac{m}{n}\pi < \overline{\alpha},$$

there exist at least two periodic point groups of n points whose angular coordinates increase by $2m\pi$ under the nth power of T.

Birkhoff then considered the particular case when the given transformation can be expressed as the product of two involutory transformations, showing that in this case the first part of the conjecture is true.

Application to the restricted problem of three bodies. In the final part of the paper Birkhoff applied the above result to the planar restricted three-body problem, the version of the problem treated in the 1870s by the American mathematical astronomer George William Hill in his work on the lunar theory—work which had famously inspired Poincaré—giving the differential equations as Hill had done (p. 33):

$$\frac{d^2x}{dt^2} - 2\frac{dy}{dt} = \frac{\partial\Omega}{\partial x}, \quad \frac{d^2y}{dt^2} - 2\frac{dx}{dt} = \frac{\partial\Omega}{\partial y}, \quad \Omega = \frac{3}{2}x^2 + \frac{1}{\sqrt{x^2+y^2}},$$

together with the equation for the Jacobian constant C,

$$\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 = 2\Omega - C.$$

He described the "actual case" of the Sun–Earth–Moon problem in which the Earth is considered to lie at the origin in the x, y plane, the Sun is at infinity in the direction of the positive x-axis and the "infinitesimal" Moon is rotating in the x, y plane at unit angular velocity with the Earth and the Sun, where the positive constant C_0 is such that the Moon can never escape from the closed region $2\Omega = C$ about the Earth, symmetric in $x = 0$ and $y = 0$. By considering values of C greater than or equal to C_0, and applying the result from the previous section, Birkhoff was led to the result he mentioned in the introduction to the manuscript and here described as a "provocative form of conclusion" (p. 35):

> *Assuming that a surface of section of the type stated exists for $C = C_0$, the non-existence for $C = C_0$ of doubly symmetric periodic orbits other than the fundamental variational periodic orbit of Hill and the corresponding retrograde orbit would imply that all possible lunar orbits whatsoever with $C = C_0$ have exactly the same mean rate of angular advance of perigee per synodic revolution.*

He further remarked in a footnote that, although the figures of the computed orbits show the initial assumption is valid, "a rigorous and mathematical proof might be a complicated and tedious matter!" In the summary he mentioned that he had "pointed out how the absence of infinitely many periodic orbits would indicate that a new *qualitative* integral exists, in addition to the usual analytic integral of Jacobi" [19, P. 600], but this remark was omitted from the manuscript.

Epilogue. The manuscript ended with a very short epilogue in which Birkhoff expressed the hope that his problems would "accelerate further advances," but admitted that he thought most of them were likely to "present difficulties which may be difficult to surmount" (p. 35).

3. CONCLUSION

So far little evidence has come to light of mathematicians responding directly to the summary of Birkhoff's lecture. Stanislaw Ulam, the Polish mathematician and emigré to the United States,[25] wrote to Birkhoff in November of 1941 to say that he had heard various reports of Birkhoff's "extremely interesting talk" and asked him for a copy of the summary.[26] And in the following January, Shizo Kakutani thanked Birkhoff for a reprint of

25 On Birkhoff's suggestion, Ulam had spent time at Harvard during 1936–1939. Later in 1939 Ulam left Poland for good in advance of the German invasion, and in 1940 was appointed to one of Birkhoff's former institutions, the University of Wisconsin-Madison, with the support of Birkhoff.

26 Letter from Ulam to Birkhoff, 25 November 1941. HUG 4213.2.2, Birkhoff Papers, Harvard University Archives.

the summary and said that he was "hoping to solve one of the problems."[27] But Kakutani did not say which one and he does not appear to have published on any of them. Had Birkhoff's manuscript been published, the situation might have been rather different. For it is only in the manuscript that the problems are set out in full and put formally into their mathematical context. The summary, being meant for a general audience, focusses on the historical rather than the mathematical detail. Indeed, the editor of *Science*, the psychologist James McKeen Cattell, told Birkhoff that he was "anxious to obtain papers on mathematical subjects" but that there were difficulties due to "the fact that the English used by mathematicians is not always understood by other scientific men," and so "complicated mathematical equations that only mathematicians can understand" must be avoided.[28] Furthermore, the fact that the summary was published in *Science* and not in a mathematical journal, and that it appeared during the War, meant it was unlikely to have had high visibility amongst mathematicians, particularly in Europe.

The manuscript is not an easy read and although Birkhoff makes several references to material in *Dynamical Systems* for purposes of clarification not everyone found the latter easy reading either. Walter Gottschalk, who became one of the leading exponents of topological dynamics, had this to say:

> *Somewhere I read that G. D. Birkhoff once said that if he thought mathematics exposition to be important, he would be the world's best expositor. Birkhoff was certainly not the world's best expositor and indeed he came close to the extremum in the other direction. I think this attitude had an important delaying effect on the initial development of topological dynamics. In his American Mathematical Society Colloquium volume* **[20]**, *Birkhoff included a discussion of the topological properties of continuous flows determined by a system of first order ordinary differential equations. … The style of writing he adopted was so inadequate in clarity and precision that almost any beginning reader had to be discouraged from continuing. It was not at all clear what the theorems were and the offered proofs were largely suggestive intuitive discussions* **[26]**.

It must also be said that Gottschalk himself was not always an easy read either.[29] Nevertheless, Gottschalk's criticisms did chime with the Russian view. In 2002 George Lorentz

---

27 Letter from Kakutani to Birkhoff, 26 January 1942. HUG 4213.2.2, Birkhoff Papers, Harvard University Archives.

28 Letters from McKeen Cattell to Birkhoff, 5 September 1941 and 8 October 1941. HUG 4213.2, Birkhoff Papers, Harvard University Archives.

29 Paul Halmos, when reviewing *Topological Dynamics* **[27]**, the book Gottschalk wrote together with his thesis supervisor Gustav Hedlund, remarked: "The chief fault of the book is its style. The presentation is in the brutal Landau manner, definition, theorem, proof, and remark following each other in relentless succession. The omission of unnecessary verbiage is carried to the extent that no motivation is given for the concepts and the theorems, and there is a paucity of illuminating examples." And he ended his review: "Conclusion: the book is a mine of information, but you sure have to dig for it." **[29]**.

recalled that Andrey Markov Jr., one of the editors of the original Russian edition of *Dynamical Systems*, "made sarcastic corrections of some of its errors" [32, P. 196].[30] Although in their preface the Russian editors urge a critical reading of the proofs—they don't think they have found all the mistakes—they do acknowledge the correctness of the theorems. Even Jürgen Moser in the introduction to the 1966 English edition conceded that "to the modern reader the style of [the] book may appear less formal and rigorous than it is now customary" while fully acknowledging its inspirational role [20, P. III]. Thus had Birkhoff's manuscript been published when he had hoped, it still may have taken some time before mathematicians were able to rise to the challenges laid down by his problems. Whether Birkhoff was right in his assessment of the direction of travel has yet to be ascertained and further research remains to be done in order to see the extent to which his problems have been tackled, if indeed they have, and to what effect.

ACKNOWLEDGMENTS

I am very grateful to the staff at Harvard University Archives for their help in negotiating the many metres of Birkhoff's archive. I also thank Jeremy Gray and Reinhard Siegmund-Schultze for their valuable comments and suggestions on an earlier version of this paper.

REFERENCES

[1] W. Ambrose, Representation of ergodic flows. *Ann. of Math. (2)* **42** (1941), 723–739.

[2] W. Ambrose and S. Kakutani, Structure and continuity of regular flows. *Duke Math. J.* **91** (1942), 25–42.

[3] D. Aubin and G. D. Birkhoff, Dynamical systems (1927). In *Landmark writings in Western mathematics, 1640–1940*, edited by I. Grattan-Guinness, pp. 871–881, Elsevier, Amsterdam, 2005.

[4] J. E. Barrow-Green, *Poincaré and the three body problem*. Amer. Math. Soc./Lond. Math. Soc., Providence, 1997.

[5] J. E. Barrow-Green, Gösta Mittag-Leffler and the foundation and administration of Acta Mathematica. In *Mathematics unbound: the evolution of an International Mathematical Research Community, 1800–1945*, edited by K. H. Parshall and A. C. Rice, pp. 265–378, Amer. Math. Soc./Lond. Math. Soc., Providence, RI, 2002.

[6] J. E. Barrow-Green, The dramatic episode of Sundman. *Historia Math.* **37** (2010), 164–203.

[7] J. E. Barrow-Green, An American goes to Europe: Three letters from Oswald Veblen to George Birkhoff in 1913/14. *Math. Intelligencer* **33** (2011), 37–47.

30 The 1966 English edition also includes corrections.

[8] G. D. Birkhoff, Proof of Poincaré's geometric theorem. *Trans. Amer. Math. Soc.* **14** (1913), 14–22; Démonstration du dernier théorème de géométrie de Poincaré, *Bull. Soc. Math. France* **42** (1914), 1–12.

[9] G. D. Birkhoff, Dynamical systems with two degrees of freedom. *Trans. Amer. Math. Soc.* **5** (1917), 199–300.

[10] G. D. Birkhoff, Surface transformations and their dynamical applications. *Acta Math.* **43** (1920), 1–119.

[11] G. D. Birkhoff, An extension of Poincaré's last geometric theorem. *Acta Math.* **47** (1925), 297–311.

[12] G. D. Birkhoff, On the periodic motions of dynamical systems. *Acta Math.* **50** (1927), 359–379.

[13] G. D. Birkhoff, Stability and the equations of dynamics. *Amer. J. Math.* **49** (1927), 1–38.

[14] G. D. Birkhoff, A remark on the dynamical rôle of Poincaré's last geometric theorem. *Acta Litt. Sci. Sect. Sci. Math., Szeged* **4** (1928), 6–11.

[15] G. D. Birkhoff, Einige Probleme der Dynamik. *Jahresber. Dtsch. Math.-Ver.* **38** (1929), 1–16.

[16] G. D. Birkhoff, Nouvelles recherches sur les systemes dynamiques. *Mem. Pontif. Acad. Sci. Novi Lyncaei* **1** (1935), 85–216.

[17] G. D. Birkhoff, Sur le problème restreint des trois corps. *Ann. Sc. Norm. Super. Pisa* **4** (1935), 267–306.

[18] G. D. Birkhoff, Sur le problème restreint des trois corps. *Ann. Sc. Norm. Super. Pisa* **5** (1936), 1–42.

[19] G. D. Birkhoff, Some unsolved problems of theoretical dynamics. *Science* **94** (1941), 598–600.

[20] G. D. Birkhoff, *Dynamical systems*. American Mathematical Society, Providence, 1966.

[21] G. D. Birkhoff, The restricted problem of three bodies. *Rend. Circ. Mat. Palermo* **39** (1915), 265–334.

[22] G. D. Birkhoff, Une généralisation à n dimensions du dernier théorème de géométrie de Poincaré. *C. R. Acad. Sci.* **192** (1931), 196–198.

[23] M. Brown and W. D. Neumann, Proof of the Poincaré–Birkhoff fixed point theorem. *Michigan Math. J.* **24** (1977), 21–31.

[24] L. Feraud, On Birkhoff's Pfaffian systems. *Trans. Amer. Math. Soc.* **32** (1930), 817–831.

[25] C. Golé and G. R. Hall, Poincaré's proof of Poincaré's last geometric theorem. Twist mappings and their applications. *IMA Math. Appl.* **44** (1992), 135–151.

[26] W. Gottschalk, The early history of general topological dynamics. In *Gottschalk's Gestalts* #12, Infinite Vistas Press, 2000.

[27] W. Gottschalk and G. Hedlund, *Topological dynamics*. American Mathematical Society, Providence, 1955.

[28] J. Hadamard, Notice nécrologique sur George David Birkhoff. *C. R. Acad. Sci.* **220** (1945), 719–721.

[29] P. Halmos, Topological dynamics. *Bull. Amer. Math. Soc.* **61** (1955), 584–588.

[30] H. P. Hudson, "Life begins …". *Univ. Chic. Mag.* (1941), 6–11.

[31] B. Koopman, Birkhoff on dynamical systems. *Bull. Amer. Math. Soc.* **36** (1930), 162–166.

[32] G. G. Lorentz, Mathematics and Politics in the Soviet Union from 1928 to 1953. *J. Approx. Theory* **116** (2002), 169–223.31.

[33] R. S. MacKay and J. D. Meiss, *Hamiltonian dynamical systems*. Adam Hilger, Bristol and Philadelphia, 1987.

[34] A. G. Maier, Sur un problème de Birkhoff. *C. R. Acad. Sci. URSS* **55** (1947), 473–475.

[35] M. Morse and G. David, Birkhoff and his mathematical work. *Bull. Amer. Math. Soc.* **52** (1946), 357–391.

[36] J. C. Oxtoby and S. M. Ulam, Measure-preserving homeomorphisms and metrical transitivity. *Ann. of Math. (2)* **42** (1941), 874–920.

[37] H. Poincaré, Sur le problème des trois corps et les équations de la dynamique. *Acta Math.* **13** (1890), 1–270.

[38] H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste, vol. III*. Gauthier-Villars, Paris, 1899.

[39] H. Poincaré, Sur un théorème de géométrie. *Rend. Circ. Mat. Palermo* **33** (1912), 375–407.

[40] D. van Dalen, *L. E. J. Brouwer. Topologist, Intuititionist, Philosopher*. Springer, London, 2013.

[41] H. Whitney, Regular families of curves. *Ann. of Math. (2)* **34** (1933), 241–279.

[42] N. Wiener, *Ex-prodigy: my childhood and youth*. The MIT Press, Cambridge, 1953.

[43] E. B. Wilson, George David Birkhoff. *Science* **102** (1945), 578–580.

[44] A. Wintner, *The analytical foundations of celestial mechanics*. Princeton University Press, Princeton, 1941.

[45] J. D. Zund, George David Birkhoff and John von Neumann: a question of priority and ergodic theorems, 1931–1932. *Historia Math.* **29** (2002), 138–156.

JUNE BARROW-GREEN

School of Mathematics and Statistics, Faculty of STEM, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK, june.barrow-green@open.ac.uk

SOME USES AND ASSOCIATIONS OF MATHEMATICS, AS SEEN FROM A DISTANT HISTORICAL PERSPECTIVE

ANNETTE IMHAUSEN

ABSTRACT

The article presents the evolution of mathematics and its various uses in ancient Egypt.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 01A16; Secondary 01A11, 01A80

KEYWORDS

History of mathematics, ancient Egypt

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5772–5782
DOI 10.4171/ICM2022/170

1. INTRODUCTION

Today mathematics constitutes an integral part of our lives on a variety of levels. For those present at this conference (and many others), the discipline of mathematics is their chosen profession which is subdivided in a multitude of individual areas. In addition, scientists other than those of the mathematical sciences also use its concepts and methods in their professional lives whenever they work with data that can be represented numerically. Thus, the education in these sciences often includes mathematical training that is specifically tailored to their needs. Furthermore, mathematics also plays an important role in everybody's daily life using numerical practices and mathematical measuring in many contexts, for example, in structuring the day by means of measuring time or by quantifying the many things like items of food or other objects. Although, as many mathematicians have encountered from nonmathematicians when asked what they do for a living, some seem to take pride in "always having been bad at mathematics." But even for these people, it would be hard to imagine them getting by without it. The unenthusiastic statement thus may less reflect on the subject itself than on the methods used in teaching it. Several attempts to present the appeal of mathematics to a general audience have succeeded, for example, the volume *The Mathematical Experience* (1981) by Philip Davis and Reuben Hersh [4] which won the National Book Award in 1983. These days, mathematics influences various areas of our lives as has been demonstrated with numerous examples by the Cambridge mathematician Tom Körner in his book *The Pleasures of Counting* (1996) [8]. While the examples chosen by Tom Körner mostly originate from a more modern context the history of mathematics has been traced back to begin as early as in the 3rd millennium BCE with the invention of writing in Egypt and Mesopotamia.

