Mathematisches Forschungsinstitut Oberwolfach

Report No. 10/2025

DOI: 10.4171/OWR/2025/10

Data Assimilation: From Mathematical and Statistical Foundations to Applications

Organized by
Jana de Wiljes, Ilmenau
Youssef Marzouk, Cambridge MA
Aretha Teckentrup, Edinburgh

23 February – 28 February 2025

Abstract. Data assimilation, where predictions from a dynamical system are updated sequentially based on new and incomplete observations, is increasingly finding applications in many areas of science and technology. This workshop brought together a collection of scientists from dynamical systems, statistics, machine learning, applied probability, uncertainty quantification, and mathematical modelling, as well as practitioners in the field.

Mathematics Subject Classification (2020): 37M10, 62F15, 62L10, 65K10, 65M32.

Introduction by the Organizers

The recent explosion of available data, driven by the increase in large-scale scientific experiments and the development of sensor technology, means that there is a pressing need to develop new algorithms for the seamless integration of observed data with sophisticated mathematical models. We require tools to inform decisions, assess risk, and formulate policies based on available evidence.

Problems of interest in applications mathematically often fall into the category of inverse problems, where one is interested in learning parameters such as physical quantities or initial conditions from noisy indirect observations, or data assimilation, where forecasts from a dynamical system are updated sequentially (and recursively) based on new partial observations. Whereas purely data-driven approaches are suitable in application areas where little is known about the processes generating the data, many application areas such as geological exploration,

climate and weather predictions, and personalized medicine require the integration of complex mathematical models with the observed data to provide accurate inferences.

While the term inverse problems usually refers to static settings, where the unknown to be inferred is a fixed parameter, data assimilation is used to refer to dynamic problems, where the unknown to be inferred is the time-evolving state of a dynamical system. Yet the two areas are intricately linked, and share many methodologies and challenges. In fact, inverse problems can be viewed as a special case of data assimilation, where the latter crucially adds dynamical complexity and often the need for online and recursive algorithms. Focusing on the common challenges and opportunities, this workshop brought together researchers working in areas crucial to advancements in data assimilation and inverse problems, including Bayesian inference, Monte Carlo methods, non-linear filtering, dynamical systems, reduced order modelling, and optimal transport.

The workshop was organized by organized by Jana de Wiljes (Ilmenau, Germany), Youssef Marzouk (Cambridge MA, USA), and Aretha Teckentrup (Edinburgh, UK). The meeting was attended by 22 participants, and represented a broad range of mathematical subject areas as well as numerous application areas from the natural sciences. The workshop is a highlight in the calendar of events in this area and was enthusiastically endorsed by all participants. The field of data assimilation has undergone major developments since the last MFO workshop on this topic in 2022. We mention in particular an emerging strong interplay between data assimilation and machine learning and mathematical statistics. A further current hot topic has been optimal transport and its interactions with computational statistics. The strong trend towards novel applications in, e.g., pharmacology, cognitive science, space weather and biology continued. A total of 16 talks were presented during the workshop. The talks were selected such as to cover the interplay between data assimilation and machine learning (Marc Bouquet, Alberto Carassi, Oana Lang, Sven Wang), novel mathematical developments on data assimilation algorithms (Joaquin Miguez, Hans Reimann, Daniel Sanz-Alonso, Xin Tong, Peter Jan van Leeuwen), theoretical and practical aspects of dynamical systems (Nisha Chandramoorthy, Olga Mula, Sahani Phatiraja, Elisabeth Ullmann), computational methods for Bayesian inference and their theoretical analysis (Elliot Addy, Aimee Maurais, Benjamin Zanger).

On Tuesday evening, there was a discussion group triggered by the many different facets of interactions between data assimilation and machine learning presented during the talks in the first days of the workshop. Participants discussed how the two fields can best collaborate and benefit each other, which lead to a more general discussion about the future of the field of data assimilation and initiatives that could help propel the field forward. In particular, it was discussed that a shared website, sharing among others things code written by researchers in the field, would benefit early career researchers and enable a more thorough comparison between various methodologies.

Workshop: Data Assimilation: From Mathematical and Statistical Foundations to Applications

Table of Contents

Abstracts

Lengthscale-informed sparse grids for high-dimensional Gaussian process emulation

Elliot Addy

(joint work with Jonas Latz and Aretha Teckentrup)

Gaussian process emulation is a popular method of surrogate modeling, in which the broad aim is to cheaply approximate outputs of computer models. Using data in the form of model runs, the mean function of a posterior Gaussian process is taken as an approximation for the parameter-to-output map, and the posterior covariance a measure of the uncertainty in these predictions. By drawing connections to scattered data approximation, we are able to develop error bounds for functions contained in the Native space of the chosen covariance kernel [1].

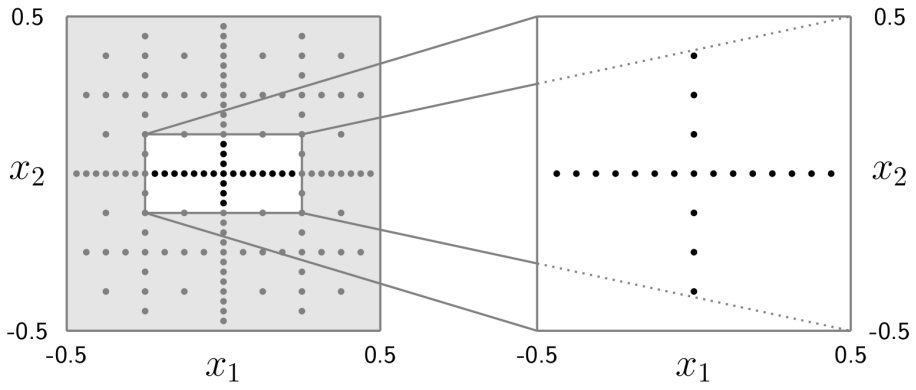


Figure 1. Level 4 isotropic sparse grid (left) versus level 4 lengthscale informed sparse grid (right). For functions of lengthscale $\boldsymbol{\lambda} = (2, 4)$ in the horizontal and vertical directions, respectively, both designs are shown to have comparable error estimates when used in conjunction with their associated Matern kernels.

Emulators, however, fall prey to the so called 'curse of dimensionality'; as we increase the dimension of parameter space, the number of model evaluations, in general, needs to increase exponentially in order to maintain the same error guarantees. Consequently, much work is done to exploit functions with known structure to mitigate this dimension dependence on the error. For example, sparse grid methods, when used in conjunction with product kernels, have been developed to efficiently approximate functions in Sobolev spaces with dominating mixed smoothness, with error bounds depending only logarithmically on the dimension [2]. Furthermore, due to an induced Kronecker structure, fast inversions of covariance matrices on sparse grid designs are possible [3]. Still, in practice, this

dimension dependence becomes prohibitive beyond $d \approx 10$, and as such, additional assumptions are required if we wish to consider higher dimensional settings. For this purpose, anisotropic methods have been developed. By accounting for structural anisotropy, dimension dependence in the rate of convergence can be further reduced - or even eliminated entirely [4, 5]. These approaches often require demanding smoothness conditions on the underlying model, and as such cannot be used in many circumstances in which anisotropy may be present.

In this work, we have developed a novel sparse grid construction, lengthscale-informed sparse grids (LISG), in which the aim is to instead exploit anisotropy in the lengthscale parameter of Matern kernel functions. By sampling on LISG designs, and employing appropriately stretched covariance kernels, we are able to emulate arbitrarily high dimensional functions when assuming the lengthscale grows sufficiently quickly with the dimension, without having to assume further regularity. In Figure 1, we see a comparison between standard isotropic and lengthscale-informed sparse grids used in kernel interpolation.

References

[1] H. Wendland, *Scattered Data Approximation*, Cambridge University Press (2004).
[2] F. Nobile, R. Tempone, S. Wolfers, *Sparse approximation of multilinear problems with applications to kernel-based methods in UQ*, Numerische Mathematik **139** (2016), 247–280.
[3] M. Plumlee, *Fast Prediction of Deterministic Functions Using Sparse Grid Experimental Designs*, Journal of the American Statistical Association **109**(508) (2014), 1581–1591.
[4] C. Rieger, H. Wendland, *Sampling inequalities for anisotropic tensor product grids*, IMA Journal of Numerical Analysis **40**(1) (2019), 285-321.
[5] F. Nobile, R. Tempone, C. G. Webster, *An Anisotropic Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data*, SIAM Journal on Numerical Analysis **46**(5) (2008), 2411-2442.

Are ensemble-based data assimilation methods really necessary for accurate filtering?

Marc Bocquet

We investigate the ability to discover data assimilation (DA) schemes meant for chaotic dynamics with deep learning. The focus is on learning the analysis step of sequential DA, from state trajectories and their observations, using a simple residual convolutional neural network, while assuming the dynamics to be known. Experiments are performed with low-order dynamics which display spatiotemporal chaos and for which solid benchmarks for DA performance exist. The accuracy of the states obtained from the learned analysis approaches that of the best possibly tuned ensemble Kalman filter, and is far better than that of variational DA alternatives. Critically, this can be achieved while propagating even just a single state in the forecast step. We investigate the reason for achieving ensemble filtering accuracy without an ensemble. We diagnose that the analysis scheme actually identifies key dynamical perturbations, mildly aligned with the unstable subspace, from the forecast state alone, without any ensemble-based covariances representation. This reveals that the analysis scheme has learned some multiplicative ergodic

theorem associated to the DA process seen as a non-autonomous random dynamical system. This also suggests building a new class of efficient deep learning-based ensemble-free DA algorithms.

