Mathematisches Forschungsinstitut Oberwolfach

Report No. 12/2019

DOI: 10.4171/OWR/2019/12

Uncertainty Quantification

Organised by
Oliver Ernst, Chemnitz
Fabio Nobile, Lausanne
Claudia Schillings, Mannheim
Tim Sullivan, Berlin

10 March – 15 March 2019

Abstract. Uncertainty quantification (UQ) is concerned with including and characterising uncertainties in mathematical models. Major steps comprise proper description of system uncertainties, analysis and efficient quantification of uncertainties in predictions and design problems, and statistical inference on uncertain parameters starting from available measurements. Research in UQ addresses fundamental mathematical and statistical challenges, but has also wide applicability in areas such as engineering, environmental, physical and biological applications. This workshop focussed on mathematical challenges at the interface of applied mathematics, probability and statistics, numerical analysis, scientific computing and application domains. The workshop served to bring together experts from those disciplines in order to enhance their interaction, to exchange ideas and to develop new, powerful methods for UQ.

Mathematics Subject Classification (2010): 60xx, 62xx, 65xx, 28xx, 35xx, 37xx, 91xx, 49xx, 46xx, 47xx.

Introduction by the Organizers

Uncertainty quantification (UQ) is a research area at the interface of applied mathematics, probability and statistics, numerical analysis, scientific computing, and typical application domains such as geosciences and engineering. UQ has assumed increasing prominence over the last 10–15 years and concerns itself with the proper incorporation of all forms of error and uncertainty in scientific inferences about such applications. Major steps comprise proper description of system uncertainties, analysis and efficient quantification of uncertainties in predictions and design

problems, and statistical inference on uncertain parameters starting from available measurements. Research in this area addresses fundamental mathematical and statistical challenges, but has also wide applicability in areas such as engineering, environmental, physical and biological applications.

The workshop *Uncertainty Quantification*, organised by Oliver Ernst (Chemnitz), Fabio Nobile (Lausanne), Claudia Schillings (Mannheim), and Tim Sullivan (Berlin), was the first full workshop on this theme to be hosted by the MFO. In comparison to major international gatherings such as the SIAM Conferences on Uncertainty Quantification (2012, 2014, 2016, 2018), this workshop featured a focus upon European-based researchers in the field (with contributions also from North America and Australasia) and emphasised meaningful interactions between mathematicians and statisticians.

The workshop had a positive and productive atmosphere, with many intense exchanges between participants of differing disciplinary and methodological backgrounds, as had been hoped for by the organisers. The principal mathematical themes of the workshop were

- forward uncertainty propagation,
- Bayesian inverse problems, and especially posterior consistency and contraction,
- data assimilation and filtering,
- statistical perspectives,
- data science perspectives,
- sampling methodology,
- optimization under uncertainty,
- interplay of probability and numerical analysis.

An intensive but flexible schedule was adopted to properly encompass this broad range of topics: four longer 45-minute talks in the mornings, and four shorter 30-minute talks in the afternoons; on the Tuesday afternoon, PhD students presented their research in a "blitz session" of 10-minute talks that were particularly well received.

Forward Uncertainty Propagation. The basic task of solving differential equations with random inputs is a mature but active area of research. Here numerical methods for stochastic ordinary differential equations (SDEs) were presented by Lang, who analysed new methods for stochastic Hamiltonian systems and Fischer, who introduced a new construction principle for higher-order solvers for SDEs based on spectral deferred correction. A fascinating presentation on random PDEs on evolving hypersurfaces motivated by applications in cell biology was given by Djurdjevac.

Bayesian Inversion, Consistency, and Contraction. A number of talks considered, in addition to the well-posedness of Bayesian procedures with deterministic (Latz) and randomised (Teckentrup) forward models, their consistency and contraction properties. These properties are relatively standard in the finite-dimensional setting (the Bernstein–von Mises theorem), but are much more subtle in the function-space context that is frequently of interest in UQ, as explained by Nickl in the

workshop's opening talk. The practical importance of these considerations was emphasised by Fox, who argued that industrial partners assess the worth of a UQ study by its frequentist accuracy, as measured against withheld data. Frequentist consistency properties are closely related to the small-noise limits of Bayesian posteriors and their points of maximum probability (maximum a posteriori estimators); for measures on infinite-dimensional spaces, such points are non-trivial to characterise, and their analysis was addressed in the talks of Agapiou, Dashti, and Lie. The question of frequentist coverage of credible sets in linear Bayesian inverse problems in the small-noise limit was taken up again by Kekkonen, who described Bernstein–von-Mises theorems for functionals of their solution. Bochkina showed how Bernstein–von-Mises results can be generalised to linear inverse problems which are misspecified in the sense that the true solution need not lie in the support of the prior distribution on the parameter space.

Data Assimilation. Closely related to SDEs are the filtering problems arising in Data Assimilation, where functionals of stochastic dynamical systems are to be inferred given noisy partial observations in order to merge the evolution of a physical system with observational data. A survey of mean-field approaches to the standard filtering problem, attractive due to its amenability to Monte Carlo techniques, was given by Reich. In the nonlinear high-dimensional setting, Marzouk presented an approach for extending the ensemble Kalman filter based on transport maps constructed using convex optimisation techniques, and de Wiljes derived stability as well as pathwise accuracy bounds for the ensemble Kalman–Bucy filter, while Dubinkina proposed an alternative approach to variational data assimilation based on numerical shadowing techniques. Progress in continuing work on applying the ensemble Kalman filter for solving Bayesian inverse problems was reported by Weissmann, who extended previous work utilising a continuous-time limit of the analysis/update step that incorporates noisy observations.

Statistical Perspectives. A primary aim of the workshop was to foster better connections between the mathematical and statistical communities, especially in the European context, and many speakers offered primarily statistical perspectives on UQ (Nickl, Dette, Simoni, Schlather, Klebanov, and Kekkonen). Many participants commented favourably on the active engagement of the statistical community in this UQ workshop; conversely, the statisticians found the problems of interest in UQ to be interesting and wholly non-trivial from a statistical point of view, and also remarked positively on the welcoming and open-minded atmosphere of the workshop.

Data Science. Particularly under the banner of Deep Learning, data science is presently a hot topic that shares some common interests with UQ. Schwab and Gottschalk offered contributions on the approximation properties of deep neural networks versus (more traditional in the UQ community) sparse generalised polynomial chaos expansions, and on uncertainty in semantic segmentation of images.

Sampling Methodology. Markov chain Monte Carlo (MCMC) methods for sampling from the posterior distribution of a Bayesian inverse problem in a function space setting is currently one of the most active and challenging UQ research

topics, requiring tools from statistics, probability theory, functional analysis and numerical analysis. Consequently, it constituted one of the major themes of the workshop, covering issues such as non-Gaussian priors (Hosseini), the effects of inexact evaluation of the transition kernel (Rudolf), robustness in the small-noise limit (Sprungk), MCMC with quasi-stationary target distributions (Wang), non-reversible non-smooth dynamics for faster exploration of the posterior (Ottobre), and the extension of the Stein discrepancy framework for assessing MCMC convergence to heavy-tailed target distributions (Mackey). Multilevel techniques for accelerating Monte Carlo sampling by recursively combining model approximations of reduced accuracy as control variates were the subject of the lectures by Haji-Ali in the context of evaluating risk measures and by Chernov, for estimating the covariance and higher moments of a random field. Multilevel ideas are also at the heart of a method presented by Bachmayr involving periodisation of the covariance kernel of stationary random fields in combination with Meyer wavelets to arrive at alternatives to the standard Karhunen–Loève expansion, thereby yielding improved convergence rates for Hermite expansions of the solution of random diffusion equations. Alternative representation of random fields was also the topic of Kuo, who proposed expansions in periodic functions of random variables which allow the application of highly efficient quasi-Monte Carlo sampling techniques.

Optimisation. UQ methods for optimisation problems aim to minimise the uncertainty/risk in decision and design processes and optimal control. Kouri and Pichler offered talks on the analysis of the resulting optimisation problems, including a discussion of appropriate risk measures in the optimisation framework. In the framework of stochastic control, Tempone presented efficient UQ methods to deal with high-dimensional parameter spaces. A closely related field is experimental design under uncertainty, which was addressed by Kostina in the context of a mixed-integer optimal control problem.

Interplay of Probability and Numerical Analysis. A newly emerging research area known as *Probabilistic Numerical Analysis* advocates viewing standard numerical tasks such as quadrature or equation solving as problems of Bayesian inference. Treating the solution of linear systems of equations by Krylov subspace methods in this way was the subject of the contribution of Cockayne. On the same theme, but more comprehensively, Owhadi gave a fascinating overview of the connections between interpolation, inference and learning as methods of estimating based on partial information connected through the perspective of game theory.

The deep mutual engagement between the mathematical and statistical communities at this workshop was very strongly praised, and bodes well for the UQ community's future. Other aspects of the workshop that attracted positive comments from the participants included the emphasis given to the contributions of younger researchers, e.g. through the 10-minute "blitz" talks on Tuesday afternoon, and the high number of female attendees and speakers.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows".

Workshop: Uncertainty Quantification

Table of Contents

Abstracts

Bayesian Uncertainty Quantification for PDE constrained regression problems

Richard Nickl

We discuss recent results that validate the use of Bayesian methods in infinite-dimensional statistical inverse problems arising with partial differential equations (PDEs). Particularly we prove Bernstein-von Mises theorems which show that the posterior distributions are approximated by certain infinite-dimensional Gaussian distributions whose covariance depends on the information operator characterising the inverse problem, and which can be used to show that Bayesian uncertainty quantification methods (such as credible sets) have objective scientific meaning in the frequentist large sample size scenario. The key examples discussed are a) inferring coefficients of a partial differential operator from noisy observations of a solution of a PDE involving the operator [2], b) inferring an image whose geodesic X-ray transform is observed under additive Gaussian white noise [1], and c) inferring the unknown drift vector field of a multi-dimensional diffusion process based on observations of the trajectory of the solution of the stochastic differential equation describing the diffusion [3].

References

[1] F. Monard, R. Nickl, G. P. Paternain, *Efficient Nonparametric Bayesian inference for X-ray transforms*, Annals of Statistics **47** (2019) 1113–1147.

[2] R. Nickl, *Bernstein - von Mises theorems for statistical inverse problems I: Schrödinger equation,* Journal of the European Mathematical Society, to appear (2019).

[3] R. Nickl, K. Ray, *Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions.* arXiv (2018).

Bayesian methods in inverse problems, and polynomial acceleration of MCMC

Colin Fox

Two under-developed methods that could greatly increase the quality of computational Bayesian methods for uncertainty quantification (UQ) are the mid-level and high-level models for unknown fields in inverse problems, and Markov chain Monte Carlo methods with better than geometric convergence.

The Bayesian formulation for an inverse problem posits the joint distribution for the 'state of nature' $\mathbf{x}$, measured data $\tilde{\mathbf{d}}$, and (hyper)parameters θ. Inference is then focused on the posterior conditional distribution

$$[\mathbf{x}, \theta|\tilde{\mathbf{d}}] = \text{const}_{\mathbf{x}}[\tilde{\mathbf{d}}|\mathbf{x}, \theta][\mathbf{x}|\theta][\theta]$$

with the right-hand side giving the usual expansion as a product of distributions produced in a hierarchical model. Some issues arise in practice: the modelled 'state of nature' is a proxy for the actual processes in nature; evaluating $[\tilde{\mathbf{d}}|\mathbf{x}, \theta]$ requires

simulating the forward map that is a complex non-linear computer model; the physics and geometry are uncertain – this is often the greatest source of uncertainty in practical inverse problems; and the data is sparse and uncertain.

Optimization-based methods seek a 'best' solution by evaluating

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} -\log\left[\tilde{\mathbf{d}}|\mathbf{x}\right] + \lambda R(\mathbf{x})$$

for some regularizing functional R, and regularizing parameter $\lambda > 0$. Computational optimization may be performed by the stationary linear relaxation algorithms of the 1950s, or the modern sophisticated algorithms such as Chebyshev polynomial acceleration, PCG, GMRES, multigrid, and fast-multipole methods. Models for unknown $\mathbf{x}$ are usually continuous valued, from Hilbert spaces, affording the convenience of gradients and projections. These models are *low-level* [5] since they can represent any field but are convenient for stating local information, only. The resulting *point estimates* suffer from well-known problems such as stochastic bias and poor predictive distributions [6].

Valid quantification of uncertainties requires correctly tracking distributions, as in the sample-based Bayesian approach, that evaluates *expectations* of any function f via the Monte Carlo estimate

$$\mathrm{E}\left[f(\mathbf{x})\right] \approx \frac{1}{N}\sum f(\mathbf{x}_i), \qquad \mathbf{x}_i \sim [\mathbf{x}|\tilde{\mathbf{d}}].$$

Computational methods for sampling are dominated by Markov chain Monte Carlo (MCMC), and Gibbs sampling, that are essentially the stationary iterative methods from the 1950s, though more efficient algorithms do exist [3]. Sampling algorithms operate on measure spaces, allowing models with continuous and discrete components, that can be low-, mid-, and high-level [5]. In particular, high-level models allow *counting* the number of objects in a scene [1], something that is practically impossible with low-level function-space representations.

We seek to draw on the best of the optimization and inference approaches, by utilizing the measure-space models used in inference, but accelerate sampling by adapting the sophisticated algorithms developed for optimization.

The advantages of mid- and high-level models can be demonstrated in an example of imaging from limited-angle X-ray absorption data, in an application of automated inspection of solder joints in the electronics industry. Traditional processing of this data first forms a high-resolution, linear-space reconstruction of the X-ray absorption which is then passed to a classification algorithm. The inverse problem is highly ill-posed, with about 15 times too little data. The left panel in Fig. 1 shows a basic difficulty with linear space representations when using limited-angle X-ray data, that reconstruction of a circular object (a) results in 'cones' (b) [2]. However, it is clear that the reconstruction is unphysical since the density of 'solder' decreases in the cones. Problems caused by these unphysical linear-space representations can be avoided by using a mid-level representation [5] of the solder as an object with a bounding surface, and roughly constant properties within the solder, as depicted in Fig. 1 (centre). Then the surface is well

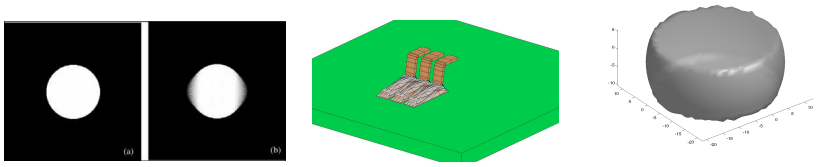

FIGURE 1. Left: Two images from limited-angle X-ray tomography; (a) true image, (b) TV reconstruction showing 'coneheads'. Centre: Surface representation of solder on gull-wing leads. Right: Sample-based estimate of a BGA using a CSG representation.

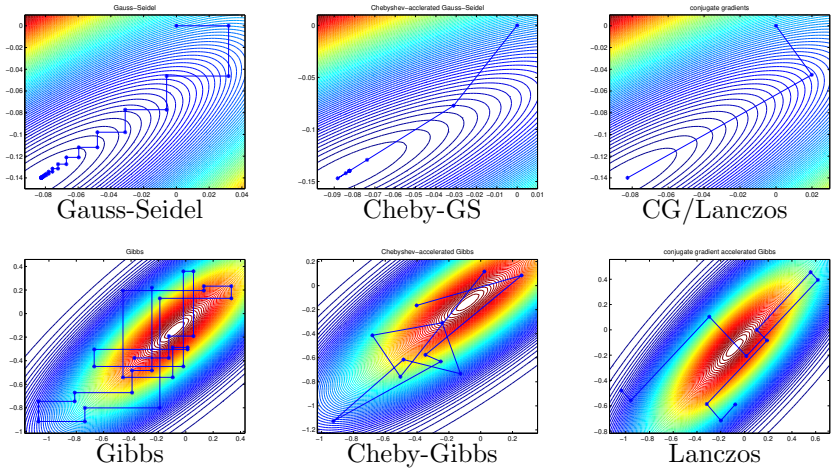

FIGURE 2. Top row shows the contours of a quadratic function $\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$ and the iteration paths for the optimizers: Gauss-Seidel (left), Chebyshev accelerated Gauss-Seidel (mid), and conjugate gradients or Lanczos (right). Bottom row shows contours of the normal density $k \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}\right\}$, and the paths of the equivalent accelerated Gibbs samplers.

defined, and the Fisher information, as appears in the Bernstein von Mises theorem, shows that about 5 times more data has been measured than is required. The ill-posedness was illusory. The right panel in Fig. 1 shows a reconstruction of a BGA using a constructive solid geometry (CSG) representation with a random number of primitives. This algebraic representation on a countably-infinite state space poses significant computational challenges for sampling, though provides excellent reconstructions with features at scales smaller than the 'resolution' of limited-angle X-ray tomography.

Polynomial acceleration and Krylov-space methods can be applied to MCMC, exactly for some finite state-space cases and also for Gibbs sampling of multivariate normal distributions. The latter case is summarized in Fig. 2.

In the quadratic case, these optimizers are usually described as solvers of $\mathbf{Ax} = \mathbf{b}$. The iteration operators, and error polynomials, are identical for the equivalent samplers and solvers, implying that the optimal solver is the optimal sampler. In particular, the Chebyshev accelerated Gibbs sampler has the optimal rate of convergence, in distribution, amongst non-stationary iterations that do not depend on the residual, while the Lanczos sampler is a 'perfect' sampler that generates independent samples in a finite number of steps (in exact arithmetic).

Research is ongoing to extend these results to general, non-normal distributions.

References

[1] Al-Awadhi, F., Jennison, C. and Hurn, M., 2004. Statistical image analysis for a confocal microscopy two-dimensional section of cartilage growth. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1), pp.31-49.

[2] Chen, Z., Jin, X., Li, L. and Wang, G., 2013. A limited-angle CT reconstruction method based on anisotropic TV minimization. *Physics in Medicine & Biology*, 58(7), pp.2119-2141.

[3] S. Dolgov, K. Anaya-Izquierdo, C. Fox, and R. Scheichl, Approximation and sampling of multivariate probability distributions in the tensor train decomposition, *arXiv preprint* arXiv:1810.01212 (2018).

[4] Fox, C. and Parker, A.E., 2017. Accelerated Gibbs sampling of normal distributions using matrix splittings and polynomials. *Bernoulli* 23(4B), pp.3711-3743.

[5] Hurn, M.A., Husby, O. and Rue, H., 2003. Advances in Bayesian image analysis. In *Highly structured stochastic systems* (pp. 301-322). Oxford University Press.

[6] Kaipio, J.P. and Fox, C., 2011. The Bayesian framework for inverse problems in heat transfer. *Heat Transfer Engineering*, 32(9), pp.718-753.

MAP estimators and their consistency for nonparametric Bayesian inverse problems

Masoumeh Dashti

(joint work with Sergios Agapiou, Martin Burger and Tapio Helin)

We consider the inverse problem of recovering an unknown functional parameter $u \in X$ from a noisy and indirect observation $y \in Y$. We work in a framework in which X is an infinite-dimensional separable Banach space, while $Y = \mathbb{R}^J$. In particular, we consider the additive noise model

$$y = \mathcal{G}(u) + \xi, \tag{1}$$

where ξ is mean zero Gaussian observational noise, with a positive definite covariance matrix $\Sigma \in \mathbb{R}^{J\times J}$, $\xi \sim N(0, \Sigma)$. Here the possibly nonlinear operator $\mathcal{G} : X \to Y$ describes the system response, connecting the observation y to the unknown parameter u. More specifically, $\mathcal{G}$ captures both the forward model and the observation mechanism and is assumed to be known.

We adopt a Bayesian approach where prior information is encoded in the *prior* distribution μ_0 on the unknown u, and the Bayesian methodology is used to obtain

the *posterior* distribution μ^y on $u|y$ (provided that $\mathcal{G}$ satisfies some mild conditions [9, 4]), in the form

$$\mu^y(du|y) \propto \exp\big(-\Phi(u;y)\big)\mu_0(du), \tag{2}$$

where

$$\Phi(u;y) = \frac{1}{2}\|\Sigma^{-\frac{1}{2}}\big(y-\mathcal{G}(u)\big)\|_Y^2.$$

We are interested in characterising the modes of the posterior measure or the maximum a posteriori (MAP) estimates. These are of interest since the characterisation of the full posterior measure is in general very difficult or infeasible in the high- or infinite-dimensional case.

In finite dimensional contexts, especially when working with continuous probability distributions, the definition of a mode is straightforward as a maximiser of the probability density function. If the prior has probability density function of the form

$$\pi(u) \propto \exp(-W(u)),$$

for a suitable positive function $W : X \to [0,\infty)$, the density of the posterior is

$$\pi^y(u) \propto \exp(-I(u;y)),$$

where

$$I(u;y) = \Phi(u;y) + W(u). \tag{3}$$

In this case, a MAP estimator can be interpreted as a classical estimator of the unknown, arising from Tikhonov regularisation with penalty term given by the negative log-density of the prior [5].

In the infinite dimensional setting things are less straightforward due to the lack of a uniform reference measure. an intuitive approach is looking at centre points of small balls in X which have maximal probability at the vanishing limit of their radius. The definition given in [3] grows out of this intuition: A point $\hat{u} \in X$ is a mode of the measure μ if it satisfies

$$\lim_{\epsilon\to 0} \frac{\mu(B_\epsilon(\hat{u}))}{\sup_{u\in X}\mu(B_\epsilon(u))} = 1.$$

They then show in the case of Bayesian inverse problems with Gaussian prior measures that such modes are characterised by the minimisers of the Tikhonov functional

$$I(u;y) = \Phi(u;y) + \|u\|_Z^2,$$

where Z is the Cameron-Martin space of the Gaussian prior measure. Later, a weaker notion of mode, which on the other hand allowed for study of the more general class of differentiable priors with continuous logarithmic derivative, was introduced by Helin and Burger [6]: A point $\hat{u} \in X$ is called a *weak mode* of μ if

$$\lim_{\epsilon\to 0} \frac{\mu(B_\epsilon(\hat{u}-h))}{\mu(B_\epsilon(\hat{u}))} \le 1,$$

for all $h \in E$, where E is a dense subspace of X. These were then shown to coincide with strong modes under certain conditions on the prior and forward operator $\mathcal{G}$ by Lie and Sullivan [8].

We consider continuous, but not necessarily differentiable, prior measures. We show that, for locally Lipschitz forward operators, weak modes of the posterior coincide with minimisers of an appropriate Tikhonov functional [1]. One important class of such priors, due to their sparsity promoting properties, is the class of p-Besov priors with $p = 1$. These priors were proposed and shown to be discretisation invariant in [7] and a well-posedness theory of the posterior was developed in [2]. These priors are defined by a wavelet expansion with random coefficients, motivated by a formal density of the form

$$\pi(u) \propto \exp(-\|u\|^p_{B^s_{pp}}),$$

where $\|\cdot\|_{B^s_{pp}}$ is the Besov space norm with regularity parameter s and integrability parameters p. For $p = 1$, we show that the modes of the posterior coincide with the minimisers of

$$I(u; y) = \Phi(u; y) + \|u\|_{B^s_{11}}.$$

Such characterisation then allows for the study of the weak consistency of the posterior: As the number of data points increases, or as the observational noise decreases, the modes of the posterior converge in probability to the underlying truth provided that the forward operator is injective.

References

[1] S. Agapiou, M. Burger, M. Dashti, and T. Helin. Sparsity-promoting and edge- preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. *Inverse Problems*, 34(4):045002, 2018.

[2] M. Dashti, S. Harris, and A. Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183?200, 2012.

[3] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9):095017, 27, 2013.

[4] M. Dashti and A.M.Stuart. The Bayesian Approach to Inverse Problems. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of Uncertainty Quantification.* Springer, 2015.

[5] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science and Business Media, 1996.

[6] T. Helin and M. Burger. Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Problems*, 31(8):085009, 2015.

[7] M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87?122, 2009.

[8] H. C. Lie and T. J. Sullivan. Equivalence of weak and strong modes of measures on topological vector spaces. *Inverse Problems*, 34(11):115013, 2018.

[9] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451?559, 2010.

Functional data analysis in the Banach space of continuous functions

Holger Dette
(joint work with K. Kokot and A. Aue)

Most of functional data analysis is based on Hilbert space-based methodology for which there exists by now a fully fledged theory. Since all functions utilized for practical purposes are at least continuous, and often smoother we develop in this talk methodology for functional data in the space of continuous functions.