My own research has always focused on ancient Egyptian mathematics, which can provoke an ambivalent reaction from historians of mathematics. On the one hand, Egypt (along with Mesopotamia) provides one of the earliest types of evidence for mathematical concepts and techniques in history. On the other hand, ancient Egyptian mathematics has been judged by some of its more recent readers as lagging behind its potential, and the occasional judgment went as far as blaming their clumsy mathematical techniques for a lack of development in other areas that would have depended on them. While the history of mathematics was initially often performed as a judgmental assessment of mathematical sources read from a modern point of view, (most) historians and (most) mathematicians alike have since realized that it is much more rewarding to attempt learning about mathematical concepts of a historic period as they were used and practiced in their time. It may be trivial to solve an ancient mathematical problem with a modern mathematical toolbox, and thus equally boring. However, to learn how this problem was thought of in its time and tackled with the then available tools is as complex as it is rewarding. This has consequences for the methodology deemed appropriate in the history of mathematics. It also has consequences for the evaluation of ancient mathematical cultures. While this has been the focus of some of my earlier work [6], this contribution has a different aim, trying to bridge gaps between periods and cultures. Because then as today the practice of mathematics has had multiple

aspects as described in the initial sentences and it is through an analysis of these aspects that parallels as well as differences between ancient and modern mathematics can be seen. My contribution will look at some of these aspects from the mathematics of ancient Egypt.

2. INVENTION OF NUMBER SYSTEM AND ITS EARLY USES

The earliest evidence we have for ancient Egyptian numbers comes from a graveyard in Abydos. The stratification of society of ancient Egypt can be traced by the sizes and contents of tombs in graveyards. Cemetery U at Abydos, situated to the north of the Early Dynastic royal tombs, contains approximately 600 tombs and covers the entire Predynastic period; see Hartung [5, P. 317]. The tomb Uj of that cemetery has become rather well known in Egyptology for the about 100 incised tags that were found within it. The incisions were either pictograms resembling later hieroglyphic signs or groups of strokes or a coiled rope resembling later Egyptian number notation. All of the tags had a hole, which points to them having been attached to some sort of goods about which they would provide information of some kind. They are considered the first evidence of writing from ancient Egypt; see Baines [1]. Despite the difficulties of interpreting these earliest sources, the tags may point to a parallel to Mesopotamia in the invention of writing and the number system, where this was prompted by administrative needs; see Robson [15]. In case of the tags found in the tomb Uj, it is possible to interpret them as some sort of administrative practice denoting, for instance, owner or provenance and quantities, a practice that was presumably first used in daily life and then as well in the context of burials. At this point it is noteworthy that as in Mesopotamia, the first evidence of writing from ancient Egypt includes as well the first evidence of numbers.

Evidence from the Old Kingdom indicates that the use of what we think of as mathematics was located in the context of administration. While there are no mathematical texts from this period in ancient Egypt, two archives of that time show that numbers and metrological units were used to audit resources of larger structures like temples; see the work of Posener-Kriéger and Demichelis [12,13]. Later evidence points to an ongoing tradition of this administrative context for the practice and development of mathematics. The Lahun Papyri, the largest papyri find from the Middle Kingdom, include a large number of accounts; see Collier and Quirke [3]. Evidence from all periods of ancient Egypt indicates that numbers, metrology, and mathematical practices were used to monitor and audit resources of various kinds (goods as well as manpower). The Egyptian systems of numeration and metrologies constitute the basis of this administrative oversight.

However, even then there were further uses of numbers. An early object that displays the Egyptian number system is the ritual mace-head of king Narmer, the last king of dynasty 0 (c. 3000 BCE). The decoration of the mace-head includes the picture of a tribute of 400,000 oxen, 1,422,000 smaller cattle, and 120,000 prisoners that were presented to king Narmer. The mace-head illustrates that the number system with its symbols to indicate powers of ten up to one million was in use as early as in dynasty 0. The large numbers indicated on the mace-head likely do not indicate quantities of an actual tribute, but were probably meant to represent the power of this king through their size. These first early examples of numbers

both originate from a royal context. The restricted use of writing and numbers in the service of the Egyptian king and his institutions is another characteristic that has to be kept in mind when analyzing the mathematics of ancient Egypt.

In the control of the king's resources, a professional group emerged whose most common name denotes its crucial skill: scribes. Scribes were active in the administration. The complex system of responsibilities is reflected in a variety of individual titles that are attested since the Old Kingdom; see Jones [7]. The basic requirement for the profession of scribe was the ability to read and write. It can be assumed that this was passed down from father to son. Likewise, it can be assumed that "writing" included not only the handling of script but also the handling of numbers. The following sections of this paper will outline the use of mathematics by scribes in the Old, Middle, and New Kingdoms as well as mathematical applications in Egyptian culture beyond the realm of administration.

3. THE OLD KINGDOM: EARLY EVIDENCE OF MATHEMATICAL CONCEPTS AND PRACTICES

The first period that is recognized as a period of remarkable cultural achievements in ancient Egypt is the Old Kingdom (c. 2686–2160 BCE). During the Old Kingdom, Egyptian kings were buried in pyramids. Several kings of the 5th Dynasty chose an area northwest of the modern village of Abusir for their pyramids. In the late 19th century, papyri were purchased by several museums that came from the administration of the cult for one of these kings. Taken together, these papyri constitute the largest papyrus find from the Old Kingdom. The Abusir papyri contain, among other things, accounts and papyri from the cult operations that took place there, such as service lists of priests, ration lists, lists of sacrifices for temples, and others. They are the oldest preserved papyri that are extant. They also contain information about the further development of Egyptian mathematics in the form of the creation of metrological systems. The ration lists contain quantities of grain, meat, and beer, each given in the corresponding units. Also attested in these papyri is the use of tables, which are recognizable as such not only by tabular arrangement of entries, but also by formatting with rows and columns marked by line and column headings. From these texts we obtain information about the activity of the scribes who, at temples or at the royal court, prepared these accounts and lists and through their work performed the administration of the empire, its resources and especially its goods produced at the lower levels.

Further information about these persons, who were responsible for the control of important parts of the Egyptian productions, for example, of bread and beer, by means of mathematical techniques, can be obtained from their tombs. There we find representations of the production processes, such as the bakery and brewery, in which the scribes overseeing them are a central part. For example, in the tomb of Nianchchnum and Chnumhotep, the production of bread and beer is depicted from the allocation of the grain to the delivery of the products; see Moussa and Altenmüller [10, PLATES 23, 26, 34–35]. When the grain is allotted, it is shown how one person measures the flour, another counts the amount of flour that is taken, a scribe keeps a record of it, and one person receives it. This is followed by the

depiction of the production of bread and beer from the flour. Finally, the delivered products are measured and noted by a scribe. In the depiction of the delivery, the bakers are shown bringing the products, as well as an auditing process, the result of which is noted down by a scribe.

This detailed audit of resources is also evident in the later mathematical texts of the Middle Kingdom. In these texts, a large number of problems deal with the management and control of resources; in particular, the number of bread-and-beer-problems, which at 19 accounts for slightly less than 1/5 of all mathematical problems, is striking. Despite the lack of evidence in form of mathematical texts of the Old Kingdom, the papyri and representation of daily life practices provide ample evidence for the further advancing of mathematics in ancient Egypt. Metrological systems for a variety of goods were created and used and thus a mathematical record of resources established.

Apart from the depictions of scribes in the context of executing numerical processes, the tombs of high-ranking officials also contain the so-called autobiographies of the tomb-owner. Ancient Egyptian autobiographies differ fundamentally from modern autobiographies. An ancient Egyptian autobiography refers to a text written in the 1st person singular describing the career and proper actions of the tomb's owner. It served to vindicate the tomb owner and represented a guarantee that he would pass the judgment of the dead by proving that he had lived his life according to the moral principles of ancient Egypt. Consequently, no negative events appear in the autobiography, the path of life is described as a continuous ascent to ever new responsibilities. In comparison to modern autobiography, personal areas of life, such as family and the expression of emotions, are missing.

Egyptian autobiographies are attested throughout pharaonic history and allow us to trace changes in the perception and self-assessment of the scribes' profession. By describing the professional activities of scribes, they also provide evidence for the role of mathematics in the life of a scribe. For the Old Kingdom, the autobiography of Weni from the 6th Dynasty is one of the most remarkable examples—not least because Weni served successively under three pharaohs, Teti, Pepi I, and Merenre. Weni is thought to have been active for a period of 70 years, his career must have begun in his youth. The text of his autobiography was written in carefully executed hieroglyphics on a stone slab that once formed the wall of a one-room burial chapel in Abydos (for a translation cf. Lichtheim [9, PP. 18–23]). The autobiography of Weni lists services performed by Weni for the king and rewards that Weni received in return. These rewards were either in the form of material goods or in the form of a promotion to a higher position. Particular emphasis is placed on activities in which Weni acted entirely on his own—demonstrating a special form of trust in him by the king. In this autobiography, explicit reference is made at least once to mathematical activities. Weni says of himself that he counted everything that could be counted in Upper Egypt twice for the king and that he counted every activity twice that had to be done for the Upper Egyptian residence as well. Through their accounting, the scribes monitored all kinds of resources for their king which placed themselves in a position of power. During the Old Kingdom, the autobiographies document that the level of success of a scribe was measured by his proximity to the king whom he served.

Summing up, the evidence from the Old Kingdom indicates the proliferation of using mathematics in the organization of resources and the evolution of a professional group that used mathematical practices. Some members of that group held exalted positions within Egyptian society as is indicated by the size and endowment of their tombs.

During the First Intermediate Period (2160–2055 BCE), the central rule of Egypt by a single king broke down. This breakdown seems to have been affected by several causes, which taken together could not be overcome by the former royal authority. The Egyptian literature of the following Middle Kingdom successfully attempted to reestablish royal authority by presenting how the new kings had overcome this dark age and its difficulties. The description of the First Intermediate Period given from a later royal point of view has led Egyptologists initially to assess this period as a dark one, associated with social and political instabilities, an assessment that probably holds for some years of the First Intermediate Period. Its beginning, with famines caused by climatic changes and the failure of the former king to maintain control over all of Egypt, must have been a drastic change for the Egyptian population. However, the autobiographies of some local nomarchs indicate that these problems were then mastered. During the First Intermediate Period, while the central administrative framework was lacking, the individual nomarchs presumably continued their administrative roles towards the population of their towns or regions. The individual success of a nomarch, as it is expressed in the autobiographies, was measured through his ability to ensure social and economic stability within his own region and through his conduct towards the weaker members of society. In the work of an official, mathematical knowledge must have played an important role in order to assess (for example) the available grain rations. Using this knowledge, however, was no longer perceived as a service to the king. Instead, the nomarchs saw themselves as installed through the power of the gods.

4. THE DISCIPLINE OF MATHEMATICS IN THE MIDDLE KINGDOM (AND AFTER)

As is well known, several papyri about mathematics are extant from the Middle Kingdom (c. 2055–1650 BCE) and the Second Intermediate Period (c. 1650–1550 BCE). The most famous of them are the Rhind Mathematical Papyrus [11] and the Moscow Mathematical Papyrus [16]. The Rhind papyrus, named after Alexander Henry Rhind, the Scottish lawyer who purchased the text, is kept today in London in the British Museum. The Moscow Mathematical Papyrus is named after the city of its current location in the Pushkin State Museum of Fine Arts.

Their initial appearance at this point in time may, of course, be due to the haphazard circumstances of preservation. However, they may also reflect the conscious attempt of the Middle Kingdom rulers to reestablish control over administrative structures that they had lost during the First Intermediate Period. Egyptian mathematical texts may contain mathematical problems and their solutions and mathematical tables used for fraction reckoning and conversion of measures. The Rhind Mathematical Papyrus, being in two pieces, has the inventory numbers BM 10057 and BM 10058. BM 10057 measures 295.5 cm by 32 cm,

BM 10058 199.5 cm by 32 cm. The gap between both pieces is assumed to be approximately 18 cm. The Rhind Mathematical Papyrus provides a corpus of some 70 problems and several tables. Most of the problems are grouped together according to their content.

The Moscow Mathematical Papyrus is the second largest extant source text. While its total length is approximately 544 cm, its height is only 8 cm. It consists of one big piece and nine little fragments. The Moscow Papyrus contains 25 problems, of which the first three are too damaged to determine more than a probable type of problem. In addition, unlike the problems of the Rhind Papyrus, those of the Moscow papyrus were not arranged in groups of problems according to their content but seem to be written down in no apparent order. The Moscow papyrus also holds two duplicate problems in numbers 8 (identical with problem no. 5) and 13 (identical with problem no. 9). However, among the problems of the Moscow papyrus are two of the most interesting for historians of mathematics, problem no. 10 about the area of a curved surface and problem no. 14 about the volume of a truncated pyramid.

Egyptian mathematical problem texts are written in a distinct style: A problem text begins by stating a mathematical problem (title). After the type of problem is announced, some specific data in the form of numerical values are given, thus specifying the problem to one concrete instance or object. This is followed by instructions (the procedure) for its solution. Title, and specifications of the problem and the following instructions are expressed in prose, using no mathematical symbolism. The title (and other parts of the text) may be accentuated by the use of red ink. Each instruction usually consists of one arithmetic operation (addition, subtraction, multiplication, division, halving, squaring, extraction of the square root, calculation of the inverse of a number) and the result of it is given. The instructions always use the specific numerical values assigned to the problem. Abstract formulas, or equations with variables did not exist. This style, which is also used in Mesopotamian problem texts, has been characterized by Jim Ritter as rhetoric, numeric, and algorithmic [14, P. 44].

Historiographic assessment of Egyptian mathematics has followed two paths so far, first the description of the mathematics found in these texts by means of modern mathematical terminology. In this line of inquiry, Egypt was praised to provide very early evidence for "algebraic equations" and other early mathematical knowledge, like that of calculating the area of certain geometric shapes. In comparison with contemporary evidence from Mesopotamia, however, it fell short and, compared with later evidence from ancient Greece, it was lacking the feature of general mathematical theorems and their proofs. The second, more sophisticated line of inquiry of research on ancient Egyptian (and Mesopotamian) mathematics attempts to understand mathematical structures within the source texts, e.g., by assessing and comparing the procedures used in solving mathematical problems. Again, Mesopotamia has fared somewhat better than Egypt, which may, however, be the result of the very different quantities of sources available. This line of inquiry is not yet exhausted at this point.