Merging DA and ML at various degree: examples from DA for Arctic Sea ice and for ocean biogeochemistry

Alberto Carrassi

In recent years, data assimilation (DA), and more generally the climate science modelling enterprise have been influenced by the rapid advent of artificial intelligence, in particular machine learning (ML), opening the path to various form of ML-based methodology. In this talk we will schematically show how ML can be included in the prediction and DA workflow in different ways with various degrees of integration within each other. In a so-called "non-intrusive" ML, we will show how ML can be used to supplement a chaotic system and help predicting the local instabilities and/or abrupt regime's changes. DA and ML can also be placed side by side in an iterative approach alternating a DA step that assimilate sparse and noisy data, and a ML step whereby the data-driven model is further optimised against the analyses outputted from the DA. In a further level of fusion ML can finally be used to within hybrid ML-DA methods in which ML is used to cope with some limitations in DA approaches. In particular we shall show an innovative formulation of the EnKF that embodies a variational autoencoder enabling the EnKF to (i) handle non-Gaussian observations, and, (ii) respecting physical balances. Using a set of idealised model and observational scenarios, we will show numerical results for all of the above-mentioned possibilities. We will focus on, and will be motivated by, problems originated in diverse areas of climate science, namely chaotic systems such as the atmosphere and the highly nonlinear and non-Gaussian DA for Arctic Sea ice and ocean biogeochemistry.

Learning ergodic dynamics from data

Nisha Chandramoorthy
(joint work with Jeongjin Park and Youssef Marzouk)

We are interested in surrogate modeling of a dynamical system in a way that preserves an underlying physical measure. In certain chaotic systems, we prove that when Jacobian information is added to the loss function, regression for short-term dynamics leads to statistical accuracy, i.e., the surrogate models learned with first order derivatives of the short-term dynamics can provably sample the physical measure (long-term data distribution). In the second half of the talk, we study the problem of sampling from an unknown probability density in the presence of its *score* or gradient of log density. This method for sampling, called Score Operator Newton, is based on writing down an infinite-dimensional Newton-Raphson iteration for the zero of a score-residual operator. The method of derivation applies also

to target measures that do not have densities with respect to Lebesgue but have absolutely continuous conditionals on a lower-dimensional manifold. Such measures, e.g., appear as physical measures in certain chaotic dynamical systems. With this modified derivation, we thus connect the two halves of this talk, physical surrogate modeling of dynamical systems and dynamical algorithms for sampling/Bayesian inference together by discussing a notion of conditional score for these singular measures. We propose a recursive algorithm to compute these conditional scores, and then apply the abovementioned Newton-Raphson iteration on the unstable manifold for sampling from Bayesian filtering distributions. Continuing with dimension reduction for sampling in the dynamical systems context, we next present two preliminary ideas for dimension reduction of generative models. Overall, our collection of results for sampling and generative modeling for dynamical systems and using dynamical systems theory fall into the broad category of problems at the intersection of dynamics with statistical learning.

We consider a chaotic map F that preserves a physical measure μ, and a learned model $F_{\text{nn}} = \text{argmin}_h E_{x\sim\mu}\ell(x)$. A neural model F_{nn} obtained with the square loss, $\ell(x,h) = \|h(x) - F(x)\|^2$ can produce unphysical orbits in the long term, as illustrated on the Lorenz '63 model in Figure 1 (center). However, we consider a modified loss function to $\ell(x,h) = \|h(x) - F(x)\|^2 + \lambda\|dh(x) - dF(x)\|^2$, where df represents the differential map (Jacobian) of f at x. Now the neural models that generalize well under this loss, which we will refer to as the Jacobian loss, is able to reproduce the attractor accurately (Figure 1, right). We numerically verify that moments of various quantities, and the empirically computed Wasserstein distance between the generated samples and the samples from the Lorenz equations match accurately. Our main result is to explain this observation. Specifically, we prove that for the class of uniformly hyperbolic maps F, C^1 matching (training with Jacobian loss) with high probability leads to statistical accuracy when shadowing orbits are typical (that is, shadowing orbits distribute according to μ). On the other hand, we do not have a shadowing orbit existence guarantee for the mean squared loss. This result implies that we do not need to add more derivatives to the loss function to get statistical accuracy of learned models. When shadowing orbits are not typical, generalization error of the Jacobian loss being small also does not imply statistical accuracy. The result provides a theoretical justification for learning statistically accurate models with regression for the one-step function F.

We remark, in light of other results discussed in this workshop by Sanz-Alonso and Bocquet, that for DA applications, statistical accuracy of the surrogate models may not be needed. However, ensuring the ability of a learned model to sample from the true underlying physical measure will improve the reliability of learned climate models, and further, may lead to lesser effort in the analysis step of a DA procedure. In practice, the Jacobian is difficult to obtain and train with in high-dimensional models, and hence may be replaced with some partial Jacobian information, e.g., Jacobian-random vector products. Although our theoretical results use the full Jacobian, we have found in practice that such partial information

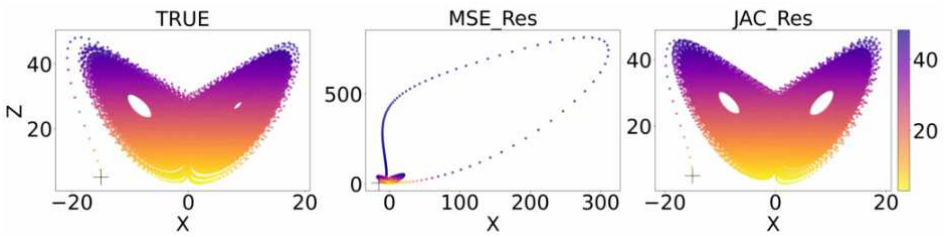


FIGURE 1. Here F is the time integration over one time step of the Lorenz '63 vector field at the standard choice of parameter values that leads to chaotic orbits. Left: true orbit of a random initial condition ('+' sign) simulated using RK4 time integration of the Lorenz '63 vector field. Center: an orbit of a neural model trained to approximate F with low mean squared error on test points. Right: an orbit generated by a neural network trained to approximate F with low jacobian loss (mean squared error in F and dF).

when used in a seismic model results in good approximations of Lyapunov exponents and other statistical measures. Overall, our results must be interpreted not as a practical method for how to train a chaotic surrogate model, but rather as a way to combine ergodic theory with statistical learning theory to obtain generalization bounds for learning physically correct/statistically accurate chaotic systems using only short-term dynamics during training.

Next, we discussed a new method, Score Operator Newton, for sampling from a target density, say, ρ^ν, corresponding a measure ν. We construct a transport map, which is an invertible function that determines a coupling between two distribution. In this case, we construct a transport map, T, such that $T_\sharp \mu = \nu$, where μ is an easy-to-sample reference distribution with a corresponding density ρ_μ. In our work, we derive a new construction of a transport map starting with the change of variables formula for the densities, given by $\rho_\nu = \rho_\mu \circ T^{-1}/|\det dT| \circ T^{-1}$. By taking the logarithm and differentiating, we can see that a transport map satisfies a functional equation that matches the score of the target with the score of the density on the left hand side. Without being explicit, we will write this equation as $\mathcal{G}(s_\mu, T) = s_\nu$, where s_π is the score associated with a measure π. Thus, the problem of finding a T that satisfies the score equation is equivalent to solving for the zero function of $\mathcal{R}(T) = \mathcal{G}(s_\mu, T) - s_\nu$. We derive a Newton-Raphson method in infinite-dimensions (such infinite-dimensional Newton-Raphson methods appear in nonlinear elliptic PDE theory and in KAM theory in dynamical systems). We show that each Newton-Raphson update involves solving a PDE, $\nabla \circ \mathcal{L}\phi := \nabla \circ (\nabla^2 \phi + s_\nu \cdot \nabla \phi)$, for different right hand sides, to obtain an update to T of the form $T \to (\mathrm{Id} + \nabla\phi) \circ T$. We suggest solving these PDEs with Feynman-Kac formulae and draw connections of the solutions ϕ with *nudging* introduced by Joaquin Miguez and Prashant Mehta at the workshop. Preliminary

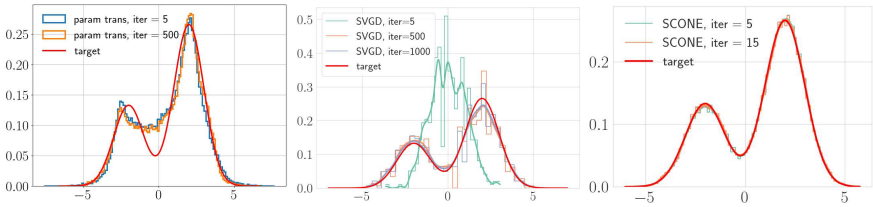


FIGURE 2. Left: parametric monotone transport maps from Parno et al 2022. Center: Stein Variational Gradient Descent from Liu and Wang 2016. Right: our transport map, based on Score Operator Newton iterations; the computational cost of the three methods per iteration is kept the same by suitable choice of hyperparameters of each scheme.

numerical results (Figure 2) suggest that the T obtained after a few iterations of this Score Operator Newton method approximates the target distribution more accurately at the same computational cost, when compared to parameterizing among transport maps and solving a regression problem (left) or when compared to a Stein Variational gradient flow transport (nonparametric transport; Right). SCONE transport may exhibit divergence however, like Newton methods in finite dimensions.