We concentrate on the sapce $C(T)$, the space of continuous functions on the compact interval $T = [0, 1]$ equipped with the sup-norm $\|f\| = \sup_{t\in T} |f(t)|$. If μ_1 and μ_2 are the mean functions corresponding to two samples we are interested in testing hypotheses of the form

$$H_0 \colon \|\mu_1 - \mu_2\| \leq \Delta \qquad \text{and} \qquad H_1 \colon \|\mu_1 - \mu_2\| > \Delta, \tag{1}$$

where $\Delta \geq 0$ denotes a pre-specified constant. The classical case of testing perfect equality, obtained by the choice $\Delta = 0$, is therefore a special case of (1). It turns out that from a mathematical point of view the problem of testing relevant (i.e., $\Delta > 0$) hypotheses is substantially more difficult than the classical problem (i.e., $\Delta = 0$). In particular, it is not possible to work with stationarity under the null hypothesis, making the derivation of a limit distribution of a corresponding test statistic or the construction of a bootstrap procedure substantially more difficult. If $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent stationary samples in $C([0, 1])$ with mean functions μ_1 and μ_2 a test for the hypotheses (1) can be based on the statistic

$$\hat{d}_\infty = \|\bar{X}_m - Y_n\|,$$

which is a natural estimate of the unknown distance

$$d_\infty = \|\mu_1 - \mu_2\|.$$

One of our main results establishes the asymptotic distribution of $\hat{d}_\infty$.

Theorem 1. *If $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are sampled from independent stationary time series $(X_j \colon j \in \mathbb{N})$ and $(Y_j \colon j \in \mathbb{N})$ in $C([0, 1])$ with mean functions μ_1 and μ_2, respectively. satisfying the conditions*

- (A1) *There is a constant K such that $\mathbb{E}[\|X_j\|^{2+\nu}] \leq K$ and $\mathbb{E}[\|Y_j\|^{2+\nu}] \leq K$ for some $\nu > 0$.*
- (A2) *There is a real-valued random variable M with $\mathbb{E}[M^2] < \infty$ such that*

$$|X_j(t) - X_j(t')| \leq M|t - t'| \ , \ j = 1, \ldots, m$$
$$|Y_j(t) - Y_j(t')| \leq M|t - t'| \ , \ j = 1, \ldots, n$$

  holds almost surely for all $t, t' \in T$.
- (A3) *$(X_j \colon j = 1, \ldots, m)$ and $(Y_j \colon j = 1, \ldots, n)$ are φ-mixing with exponentially decreasing mixing coefficients, that is, there is a constant $a \in [0, 1)$ such that $\varphi(k) \leq a^k$ for any $k \in \mathbb{N}$.*

then as $m, n \to \infty$ *and* $m/(m+n) \to \lambda \in (0,1)$

$$(2) \quad T_{m,n} = \sqrt{n+m}(\hat{d}_\infty - d_\infty) \xrightarrow{\mathcal{D}} T(\mathcal{E}) = \max\Big\{ \sup_{t\in\mathcal{E}^+} Z(t), \sup_{t\in\mathcal{E}^-} -Z(t)\Big\},$$

where the centered Gaussian process Z *with covariance structure*

$$C(s,t) = \mathrm{Cov}(Z(s), Z(t)) = \frac{1}{\lambda} C_1(s,t) + \frac{1}{1-\lambda} C_2(s,t),$$

$$C_1(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(X_1(s), X_{1+i}(t)) \ , \ C_2(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(Y_1(s), Y_{1+i}(t))$$

and the sets $\mathcal{E}^+$ *and* $\mathcal{E}^-$ *are defined in*

$$\mathcal{E}^\pm = \big\{ t \in [0,1] \colon \mu_1(t) - \mu_2(t) = \pm d_\infty \big\}$$

As the asymptotic distribution in Theorem 1 is not distribution free, we also develop estimates of the extremal sets $\mathcal{E}^+$ and $\mathcal{E}^-$ which are defined by

$$\hat{\mathcal{E}}^+_{m,n} := \Big\{ t \in [0,1] \ \Big| \ \bar{X}_m(t) - \bar{Y}_n(t) \geq \hat{d}_\infty - c\frac{\log(m+n)}{\sqrt{m+n}} \ \Big\}$$

$$\hat{\mathcal{E}}^-_{m,n} := \Big\{ t \in [0,1] \ \Big| \ \bar{X}_m(t) - \bar{Y}_n(t) \leq -\hat{d}_\infty + c\frac{\log(m+n)}{\sqrt{m+n}} \ \Big\}$$

The second main result established consistency of these estimates with respect to the Hausdorff distance d_H.

Theorem 2. *Let the assumptions of Theorem 1 be satisfied, then*

$$d_H(\hat{\mathcal{E}}^\pm_{m,n}, \mathcal{E}^\pm) \xrightarrow{\mathbb{P}} 0.$$

The results of Theorem 1 and 2 are used to develop a multiplier bootstrap procedure, to generate critical values for the test, which rejects the null hypotheses for large vales the statistic $\hat{d}_\infty$. We prove that this test has asymptotic level α and is consistent.

Rates of contraction of posterior distributions with product priors: beyond Gaussianity

Sergios Agapiou

(joint work with Masoumeh Dashti, Tapio Helin)

We are interested in performing Bayesian inference on functional unknowns. In particular, we are interested in unknowns which exhibit blocky structure (large areas of constant value) as well as sparsity in an appropriate expansion and this should be taken into account in our choice of the prior.

We construct priors on a function u, by randomising the coefficients in a series expansion. For example, if we work in $L_2[0,1]$, we let (ψ_ℓ) be an orthonormal basis and write

$$u(x) = \sum_{\ell=1}^{\infty} \gamma_\ell \xi_\ell \psi_\ell(x),$$

where (γ_ℓ) is a deterministic decaying scaling sequence and (ξ_ℓ) is a sequence of independent and identically distributed real random variables. The scalings γ_ℓ determine the regularity of draws, while to indeed get priors which promote sparsity and blocky structure we need to choose appropriately the distribution of the ξ_ℓ and the basis (ψ_ℓ). A standard choice in the Bayesian inverse problems community giving priors with these properties, is Laplace-distributed coefficients coupled with appropriate wavelet bases, [7]. We have previously studied such priors from the point of view of MAP estimators in [1].

In this talk we consider priors on a separable Banach space X constructed via random series expansions with ξ_ℓ having a mean-zero p-exponential distribution, where $p \in [1,2]$. For $p = 1$ we get Laplace-type priors while for $p = 2$ Gaussian priors. We call these priors p-exponential.

We study rates of contraction of posterior distributions based on p-exponential priors, in nonparametric settings where we observe $X^{(n)} \sim P_\theta^{(n)}$ and our aim is to infer the functional parameter $\theta \in \Theta$. Such rates, assume that the observations are generated from a fixed underlying value of the unknown, θ_0, and measure the concentration rate of the posterior distribution around this underlying value in the limit $n \to \infty$, corresponding to infinitely informative data.

More specifically, we generalize the Gaussian contraction theory of Aad van der Vaart and Harry J. van Zanten in [5], which is based on the concentration properties of Gaussian measures. To this end we study the concentration of p-exponential measures and use then use the techniques of [5].

We give a short overview of the general posterior contraction theory developed in [3, 4], where general here refers to general models and general priors. This theory gives conditions for a rate ϵ_n to be a rate of contraction: the prior needs to put a certain minimum mass around the truth and we need to have subsets $\Theta_n \subset \Theta$ which contain the bulk of the prior mass and have appropriately bounded complexity.

We start the study of the concentration properties of p-exponential priors, by determining their space of admissible shifts Q, that is the space of translations which are such that the translated and the original measures are equivalent. This space is a weighted ℓ_2-type space, with weights given by the scalings γ_ℓ. We also determine the Radon-Nikodym derivative between the translated and the original measure, which involves weighted ℓ_p-type terms.

We next define another space Z which is a weighted ℓ_p-type space, again with weights given by γ_ℓ and where p is the exponent of the prior. For Gaussian measures, $p = 2$, $Z = Q$ and these spaces are identified to the Reproducing Kernel Hilbert Space, while for $p \in [1,2)$ we have $Z \subset Q$. We show a lower bound on the loss of probability occured by shifting from balls in X centered at the origin

to balls centered around an $h \in Z$. This loss is exponential in the Z-norm of the center. Based on this lower bound and approximation, we can define the so called concentration function $\phi_w(\cdot)$ of a p-exponential measure, for a fixed $w \in X$. The concentration function at an $\epsilon > 0$ controls the probability of balls in X centered at w.

We also present a concentration inequality, which is based on a two-level concentration inequality due to Talagrand [6]. Eventhough it is straightforward to see that both the spaces Q and Z have zero measure, this inequality shows that the bulk of the mass of a p-exponential prior is contained in $\epsilon B_X + M^{\frac{p}{2}} B_Q + M B_Z$ for small $\epsilon > 0$ and large $M > 0$.

We then present our contraction result for general models for p-exponential priors. This result says that for a fixed $w_0 \in X$, if a rate ϵ_n satisfies $\phi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$, then the prior puts a certain minimum mass around w_0, and we can find sets $X_n \subset X$ which contain the bulk of the prior mass and have bounded complexity.

The proof of this theorem follows in a relatively straightforward way using our previous considerations. Its assertions point to the conditions of the aforementioned general contraction results in [3, 4] for obtaining rates of contraction. Compared to the Gaussian case $p = 2$, there is a certain intricacy in the complexity bound when $p \in [1, 2)$, due to the space Z being a strict subset of the space Q. Whether this intricacy affects the obtained rates or not, depends on the quality of approximation of Q by Z in the norm of X.

Finally, we apply our contraction result to obtain rates of posterior contraction in the white noise model in $L_2[0, 1]$, with α-regular p-exponential priors, under Sobolev-type regularity of the truth. In this case the intricate complexity bound is benign, and the rates we obtain are minimax optimal when the regularity of the prior matches the regularity of the truth. We can get similar rates in other separable Hilbert space settings, such as nonparametric binary regression or nonparametric regression in L_2.

A preprint of the article containing this work can be found in [2]. Possible future directions include studying the frequentist performance of credible sets with p-exponential priors, or studying contraction rates for heavier-tailed product priors.

References

[1] S. Agapiou, M. Burger, M. Dashti, and T. Helin. *Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems*, Inverse Problems **34** (2018), 045002.

[2] S. Agapiou, M. Dashti, and T. Helin. *Rates of contraction of posterior distributions based on p-exponential priors*, arXiv:1811.12244.

[3] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart, *Convergence rates of posterior distributions*, The Annals of Statistics **28** (2000),c500–531.

[4] S. Ghosal and A. Van Der Vaart, *Convergence rates of posterior distributions for noniid observations*, The Annals of Statistics **35** (2007), 192–223.

[5] A. Van Der Vaart and J. H. Van Zanten, *Rates of contraction of posterior distributions based on gaussian process priors*, The Annals of Statistics **36** (2008), 1435–1463.
[6] M. Talagrand. *The supremum of some canonical processes*, American Journal of Mathematics **116** (1994), 283–325.
[7] M. Lassas, E. Saksman, and S. Siltanen. *Discretization-invariant Bayesian inversion and Besov space priors*, Inverse Problems and Imaging **3** (2009), 87–122.

Drift-preserving numerical integrators for stochastic Hamiltonian systems

Annika Lang
(joint work with Chuchu Chen, David Cohen, Raffaele D'Ambrosio)

For a positive integer m and a smooth potential $V\colon \mathbb{R}^m \to \mathbb{R}$, let us consider the separable Hamiltonian

$$H(p,q) = \frac{1}{2}\sum_{j=1}^{m} p_j^2 + V(q).$$

The goal of this presentation is to introduce a new approximation scheme for the stochastic Hamiltonian system

$$\begin{aligned} \mathrm{d}q(t) &= \nabla_p H(p(t),q(t))\,\mathrm{d}t \\ \mathrm{d}p(t) &= -\nabla_q H(p(t),q(t))\,\mathrm{d}t + \Sigma\,\mathrm{d}W(t) \end{aligned}$$

with initial condition $(p(0),q(0))$ driven by an additive $\mathbb{R}^d$-valued Wiener process W in the p component. This system simplifies to

$$\begin{aligned} \mathrm{d}q(t) &= p(t)\,\mathrm{d}t \\ \mathrm{d}p(t) &= -V'(q(t))\,\mathrm{d}t + \Sigma\,\mathrm{d}W(t). \end{aligned}$$

By the Itô formula, one shows that p and q satisfy for all $t>0$ the following *trace formula* for the energy

$$\mathbb{E}\left[H(p(t),q(t))\right] = \mathbb{E}\left[H(p_0,q_0)\right] + \tfrac{1}{2}\mathrm{tr}\left(\Sigma^\top\Sigma\right)t.$$

For an equidistant time discretization $(t_0,\dots,t_N)$ with step size h, define the new *approximation scheme*

$$\begin{aligned} \Psi_{n+1} &= p_n + \Sigma\Delta W_n - \frac{h}{2}\int_0^1 V'(q_n + sh\Psi_{n+1})\,\mathrm{d}s \\ q_{n+1} &= q_n + h\Psi_{n+1} \\ p_{n+1} &= p_n + \Sigma\Delta W_n - h\int_0^1 V'(q_n + sh\Psi_{n+1})\,\mathrm{d}s \end{aligned}$$

with $(p_0,q_0) = (p(0),q(0))$. This scheme preserves the trace formula, i.e., for all $n=1,\dots,N$

$$\mathbb{E}\left[H(p_n,q_n)\right] = \mathbb{E}\left[H(p_0,q_0)\right] + \tfrac{1}{2}\mathrm{tr}\left(\Sigma^\top\Sigma\right)t_n.$$

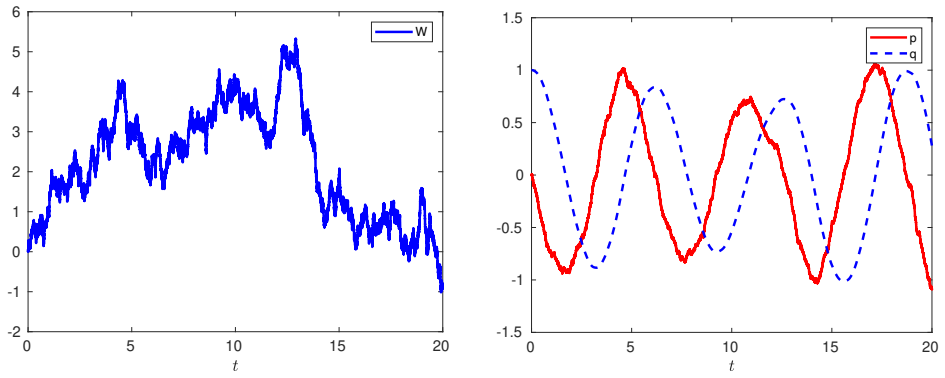

FIGURE 1. Sample path of a Wiener process (left) and the corresponding solution of the linear stochastic oscillator (right).

Furthermore, the scheme converges *strongly* and *weakly* with order 1. More specifically, for sufficiently small $h > 0$, one proves

$$\left(\mathbb{E}\left[|q(t_n) - q_n|^2\right]\right)^{1/2} + \left(\mathbb{E}\left[|p(t_n) - p_n|^2\right]\right)^{1/2} \leq Ch$$

and for sufficiently smooth test functions f

$$|\mathbb{E}\left[f(p(t_n), q(t_n))\right] - \mathbb{E}\left[f(p_n, q_n)\right]| \leq Ch.$$

It is clear from the trace formula that the choice $f = H$ leads to no weak error. In numerical experiments one observes further for specific test functions weak convergence of order 2.

The approximation of the expectation by a multilevel Monte Carlo methods decreases the computational work to essentially h^{-2} compared to h^{-3} for a standard Monte Carlo approximation while preserving the accuracy.

Numerical examples including the linear stochastic oscillator, the stochastic mathematical pendulum, and a double well potential confirm the theoretical results. A sample path of a Wiener process and the corresponding solution for the linear stochastic oscillator is shown in Figure 1.

Multilevel Multilevel Monte Carlo approximation of moments, covariance functions and the maximum entropy reconstruction

ALEXEY CHERNOV

(joint work with Claudio Bierig, Erik M. Schetzke)

The Multilevel Monte Carlo Method (MLMC) is a technique for efficient computation of observable's statistics by approximate sampling in the case when generation of samples of different accuracy is possible. The method is particularly advantageous for complex (possibly non-smooth/nonlinear) problems with low regularity, typically resulting in high memory and CPU time demands. The idea is based on

the observation that coarse sample approximations can be used as control variates for more accurate sample approximations and thereby reduce the variance of the Monte Carlo estimator [1, 2, 3].

In this talk we explain how to use the MLMC framework to estimate the pointwise variance, pointwise higher order centred moments and the covariance of a random field. The estimators are unbiased and have the form of multilevel telescoping sums. Once (generalized) moments of arbitrary order are computed via MC or MLMC simulations, they can be used for reconstruction of the density function by the maximum entropy method. We discuss this methodology as well.

To set up the stage, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and H be a separable Hilbert space. We are interested in estimating probabilistic characteristics of random fields $X : \Omega \to H$. For example, if X is a solution of a second order random elliptic PDE in a physical domain D, then H is the Sobolev space $H^1(D)$, or a closed subspace thereof, cf. e.g. [2]. For a more involved example modelling contact with rough random obstacles see [4, 5, 6]. In the case when X is the value of a linear on nonlinear functional of the computed solution (e.g. a point value, or a volume/surface average) we may have $H = \mathbb{R}$, cf. [3]. It is well known, that the expectation $\mathbb{E}[X]$ can be well approximated by the sample average

$$E_M[X] = \frac{1}{M}\sum_{i=1}^{M} X^i, \qquad \text{where } X^i \text{ are i.i.d. samples of } X. \tag{1}$$

Typically (this also applies to the above examples) practical evaluation of samples X^i of X involves some sort of approximation procedure, e.g. discretization of a PDE, at a prescribed accuracy level ℓ. In other words, only an approximate random field $X_\ell \approx X$ is practically available for sampling. This introduces an additional bias in the estimate so that it holds that

$$\|\mathbb{E}[X] - E_M[X_L]\|^2_{L^2(\Omega,\mathbb{P};H)} = \|\mathbb{E}[X - X_L]\|^2_H + \frac{1}{M}\mathcal{V}(X_L) \tag{2}$$

where $\mathcal{V}(X) = \|X - \mathbb{E}[X]\|^2_{L^2(\Omega,\mathbb{P};H)}$ is an *aggregated variance* (it coincides with the standard variance of X if X is real-valued, i.e. $H = \mathbb{R}$). Relation (2) clearly demonstrates that any a priori prescribed error tolerance ε can be achieved if the approximation error level L and the number of samples M are appropriately balanced.

As indicated in [1, 2, 3] the same accuracy ε can be achieved at a reduced computational cost if the following multilevel estimator is used instead of (1)

$$E^{\mathrm{ML}}[X] = \sum_{\ell=1}^{L} E_{M_\ell}[X_\ell - X_{\ell-1}], \qquad X_0 \equiv 0 \tag{3}$$

with the variable number of samples M_ℓ at different levels $\ell = 1, \ldots, L$. Importantly, the samples of the level correction $X^i_\ell - X^i_{\ell-1} =: Y^i_\ell$ are built upon two *correlated* approximate realizations of X: X^i_ℓ and $X^i_{\ell-1}$ as approximations for the same sample X^i at two consecutive levels. If, moreover, Y^i_ℓ and Y^j_k are independent

for $k \neq \ell$, the mean squared error admits the representation

$$\|\mathbb{E}[X] - E^{ML}[X]\|^2_{L^2(\Omega,\mathbb{P};H)} = \|\mathbb{E}[X - X_L]\|^2_H + \sum_{\ell=1}^{L} \frac{1}{M_\ell}\mathcal{V}(Y_\ell). \tag{4}$$

The reduction of the computational cost is justified by the following observation: when $X_\ell \to X$ in an appropriate sense, it follows that $\mathcal{V}(Y_\ell) \to 0$ as $\ell \to \infty$. Hence, the optimal balance in the right-hand side of (4) is achieved when $\{M_\ell\}$ a rapidly decreasing sequence, we refer to [1, 2, 3] and [4, Sect. 3] for further details. On the other hand, the estimator (3) can be viewed as a variance reduction technique, where the level corrections Y_ℓ serve as control variates for $Y_{\ell+1}$.

If realizations of X are functions in a physical domain D, then $\mu = \mathbb{E}[X]$ is a pointwise average $\mu = \mu(x)$, $x \in D$. The next quantity that we are interested in is the pointwise variance $v(x) = \mathbb{V}\mathrm{ar}[X(x)]$ where

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{5}$$

Mimicking the hierarchical structure in (3), we propose in [4] the following pointwise multilevel sample variance estimator

$$\mathrm{Var}^{\mathrm{ML}}[X] := \sum_{\ell=0}^{L} \Big\{ \mathrm{Var}_{M_\ell}[X_\ell] - \mathrm{Var}_{M_\ell}[X_{\ell-1}] \Big\} \tag{6}$$

with the unbiased sample variance

$$\mathrm{Var}_M[X] := \frac{1}{M-1}\sum_{i=1}^{M}(X^i - E_M[X])^2. \tag{7}$$

Similarly as in (3), the expression in (6) in curly brackets involves two *correlated* approximate realizations of X: X^i_ℓ and $X^i_{\ell-1}$ as approximations for the same sample X^i at two consecutive levels. This structure results in the error representation [4, (5.5)] that is similar to (4). The error/cost estimates [4, Theorems 5.2, 7.4] demonstrate, in particular, that $\mathbb{E}[X]$ and $\mathbb{V}\mathrm{ar}[\mathrm{X}]$ can be estimated by (3) and (6) to the same accuracy with nearly the same computational cost. In [5] these results were extended to multilevel approximations of the centred moments of arbitrary order r.

The situation is somewhat different when (auto)covariance functions are estimated. Here the issue is that handling the natural one-level estimator

$$\mathrm{Cov}_M[X_\ell, X_\ell] = \frac{1}{M-1}\sum_{i=1}^{M}(X^i_\ell - E_M[X_\ell]) \otimes (X^i_\ell - E_M[X_\ell]) \tag{8}$$

leads to the asymptotic complexity $\mathcal{O}(N_\ell^2)$ (e.g. due to storage requirements), when the representation of one sample X^i_ℓ involves $\mathcal{O}(N_\ell)$ unknowns. In [7] we propose a simple sparse tensor multilevel covariance estimator that overcomes this difficulty and does not require complicated tools, as wavelet projections, cf. [2].

Suppose now that X is real-valued, i.e. $H = \mathbb{R}$, and $\{\phi_1, \ldots, \phi_R\}$ is a linearly independent family, e.g. algebraic or trigonometric polynomials. The previously

described techniques enable estimation of (generalized) moments $\mu_k = \mathbb{E}[\phi_k(X)]$. In [6] we address the question of approximate recovery of the unknown density ρ of X from its first R generalized moments, and the convergence of the approximation $\rho_R \to \rho$ as $R \to \infty$. Notably, if the reconstruction ρ_R exists, it is usually non-unique. To gain a unique reconstruction ρ_R we select the one that, in addition to the moment constraints, maximizes the Shannon entropy. Such densities can be equivalently characterized as elements of exponential families

$$\begin{cases} \rho_R \propto \exp\left(\sum_{k=1}^{R} \lambda_k \phi_k(x)\right), & \lambda_k \in \mathbb{R}, \\ \mu_k = \int \phi_k(x)\rho_R(x)\,dx, & k = 1, \dots, R. \end{cases} \tag{9}$$

Here the coefficients λ_k should be determined so that the moment constraints in the second line are satisfied. The nonlinear system of equations (9) can be solved e.g. by the Newton-Raphson method. In [6] we provide a complete *a priori* asymptotic convergence theory for the maximum entropy approach combined with the MLMC approximation of the generalized moments. Our numerical tests demonstrate that the regularity assumptions can possibly be relaxed. An interesting open question is how to select the concrete number of moments R in practical computations: an overestimation may lead to overfitting and thereby to inaccurate reconstructions.

References

[1] M.B. Giles, *Multilevel Monte Carlo path simulation*, Oper. Res. **56**(3) (2008), 607–617

[2] A. Barth, C. Schwab, N. Zollinger, *Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients*, Numer. Math. **119**(1) (2011), 123–161

[3] K.A. Cliffe, M.B. Giles, R. Scheichl, A.L. Teckentrup, *Multilevel Monte Carlo methods and appli cations to elliptic PDEs with random coefficients* Comput. Vis. Sci. **14**(1) (2011), 3–15

[4] C. Bierig and A. Chernov, *Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems*, Numer. Math., **130** (2015), 579–613

[5] C. Bierig and A. Chernov, *Estimation of arbitrary order central statistical moments by the multilevel Monte Carlo method*, J. Stoch. PDE Ana. Comp., 4 (2016), 3–40

[6] C. Bierig and A. Chernov, *Approximation of probability density functions by the Multilevel Monte Carlo Maximum Entropy method*, J. Comput. Physics, **314** (2016), 661–681

[7] A. Chernov and E. Schetzke, *A simple, bias-free approximation of covariance functions by the Multilevel Monte Carlo method having nearly optimal complexity*, in prep. (2019)

Bayesian Estimation and Comparison of Conditional Moment Models

Anna Simoni
(joint work with Siddhartha Chib, Minchul Shin)

We consider models in which the unknown distribution of the outcomes is non specified except for a set of conditional moment restrictions of the type

$$\mathbf{E}^P[\rho(X,\theta)|Z] = 0, \tag{1}$$

where $\rho(X,\theta)$ is a known vector-valued function of a random vector X and an unknown parameter vector θ, and P is the unknown conditional distribution of X given a random vector Z. Models of this type arise in many different situations (*e.g.* in causal inference – where Z is typically a vector of instrumental variables, or in models of heteroskedasticity of unknown form), and provide a useful semi-parametric modeling of outcomes under minimal probabilistic assumptions. In this article, we provide a Bayesian analysis of such problems, extending and completing the theory developed by [3] for models with unconditional moment conditions. The prior-posterior analysis is based on the nonparametric exponentially tilted empirical likelihood (ETEL) function, constructed to satisfy a sequence of unconditional moments which are obtained from the conditional moments by an increasing (in sample size) vector of approximating functions.