The contents of the problems enforce the impression that the context of ancient Egyptian mathematics remained within the area of administration. However, while most of the problems can be understood as mathematical solutions to actual administrative tasks, several problems point to an awareness of mathematical knowledge as such, for example, prob-

lem 79 of the Rhind Papyrus that asks to compute the total of a number of items comprised of a house and cats, mice and cereal found within. Likewise, the phrasing of problem 67 of the Rhind Papyrus, ostensibly computing the produce of a herdsman, points to the existence of a different type of mathematical setting, comparable with so-called recreational mathematics.

Due to the fragmentary state of preservation, only two titles of mathematical papyri are known. Of these two, one fits the assessment of mathematics as a tool in administration. Fragment UC32162 of the Lahun Mathematical Papyri contains fragments of two problems, a calculation of areas and a calculation of the produce of a fowler. Before the text of these problems a title reading "[Method] of calculating the matters of accounting" is extant. The other title is found at the beginning of the Rhind Mathematical Papyrus, and reads "Method of calculating for inquiring into nature, and for knowing all that exists, [every] mystery, […] every secret" which seems to point to an appreciation of mathematical knowledge that exceeds its simple utility in administration.

Thus, while the disciplinary context of mathematics in ancient Egypt remains within administration, the recognition that its application may be wider than the accounting of resources is indicated by the content and title of the Rhind Papyrus. Sources from later periods provide examples of these further applications.

5. FURTHER USES OF MATHEMATICAL CONCEPTS AND PRACTICES IN ANCIENT EGYPT

After the Second Intermediate Period there is a lack of sources as far as mathematical texts are concerned. However, mathematics still features quite prominently in the lives of the scribes as literary texts indicate. Instead of school books of individual subjects, a variety of texts, which were presumably circulated among New Kingdom scribes, is extant. They include compositions describing the superiority of the scribal profession over any other profession, eulogies to scribal teachers, and model letters. This corpus of texts is referred to as the *Late Egyptian Miscellanies*. At least some of these texts include references to mathematical practices. The theme of scribal superiority above all other professions is the topic of the following excerpt, section 4,2–5,7 of Papyrus Lansing, which was titled "All callings are bad except that of the scribe" by its first translator, Ricardo Caminos **[2, PP. 384–385]**. Reference to scribal work is made twice within this section, first when describing how the profit of the merchants is taken away by the tax-people (scribes!), and again at the end of the passage when the profession of the scribe is compared to the aforementioned professions: "But the scribe, it is he that reckons the produce of all those." Both of these references are with respect to the mathematical abilities of the scribe, who has to calculate the taxes of the merchants before carrying them off and who also calculates the output of the other professions, presumably to determine their taxes. These references provide evidence for the ongoing use of mathematics in administration as well as the gains that those proficient in it were to expect. Likewise, mathematics features in the text of Papyrus Anastasi I, a fictional letter from the context of a learned debate between two scribes. The debate includes a set of mathematical problems: the calculation of bricks needed to construct a ramp; the number of

workers needed to transport an obelisk; the number of workers needed to move sand when a colossal statue has to be erected in a given time; and the calculation of rations for a military excursion. Although these problems are phrased like their earlier counterparts of the mathematical texts, the numerical information given in Papyrus Anastasi I does not suffice to actually solve these problems. Their intention may have been to remind the numerate reader of his mathematical education. To us, the text is a source that provides us with an idea of the variety of numerate tasks that a scribe had to master. In addition, it also informs us about areas of their profession that scribes thought of as meaningful and valuable. Thus, although there is practically no evidence for mathematical texts, administrative documents and evidence from literary texts leave no doubt that mathematics continued to play an important role in the scribal profession during the New Kingdom.

In addition, by this time mathematics had also acquired another function in ancient Egyptian society. Not only did it provide the means to perform administrative tasks, but it did so in a way that was considered to fulfill the requirements of acting according to the Egyptian moral code. The normative framework of this moral code is expressed by the Egyptian term *Maat*. This term comprises the idea, that there is a certain perfect order of the cosmos and everything in it. Therefore, the term *Maat* is closely linked or may express ideas of truth, order and justice. From the idea of a perfect order of the cosmos then follow certain codes of conduct to which every Egyptian was supposed to adhere in his or her daily life. For some literate members of Egyptian society, the respective rules were explicitly stated in a genre of ancient Egyptian literature called teachings or instructions. Extant Teachings (with settings from the Old Kingdom on) provide examples for scribes (*Loyalist Teaching*, *Teaching of Khety*, *Instruction of Any* and *Instruction of Amenemope*), viziers (*Instruction addressed to Kagemni* and *Teaching of the Vizier Ptahhotep*), princes (*Instruction of Prince Hardjedef*) or even kings (*Teaching for King Merikare*, *Teaching of King Amenemhat*). From the four examples of teachings addressed to scribes, *The Teaching of Amenemope*, includes several references to mathematics beginning with the introduction of its fictional author Amenemope, which identifies him as the person who controls the measuring of agricultural affairs including the registration of land and audit of the vessels used to measure grain. The authority of the scribe, formally provided by his being in the service of the king, de facto originates from his numerical and metrological expertise, which enable him to execute the tasks mentioned in his introduction. Further references to numerical and metrological duties occur throughout the following 30 chapters of instructions. It is explicitly mentioned that a scribe must not "falsify the temple rations" (Chapter 5), "move the markers on the borders of fields, or shift the position of the measuring-cord," "be greedy for a cubit of land, or encroach on the boundaries of a widow" (Chapter 6), "move the scales or alter the weights, or diminish the fractions of the measure," "desire a measure of the fields, or neglect those of the treasury," "make for himself deficient weights" (Chapter 16), "disguise the measure, so as to falsify its fractions," and "force it (the measure) to overflow, or let its belly be empty" (Chapter 17). The teaching illustrates, on the one hand, the role that was by now assigned to mathematics, i.e., to provide justice, and, on the other hand, it indicates the awareness that consisted in a dishonest scribe who would falsify mathematical results. In depictions of metrological

practices and administrative texts that document them, this awareness is also reflected, since it seems to have been the rule that there was rarely a single scribe entrusted with measuring and recording the respective results, but usually a group of scribes who would then check each other's work. In the context of this ensuring of justice, it is noteworthy that the issue of setting numerical values, e.g., the amounts of produce that were expected of a worker, is never discussed nor addressed explicitly. The king was simply expected to execute his power according to the rules of *Maat*.

Mathematical practices also were used in another context that provides evidence for the concurrence of mathematics and justice in ancient Egypt, namely the judgement of the dead. In order to prove worthy for the afterlife, the deceased had to pass judgement of his way of life. In order to be successful, the deceased first had to recite sins that he did not commit during his lifetime. Then his heart was weighed on a balance against a feather, a symbol of the goddess Maat. If the balance showed equilibrium between the heart and the feather, the judgement was passed successfully and the dead would be presented to Osiris. It was the mathematical operation of weighing that ensured a just judgement of the deceased.

6. CONCLUSION

From its first beginnings in Egypt and Mesopotamia, mathematics as a discipline has made immense progress, the history of which is traced in the history of mathematics. If this is done in a historically correct way, it can provide fascinating glimpses into a variety of mathematical cultures. At the same time, mathematics has at all times remained a key element in our daily lives. The two aspects have always been intertwined.

In using mathematics for daily life purposes, the aspect of justice and reliability are key characteristics that can be demonstrated as early as in ancient Egypt, and presumably ever since. However, as history and especially the crisis of recent years indicate, using mathematics does not in itself guarantee success. Ancient Egyptian sources indicate an awareness of the potential weaknesses of measuring on an abstract level (as seen in the teachings) as well as in practice (as seen by the depiction of several scribes that are meant to perform the measuring and thus are meant to check each other). The success of mathematical practice therefore depended on two aspects, the quality of a mathematical technique that was developed to solve a given problem and the quality of its execution by its practitioners.

REFERENCES

[1] J. Baines, The earliest Egyptian writing: development, context, purpose. In *The first writing. Script invention as history and process*, edited by S. D. Houston, pp. 150–189, Cambridge University, Cambridge, 2004.

[2] R. A. Caminos, *Late-Egyptian miscellanies (Brown Egyptological Studies I)*. Oxford University Press, London, 1954.

[3] M. Collier and S. Quirke, *The UCL Lahun Papyri: accounts*. BAR. Int. Ser. 1471, Archaeopress, London, 2006.

[4] P. Davis and R. Hersh, *The mathematical experience*. Birkhäuser, Boston, 1981.

[5] U. Hartung, Cemetery U at Umm el-Qaab and the funeral landscape of the Abydos region in the 4th millennium BC. In *Desert and the Nile. Prehistory of the Nile Basin and the Sahara Papers in honour of Fred Wendorf*, edited by J. Kabaciński, M. Chłodnicki, M. Kobusiewicz, and M. Winiarska-Kabacińska, pp. 313–338, Stud. Afr. Archaeol. 15, Poznań Archaeological Museum, Poznań, 2018.

[6] A. Imhausen, *Mathematics in Ancient Egypt. A contextual history*. Princeton University Press, Princeton, 2016.

[7] D. Jones, *An index of Ancient Egyptian titles, epithets and phrases of the Old Kingdom*. Archaeopress, Oxford, 2000.

[8] T. W. Körner, *The pleasures of counting*. Cambridge University Press, Cambridge, 1996.

[9] M. Lichtheim, *Ancient Egyptian literature, Volume I: The Old and Middle Kingdoms*. University of California Press. Berkeley and Los Angeles, 1975.

[10] A. Moussa and H. Altenmüller, *Das Grab des Nianchchnum und Chnumhotep*. Philipp von Zabern, Darmstadt, 1977.

[11] T. E. Peet, In *The Rhind Mathematical Papyrus, British Museum 10057 and 10058*, The University Press of Liverpool and Hodder & Stoughton, London, 1923.

[12] P. Posener-Kriéger, *Les archives du temple funéraire de Néferirkaré-Kakaï: Les papyrus d'Abousir. Traduction et commentaire*. Bibl. Étude 65, Institut français d'archéologie orientale du Caire, Cairo, 1976.

[13] P. Posener-Kriéger and S. Demichelis, *I papiri di Gebelein: scavi G. Farina 1935*. 2004. *I papiri di Gebelein: scavi G. Farina 1935*. Turin: Ministero per i Beni e le Attività Culturali, Soprintendenza al Museo delle Antichità Egizie, Turin, 2004.

[14] J. Ritter, Chacun sa verité: les mathématiques en Égypte et en Mésopotamie. In *Éléments d'histoire des sciences*, edited by M. Serres, pp. 39–61, Bordas, Paris, 1989.

[15] E. Robson, Literacy, numeracy, and the state in early Mesopotamia: social and cultural practices. In *Literacy and the state in the Ancient Mediterranean*, edited by K. Lomas, R. D. Whitehouse, and J. B. Wilkins, pp. 37–50, Accordia Research Institute, London, 2007.

[16] W. W. Struve, *Mathematischer Papyrus des Staatlichen Museums der Schönen Künste in Moskau*. Quellen Stud. Gesch. Math., Abt. A: Quellen 1, Springer, Berlin, 1930.

ANNETTE IMHAUSEN

Arbeitsgruppe Wissenschaftsgeschichte, Historisches Seminar, Goethe-Universität Frankfurt, Norbert-Wollheim-Platz 1, 60323 Frankfurt am Main, Germany, annette.imhausen@normativeorders.net

THE HISTORY AND HISTORIOGRAPHY OF THE DISCOVERY OF CALCULUS IN INDIA

K. RAMASUBRAMANIAN

ABSTRACT

Weaving through the emergence and convergence of various mathematical ideas that led towards the discovery of calculus in India provides an enthralling experience for aficionados of mathematics and its diverse history. This article attempts to briefly capture some of the milestones in the journey made by Indian mathematicians through two eras that paved the way for the discovery of infinite series for π and some of the trigonometric functions in India around the middle of the 14th century. In the first part we shall discuss the developments during what may be called the classical period, starting with the work of Āryabhaṭa (c. 499 CE) and extending up to the work Nārāyaṇa Paṇḍita (c. 1350). The work of the Kerala School starting with Mādhava of Saṅgamagrāma (c. 1340), which has a more direct bearing on calculus, will be dealt with in the second part. The third part recounts the story of the 19th century European discovery of infinite series in India which seems to have struck a wrong note among the targeted audience in Europe with a serious cascading effect.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 01A32; Secondary 01A85

KEYWORDS

Discovery of calculus, infinite series, Kerala school of mathematics, Mādhava, Āryabhaṭa, Charles Whish, historiography

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5784–5812
DOI 10.4171/ICM2022/77

1. INTRODUCTION

Couched in sublime poetry in a variety of rhythmic meters and codified in the classical Sanskrit language, a journey through the history of mathematics in India could be extremely fascinating and at the same time quite challenging too. The journey would indeed be highly enriching to those who have mastered the language and understood the subtlety of expressions and figures of speech employed in it. However, for those untrained in the nuances of such a knowledge system, it would be difficult to appreciate the beautiful blend of mathematics and poetry—usually characterized with brevity without sacrificing the perspicuity—that we find in most of the texts composed over the last two millennia. The distinct style adopted by the Indian mathematicians for practising (thinking, codifying, transmitting, etc.) mathematics, by directly plunging into results without much mathematical elaborations, has been succinctly and beautifully brought out by A. A. K. Ayyangar in his article **[17, P. 4.101]**:[1]

> *The Hindu mind has always shown peculiar aptitude for fundamental thinking, digging down into the depths of thought with the minimum of external equipment, while other minds are after heavy superstructures with complicated scaffolding, tools and machinery. One extra-ordinary illustration of this trait of the Hindu mind we have in Ramanujan.*

Perhaps being fascinated by this peculiar way of doing mathematics by Hindus, using poetic verses, and aphoristic expressions, some of the of European scholars who were serving the British establishment in various capacities—starting from the final decades of the 18th century—embarked on their journey to study the civilizational basis of India, and the route adopted by Indians to excel in mathematics and astronomy, besides arts, architecture, aesthetics, philosophy and other disciplines.[2]

One such European scholar who got deeply attracted towards the mathematics and astronomy of the Kerala School was the then civil servant of the East India Company, Charles M. Whish (1792–1833). Having been posted at the Malabar region of Kerala for more than a decade, Whish started interacting with the local pundits and gained proficiency in both the local language Malayalam and Sanskrit. He also began to communicate some of his fascinating findings concerning the breakthroughs made by the native astronomers of Kerala, by way of both authoring papers and sharing them with the Madras Literary Society. A remarkable paper of his carrying the details of signal contributions made by the Kerala School of mathematicians, which flourished during the medieval period (14–16 centuries CE), got published in the Transactions of Royal Asiatic Society of Great Britain and Ireland in 1834—unfortunately, only posthumously—due to his premature death in 1833.