The work on Score Operator Newton and surrogate modeling in chaotic systems have been published respectively in

- Chandramoorthy, N., Schaefer, F. T., & Marzouk, Y. M. (2024, April). Score Operator Newton transport. AISTATS 2024 (pp. 3349–3357). PMLR.
- Park, J., Yang, N., & Chandramoorthy, N. (2024). When are dynamical systems learned from time series data statistically accurate?. NeuRIPS 2024

Data assimilation with generative models: Refining nonlinear signal calibration with diffusion processes

OANA LANG

(joint work with Alexander Lobbe and Dan Crisan)

Data assimilation integrates real-world data into mathematical models to improve their accuracy and predictive capabilities. It is widely used in fields like meteorology, oceanography, and environmental science, where dynamic systems require continuous updates. However, accurately calibrating these models remains challenging, particularly when dealing with sparse or noisy data.

In this work, we introduce a novel approach using diffusion generative models to improve model calibration before data assimilation. By leveraging their ability to generate statistically consistent synthetic data for high-dimensional phenomena, we achieve a more accurate and robust initial calibration. This enhances the

data assimilation process, leading to improved model performance and predictive accuracy.

Generative models are a class of machine learning models designed to approximate an unknown data distribution based on a given dataset of samples. An important subclass refers to *diffusion models*, which iteratively map training data to a well-defined distribution (such as a Gaussian) through a process analogous to diffusion. In this work, we use a variant known as the Diffusion Schrödinger Bridge (DSB), where both the forward and the backward dynamics are learnt via a neural network. Once trained, the model generates samples from the target distribution by simulating the backward diffusion process, starting from Gaussian-distributed samples.

We adopt the framework of stochastic nonlinear filtering to describe the data assimilation methodology in general and its specific application in our work. We define two processes, X and Z, on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where X represents the *signal process* or *truth*, and Z denotes the *observation process*. In this study, X corresponds to the pathwise solution of a rotating shallow water system (1), approximated using a high-resolution numerical method. The pair (X, Z) forms the foundation of the nonlinear filtering problem, which seeks to approximate the posterior distribution of the signal X_t given the observations $Z_1, Z_2, \ldots, Z_t$. The posterior distribution at time t is denoted by π_t.

Let d_X and d_Z denote the dimensions of the state and observation spaces, respectively. In many real-world applications, such as weather prediction, d_X is extremely large, typically $d_X = O(10^9)$. Performing data assimilation (DA) on such high-dimensional models requires supercomputing resources, which is why some of the world's most advanced supercomputers are employed in meteorological centers. Here, we propose an alternative approach: instead of working with the full signal X_t, we introduce an approximate model X_t^c computed on a coarser grid. In our example, the signal is denoted by X_t^f to highlight that it evolves on a finer grid than its proxy X_t^c, which is constructed on a coarser grid. Naturally, X_t^f and X_t^c will exhibit different dynamics, as small-scale effects are lost in the coarser representation. This is where generative modeling plays a crucial role. To account for small-scale influences, we introduce a stochastic term in the equation governing X_t^c. This term must be calibrated using data recorded from X_t^f *prior* to applying data assimilation. The two diagrams in Figure 1 illustrate this process.

In some instances within the field, a generative model is used to approximate the posterior distribution offline, which is computationally expensive when performed iteratively. In contrast, we do not replace the forecast and assimilation steps with generative models, but instead apply a diffusion model methodology to calibrate the signal *before* data assimilation begins.

In the assimilation step, we use a particle filter with tempering and jittering. Particle filters are well-suited for complex, multimodal distributions. Unlike traditional methods, such as the Kalman filter, which assume linearity and Gaussian noise, particle filters represent the posterior distribution using random samples or *particles*. Each particle is assigned a weight based on how well it fits the observed

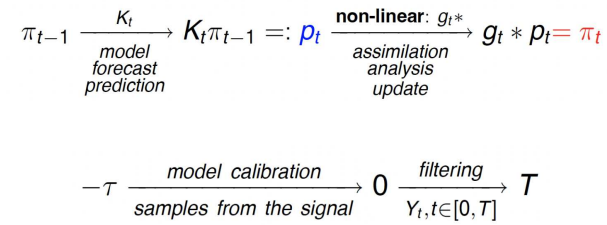


FIGURE 1. **Data assimilation and model calibration.** *Top:* The predictive distribution evolves forward using a forecast model and is subsequently updated through a nonlinear operation incorporating observed data. *Bottom:* Before time 0, the forecast model undergoes calibration. The filtering process starts only after calibration is complete.

data. As new observations are incorporated, particles are propagated through the model and their weights are updated according to the likelihood of the observed data, with particles matching the observations receiving higher weights.

We base our analysis on a stochastic and non-dimensional version of the rotating shallow water model given by:

$$\begin{aligned} \mathrm{d}_t\mathbf{u} + \mathcal{R}(\mathbf{u},\eta) &= 0 \\ \mathrm{d}_t\eta + \mathcal{P}(\eta,\mathbf{u}) &= 0 \end{aligned} \tag{1}$$

where $\mathbf{u}(x,t) = (u(x,t), v(x,t))$ is the horizontal fluid velocity vector field and $\eta(x,t)$ is the height of the fluid column. The operator $\mathcal{R} : C^1(\Omega,\mathbb{R}^2) \times C^1(\Omega,\mathbb{R}) \to C^0(\Omega,\mathbb{R}^2)$, where Ω is a spatial domain, governs the dynamics of the velocity vector field:

$$\mathcal{R}(\mathbf{u},\eta) := (\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u} + \frac{f}{\mathrm{Ro}}\hat{\mathbf{z}}\times\mathbf{u} + \frac{1}{\mathrm{Fr}^2}\boldsymbol{\nabla}(\eta - b) - \nu\Delta\mathbf{u} - \mathbb{F} - \mathbb{B}.$$

The operator $\mathcal{P} : C^1(\Omega,\mathbb{R}) \times C^1(\Omega,\mathbb{R}^2) \to C^0(\Omega,\mathbb{R})$ governs the evolution of the height of the fluid η

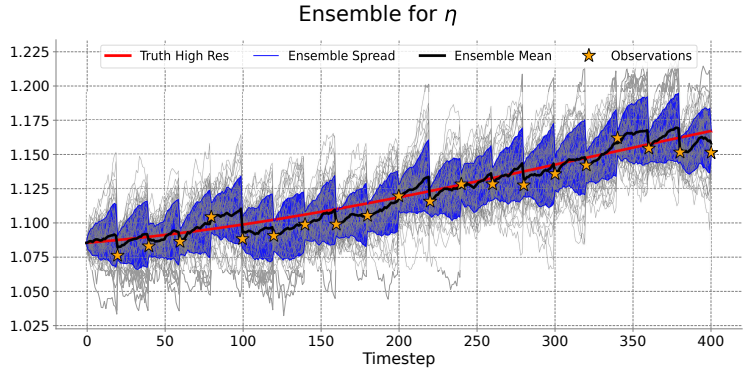
$$\mathcal{P}(\eta,\mathbf{u}) := \boldsymbol{\nabla}\cdot(\eta\mathbf{u}).$$

Here $f \in \mathbb{R}$ is the Coriolis parameter, $f = 2\Theta\sin\varphi$ where Θ is the rotation rate of the Earth and φ is the latitude; $f\hat{z}\times\mathbf{u} = (-fv, fu)^T$, where $\hat{z}$ is a unit vector pointing away from the centre of the Earth; $\mathrm{Fr} = \frac{U}{\sqrt{gH}}$ is the Froude number (dimensionless) which is connected to the stratification of the fluid flow. In this case U is a typical scale for horizontal speed and H is the typical vertical scale, while g is the gravitational acceleration; $\mathrm{Ro} = \frac{U}{f_0 L}$ is the Rossby number (also dimensionless) which describes the effects of rotation on the fluid flow: a small Rossby number ($\mathrm{Ro} \ll 1$) suggests that the rotation term dominates over the advective terms; $b(x,t)$ is the bottom topography function and ν is the viscosity

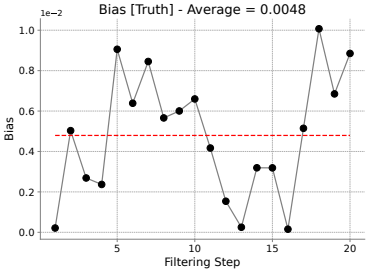
coefficient. For more exact details regarding the wind forcing $\mathbb{F}$ and the bottom friction $\mathbb{B}$, please see [3]. The full stochastic version that gives our signal is

$$
\begin{aligned}
&\mathrm{d}\mathbf{u} + \mathcal{R}(\mathbf{u},\eta)\,\mathrm{d}t + \sum_i \left[(\boldsymbol{\xi}_i\cdot\boldsymbol{\nabla})\mathbf{u} + \boldsymbol{\nabla}\boldsymbol{\xi}_i\cdot\mathbf{u} + \frac{f}{\mathrm{Ro}}\hat{\mathbf{z}}\times\boldsymbol{\xi}_i\right]\circ\mathrm{d}W_t^i = 0 \\
&\mathrm{d}\eta + \mathcal{P}(\eta,\mathbf{u})\,\mathrm{d}t + \sum_i \boldsymbol{\nabla}\cdot(\eta\boldsymbol{\xi}_i)\circ\mathrm{d}W_t^i = 0
\end{aligned}
\tag{2}
$$

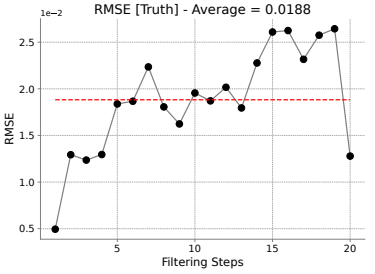
where $\circ$ denotes Stratonovich integration and W^i are standard i.i.d. Brownian motions.