Let $X := (X_1', X_z')'$ be an $\mathbb{R}^{d_x}$-valued random vector and $Z := (Z_1', X_z')'$ be an $\mathbb{R}^{dz}$-valued random vector. The vectors Z and X have elements in common if the dimension of the subvector X_z is non-zero. Moreover, we denote $W := (X', Z_1')' \in \mathbb{R}^{d_w}$ and its (unknown) joint distribution by P. By abuse of notation we use P also to denote the associated conditional distribution. We suppose that we are given a random sample $w_{1:n} = (w_1, \ldots, w_n)$ of W. Hereafter, we denote by $\mathbf{E}^P[\cdot]$ the expectation taken with respect to P and by $\mathbf{E}^P[\cdot|\cdot]$ the conditional expectation taken with respect to the conditional distribution associated with P.

Under certain circumstances (see *e.g.* [1, 2]), conditional moment restrictions are equivalent to a countable number of unconditional moment restrictions. The equivalent set of unconditional moments are obtained through approximating functions. Let $q^K(z) = (q_1^K(z), \ldots, q_K^K(z))'$, $K > 0$, denote a K-vector of functions of Z, for instance, splines, truncated power series, or Fourier series. We assume that these functions are good approximating functions in the following sense. For all K, we assume that $\mathbf{E}^P[q^K(Z)'q^K(Z)]$ is finite, and for any $a(z)$ with $\mathbf{E}^P[a(Z)^2] < \infty$ there are $K \times 1$ vectors γ_K such that as $K \to \infty$,

$$\mathbf{E}^P[(a(Z) - q^K(Z)'\gamma_K)^2] \to 0.$$

If $\mathbf{E}^P[\rho(X,\theta)'\rho(X,\theta)] < \infty$, then [4, Lemma 2.1] established that: (I) if equation (1) is satisfied with $\theta = \theta_*$ then $\mathbf{E}^P[\rho(X,\theta_*) \otimes q^K(z)] = 0$ for all K; (II) if equation (1) is not satisfied then $\mathbf{E}^P[\rho(x,\theta_*) \otimes q^K(z)] \neq 0$, $\forall K$ large enough. It follows that we can construct our Bayesian inference based on the expanded moment functions

$$g(W,\theta) := \rho(X,\theta) \otimes q^K(Z),$$

and on the unconditional moment conditions

$$\mathbf{E}^P[\rho(X,\theta)\otimes q^K(Z)]=0 \tag{2}$$

which are equivalent to (1) as $K\to\infty$.

For a given value of K, the prior on (θ,P) is specified as $\pi(\theta)\pi(P|\theta,K)$, where the prior on θ is standard. Our default prior on θ is a product of independent student-t distributions with 2.5 degrees of freedom on each component of θ. Our prior on P is a mixture of uniform probability densities restricted to satisfy the expanded moment restrictions $\mathbf{E}^P[g(W,\theta)]=0$, given (θ,K), and is an extension of the prior proposed by [5] for the unconditional moment condition models. As in the case of the unconditional moments problem, the posterior distribution of θ, after marginalization over the nonparametric prior on P, has the form

$$\pi(\theta|w_{1:n},K)\propto\pi(\theta)p(w_{1:n}|\theta,K)$$

where

$$p(w_{1:n}|\theta,K)=\prod_{i=1}^n\widehat{p}_i(\theta)$$

is the Exponential Tilting (ET) empirical likelihood (ETEL) for a given K, and $\{\widehat{p}_i(\theta),i=1,\ldots,n\}$ are the probabilities that minimize the KL divergence between the probabilities $(p_1,\ldots,p_n)$ assigned to each sample observation and the empirical probabilities $(\frac{1}{n},\ldots,\frac{1}{n})$, subject to the conditions that the probabilities $(p_1,\ldots,p_n)$ sum to one and that the expectation under these probabilities satisfy the given unconditional moment conditions (2). That is, for every θ, $\{\widehat{p}_i(\theta),i=1,\ldots,n\}$ solve the following constrained maximization problem:

$$\max_{p_1,\ldots,p_n}\sum_{i=1}^n[-p_i\log(np_i)]\qquad\text{subject to: }\sum_{i=1}^n p_i=1,\quad\sum_{i=1}^n p_i g(w_i,\theta)=0,\quad p_i\geq 0.$$

These probabilities are computed conveniently from the dual representation as

$$\widehat{p}_i(\theta):=\frac{e^{\widehat{\lambda}(\theta)'g(w_i,\theta)}}{\sum_{j=1}^n e^{\widehat{\lambda}(\theta)'g(w_j,\theta)}}\qquad(i=1,\ldots,n)$$

where $\widehat{\lambda}(\theta)=\arg\min_{\lambda\in\mathbb{R}^{dK}}\frac{1}{n}\sum_{i=1}^n e^{\lambda' g(w_i,\theta)}$ is the estimated tilting parameter. Therefore, by multiplying the ETEL function by the prior density of θ, the posterior distribution takes the form

$$\pi(\theta|w_{1:n},K)\propto\pi(\theta)\prod_{i=1}^n\frac{e^{\widehat{\lambda}(\theta)'g(w_i,\theta)}}{\sum_{j=1}^n e^{\widehat{\lambda}(\theta)'g(w_j,\theta)}}.$$

In the paper, we study asymptotic properties of the posterior distribution of θ from a frequentist point of view. This means that we admit the existence of a true value θ_* of the parameter of interest θ and a true value P_* of the data distribution P. A powerful result of our theory shows that, subject to a growth rate condition on the number of approximating functions, the Bayesian posterior distribution of the local parameter $h:=\sqrt{n}(\theta-\theta_*)$ satisfies the Bernstein-von Mises theorem.

Theorem 1 (Bernstein - von Mises)**.** *Under Assumptions 3.1-3.7 in the paper, if* $K \to \infty$*,* $\zeta(K)K^2/\sqrt{n} \to 0$*, and if for any* $\delta > 0$*,* $\exists \epsilon > 0$ *such that as* $n \to \infty$

$$P\left(\sup_{\|\theta-\theta_*\|>\delta} \frac{1}{n}\sum_{i=1}^{n}(\ell_{n,\theta}(w_i) - \ell_{n,\theta_*}(w_i)) \leq -\epsilon\right) \to 1, \tag{3}$$

then the posterior distribution of the local parameter h *converges in total variation towards a random Normal distribution, that is,*

$$\sup_{B} \left|\pi(\sqrt{n}(\theta-\theta_*) \in B|w_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_*},V_{\theta_*}}(B)\right| \xrightarrow{p} 0 \tag{4}$$

where $B \subseteq \Theta$ *is any Borel set,* $\Delta_{n,\theta_*} := -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_{\theta_*} D(z_i)'\Sigma(z_i)^{-1}\rho(x_i,\theta_*)$ *is bounded in probability and* $V_{\theta_*} := \left(\mathbf{E}^P[D(Z)'\Sigma(Z)^{-1}D(Z)]\right)^{-1}$*.*

In the theorem we have used the notation $\ell_{n,\theta}(w_i) := \log \widehat{p}_i(\theta)$, $\rho_\theta(X,\theta) := \frac{\partial \rho(X,\theta)}{\partial \theta'}$, $D(z) := \mathbf{E}^P[\rho_\theta(X,\theta_*)|z]$, and $\Sigma(z) := \mathbf{E}^P[\rho(X,\theta_*)\rho(X,\theta_*)'|z]$. Moreover, $\zeta(K)$ is a constant scalar such that $\sup_{z\in\mathcal{Z}} \|q^K(z)\| \leq \zeta(K)$ for each K. The asymptotic covariance V_{θ_*} of the posterior distribution coincides with the semi-parametric efficiency bound given in [2] for conditional moment condition models.

In the paper we also show that a Bernstein-von Mises theorem holds even when the set of conditional moments contains misspecified moment conditions. Large-sample theory for comparing different conditional moment models, that are non-nested (or nested) and misspecified, based on the behavior of the model marginal likelihoods, is also developed. We prove an important result that the marginal likelihood criterion selects the model that is less misspecified, that is, the model that is closer to the unknown true distribution in terms of the Kullback-Leibler divergence. Several examples with simulated and real data are used to illustrate the large practical ramifications of the framework and results.

References

[1] H.J. Bierens, *Consistent model specification tests*, Journal of Econometrics, **20** (1982), 105–134.

[2] G. Chamberlain, *Asymptotic Efficiency in Estimation with Conditional Moment Restrictions*, Journal of Econometrics, **34** (1987), 305–334.

[3] S. Chib, M. Shin and A. Simoni, *Bayesian Estimation and Comparison of Moment Condition Models*, Journal of the American Statistical Association, **113** (2018), 1656–1668.

[4] S.G. Donald, G.W. Imbens and W. K. Newey, *Empirical likelihood estimation and consistent tests with conditional moment restrictions*, Journal of Econometrics, **117** (2003), 55–93

[5] S.M. Schennach, *Bayesian Exponentially Tilted Empirical Likelihood*, Biometrika, **92** (2005), 31–46.

Continuous time state and parameter estimation from noisy state increments

Sebastian Reich
(joint work with Nikolas Nüsken, Paul Rozdeba)

We consider the problem of inferring the states and parameters of a stochastic differential equation (SDE)

$$\mathrm{d}X_t = f(X_t,\theta)\mathrm{d}t + G\mathrm{d}W_t \tag{1}$$

from a given initial state $X_0 = x_0$ and noisy and partial observations Y_t of the form

$$Y_t = HX_t + R^{1/2}V_t. \tag{2}$$

Here W_t and V_t denote uncorrelated Brownian motion. Such inference problems arise, for example, when a particle with position X_t moves under a stochastic velocity field given by the right hand side of (1) and in case those velocities fields are only partially and noisily observed.

More specifically, since (2) implies

$$\mathrm{d}Y_t = H\mathrm{d}X_t + R^{1/2}\mathrm{d}V_t = Hf(X_t,\theta)\mathrm{d}t + HG\mathrm{d}W_t + R^{1/2}\mathrm{d}V_t, \tag{3}$$

the noise in the observation process is correlated with the model errors in (1). The implications of such correlations on the law π_t of X_t conditioned on the observations $Y_{[0,t]}$ up to time t and a given parameter value θ have been discussed in [1]. In particular, the associated Kushner–Stratonovitch equation for the time evolution of π_t is given in weak form by

$$\mathrm{d}\overline{\phi}_t = \overline{\mathcal{L}\phi}_t\mathrm{d}t + \left\{\overline{\phi h}_t - \overline{\phi}_t\overline{h}_t\right\}^{\mathrm{T}} C^{-1}\left\{\mathrm{d}Y_t - \overline{h}_t\mathrm{d}t\right\}, \tag{4}$$

where $\overline{\phi}_t = \pi_t[\phi]$ denotes the expectation value of an observable ϕ with respect to the conditional density π_t, $C = GG^T + R$,

$$h_t(x) = Hf(x,\theta) - GG^{\mathrm{T}}\nabla_x \log \pi_t(x), \tag{5}$$

and $\mathcal{L}$ denotes the generator of (1).

In my talk I have discussed McKean–Vlasov mean-field equations of the form

$$\mathrm{d}\widetilde{X}_t = f(\widetilde{X}_t,\theta)\mathrm{d}t + G\mathrm{d}W_t + K_t(\widetilde{X}_t)\circ \mathrm{d}I_t + \Omega_t(X_t) \tag{6}$$

such that $\pi_t = \mathrm{law}\,(\widetilde{X}_t)$. Here the matrix-valued K_t is called the gain, I_t is the innovation and Ω_t denotes a Ito correction term. Such mean-field approaches to the standard filtering problem have first been discussed in [3] and in the specific form (6) in [4]. The ensemble Kalman–Bucy filter [2] fits into the framework (6) with K_t independent of $\widetilde{X}_t$ and, more specifically,

$$K_t = (P_t^{xf} + GG^{\mathrm{T}})C^{-1} \tag{7}$$

for the correlated filtering problem (1)–(2) and $\Omega \equiv 0$ [5]. Here P_t^{xf} denote the correlation between X_t and $f(X_t,\theta)$. An appropriate innovation process is given

by

$$I_t = Y_t - \left\{ HX_t + R^{1/2}U_t \right\}, \tag{8}$$

where U_t denotes Brownian motion independent of W_t and V_t. Note that the special case $H = I$ and $R = 0$ leads to $\mathrm{d}\widetilde{X}_t = \mathrm{d}Y_t$ and hence $\widetilde{X}_t = Y_t$ for all $t \geq 0$.

Such mean-field equations are computationally attractive as they can easily be implemented using Monte–Carlo techniques and lead to robust and in many situations sufficiently accurate algorithms. Furthermore, the combined state and parameter estimation problem can be treated by a simple state augmentation $z = (x^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}$.

See [5] for more details on the numerical implementation of (6) and applications to state-parameter estimation problems in the context of averaging, homogenisation and non-parametric drift estimation.

References

[1] A. Bain and D. Crisan, *Fundamental of Stochastic Filtering*, Springer–Verlag, 2009.

[2] K. Bergemann and S. Reich, *An ensemble Kalman–Bucy filter for continuous data assimilation*, Meteorolog. Zeitschrift **21** (2012), 213–219.

[3] D. Crisan and J. Xiong, *Approximate McKean–Vlasov representation for a class of SPDEs*, Stohcastics **82** (2010), 53–68.

[4] T. Yang, P. Mehta and S. Meyn, *Feedback particle filter*, IEEE Trans. Automatic Control **58** (2013), 2465–2480.

[5] N. Nüsken, S. Reich and P. Rozdeba, *State and parameter estimation from observed signal increments*, technical report, SFB 1294 *Data Assimilation*, 2019.

On coupling methods for nonlinear ensemble filtering

Youssef Marzouk

(joint work with Ricardo Baptista, Alessio Spantini)

We consider the Bayesian filtering problem for high dimensional non-Gaussian state-space models with challenging nonlinear dynamics, and sparse observations in space and time. While the ensemble Kalman filter (EnKF) yields robust ensemble approximations of the filtering distribution, it is limited by linear forecast-to-analysis transformations. To generalize the EnKF, we propose a methodology that transforms the non-Gaussian forecast ensemble at each assimilation step into samples from the current filtering distribution via a sequence of local nonlinear couplings. These couplings are based on transport maps that can be computed quickly using convex optimization, and that can be enriched in complexity to reduce the intrinsic bias of the EnKF. We discuss the low-dimensional structure inherited by the transport maps from the filtering problem, including decay of correlations, conditional independence, and local likelihoods. We then exploit this structure to regularize the estimation of the maps in high dimensions and with a limited ensemble size.

References

[1] Alessio Spantini, Ricardo Baptista, Youssef Marzouk. "Coupling techniques for nonlinear ensemble filtering". arXiv:1907.00389 (2019)

Computational approximations in Bayesian inverse problems

Aretha Teckentrup
(joint work with Han Cheng Lie, Tim Sullivan, Andrew Stuart)

We are interested in the inverse problem of estimating unknown parameters in a mathematical model from observed data. We follow the Bayesian approach, in which the solution to the inverse problem is the probability distribution of the unknown parameters conditioned on the observed data, the so-called *posterior* distribution. Specifically, this talk considers the stability of the posterior distribution when an accurate but computationally intractable model is replaced by a *surrogate* (or emulator). A particular focus is on stochastic surrogates, which often arise in practice. An expensive model such as the solution of a PDE may replaced by a kriging/Gaussian process (GP) model [1]. Or, in the realm of "big data", a residual vector of prohibitively high dimension may be randomly subsampled or orthogonally projected onto a randomly-chosen low-dimensional subspace [1, 2].

Mathematically, the inverse problem of interest is to determine the parameters $u \in U$ from the noisy data $y \in \mathbb{R}^{d_y}$ given by $y = \mathcal{G}(u) + \eta$, where U is a separable Banach space. Given the prior measure μ_0 and data y, we are interested in the posterior distribution μ^y on the conditioned random variable $u|y$, which can be characterised by Bayes' Theorem as follows:

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z}\exp\big(-\Phi(u)\big), \qquad Z = \mathbb{E}_{\mu_0}\Big(\exp\big(-\Phi(u)\big)\Big),$$

where the form of the negative *log-likelihood* Φ depends on the distribution of η.

In many practical applications, Φ is computationally too expensive or impossible to evaluate exactly; one therefore often uses an approximation Φ_N of Φ. This leads to an approximation μ_N^y of the exact posterior μ^y, and a key desideratum is convergence, in a suitable sense, of μ_N^y to μ^y as the approximation error $\Phi_N - \Phi$ tends to zero. Convergence in the Hellinger distance can be shown [1, 2], for a broad class of prior distributions and approximations schemes Φ_N, giving bounds on (moments of) the Hellinger distance between μ_N^y and μ^y in terms of $\|\Phi_N - \Phi\|$ (in a suitably chosen norm). To obtain convergence rates in terms of the discretisation parameter N, we need to use the specific form of Φ_N. For approximations by GP models, convergence rates in N follow from results in scattered data approximation [1]. For random projection methods in the data space, convergence rates follow from Monte Carlo estimates [2].

References

[1] Stuart, A. M., Teckentrup, A. L., 2018. *Posterior consistency for Gaussian process approximations of Bayesian posterior distributions*. Mathematics of Computation, 87(310), pp.721-753.

[2] Lie, H.C., Sullivan, T.J., Teckentrup, A.L., 2018. *Random forward models and log-likelihoods in Bayesian inverse problems*. SIAM/ASA Journal on Uncertainty Quantification, 6(4), pp.1600-1629.

Sampling from some high dimensional distributions

Martin Schlather

(joint work with Felix Ballani, Marco Oesting, Claudia Schillings)

A max-stable process Z on $\mathbb{R}^d$ with standard Fréchet margins is characterized by the equality $Z =^d \frac{1}{n}\bigvee_{i=1}^n Z_i$ for any $n \in \mathbb{N}$ and i.i.d. $Z_i \sim Z$. A subclass is given by the mixed moving maxima processes that can be represented by

$$Z(x) = \bigvee_{(u,y,S)\in\Pi} uS(x-y)$$

where Π is a Poisson point process with intensity $u^{-2}du \times dy \times dH$ for a probability measure H on some measurable function space $\mathbb{H}$ of non-negative functions, and $\int_{\mathbb{R}^d}\int_{\mathbb{H}} S(y)\, H(dS)dy = 1$. Since

$$Z(x) =^d \bigvee_{(u,y,S)\in\tilde{\Pi}} S(x-y)/f(y,S)$$

for a Poisson point process $\tilde{\Pi}$ with intensity $u^{-2}du \times f(y,S)dy\, H(dS)$, the question of an optimal f appears with respect to sampling from Z. [2] show that an efficient sampling from Z can be obtained by choosing $f(y,S) = c^{-1}\sup_{x\in K} S(x-y)$ and $c = \int_{\mathbb{R}^d}\int_{\mathbb{H}} \sup_{x\in K} S(x-y)H(dS)dy$.

The construction of a random coin model $\tilde{Z}$ on $\mathbb{R}^d$ is related to that of a mixed moving maximum process. Let Π a Poisson point process on $\mathbb{R}^d \times \mathbb{H}$ with intensity $\lambda dy \times dH$ for $\lambda > 0$ and some probability measure H on a measurable function space $\mathbb{H}$ with

$$\int_{\mathbb{R}^d}\int_{\mathbb{H}} S^2(y)H(dS)dy = 1$$

and random sign, i.e. $H(S \in A) = H(-S \in A)$ for all measurable subsets $A \subset \mathbb{H}$. Then the random field

$$\tilde{Z}(x) = \lambda^{-1/2}\sum_{(y,S)\in\Pi} S(x-y).$$

has expectation zero, variance 1, and correlation function

$$C(x,y) = \int_{\mathbb{R}^d}\int_{\mathbb{H}} S(t-x)S(t-y)H(dS)dt.$$

As $\lambda \to \infty$, $\tilde{Z}$ converges to a Gaussian random field. Assume for simplicity that S is deterministic except for the sign. Let f be any density function on $\mathbb{R}^d$. Then

$$Z^*(x) = \sum_{(y,S)\in\Pi} S(x-y)/\sqrt{f(y)}$$

has the first two moments in common with $\tilde{Z}$, independently of f, but not the higher moments. Following the Berry-Esseen theorem we search for a density f such that the third moments of Z^* become minimal on a fixed compact set K with respect to a given L_p norm. Solutions can be obtained for special settings, e.g. when K is finite. [1] present a specification in case of the Brown-Resnick process.

References

[1] M. Oesting, M. Schlather and C. Schillings, *Sampling Sup-Normalized Spectral Functions for Brown-Resnick Processes*, Submitted (2019).

[2] M. Oesting. M. Schlather and C. Zhou *Exact and fast simulation of max-stable processes on a compact set using the normalized spectral representation*, Bernoulli **24** (2018), 1497–1530.

Random PDEs on moving hypersurface

Ana Djurdjevac

(joint work with L. Church, C. Elliott, R. Kornhuber, T. Ranner)

Our goal is to include and treat the uncertainty in models, which comes from geometry. More precisely, we consider: partial differential equations (PDEs) with random coefficients posed on an (evolving) deterministic hypersurface and a PDE posed on a randomly evolving domain. Our main motivation for considering these types of problems comes from the cell-biology.

In the first part we analyse the advection-diffusion equation with random coefficients that is posed on an evolving hypersurface $\{\Gamma_t\}$

$$\partial^\bullet u - \nabla_\Gamma \cdot (\alpha(\omega)\nabla_\Gamma u) + u\nabla_\Gamma \cdot \mathbf{w} = f.$$

We will consider both cases, uniformly bounded and log-normal distributions of the coefficient α. We will define the solution space and prove the well-posedness using Banach - Nečas - Babuška theorem [1]. Furthermore, we will introduce and analyse the evolving surface finite element discretization of the equation, introduced by Dziuk and Elliott. In the uniformly bounded case, we will show unique solvability of the resulting semi-discrete problem and prove optimal error bounds for the semi-discrete solution and Monte Carlo of its expectation [2]. The plan for the future work is to consider the numerical analysis in the log-normal case.

In the last part we will study the case when the velocity of a hypersurface is the uniformly bounded random field. First we will consider a heat equation posed on a flat domain that evolves by a given random velocity field [4]. To deal with this problem, we will use the domain mapping method which will result in a parabolic equation with random coefficients on a fixed cylindrical domain. Then we will study an elliptic PDE on a random curved domain [3]. For this problem

we will again apply the domain mapping method. This includes computations of geometric quantities of functions given over parametrised hypersurfaces in terms of quantities of the reference surface and derivatives of the domain mapping and corresponding pull-back function. Natural questions that arise are: what could one say about parabolic PDEs posed on a randomly evolving curved domain and how rough can be the random velocity which defines the evolution of the hypersurface.

References

[1] A. Djurdjevac, *Advection-diffusion equations with random coefficients on evolving hypersurfaces*, IFB **19(4)** (2017), 525–552.

[2] A. Djurdjevac, C. Elliott, R. Kornhuber, T. Ranner *Evolving Surface Finite Element Methods for Random Advection-Diffusion Equations*, SIAM/ASA J. Uncertainty Quantification **6(4)** (1990), 656–1684.

[3] L. Church, A. Djurdjevac, C. Elliott, *A domain mapping approach for elliptic equations posed on random bulk and surface domains*, arxiv arXiv:1903.06450, (2019).

[4] A. Djurdjevac, *Random moving domain* , arxiv arXiv:1808.06970, (2018).

Higher-order methods for stochastic differential equations

Lisa Fischer

(joint work with Matthew R. Christian, Sebastian Götschel, Michael L. Minion)

Stochastic differential equations (SDE) are used for modeling real-world problems such as biological or physical processes which are often stochastic in nature. The most common methods to numerically approximate solutions to SDEs are the Euler-Maruyama (EM) and Milstein scheme, which are of low-order [5]. While more advanced methods have been developed, they as well are typically limited with respect to their convergence order [5, 9, 7, 1, 4, 8].