---

1 Ayyangar, who came out in flying colors, with his Master's degree at the age of 18 years, has done remarkable research particularly with respect to second-order indeterminate equations.

2 See, for instance, **[12]**.

It is this paper, which for the first time brings to the notice of European scholars the discovery of the infinite series for π, and some of the trigonometric functions by the Kerala mathematicians, almost three centuries before their advent in Europe. Strangely, this paper of Whish, instead of generating curiosity, discussion, and excitement among the European scholars, remained largely disregarded for almost a century. This deafening silence—along with the discount of its contents, among the historians of mathematics in the West—got broken only in the decades to follow from the 1940s, when some of the Indian mathematicians such as C. T. Rajagopal, Mukunda Marar, and others brought to fore the sophisticated mathematics produced by this school in the form of a series of articles [22, 23, 25, 26]. During the same period, Ramavarma Thampuran and Akhileswara Ayyar also brought out an edition of the first part (dealing with mathematics) of seminal text of Kerala astronomy and mathematics, *Yuktibhāṣā* (c. 1530), along with detailed explanations in Malayalam [28].

The Kerala School that we refer to in this article commences with Mādhava of Saṅgamagrāma (c. 1340–1420), the originator of this *guru-paraṃparā* or "lineage of teachers." His followers include Dāmodara, Parameśvara, Nīlakaṇṭha Somayājī, Jyesṭhadeva, Śaṅkara Vāriyar, and others. Though Mādhava's works containing the infinite series are not available to us, the later mathematicians in this tradition unanimously ascribe the series to Mādhava. In some of the recent studies, it has been convincingly argued by modern scholars that these series expansions for π and other trigonometric functions, and the evaluation of derivatives of various functions (while computing instantaneous velocities) rely indispensably on the central ideas of infinitesimal calculus, which include local approximation by linear function (see Section 3.4 of the present article).[3]

It is, however, important to understand that these breakthroughs achieved in the Kerala School of Mathematics cannot be narrowed to only the scope of work made in a span of two centuries. It is the continuum of mathematical ideas evolved by various Indian mathematicians spanning over nine centuries before—starting at least from the time of Āryabhaṭa (5th century)—till the dawn of the Kerala School that has led to the convergence point which has led Mādhava (14th century) to invent infinitesimal methods, thereby marking the advent of the discipline of calculus, though largely restricted to the consideration of the circular functions.[4]

This paper attempts to string the pearl of ideas and breakthroughs through the history of mathematics in India that led to this advent. The evolution of poignant ideas is traced in two parts. The first part, covered in Section 2, deals with precalculus breakthroughs and the germinating ideas for calculus that were intuitively apprehended in India well before Mādhava came on the scene. The second part, dealt with in Section 3, captures the discovery of calculus in the Kerala School. Section 4 of this paper recounts the story of how the revelations of the work of the Kerala School brought out by Whish seems to have struck a wrong note and alarmed some of the leading figures in the British academic establishment which led to the denigration and suppression of this work for almost a century.

---

3 The reader is also referred to the articles [13, 14, 24].

4 For a detailed discussion on this evolution readers may refer to [15, 27].

2. DEVELOPMENTS IN THE CLASSICAL ERA OF INDIAN MATHEMATICS

In this section, we shall consider some of the ideas and methods developed in Indian mathematics, during the period 450–1350 CE, which have a bearing on the later work of the Kerala School. In particular, we shall focus on the following topics: the notion and mathematics of zero and infinity; iterative approximations for irrational numbers; summation of powers of natural numbers; the discrete form of the harmonic equation for the sine function given by Āryabhaṭa; and the emergence of the notion of instantaneous velocity of a planet in astronomy.

2.1. Notion of zero and infinity

2.1.1. Philosophical and cultural context of zero and infinity

Select passages in *Upaniṣads*, as well as contemporary Buddhist and Jaina philosophy, point to the philosophical and cultural context that has possibly led to the development of the fundamental and intriguing concepts such as void and the infinite which later got incorporated in mathematics as zero and infinity. In this section, we present quotes from different ancient literature in this regard.

The *śānti-mantra* of the *Īśāvāsyopaniṣad* refers to the ultimate absolute reality, the *Brahman*, as *pūrṇa*, the perfect, complete or full. Talking of how the universe emanates from the *Brahman*, it states:

> पूर्णमदः पूर्णमिदं पूर्णात्पूर्णमुदच्यते।
> पूर्णस्य पूर्णमादाय पूर्णमेवावशिष्यते॥
> *pūrṇamadaḥ pūrṇamidaṃ pūrṇātpūrṇamudacyate।*
> *pūrṇasya pūrṇamādāya pūrṇamevāvaśiṣyate॥*
>
> That (*Brahman*) is *pūrṇa*; this (the universe) is *pūrṇa*; [this] *pūrṇa* emanates from [that] *pūrṇa*; even when *pūrṇa* is drawn out of *pūrṇa*, what remains is also *pūrṇa.*

In the *Kṛṣṇa-Yajurveda Taittirīya-Brāhmaṇa* (*Kāṭhaka* 3.49), we have the word *śūnya* (generally employed to mean zero in mathematics) appearing in the form of a compound word with a negative particle (*nañ*) tagged to it. This is in the context of describing the glory of the sun:

> वेदैरशून्यस्त्रिभिरेति सूर्यः।
> *vedairaśūnyastribhireti sūryaḥ।*

Pāṇini's *Aṣṭādhyāyī* (c. 500 BCE) has the notion of *lopa* which functions as a null-morpheme. *Lopa* appears in several *sūtras*, starting with

> अदर्शनं लोपः। (1.1.60).
> *adarśanaṃ lopaḥ।*
>
> That which gets voided is [termed] *lopaḥ*.

The word *śūnya* also appears twice as a symbol in Piṅgala's *Chandaḥ-śūtra* (c. 300 BCE). In Chapter VIII, while enunciating an algorithm for evaluating any positive integral power of 2 in terms of an optional number of squaring and multiplication (duplication) operations, *śūnya* is used as a marker:

> रूपे शून्यम्। द्विः शून्ये। (8.29-30).
> *rūpe śūnyam। dviḥ śūnye।*
>
> If you get one (*rūpe*) [as the remainder after doing modulo 2 arithmetic] place zero [as the marker]. If you get zero [as the remainder] place two.

Different schools of Indian philosophy have related notions such as the notion of absence (*abhāva*) in Nyāya School, and the *śūnyavāda* among the Bauddhas.

2.1.2. The mathematics of zero

The *Brāhmasphuṭa-siddhānta* (c. 628 CE) of Brahmagupta seems to be the first available text that thoroughly discusses the mathematics of zero. While describing arithmetic, the six operations with zero (*śūnya-parikarma*) are also discussed in Chapter XVIII on algebra (*kuṭṭakādhyāya*). While zero divided by zero is stated to be zero, any other quantity divided by zero is said to be *taccheda* (that with zero denominator). Of the six verses, two are presented below and the rest are paraphrased here [5, pp. 309–310]:

> धनयोर्धनमृणमृणयोः धनर्णयोरन्तरं समैक्यं खम्।
> ऋणमैक्यं च धनमृणधनशून्ययोः शून्यम्॥ ...
> खोद्धृतमृणं धनं वा तच्छेदं खमृणधनविभक्तं वा।
> ऋणधनयोर्वर्गः स्वं खं खस्य पदं कृतिर्यत् तत्॥
> *dhanayordhanamṛṇamṛṇayoḥ dhanarṇayorantaraṃ samaikyaṃ kham।*
> *ṛṇamaikyaṃ ca dhanamṛṇadhanaśūnyayoḥ śūnyam॥ ...*
> *khoddhṛtamṛṇaṃ dhanaṃ vā tacchedaṃ khamṛṇadhanavibhaktaṃ vā।*
> *ṛṇadhanayorvargaḥ svaṃ khaṃ khasya padaṃ kṛtiryat tat॥*
>
> … [The sum of] positive (*dhana*) and negative (*ṛṇa*), if they are equal, is zero (*kham*). The sum of a negative and zero is negative, of a positive and zero is positive and of two zeros, zero (*śūnya*). … Negative subtracted from zero is positive, and positive from zero is negative. Zero subtracted from negative is negative, from positive is positive, and from zero is zero (*ākāśa*).
>
> … The product of zero and a negative, of zero and a positive, or of two zeroes is zero. A zero divided by zero is zero. … A positive or a negative divided by zero is that with zero denominator (*taccheda*). The square (*kṛti*) of a positive or negative number is positive; the square and square-root (*padam*) of zero is zero.

Bhāskarācārya (c. 1150), while discussing the mathematics of zero in his work *Bīja-gaṇita*, explains that infinity (*ananta-rāśi*) which results when some number is divided by zero is called *khahara*. He also graphically describes [4, P. 6] the characteristic property of infinity that it is unaltered even if a huge quantity (*bahu*) is added to or taken away from it with a beautiful simile:[5]

> खहरो भवेत् खेन भक्तश्च राशिः॥ ...
> अस्मिन्विकारः खहरे न राशावपि प्रविष्टेष्वपि निःसृतेषु।
> बहुष्वपि स्याल्लयसृष्टिकालेऽनन्तेऽच्युते भूतगणेषु यद्वत्॥
> *khaharo bhavet khena bhaktaśca rāśiḥ॥ ...*
> *asminvikāraḥ khahare na rāśāvapi praviṣṭeṣvapi niḥsṛteṣu।*
> *bahuṣvapi syāllayasṛṣṭikāle'nante'cyute bhūtagaṇeṣu yadvat॥*
>
> A quantity divided by zero will be (called) *khahara* (an entity with zero as divisor). … In this quantity, *khahara*, there is no alteration even if many are added or taken out, just as there is no alteration in the Infinite (*ananta*), Infallible (*acyuta*) [Brahman] even though many groups of beings enter in or emanate from [It] at times of dissolution and creation.

From the above illustrations it is discernible that Indian mathematicians began dabbling with the notions of zero and infinity in varied mathematical contexts.

2.2. Irrationals and iterative approximations

2.2.1. Approximation for surds in *Śulbasūtras*

Śulbasūtras (c. 800 BCE) that form a part of *Kalpasūtras* (one of the six *Vedāṅgas*) are essentially manuals that contain systematic procedures (algorithms) for the exact construction of altars that were laid out on leveled ground by manipulating cords of various lengths tied to a gnomon. The manuals also contain certain other mathematical details that are relevant to the construction, and are composed in the form of short, cryptic phrases—usually prose, although sometimes including verses—called *sūtras* (literally "string" or "rule, instruction"). The term for the measuring-cords called *śulba* got associated with the name to this set of texts as the *Śulbasūtras* or "Rules of the cord." Starting with simple shapes involving symmetrical figures such as squares and rectangles, triangles, trapezia, rhomboids, and circles, the texts move on to discuss the construction of complex shaped figures such that of falcon. Frequently, one also finds problems pertaining to transformation of one shape into another. Hence, the *Śulbasūtra* rules often involve what we would call area-preserving transformations of plane figures, and thus include the earliest known Indian versions of certain geometric formulas and constants. More interestingly, *Baudhāyana-śulvasūtra* gives the following approximation for $\sqrt{2}$ [33, (1.61–2), P. 19]:

---

5 This simile can be better appreciated by those who are reasonably familiar with the fundamental tenets of Hinduism and its philosophy.

प्रमाणं तृतीयेन वर्धयेत्तच्च चतुर्थेनात्मचतुस्त्रिंशोनेन। सविशेषः।
pramāṇaṃ tṛtīyena vardhayettacca caturthenātmacatustriṃśonena। saviśeṣaḥ।

The measure [of the side] is to be increased by its third and this [third] again by its own fourth less the thirty-fourth part [of the fourth]. That is the approximate diagonal (*saviśeṣa*).

$$\begin{aligned}\sqrt{2} &\approx 1 + \frac{1}{3} + \frac{1}{3\cdot 4} - \frac{1}{3\cdot 4\cdot 34} \\ &= \frac{577}{408} \\ &\approx 1.4142156. \end{aligned} \tag{1}$$

The above approximation is accurate to 5 decimal places. From certain other prescriptions **[33, (1.58), P. 19]** given in this text, one could discern the approximation for π to be given as $\pi \approx 3.0883$.

2.2.2. Approximation for π by Āryabhaṭa

Āryabhaṭa (c. 499) gives the following approximate value for π: [6]

चतुरधिकं शतमष्टगुणं द्वाषष्टिस्तथा सहस्राणाम्।
अयुतद्वयविष्कम्भस्यासन्नो वृत्तपरिणाहः॥
caturadhikaṃ śatamaṣṭaguṇaṃ dvāṣaṣṭistathā sahasrāṇām।
ayutadvayaviṣkambhasyāsanno vṛttapariṇāhaḥ॥

One hundred plus four multiplied by eight and added to sixty-two thousand: This is the approximate measure of the circumference of a circle whose diameter is twenty-thousand.

Thus as per the above verse, $\pi \approx \frac{62832}{20000} = 3.1416$.

It appears that Indian mathematicians (at least in the Āryabhaṭan tradition) employed the method of successive doubling of the sides of a circumscribing polygon—starting from the circumscribing square leading to an octagon, etc.—to find successive approximations to the circumference of a circle. This method has been described in the later Kerala texts *Yuktibhāṣā* (c. 1530) of *Jyeṣṭhadeva* and the *Kriyākramakarī* commentary (c. 1535) of Śaṅkara Vāriyar on the *Līlāvatī*, of Bhāskarācārya.

2.3. Summation of geometric series

The result obtained by summing the geometric series $1 + 2 + 2^2 + \cdots + 2^n$ is stated in Chapter VIII of Piṅgala's *Chandaḥ-sūtra* (c. 300 BCE). It is quite remarkable that Piṅgala also gives a systematic algorithm for evaluating any positive integral power of a number (2 in this context) in terms of an optimal number of squaring and multiplication operations.

---

6 **[2, P. 45]**, *Gaṇitapāda*, verse 10.

Mahāvīrācārya (c. 850), in his *Gaṇita-sāra-saṅgraha* gives the sum of a geometric series and also explains Piṅgala's algorithm for finding the required power of the common ratio between the terms of the series **[16, PP. 28–29]**:

> पदमितगुणहतिगुणितप्रभवः स्याद्गुणधनं तदाद्यूनम्।
> एकोनगुणविभक्तं गुणसङ्कलितं विजानीयात्॥
> *padamitaguṇahatiguṇitaprabhavaḥ syādguṇadhanaṃ tadādyūnam*।
> *ekonaguṇavibhaktaṃ guṇasaṅkalitaṃ vijānīyāt*॥
>
> The first term when multiplied by the product of the common ratio (*guṇa*) taken as many times as the number of terms (*pada*) [in the series], gives rise to the *guṇadhana*. This *guṇadhana*,[7] when diminished by the first term and divided by the common ratio less one, is to be understood as the sum of the geometrical series (*guṇa-saṅkalita*).