(A) Ensemble Evolution



(B) Ensemble Bias



(C) Ensemble RMSE

FIGURE 2. Results of the filtering experiment conducted over 400 timesteps. The assimilated variable is η, with an ensemble of 50 particles and an assimilation window of 20 forecast timesteps. At the observed grid location, we present (A) the ensemble evolution in comparison to the true deterministic fine-grid trajectory, (B) the ensemble bias over time, and (C) the ensemble RMSE.

Results of the filtering experiment, using the calibrated stochastic signal and a particle filtering methodology introduced above, are presented in Figure 2. The particle filter yields good results, even though it operates on a coarser scale than

the true signal. This is achieved by utilizing the filter's robustness to misspecifications in the transition model and carefully calibrating the unresolved scales.

References

[1] D. Crisan, A. Lobbe, O. Lang, *Generative Modelling of Stochastic Rotating Shallow Water Noise* (arXiv:2403.10578).
[2] D. Crisan, O. Lang, A. Lobbe, P.J. van Leeuwen, R. Potthast, *Noise calibration for the stochastic rotating shallow water model*, Foundations of Data Science (2023).
[3] O. Lang, D. Crisan, P. J. van Leeuwen, R. Potthast, *Bayesian Inference for Fluid Dynamics: A Case Study for the Stochastic Rotating Shallow Water Model*, Frontiers in Applied Mathematics and Statistics **8** (2022).

Likelihood-driven dynamic measure transport: A natural fit for data assimilation?

Aimee Maurais
(joint work with Youssef Marzouk)

In the Bayesian approach to inverse problems and data assimilation we seek the posterior distribution of unknown parameters given noisy, often indirect measurements. Fundamental to our ability to make use of a Bayesian posterior is our ability to obtain samples from it, whether for estimating expectations via Monte Carlo or for performing uncertainty quantification. And yet, for all but the most simple distributions, sampling is a highly non-trivial computational task and continues to be the subject of extensive research. One powerful approach to sampling is *measure transport* [1]*:* given a posterior distribution π_1 on $\mathbb{R}^d$ and a prior distribution π_0 on $\mathbb{R}^d$ from which we can sample, the idea is to find $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\sharp \pi_0 = \pi_1$, i.e., if $X_0 \sim \pi_0$, then $T(X_0) \sim \pi_1$. The attractiveness of the transport framework is that obtaining samples from π_1 is as simple as sampling from π_0 and applying T. The challenge of the transport framework lies in identifying a suitable map T: such maps are not unique and can be difficult to learn in high dimensions or when the target distribution has challenging features such as multimodality.

An alternative to searching for a single, highly expressive "one-shot" transport map between π_0 and π_1 is to use *dynamics* to define a transport incrementally, e.g., via the flow map induced by trajectories of an ODE or the coupling induced by sample paths of an SDE; see 1. In either case, the idea is to apply dynamics which will transform some initial state $X_0 \sim \pi_0$ to a state $X_S \sim \pi_{X_S} \approx \pi_1$ for some time $S > 0$. Dynamic approaches to transport are state-of-the-art in generative modeling [2, 3, 4, 5, 6], wherein samples from both π_0 and π_1 are almost always required for training. In Bayesian and other density-driven sampling settings, where π_1 is only known through its unnormalized density, there are a number of dynamic sampling algorithms which have their grounding as *gradient flows* of functionals on spaces of probability measures. Most well-known algorithms in this vein (e.g., [7, 8, 9, 10]) use some form of the Wasserstein geometry to obtain

particle dynamics which must, in principle, be run for *infinite time* in order to ensure correct sampling from π_1.

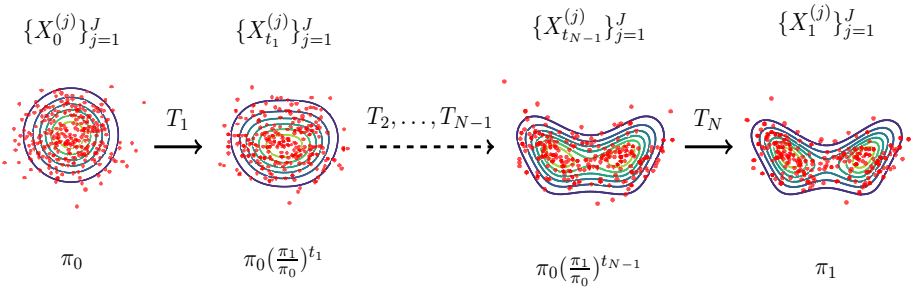


FIGURE 1. We employ a finite-time dynamic sampling scheme in this work, deriving a mean-field ODE which approximately transports a reference π_0 to a target π_1 in unit time along the path $\pi_t \propto \pi_0^{1-t}\pi_1^t$, $t \in [0,1]$. In discrete time this approach can be viewed as one of obtaining incremental transport maps $T_1, \ldots, T_N$ along a discretization of the path $(\pi_t)_{t\in[0,1]}$.

In this talk we introduce a dynamic sampling approach based on an ODE which transports samples from π_0 to π_1 in unit time such that the time-dependent distribution of the samples is the geometric mixture $\pi_t \propto \pi_0^{1-t}\pi_1^t = \pi_0(\frac{\pi_1}{\pi_0})^t$, $t \in [0,1]$. [11, 12]. Our method identifies gradient velocity fields, via solution of Poisson equations, which will cause the distribution of the samples to follow this path. The gradient structure of the velocity field has an optimal transport interpretation, while this choice of path lends Fisher-Rao gradient flow structure to the sampler. On a practical level, our approach of solving the weak form of the Poisson equations in reproducing kernel Hilbert space yields tractable interacting particle systems which are gradient-free and only require samples from the prior and access to evaluations of the unnormalized likelihood for use, making them suitable choices for ensemble data assimilation. Moreover, our approach suggests a general framework for gradient-free ODE transport between prior and posterior which can be employed for many different combinations of distribution paths and features for representing the velocity field. This flexibility enables, for instance, the use of random Fourier features in the place of explicit kernel functions, which can lead to dramatic reductions in computational complexity, and is arguably essential in situations when the traditionally used path of distributions, the geometric mixture, features "teleportation of mass" behavior. We conclude the talk by discussing modifications to the framework which we view as essential for scaling the ODE transport approach to high-dimensional problems, including exploitation of sparse conditional dependence structure (localization) and the ability to automatically pick paths of distributions which are suitable for transport and do not feature teleportation behavior.

References

[1] Y. Marzouk, T. Moselhy, M. Parno, A. Spantini, *Sampling via Measure Transport: An Introduction*, Handbook of Uncertainty Quantification (2016), 1–41.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, *Score-based generative modeling through stochastic differential equations*, International conference on learning representations (2021).

[3] V. De Bortoli, J. Thornton, J. Heng, A. Doucet, *Diffusion Schrodinger Bridge with Applications to Score-Based Generative Modeling*, Advances in Neural Information Processing Systems **34** (2021), 17695–17709.

[4] X. Liu, C. Gong, Q. Liu, *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow*, The Eleventh International Conference on Learning Representations, (2023).

[5] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, M. Le, *Flow Matching for Generative Modeling*, The Eleventh International Conference on Learning Representations, (2023).

[6] M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, *Stochastic Interpolants: A Unifying Framework for Flows and Diffusions*, (arXiv:2303.08797).

[7] Q. Liu, D. Wang, *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm*, Advances in Neural Information Processing Systems **29** (2016).

[8] A. Garbuno-Inigo, F. Hoffmann, W. Li, A. M. Stuart, *Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler*, SIAM Journal on Applied Dynamical Systems **19**(1) (2020), 412–441.

[9] A. Garbuno-Inigo, N. Nusken, S. Reich, *Affine Invariant Interacting Langevin Dynamics for Bayesian Inference*, SIAM Journal on Applied Dynamical Systems (2020).

[10] S. Reich, S. Weissmann, *Fokker–Planck Particle Systems for Bayesian Inference: Computational Approaches*, SIAM/ASA Journal on Uncertainty Quantification **9**(2) (2021), 446–482.

[11] A. Maurais, Y. Marzouk, *Continuous-Time Transport: Homotopy-Driven Sampling and a New Interacting Particle System*, NeurIPS 2023 Workshop on Optimal Transport and Machine Learning (2023).