If higher accuracy approximations are required, e.g, to reduce bias in Monte Carlo estimators, this leads to high computational effort. Here, we investigate a Spectral Deferred Correction (SDC) based method to numerically solve SDEs given in Stratonovich form. SDC methods, first introduced in [2], are well known for generating high-order solutions for ODEs. We use the Wong-Zakai approximation [11, 12] and a smooth approximation of Brownian bridges, and apply SDC to the resulting equation. The SDC method is an iterative scheme, that computes corrections to the current solution $x_{[j]}$, i.e., $x_{[j+1]} = x_{[j]} + \delta_{[j]}$ with $\delta_{[j]}(t_0) = 0$. The method is very flexible with respect to time discretization enabling time adaptivity, parallel-in-time algorithms, and multilevel approaches [3, 10]. Figure 1 illustrates the method. Starting from the Picard integral formulation of the initial value problem on a single time step $[t, t + \Delta t]$, M collocation points t_i are chosen, e.g., a Lobatto time grid, as

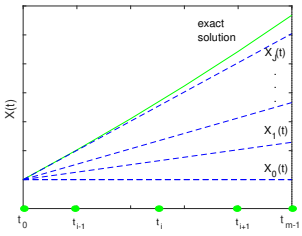

Figure 1. Illustration of the SDC method using a constant initialization, i.e., $x_{[0]}(t) = x(0)$ for all t.

well as an initial solution, e.g., constant initialization. Since the initial value is known, the correction at t_0 is 0 for all iterations j. Given the Stratonovich form $X_t = X(t_0) + \int_{t_0}^{t} a(s, X_s)ds + \int_{t_0}^{t} b(s, X_s) \circ dW_s$, we can derive the correction scheme based on [2] as

$$\begin{aligned}\delta_{[j]}(t_i) = \delta_{[j]}(t_{i-1}) &+ \int_{t_{i-1}}^{t_i} \left[a(s, x_{[j]}(s) + \delta_{[j]}(s)) - a(s, x_{[j]}(s))\right] ds \\ &+ \int_{t_{i-1}}^{t_i} \left[b(s, x_{[j]}(s) + \delta_{[j]}(s)) - b(s, x_{[j]}(s))\right] \circ dW_s \\ &+ \int_{t_{i-1}}^{t_i} a(s, x_{[j-1]}(s))ds + \int_{t_{i-1}}^{t_i} b(s, x_{[j-1]}(s)) \circ dW_s + x_{[j]}(t_{i-1}) - x_{[j]}(t_i). \end{aligned} \tag{1}$$

The first two integrals of equation (1) can be approximated by low-order schemes (explicit, implicit, or semi-implicit). The more interesting parts are the two last integrals where the calculation of the integral of the drift term $a(t, x(t))$ and of the diffusive term $b(t, x(t))$ are required; for this we use spectral quadrature. For the drift term the approximation is well known, $\int_{t_{i-1}}^{t_i} a(s, x_{[j]}(s))ds \approx \sum_{k=0}^{M-1} S_{i,k}\ a(t_k, x_{[j]}(t_k)), i = 1, \ldots, M$, with S being the spectral integration matrix involving Lagrange polynomials. To apply spectral quadrature to the diffusive term as well, we replace the Wiener process by a smooth approximation. More specifically, we smoothly approximate the Brownian bridge between samples from the Wiener process at t and $t + \Delta t$ by its truncated Fourier series,

$$B_K(t) = \frac{t\eta_0}{\sqrt{(\Delta T)}} + \frac{\sqrt{2\Delta T}}{\pi} \sum_{k=1}^{K} \frac{\sin(k\pi t/\Delta T)}{k} \eta_k,$$

$\eta_k \sim \mathcal{N}(0, 1)$ i.i.d., for all $k = 0, \ldots, K$. This transforms the SDE to the ODE

$$\frac{dx(t)}{dt} = a(t, x(t)) + b(t, x(t)) \frac{dB_K(t)}{dt}.$$

Note that while the Brownian bridge approximation contains random variables, they are not time-dependent, such that in principle the usual spectral quadrature can be used, analogously to the drift term. Small time steps, or using a large truncation index K, leads to highly oscillatory terms in the expansion and requires specialized quadrature.

To illustrate convergence of the method consider the following simple linear scalar example,

$$dX_t = X_t dW_t, \tag{2}$$

with initial value $X_0 = 1$ and exact solution $X_t = X_0 \exp(W_t)$. Note that as the drift term is a multiple of the diffusive term, the B_0 term of the Brownian bridge approximation is sufficient. One can show that every two sweeps the convergence

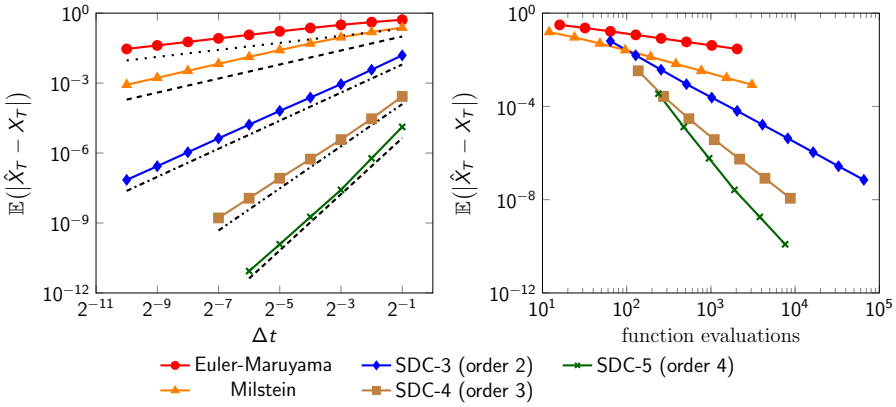

FIGURE 2. Strong convergence error (left) and computational effort (right) of SDC-n (n Lobatto nodes), EM, and Milstein.

order increases by 1. When using the linear term of the Brownian bridge approximation only, the convergence order is limited by half the order of the underlying quadrature rule. Figure 2 (left) shows the strong convergence order of the EM, Milstein and SDC-n, $n = 3, 4, 5$ collocation points, methods, whereas the dashed and dotted lines represent the theoretical convergence order. For reaching the same accuracy (below 10^{-2}), the computational effort shown in figure 2 (right) of SDC is smaller than the cost of EM and Milstein.

In the future, we will extend the presented work to real-world problems such as problems with non-commutative noise, and jump-diffusion SDEs. Looking at randomized Bayesian inverse problems, efficient solvers are required if accurate estimations of the system's parameters are important. Here, the strong convergence order of the numerical method translates into a convergence order for the posterior estimate [6]. In view of multilevel Monte-Carlo sampling, multilevel SDC approaches and parallel-in-time algoithms can be beneficial.

REFERENCES

[1] E. Buckwar, A. Rößler, R. Winkler, *Stochastic Runge–Kutta methods for Itô SODEs with small noise*, SIAM Journal on Scientific Computing **32,4** (2010), 1789–1808.

[2] A. Dutt, L. Greengard, V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT Numerical Mathematics **40,2** (2000), 241–266.

[3] M. Emmett, M.L. Minion, *Toward an efficient parallel in time method for partial differential equations*, Communications in Applied Mathematics and Computational Science **7,1** (2012), 105–132.

[4] M. Kamrani, N. Jamshidi, *Implicit Milstein method for stochastic differential equations via the Wong-Zakai approximation*, Numerical Algorithms **79,2** (2018), 357–374.

[5] P.E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag (1992).

[6] H.C. Lie, T.J. Sullivan, A.L. Teckentrup, *Random forward models and log-likelihoods in Bayesian inverse problems*, SIAM/ASA Journal on Uncertainty Quantification **6,4** (2018), 1600–1629.
[7] G.N. Milstein, *Numerical integration of stochastic differential equations*, Springer Science & Business Media **313** (1994).
[8] C. Rackauckas, Q. Nie*Stability-Optimized High Order Methods and Stiffness Detection for Pathwise Stiff Stochastic Differential Equations*, arXiv preprint arXiv:1804.04344 (2018).
[9] A. Rößler, *Runge–Kutta methods for Itô stochastic differential equations with scalar noise*, BIT Numerical Mathematics **46, 1** (2006), 97–110.
[10] R. Speck et al., *A multi-level spectral deferred correction method*, BIT Numerical Mathematics **55,3** (2015), 843–867.
[11] H. J. Sussmann, *On the Gap Between Deterministic and Stochastic Ordinary Differential Equations*, The Annals of Probability **1** (1978), 19–41.
[12] E. Wong, M. Zakai, *On the Convergence of Ordinary Integrals to Stochastic Integrals*, The Annals of Mathematical Statistics **5** (1965), 1560–1564.

On the well-posedness of Bayesian inverse problems: The Gaussian noise case

Jonas Latz

In this work, we introduce a weaker concept of well-posedness for Bayesian inverse problems (BIPs) and show that a popular class of BIPs is well-posed according to this concept. In particular, we are given an inverse problem

$$\text{Find } \theta^\dagger \in X : y = \mathcal{G}(\theta^\dagger) + \eta$$

where y is observed data and $\eta \sim \mu_{\text{noise}} := \mathrm{N}(0, \Gamma)$ is measurement noise that is modelled as a non-degenerate Gaussian random variable. Data and noise take values in the finite dimensional data space $Y := \mathbb{R}^k$, equipped with a norm $\|\cdot\|_Y$. The parameter space X is a Radon space.

We apply a Bayesian approach. Hence, we assume that $\theta \sim \mu_0$ is uncertain and modelled as a random variable. Then, we use Bayes' formula to obtain the posterior measure

$$\mu^y := \mathbb{P}(\theta \in \cdot | \mathcal{G}(\theta) + \eta = y),$$

by conditioning the random variable θ with respect to the event of observing y. Numerical computability of the posterior measure has been quantified by [3] in terms of - what we will refer to by - *Lipschitz well-posedness*. A Bayesian inverse problem is Lipschitz well-posed, if the posterior measure exists, if it is unique, and if the map

$$f : (Y, \|\cdot\|_Y) \to (\mathrm{Prob}(X, \mu_0), d_{\text{Hel}}), y \mapsto \mu^y$$

is locally Lipschitz continuous. Here, $(\mathrm{Prob}(X, \mu_0), d_{\text{Hel}})$ is the metric space of probability measures μ on $(X, \mathcal{B}X)$, where $\mu \ll \mu_0$, and the metric is the Hellinger distance. This concept is fairly strong, considering the fact that the continuity condition is only a stability criterion: One may find contradictory that a continuous, non-Lipschitz transformation of the data can turn a well-posed problem into an ill-posed problem. Moreover, there are practical non-Lipschitz well-posed Bayesian inverse problems. Such are the Bayesian elliptic inverse problem with

certain geometric prior measures, see [1]. In both cases, one can show continuity in the posterior measure under perturbations in the data. This motivates us to consider a weaker concept of well-posedness.

A Bayesian inverse problem is *well-posed*, if the posterior measure μ^y exists, if it is unique, and if the aforementioned map f is continuous. We give easily verifiable assumptions in [2] that imply well-posedness. Indeed, the assumptions are fulfilled by any Bayesian inverse problem described in the beginning of this abstract: the data space is a finite dimensional Banach space, the measurement noise is additive, non-degenerate Gaussian and the forward map $\mathcal{G} : X \to Y$ is measurable.

This has several practical implications: A thorough analysis of the forward operator is not necessary to make well-posedness statements. Moreover, the well-posedness statement is independent of the prior probability measure. Hence, complicated hierarchical priors (such as Deep Gaussian processes), as well as discrete priors (in discretisations and mixed-integer inverse problems), and combinations of these (e.g. transdimensional estimations) are covered.

References

[1] M.A. Iglesias, K. Lin, and A.M. Stuart, *Well-posed Bayesian geometric inverse problems arising insubsurface flow*, Inverse Probl., **30(11)** (2014), 114001.
[2] J. Latz, *On the well-posedness of Bayesian inverse problems*, arXiv: **1902.10257** (2019).
[3] A.M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica **19** (2010), 451–559.

Quasi-stationary Monte Carlo

Andi Q. Wang
(joint work with Martin Kolb, Divakar Kumar, Murray Pollock, Gareth Roberts, David Steinsaltz)

In recent years there has been significant interest in developing Monte Carlo methods based on *continuous-time* processes. Examples include diffusions (MALA and ULA, [5]), ODE dynamics (HMC, [3]), inhomogeneous Poisson processes, (BPS [2] and Zig-Zag [1]). This talk concerns quasi-stationary Monte Carlo (QSMC) methods, which draw on both diffusion processes and inhomogeneous Poisson processes.

Recall that for a Markov process X killed at a time τ_∂, a probability measure μ is *quasi-stationary* if for all $t \geq 0$,

$$\mathbb{P}_\mu(X_t \in \cdot \,|\tau_\partial > t) = \mu(\cdot).$$

The idea behind QSMC is to construct a killed Markov process whose quasi-stationary distribution coincides with a target distribution π of interest. In [4] it was shown that such methods can make use of subsampling techniques without error, leading to unbiased algorithms for posterior sampling which scale well as the number of observations in the Bayesian experiment grows.

In this talk I will discuss theoretical advances in QSMC. I will give sufficient conditions for the quasi-stationary distribution of a killed diffusion to coincide with a given target density π. I will also describe a stochastic approximation approach to the simulation of quasi-stationary distributions of killed diffusions. The details can be found in [6] and [7], respectively.

References

[1] J. Bierkens, P. Fearnhead, G. O. Roberts, *The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data*, The Annals of Statistics, **47**(3) (2019), 1288–1320.

[2] A. Bouchard-Côté, S. J. Vollmer, A. Doucet, *The Bouncy Particle Sampler: A Nonreversible Rejection-Free Markov Chain Monte Carlo Method*, Journal of the American Statistical Association, **113**(522) (2018), 855–867.

[3] R. M. Neal, *MCMC using Hamiltonian dynamics*. In S. Brooks, A. Gelman, G. Jones, X.-L. Meng, (eds.), Handbook of Markov Chain Monte Carlo. Chapman & Hall, CRC Press, (2010).

[4] M. Pollock, P. Fearnhead, A. M. Johansen, G. O. Roberts, *The Scalable Langevin Exact Algorithm: Bayesian inference for big data*. arXiv Preprint, arXiv 1609.0343, (2016).

[5] G. O. Roberts, J. S. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **60**(1) (1998), 255–268.

[6] A. Q. Wang, M. Kolb, G. O. Roberts, D. Steinsaltz, *Theoretical properties of quasi-stationary Monte Carlo methods*, The Annals of Applied Probability, **29**(1) (2019), 434–457.

[7] A. Q. Wang, G. O. Roberts, D. Steinsaltz, . An Approximation Scheme for Quasistationary Distributions of Killed Diffusions, arXiv preprint, arXiv 1808.07086, (2018).

Analysis of the ensemble Kalman inversion

Simon Weissmann

(joint work with Dirk Bloemker, Claudia Schillings, Philipp Wacker)

The goal in the inverse problem setting is to recover an unknown $u \in \mathcal{X}$ from (typically a finite number of) observations $y \in \mathbb{R}^k$ where

$$y = \mathcal{G}(u) + \eta.$$

Here $\mathcal{X}$ is a separable Hilbert space, $\mathcal{G} : \mathcal{X} \to \mathbb{R}^k$ is a nonlinear, continuous operator and $\eta \sim \mathcal{N}(0, \Gamma^{-1})$ denotes the observational noise. In Bayesian regularization (u, y) is viewed as a jointly varying random variable and, assuming that η is independent of u according to μ_0, the solution to the inverse problem is the random variable $u|y$ characterized by

$$\mu(du) \propto \exp(-\frac{1}{2}\|\Gamma^{-\frac{1}{2}}(y - \mathcal{G}(u))\|^2_{\mathbb{R}^k})\, \mu_0(du).$$

In many cases it is not possible to compute the posterior distribution in a closed form or even simulate straightforwardly from this distribution.

The ensemble Kalman filter (EnKF) is a widely used metheodology for data assimilation problems and has also been generalized to inverse problems. We are interested in studying properties of ensemble Kalman inversion (EKI). We view the method as a derivative free optimization method for the least-squares misfit

functional. This opens up the perspective to use the method in various areas of applications such as imaging, groundwater flow problems, biological problems and machine learning.

Based on recent results for the analysis of the EKI for linear forward problems [2], we focus on the generalization of the ideas to the case of perturbed observations, i.e.

$$y^{(j)} = y + \eta^{(j)}.$$

The analysis is based on the continuous time limit of the algorithm, which leads to the following system of stochastic differential equations (SDEs). Firstly let $\mathcal{G}(\cdot) = A\cdot$, for some $A \in \mathcal{L}(\mathcal{X}, \mathbb{R}^k)$, be a bounded linear operator, for $u = \{u^{(j)}\}_{j=1}^J$, $u^{(j)} \in \mathcal{X}$ where J is the number of ensemble members. Then we define the empirical covariance operator

$$C(u) = \frac{1}{J}\sum_{k=1}^{J}(u^{(k)} - \bar{u}) \otimes (u^{(k)} - \bar{u}),$$

and consider the system of SDEs

$$du^{(j)} = C(u)A^*\Gamma^{-1}(y - Au^{(j)})\, dt + C(u)A^*\Gamma^{-\frac{1}{2}}\, dW^{(j)},$$

where $\{W^{(j)}\}_{j=1}^J$ are independent Brownian motions in $\mathbb{R}^k$.

We have shown well-posedness of the scheme, i.e. existence and uniqueness of strong solutions of the limiting system of SDEs. Furthermore, we have quantified the ensemble collapse in the observation space $\mathbb{R}^k$ as well as in the parameter space $\mathcal{X}$ in the sense of moments and almost sure convergence with **a** given rate.

In **the** case of non perturbed observations, by using variance inflation we were able to prove convergence to the truth in the observation space**,** in the sense of L^2-convergence and almost sure convergence with a given rate.

There remain open and interesting questions related to this work:

- Verification of the continuous time limit of the ensemble Kalman filter.
- Extension to nonlinear forward operators**.**
- Study of the empirical covariance of the particles, in particular the quantification of the convergence rate of the smallest eigenvalues.

This will be subject to future work.

References

[1] D. Blomker, C.Schillings, P. Wacker and S. Weissmann, *Well posedness and convergence analysis of the ensemble Kalman inversion*, ArXiv e-prints (submitted to: Inverse Problems), 2018.

[2] C. Schillings and A. M. Stuart, *Analysis of the ensemble Kalman filter for inverse problems.*, SINUM, 55, pp. 1264-1290, 2017.

[3] M. A. Iglesias, K. J. H. Law and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 2013.

[4] D. Blomker, C. Schillings and P. Wacker, *A strongly convergent numerical scheme from EnKF continuum analysis*, SIAM Journal on Numerical Analysis, 2018.

[5] A. M. Stuart, *Inverse problems: A bayesian perspective*, Acta Numerica, Cambridge University Press, 2010.
[6] X. Mao, *Stochastic differential equations and applications*, Horwood Pub., 2008.
[7] R. Khasminskii, *Stochastic stability of differential equations*, Springer, Stochastic Modelling and Applied Probability; 66, 2012.

Accuracy and stability of EnKBF for finite ensemble sizes in a non-linear setting

Jana de Wiljes
(joint work with Sebastian Reich, Wilhelm Stannat, Xin Tong)

In the context of nonlinear high dimensional filtering problems ensemble based techniques such as the Ensemble Kalman Filter are still consider state of the art despite the lack of mathematical foundation in this setting. Here the continuous-time filtering problem is considered, i.e, the aim is to approximate the conditional distribution in X_t given by

$$\mathrm{d}X_t = f(X_t)\mathrm{d}t + \sqrt{2}C\mathrm{d}W_t \tag{1}$$

for given observations Y_s, $s \in [0,t]$ that evolve according to

$$\mathrm{d}Y_t = h(X_t)\mathrm{d}t + R^{1/2}\mathrm{d}B_t. \tag{2}$$

where $f : \mathbb{R}^{N_x} \to \mathbb{R}^{N_x}$ is a Lipschitz-continuous drift, $C \in \mathbb{R}^{N_x \times N_w}$, $R \in \mathbb{R}^{N_y \times N_y}$ and $W_t \in \mathbb{R}^{N_w}$ and $B_t \in \mathbb{R}^{N_y}$ denote independent Brownian motion. For the following deterministic variant of the Ensemble Kalman Bucy Filter

$$\mathrm{d}X_t^i = f(X_t^i)\mathrm{d}t + D(P_t^M)^{-1}(X_t^i - \bar{x}_t^M)\mathrm{d}t - \frac{1}{2}Q_t^M R^{-1}\left(h(X_t^i)\mathrm{d}t + \bar{h}_t^M\mathrm{d}t - 2\mathrm{d}Y_t\right) \tag{3}$$

lower and upper bounds of the covariance matrix P_t^M are derived for the fully-observed case with small measurement noise ϵ and $M > N_x$. These stability results are then used to obtain point and path-wise accuracy results in terms of ability to track the true solution:

$$\mathbb{E}\left[\sup_{t\le T} E_t\right] \le C\varepsilon^{\frac{1}{2}-\eta} \tag{4}$$

with $\eta \in \left(0, \frac{1}{4}\right)$. These results are extended to the case where the ensemble size is larger than the state space, i.e., $N_x > M$ for a localised deterministic Ensemble Kalman Bucy Filter

$$dX_t^i = f(X_t^i)dt + C(P_t^L)^{\dagger}(X_t^i - \bar{x}_t) - \frac{1}{2}P^L(t)R^{-1}(HX^i(t)dt + H\bar{x}(t)dt - 2dY(t)).$$

where

$$P_t^L = \frac{1}{M-1}\sum (X_t^i - \bar{x}_t)(X_t^i - \bar{x}_t)^T \circ D_L \tag{5}$$

with D_L being an appropriate localization matrix.

References

[1] J. de Wiljes, S. Reich, W. Stannat *Long-Time Stability and Accuracy of the Ensemble Kalman–Bucy Filter for Fully Observed Processes and Small Measurement Noise*, SIAM J. APPLIED DYNAMICAL SYSTEMS **17** (2018), 1152–1181.

[2] J. Wiljes and X. Tong, *Analysis of a localised Ensemble Kalman Bucy Filter for a nonlinear setting*, in preparation.

Numerical Shadowing for Data Assimilation

Svetlana Dubinkina

(joint work with Bart de Leeuw)

A data assimilation method combines a solution of a physical system model with measurement data to obtain an improved estimate for the state of a physical system. We study problems for which the system model is discrete in both space and time, and may be contaminated by model errors

$$x_{n+1} = F_n(x_n) + Q_n, \quad x_n \in \mathbb{R}^m, \quad n = 0, \dots, N-1, \tag{1}$$

where $F_n : \mathbb{R}^m \to \mathbb{R}^m$ and Q_n is some unknown model error drawn from a Gaussian distribution with zero mean and covariance matrix C_{m}. We assume F_n to be $\mathcal{C}^3$ for all n. Let the sequence $\boldsymbol{X} := \{\mathcal{X}_0, \dots \mathcal{X}_N\}$ be a distinguished orbit of (1), referred to as the true solution of the model, and presumed to be unknown. Suppose we are given a sequence of noisy observations $\boldsymbol{y} := \{y_0, \dots y_N\}$ related to $\boldsymbol{X}$ via

$$y_n = H(\mathcal{X}_n) + \xi_n, \qquad y_n \in \mathbb{R}^d, \quad n = 0, \dots, N, \tag{2}$$

where $H : \mathbb{R}^m \to \mathbb{R}^d$, $d \leq m$, is the observation operator, and the noise variables ξ_n are drawn from a normal distribution with zero mean and known observational error covariance matrix C_{o}. The goal of data assimilation is to find $\boldsymbol{u} = \{u_0, u_1, \dots, u_N\}$, $u_n \in \mathbb{R}^m$, such that the differences $\|y_n - H(u_n)\|$ and $\|u_{n+1} - F_n(u_n)\|$, $n = 0, \dots, N$ are small in an appropriately defined sense.

The data assimilation problem may be solved using variational data assimilation approach. Variational data assimilation methods are based on minimization of a cost function. In [1], we proposed a strong constraint data assimilation method that, instead of minimizing a cost function, searches for a zero of the cost operator $G(\boldsymbol{u})$ defined as

$$G(\boldsymbol{u}) = \begin{pmatrix} G_0(\boldsymbol{u}) \\ G_1(\boldsymbol{u}) \\ \vdots \\ G_{N-1}(\boldsymbol{u}) \end{pmatrix}, \qquad G_n(\boldsymbol{u}) = u_{n+1} - F_n(u_n), \quad n = 0, \dots, N-1. \tag{3}$$

Instead of solving directly for the initial condition as in variational data assimilation, we solve for the whole orbit at once. This approach is motivated by research on numerical shadowing methods.

Suppose $\boldsymbol{u}$ is an ε-orbit in a neighborhood of a hyperbolic set for F, namely $\|G_n(\boldsymbol{u})\| < \varepsilon$, $n = 0, \dots, N-1$, where $\|\cdot\|$ is a norm in $\mathbb{R}^m$. The shadowing lemma

(e.g. Theorem 18.1.2 of [2]) states that, for every $\delta > 0$ there exists $\varepsilon > 0$ such that $\boldsymbol{u}$ is δ-shadowed by an orbit of F, i.e. there exists an orbit $\boldsymbol{x}$ satisfying $G(\boldsymbol{x}) = 0$ such that $\|u_n - x_n\| < \delta$ for all $n = 0, \ldots, N$. For example, let F be the exact time-τ flow map of an autonomous ODE $\dot{x} = f(x)$. If the components of $\boldsymbol{u}$ are the iterates of a numerical integrator with local truncation error bounded by ε, then these define an ε-orbit of F. Shadowing refinement [3] employs the pseudo-orbit as an initial guess for $G(\boldsymbol{u}) = 0$ and, as opposed to proving the existence of a nearby zero of G, iteratively refines the pseudo-orbit to obtain an improved approximation of a true solution. The inverse problem to shadowing is to determine an optimal initial condition u_0 for a numerical integration, such that the numerical iterates $\boldsymbol{u}$ δ-shadow a desired orbit of $\dot{x} = f(x)$.