If a is the first term and r the common ratio, then what is stated in the verse above may be expressed as

$$a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a(r^n - 1)}{(r - 1)}. \tag{2}$$

Vīrasena (c. 816), in his commentary *Dhavalā* on the *Ṣaṭkhaṇḍāgama*, has made use of the sum of the following infinite geometric series in his evaluation of the volume of the frustum of a right circular cone:[8]

$$1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots + \left(\frac{1}{4}\right)^n + \cdots = \frac{4}{3}. \tag{3}$$

The proof of the above result is outlined by Nīlakaṇṭha Somayājī in his *Āryabhaṭīya-bhāṣya*. Nīlakaṇṭha presents this discussion in the context of deriving an approximation for a small arc in terms of the corresponding chord in a circle. More details are presented in Section 3.1 of the article.

2.4. Āryabhaṭa's computation of Rsine-differences

In the mathematical section of *Āryabhaṭīya* (c. 499), Āryabhaṭa presents two different methods for the computation of tabular Rsine values. While the first is the usual geometric method, the second is an ingenious method which is based on computing the Rsine-differences employing the important property that the second-order differences of Rsines are proportional to the Rsines themselves:[9]

> प्रथमाच्चापज्यार्धाद्यैरूनं खण्डितं द्वितीयार्धम्।
> तत्प्रथमज्यार्धांशैस्तैस्तैरूनानि शेषाणि॥

---

7 This is a technical term employed to refer to ar^n in (2).

8 See, for instance, **[29, PP. 203–205]**.

9 **[2, P. 51]**, *Gaṇitapāda*, verse 12.

prathamāccāpajyārdhādyairūnaṃ khaṇḍitaṃ dvitīyārdham |
tatprathamajyārdhāṃśaistaistairūnāni śeṣāṇi ||

The first Rsine divided by itself and then diminished by the quotient will give the second Rsine-difference. The same first Rsine, diminished by the quotients obtained by dividing each of the preceding Rsines by the first Rsine, gives the remaining Rsine-differences.

Let the quadrant be divided into 24 equal parts, and let J_i denote $R\sin(i\alpha)$ where $\alpha = 225'$ for $i = 1, 2, \ldots, 24$. Now $J_1 = R\sin(225')$, $J_2 = R\sin(450')$, $\ldots$, $J_{24} = R\sin(90°)$, are the 24 Rsines. Let $\Delta_1 = J_1$, $\Delta_2 = J_2 - J_1$, $\ldots$, $\Delta_k = J_k - J_{k-1}$, be the first-order Rsine-differences. Then, the prescription given in the above verse may be expressed as

$$\Delta_2 = J_1 - \frac{J_1}{J_1} \tag{4}$$

$$= \Delta_1 - \frac{J_1}{J_1}. \tag{5}$$

In general,

$$\Delta_{k+1} = \Delta_k - \frac{J_k}{J_1} \quad (k = 1, 2, \ldots, 23). \tag{6}$$

Since Āryabhaṭa also takes $\Delta_1 = J_1 = R\sin(225') \approx 225'$, the above relations reduce to

$$\Delta_2 = 224', \tag{7}$$

$$\Delta_{k+1} - \Delta_k = \frac{-J_k}{225'} \quad (k = 1, 2, \ldots, 23). \tag{8}$$

The renowned mathematician David Mumford refers to the above equation as "the differential equation for the sine function in its finite difference form" [24].

2.5. Instantaneous velocity of a planet (*tātkālika-gati*)

In Indian astronomy, the motion of a planet is computed by making use of two corrections: the *manda-saṃskāra* which essentially corresponds to the equation of center and the *śīghra-saṃskāra* which corresponds to the conversion of the heliocentric longitudes to geocentric longitudes.

In Figure 1, C is the center of a circle on which the mean planet P_0 is located; CU is the direction of the *ucca* (aphelion or apogee as the case may be); P is the true planet which lies on the epicycle of (variable) radius r centered at P_0, such that P_0P is parallel to CU. If M is the mean longitude of a planet, α the longitude of the *ucca*, then the correction (*manda-phala*) $\Delta\mu$ is given by

$$R\sin(\Delta\mu) = \left(\frac{r}{K}\right) R\sin(M - \alpha). \tag{9}$$

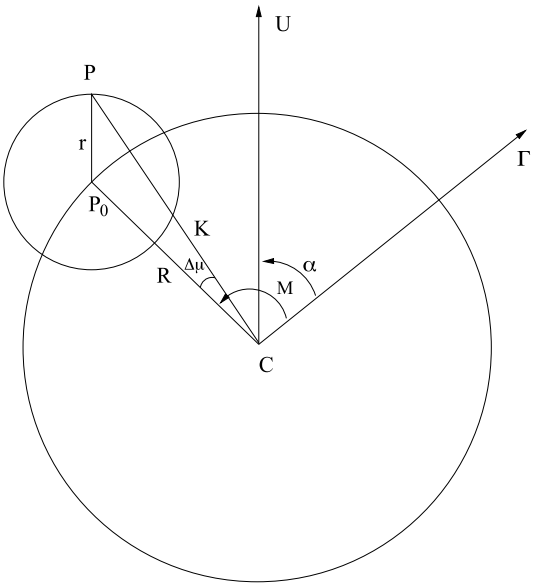


FIGURE 1
Manda correction.

Here K is the *karṇa* (hypotenuse) or the (variable) distance of the planet from the center of the concentric. The texts on Indian astronomy while giving the *manda-phala*, present the following formula:

$$R\sin(\Delta\mu) = \left(\frac{r_0}{R}\right) R\sin(M-\alpha), \tag{10}$$

where r_0 is the tabulated (or mean) radius of the epicycle in the measure of the concentric circle of radius R.

Thus there seems to have been an implicit understanding among the Indian astronomers in accepting this model that the true planet P moves on the variable epicycle of radius r in a way such that the following equation is satisfied:

$$\frac{r}{K} = \frac{r_0}{R}. \tag{11}$$

For small r, the left-hand side of (10) is usually approximated by the arc itself. Thus we have

$$\Delta\mu = \left(\frac{1}{R}\right)\left(\frac{r_0}{R}\right) R\sin(M-\alpha). \tag{12}$$

The *manda*-correction is to be applied to the mean longitude M, to obtain the true or *manda*-corrected longitude μ given by

$$\begin{aligned}\mu &= M - \Delta\mu \\ &= M - \left(\frac{r_0}{R}\right)\left(\frac{1}{R}\right) R\sin(M-\alpha).\end{aligned} \tag{13}$$

If n_m and n_u are the mean daily motions of the planet and the *ucca*, then the true longitude of the planet on the next day may be expressed as

$$\mu + n = (M + n_m) - \left(\frac{r_0}{R}\right)\left(\frac{1}{R}\right) R \sin(M + n_m - \alpha - n_u). \tag{14}$$

Thus the true daily motion (n), obtained by finding the difference of the two equations (13) and (14) is given by

$$n = n_m - \left(\frac{r_0}{R}\right)\left(\frac{1}{R}\right)\left[R \sin\{(M - \alpha) + (n_m - n_u)\} - R \sin(M - \alpha)\right]. \tag{15}$$

The second term in the above is the correction to mean daily motion (*gati-phala*), which strictly involves evaluating the rate of change of the sine function. While an expression for this has been pursued by Bhāskara I (c. 629) in his *Mahābhāskarīya*, the correct formula for the true daily motion of a planet, employing the Rcosine as the "rate of change" of Rsine, seems to have been first given by Muñjāla (c. 932) in his short manual *Laghumānasa* [18, p. 125] and also by Āryabhaṭa II (c. 950) in his *Mahā-siddhānta* [20, p. 58]:

> कोटिफलघ्नी भुक्तिर्गज्याभक्ता कलादिफलम्॥
> *koṭiphalaghnī bhuktirgajyābhaktā kalādiphalam*॥
>
> The *koṭiphala* multiplied by the [mean] daily motion and divided by the radius gives the minutes of the correction [to the rate of the motion].

Essentially, the above verse gives the true daily motion in the form

$$n = n_m - (n_m - n_u)\left(\frac{r_0}{R}\right)\left(\frac{1}{R}\right) R \cos(M - \alpha). \tag{16}$$

Bhāskarācārya (c. 1150) in his *Siddhānta-śiromaṇi* clearly distinguishes the true daily motion from the instantaneous rate of motion [32]. And he gives the Rcosine correction to the mean rate of motion as the instantaneous rate of motion. He further emphasizes the fact that the velocity is changing every instant and this is particularly important in the case of the moon because of its rapid motion [27, pp. 225–227].

3. KERALA SCHOOL OF MATHEMATICS AND ASTRONOMY

The banks of the river Nīlā in the south Malabar region of Kerala witnessed for over 300 years, beginning from about the mid-14th century, what may arguably be considered the golden age of Indian mathematics. The Kerala School of Mathematics and Astronomy pioneered by Mādhava (c. 1340–1420) of Saṅgamagrāma, extended well into the 19th century as exemplified in the work of Śaṅkaravarman (c. 1830), *Rājā* of Kaḍattanāḍu. Only a couple of astronomical works of Mādhava (*Veṇvāroha*, *Lagnaprakaraṇa* and *Sphuṭacandrāpti*) seem to be extant now. Most of his celebrated mathematical discoveries—such as the infinite series for π and the sine and cosine functions—are available only in the form of citations in later works.

Mādhava's disciple Parameśvara (c. 1380–1460) of Vaṭasseri is reputed to have carried out detailed observations for around 55 years. Though a large number of original works and commentaries written by him have been published, one of his important works on mathematics, the commentary *Vivaraṇa* on *Līlāvatī* of Bhāskarācārya, is yet to be published. Nīlakaṇṭha Somayājī (c. 1444–1550) of Kuṇḍagrāma, disciple of Parameśvara's son Dāmodara (c. 1410–1520), is the most celebrated member of Kerala School after Mādhava. Nīlakaṇṭha has cited several important results of Mādhava in his various works, the most prominent of them being *Tantrasaṅgraha* (c. 1500) and *Āryabhaṭīya-bhāṣya*. In the latter work, while commenting on the *Gaṇitapāda* of *Āryabhaṭīya*, Nīlakaṇṭha has also provided ingenious demonstrations or proofs for various mathematical formulae [21].

However, the most detailed exposition of the work of the Kerala School, starting from Mādhava, and including the seminal contributions of Parameśvara, Dāmodara, and Nīlakaṇṭha, is to be found in the famous Malayalam work *Gaṇita-yuktibhāṣā* (henceforth simply *Yuktibhāṣā*) (c. 1530) of Jyeṣṭhadeva (c. 1500–1610), who was a junior contemporary of Nīlakaṇṭha. The direct lineage from Mādhava continued at least till Acyuta Piśāraṭi (c. 1550–1621), a disciple of Jyeṣṭhadeva, who wrote many important independent works in Sanskrit, as well as a couple of commentaries in the local language Malayalam.

In the following sections we shall present an overview of the contribution of the Kerala School to the development of calculus (during the period 1350–1500), following essentially the exposition given in *Yuktibhāṣā*. In order to indicate some of the concepts and methods developed by the Kerala astronomers, we first take up the summation of infinite geometric series as discussed by Nīlakaṇṭha Somayājī in his *Āryabhaṭīya-bhāṣya*, that was alluded to just before. We then consider the derivation of binomial series expansion and the estimation of the sum of integral powers of integers, $1^k + 2^k + \cdots + n^k$ for large n, as presented in *Yuktibhāṣā*. These results constitute the basis for the derivation of the infinite series for $\frac{\pi}{4}$ and its various fast convergents given by Mādhava. Following this, we shall outline another interesting work of Mādhava on the estimation of the end-correction terms called the *antya-saṃskāra*,[10] that had enabled him to arrive at the transformation of the π-series to fast convergent ones—whose multifarious forms may be noted from a citation in Section 4.3.

3.1. Discussion of the sum of an infinite geometric series

In his *Āryabhaṭīya-bhāṣya*, while explaining the *upapatti* (rationale) behind an interesting approximation for the arc of a circle in terms of the *jyā* (Rsine) and the *śara* (Rversine), Nīlakaṇṭha presents a detailed demonstration of how to sum an infinite geometric series. The context of this discussion is Nīlakaṇṭha's pursuit to approximate the arc of a circle in terms of *jyā* (sine) and *śara* (versine). The verse that succinctly presents this approximation is the following:

---

10 Interestingly, this term in common parlance refers to the last rites to be performed.

सत्र्यंशादिषुवर्गात् ज्यावर्गाढ्यात् पदं धनुः प्रायः।
satryaṃśādiṣuvargāt jyāvargāḍhyāt padaṃ dhanuḥ prāyaḥ।

The arc is nearly (*prāyaḥ*) equal to the square root of the sum of the square of the *śara* added to one-third of it, and the square of the *jyā*.

In Figure 2, AB is the arc whose length (assumed to be small) is to be determined in terms of the chord lengths AD and BD. In the Indian mathematical literature, the arc AB, the semichord AD, and the segment BD are referred to as the *cāpa*, *jyārdha*, and *śara*, respectively. As can be easily seen from the figure, this terminology arises from the fact that these geometrical objects look like a bow, string, and arrow, respectively. Denoting them by c, j, and s, the expression for the arc given by Nīlakaṇṭha may be written as

$$c \approx \sqrt{\left(1 + \frac{1}{3}\right)s^2 + j^2}. \tag{17}$$

The proof of the above equation which has been discussed in detail by Sarasvati Amma **[29, PP. 179–182]** involves a summation of an infinite geometric series given by (19).

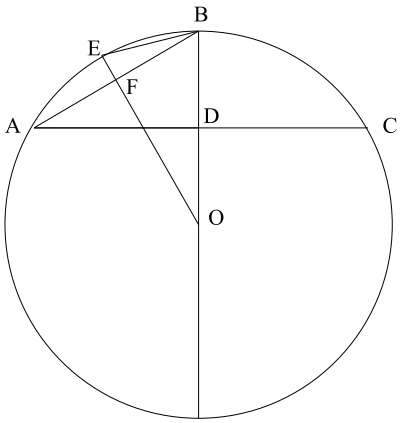


FIGURE 2
Arc-length in terms of *jyā* and *śara*.

The question that Nīlakaṇṭha poses as he commences his detailed discussion on the sum of geometric series is very important and pertinent to the current discussion. In fact, this is a general question that arises quite naturally whenever one encounters the sum of an infinite series **[1, P. 106]**:

कथं पुनः तावदेव वर्धते तावद्वर्धते च ?
kathaṃ punaḥ tāvadeva vardhate tāvadvardhate ca ?

How does one know that [the sum of the series] increases only up to that [limiting value] and that it certainly increases up to that [limiting value]?