[12] A. Maurais, Y. Marzouk, *Sampling in Unit Time with Kernel Fisher-Rao Flow*, Proceedings of the 41st International Conference on Machine Learning (2024), 35138–35162.

Nudging state-space models for Bayesian filtering under misspecified dynamics

Joaquin Miguez

(joint work with Fabian Gonzalez, O. Deniz Akyildiz and Dan Crisan)

State space models & Bayesian filtering. State-space models (SSMs) are key building blocks in many applications in signal processing, machine learning, weather forecasting, etc. In a typical SSM, the *system state* is a random sequence that evolves over time according to a Markov transition kernel and the available *observations* (data) are related to the system state by a likelihood function. The main statistical goal in SSMs is to infer the state of the system given a sequence of observations, a problem known as filtering [1].

We represent the state of the SSM by a random sequence $\{X_t\}_{t\geq 0}$. The initial state X_0 is a random variable (r.v.) with probability law π_0 and, at any time $t \geq 1$, the dynamics of the transition from X_{t-1} to X_t is modelled by a Markov kernel $K_t(x_{t-1}, \mathsf{d}x_t)$. The sequence of observations is denoted $\{Y_t\}_{t\geq 1}$ and the relationship between the state X_t and the observation Y_t is modelled by a conditional

probability density function (pdf) $p_t(y_t|X_t = x_t)$. Since in practice the observations are given, $Y_t = y_t$ for $t \geq 1$, the latter relationship is usually given in terms of a likelihood function $g_t(x_t) \propto p_t(y_t|X_t = x_t)$. With these elements, the conditional probability law of the state X_t given the data $Y_{1:t} = y_{1:t} := \{y_1, \ldots, y_t\}$ can be constructed recursively via the Chapman-Kolmogorov equation and Bayes' theorem and we denote it as π_t. The conditional law π_t is often termed the optimal, or Bayesian, filter.

The optimal filter π_t can only be computed exactly in a few specific cases. The most relevant one is the scenario where both the Markov kernels K_t and the likelihoods g_t correspond to linear relationships and Gaussian noise. Under such assumptions, π_t is Gaussian and its mean and covariance matrix can be computed recursively via the Kalman filter (KF) algorithm [2]. In most practical applications, however, the optimal filter π_t can only be approximated numerically using nonlinear KFs, particle filters (PFs) or other approximation methods [3, 4].

Model misspecification. A major challenge in Bayesian filtering is model misspecification, which occurs when the chosen family of transition models, $\{K_t\}_{t\geq 1}$, likelihood functions, $\{g_t\}_{t\geq 1}$, or both, fail to represent the statistical properties of the real-world system with sufficient accuracy. Model misspecification has been studied from different viewpoints in the literature, including outlier detection, robust filtering, parameter estimation, and the so-called *nudging* techniques.

Outlier detection [5] is, perhaps, the simplest way to manage observations which are in poor agreement with the assumed SSM. When an observation is collected, a statistical test can be run to determine whether the observed data y_t is compatible with the predicted distribution generated by the SSM and the filtering algorithm. If the test indicates that the observation is anomalous then the data can either be discarded or be processed using a *robust* procedure that mitigates the effect of the outlying data on the filter update. A fundamental problem with these approaches is that anomalous data are handled as detrimental and uninformative, under the assumption that they have not been generated by the system of interest. Very often, however, a genuine observation from the system of interest may appear as an outlier because of the misspecification of the SSM. By discarding or mitigating this observation, relevant information is wasted and model errors are reinforced.

Another classical strategy to account for modelling uncertainty is to choose not *one* SSM but a parametric family of SSMs indexed by a (possibly multidimensional) parameter θ. When a sequence of observations becomes available, the model is calibrated by tuning the parameter θ to the data according to some statistical criterion. Maximum likelihood estimation methods have been proposed [6, 7], as well as Bayesian estimation methods [8, 9, 10, 11]. While parameter estimation methods are practically indispensable, they do not solve the model misspecification problem –because the parametric family of SSMs may not be flexible enough the represent the features of the system of interest, no matter the choice of θ.

In geophysics, filtering algorithms are often referred to as *data assimilation* methods and a class of techniques collectively known as *nudging* have been devised to mitigate the model misspecification problem [12]. Nudging methods are

designed to steer (or *nudge*) a model towards the observed data over time by adding a (small) corrective term to the model dynamics. The goal is to make the model follow observed values more closely without breaking down its original dynamics. This 'definition' is vague enough to encompass a large and heterogeneous family of methods [13, 14, 15, 16].

A nudging methodology for misspecified kernels. We adopt a viewpoint of nudging as a data-informed modification of the kernels $\{K_t\}_{t\geq 1}$ of the SSM, rather than a tweak of the filtering algorithms. In particular, let $\mathcal{M}$ denote the original SSM available for a given problem or application. We introduce a family of *nudging maps* $(\alpha_t)_{t\geq 1}$ which, given the available observations $\{y_t\}_{t\geq 1}$, yield a sequence of modified (nudged) kernels $\{K_t^\alpha\}_{t\geq 1}$. These kernels, in turn, characterise a modified SSM, denoted $\mathcal{M}^\alpha$, which is therefore different from the original $\mathcal{M}$. We investigate the relative agreement of the two models, $\mathcal{M}$ and $\mathcal{M}^\alpha$, with a given data set $y_{1:T}$. This agreement is quantified by means of the marginal likelihoods, or Bayesian model evidence, of the two SSMs. Our key findings are outlined below:

- We introduce a class of "parametric nudging transformations" that satisfy some regularity conditions and admit various different practical implementations.
- For a given set of observations $y_{1:T}$, and under mild assumptions on the original model $\mathcal{M}$, we prove that the proposed nudging methodology can yield a modified model $\mathcal{M}^\alpha$ that attains a higher marginal likelihood than the base model $\mathcal{M}$. In particular, when the original model $\mathcal{M}$ is indexed by a vector of parameters θ, i.e., $\mathcal{M} \equiv \mathcal{M}_\theta$, we prove that the nudged model $\mathcal{M}_\theta^\alpha$ can attain a marginal likelihood that (a) is higher than the marginal likelihood of the model $\mathcal{M}_\theta$, with the same parameters θ, and (b) lies in a neighbourhood of the marginal likelihood attained by model $\mathcal{M}_{\theta_*}$, where θ_* is the maximum likelihood estimator of the parameters.
- We describe a specific class of nudging transformations that rely on the ability to compute the gradient of the log-likelihood function $\log g_t$ of the original model $\mathcal{M}$. We prove that the theoretical guarantees obtained for the general parametric transformations also hold for the proposed gradient-based nudging. This version of nudging is relatively easy to implement, even when $\nabla \log g_t$ is analytically intractable, using standard numerical tools.
- We apply the proposed methodology, with gradient-based nudging transformations, to the class of linear-Gaussian SSMs and explicitly obtain a nudged version of the KF (i.e., a KF for the nudged model $\mathcal{M}^\alpha$).

We present a set of numerical results that illustrate the application of the methodology for both linear and nonlinear dynamical models.

References

[1] B. Anderson, J. Moore, *Optimal filtering*, Courier Corporation (2005).
[2] R. Kalman, *A new approach to linear filtering and prediction problems* (1960).
[3] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*, Springer (2009).

[4] S. Sarkka, L. Svensson, *Bayesian Filtering and Smoothing*, vol 32, Cambridge University Press (2023).
[5] A. Blazquez-Garcia, A. Conde, U. Mori, J. Lozano, *A review on outlier/anomaly detection in time series data*, ACM Computing Surveys **54**(3) (2021), 1–33.
[6] F. LeGland, L. Mevel, *Recursive estimation in hidden Markov models*, Proceedings of the 36th IEEE Conference on Decision and Control **4** (1997), 3468–3473.
[7] V. Tadic, *Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden Markov models*, IEEE Transactions on Information Theory **56**(12) (2010), 6406–6432.
[8] C. Andrieu, A. Doucet, R. Holenstein, *Particle Markov chain Monte Carlo methods*, Journal of the Royal Statistical Society B, **72** (2010), 269–342.
[9] N. Chopin, P. E. Jacob, O. Papaspiliopoulos, *SMC2: an efficient algorithm for sequential analysis of state space models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2012).
[10] D. Crisan and J. Miguez, *Nested particle filters for online parameter estimation in discrete-time state-space Markov models*, Bernoulli **24**(4a) (2018), 3039–3086.
[11] S. Perez-Vieites, I. Marino, J. Miguez, *Probabilistic scheme for joint parameter estimation and state prediction in complex dynamical systems*, Physical Review E **98**(6) (2018).
[12] C. Cotter, S. Reich, *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press (2015).
[13] S. Lakshmivarahan, J. Lewis, *Nudging methods: A critical overview*, Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II) (2013), 27–57.
[14] X. Luo, I. Hoteit, *Ensemble Kalman filtering with residual nudging*, Tellus A: Dynamic Meteorology and Oceanography **64**(1) (2021).
[15] S. Dubinkina, H. Goosse, *An assessment of particle filtering methods and nudging for climate state reconstructions*, Climate of the Past **9**(3) (2013), 1141–1152.
[16] O. D. Akyildiz, J. Miguez, *Nudging the particle filter*, Statistics and Computing **30** (2020), 305–330.