Shadowing theory has already motivated a practical data assimilation algorithm known as pseudo-orbit data assimilation (PDA) [4], which solves a problem of minimizing $\frac{1}{2}\|G(\boldsymbol{u})\|_I^2$. In order to stay close to observations, PDA is initialized at observations and the minimization is approximately solved using a fixed number of gradient descent steps to satisfy a stopping criterion.

In [6], we formulate the shadowing-based method of [1] in a weak constraint form. Initializing at observations, we take model error into account following the Levenberg-Marquardt regularization approach and impose a stoping criterion based on data mismatch. Denoting by k the index of the Newton's iteration, we have at $k = 0$ $\boldsymbol{u} = \boldsymbol{y}$ and we seek an update $\boldsymbol{\delta}^{(k)}$ by approximately solving

$$G(\boldsymbol{u}^{(k)} + \boldsymbol{\delta}^{(k)}) = 0. \tag{4}$$

We then update using $\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \boldsymbol{\delta}^{(k)}$, where

$$\boldsymbol{\delta}^{(k)} = -C_{\mathrm{o}} G'^{T}(\boldsymbol{u}^{(k)}) \left(G'(\boldsymbol{u}^{(k)}) C_{\mathrm{o}} G'^{T}(\boldsymbol{u}^{(k)}) + \alpha^{(k)} C_{\mathrm{m}} \right)^{-1} G(\boldsymbol{u}^{(k)}), \tag{5}$$

for $\alpha^{(k)} > 0$. Under some regularity conditions and algorithms for choosing $\alpha^{(k)}$, convergence to a model orbit can be proven as $k \to \infty$ [5]. We remark that if $C_{\mathrm{o}} = I$ and we choose $\alpha^{(k)} = 0$, for all k, then (5) reduces to the strong contstraint shadowing of [1]. If $C_{\mathrm{m}} = C_{\mathrm{o}} = I$ and we choose $\alpha^{(k)} \to \infty$, for all k, then (5) reduces to the gradient descent algorithm of PDA. The Jacobian of G has a $m(N-1) \times mN$ block structure:

$$G'(\boldsymbol{u}) = \begin{bmatrix} -F_0'(u_0) & I & & & \\ & -F_1'(u_1) & I & & \\ & & \ddots & \ddots & \\ & & & -F_{N-1}'(u_{N-1}) & I \end{bmatrix}.$$

The solution $\boldsymbol{\delta}^{(k)}$ to (5) is the minimizer of

$$\frac{1}{2}\|G'\left(\boldsymbol{u}^{(k)}\right)\boldsymbol{\delta}^{(k)} + G\left(\boldsymbol{u}^{(k)}\right)\|_{C_{\mathrm{m}}}^2 + \frac{\alpha^{(k)}}{2}\|\boldsymbol{\delta}^{(k)}\|_{C_{\mathrm{o}}}^2, \tag{6}$$

where $\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \boldsymbol{\delta}^{(k)}$. At the first iteration the weak constraint shadowing is identical to the weak constraint 4DVar, when initialized at the full observations.

As the iteration proceeds it, however, becomes distinct since (6) does not stay fixed throughout the iteration.

The regularization parameter $\alpha^{(k)}$ can be determined uniquely by imposing that for some $0 < \rho < 1$ $\alpha^{(k)}$ is the smallest non-negative scalar satisfying

$$\rho^{-1}\|\delta\left(\alpha^{(k)}\right)\|_{C_\mathrm{o}} \leq \sqrt{Nd} - \|H(\boldsymbol{u}^{(k)}) - \boldsymbol{y}\|_{C_\mathrm{o}}, \tag{7}$$

where we made the dependency of the update step δ on the parameter $\alpha^{(k)}$ explicit.

For the stopping criterion we require that the distance between analysis and observations remains bounded. Denoting the principal square root of the observational precision by $C_\mathrm{o}^{-\frac{1}{2}}$, $C_\mathrm{o}^{-\frac{1}{2}}(H(\boldsymbol{X}) - \boldsymbol{y})$ is distributed according to a standard normal distribution. In particular, $\mathbb{E}(\|H(\boldsymbol{X}) - \boldsymbol{y}\|^2_{C_\mathrm{o}})/Nd = 1$ and when the number of observations is large enough we may assume $\|H(\boldsymbol{X}) - \boldsymbol{y}\|^2_{C_\mathrm{o}})/Nd \approx 1$ with high probability. Thus we stop the algorithm at the minimum k for which $\|H(\boldsymbol{u}^{(k)}) - \boldsymbol{y}\|^2_{C_\mathrm{o}})/Nd > r$ for a predefined parameter r close to 1.

We show that for fully observed linear systems, the solution is unbiased with respect to the truth and the variance is consistent with least-square minimization. We extend the weak constraint shadowing to ensemble approximation and show for fully observed linear systems that the ensemble mean is unbiased with respect to the truth. For nonlinear systems, well-possedness of the method in algorithmic time limit is guaranteed. In future, we plan to study error bounds of the shadowing method for nonlinear partially observed dynamical systems.

References

[1] B. de Leeuw, S. Dubinkina, J. Frank, A. Steyer, X. Tu and E. Van Vleck, *Projected Shadowing-based Data Assimilation*, SIAM J. Appl. Dyn. Sys. **17** (2018), 2446–2477.

[2] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, **54** (1995).

[3] C. Grebogi, S.M. Hammel, J.A. Yorke and T. Sauer, *Shadowing of physical trajectories in chaotic dynamics: Containment and refinement*, Phys. Rev. Lett., **13** (1990),1527–1530.

[4] H. Du and L.A. Smith, *Pseudo-orbit data assimilation. Part II: Assimilation with imperfect models*, Journal of the Atmospheric Sciences, **71** (2014), 483–495.

[5] M. Hanke, *A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse problems ,**13** (1997), 1–18.

[6] B. de Leeuw and S. Dubinkina, *Regularized shadowing-based data assimilation method for imperfect models and its comparison to the weak constraint 4DVar method*, submitted (2018), https://arxiv.org/abs/1810.07064

A Primal-Dual Algorithm for Large-Scale Risk Minimization

Drew P. Kouri

(joint work with Thomas M. Surowiec)

Many practical applications require the optimization of systems (e.g., partial differential equations) with uncertain inputs such as noisy problem data, unknown operating conditions, and unverifiable modeling assumptions. In this work, we formulate such problems as risk-averse stochastic optimization problems for which

we minimize a measure of risk associated with the system outputs. For many popular risk models, the resulting risk-averse objective function is not differentiable, which significantly complicates the numerical solution of the optimization problem. Methods for nonsmooth optimization often exhibit slow convergence rates (i.e., (sub)linear) and therefore are often intractable for problems in which the objective function and its derivative are expensive to evaluate. To address this challenge, we introduce a general primal-dual algorithm for solving large-scale nonsmooth risk-averse optimization problems. At each iteration of the algorithm, we approximately solve a smooth optimization problem using, e.g., a rapidly-converging derivative-based optimization method.

Let Z be a reflexive Banach space and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where Ω denotes the set of outcomes, $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra of events and $\mathbb{P} : \mathcal{F} \to [0,1]$ is a probability measure. We denote the expectation of a random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ by $\mathbb{E}[X]$. We consider optimization problems of the form

$$\min_{z \in Z_{\rm ad}} \ \mathcal{R}(F(z)) + \wp(z) \tag{1}$$

where $Z_{\rm ad} \subseteq Z$ is a nonempty, closed and convex set of admissible optimization variables, $F : Z \to L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a random loss function, $\wp : Z \to \mathbb{R}$ is a deterministic loss functional, and $\mathcal{R} : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is a risk functional. We make the following basic assumptions on the risk functional: $\mathcal{R}$ is convex, positively homogeneous and satisfies the monotonicity condition

$$\forall\, X,\, X' \in L^2(\Omega, \mathcal{F}, \mathbb{P}) \quad \text{with} \quad X \le X' \ \text{a.s.} \quad \Longrightarrow \quad \mathcal{R}(X) \le \mathcal{R}(X').$$

Under these assumptions, the Fenchel-Moreau Theorem [2] ensures that

$$\mathcal{R}(X) = \sup_{\theta \in \mathfrak{A}} \mathbb{E}[\theta X] \quad \text{where} \quad \mathfrak{A} \subseteq \{\, \theta \in L^2(\Omega, \mathcal{F}, \mathbb{P}) \,|\, \theta \ge 0 \ \text{a.s.}\,\}. \tag{2}$$

Substituting (2) into the optimization problem (1) results in the min-max problem

$$\min_{z \in Z_{\rm ad}} \ \sup_{\theta \in \mathfrak{A}} \{\ell(z,\theta) := \mathbb{E}[\theta F(z)] + \wp(z)\}. \tag{3}$$

The functional in (3) resembles the Lagrangian functional from nonlinear programming. With this as motivation, we define the generalized augmented Lagrangian functional as

$$L(z,\lambda,r) := \max_{\theta \in \mathfrak{A}} \left\{ \ell(z,\theta) - \frac{1}{2r}\mathbb{E}[(\lambda - \theta)^2] \right\} \tag{4}$$

for $z \in Z$, $\lambda \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $r > 0$. Applying results from variational analysis, one can show that $L(z,\lambda,r)$ is continuously Fréchet differentiable with respect to z and λ with partial derivatives given by

$$\nabla_z L(z,\lambda,r) = \mathbf{P}_{\mathfrak{A}}(rF(z) + \lambda) \quad \text{and} \quad \nabla_\lambda L(z,\lambda,r) = (\mathbf{P}_{\mathfrak{A}}(rF(z) + \lambda) - \lambda)/r$$

where $\mathbf{P}_{\mathfrak{A}}$ denotes the projection onto the convex set $\mathfrak{A}$ [1, 3]. Based on these properties of L, we have the following generalization of the classical method of multipliers.

Algorithm 1 Primal-Dual Risk Minimization

Initialize: Given $z_0 \in Z_{\text{ad}}$, $r_0 > 0$ and $\lambda_0 \in \mathfrak{A}$.
While("Not Converged")

(1) Compute $z_{k+1} \in Z_{\text{ad}}$ that approximately minimizes $L(\cdot, \lambda_k, r_k)$.
(2) Set $\lambda_{k+1} = \mathbf{P}_{\mathfrak{A}}(r_k F(z_{k+1}) + \lambda_k)$.
(3) Update r_{k+1}.

End While

When the iterates x_{k+1} of Algorithm 1 are ϵ_k-minimizers of $L(\cdot, \lambda_k, r_k)$, i.e.,

$$L(x_{k+1}, \lambda_k, r_k) - \inf_{x \in X_{\text{ad}}} L(x, \lambda_k, r_k) \leq \epsilon_k,$$

then we can show that any weak accumulation point of $\{x_k\}$ is an ϵ-minimizer of (1) [4, Th. 1]. Under additional conditions on the sequence $\{\epsilon_k\}$, we can further show that the entire sequence $\{\lambda_k\}$ converges to a maximizer of the dual problem

$$\max_{\theta \in \mathfrak{A}} v(\theta) \quad \text{where} \quad v(\theta) := \inf_{z \in Z_{\text{ad}}} \ell(z, \theta)$$

[4, Th. 2]. This result follows from the relationship of Algorithm 1 with the proximal point method [5]. These results typically only apply to convex problems for which we can ensure that x_{k+1} is in fact an ϵ_k-minimizer. For general nonconvex problems, we often can only ensure that the iterates x_{k+1} are ϵ_k-stationary points of $L(\cdot, \lambda_k, r_k)$. That is, if $\wp$ and F are continuously Fréchet differentiable, then

$$\langle \wp'(z_{k+1}) + \mathbb{E}[\lambda_{k+1} F'(z_{k+1})], z - z_{k+1} \rangle_{Z^*, Z} \geq -\epsilon_k \|z - z_{k+1}\|_Z \quad \forall\, z \in Z_{\text{ad}}.$$

Under additional assumptions on the continuity of $\wp'$ and F', we can prove that if x_{k+1} are ϵ_k-stationary points and $\epsilon_k \to 0$, then any weak accumulation point of $\{x_k\}$ is a stationary point of (1) [4, Th. 3].

References

[1] H. Attouch, *Viscosity solutions of minimization problems*, SIAM Journal on Optimization, 6 (1996), pp. 769–806.
[2] H. Attouch, G. Buttazzo, and G. Michaille, *Variational analysis in Sobolev and BV spaces*, vol. 6 of MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
[3] D. P. Kouri and T. M. Surowiec, *Epi-regularization for risk measuares*, submitted, 2018.
[4] ———, *A primal-dual algorithm for risk minimization*, submitted, 2018.
[5] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

Acknowledgements: Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper

do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Uncertainty quantification: dynamic optimization with risk

Alois Pichler
(joint work with Alexander Shapiro, Ruben Schlotter)

The talk addresses stochastic optimization. The following points cause particular complications in stochastic optimization:

(1) the decision variables are functions;
(2) in real applications, the random process is often not known in detail;
(3) the curse of dimensionality;
(4) the summary functions can be more complicated than usual expectation;
(5) continuous time processes.

An initial emphasize is given to point 2 above, i.e., to the fact that the process is not known in detail in usual real applications.

We elaborate that the Wasserstein distance is of crucial importance for these types of problems, as its convex dual, the Kantorovich–Rubinstein theorem, precisely describes the aberration of evaluations of expectations with respect to varying probability measures.

The subsequent step generalizes the situation to non-Markovian optimization problems in multiple stages and what is called the *process* or *nested distance* is introduced (cf. [1]). This distance allows to assign a distance to stochastic processes quantitatively. It is demonstrated that optimization problems are continuous, given reasonable regularity conditions, with respect to varying unerlying stochastic processes. This observation paves the way for replacing complicated stochastic processes by simpler processes, which are eligible for numerical computations.

It turns out, however, that the empirical measure corresponding to a stochastic process has a finite process distance, compared with the initial distance. Even for an increasing number of samples, their distance does not tend to zero. Smoothing techniques, as well-known from kernel density estimation, can be used to overcome this difficulty. It is demonstrated that the smoothed empirical process indeed converges to the genuine process and an explicit quantitative result is presented; this result follows the pattern of large deviation results. However, the result as well reveals the curse of dimensionality.

A second part of the talk involves risk measures (cf. [3]). They replace the expectation in the objective of the corresponding stochastic problem. Again, continuity results are given which allow exchanging the probability measure, i.e., replacing the measure by simpler measures, which are eligible for numerical computations and evaluations. At this point it is essential to discuss the spaces of natural domain, which are Banach spaces allowing to define the risk measure rigorously. The continuity results mentioned above then involve the norm of the dual spaces.

A particular emphasize is given to *Entropic Value-at-Risk* (cf. [2]), its definition involves the usual entropy.

The history process always allows formulating dynamic programming equations (Hamilton-Jacobi-Bellman equations) given the expectation as objective. In the context of risk averse stochastic programming, this is no longer the case unless the risk functionals can be decomposed or nested. In this context, the risk functionals can be nested in a generalized way in continuous time. Employing the Entropic Value-at-Risk, the corresponding evolution equations have an additional term which corresponds to risk. The additional term makes the infinitesimal generator non-linear, as risk is directed and distinguishes risk from profit.

It turns out that the additional non-linear term can be handled effectively in the Hamiltonian. The presented framework thus gives rise to successfully generalize stochastic and dynamic optimization to risk averse situations.

References

[1] G. Ch. Pflug and A. Pichler *Multistage Stochastic Programming*, Springer Series in Operations Research and financial Engineering (2014), DOI: 10.1007/978-3-319-08843-3.
[2] A. Pichler and R. Schlotter, *Entropy based risk measures*, European Journal of Operational Research, DOI: 10.1016/j.ejor.2019.01.016 (2019).
[3] A. Pichler and A. Shapiro, *Risk averse stochastic programming: time consistency and optimal stopping*, preprint (2019).

Meta Classification for Semantic Segmentation

Hanno Gottschalk
(joint work with M. Rottmann, P. Colling, T. Hack, F. Hüger and P. Schlicht)

1. Semantic Segmentation

Deep Learning approaches have been particularly sucessful on image perception tasks [3]. At the time of writing, deep convolutional neural networks have de facto become the industry standard. In semantic segmentation, this task is combined with localization [1, 3]. A classification problem is solved pixel wise and thereby partitions the image in a number of segments, each containing the pixels adhering to one class.

Mathematically, we consider a $n_x \times n_y$ input image X and a corresponding class matrix $Y \in \mathcal{C}^{(n_x \times n_y)}$ where $\mathcal{C}$ is the finite set of q classes. Y is also referred to as the ground truth. A convolutional neural network defines a function $f_z(y|x,w)$ depending on a (high diemensional) vector of weights w which gives the model-probabilities of observing class $y \in \mathcal{C}$ at output pixel ($z = (x,y)$ given the input image X. The model is trained on i.i.d. samples $(X^{(i)}, Y^{(i)})$ by minimizing ('learning' in machine learning lingo) the cost function given by the negative log likelihood ('cross entropy') in the weights w,

$$-\sum_{j=1}^{n}\sum_{z}\log f_z(Y_z^{(i)}|X^{(i)},w) \longrightarrow \min\,.$$

Given an (approximate) solution $\hat{w}$ to the above optimization problem, we obtain the class prediction per pixel:

$$\hat{Y}_z(X) = \arg\max_{y \in \mathcal{C}} f_z(y|X, \hat{w}).$$

We consider ground thruth class masks $M_y(Y, X) = \{z \in \{1, \ldots, n_x\} \times \{1, \ldots, n_y\} : Y_z = y\}$ and predicted masks $\hat{M}_y(X, \hat{w}) = \{z \in \{1, \ldots, n_x\} \times \{1, \ldots, n_y\} : \hat{Y}_z(X) = y\}$, $y \in \mathcal{C}$. Let $CM_y(Y, X)$ and $C\hat{M}_y(X, \hat{w})$ be respective the sets of connected components. For $k \in C\hat{M}_y(X, \hat{w})$ we define the (adjusted) intersection over unions (IoU) as

$$IoU_{\text{adj.}}(\hat{k}) = \frac{|\hat{k} \cap CM_y(Y, X)|}{\hat{k} \cup \bigcup_{k \in CM_y(Y,X): k \cap \hat{k} \neq \emptyset} k \setminus \bigcup_{k \in CM_y(Y,X): k \cap \hat{k} = \emptyset} \left(k \cap C\hat{M}_y(X, \hat{w})\right)}.$$

Here, $IoU(\hat{k}) = 1$ corresponds to a perfect prediction of the instance k, whereas $IoU(\hat{k}) = 0$ marks $\hat{k}$ as a false positive having no intersection with the ground thruth mask $CM_y(Y, M)$. The adjusted IoU is a measure of quality which can only be calculated if the ground truth $CM_y(Y, X)$ is available. Therefore, this meaure of quality is not available at the time of inference.

2. Meta Regression and Classification

Here we propose methods to predict the IoU-value at the time of inference using aggregated pixel wise dispersion measures. In the first step, we use pixel wise dispersion measures like the entropy E_z or the difference between the leading and the next highest probability D_z to create uncertainty heat maps over the predicted image.

$$E_z(x, w) = -\frac{1}{\log(q)} \sum_{y \in \mathcal{C}} f_z(y|w, x) \log f_z(y|w, x),$$

$$D_z(x, w) = 1 - f_z(\hat{y}_z(x, w)|x, w) + \max_{y \in \mathcal{C} \setminus \{\hat{y}_z(x,w)\}} f_z(y|x, w).$$

We thereafter aggregate these quantities over the boundary or inner regions of the predicted instance $\hat{k}$ of the class $y \in \mathcal{C}$. This generates a set of uncertainty features per every such instance. We complete the data set by the predicted class and information on the size and boundary length of $\hat{k}$ and thus obtain a set of covariates $\xi(\hat{k})$ which can be calculated without knowledge of the ground thruth.

Based on this set of data per instance we progress to the task of metaclassification training a LASSO [8] logistic regression model that is trained to classify false positives ($IoU = 0$) vs non false positives. We thus solve numerically

$$-\sum_{i=1}^{n} \sum_{y \in \mathcal{C}} \sum_{\hat{k} \in C\hat{M}(X^{(i)}, \hat{w})} \Big\{ \chi_{\{IoU(\hat{k})=0\}} \log\left(\tau(\xi(\hat{k})'\beta)\right) + \chi_{\{IoU(\hat{k})>0\}} \log\left(1 - \tau(\xi(\hat{k})'\beta)\right) \Big\} + \lambda \|\beta\|_1 \longrightarrow \min_{\beta}'$$

where the minimization is in β and the shrinkage hyper parameter λ is adjusted afterwards to minimize cross validation error on a separate validation data set. Here τ is the logistic function and χ_A the indicator function of the set A. Likewise, we perform a regression analysis to predict the unknown $IoU(\hat{k})$ from the known covariates $\xi(\hat{k})$ encoding - among other information - the aggreated uncertainty measures.

3. Results

We perform numerical experiments for two segmentation networks, namely the high end Google network Xception65 [4] and a less performant network designed for mobile applications Mobilenet V2 [7]. The nets are publicly available including wthe weigt set $\hat{w}$ containing 22M weights for the Xception65 and 2.3M weights for the Mobilenet V2. Both networks are evaluated on the Cityscapes data set [5] leading to a quota of non $IoU = 0$ instances of 75% and 58%, respectively. These quantities mark the guessing baseline.

Using the mataclassification approach described above, we achieve 81.51% meta classification accuracy for the strong Xception65 network and 78.65% for the Mobilenet V2. Metaclassification is thus able to enhance the metaclassification by roughly 7% and 20%, respectively, compared to the baseline set by guessing that all instances have $IoU > 0$, see Table 1.

A paremeter selection method by gradually decreasing λ is used to determine influential covariates withe $\bar{D}(\hat{k})$, the average over the probability distance found to be most influential. A small set of only 4 four parameters is empirically shown to be almost as efficient as the entire parameter set.

	Xception65		MobilenetV2	
	training	validation	training	validation
	Classification $IoU = 0, > 0$			
ACC, penalized	81.88%(±0.13%)	81.91%(±0.13%)	78.87%(±0.13%)	78.93%(±0.17%)
ACC, unpenalized	81.91%(±0.12%)	81.92%(±0.12%)	78.84%(±0.14%)	78.93%(±0.18%)
ACC, entropy baseline	76.36%(±0.17%)	76.32%(±0.17%)	68.33%(±0.27%)	68.57%(±0.25%)
	Regression $IoU_{\text{adj.}}$			
σ, all metrics	0.181(±0.001)	0.182(±0.001)	0.130(±0.001)	0.130(±0.001)
σ, entropy baseline	0.258(±0.001)	0.259(±0.001)	0.215(±0.001)	0.215(±0.001)
R^2, all metrics	75.06%(±0.22%)	74.97%(±0.22%)	81.50%(±0.23%)	81.48%(±0.23%)
R^2, entropy baseline	49.37%(±0.32%)	49.02%(±0.32%)	49.32%(±0.31%)	49.12%(±0.32%)

Table 1. Comparison of metaclassification results with two nets using only aggregated entropy quantities ('entropy baseline') and further dispersion measures with ($\lambda > 0$) and without ($\lambda = 0$) weight penalization. The number in brackets denote the standard deviations of averaged values over 10 runs with resmpled validation sets.

An extended version of this abstract can be found in [6], see also [2] for an approach to detect false negatives.

References

[1] M. Abadi, A. Agarwal, P. Barham, et al. TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 2015.

[2] R. Chan, M. Rottmann, F. Hüger, P. Schlicht and H. Gottschalk, Application of Decision Rules for Handling Class Imbalance in Semantic Segmentation, arXiv:1901.08394

[3] L.-C. Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". Corr abs/1802.02611 (2018).

[4] F. Chollet. "Xception: deep learning with depthwise separable convolutions". 2017 ieee conference on computer vision and pattern recognition (cvpr) (2017), pp. 1800–1807.

[5] M. Cordts et al. "The cityscapes dataset for semantic urban scene understanding". Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[6] M. Rottmann, P. Colling, T.-P. Hack, F. Hüger, P. Schlicht and H. Gottschalk, Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities, arXiv:1811.00648.

[7] M. Sandler, A. Howard, M-L. Zhu, A. Zhmoginov, L-C Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, arXiv:1801.04381

[8] R. Tibshirani. "Regression shrinkage and selection via the lasso". Journal of the royal statistical society: series b 58 (1996), pp. 267–288.

When it comes to sampling, is smoothness a good thing?

Michela Ottobre

As is well known, a large number of approaches to sampling for a given distribution π consists in designing a dynamics $\{x_m\}$, $n \in \mathbb{N}$ which converges (in some appropriate sense) to π; running the dynamics long enough will then provide the desired samples, approximately distributed according to π. Broadly speaking, one could either design a continuous-time dynamics and then discretize it, or come up straight away with a discrete-time dynamics. If the former approach is chosen, one needs to make sure that the discretization $\{x_n\}_n$ of the continuous-time process $\{x_t\}_t$ does still retain all the relevant properties of x_t - for example, that the invariant measure has not been altered.