Proceeding to answer the above question, Nīlakaṇṭha first states the general result

$$a\left[\left(\frac{1}{r}\right)+\left(\frac{1}{r}\right)^2+\left(\frac{1}{r}\right)^3+\cdots\right]=\frac{a}{r-1}. \tag{18}$$

Here, the left-hand side is an infinite geometric series with the successive terms being obtained by dividing by a common divisor, r, known as *cheda*, whose value is assumed to be greater than 1. He further notes that this result is best demonstrated by considering a particular case, say $a = 1$ and $r = 4$. In his own words **[1, PP. 106–107]**:

> उच्यते — एवं यः तुल्यच्छेदपरभागपरम्परायाः अनन्तायाः अपि संयोगः, तस्य अनन्तानामपि कल्प्यमानस्य योगस्य आद्यावयविनः परम्परांशच्छेदात् एकोनच्छेदांशसाम्यं सर्वत्र समानमेव। तद्यथा — चतुरंशपरम्परायामेव तावत् प्रथमं प्रतिपाद्यते।
>
> *ucyate — evaṃ yaḥ tulyacchedaparabhāgaparamparāyāḥ anantāyāḥ api saṃyogaḥ, tasya anantānāmapi kalpyamānasya yogasya ādyāvayavinaḥ paramparāṃśacchedāt ekonacchedāṃśasāmyaṃ sarvatra samānameva। tadyathā — caturaṃśaparamparāyāmeva tāvat prathamaṃ pratipādyate।*
>
> It is being explained. Thus, in an infinite (*ananta*) geometrical series (*tulyaccheda-parabhāga-paramparā*)[11] the sum of all the infinite number of terms considered will always be equal to the value obtained by dividing by a factor which is one less than the common factor of the series. That this is so will be demonstrated by first considering the series obtained with one-fourth (*caturaṃśa-paramparā*).

What is intended to be demonstrated is

$$\left[\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^3+\cdots\right]=\frac{1}{3}. \tag{19}$$

It is noted that one-fourth and one-third are the only terms appearing in the above equation. Nīlakaṇṭha first defines these numbers in terms of one-twelfth of the multiplier a referred to by the word *rāśi*. For the sake of simplicity, we take the *rāśi* to be unity:

$$3\times\frac{1}{12}=\frac{1}{4};\quad 4\times\frac{1}{12}=\frac{1}{3}. \tag{20}$$

Having defined them, Nīlakaṇṭha first obtains the sequence of results:

$$\frac{1}{3}=\frac{1}{4}+\frac{1}{(4\cdot 3)},$$
$$\frac{1}{(4\cdot 3)}=\frac{1}{(4\cdot 4)}+\frac{1}{(4\cdot 4\cdot 3)},$$
$$\frac{1}{(4\cdot 4\cdot 3)}=\frac{1}{(4\cdot 4\cdot 4)}+\frac{1}{(4\cdot 4\cdot 4\cdot 3)},$$

11 This compound word that has been coined in Sanskrit for the geometric series is very cute and merits attention. It literally means "A series of terms (*paramparā*) in which the successive ones (*parabhāga*) are obtained by the same divisor (*tulyaccheda*) [as the previous]."

and so on, which leads to the general result

$$\frac{1}{3} - \left[\frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots + \left(\frac{1}{4}\right)^n\right] = \left(\frac{1}{4}\right)^n \left(\frac{1}{3}\right). \tag{21}$$

Nīlakaṇṭha then goes on to present the following crucial argument to derive the sum of the infinite geometric series: As we sum more terms, the difference between $\frac{1}{3}$ and sum of powers of $\frac{1}{4}$ (as given by the right-hand side of the above equation) becomes extremely small, but never zero. Only when we take all the terms of the infinite series together, do we obtain the equality

$$\frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots + \left(\frac{1}{4}\right)^n + \cdots = \frac{1}{3}. \tag{22}$$

3.2. Derivation of binomial series expansion

The text *Yuktibhāṣā* presents a very interesting derivation of the binomial series for $(1+x)^{-1}$ by making iterative substitutions in a simple algebraic identity. The method given here may be summarized as follows:

Consider the product $a(\frac{c}{b})$, where some quantity a is multiplied by the multiplier c, and divided by the divisor b. Here, a is called *guṇya*, c the *guṇaka* and b the *hāra*, which are all assumed to be positive integers, with $b > c$. Now the above product can be rewritten as

$$a\left(\frac{c}{b}\right) = a - a\frac{(b-c)}{b}. \tag{23}$$

In the expression $a\frac{(b-c)}{b}$ of the equation above, if we want to replace the division by b (the divisor) by division by c (the multiplier), then we have to make a subtractive correction (called *śodhya-phala*) which amounts to the following equation:

$$a\frac{(b-c)}{b} = a\frac{(b-c)}{c} - \left(a\frac{(b-c)}{c} \times \frac{(b-c)}{b}\right). \tag{24}$$

Now, in the second term (inside parentheses) if we again replace the division by the divisor b by the multiplier c, then we have to make a subtractive-correction once again. Proceeding thus we obtain an alternating series:

$$\begin{aligned} a\frac{c}{b} = a - a\frac{(b-c)}{c} + a\left[\frac{(b-c)}{c}\right]^2 - \cdots + (-1)^{m-1} a\left[\frac{(b-c)}{c}\right]^{m-1} \\ + (-1)^m a\left[\frac{(b-c)}{c}\right]^m + \cdots . \end{aligned} \tag{25}$$

It may be noted that if we set $\frac{(b-c)}{c} = x$, then $\frac{c}{b} = \frac{1}{(1+x)}$. Hence, the series given by (25) is none other than the well-known binomial series

$$\frac{a}{1+x} = a - ax + ax^2 - \cdots + (-1)^m a x^m + \cdots ,$$

which is known to be convergent for $-1 < x < 1$.

Regarding the question of termination of the process, both texts, *Yuktibhāṣā* and *Kriyākramakarī*, clearly mention that logically there is no end to the process of generating *śodhya-phalas*.

It is also noted that the process may be terminated after having obtained the desired accuracy by neglecting the subsequent *phalas* as their magnitudes become smaller and smaller. In fact, *Kriyākramakarī* explicitly mentions that $(b - c)$ should be smaller than c, so that the successive *phalas* become smaller and smaller. In other words, the text, besides presenting a technique to turn a simple algebraic expression into an infinite series, also states the condition that would ensure the convergence of the series.

3.3. Estimation of sums of integral powers of natural numbers

The word employed in the Indian mathematical literature for summation is *saṅkalita*. *Yuktibhāṣā* gives a general method of estimating the sums of integral powers of natural numbers or *samaghāta-saṅkalita*.[12] The detailed procedure given in the text, which is tantamount to providing a proof by induction may be outlined as follows. Before proceeding further with the discussion, a brief note on the notation employed may be useful. We employ S to denote the sum with a subscript and superscript. The subscript denotes the number of terms that are being summed and the superscript denotes the nature of the numbers that are being summed. For the sum of natural numbers, we use (1) as the superscript. For squares of natural numbers, we use (2), and so on. Now, the sum of the first n natural numbers may be written as:

$$\begin{aligned} S_n^{(1)} &= n + (n-1) + \cdots + 1 \\ &= n + [n-1] + [n-2] + \cdots + \big[n - (n-2)\big] + \big[n - (n-1)\big] \\ &= n \cdot n - \big[1 + 2 + \cdots + (n-1)\big]. \end{aligned} \tag{26}$$

When n is very large, the quantity to be subtracted from n^2 is practically (*prāyeṇa*) the same as $S_n^{(1)}$, thus leading to the estimate

$$S_n^{(1)} \approx n^2 - S_n^{(1)}, \quad \text{or} \quad S_n^{(1)} \approx \frac{n^2}{2}. \tag{27}$$

The sum of the squares of the natural numbers up to n may be written as

$$S_n^{(2)} = n^2 + (n-1)^2 + \cdots + 1^2. \tag{28}$$

It can also easily be shown that

$$n S_n^{(1)} - S_n^{(2)} = S_{n-1}^{(1)} + S_{n-2}^{(1)} + S_{n-3}^{(1)} + \cdots . \tag{29}$$

For large n, we have already estimated that $S_n^{(1)} \approx \frac{n^2}{2}$. Thus, for large n, the right-hand side of (29) can be written as

$$n S_n^{(1)} - S_n^{(2)} \approx \frac{(n-1)^2}{2} + \frac{(n-2)^2}{2} + \frac{(n-3)^2}{2} + \cdots . \tag{30}$$

Thus, the excess of $n S_n^{(1)}$ over $S_n^{(2)}$ is essentially $\frac{S_n^{(2)}}{2}$ for large n, so that we obtain

$$n S_n^{(1)} - S_n^{(2)} \approx \frac{S_n^{(2)}}{2}. \tag{31}$$

---

12 The compound *sama-ghāta* in this context means the product of a number with itself.

Again, using the earlier estimate for $S_n^{(1)}$, we obtain the result

$$S_n^{(2)} \approx \frac{n^3}{3}. \tag{32}$$

Proceeding along these lines, *Yuktibhāṣā* presents an argument essentially based on mathematical induction that the summation of the kth powers of natural numbers for a large n may be written as

$$S_n^{(k)} \approx \frac{n^{k+1}}{(k+1)}. \tag{33}$$

3.4. Mādhava's infinite series for π

The infinite series for π attributed to Mādhava is cited by Śaṅkara Vāriyar in his commentaries *Kriyākramakarī* and *Yuktidīpikā*. Mādhava's quoted verse runs as follows [19, P. 379]:

> व्यासे वारिधिनिहते रूपहृते व्याससागराभिहते।
> त्रिशरादिविषमसङ्ख्याभक्तमृणं स्वं पृथक् क्रमात् कुर्यात्॥
> *vyāse vāridhinihate rūpahṛte vyāsasāgarābhihate।*
> *triśarādiviṣamasaṅkhyābhaktamṛṇaṃ svaṃ pṛthak kramāt kuryāt॥*
>
> The diameter multiplied by four and divided by unity [is found and saved]. Again the products of the diameter and four are divided by the odd numbers (*viṣama-saṅkhyā*) three, five, etc., and the results are subtracted and added sequentially [to the earlier result saved].

The words *paridhi* and *vyāsa* in the above verse refer to the circumference (C) and diameter (D), respectively. Hence the content of the verse above, expressed in the form of an equation, becomes

$$C = \frac{4D}{1} - \frac{4D}{3} + \frac{4D}{5} - \frac{4D}{7} + \cdots. \tag{34}$$

Rearranging the terms and using the notation π, we get

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots. \tag{35}$$

We shall now present the derivation of the above result as outlined in *Yuktibhāṣā* of Jyeṣṭhadeva and *Kriyākramakarī* of Śaṅkara Vāriyar. For this purpose, let us consider the quadrant OP_0P_nA of the square circumscribing the given circle (see Figure 3). Let r be the radius of the circle. Divide the side $P_0P_n(= r)$ into n equal parts (n large). Then P_0P_i $(i = 1, 2, \ldots, n)$ are the *bhujās* (sides) and $k_i = OP_i$ are the *karṇas* (hypotenuses) of the triangle to be conceived of. The points of intersection of these *karṇas* and the circle are marked as A_is.

It is straightforward to see that the *bhujās* P_0P_i, the *karṇas* k_i, and the East–West line OP_0 form right-angled triangles. Hence we have the relation

$$k_i^2 = r^2 + \left(\frac{ir}{n}\right)^2. \tag{36}$$

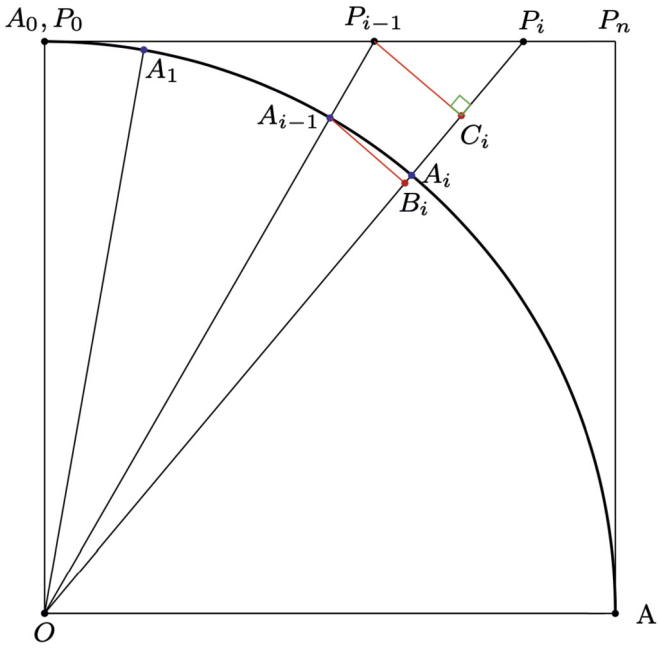


FIGURE 3
Geometrical construction used in the proof of the infinite series for π.

Considering two successive *karṇas*, and the pairs of similar triangles $OP_{i-1}C_i$ and $OA_{i-1}B_i$, and $P_{i-1}C_iP_i$ and OP_0P_i, it can be shown that the length of the segment $A_{i-1}B_i$ is given by

$$A_{i-1}B_i = \left(\frac{r}{n}\right)\left(\frac{r^2}{k_{i-1}k_i}\right). \tag{37}$$

Now the text presents the crucial argument that, when n is large, the Rsines $A_{i-1}B_i$ can be taken as the arc-bits $A_{i-1}A_i$ themselves.

> परिधिखण्डस्य अर्धज्या परिध्यंश एव।
> *paridhikhaṇḍasya ardhajyā paridhyaṃśa eva*।
>
> The Rsines (*ardhajyā*) corresponding to the arc-bits (*paridhikhaṇḍa*) are essentially the arc-bits themselves.

Recalling that A_0 will merge with P_0, we can easily see that

$$\sum_{i=1}^{n} A_{i-1}A_i = \frac{C}{8}. \tag{38}$$

Thus, one-eighth of the circumference of the circle can be written as the sum of the contributions made by the individual segment $A_{i-1}B_i$ given by (37). That is,

$$\frac{C}{8} \approx \left(\frac{r}{n}\right)\left[\left(\frac{r^2}{k_0k_1}\right) + \left(\frac{r^2}{k_1k_2}\right) + \left(\frac{r^2}{k_2k_3}\right) + \cdots + \left(\frac{r^2}{k_{n-1}k_n}\right)\right]. \tag{39}$$

It is further argued that the denominators may be replaced by the square of either of the *karṇas* since the difference is negligible. Hence we obtain:

$$\begin{aligned}\frac{C}{8} &= \sum_{i=1}^{n} \frac{r}{n}\left(\frac{r^2}{k_i^2}\right) \\ &= \sum_{i=1}^{n}\left(\frac{r}{n}\right)\left(\frac{r^2}{r^2+(\frac{ir}{n})^2}\right) \\ &= \sum_{i=1}^{n}\left[\frac{r}{n} - \frac{r}{n}\left(\frac{(\frac{ir}{n})^2}{r^2}\right) + \frac{r}{n}\left(\frac{(\frac{ir}{n})^2}{r^2}\right)^2 - \cdots\right]. \end{aligned} \tag{40}$$

In the series expression for the circumference given above, factoring out powers of $\frac{r}{n}$, the sums involved are the even powers of the natural numbers. Now, recalling the estimates that were obtained earlier (33) for these sums when n is large, we arrive at the result (35), which was rediscovered by Gregory and Leibniz almost three centuries later.