Filtering of Hamiltonian dynamics with dynamical sensor placement

Olga Mula

Hamiltonian dynamics are challenging for filtering because solutions have often low regularity, they are spatially localized, and the evolution preserves certain quantities which one would like to discover and then preserve in a numerical reconstruction. In this talk, I will present a filtering algorithm formulated in continuous time, and which can be implemented after time discretization. Two main ingredients of the algorithm are symplectic dynamical reduced order models, and a dynamic strategy to position sensors.

On connections between sequential Bayesian inference and evolutionary dynamics

Sahani Pathiraja
(joint work with Philipp Wacker)

It has long been posited that there is a connection between the dynamical equations describing evolutionary processes in biology (namely, replicator-mutator dynamics) and sequential Bayesian learning methods. This talk describes new research

in which this precise connection is rigorously established and expanded in the continuous time setting. Here we focus on a class of interacting particle methods for solving the sequential Bayesian inference problem which are characterised by a McKean-Vlasov SDE. Of particular importance is a piecewise smooth approximation of the observation path from which the discrete time filtering equations are shown to converge to a Stratonovich interpretation of the Kushner equation. This smooth formulation will then be used to draw precise connections between nonlinear filtering and replicator-mutator dynamics. Additionally, gradient flow formulations with respect to the Fisher-Rao metric will be investigated. We also demonstrate that a particular form of replicator-mutator dynamics with collaboration is beneficial for the misspecified model filtering problem, and highlight a connection to inflation in data assimilation. It is hoped this work will spur further research into exchanges between sequential learning and evolutionary biology and to inspire new algorithms in filtering and sampling.

References

[1] S. Pathiraja, P. Wacker, *Connections between sequential Bayesian inference and evolutionary dynamics* (arXiv:2411.16366)

Approaching observation noise misspecification via generalised posteriors – robust Kalman filter variants and some of their properties

Hans Reimann

Almost four decades after Zellner's notion of information optimality via Bayes' theorem as an optimal way of processing information in [1], we still frequently observe instances of approximate schemes seemingly outperforming the supposedly optimal method in a variety of contexts including Bayesian data assimilation.

In their testimony for finding novel forms of posteriors, the authors in [2, 3] break down this apparent contradiction into key factors for when this optimality may no longer hold. Among these factors, misspecified observation likelihood models that do not describe reality and its true data generating mechanisms sufficiently are pointed out as a major source of this discrepancy. Although a certain family of observation likelihoods may account best for our knowledge of a dynamic as well as suit our methods, when it is misaligned in crucial properties, Bayes' theorem is struggling to recover accurate representation. In data assimilation both, limited knowledge and limited computational capabilities are established challenges.

One of these crucial effects on inference via Bayes' theorem is tail decay of the likelihood distribution. As Bayesian learning such as in Bayesian filtering is rooted in Kullback-Leibler divergence and its information processing, mismatch in tail weight between a true data generating process and an assumed probabilistic model can deteriorate results significantly. In practice, this can be understood as extreme observations appearing more frequent than suggested by the model or simply frequent observation outliers.

The corresponding branch of robust Bayesian statistics experienced a new influx of methods with so called generalized (robust) posteriors based on the idea of generalized Bayesian inference in works of [4, 5, 6] among others. The idea is hereby similar to McAllester's idea of PAC Bayesian learning and the corresponding Gibbs posteriors [7] in that Kullback-Leibler divergence is replaced by some other optimization criteria for matching observation likelihoods to a true data generating process. Where PAC Bayes requires some arbitrary loss function between data and parameter, generalised Bayesian inference focuses on discrepancy measures between distributions with support in observation space.

In a promising line of work in [8, 9, 10], the authors explore use of novel divergence measures in the context of Bayesian online change point detection and Gaussian process regression to obtain the described outlier robustness. Utilizing diffusion score matching as an estimator for minimum diffusion Fisher divergence as investigated in [11], akin to regular score matching as established in [12, 13] with respect to Fisher divergence, it hereby provides the required generalization in an extra degree of freedom to obtain this robustness of interest while maintaining desired properties such as forms of conjugacy.

While the context of generalised Bayesian inference is yet to be fully utilized and understood in the context of Bayesian data assimilation and Bayesian filtering, initial investigations have resulted in curious first insights. Work in [14] was first via formulating a generalized particle filter, however still fairly costly in their use of β-divergences. Further exploring as well as exploiting the Gauss-Gauss conjugacy properties derived in [10], the work in [15] utilized diffusion score matching to derive a Kalman filter variant with the desired provable robustness property. The work in [16] came to similar results based on what they coined weighted observation likelihood - a form of weighted cross-entropy measure.

Both these novel Kalman filter variants are promising in that they again only required computationally cheap parameter updates to obtain closed form, analytic expressions of the Gaussian analysis, or posterior, distribution. Moreover, utilizing results in [17] we can show stability in an asymptotic steady state of the covariance matrix under usual, mild conditions as well as intuition on choice of introduced tuning parameters via expanding on results in [18]. We are hereby generally subject to the usual limitation to linear Gaussian state space system. However, as with the regular Kalman filter variants via ensemble approximations akin to the ensemble Kalman filter with perturbed observations and ensemble square root filters are readily available. It is these ensemble variants to the novel Kalman filters that are promising for filtering of non-linear signal processes. Additionally, for non-linear observation, or forward, maps, the popular local linear approximation in the celebrated local ensemble transform Kalman filter in [19] can be expanded in consistent fashion to the novel Kalman filter variants based on generalized Bayesian inference.

In summary, generalised posteriors can provide a novel and intuitive way to modify classic results in Bayesian filtering and data assimilation. The diffusion score matching based EnKF, ESRF and LETKF maintain crucial properties for ease of

application while providing robustness to tail decay mismatch and thus help reduce impact of when assumptions on the observation model do not hold. Moreover, they provide first insights and can pave the way to a more semi-parametric intuition of data assimilation when entering the much broader context of PAC Bayesian learning, e.g. with Stein discrepancy based loss measures. The diffusion score matching Kalman filter can then be understood as a special case of one such.

References

[1] A. Zellner, *Optimal information processing and Bayes's theorem*, The American Statistician **42**(4) (1988), 278–280.

[2] J. Knoblauch, J. Jewson, T. Damoulas, *Generalized variational inference: Three arguments for deriving new posteriors* (arXiv:1904.02063).

[3] J. Knoblauch, J. Jewson, T. Damoulas, *An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference*, Journal of Machine Learning Research **23**(132) (2022), 1–109.

[4] P. G. Bissiri, C. C. Holmes, S. G. Walker, *A general framework for updating belief distributions*, Journal of the Royal Statistical Society Series B: Statistical Methodology **78**(5) (2016), 1103–1130.

[5] A. Ghosh, A. Basu, *Robust Bayes estimation using the density power divergence*, Annals of the Institute of Statistical Mathematics **68** (2016), 413–437.

[6] J. Jewson, J. Q. Smith, C. Holmes, *Principles of Bayesian inference using general divergence criteria*, Entropy **20**(6) (2018).

[7] D. A. McAllester, *Some pac-bayesian theorems*, Eleventh Annual Conference on Computational Learning Theory (1998), 230–234.

[8] J. Knoblauch, T. Damoulas, *Spatio-temporal Bayesian on-line changepoint detection with model selection*, In Proceedings of the International Conference on Machine Learning (2018), 2718–2727.

[9] M. Altamirano, F.-X. Briol, J. Knoblauch, *Robust and Scalable Bayesian Online Changepoint Detection* (arXiv:2302.04759).

[10] M. Altamirano, F.-X. Briol, J. Knoblauch, *Robust and Conjugate Gaussian Process Regression* (arXiv:2311.00463).

[11] A. Barp, F.-X. Briol, A. Duncan, M. Girolami, L. Mackey, *Minimum Stein discrepancy estimators*, Advances in Neural Information Processing Systems **32** (2019).

[12] A. Hyvärinen, P. Dayan, *Estimation of non-normalized statistical models by score matching*, Journal of Machine Learning Research **6**(4) (2005).

[13] A. Hyvärinen, *Some extensions of score matching*, Computational Statistics & Data Analysis, **51**(5) (2007), 2499–2512.

[14] A. Boustati, O. D. Akyildiz, T. Damoulas, A. Johansen, *Generalised Bayesian filtering via sequential Monte Carlo*, In Advances in Neural Information Processing Systems **33** (2020), 418–429.

[15] H. Reimann, *Towards robust inference for Bayesian filtering of linear Gaussian dynamical systems subject to additive change*, Master's thesis, Universität Potsdam (2024).

[16] G. Duran-Martin, M. Altamirano, A. Y. Shestopaloff, L. Sánchez-Betancourt, J. Knoblauch, M. Jones, F.-X. Briol, K. Murphy, *Outlier-robust Kalman Filtering through Generalised Bayes* (arXiv:2405.05646).

[17] V. Solo, *Stability of the Kalman filter with stochastic time-varying parameters*, In Proceedings of 35th IEEE Conference on Decision and Control **1** (1996), 57–61.

[18] X. Gao, M. Sitharam, A. E. Roitberg, *Bounds on the Jensen gap, and implications for mean-concentrated distributions* (arXiv:1712.05267).

[19] B. R. Hunt, E. J. Kostelich, I. Szunyogh, *Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter*, Physica D: Nonlinear Phenomena **230** (2007), 112–126.