Popular continuous-time dynamics used in the context are Langevin dynamics

$$dx_t = -\nabla V(x_t)dt + dW_t \qquad \pi(x) = e^{-V(x)} \tag{1}$$

and, more recently, their *non-reversible* modifications

$$dy_t = -\nabla V(y_t)dt + A \cdot \nabla V(y_t) + dW_t \tag{2}$$

where A is any antisymmetric matrix. The second approach is instead the one taken by Markov Chain Monte Carlo (MCMC) algorithms. Typically such algorithms are reversible.

Very recently the interest has shifted towards the use of *Piecewise Deterministic Markov processes*. These processes are fundamentally different from those traditionally used, not only because they are non-reversible, but also because they are *non-smooth*. In their simplest form they indeed move along straight lines until, at

a random time (typically a non-homogeneous Poisson time), they change velocity and resume their linear motion until the next random check goes off. Because of their ODE-like behaviour these processes have a clear edge with respect to slower diffusion-like algorithms, such as the Random Walk Metropolis algorithm. However, to the best of our knowledge, such processes still remain, in general, quite difficult to simulate.

One of my recent interest is in finding a class of dynamics which is at the same time easy to simulate while retaining this lack-of-smoothness property. One possible directions, which seems promising, consists in considering a class of SDEs, so-called UFG-SDEs (UFG is an acronym for *uniformly finely generated*) which, roughly speaking, strictly contains the wider class of well known hypoelliptic diffusions. The reason why UFG-type SDEs could perform better comes from the following observation: if an SDE is of UFG-type, then there always exist a local change of coordinate such that the original SDE can be turned into a decoupled system ODE+SDE see [1]. In other words, such SDEs are not smooth in every direction, and we believe that such a lack of smoothness could in principle be used to explore the modes of a distribution faster.

References

[1] T. Cass, D. Crisan, P. Dobson and M. Ottobre. *Long-time behaviour of Degenerate diffusions: UFG-type SDE and time-inhomogeneous hypoelliptic processes*, arxiv 1805.013350

Uncertainty quantification using periodic random variables and lattice quasi-Monte Carlo rules

Frances Y. Kuo

(joint work with Vesa Kaarnioja, Ian H. Sloan)

Many studies in uncertainty quantification have been carried out under the assumption of an input random field in which a countable number of independent random variables are each uniformly distributed on an interval, with these random variables entering linearly in the input random field — the so-called "*affine*" model, see e.g., [1, 2, 3, 4, 5]. Here we propose an alternative model of the random field, in which the random variables have the same uniform distribution on an interval, but the random variables enter the input field as *periodic* functions. The field is constructed in such a way as to have *the same mean and covariance function* as the affine random field. Higher moments differ from the affine case, but in general the periodic model seems no less desirable.

More precisely, we consider a random field on a bounded physical domain $D \subseteq \mathbb{R}^d$, where $d = 1, 2$ or 3, with Lipschitz boundary ∂D. The affine model of the random field takes the form

$$A(\boldsymbol{x}, \omega) = \overline{a}(\boldsymbol{x}) + \sum_{j=1}^{\infty} Y_j(\omega)\, \psi_j(\boldsymbol{x}), \quad \boldsymbol{x} \in D,\ \omega \in \Omega, \tag{1}$$

with $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. Here $(Y_j)_{j=1}^{\infty}$ are i.i.d. random variables uniformly distributed on $[-\frac{1}{2}, \frac{1}{2}]$, while $\overline{a}$ and $(\psi_j)_{j=1}^{\infty}$ are real-valued L_∞ functions on D satisfying $\sum_{j=1}^{\infty} \|\psi_j\|_{L_\infty} < \infty$. With these definitions the sum in (1) converges uniformly on D for all values of Y_j, and the random field is pointwise well defined, with mean

$$\mathbb{E}[A(\boldsymbol{x}, \cdot)] \;=\; \overline{a}(\boldsymbol{x}) + \sum_{j=1}^{\infty} \mathbb{E}[Y_j]\, \psi_j(\boldsymbol{x}) \;=\; \overline{a}(\boldsymbol{x}),$$

and covariance function

$$\mathrm{cov}(A)(\boldsymbol{x}, \boldsymbol{x}') \;=\; \sum_{j=1}^{\infty} \sum_{j'=1}^{\infty} \mathbb{E}[Y_j\, Y_{j'}]\, \psi_j(\boldsymbol{x})\, \psi_{j'}(\boldsymbol{x}') \;=\; \frac{1}{12} \sum_{j=1}^{\infty} \psi_j(\boldsymbol{x})\, \psi_j(\boldsymbol{x}').$$

The fact that the random variables in (1) occur linearly seems to be a result of history, rather than something imposed by modeling assumptions. Suppose instead that we replace Y_j in (1) by $\Theta(Y_j)$, where $\Theta\colon [-\frac{1}{2}, \frac{1}{2}] \to \mathbb{R}$ is a continuous function with the properties that

$$\int_{-1/2}^{1/2} \Theta(y)\, \mathrm{d}y \;=\; 0 \qquad \text{and} \qquad \int_{-1/2}^{1/2} \Theta^2(y)\, \mathrm{d}y \;=\; \frac{1}{12}.$$

Then (1) is replaced by a random field with *exactly the same mean and covariance*.

Specifically, we explore a *periodic* choice of Θ, namely

$$\Theta(y) \;=\; \frac{1}{\sqrt{6}} \sin(2\pi y) \qquad \text{for} \quad y \in [-\tfrac{1}{2}, \tfrac{1}{2}].$$

The motivation for this choice is that the random field is now a *1-periodic function of the random variables*, and periodic spaces are known to be especially advantageous in the context of *lattice quasi-Monte Carlo rules*. By using the periodic model of random fields instead of the affine model, it is possible to carry out lattice rule calculations of expected values in high dimensions with *higher order convergence rates*.

We apply the new periodic model of the random field in the context of computing expected values of a quantity of interest arising from an elliptic PDE with random coefficients. Following [4], the parametric weak problem is: for $f \in H^{-1}(D)$ and $\boldsymbol{y} \in [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$, find $u(\cdot, \boldsymbol{y}) \in H_0^1(D)$ such that

$$(2) \quad \int_D a(\boldsymbol{x}, \boldsymbol{y})\, \nabla u(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla \phi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \;=\; \int_D f(\boldsymbol{x})\, \phi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad \text{for all} \quad \phi \in H_0^1(D).$$

The quantity of interest is the expected value of some bounded linear functional $G(\cdot)$ of the PDE solution $u(\cdot, \boldsymbol{y})$, taken over the parameter space:

$$(3) \qquad \mathbb{E}[G(u)] \;=\; \lim_{s \to \infty} \int_{[-\frac{1}{2}, \frac{1}{2}]^s} G(u(\cdot, (y_1, \ldots, y_s, 0, 0, \ldots)))\, \mathrm{d}y_1 \cdots \mathrm{d}y_s.$$

However, instead of the affine model in [4], the diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$ is now 1-periodic with respect to $\boldsymbol{y}$, as given by

$$a(\boldsymbol{x}, \boldsymbol{y}) = \overline{a}(\boldsymbol{x}) + \frac{1}{\sqrt{6}} \sum_{j=1}^{\infty} \sin(2\pi y_j)\, \psi_j(\boldsymbol{x}). \tag{4}$$

We approximate the expected value by truncating the series in (4) to s terms, discretizing the PDE (2) using a (piecewise linear) finite element method with meshwidth h, and then approximating the (now s-dimensional) integral (3) using lattice rules with n points.

The key step in our analysis is to obtain a new regularity estimate of the PDE solution with respect to the parametric variable $\boldsymbol{y}$ within the new periodic framework. Under a number of assumptions, including $f \in H^{-1+t}(D)$ with $0 \le t \le 1$, $G \in H^{-1+t'}(D)$ with $0 \le t' \le 1$, and

$$\sum_{j=1}^{\infty} \|\psi_j\|_{L_\infty}^p < \infty \qquad \text{for some} \quad 0 < p < 1,$$

we obtain the overall error bound of

$$\mathcal{O}(s^{-2/p+1} + h^{t+t'} + n^{-1/p}),$$

with the implied constant independent of s, h, and n. In particular, the lattice integration error of $\mathcal{O}(n^{-1/p})$ beats the $\mathcal{O}(n^{-\min(1/p-1/2,1-\delta)})$ rate in [4, 5], $\delta > 0$, for the affine model with *randomly shifted lattice rules*, and matches the rate in [2] for the affine model with *interlaced polynomial lattice rules*. The dimension truncation error of $\mathcal{O}(s^{-2/p+1})$ matches the rate in [3]. The finite element error of $\mathcal{O}(h^{t+t'})$ can potentially be improved by using higher order elements.

In summary, lattice rules can achieve higher order convergence rates in the periodic setting, either as deterministic rules or randomized rules. Lattice rules are elegant and easy to implement: all we require is an integer generating vector and modulo arithmetic. We see great potential for lattice rules to be widely used for uncertainty quantification under the new periodic model of the random field.

References

[1] A. Cohen, R. DeVore, and Ch. Schwab, *Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs*, Found. Comput. Math. **10** (2010), 615–646.

[2] J. Dick, F. Y. Kuo, Q. T. Le Gia, D. Nuyens, and Ch. Schwab, *Higher order QMC Petrov–Galerkin discretization for affine parametric operator equations with random field inputs*, SIAM J. Numer. Anal. **52** (2014), 2676–2702.

[3] R. N. Gantner, *Dimension truncation in QMC for affine-parametric operator equations*, in: Monte Carlo and Quasi-Monte Carlo Methods 2016 (A. B. Owen and P. W. Glynn, eds), Springer, 2018, pp. 249–264.

[4] F. Y. Kuo, Ch. Schwab, and I. H. Sloan, *Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients*, SIAM J. Numer. Anal. **50** (2012), 3351–3374.

[5] Ch. Schwab, *QMC Galerkin discretization of parametric operator equations*, in: Monte Carlo and Quasi-Monte Carlo Methods 2012 (J. Dick, F. Y. Kuo, G. W. Peters and I. H. Sloan, eds.), Springer Verlag, Heidelberg, 2013, pp. 613–629.

Multilevel representations of stationary Gaussian random fields and efficient sampling methods

Markus Bachmayr

(joint work with Albert Cohen, Ivan Graham, Giovanni Migliorati, Van Kien Nguyen, Robert Scheichl)

Circulant embedding is a technique for drawing samples of stationary Gaussian random fields on uniform spatial grids. Using the fast Fourier transform (FFT), the method requires only $\mathcal{O}(N \log N)$ operations to create a sample in N points. This requires the symmetric Toeplitz covariance matrix corresponding to the grid points to be embedded into a positive definite circulant matrix. For ensuring positive definiteness, in many cases it suffices to choose the domain before embedding sufficiently large, exploiting the spatial decay of the covariance function. In the important case of covariance functions $\rho_{\lambda,\nu}$ from the Matérn family with smoothness parameter ν and correlation length λ, it has recently been shown in [4] that an extension of the dimensions of the original domain by a factor proportional to $\lambda\sqrt{\nu}\max\{\log\nu, |\log(h/\lambda)|\}$, with $\nu \geq 1/2$ and grid size $h > 0$, suffices to ensure positive definiteness of the resulting circulant. It was also conjectured in [4] that the eigenvalues of this circulant exhibit, uniformly in N, the same asymptotic decay as the exact Karhunen-Loève (KL) eigenvalues of $\rho_{\lambda,\nu}$; this property is particularly relevant for QMC integration methods. In [2], we prove this conjecture to hold true up to a modification by a factor of order $\mathcal{O}(|\log h|)$.

The embedding into a circulant can also be interpreted as a periodisation of the underlying covariance function by a simple reflection. Alternatively, as considered in [1] (see also [3] for a similar approach), a periodisation based on a smooth truncation of $\rho_{\lambda,\nu}$ leads to a grid-independent periodic random field. We show in [2] that it suffices in this case to extend the domain by a factor proportional to $\lambda\max\{\sqrt{\nu}\log\nu, 1/\sqrt{\nu}\}$ for any $\lambda, \nu > 0$. Numerical experiments show that the grid-independence of this factor leads to substantial computational savings already for moderate h. By this periodisation, as described in [1], one also obtains an alternative to the standard KL expansion on complicated domains, where the decay of the eigenvalues (which can be computed by FFT) is preserved, but the eigenfunctions are explicitly known restrictions of trigonometric functions. Applying the square root of the covariance operator to periodic Meyer wavelets in this representation, in [1] we obtain wavelet expansions for a class of random fields including the Matérn family, which yield improved convergence rates for tensor Hermite polynomial approximations of solutions of random diffusion equations with lognormal coefficients.

References

[1] M. Bachmayr, A. Cohen, G. Migliorati, *Representations of Gaussian Random Fields and Approximation of Elliptic PDEs with Lognormal Coefficients*, J Fourier Anal Appl **24** (2017), 621–649.

[2] M. Bachmayr, I. G. Graham, V. K. Nguyen, R. Scheichl, *Unified Analysis of Periodisation-Based Sampling Methods for Matérn Covariances*, in preparation (2019).

[3] T. Gneiting, H. Ševčíková, D. B. Percival, M. Schlather, Y. Jiang, *Fast and Exact Simulation of Large Gaussian Lattice Systems in $\mathbb{R}^2$: Exloring the Limits*, J Comp Graph Stat **15** (2006), 483–501.

[4] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, I. H. Sloan, *Analysis of Circulant Embedding Methods for Sampling Stationary Random Fields*, SIAM J Numer Anal **56** (2018), 1871–1895.

Glorified linear interpolation

Houman Owhadi

Although numerical approximation, statistical inference and learning are traditionally seen as entirely separate subjects, they are intimately connected through the common purpose of making estimations with partial information. This talk is an invitation to explore these connections from the consolidating perspective of game/decision theory and it is motivated by the suggestion that these confluences might not just be objects of curiosity but can constitute a pathway to simple solutions to fundamental problems in all three areas. We illustrate this point through problems related to numerical homogenization, operator adapted wavelets, computation with dense kernel matrices and to the kernel selection/design problem in Machine Learning. In these interplays, accurate reduced/multiscale models (for PDEs) can be identified as optimal bets for adversarial games describing the process of computing with partial information. Moreover, efficient kernels (for ML) can be selected by using relative energy content at fine scales (with a notion of scale corresponding to the number of data points) as an ordering criterion leading to the identification of (data driven) flows in kernel spaces (Kernel Flows), that (1) enable the design of bottomless networks amenable to some degree of analysis (2) appear to converge towards kernels with good generalization properties.

This talk covers joint work with F. Schäfer, C. Scovel, T. Sullivan, G. R. Yoo and L. Zhang.

A Bayesian Conjugate Gradient Method

Jon Cockayne

(joint work with Chris J. Oates, Ilse Ipsen, Mark Girolami)

A fundamental task in numerical computation is the solution of large linear systems of the form

$$A\boldsymbol{x}^* = \boldsymbol{b} \tag{1}$$

where $A \in \mathbb{R}^{d\times d}$ and $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^d$. The conjugate gradient (CG) method is an iterative method which offers rapid convergence to the solution, particularly when an effective preconditioner is employed. However, for more challenging systems a substantial error can be present even after many iterations have been performed. The estimates obtained in this case are of little value unless further information can be provided about the numerical error. Probabilistic numerical methods are a class of numerical methods that return probability measures designed to provide

probabilistic uncertainty quantification for numerical error. In [1] we develop a probabilistic numerical method for the task of solving linear systems.

A Bayesian approach is employed. To this end, we assume that $\boldsymbol{x}^*$ is a-priori unknown and model this uncertainty with a prior distribution. For a generic Gaussian prior, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_0, \Sigma_0)$, introduce an as-yet arbitrary set of search directions $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m \in \mathbb{R}^d$. Let $S_m \in \mathbb{R}^{d\times m}$ denote the matrix with these search directions as its columns. Then it is straightforward to show that

$$\boldsymbol{x} | S_m^\top A\boldsymbol{x} = S_m^\top \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{x}_m, \Sigma_m).$$

where $\boldsymbol{x}_m, \Sigma_m$ have closed-form expressions that involve the inversion of the matrix $\Lambda_m = S_m^\top A\Sigma_0 A^\top S_m$. The Bayesian conjugate gradient method (BayesCG) follows a similar procedure to classical iterative methods such as CG by iteratively constructing search directions $\boldsymbol{s}_m$ so that $\Lambda_m = I$, i.e. so that the search directions are $A\Sigma_0 A^\top$-orthonormal. To be specific, the BayesCG search directions are defined as

$$\begin{aligned} \tilde{\boldsymbol{s}}_1 &= \boldsymbol{r}_0 && \\ \tilde{\boldsymbol{s}}_m &= \tilde{\boldsymbol{r}}_m - \langle \boldsymbol{s}_{m-1}, \boldsymbol{r}_m \rangle_{A\Sigma_0 A^\top} \boldsymbol{s}_{m-1} && m = 2, \ldots, d \\ \boldsymbol{s}_m &= \tilde{\boldsymbol{s}}_m / \|\tilde{\boldsymbol{s}}_m\|_{A\Sigma_0 A^\top} && m = 1, \ldots, d \end{aligned} \tag{2}$$

where $\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{x}_m$. It can then be shown that these search directions are $A\Sigma_0 A^\top$-orthonormal, and moreover that with these search directions $\boldsymbol{x}_m, \Sigma_m$ can be computed iteratively as

$$\begin{aligned} \boldsymbol{x}_m &= \boldsymbol{x}_{m-1} + \Sigma_0 A^\top \boldsymbol{s}_m (\boldsymbol{s}_m^\top \boldsymbol{r}_m) \\ \Sigma_m &= \Sigma_{m-1} - \Sigma_0 A^\top \boldsymbol{s}_m \boldsymbol{s}_m^\top A \Sigma_0. \end{aligned}$$

The posterior mean $\boldsymbol{x}_m$ has interesting optimality properties when computed in this way. Define the m^{th} Krylov Subspace generated by a matrix $B \in \mathbb{R}^{d\times d}$ and a vector $\boldsymbol{v} \in \mathbb{R}^d$ as

$$K_m(B, \boldsymbol{v}) = \operatorname{span}(\boldsymbol{v}, B\boldsymbol{v}, \ldots, B^{m-1}\boldsymbol{v}).$$

Then, introducing the affine space $K_m^* = \boldsymbol{x}_0 + \Sigma_0 A^\top K_m(A\Sigma_0 A^\top, \boldsymbol{r}_0)$, we can show that $\boldsymbol{x}_m$ is optimal in the following sense:

$$\boldsymbol{x}_m = \operatorname*{arg\,min}_{\boldsymbol{x} \in K_m^*} \|\boldsymbol{x} - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}. \tag{3}$$

This optimality result also allows us to bound the contraction of the relative error of the posterior mean:

$$\frac{\|\boldsymbol{x}_m - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}}{\|\boldsymbol{x}_m - \boldsymbol{x}^*\|_{\Sigma_0^{-1}}} \leq 2 \left(\frac{\sqrt{\kappa(\Sigma_0 A^\top A)} - 1}{\sqrt{\kappa(\Sigma_0 A^\top A)} + 1} \right)^m \tag{4}$$

which, importantly, is exponential in m.

Eq. (3) also leads to an interpretation of BayesCG as a generalisation of CG, in the sense that when $\Sigma_0 = A^{-1}$ the CG iterate is recovered as the posterior mean. This can be seen by comparing Eq. (3) to the similar optimality result for the

solution estimate produced by CG [2]. Note however that this is not a practical choice of prior, as computing Σ_m would then require computation of A^{-1}.

This leads to the critical question of prior choice. We particularly advocate the idea of using a preconditioner to construct a prior covariance. Given a preconditioner P which is such that $\kappa(P^{-1}A) \ll \kappa(A)$, the choice $\Sigma_0 = (P^\top P)^{-1}$ accelerates the rate of convergence of the posterior mean, as is clear from Eq. (4). However, it can be shown both theoretically and empirically that the rate of contraction of the posterior covariance is essentially linear regardless of choice of Σ_0, implying that the uncertainty quantification provided by the posterior will be conservative in general. This mismatch is because the search directions in Eq. (2) depend implicitly on $\boldsymbol{x}^*$, in a way that violates the linearity assumption exploited to produce a conjugate Gaussian posterior mean. Correcting for this issue is the subject of ongoing research.

References

[1] J. Cockayne, C.J. Oates, I. Ipsen, M. Girolami, *A Bayesian conjugate gradient method*, Bayesian Anal., **to appear** (2019).
[2] G.H. Golub, C.F. Van Loan, *Matrix computations*, Johns Hopkins University Press (1996).

Objective priors in the empirical Bayes framework

Ilja Klebanov
(joint work with Alexander Sikorski, Christof Schütte, Susanna Röblitz)

The choice of the prior distribution for Bayesian inference is a controversial issue in the statistics community. When no prior information on the parameter θ is available, one may want to choose a prior that is in some sense non-informative, thus letting the information within the posterior be dominated by the measurement x. Here we assume that the parametric model

$$\mathcal{M} = \{p(x|\theta) : \theta \in \Theta,\, x \in \mathcal{X}\}$$

is given. *Objective priors* [1, 2] provide one attempt of defining such priors, which besides non-informativity have the desirable property of being invariant under transformations of the parameter (reparametrization). This idea follows the reasoning of Shore and Johnson [6], that

> "[...] reasonable methods of inductive inference should lead to consistent results when there are different ways of taking the same information into account (for example, in different coordinate systems)."

The key idea behind reference priors [2], a commonly used objective Bayes method, is to use information-theoretic concepts such as *mutual information* $\mathcal{I}[\mathcal{M}]$ and

missing information $\mathcal{I}[\mathcal{M}^\infty]$, which are transformation-invariant by nature,[1]

$$\mathcal{I}[\mathcal{M}] = D_{\mathrm{KL}}\big(p(x,\theta) \,\|\, p(x)p(\theta)\big),$$
$$\mathcal{I}[\mathcal{M}^\infty] = \lim_{k\to\infty} \mathcal{I}[\mathcal{M}^{(k)}], \qquad \mathcal{M}^{(k)} = \{p(x^{(k)}|\theta) : \theta \in \Theta,\, x^{(k)} \in \mathcal{X}^k\},$$

where D_{KL} denotes the Kullback-Leibler divergence and $x^{(k)}$ denotes the measurement corresponding to k (artificial) independent replications of x. Maximizing the missing information $\mathcal{I}[\mathcal{M}^\infty]$ as a function of the prior can be argued to favor non-informativity of the prior, see [1].

In the case when information is available in the form of data $X = \{x_1, \ldots, x_M\}$ stemming from the hierarchical model

$$\theta_m \overset{\text{iid}}{\sim} p(\theta), \qquad x_m \sim p(x|\theta_m), \qquad m = 1, \ldots, M,$$

– the reader may think of large clinical studies with patient-specific parametrizations – *empirical Bayes methods* [3] provide an alternative approach for choosing the prior. Roughly speaking, these methods proceed by first estimating the prior distribution from X and then applying Bayes' rule for each parameter θ_m and measurement x_m (each patient) separately.[2] From now on we will focus on the first step, the estimation of the prior, where we will denote such estimates by the letter π. The (conceptually) simplest approach is to maximize the likelihood $L(\pi)$ of the prior,

$$\pi_{\mathrm{MLE}} = \arg\max_{\mathcal{P}} L(\pi), \qquad L(\pi) = p(X|\pi) = \prod_{m=1}^{M} \int p(x_m|\theta)\, \pi(\theta)\, \mathrm{d}\theta.$$

However, in the non-parametric case, i.e. when no parametric form of the prior distribution is assumed, this leads to overfitting. In fact, π_{MLE} is provably a discrete distribution with at most M nodes [7]. A natural workaround is regularization or penalization, where a penalty term $\Phi(\pi)$ is subtracted from the log-likelihood, resulting in the so-called *maximum penalized likelihood estimate* (MPLE)

$$\pi_{\mathrm{MPLE}} = \arg\max_{\mathcal{P}} \log L(\pi) - \gamma\Phi(\pi).$$

Here, $\gamma > 0$ is the smoothness parameter which balances the goodness of fit with the smoothness or non-informativity of the prior. Typical penalty terms are

$$\Phi_1(\pi) = \|\pi\|_{L^2}, \qquad \Phi_2(\pi) = \int \frac{\pi'(\theta)^2}{\pi(\theta)}\, \mathrm{d}\theta, \qquad \Phi_3(\pi) = \int \pi(\theta)\, \log \pi(\theta)\, \mathrm{d}\theta.$$

However, all these penalty terms are variant under reparametrizations, resulting in inconsistent estimates of the prior: If $\varphi\colon \Theta \to \tilde{\Theta}$ is a reparametrization and the estimation process is performed in the transformed parameter space $\tilde{\Theta}$ in place of

[1] Note that $\mathcal{I}[\mathcal{M}^{(k)}]$ typically tends to infinity for $k \to \infty$, which requires a technical workaround not discussed here, see e.g. [2].

[2] One might object that such an approach uses the data twice – more precisely the prior estimate should be used for any potential new measurement x_{M+1} or, alternatively, the estimation of the mth parameter should rely on a prior estimated from all but the mth measurement. Since we are mainly interested in the estimation of the prior, we will omit this detail here.