3.5. Derivation of end-correction terms (*antya-saṃskāra*)

It is well known that the series given by (35) for $\frac{\pi}{4}$ is an extremely slowly converging series. Mādhava seems to have found an ingenious way to circumvent this problem with a technique known as *antya-saṃskāra*. The nomenclature stems from the fact that a correction (*saṃskāra*) is applied towards the end (*anta*) of the series after we terminate it, by considering only a certain number of terms from the beginning. We can, of course, terminate the series at any term we desire, provided we find a correction $\frac{1}{a_p}$ to be applied, that happens to be a good approximation for the rest of the truncated terms in the series. This seems to have been the thought process that has gone in in discovering this *antya-saṃskāra* technique.

Suppose we terminate the series after the term $\frac{1}{p}$ and consider applying the correction term $\frac{1}{a_p}$, then

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + (-1)^{\frac{p-3}{2}}\frac{1}{p-2} + (-1)^{\frac{p-1}{2}}\frac{1}{p} + (-1)^{\frac{p+1}{2}}\frac{1}{a_p}. \tag{41}$$

Three successive approximations to the correction divisor a_p given by Mādhava may be expressed as:

$$\begin{aligned} a_p(1) &= 2(p+2), \\ a_p(2) &= (2p+2) + \frac{4}{(2p+2)}, \\ a_p(3) &= (2p+2) + \frac{4}{2p+2+\frac{16}{2p+2}}. \end{aligned} \tag{42}$$

Yuktibhāṣā contains a detailed discussion on how these correction terms of successive orders are arrived at. While the discussion in the text goes only up to the three terms as above, presumably because the expressions become increasingly cumbersome, the idea that

the partial quotients of the continued fraction

$$(2p+2)+\cfrac{2^2}{(2p+2)+\cfrac{4^2}{(2p+2)+\cfrac{6^2}{(2p+2)+\cdots}}}$$

serve as correction factors to higher and higher orders is seen to be inherently present in the reasoning. A graph depicting the variation of error in the estimate of π using the three successive end-corrections by truncating the series at different values of p is shown in Figure 4. It may be noted that, when we use the third-order end-correction, by just considering about 25 terms in the series, we are able to obtain the π value correct to 10 decimal places.

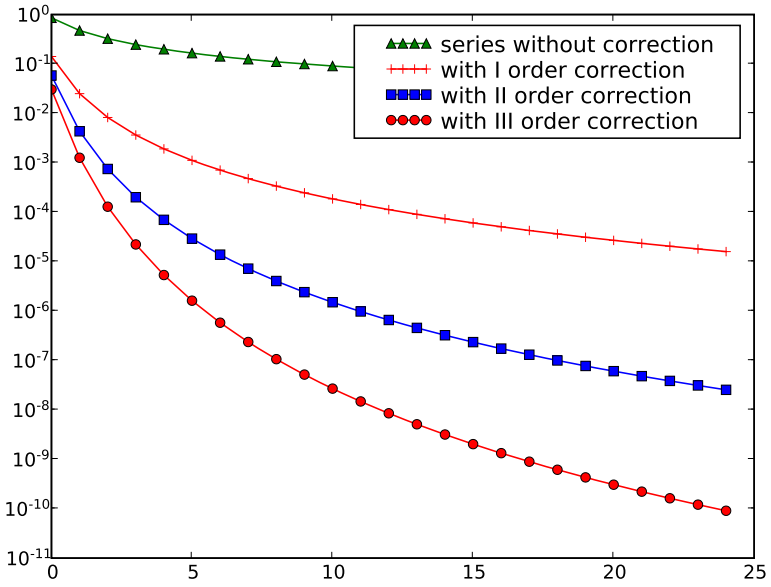


FIGURE 4
Graph depicting the accuracy that is obtained in estimating the value of π by truncating the series at different values of p and employing the three corrections given by (42).

The following accurate value of π (correct to 11 decimal places), given by Mādhava, has been cited by Nīlakaṇṭha in his *Āryabhaṭīya-bhāṣya* and by Śaṅkara Vāriyar in his *Kriyākramakarī*:[13]

13 **[1, P. 42]**, comm. on *Gaṇitapāda* verse 10; **[19, P. 377]**.

विबुधनेत्रगजाहिहुताशनत्रिगुणवेदभवारणबाहवः।
नवनिखर्वमिते वृतिविस्तरे परिधिमानमिदं जगदुर्बुधाः॥
vibudhanetragajāhihutāśanatriguṇavedabhavāraṇabāhavaḥ।
navanikharvamite vṛtivistare paridhimānamidaṃ jagadurbudhāḥ॥

The π value given above is

$$\pi \approx \frac{2827433388233}{9 \times 10^{11}} = 3.141592653592\ldots \tag{43}$$

The 13-digit number appearing in the numerator has been specified using object-numeral (*bhūta-saṅkhyā*) system, whereas the denominator is specified by word numerals.[14]

4. HISTORIOGRAPHY OF THE INCEPTION OF CALCULUS IN INDIA

4.1. Brief note on Charles Whish and his collections

Charles Matthew Whish (1794–1833), as noted earlier, was instrumental in first bringing to the notice of modern mathematical scholarship the achievements of the Kerala School through his historic paper that got posthumously published in TRAS (1934) [36]. The fact that Whish had discovered them more than a decade before the paper got published is evident from the correspondence between John Warren and George Hyne that has been noted down by the former in his *Kālasaṅkalita* [35]. It may also be mentioned here that the collection of manuscripts that Whish had made—which the author of this paper had an occasion to look at—amply demonstrates the fact that he was interested not only in astronomy and mathematics, but also in a wide variety of topics that includes vedic literature, *itihāsas* and *purāṇas*. Fortunately, these manuscripts were deposited in the Royal Asiatic Society of Great Britain and Ireland in July 1836 by his brother, and are still well preserved in the Royal Asiatic Society, London.

The personal notes (see Figure 5) found in various manuscripts in Whish's collection also reveal that during his stay in South Malabar, he had got in touch with several scholars, and read some of the Sanskrit and Malayalam texts with them. Given his abiding interest to acquire scholarship in a variety of fields by familiarizing with the culture, language, and knowledge systems of India—and also share it back with his counterparts in Europe—it is highly unfortunate that Charles Whish suffered a premature death in 1833 at the age of 38 years.[15]

---

14 In the *bhūta-saṅkhyā* system, *vibudha* = 33, *netra* = 2, *gaja* = 8, *ahi* = 8, *hutāśana* = 3, *tri* = 3, *guṇa* = 3, *veda* = 4, *bha* = 27, *vāraṇa* = 8, *bāhu* = 2. In word numerals, *nikharva* represents 10^{11}. Hence, *nava-nikharva* = 9×10^{11}.

15 The list of European tombs in the district of Cuddapah prepared by C. H. Mounsey in 1893 mentions: "Sacred to the memory of C. M. Whish, Esquire of the Civil Service, who departed this life on the 14th April 1833, aged 38 years."

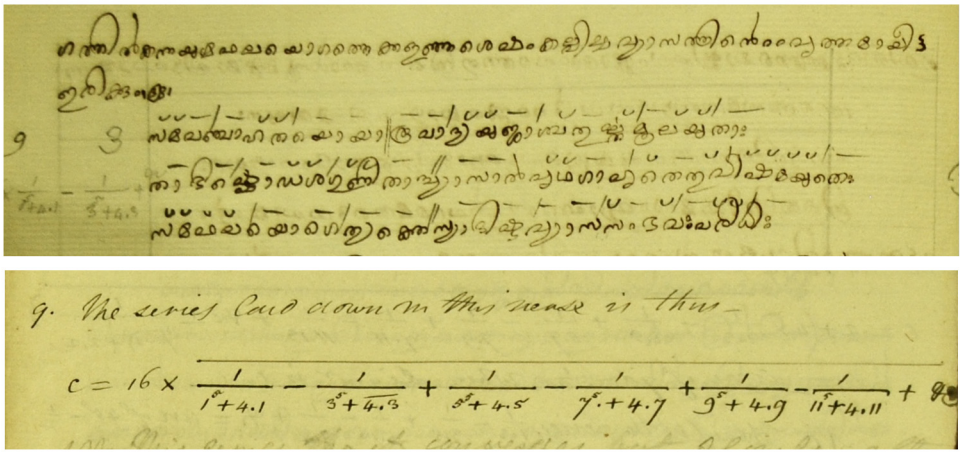


FIGURE 5
Excerpts from Whish's manuscript showing the verses in Malayalam along with his metrical markings and a portion from his mathematical notes in English (Courtesy: RAS, London).

4.2. About *Kālasaṅkalita*

Kālasaṅkalita, published in 1825 by John Warren who was the director of the Madras observatory for sometime, is a compendium of the different methods employed by the *pañcāṅga*-makers for reckoning time. The main purpose of preparing this text was to facilitate a comparison of the European and Indian chronologies, as is mentioned in the preface: "… their chief object being merely to explain the various modes according to which the Natives of India divide time, in these southern provinces, and to render their Kalendars intelligible. These may therefore be properly considered rather as instruments contrived for Chronological purposes, than as Astronomical Tracts."

It turns out that the text is useful in several other respects as well, especially from a historical perspective. Among other things, the one which is of particular interest to us in this paper is the exchange of ideas that took place among the three civil servants of the East India Company, namely, Warren, Whish, and Hyne, particularly with regard to the invention of the infinite series expansion by the "Natives."

4.3. Extracts from the exchanges between Whish, Hyne, and Warren

In the Second Memoir of *Kālasaṅkalita* on the Hindu Lunisolar year, before commencing his discussion on *śaṇku*[16] and the diurnal problems associated with it, John Warren notes:

16 The term *śaṇku* refers to a very simple contrivance, yet a powerful tool that has been extensively employed by Indian astronomers – right from the period of the *Śulvasūtras* (c. 800 BCE) – to carry out a variety of experiments related to shadow measurements.

Before entering into the resolution of the Problems which depend on the length of the Meridian shadow, it is proper to enquire …
Of their manner of resolving geometrically the ratio of the diameter to the circumference of a circle, I never saw any Indian demonstration: the common opinion, however, is that they approximate it in the manner of the ancients, by exhaustion; that is, by means of inscribed and circumscribed Polygons. However, a Native Astronomer who was a perfect stranger to European Geometry, gave me the well-known series $1 - \frac{1}{3} + \frac{1}{5} + \cdots$. *This person reduced the five first terms of the series before me, which he called Bagah Anoobanda, or Bagah Apovacha; to shew that he understood its use. This proves at least that the Hindus are not ignorant of the doctrine of series …*

This passage clearly indicates that John Warren is confronting a dilemma: on the one hand, he has met "a Native Astronomer who was a perfect stranger to European Geometry" giving the well-known series $1 - \frac{1}{3} + \frac{1}{5} + \cdots$ and, on the other hand, "he never saw any Indian demonstration" of the series. To the above passage, Warren appends a note where he mentions:

I owe the following note to Mr. Hyne's favor: "The Hindus never invented the series; it was communicated with many others, by Europeans, to some learned Natives in modern times. Mr. Whish sent a list of the various methods of demonstrating the ratio of the diameter and circumference of a Circle employed by the Hindus to the literary society, being impressed with the notion that they were the inventors. I requested him to make further inquiries, and his reply was that he had reasons to believe them entirely modern and derived from Europeans, observing that not one of those who used the Rules could demonstrate them. Indeed, the pretensions of the Hindus to such a knowledge of geometry, is too ridiculous to deserve refutation." I join in substance Mr. Hyne's opinion, but do not admit that the circumstance that none of the Sastras mentioned by Mr. Whish, who used the series could demonstrate them, would alone be conclusive.

John Warren returns to this issue in "Fragments II" attached at the end of his treatise *Kālasaṅkalita*, entitled *"On certain infinite Series collected in different parts of India, by various Gentlemen, from Native Astronomers."— Communicated by George Hyne, Esq. of the H. C.'s Medical Service*, which we reproduce below:

"My Dear Sir,
I have great pleasure in communicating the Series, to which I alluded …

$$C = 4D\left(1 - \frac{1}{3} + \frac{1}{5} - \cdots\right), \tag{44}$$

$$C = \sqrt{12D^2} - \frac{\sqrt{12D^2}}{3 \cdot 3} + \frac{\sqrt{12D^2}}{3^2 \cdot 5} - \frac{\sqrt{12D^2}}{3^3 \cdot 7} + \cdots, \tag{45}$$

$$C = 2D + \frac{4D}{(2^2-1)} - \frac{4D}{(4^2-1)} + \frac{4D}{(6^2-1)} - \cdots, \tag{46}$$

$$C = 8D\left[\frac{1}{(2^2-1)} + \frac{1}{(6^2-1)} + \frac{1}{(10^2-1)} + \cdots\right], \tag{47}$$

$$C = 8D\left[\frac{1}{2} - \frac{1}{(4^2-1)} - \frac{1}{(8^2-1)} - \frac{1}{(12^2-1)} - \cdots\right], \tag{48}$$

$$C = 3D + \frac{4D}{(3^3-3)} - \frac{4D}{(5^3-5)} + \frac{4D}{(7^3-7)} - \cdots, \tag{49}$$

$$C = 16D\left(\frac{1}{1^5+4.1} - \frac{1}{3^5+4.3} + \frac{1}{5^5+4.5} - \cdots\right). \tag{50}$$

I am, my dear Sir, most sincerely, your's,

MADRAS, *17th August 1825.* *G. HYNE."*

Based on the nature of exchanges recorded by Warren in 1825, it is quite clear that:

1. Whish was convinced that the infinite series were discovered by the "Natives."
2. Hyne was convinced that the infinite series were NOT discovered by the "Natives" but was only borrowed, and that the Hindus were merely pretending as originators of the series.
3. Warren decides to go with the opinion of Hyne, though initially he felt that the latter's argument is not "conclusive."

Under such circumstances, with a lot of communication back and forth, one could only imagine how challenging it would have been for Whish[17] to swim against the current, and place on record his own understanding regarding the knowledge of the infinite series, or of their demonstration in the Indian astronomical tradition. The mere fact the paper authored by him in 1820s got accepted for publication in the 1830s posthumously, stands testimony to his courage, perseverance, assiduity, and tenacity with which he would wear down his opponents.

One of the remarkable statements in the paper of Whish that is of particular interest to us in the present context is: "A further account of the Yucti-Bhāshā, the demonstrations of the rules for the quadrature of the circle by infinite series, with the series for the sines, cosines, and their demonstrations, will be given in a separate paper." Unfortunately, Whish did not survive to publish this paper with demonstrations from *Yuktibhāṣā*, which could have silenced all those who doubted whether these series listed by them were discovered by Indians.