[20] S. Reich, *A nonparametric ensemble transform method for Bayesian inference*, SIAM Journal on Scientific Computing **35**(4) (2013).
[21] S. Reich, C. Cotter, *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press (2015).

Ensemble Kalman inversion for high dimensional problems

Xin Tong

Ensemble Kalman inversion (EKI) is an ensemble-based method to solve inverse problems. Its gradient-free formulation makes it an attractive tool for problems with involved formulation. However, EKI suffers from the "subspace property", i.e., the EKI solutions are confined in the subspace spanned by the initial ensemble. It implies that the ensemble size should be larger than the problem dimension to ensure EKI's convergence to the correct solution. Such scaling of ensemble size is impractical and prevents the use of EKI in high dimensional problems. To address this issue, we propose two novel approaches using localization and dropout regularization to mitigate the subspace problem. We prove that these methods converge in the small ensemble settings, and the computational cost of the algorithm scales linearly with dimension. We also show that they reach the optimal query complexity, up to a constant factor. Numerical examples demonstrate the effectiveness of our approach.

Uncertainty quantification analysis of bifurcations of the Allen–Cahn equation with random coefficients

Elisabeth Ullmann
(joint work with Chiara Piazzola, Christian Kuehn)

We consider the Allen–Cahn equation in a domain $D \subset \mathbb{R}^d$, $d = 1, 2, 3$, together with suitable boundary and initial conditions,

$$\begin{aligned} \partial_t u(\boldsymbol{x}, t) &= \Delta u(\boldsymbol{x}, t) + p u(\boldsymbol{x}, t) - u(\boldsymbol{x}, t)^3, && \boldsymbol{x} \in D, \\ u(\boldsymbol{x}, t) &= 0, && \boldsymbol{x} \in \partial D, t > 0, \\ u(\boldsymbol{x}, 0) &= u_0(\boldsymbol{x}), && \boldsymbol{x} \in \partial D, \end{aligned} \tag{1}$$

where $u\colon D \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a function, $p \in \mathbb{R}$ is a given parameter, and $u_0\colon D \to \mathbb{R}$ is a given function. (1) is a prototypical model problem in the dynamics of nonlinear partial differential equations (PDEs) [1]. It is well known that the dynamics of (1) changes qualitatively according to variations of the so-called *bifurcation parameter* p, which induces supercritical pitchfork bifurcations [2, 3].

In our work [4] we go beyond the state-of-the-art by introducing a random coefficient in the linear reaction part of the Allen–Cahn equation, thereby accounting

for random, spatially-heterogeneous effects. That is, we consider

$$
\begin{aligned}
&\partial_t u(\boldsymbol{x},t) = \Delta u(\boldsymbol{x},t,\omega) + q(\boldsymbol{x},\omega)u(\boldsymbol{x},t,\omega) - u(\boldsymbol{x},t,\omega)^3, && \boldsymbol{x}\in D,\\
&u(\boldsymbol{x},t,\omega) = 0, && \boldsymbol{x}\in \partial D, t>0,\\
&u(\boldsymbol{x},0,\omega) = u_0(\boldsymbol{x}), && \boldsymbol{x}\in \partial D,\ \mathbb{P}\text{-a.s.},
\end{aligned} \tag{2}
$$

where $(\Omega, \mathcal{A}, \mathbb{P})$ is a suitable probability space associated with the random coefficient function $q\colon D \times \Omega \to \mathbb{R}$. Importantly, we assume a spatially constant, deterministic mean value of the random coefficient, that is,

$$
q(\boldsymbol{x},\omega) = p + g(\boldsymbol{x}, \boldsymbol{Y}(\omega)), \tag{3}
$$

where $p \in \mathbb{R}$ is deterministic, $\boldsymbol{Y}\colon \Omega \to \Gamma \subset \mathbb{R}^N$, $N \in \mathbb{N}$, is a random vector with independent components Y_i, $i = 1, \ldots, N$ (so called *finite-dimensional noise*), $g\colon D\times\Gamma \to \mathbb{R}$ is uniformly bounded, that is, $\exists\, \overline{g} \in \mathbb{R}$, s.t. $\mathbb{P}(\{\omega \in \Omega\colon |g(\boldsymbol{x}, \boldsymbol{Y}(\omega))| \leq \overline{g}\ \forall \boldsymbol{x} \in D\}) = 1$, and $\mathbb{E}[g(\boldsymbol{x}, \boldsymbol{Y})] = 0$ for all $\boldsymbol{x} \in D$. We show that the mean value p of q in (3) is in fact a bifurcation parameter in the Allen-Cahn equation with random coefficients (2). Moreover, we show that the bifurcation points and bifurcation curves become random objects. We consider two distinct modelling situations: (i) for a spatially homogeneous coefficient q in (2) we derive analytical expressions for the distribution of the bifurcation points and show that the bifurcation curves are random shifts of a fixed reference curve; (ii) for a spatially heterogeneous coefficient q in (2) we employ a generalized polynomial chaos expansion to approximate the statistical properties of the random bifurcation points and bifurcation curves. We present numerical examples in 1D physical space ($d = 1$), where we combine the popular software package Continuation Core and Toolboxes (CoCo) [5] for numerical continuation and the Sparse Grids Matlab Kit [6] for the polynomial chaos expansion. Our exposition addresses both, dynamical systems and uncertainty quantification, highlighting how analytical and numerical tools from both areas can be combined efficiently for the challenging uncertainty quantification analysis of bifurcations in random differential equations. Moreover, we systematically build a framework for the bifurcation analysis of nonlinear PDEs with random coefficients, for example, the Allen–Cahn equation with a random diffusion coefficient a and a random coefficient b in front of the cubic nonlinearity,

$$
\partial_t u = \nabla \cdot (a \nabla u) + qu - bu^3. \tag{4}
$$

Looking forward, we mention one open question at the intersection of PDE analysis and uncertainty quantification: Which parameters in the PDE with random coefficients are bifurcation parameters? While this question is reasonably well studied for certain PDEs with deterministic coefficients, such as the Allen–Cahn equation, it is quite open for PDEs with random coefficients. For example, in our work [4] we consider a deterministic bifurcation parameter, namely the mean value of q in (2). Other bifurcation parameters are possible, for example, the standard deviation of q, or the probability $\mathbb{P}(q > q_{\max})$, where $q_{\max} \in \mathbb{R}$ is a given exceedance level. Finally, it is also possible to consider the random field q in (2) as bifurcation parameter, that is, we study a *random* bifurcation parameter.

References

[1] C. Kuehn, *PDE dynamics*, Mathematical Modeling and Computation, 23, Society for Industrial and Applied Mathematics (2019).

[2] N. Chafee, E. F. Infante, *A bifurcation problem for a nonlinear partial differential equation of parabolic type*, Applicable Analysis (1974/75), 17–37.

[3] J. C. Robinson, *Infinite-dimensional dynamical systems. An introduction to dissipative parabolic PDEs and the theory of global attractors*, Cambridge Texts in Applied Mathematics, Cambridge University Press (2001).

[4] C. Kuehn, C. Piazzola, E. Ullmann, *Uncertainty quantification analysis of bifurcations of the Allen–Cahn equation with random coefficients*, Physica D: Nonlinear Phenomena **470** (2024).

[5] H. Dankowicz, F. Schilder, *Recipes for continuation*, Computational Science & Engineering, 11, Society for Industrial and Applied Mathematics (2013).

[6] C. Piazzola, L. Tamellini, *Algorithm 1040: The Sparse Grids Matlab Kit - a Matlab implementation of sparse grids for high-dimensional function approximation and uncertainty quantification*, ACM Transactions on Mathematical Software **50**(1) (2024).

[7] K. Lux, P. Ashwin, R. Wood, C. Kuehn, *Assessing the impact of parametric uncertainty on tipping points of the Atlantic meridional overturning circulation*, Environmental Research Letters **17**(7) (2022).

[8] K. Lux, C. Kuehn, *Uncertainty quantification of bifurcations in random ordinary differential equations*, SIAM Journal on Applied Dynamical Systems **20**(4) (2021), 2295–2334.

[9] H. Kielhöfer, *Bifurcation theory. An introduction with applications to partial differential equations*, Applied Mathematical Sciences, 2nd edition, Springer (2012).

[10] S. Bartels, *Numerical methods for nonlinear partial differential equations*, Springer Series in Computational Mathematics, Springer (2015).

Combining nonlinear data assimilation and generative machine learning methods for fast high-dimensional Bayesian inference

Peter Jan van Leeuwen

Recently, strong progress has been made towards the solution of fully nonlinear Bayesian Inference, also called data assimilation, in high-dimensional geophysical systems, such as the atmosphere and ocean. Examples are the Particle Flow Filter (Pulido and Van Leeuwen, 2019), a recently developed fully nonlinear and efficient sequential Monte Carlo filter, and generative diffusion methods from machine learning. We explain these two methods and show that both have issues, which can be largely solved by combining them in an optimal way.