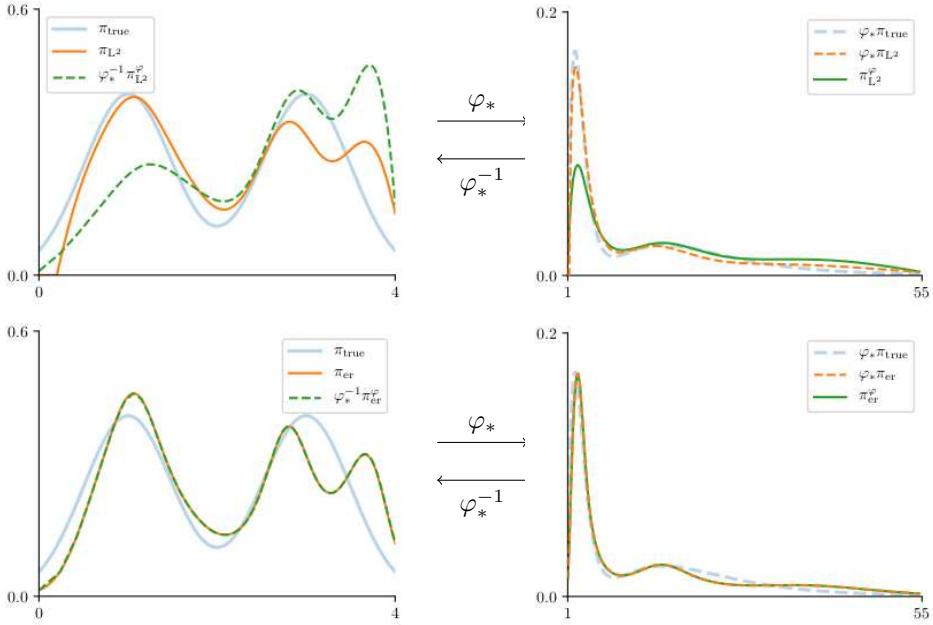

FIGURE 1. MPLE using Tikhonov regularization (denoted by π_{L^2}) is not invariant under reparametrizations, $\pi^{\varphi}_{\text{est}} \neq \varphi_{\#}\pi_{\text{est}}$. The missing information, on the other hand, is transformation-invariant, hence the empirical reference prior is consistent under reparametrizations, $\pi^{\varphi}_{\text{er}} = \varphi_{\#}\pi_{\text{er}}$.

Θ, the resulting estimate $\pi^{\varphi}_{\text{est}}$ can differ from the pushforward of the estimate in the original space, $\pi^{\varphi}_{\text{est}} \neq \varphi_{\#}\pi_{\text{est}}$, see Figure 1. In order to overcome this obstacle and motivated by the definition of reference priors [2], we suggest using the missing information $\mathcal{I}[\mathcal{M}^{\infty}](\pi)$ as a penalty term for MPLE, thus combining the objective Bayesian and empirical Bayes methodologies. The resulting prior estimate will be referred to as the *empirical reference prior* and is given by[3]

$$\pi_{\text{er}} = \arg\max_{\mathcal{P}} \log L(\pi) + \gamma \mathcal{I}[\mathcal{M}^{\infty}](\pi).$$

As visualized in Figure 1, this estimate is consistent under reparametrizations.

Under the assumption of asymptotic normality [4], the penalty term can be simplified to $\Phi(\pi) = D_{\text{KL}}(\pi \,\|\, \pi_J)$, where π_J denotes the Jeffreys prior [5].

Similar to reference priors, the generalization of this approach to dimensions $d > 1$ of the parameter space is not unambiguous [4].

[3]Again, we omit the technical workaround necessary to handle the limit $\mathcal{I}[\mathcal{M}^{\infty}] = \lim_{k\to\infty} \mathcal{I}[\mathcal{M}^{(k)}]$. For details see [4].

References

[1] J. M. Bernardo, *Reference posterior distributions for Bayesian inference*, Journal of the Royal Statistical Society: Series B (Methodological) 41.2 (1979): 113-128.

[2] J. O. Berger et al., *The formal definition of reference priors*, The Annals of Statistics 37.2 (2009): 905-938.

[3] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall/CRC, 2010.

[4] I. Klebanov et al., *Objective priors in the empirical Bayes framework*, arXiv preprint arXiv:1612.00064.

[5] H. Jeffreys, *An invariant form for the prior probability in estimation problems*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 186.1007 (1946): 453-461.

[6] J. Shore and R. Johnson, *Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy*, IEEE Transactions on information theory 26.1 (1980): 26-37.

[7] B. M. Lindsay, *Mixture models: theory, geometry and applications*, NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics and the American Statistical Association, 1995.

Modes of measures on infinite-dimensional topological vector spaces

Han Cheng Lie

(joint work with T. J. Sullivan)

In Bayesian inference, maximum a posteriori (MAP) estimators are modes of the posterior probability measure. Together with the posterior mean and median, they provide imperfect but nevertheless useful descriptions of the posterior. When the unknown parameter lives in a finite-dimensional topological vector space, then Lebesgue measure provides a canonical reference measure; if in addition the posterior measure of interest is absolutely continuous with respect to Lebesgue measure, then a mode of the posterior measure may be characterised as a maximiser of the Lebesgue density. In contrast, if the parameter lives in an infinite-dimensional topological vector space, then there is no canonical reference measure, and thus the notion of a mode as a maximiser of a Lebesgue density no longer applies. This raises the question of how to define a mode of a measure on an infinite-dimensional topological vector space as an object intrinsic to the measure.

Dashti et al. [2] considered the setting of a posterior measure μ on a real separable Banach space $(X, \|\cdot\|)$, given a centred Gaussian prior μ_0 supported on X, and given standard assumptions on the observational noise and misfit. A (strong) mode of the posterior μ was defined to be any vector $x \in X$ such that

$$\lim_{r \downarrow 0} \frac{\sup_{z \in X} \mu(B(z,r))}{\mu(B(x,r))} = 1, \tag{1}$$

where $B(z,r) = \{z \in X \ : \ \|z - x\| < r\}$. The interpretation of (1) is that, in the small radius limit, translating a ball of radius r away from a mode x cannot yield an increase in measure. Furthermore, the presence of the supremum in (1) implies

that this domination over all translations is uniform. One of the key results in [2] is to show that strong modes must belong to the Cameron-Martin space of μ_0.

In [3], Helin and Burger considered the setting of arbitrary priors μ_0 supported on a real separable Banach space X. They defined a weak mode with respect to a dense proper subspace E of X as some $x \in X$ such that

$$\lim_{r \downarrow 0} \frac{\mu(B(z,r))}{\mu(B(x,r))} \leq 1, \quad \forall z \in x - E. \tag{2}$$

As is the case in (1), (2) captures the property that the measure of a small ball centred at a mode dominates that of translates of the ball. Unlike strong modes however, only translations by vectors in E are permitted for E-weak modes, and the domination is not uniform. One motivation for considering dense proper subspaces E is to establish a connection to the work of Dashti et al. considered above: if one chooses the prior to be a centred Gaussian, then one can choose the set E to be the Cameron-Martin space of μ_0.

One of the questions raised by Helin and Burger was to determine whether there exist conditions for an E-weak mode to be a strong mode. This problem was motivated by the fact that any strong mode is also an E-weak mode for any arbitrary dense proper subspace E. The significance of this problem was emphasised in the recent work of Agapiou et al. [4], in which it is suggested that the E-weak mode may be more practical to work with in certain applications.

In this talk, we present our solution to the problem of equivalence of modes, as presented in [1]. The first step is to introduce a new type of mode, hereafter referred to as an 'E-strong mode', which is a vector $x \in X$ for which

$$\lim_{r \downarrow 0} \frac{\sup_{z \in x - E} \mu(B(z,r))}{\mu(B(x,r))} \leq 1. \tag{3}$$

We show that if E is topologically dense in X, then an E-strong mode is a strong mode. This result relies on our result that for every $r > 0$, the evaluation map $z \mapsto \mu(B(z,r))$ is lower semicontinuous on X. Furthermore, we introduce a new concept: given some $x \in X$ and $\emptyset \neq E \subseteq X$, the pair (x, E) is said to satisfy the 'uniformity condition' if there exists a $v^* \in E$ and a $0 < r^* < 1$ such that

$$\mu(B(x - v^*, r)) = \sup_{z \in x - E} \mu(B(z,r)), \quad \forall r \in (0, r^*). \tag{4}$$

We show that for a given E-weak mode x, if (x, E) satisfies the uniformity condition, then x is an E-strong mode. Thus, topological density of E and the uniformity condition suffice for an E-weak mode to be a strong mode.

We further investigate the uniformity condition by presenting an example of a finite measure on $X = \mathbb{R}^2$ with dense subset $E = \mathbb{Q}^2$, where the measure does not satisfy the uniformity condition and has countably many E-weak modes but no E-strong mode. This example shows that the uniformity condition cannot be removed in general, and shows that it is possible for measures to have no strong mode. We also show that if the measure μ satisfies Anderson's inequality, i.e. if

for any symmetric, convex, measurable set A it holds that

$$\mu(A + x) \leq \mu(A), \quad \forall x \in X,$$

then for any $\emptyset \neq E \subseteq X$ it holds that $(0, E)$ satisfies the uniformity condition. Finally, we discuss extensions of the preceding results to metrisable topological vector spaces and possibly infinite measures.

References

[1] H.C. Lie and T.J. Sullivan, *Equivalence of weak and strong modes of measures on topological vector spaces*, Inverse Problems **34** (2018), 115013.

[2] M. Dashti, K.J.H. Law, A.M. Stuart and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems **29** (2013), 095017.

[3] T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems **31** (2015), 085009.

[4] S. Agapiou, M. Burger, M. Dashti and T. Helin, *Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems*, Inverse Problems **34** (2018), 045002.

Bernstein–von Mises Theorems for Linear Inverse Problems

Hanne Kekkonen

(joint work with Matteo Giordano)

We consider the continuous linear inverse problem of recovering an unknown function $f^\dagger$ from a noisy indirect measurement

$$M_\varepsilon = Af^\dagger + \varepsilon \mathbb{W}, \tag{1}$$

where $\mathbb{W}$ is assumed to be Gaussian white noise in a separable Hilbert space W_2. Note that while $\mathbb{W}$ can be defined by its actions on W_2, it does not take values there almost surely.

We follow the Bayesian approach to inverse problems, choosing a centred Gaussian prior Π for the unknown. The solution to the statistical inverse problem is then the conditional distribution of f given M_ε, and the mean or mode of this posterior distribution can be used as a point estimator. The main appeal of the method is, however, that it automatically delivers a quantification of uncertainty in the reconstruction, obtained through credible sets, i.e. regions of the parameter space with specified high posterior probabilities.

We are interested whether this uncertainty quantification is objectively valid. The central question is: Do credible sets have the correct frequentist coverage in the small noise limit? That is, do we have, for some $C = C(M_\varepsilon)$,

$$\Pi\Big(f \in C \,|\, M_\varepsilon\Big) \approx 1 - \alpha \quad \Leftrightarrow \quad \mathbb{P}\Big(f^\dagger \in C(M_\varepsilon^\dagger)\Big) \approx 1 - \alpha, \tag{2}$$

with a small $\alpha \in (0, 1)$ as $\varepsilon \to 0$? The importance of the above questions is not restricted just to the Bayesian paradigm. In linear Bayesian inverse problems with Gaussian priors the conditional mean estimator coincides with a Tikhonov regulariser $\bar{f}$ with a Cameron-Martin space norm penalty. Thus, if (2) holds for

a credible set C centred at the posterior mean, we can use C as an (asymptotic) frequentist confidence region for the Tikhonov regulariser $\bar{f}$.

Obtaining optimal contraction rates is not enough to answer the above question even in the parametric case. For finite-dimensional models the Bernstein–von Mises (BvM) theorem implies that credible sets coincide asymptotically with frequentist confidence regions. Understanding the frequentist properties of nonparametric credible sets presents a more delicate matter. It is well known that the BvM phenomenon may fail to hold even in a simple nonparametric regression model, where credible balls in L^2 can have null asymptotic coverage.

In paper [3] we follow the semiparametric approach presented in [1] and extend the results to linear inverse problems of the general form (1). In particular, we prove BvM theorems for functionals $\langle f, \psi \rangle$, with a large family of test functions ψ, which entails the convergence of $\varepsilon^{-1}\langle f, \psi \rangle \,|\, M_\varepsilon$ to a limiting Gaussian process with optimal covariance structure that recovers the semiparametric information lower bound. As a consequence, we deduce the statistical efficiency of plug-in Tikhonov regularisers $\langle \bar{f}, \psi \rangle$ and that credible intervals centred at such estimators constitute asymptotically valid and optimal confidence intervals. The applicability of the general theory is illustrated by deriving sufficient conditions on the test functions for the BvM phenomenon to occur in case of recovering an unknown source function in elliptic boundary value problems (BVP).

We also show for the elliptic BVP example, in which the properties of the crucial 'inverse Fisher information' operator $(A^*A)^{-1}$ are well-understood, that the techniques employed previously can be refined to further relax the assumptions on the test functions to depend only on the smoothing properties of A. Finally, by requiring a slightly stronger smoothness, we adapt the program laid out in [2] to the problem at hand, and obtain a nonparametric BvM theorem which implies that certain nonparametric credible sets built around the Tikhonov regulariser $\bar{f}$ have asymptotically correct coverage and optimal diameter. Note that we do not make additional assumptions about the smoothness of f. Instead of assuming a source condition to achieve convergence in a desired space, we study the convergence in a larger space which is defined by the smoothness of ψ.

References

[1] F. Monard, R. Nickl and G. Paternain, *Efficient nonparametric Bayesian inference for X-ray transforms*, The Annals of Statistics **47**, 2 (2019), 1113–1147.

[2] R. Nickl, *Bernstein-von Mises theorems for statistical inverse problems I: Schrödinger equation*, arXiv:1707.01764 (2017).

[3] M. Giordano and H. Kekkonen, *Bernstein-von Mises theorems and uncertainty quantification for linear inverse problems*, arXiv:1811.04058 (2018).

Perturbation theory of Markov chains and Monte Carlo within Metropolis

Daniel Rudolf
(joint work with Felipe Medina-Aguayo and Nikolaus Schweizer)

Perturbation theory for Markov chains addresses the question how small differences in the transition mechanism of Markov chains are reflected in differences between their distributions after a certain number of steps. In the statistical analysis of datasets the approximate simulation/sampling of posterior distributions is of particular interest. For this goal approximations of transition kernels can be used (and considered as perturbations) which might lead to implementable algorithms and a reduction of the computational cost. However, they also might change the invariant distribution, such that posterior and stationary distribution do not coincide anymore. By using the perturbation theory we derive estimates of the bias of such approximations of geometrically ergodic Markov chains. We illustrate the result by considering a Monte Carlo within Metropolis (MCWM) algorithm. Here it is essential to control the error in the approximation of the likelihood function and, in addition to that, we need to verify a stability condition of the MCWM Markov chain, either by proving a Lyapunov-type condition, or by a restriction to the center of the state space. The talk was based on [1, 2] and we refer for further details to that papers.

References

[1] F. Medina-Aguayo, D. Rudolf and N. Schweizer, *Perturbation bounds for Monte Carlo within Metropolis via restricted approximations*, (2018) https://arxiv.org/abs/1809.09547
[2] D. Rudolf and N. Schweizer, *Perturbation theory for Markov chains via Wasserstein distance*, Bernoulli **24** (2018), 2610–2639.

Noise-Level Robust Sampling Methods for Bayesian Inverse Problems

Björn Sprungk
(joint work with Daniel Rudolf, Claudia Schillings, Philipp Wacker)

The Bayesian approach to inverse problems and the corresponding numerical methods for sampling or integrating w.r.t. the resulting posterior measure has drawn a lot of attention in recent years. For instance, there has been an intense research on dimension-independent sampling methods suitable for Bayesian inference in high- or even infinite-dimensional spaces. Also the development of multilevel techniques or the employment of surrogate methods for reducing the computational burden of expensive likelihood models has been a fruitful field of research. However, a third important computational challenge has drawn less attention so far, namely, the case of a concentrated posterior measure. Such a situation occurs if the observational data is highly informative, e.g., due to a small noise or a large amount of data. From a statistical point of view a concentrated posterior is a rather desirable situation but for (naïve) numerical methods this can pose a serious challenge.

In this talk, we provide some first analysis of several sampling methods such as Markov chain Monte Carlo, importance sampling, and quasi Monte Carlo for the case of concentrated posterior measures. In particular, we show that standard algorithms based on the prior measure yield a decreasing efficiency as the posterior becomes more concentrated. As a remedy we propose to use of the Laplace approximation of the posterior in order to construct noise-level robust sampling methods. The Laplace approximation is a Gaussian measure centered at the MAP estimate of the posterior with a covariance based on the local curvature of the log posterior density at the MAP. It is derived from the classical Laplace method for asymptotic approximations of integrals. We show that the posterior converges to its Laplace approximation in the small noise limit in Hellinger distance. This implies that the density of the posterior w.r.t. its Laplace approximation converges to one, whereas the density of the posterior w.r.t. the prior deterioates to a Dirac function as the observational noise decays. This suggests that numerical methods based on the Laplace approximation behave well in the small noise limit. In fact, we prove for the case of a concentrating posterior measure that: (1) Metropolis–Hastings algorithms employing the Laplace approximation in their proposal kernel have a nondegenerating mean acceptance rate and mean squared jump distance, (2) the error of importance sampling based on the Laplace approximation decreases with the noise-level, (3) shifted lattice rules with random shifts, when transformed using the Laplace approximation, have a nondegenerating relative error for computing the normalization constant of the posterior.

References

[1] D. Rudolf and B. Sprungk, *On a generalization of the preconditioned Crank-Nicolson Metropolis algorithm*, Found. Comput. Math. **18**:2 (2018), 309–343.

[2] C. Schillings, B. Sprungk and P. Wacker. On the Convergence of the Laplace Approximation and Noise-Level-Robustness of Laplace-based Monte Carlo Methods for Bayesian Inverse Problems. arXiv:1901:03958 (2019).

Multilevel Monte Carlo for Efficient Risk Estimation

Abdul-Lateef Haji-Ali
(joint work with Michael B. Giles)

We investigate the problem of computing a nested expectation of the form $\mathbb{P}[\mathbb{E}[X|Y] \geq 0] = \mathbb{E}[\mathrm{H}(\mathbb{E}[X|Y])]$ where H is the Heaviside function. This nested expectation appears, for example, when estimating the probability of a large loss and other risk measures of a financial portfolio. The typical method is a "Uniform Monte Carlo" [4] method where the outer expectation is estimated with M i.i.d. samples of Y and the inner expectation is estimated with N i.i.d. samples of X conditioned on each value of Y, i.e.,

$$\mathbb{E}[\mathrm{H}(\mathbb{E}[X|Y])] \approx \frac{1}{M}\sum_{m=1}^{M} \mathrm{H}\Big(\widehat{\mathrm{E}}_N\Big(Y^{(m)}\Big)\Big).$$

where $Y^{(m)}$ is the m^{th} sample of Y and

$$\widehat{\mathrm{E}}_N(y) \coloneqq \frac{1}{N}\sum_{n=1}^{n} X^{(n)}(y) \approx \mathbb{E}[X \mid Y = y]$$

where $X^{(n)}(y)$ is the n^{th} sample of X given $Y = y$. To achieve a root mean-squared error ε, the computational complexity of Uniform Monte Carlo is $\mathcal{O}(\varepsilon^{-3})$. On the other hand, Multilevel Monte Carlo [2] uses correlated estimators with increasing number of inner samples to reduce the overall computational complexity by reducing the number of Y samples with a large number of inner samples. The MLMC estimator can be written as

$$\sum_{\ell=0}^{L} \frac{1}{M_\ell}\sum_{m=1}^{M_\ell} \mathrm{H}\Big(\widehat{\mathrm{E}}_{N_\ell}\Big(Y^{(\ell,m)}\Big)\Big) - \mathrm{H}\Big(\widehat{\mathrm{E}}_{N_{\ell-1}}\Big(Y^{(\ell,m)}\Big)\Big).$$

where $N_\ell = N_0 2^\ell$ for some $N_0 > 0$ and $\mathrm{H}\Big(\widehat{\mathrm{E}}_{N_{-1}}(\cdot)\Big) \coloneqq 0$. Here again $\{Y^{(\ell,m)}\}_{\ell,m}$ are i.i.d. samples of Y. The cost of computing $\mathrm{H}\Big(\widehat{\mathrm{E}}_{N_\ell}\big(Y^{(\ell,m)}\big)\Big)$ is $\mathcal{O}(N_\ell) = \mathcal{O}(2^\ell)$ and one can show that

$$\mathbb{E}\Big[\mathrm{H}(\widehat{\mathrm{E}}_{N_\ell}(Y)) - \mathrm{H}(\mathbb{E}[X \mid Y])\Big] = \mathcal{O}(2^{-\ell})$$

$$\text{and} \quad \mathrm{Var}\Big[\mathrm{H}(\widehat{\mathrm{E}}_{N_\ell}(Y)) - \mathrm{H}(\mathbb{E}[X \mid Y])\Big] = \mathcal{O}(2^{-\ell/2}),$$

cf. [3, 4]. Hence, the optimal computational complexity of the MLMC method is $\mathcal{O}(\varepsilon^{-5/2})$, cf. [2].

The number of inner samples need not be large for certain values of Y for which $|\mathbb{E}[X \mid Y]|$ is large since even a rough approximation $\widehat{\mathrm{E}}_{N_\ell}(Y) \approx \mathbb{E}[X \mid Y]$ would still yield $\mathrm{H}(\widehat{\mathrm{E}}_{N_\ell}(Y)) = \mathrm{H}(\mathbb{E}[X \mid Y])$ exactly. This is the idea of adaptive sampling [1], where the number of inner samples N_ℓ is not chosen uniformly with respect to all samples of Y. Instead, for a given value of Y and depending on relatively cheap Monte Carlo approximations of $\mathbb{E}[X \mid Y]$ and $\mathrm{Var}[X \mid Y]$, a small number of inner samples $N_\ell(Y)$ is sampled which still maintains a high accuracy of the estimate $\mathrm{H}\Big(\widehat{\mathrm{E}}_{N_\ell(Y)}(Y)\Big)$ with high probability. In particular, we prove that there is a such choice of $N_\ell(Y)$ satisfying

$$\mathbb{E}[N_\ell(Y)] = \mathcal{O}(2^\ell)$$

$$\text{and} \quad \mathrm{Var}[\mathrm{H}(\widehat{\mathrm{E}}_{N_\ell(Y)}(Y)) - \mathrm{H}(\mathbb{E}[X \mid Y])] = \mathcal{O}(2^{-\ell}).$$

The result is that an MLMC method with such a choice for the number of inner samples has an average computational complexity of $\mathcal{O}(\varepsilon^{-2}|\log\varepsilon|^2)$.

The theoretical analysis is presented in this talk and is verified by numerical experiments on a simple model problem. We also discuss a stochastic root-finding algorithm that, combined with the presented MLMC methods, can be used to compute other risk measures such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), with the latter being achieved with $\mathcal{O}(\varepsilon^{-2})$ computational complexity.

References

[1] M. Broadie, Y. Du, and C. C. Moallemi. Efficient risk estimation via nested sequential simulation. *Management Science*, 57(6):1172–1194, 2011.
[2] M. B Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.
[3] M. B. Giles and A.-L. Haji-Ali. Multilevel nested simulation for efficient risk estimation. *SIAM/ASA Journal on Uncertainty Quantification* (to appear) arXiv:1802.05016.
[4] M. B G. and Sandeep Juneja. Nested simulation in portfolio risk measurement. *Management Science*, 56(10):1833–1848, 2010.

Pricing American options by exercise rate optimization and markovian projections

Tempone, Raúl

(joint work with Christian Bayer, Juho Häppolä, Sören Wolfers)

This talk addresses the problem of pricing American basket options in a multivariate setting, which includes among others, the Bachelier and the Black-Scholes models. In high dimensions, nonlinear partial differential equation methods for solving the problem become prohibitively costly due to the curse of dimensionality.

In the first part, we present a novel method [1] for the numerical pricing of American options based on Monte Carlo simulation and the optimization of exercise strategies. Previous solutions to this problem either explicitly or implicitly determine so-called optimal *exercise regions*, which consist of points in time and space at which a given option is exercised. In contrast, our method determines the *exercise rates* of randomized exercise strategies. We show that the supremum of the corresponding stochastic optimization problem provides the correct option price. By integrating analytically over the random exercise decision, we obtain an objective function that is differentiable with respect to perturbations of the exercise rate even for finitely many sample paths. The global optimum of this function can be approached gradually when starting from a constant exercise rate. Numerical experiments on vanilla put options in the multivariate Black–Scholes model and a preliminary theoretical analysis underline the efficiency of our method, both with respect to the number of time-discretization steps and the required number of degrees of freedom in the parametrization of the exercise rates. Finally, we demonstrate the flexibility of our method through numerical experiments on max call options in the classical Black–Scholes model, and vanilla put options in both the Heston model and the non-Markovian rough Bergomi model.