---

17 It may also be recalled that Whish was hardly 30-year old in 1825, whereas George Hyne and Warren were seniors. Warren was the director of Madras Observatory around 1805 and Hyne was a senior member of Madras Literary Society who was appointed as the first Secretary of the Committee of Public Instruction by the Madras Government.

More striking and intriguing development connected with Whish's paper than what is narrated above, is the kind of consensus that seems to have emerged among the European indologists and historians of mathematics and astronomy to undermine and suppress it for almost a hundred years since its publication in 1834. Either the work itself was not referenced in their writings, or even if it were, some of the well-established mathematicians, such as Augustus De Morgan and scholar administrators such as Charles P. Brown dismissed it—far more strongly than was done by Hyne—by castigating it as "hoax" and "forgery" [11], [6, PP. 48–49].

Not providing reference to this paper of Whish on the contributions of the Kerala School, or discussing its contents, is certainly not out of ignorance, which is perfectly understandable. But strangely it seems to be a volitional act! See, for instance, the scholarly monograph of Geroge Thibaut (in German) on Indian Astronomy, Astrology, and Mathematics [34, P. 2] which makes note of 1827 article of Whish, on the Greek origin of the Hindu Zodiac. However, it mysteriously fails to mention this 1934 paper of Whish, though the paper is germane to the subject of his discussion. We present below a clip (Figure 6) of the relevant section from Thibaut's volume, along with a concise translation (done with the help Google).

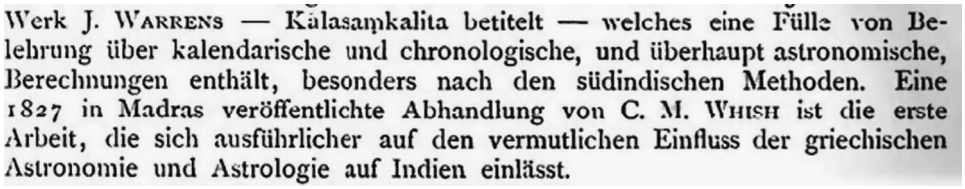

Werk J. Warrens — Kālasaṃkalita betitelt — welches eine Fülle von Belehrung über kalendarische und chronologische, und überhaupt astronomische, Berechnungen enthält, besonders nach den südindischen Methoden. Eine 1827 in Madras veröffentlichte Abhandlung von C. M. Whish ist die erste Arbeit, die sich ausführlicher auf den vermutlichen Einfluss der griechischen Astronomie und Astrologie auf Indien einlässt.

FIGURE 6
A clip of the relevant section from Thibaut's volume

> *J. WARREN'S work entitled Kālasaṅkalita, which contains a wealth of instruction on calendar and chronological, and generally astronomical, calculations, especially according to the South Indian methods. A treatise by C. M. WHISH, published in Madras in 1827, is the first to delve into the probable influence of Greek astronomy and astrology on India.*

Similarly, the popular translation of *Sūryasiddhānta* by Ebenezer Burgess [7, P. 174], and the review article by John Burgess of the European studies of Indian astronomy in the 18th and 19th centuries [8, PP. 746–750] do not refer to the 1934 paper of Whish while they take note of his other contributions.

Furthermore, David Eugene Smith (1860–1944), in his seminal two-volume history of mathematics completed in 1925, simply refers to the article of Whish, but does not touch upon its content except for noting that it deals with Indian values for π. Thus we find an interesting period of almost a century in European historiography where either both the title and the content, or at least the content of Whish's article remained an untouchable!

Fortunately, the references given by David Smith [31, P. 309] caught the attention of the renowned historian of Indian mathematics, Bibhutibhusan Datta, who drew attention to

the various infinite series mentioned in Whish's article in an article published in 1926 [10]. This was followed a decade later by Datta's colleague, Avadesh Narayan Singh, who referred to the various manuscripts of the Kerala texts which discuss these infinite series [30]. And the next decade finally saw the publication of a series of articles by C. T. Rajagopal and his collaborators and the edition of the mathematics part of *Yuktibhāṣā* by Ramavarma Thampuran and Akhileswara Ayyar (for details, please, see [22, 23, 25, 26, 28]).

5. CONCLUDING REMARKS

It is quite evident from the above mathematical and historical discussions that the mathematicians of the Kerala School, around the 14th century, had clearly mastered the technique of handling the infinitesimal, the infinite and the notion of limit—the three pillars on which the edifice of calculus rests upon. The context and purpose for which the Kerala mathematicians developed these techniques are different from those in which they got developed in Europe a couple of centuries later. It must also be mentioned here that the Kerala mathematicians had restricted their discussions to the quadrature of a circle and certain trigonometric functions.[18] However, their mathematical formulation of the problem involving the "infinitesimally" small and summing up the "infinite" number of the resulting infinitesimal contributions, along with a clear understanding of the mathematical subtleties involved in it, are not in any way fundamentally different from the way it would be formulated or understood today.

While there were a number of European mathematicians and indologists who expressed their appreciations for the contributions made by Indians, the historiography captured in Section 4, in no uncertain terms reveals that there were many others who promulgated their views and tried to suppress the discovery of Kerala mathematicians, by brazenly discounting their work.[19] The cascading effect of it has resulted in some well-known authors producing books even in 1930s—almost a century after the publication of the Whish's historic paper—containing descriptions such as "… the Hindus may have inherited some of the bare facts of Greek science, but not the Greek critical acumen. Fools rush in where angels fear to tread [9][20] …" that are quite misleading, derailing, and damaging. It is perhaps a fitting tribute to Whish that today at least most historians of mathematics are aware of this "neglected chapter" in the history of mathematics. For this reason, the following statement by David Mumford is quite relevant [24]:

> *It is high time that the full story of Indian mathematics from Vedic times through 1600 became generally known. I am not minimizing the genius of the Greeks and*

---

18 The mathematicians of Europe, however, took a different approach to the subject, by considering an arbitrary curve for analysis, and by providing formal definitions and generalized treatment to the topic.

19 The episode essentially reminds us of the important lesson: if we look through a malicious goggle, then even the genuine narratives may sound to be an elaborate hoax!

20 Quoted by A. A. K. Ayyangar in his article [3].

their wonderful invention of pure mathematics, but other peoples have been doing math in different ways, and they have often attained the same goals independently. Rigorous mathematics in the Greek style should not be seen as the only way to gain mathematical knowledge.

… the muse of mathematics can be wooed in many different ways and her secrets teased out of her. And so they were in India …

Apart from the topics discussed in the present article, several other ideas of calculus seem to have been employed by Indian astronomers in their studies related to planetary motion. For instance, one of the verses in the second chapter of *Tantrasaṅgraha* deals with the derivative of the inverse sine function.[21] We would also like to refer the reader to the literature for the very interesting proof of the sine and cosine series given in the *Yuktibhāṣā*. As has been remarked recently by Divakaran **[15, P. 335]** that, unlike the derivation that was given by Newton, which involved "guessing" successive terms "from their form," the *Yuktibhāṣā* approach of "integrating the difference/differential equation for sine and cosine is entirely different and very modern", which has also been briefly touched upon by Mumford in his article cited above.

For most of us who have got trained completely in the modern scheme of education, it may be hard to imagine doing mathematics without the "luxury" of expressing things "neatly" in symbolic forms. It is equally hard to think of expressing power series for trigonometric functions, derivatives of functions, and the like, purely in metrical forms. But that is how knowledge seems to have been preserved and handed down from generation to generation in India for millennia starting from Vedic age till the recent past. It only proves the point: equations may be handy but not essential; notations may be useful, but not indispensable. Formal definitions and structures are certainly valuable and helpful, but the absence of them does not inhibit or stagnate the birth and development of mathematical ideas. After all, mathematics is mathematics irrespective of how, where, and why it is practiced!

ACKNOWLEDGMENTS

The author would like to profusely thank Prof. M. D. Srinivas for useful discussions on the topic. He would like to place on record his sincere gratitude to the Ministry of Education, Government of India, for the generous support extended to carry out research activities on Indian science and technology by way of initiating the Science and Heritage Initiative (SandHI) at IIT Bombay. His thanks are also due to his doctoral student D. G. Sooryanarayan for readily and enthusiastically helping in the preparation of this article.

---

21 In fact, Nīlakaṇṭha ascribes this verse—dealing with the instantaneous velocity (*tātkālikagati*) of the moon—to his teacher Dāmodara in his *Jyotirmīmāṃsā*. While the details of how Dāmodara, or someone else before him, arrived at this result is not evident to us, one thing is quite clear—the astronomers were adept at dealing with the derivatives of basic trigonometric functions.

REFERENCES

[1] *Āryabhaṭīya* of Āryabhaṭa, edited with *Āryabhaṭīyabhāṣya* of Nīlakaṇṭha Somayājī by K. Sambasiva Sastri, Trivandrum Sanskrit Series 101, Trivandrum, 1930.

[2] *Āryabhaṭīya* of Āryabhaṭa, edited by K. S. Shukla and K. V. Sarma, Indian National Science Academy, New Delhi, 1976.

[3] A. A. K. Ayyangar, Peeps into India's mathematical past. *The Half-Yearly J. Mysore University*, **V** (1945), 101.

[4] *Bījagaṇita* of Bhāskarācārya, edited by Muralidhara Jha, Benaras Sanskrit Series 159, Benaras, 1927, *Vāsanā* on *Khaṣaḍvidham* 3.

[5] *Brāhmasphuṭasiddhānta* of Brahmagupta, edited with his own commentary by Sudhakara Dvivedi, Medical Hall Press, Benaras, 1902, verses 18.30–35.

[6] C. P. Brown, *Carnatic Chronology The Hindu and Mahomedan Methods of Reckoning Time Explained*. London, 1863.

[7] E. Burgess, *Translation of the Sūryasiddānta*. The American Oriental Society, New Haven, 1860.

[8] J. Burgess, Notes on Hindu Astronomy and the History of our knowledge of it, *J. R. Asiat. Soc.* **25** (1893), 717–761.

[9] T. Dantzig, *Number, the language of science; a critical survey written for the cultured non-mathematician*. Macmillan, New York, 1930.

[10] B. Datta, Hindu values of π. *J. Asiat. Soc. Bengal* **22** (1926), 25–42.

[11] A. De Morgan, Article on "Viga Ganita" (sic). In *The Penny Cyclopaedia*, Vol. XXVI, pp. 318–326, Charles Knight & Co, London, 1843.

[12] Dharampal, *Indian science and echnology in the eighteenth century*. Dharampal Classics Series, Rashtrotthana Sahitya, Bengaluru, in collaboration with Centre for Policy Studies, Chennai, 2021.

[13] P. P. Divakaran, The first textbook of calculus: *Yuktibhāṣā*. *J. Indian Philos.*, **35** (2007), 417–433.

[14] P. P. Divakaran, Calculus in India: The historical and mathematical context. *Current Sci.* **99** (2010), no. 3, 8–14.

[15] P. P. Divakaran, *The mathematics of India: concepts, methods, connections*. Springer, 2018.

[16] *Gaṇitasārasaṅgraha* of Mahāvīrācārya, edited by Lakshmi Chand Jain, Jaina Saṃskṛti Saṃrakshaka Saṃgha, Sholapur, 1963, verse 2.93.

[17] A. A. Krishnaswami Ayyangar, Ramanujan the mathematician. In *Book of Commemoration of Gopalakrishnamacharya*, Madras, 1942.

[18] *Laghumānasa* of Muñjāla, edited by K. S. Shukla, Indian National Science Academy, New Delhi, 1990, verse 3.4.

[19] *Līlāvatī of Bhāskara II, Ed. with commentary Kriyākramakarī of Śaṅkara Vāriyar* by K. V. Sarma, Vishveshvaranand Vedic Research Institute, Hoshiarpur, 1975.

[20] *Mahāsiddhānta* of Āryabhaṭa II, edited by Sudhakara Dvivedi, Benaras Sanskrit Series 148, Varanasi 1910, verse 3.15.

[21] K. Mahesh, D. G. Sooryanarayan, and K. Ramasubramanian, Elegant dissection proofs for algebraic identities in Nīlakaṇṭha's *Āryabhaṭīyabhāṣya*. *Indian J. Hist. Sci.* **56** (2021), no. 2, 1–17.

[22] K. Mukunda Marar, Proof of Gregory's series. *Teach. Mag.* **15** (1940), 28–34.

[23] K. Mukunda Marar and C. T. Rajagopal, On the Hindu quadrature of the circle. *J. Bombay Branch R. Asiat. Soc.* **20** (1944), 65–82.

[24] D. Mumford, Mathematics in India (a review article). *Notices Amer. Math. Soc.* **57** (2010), no. 3, 385–390.

[25] C. T. Rajagopal, A neglected chapter of Hindu mathematics. *Scripta Math.* **15** (1949), 201–209.

[26] C. T. Rajagopal and A. Venkataraman, The sine and cosine power series in Hindu mathematics. *J. R. Asiat. Soc. Bengal, Sci.* **15** (1949), 1–13.

[27] K. Ramasubramanian and M. D. Srinivas, Development of calculus in India. In *Studies in history of Indian mathematics*, edited by C. S. Seshadri, pp. 201–286, Hindustan Book Agency, New Delhi, 2010.

[28] Ramavarma (Maru) Tampuran and A. R. Akhileswara Ayyar, *Yuktibhāṣā Part I*. Mangalodayam Ltd. Trichur, 1948.

[29] T. A. Sarasvati Amma, *Geometry in Ancient and Medieval India*. Motilal Banarsidass, Delhi, 1979, Rep. 2007.

[30] A. N. Singh, On the Use of Series in Hindu Mathematics. *Osiris* **1** (1936), 606–628.

[31] D. E. Smith, *History of Mathematics*, Vol. II. Dover, New York, 1958 (rep. of 1925 edition).

[32] M. S. Sriram, *Grahagaṇitādhyāya* of Bhāskarācārya's *Siddhāntaśiromaṇi*. In *Bhāskara-prabhā*, edited by K. Ramasubramanian, Takao Hayashi, Clemency Montelle, pp. 197–231, Springer, Singapore, 2019.

[33] *The Śulbasūtras of Baudhāyana, Āpastamba, Kātyāyana and Mānava (With Text, English Translation and Commentary)*, edited by S. N. Sen and A. K. Bag, Indian National Science Academy, New Delhi, 1983.

[34] G. Thibaut, *Astronomie, Astrologie und Mathematik*. De Gruyter, Strassbourg, 1899.

[35] J. Warren, *Kālasaṅkalita*. College Press, Madras, 1825.

[36] C. M. Whish, On the Hindu quadrature of the circle, and the infinite series of the proportion of the circumference to the diameter exhibited in the four Shastras, the Tantrasangraham, Yucti Bhasa, Carana Paddhati and Sadratnamala. *Trans. R. Asiat. Soc. G.B.* **3** (1834), 509–523.

K. RAMASUBRAMANIAN

Indian Institute of Technology Bombay, Mumbai, India, ram@hss.iitb.ac.in

LIST OF CONTRIBUTORS

https://ems.press

ISBN Set 978-3-98547-058-7

ISBN Volume 7 978-3-98547-065-5