In the last part of the presentation, we proposed to use a stopping rule that depends on the dynamics of a low-dimensional Markovian projection of the given basket of assets [2]. It is shown that the ability to approximate the original value function by a lower-dimensional approximation is a feature of the dynamics of the system and is unaffected by the path-dependent nature of the American basket option. Assuming that we know the density of the forward process and

using the Laplace approximation, we first efficiently evaluate the diffusion coefficient corresponding to the low-dimensional Markovian projection of the basket. Then, we approximate the optimal early-exercise boundary of the option by solving a Hamilton-Jacobi-Bellman partial differential equation in the projected, low-dimensional space. The resulting near-optimal early-exercise boundary is used to produce an exercise strategy for the high-dimensional option, thereby providing a lower bound for the price of the American basket option. A corresponding upper bound is also provided. These bounds allow to assess the accuracy of the proposed pricing method. Indeed, our approximate early-exercise strategy provides a straightforward lower bound for the American basket option price. Following a duality argument due to Rogers, we derive a corresponding upper bound solving only the low-dimensional optimal control problem. Numerically, we show the feasibility of the method using baskets with dimensions up to fifty. In these examples, the resulting option price relative errors are only of the order of few percent.

References

[1] C. Bayer, R. Tempone and S. Wolfers, *Pricing American Options by Exercise Rate Optimization*, *arXiv:1809.07300* (2018).

[2] C. Bayer, J. Häppölä and R. Tempone, *Implied Stopping Rules for American Basket Options from Markovian Projection*, Quantitative Finance, **19**(3) (2019), 371–390.

Bernstein – von Mises theorem and nonregular constrained ill-posed models

Natalia Bochkina

(joint work with Peter J. Green)

A broad class of statistical models is considered that can be misspecified and ill-posed, under a positivity constraint, from a Bayesian perspective. This provides a flexible and interpretable framework for their analysis, but it is important to understand robustness of the chosen Bayesian model and its effect on the resulting solution, especially in the ill-posed case where in the absence of prior information the solution is not unique. We address this problem by studying the local behaviour of the posterior distribution around the point of its concentration as the observation error decrease, from the frequentist perspective. The results apply to misspecified models which allows, for instance, to evaluate the effect of model approximation on statistical inference. Emission tomography is taken as a canonical example for study, but our results hold for a wider class of generalised linear inverse problems with constraints. This work extends the results of [1] to ill-posed models.

References

[1] Natalia A. Bochkina and Peter J. Green *The Bernstein-von Mises theorem and non-regular models*, Annals of Statistics (2014), V. 42, N. 5, 1850-1878.

Prior aware Metropolis-Hastings for posterior measures with non-Gaussian priors

Bamdad Hosseini

Markov Chain Monte Carlo (MCMC) algorithms are perhaps the most common methods for extraction of information from high-dimensional probability measures with wide spread applications in uncertainty quantification, applied probability and statistics. The main focus of my talk was the development of a class of MCMC algorithms for probability measures on Banach spaces that are absolutely continuous with respect to non-Gaussian measures. Almost all MCMC algorithms in the literature somehow use a Gaussianity assumption which is very restrictive in high dimensions. As far as I know my algorithm is the first of its kind that does not use any assumptions of Gaussianity.

Often times in practice we need to simulate samples from a target probability measures ν on a Banach space X. For example suppose we need to approximate integrals of the form

$$\int_X f(u)d\nu(u), \tag{1}$$

for some function $f : X \mapsto \mathbb{R}$. If X is high dimensional we can't use standard quadrature. Now if we had access to a sequence of random variables $\{u^{(j)}\}_{j=1}^{\infty}$ that were distributed according to ν then we could approximate (1) by

$$\frac{1}{N}\sum_{j=1}^{N} f(u^{(j)}),$$

for some large N. The goal of MCMC is precisely to generate such a sequence of random variables. In fact, MCMC algorithms generate a Markov chain $\{u^{(j)}\}_{j=1}^{\infty}$ so that the law of $u^{(j)}$ converges in some sense to ν as $j \to \infty$.

In my talk I introduced a new class of MCMC algorithms under the subcategory of Metropolis-Hastings (MH) algorithms as outlined in my preprint [1]. I started by assuming that the target measure ν is given in terms of an underlying prior measure μ via its Radon-Nikodym derivative

$$\frac{d\nu}{d\mu}(u) = \frac{1}{Z}\exp(-\Psi(u)),$$

where Z is a normalizing constant and $\Psi : X \mapsto \mathbb{R}$ is the negative log-density of ν with respect to μ or simply referred to as "the potential". I refer to μ as the prior in analogy with Bayesian inverse problems [2].

Then our ultimate goal is to construct a Markov transition kernel $\mathcal{P}$ (viewed as an operator on probability measures) so that

$$\mathcal{P}\nu = \nu, \qquad \text{and} \qquad d(\mathcal{P}^n\delta_{u^{(0)}}, \nu) \leq \gamma^n d(\delta_{u^{(0)}}, \nu), \tag{2}$$

where $d : P(X) \times P(X) \mapsto \mathbb{R}$ is an appropriate distance on $P(X)$ the space of probability measures on X, $\gamma \in (0, 1)$ is a uniform constant and $\delta_{u^{(0)}}$ is a point

mass at some initial point $u^{(0)} \in X$. Basically, if (2) holds then $u^{(j)} = \mathcal{P}^j \delta_{u^{(0)}}$ forms a Markov chain that can be used to approximate the integral in (1).

I introduced such a kernel using the MH recipe by assuming the form

$$\mathcal{P}(u, dv) = \mathcal{Q}(u, dv)\alpha(u, v) + d\delta_u(v) \int_X (1 - \alpha(u, w))\mathcal{Q}(u, dw),$$

where $\mathcal{Q}$ is the MH proposal kernel (it is a Markov transition kernel itself) and $\alpha : X \times X \mapsto [0, 1]$ is the acceptance ratio. The above definition of $\mathcal{P}$ gives rise to the well-known MH algorithm:

(1) Start with some point $u^{(0)}$ and set $j = 1$.
(2) At step j propose $v \sim \mathcal{Q}(u^{(j-1)}, dv)$ conditioned on $u^{(j-1)}$.
(3) Accept v as a new point in the Markov chain with probability $\alpha(u^{(j-1)}, v)$ and set $u^{(j)} = v$.
(4) Otherwise set $u^{(j)} = u^{(j-1)}$.
(5) Set $j = j + 1$ and go to Step 2.

The basic idea of my algorithm is that if we pick the right $\mathcal{Q}$ and α then the resulting $\mathcal{P}$ will have the desirable properties in (2). In fact, I show in [1] and together with J. E. Johndrow in [3] that if we take

$$\alpha(u, v) = \min\{1, \exp(\Psi(u) - \Psi(v))\}$$

and pick $\mathcal{Q}$ so that

$$\mathcal{Q}(u, dv)d\mu(v) = \mathcal{Q}(v, du)d\mu(u),$$

in the sense of probability measures on $X \times X$, then $\mathcal{P}$ satisfies (2) (we need a few more technical assumptions for this to hold that I will skip for brevity). Note that the equation for $\mathcal{Q}$ only depends on the prior μ and does not involve the target ν, hence the name "prior aware Metropolis-Hastings".

Choosing the correct $\mathcal{Q}$ is often quite difficult in practice so I spent the rest of my talk introducing a concrete numerical example from [1] where I derive the kernel $\mathcal{Q}$ for a certain type of prior μ. I won't discuss the details of the example here since it can be accessed in the preprint but it involves recovering an unknown function from its convolution with a smoothing kernel. This is a classic inverse problem called deconvolution which is often used for benchmarking in the inverse problems literature. I showed that my algorithm works very well with some classes of highly non-Gaussian priors based on the gamma distribution even in cases where X is a function space and therefore infinite-dimensional.

References

[1] B. Hosseini, *Two Metropolis-Hastings algorithms for posterior measures with non-Gaussian priors in infinite dimensions*, arXiv:1804.07833, (2018).
[2] A. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numerica, **19** (2010) 451–559.
[3] B. Hosseini and J. E. Johndrow, *Convergence and perturbation theory for an infinite-dimensional Metropolis-Hastings algorithm with self-decomposable priors*, arXiv:1810.00297, (2018).

Measuring Sample Quality with Diffusions

Lester Mackey
(joint work with Jackson Gorham, Andrew B. Duncan, Sebastian J. Vollmer)

Stein's method for measuring convergence to a continuous target distribution relies on an operator characterizing the target and Stein factor bounds on the solutions of an associated differential equation. While such operators and bounds are readily available for a diversity of univariate targets, few multivariate targets have been analyzed. In [1], we introduce a new class of characterizing operators based on Ito diffusions and develop explicit multivariate Stein factor bounds for any target with a fast-coupling Ito diffusion. As example applications, we develop computable and convergence-determining diffusion Stein discrepancies for log-concave, heavy-tailed, and multimodal targets and use these quality measures to select the hyperparameters of biased Markov chain Monte Carlo (MCMC) samplers, compare random and deterministic quadrature rules, and quantify bias-variance tradeoffs in approximate MCMC. Our results establish a near-linear relationship between diffusion Stein discrepancies and Wasserstein distances, improving upon past work even for strongly log-concave targets. The exposed relationship between Stein factors and Markov process coupling may be of independent interest.

References

[1] Gorham, J. and Duncan, A. and Vollmer, S. and Mackey, L., *Measuring Sample Quality with Diffusions*, Annals of Applied Statistics, in press.

Optimal Experimental Design Problem as Mixed-integer Optimal Control Problem

Ekaterina Kostina
(joint work with Hans Georg Bock)

We consider a dynamic process, that is described by a system of ordinary differential equations (ODE)

$$\dot{x} \quad = \quad f(x(t), u(t), p),\ t \in [0, T], x(0) = x_0(p),$$

where $x : [0, T] \to \mathbb{R}^{n_p}$ denotes the states of the process, $p \in \mathbb{R}^{n_p}$ is a vector of unknown parameters, $u : [0, T] \to\in \mathbb{R}^{n_u}$ denotes the vector of control functions, which describe experimental conditions. Further, the states and parameters may satisfy possible boundary or interior type constraints.

We assume that at times $0 \le t_1 < t_2 < ... < t_M$ we get the measurement data η from the measurement devices. The measurements are assumed to be equal to the model response $h(\cdot)$ subject to measurement errors, which are normally distributed with zero mean and known variances: $\eta_{ij} = h_i(x(t_j)) + \varepsilon_{ij}$, $\varepsilon_{ij} \in \mathcal{N}(0, \sigma_{ij}^2)$, $i = 1, ..., n_h$, $j = 1, ..., M$.

In order to estimate the unknown parameters p we solve the following parameter estimation problem, for given controls $u(\cdot)$ and *sampling design* ω_{ij}

$$\begin{aligned} \min_{x(\cdot),p} \quad & \frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{n_h}\omega_{ij}\frac{\big(\eta_{ij}-h_i(x(t_j))\big)^2}{\sigma_{ij}^2},\ \omega_{ij}\in\{0,1\},\forall i,j, \\ \text{s.t.} \quad & \dot{x}=f(x(t),u(t),p),\ t\in[0,T],\ x(0)=x_0(p). \end{aligned}$$

Here, the *sampling* design variables ω_{ij} indicate at when the measurements should be taken and by which measurement device.

For a given control function $u(\cdot)$ and given sampling design ω_{ij} we evaluate the information matrix $F\in\mathbb{R}^{n_p\times n_p}$:

$$F(T) \quad := \quad \sum_{j=1}^{M}\sum_{i=1}^{n_h}\omega_{ij}y(t_j)y(t_j)^T,\ y(t_j)=\left[\frac{\partial h_i}{\partial x}(x(t_j))G^p(t_j)\right]^T,$$

where the derivative $\frac{\partial x(t_j)}{\partial p}=:G^p(t_j)$ solves the variational differential equation:

$$\dot{G^p} \quad = \quad \frac{\partial f}{\partial x}(x(t),u(t),p)G^p(t)+\frac{\partial f}{\partial p}(x(t),u(t),p), G^p(0)=0,$$

or the covariance matrix $C(T)=F(T)^{-1}$.

Now, we can formulate the optimal experimental design (OED) problem: Determine sampling design variables ω_{ij} and experimental control functions $u(\cdot)$ by solving the nonlinear mixed integer control problem (MIOCP):

$$\begin{aligned} \max_{x(\cdot),u(\cdot),\omega} \quad & \Phi(F(T)) \text{ resp. } \min_{x(\cdot),u(\cdot),\omega}\Psi(C(T)), \\ \text{s.t.} \quad & \dot{x}=f(x,u,p), x(0)=x_0(p), \\ & F(T)=\sum_{j=1}^{M}\sum_{i=1}^{n_h}\omega_{ij}y(t_j)y(t_j)^T, y(t_j)=\left[\frac{\partial h_i}{\partial x}(x(t_j))G^p(t_j)\right]^T, \\ & \text{resp.} \\ & C(T)=-\sum_{j=1}^{M}\sum_{i=1}^{n_h}\omega_{ij}z(t_j)z(t_j)^T, z(t_j)=\left[\frac{\partial h_i}{\partial x}(x(t_j))G^p(t_j)C(t_j)\right]^T, \\ & \dot{G^p}=\frac{\partial f}{\partial x}(x(t),u(t),p)G^p(t)+\frac{\partial f}{\partial p}(x(t),u(t),p),\ G^p(0)=0, \\ & \omega_{ij}\in\{0,1\},\forall i,j,\ \text{ further possible constraints on } \omega_{ij},u(\cdot),x(\cdot). \end{aligned}$$

The usual choice of the cost function Φ resp. Ψ is the trace, minimal resp. maximal eigenvalue or minimal resp. maximal diagonal element of the corresponding matrix.

Although OED is a complex non-standard nonlinear mixed-integer optimal control problem, the efficient numerical methods have been developed, see, e.g., [1]. The characteristic features of these numerical methods are that the methods are structure exploiting, direct, all-at-once optimization algorithms for dynamic models, based on multiple shooting. Treatment of integer sampling decisions can be

performed by very efficient methods for nonlinear mixed integer optimal control problems, e.g., [5].

Our aim is to further improve the developed numerical methods. That is why we are interested in an asymptotic analysis of the OED problem from the point of view of the Pontryagin's Maximum Principle. For this purpose, we consider a "relaxed" version of a sampling design problem. The relaxed cost function in the parameter estimation problem now takes the form: $\min \frac{1}{2} \int_0^T \sum_{i=0}^{n_h} w_i(t) \frac{\left(\eta_i(t) - h_i(x(t))\right)^2}{\sigma_i(t)} dt$. Here, we have introduced a new control $w_i(t)$ which is the *derivative* of the relaxed sampling design ω_{ij}.

The corresponding relaxed OED control problem reads: Find $x(\cdot)$, $u(\cdot)$, $w(\cdot)$ such that

$$\begin{aligned} \max \quad & \Phi(F(T)) \text{ resp. } \min \Psi(C(T)), \\ \text{s.t.} \quad & \dot{x} = f(x,u,p), x(0) = x_0(p), \end{aligned}$$

$$\dot{G}^p = \frac{\partial f}{\partial x}(x(t),u(t),p)G^p(t) + \frac{\partial f}{\partial p}(x(t),u(t),p),\ G^p(0) = 0,$$

$$\dot{F} = \sum_{i=1}^{n_h} w_i(t) y_i(t) y_i(t)^T, F(0) = 0,\ y_i(t) = \left[\frac{\partial h_i}{\partial x}(x(t))G^p(t)\right]^T,$$

$$\text{resp. } \dot{C} = -\sum_{i=1}^{n_h} w_i(t) z_i(t) z_i(t)^T, C(0) = C_0,\ C_0 \text{ a priori information},$$

$$z_i(t) := \left[\frac{\partial h_i}{\partial x}(x(t))G^p(t)C(t)\right]^T, \text{ further constraints on } w(\cdot), u(\cdot), x(\cdot).$$

In this problem, the dynamic controls $u(t)$ are usually nonlinear; but the new "sampling" controls $w(t)$ enters linearly the right hand side of the ODE system. Since the cost function (e.g. the trace or the maximal eigenvalue) is monotonous in w, additional *constraints* on the sampling design $w(\cdot)$ are *required* in order to avoid unboundedness of the cost function.

The most obvious restrictions are constraints on the measurement costs, e.g., $\int_0^T \sum_i^{n_h} \text{cost}_i(t) w_i(t) dt \leq M$, which can be also interpreted as the l_1 penalty on $w(\cdot)$.

Application of the Pontryagin's Maximum Principle yields the conclusion, thar optimal sampling design $w(\cdot)$ has a finite support on $[0,T]$, which is in agreement with the theoretical results, which state, namely, that at most $n_p(n_p+1)/2$ measurement points are necessary to get an optimal information or covariance matrix, see, e.g., [2, 3, 4, 6]. In other words, natural l_1 constraint resp. penalization leads to sparse sampling design. Using the Pontryagin's Maximum Principle allows to understand the structure of the sampling design, thus we may further

improve numerical methods by new strategies for computing intelligent approximation strategies of integer sampling design and by adaptive refinement of design grids.

References

[1] Bauer, I., Bock, H.G., Körkel, S., and Schlöder, J.P., *Numerical methods for optimum experimental design in DAE systems*, Journal of Computational and Applied Mathematics, **120**(2000), 1 – 25.

[2] La, H., Potschka, A., and Bock, H., *Partial stability for nonlinear model predictive control*, Automatica, **78** (2017), 14 – 19.

[3] Neitzel, I., Pieper, K., Vexler, B., and Walter D., *A sparse control approach to optimal sensor placement in PDE constrained parameter estimation problems*, submitted, 2018.

[4] Pukelsheim, F., *Optimal design of experiments*, John Wiley & Sons Inc., New York, 1993. Sager, S., *Numerical methods for mixed-integer optimal control problems*, PhD Thesis, Ruprecht-Karls-Universität Heidelberg, Heidelberg, 2006.

[5] Sager, S., *Numerical methods for mixed-integer optimal control problems*, PhD Thesis, Ruprecht-Karls-Universität Heidelberg, 2006.

[6] Ucinski, D., *Optimal measurement methods for distributed parameter system identification*, Systems and Control Series, CRC Press, Boca Raton, FL, 2005.

Reporter: Jon Cockayne

Participants

Prof. Dr. Sergios Agapiou
Department of Mathematics and Statistics
University of Cyprus
1 University Avenue
2109 Nicosia
CYPRUS

Prof. Dr. Markus Bachmayr
Institut für Mathematik
Johannes-Gutenberg Universität Mainz
Staudingerweg 9
55128 Mainz
GERMANY

Prof. Dr. Andrea Barth
Institut für Angewandte Analysis und Numerische Simulation
Abteilung für Computational Methods for
Uncertainty Quantification
Universität Stuttgart
Allmandring 5b
70569 Stuttgart
GERMANY

Dr. Natalia Bochkina
School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
King's Buildings, Mayfield Road
Edinburgh EH9 3FD
UNITED KINGDOM

Prof. Dr. Alexey Chernov
Institut für Mathematik
Carl von Ossietzky Universität Oldenburg
26111 Oldenburg
GERMANY

Prof. Dr. Oxana A. Chkrebtii
Department of Statistics
Ohio State University
323 Cockins Hall
1958 Neil Avenue
Columbus OH 43210-1247
UNITED STATES

Dr. Jon Cockayne
Department of Statistics
University of Warwick
Zeeman Building
Coventry CV4 7AL
UNITED KINGDOM

Dr. Masoumeh Dashti
Department of Mathematics
University of Sussex
Falmer
Brighton BN1 9RH
UNITED KINGDOM

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Dr. Jana de Wiljes
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Straße 24-25
14476 Potsdam
GERMANY

Dr. Ana Djurdjevac
ZEDAT
Freie Universität Berlin
Fabeckstrasse 32
14195 Berlin
GERMANY

Dr. Svetlana Dubinkina
Centrum voor Wiskunde en Informatica
Science Park 123
1098 XG Amsterdam
NETHERLANDS

Dr. Matthew Dunlop
Courant Institute of Mathematical Sciences
New York University
251, Mercer Street
New York, NY 10012
UNITED STATES

Dr. Martin Eigel
Weierstraß-Institut für Angewandte Analysis und Stochastik
Mohrenstrasse 39
10117 Berlin
GERMANY

Prof. Dr. Oliver Ernst
Fakultät für Mathematik
Technische Universität Chemnitz
Postfach 964
09009 Chemnitz
GERMANY

Lisa Fischer
Konrad Zuse Institute Berlin (ZIB)
Numerical Mathematics
Computational Systems Biology
Takustraße 7
14195 Berlin
GERMANY

Prof. Dr. Colin Fox
Department of Physics
University of Otago
P.O. Box 56
Dunedin 9054
NEW ZEALAND

Prof. Dr. Hanno Gottschalk
FB C: Mathematik und Naturwissenschaften
Bergische Universität Wuppertal
Gaußstrasse 20
42119 Wuppertal
GERMANY

Prof. Dr. Heikki Haario
Department of Mathematics and Physics
Lapeenranta University of Technology
Box 20
53851 Lappeenranta
FINLAND

Dr. Abdul-Lateef Haji Ali
Mathematical Institute
Oxford University
Woodstock Road
Oxford OX1 3LB
UNITED KINGDOM

Dr. Bamdad Hosseini
Department of Mathematics
California Institute of Technology
Pasadena, CA 91125
UNITED STATES

Dr. Yoshihito Kazashi
EPFL SB MATH CSQI
MA B2 434 (Bâtiment MA)
Station 8
1015 Lausanne
SWITZERLAND

Dr. Hanne Kekkonen
Department of Applied Mathematics and Theoretical Physics (DAMTP)
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
UNITED KINGDOM

Dr. Ilja Klebanov
Konrad-Zuse-Zentrum für
Informationstechnik Berlin (ZIB)
Takustrasse 7
14195 Berlin
GERMANY

Prof. Dr. Ekaterina A. Kostina
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Dr. Drew Philip Kouri
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
UNITED STATES

Prof. Frances Y. Kuo
School of Mathematics and Statistics
The University of New South Wales
Sydney NSW 2052
AUSTRALIA

Dr. Annika Lang
Department of Mathematical Sciences
Chalmers University of Technology and
the
University of Gothenburg
412 96 Göteborg
SWEDEN

Jonas Latz
Zentrum Mathematik
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Dr. Han Cheng Lie
Konrad-Zuse-Zentrum für
Informationstechnik Berlin (ZIB)
Takustrasse 7
14195 Berlin
GERMANY

Prof. Dr. Lester Mackey
Department of Mathematics
Stanford University
Stanford, CA 94305-2125
UNITED STATES

Prof. Dr. Youssef M. Marzouk
Department of Aeronautics and
Astronautics
Massachusetts Institute of Technology
Room 37-451
77 Massachusetts Avenue
Cambridge MA 02139
UNITED STATES

Prof. Dr. Richard Nickl
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
UNITED KINGDOM

Prof. Dr. Fabio Nobile
CSQI - MATH
École Polytechnique Fédérale de
Lausanne
Station 8
1015 Lausanne
SWITZERLAND

Dr. Michela Ottobre
Department of Mathematics
Heriot-Watt University
Riccarton
Edinburgh EH14 4AS
UNITED KINGDOM

Prof. Dr. Houman Owhadi
Department of Applied Mathematics
California Institute of Technology
Rm. 217-50
Pasadena, CA 91125
UNITED STATES

Prof. Dr. Alois Pichler
Fakultät für Mathematik
Technische Universität Chemnitz
Postfach 964
09009 Chemnitz
GERMANY

Prof. Dr. Sebastian Reich
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Straße 24-25
14476 Potsdam
GERMANY

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Dr. Daniel Rudolf
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Robert Scheichl
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Martin Schlather
Lehrstuhl für Stochastik und ihre Anwendungen
Universität Mannheim
68131 Mannheim
GERMANY

Prof. Dr. Christoph Schwab
Seminar für Angewandte Mathematik
ETH Zentrum, HG G 57.1
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Anna Simoni
École Nationale de la Statistique et de l'Administration Économique
5, Avenue Henry Le Châtelier
91764 Palaiseau Cedex
FRANCE

Dr. Björn Sprungk
Institut fü Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Claudia Strauch
Fakultät für Mathematik
Universität Mannheim
68131 Mannheim
GERMANY

Prof. Dr. Tim Sullivan
Fachbereich Mathematik und Informatik
Freie Universität Berlin
Arnimallee 6
14195 Berlin
GERMANY

Dr. Aretha Teckentrup
School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
Edinburgh EH9 3FD
UNITED KINGDOM

Prof. Dr. Raul F. Tempone
4700 King Abdullah University of Science
and Technology (KAUST)
Mathematical and Computer Sciences
Thuwal 23955-6900, Jeddah
SAUDI ARABIA

Prof. Dr. Elisabeth Ullmann
Zentrum Mathematik
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Andi Q. Wang
Department of Statistics
University of Oxford
24-29 St. Giles
Oxford OX1 3LB
UNITED KINGDOM

Simon Weißmann
Fakultät für Mathematik und Informatik
Universität Mannheim
68131 Mannheim
GERMANY