The Particle Flow Filter. The PPF uses the idea of a particle flow that iteratively pushes forward a set of interacting particles from the prior $p(\mathbf{x})$ to samples from the posterior $p(\mathbf{x}|\mathbf{y})$, in which $\mathbf{y}$ is the observation vector, without any reweighting or resampling strategies. Specifically, a set of N_p particles $\{\mathbf{x}^i\}_{i=1}^{N_p}$ is moved through state space via a gradient velocity field $\mathbf{f} : \Re^{n_x} \rightarrow \Re^{n_x}$ over pseudo-time s as:

$$d\mathbf{x}^i = \mathbf{f}(\mathbf{x}^i)ds \tag{1}$$

where we suppressed the dependence of $\mathbf{f}$ on the other particles with index $j \neq i$. This evolution can be viewed as Stein variational gradient descent. In practice, an Euler discretization of Equation (1) can be used:

$$\mathbf{x}_{s+1}^i = \mathbf{x}_s^i + \Delta s \mathbf{f}(\mathbf{x}_s^i) \tag{2}$$

where Δs is the pseudo-time step and $\mathbf{x}_s^i$ is the i^{th} particle at pseudo time s.

Given an intermediate pdf $q_s(\mathbf{x})$ formed by the particles at pseudo-time s, the goal is to find an optimal velocity field $\mathbf{f}$ such that the distance between $q_s(\mathbf{x})$ and the posterior pdf $p(\mathbf{x}|\mathbf{y})$ decrease with pseudo-time. Here, the distance between $q_s(\mathbf{x})$ and $p(\mathbf{x}|\mathbf{y})$ is measured through Kullback-Leibler (KL) divergence:

$$KL(q_s(\mathbf{x})||p(\mathbf{x}|\mathbf{y})) = \int q_s(\mathbf{x}) \log\left(\frac{q_s(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}\right) d\mathbf{x} \tag{3}$$

The velocity field $\mathbf{f}$ is chosen to optimally decrease the KL divergence of the intermediate pdf and the posterior pdf by solving the following problem

$$\mathbf{f}^* = arg\max_{\mathbf{f}\in\mathcal{H}} \left\{ -\frac{d}{ds} KL(q_s(\mathbf{x})||p(\mathbf{x}|\mathbf{y})) \right\} \tag{4}$$

where $\mathcal{H}$ is an infinite dimensional Hilbert space.

In practice, this optimization problem is still intractable because there is an infinite number of velocity fields that solve this equation. By constraining the functions within the unit ball of a reproducing kernel Hilbert space (RKHS) with a kernel function $\mathbf{K}(\cdot,\cdot) : \Re^{n_x} \times \Re^{n_x} \to \Re$, Hu and Van Leeuwen (2021) showed that the optimal velocity field $\mathbf{f}$ is

$$\mathbf{f}(\mathbf{x}_s^i) = \frac{1}{N_p}\mathbf{D}\sum_{j=1}^{N_p}\left[\mathbf{K}(\mathbf{x}_s^i, \mathbf{x}_s^j)\nabla_{\mathbf{x}_s^j} \log p(\mathbf{x}_s^j|\mathbf{y}) + \nabla_{\mathbf{x}_s^j} \cdot \mathbf{K}(\mathbf{x}_s^i, \mathbf{x}_s^j)\right] \tag{5}$$

where $\mathbf{D}$ is a positive-definite matrix that can be chosen to accelerate convergence. The kernel $\mathbf{K}$ characterizes the distance between any two particles. Typically, a diagonal and isotropic scalar Gaussian kernel is adopted, $\mathbf{K}(\mathbf{x}_s^i, \mathbf{x}_s^j) = K(\mathbf{x}_s^i, \mathbf{x}_s^j)\mathbf{I}_{n_x}$, where $\mathbf{I}_{n_x}$ is an identity matrix of size $n_x \times n_x$,

$$K(\mathbf{x}_s^i, \mathbf{x}_s^j) = \exp\left(-\frac{1}{2}(\mathbf{x}_s^i - \mathbf{x}_s^j)^T\mathbf{A}(\mathbf{x}_s^i - \mathbf{x}_s^j)\right) \tag{6}$$

and $\mathbf{A}$ is a symmetric matrix, often chosen similarly to $\mathbf{D}$.

To implement the update in equation (5), the logarithm of the posterior gradient must first be determined via Bayesx Theorem, as:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{y}\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{7}$$

with $p(\mathbf{y}|\mathbf{x})$ the likelihood, which is assumed to be known in closed form. For instance, Gaussian observation errors lead to

$$\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) = \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y} - H(\mathbf{x})) \tag{8}$$

where H is an observation operator that maps state space variables to the observational space, $\mathbf{H}$ its gradient, and $\mathbf{R}$ is the observation error covariance matrix.

For fully nonlinear data-assimilation methods the gradient of the prior is not easily determined. The problem is that we only have a representation of the prior in terms of a set of discrete particles, such that taking the log and a gradient is

not well defined. Assuming a Gaussian or Gaussian mixture for the prior can work well (e.g. Hu and Van Leeuwen, 2021), but is not exact and it is easy to envision cases where this would be inaccurate.

An alternative: generative diffusion methods. An alternative way to solve the Bayesian inference problem is to use generate diffusion methods from machine learning. Diffusion sampling relies on an Ornstein-Uhlenbeck process to transform samples from the prior into a standard Gaussian distribution $p_G(\mathbf{x}) = N(0, \mathbf{I})$. For this pdf we find that $\nabla_{\mathbf{x}} \log p_G(\mathbf{x}) = -\mathbf{x}$, such that the evolution in pseudo time for each sample $\mathbf{x}_i$ becomes:

$$d\mathbf{x}_i = -b(\tau)\mathbf{x}_i d\tau + \sigma(\tau) d\mathbf{W}_i \tag{9}$$

For suitably chosen $b(\tau)$ and $\sigma(\tau)$ this generates a set of samples from $N(0, \mathbf{I})$, from our original samples from $p(\mathbf{x})$. The key insight is that this process is reversible– starting from samples drawn from $N(0, \mathbf{I})$, we can reconstruct samples from the original distribution $p(\mathbf{x})$ by reversing the Ornstein-Uhlenbeck process.

To discretize the process in pseudo-time, we define a sequence $[0, \tau_1, \tau_2, ..., 1]$. The transition density between steps follows

$$q(\mathbf{x}_{t+1}|\mathbf{x}_0) = N(\gamma \mathbf{x}_0, \beta^2 \mathbf{I}) \tag{10}$$

where $\gamma(\tau)$ and $\beta(\tau)$ depend on the drift term$b(\tau)$ and noise function $\sigma(\tau)$ (Bao et al., 2023). Rewriting the evolution equation, we obtain

$$d\mathbf{x}_i = \left(-b(\tau)\mathbf{x}_i + \nabla_{\mathbf{x}_i} \log q(\mathbf{x})\right) d\tau + \boldsymbol{\sigma}(\tau) d\mathbf{W}_i \tag{11}$$

Running this equation backward in time reconstructs samples from $p(\mathbf{x})$. Since the transition densities $q(\mathbf{x}_{t+1}|\mathbf{x}_0)$ are Gaussian, all terms are computationally tractable, making diffusion sampling an efficient method for high-dimensional Bayesian inference.

So far, we have established a method to generate new samples from an existing distribution $p(\mathbf{x})$. However, Bayesian inference requires generating samples from the posterior distribution $p(\mathbf{x}|\mathbf{y})$ where $\mathbf{y}$ represent new observations. To achieve this, we need to incorporate the likelihood term into our sampling procedure. Since the posterior distribution is related to the prior through Bayes' theorem, its gradient can be decomposed as $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x})$. Bao et al. (2023) propose incorporating this additional likelihood term into the backward evolution equation by adding a new forcing term:

$$d\mathbf{x}_i = \left[-b(\tau)\mathbf{x}_i + \nabla_{\mathbf{x}_i} \log q(\mathbf{x}) + h(\tau)\nabla_{\mathbf{x}_i} \log p(\mathbf{y}|\mathbf{x})\right] d\tau + \boldsymbol{\sigma}(\tau) d\mathbf{W}_i \tag{12}$$

Here, $h(\tau)$ is a function that smoothly transitions from 0 at the start of the backward integration to 1 at the end. This ensures that the likelihood term is gradually incorporated into the sampling process, preventing abrupt adjustments and allowing for a more stable convergence to the posterior distribution. While this method has proven effective for certain monotonic nonlinear observation operators (Bao et al., 2023), it faces challenges when the observation operator is highly nonlinear. Furthermore, results can be sensitive to the choice of $h(\tau)$, leading to stability issues.

A new solution: merging the two As we have seen, both the Particle Flow Filter and generative diffusion methods have issues solving high-dimensional nonlinear Bayesian inference problems. The new idea is to combine the two - use the generative diffusion method to generate an analytical (!) expression of the gradient of the log of the prior pdf, which is then used in the Particle Flow Filter. Initial results are encouraging, but a few outstanding issues remain: 1) How to ensure that the gradient of the log of the prior pdf from the diffusion method leads to physically realistic samples, 2) is it possible to properly include the likelihood in the diffusion method, in which case we do not need the Particle Flow method, and finally 3) how can we accelerate the convergence rate of these methods?

References

[1] F. Bao, Z. Zhang, G. Zhang, *An ensemble score filter for tracking high-dimensional dynamical systems* (Arxiv:2309.00983v1).
[2] C.-C. Hu, P. J. van Leeuwen, *A particle flow filter for high-dimensional system applications*, Quarterly Journal of the Royal Meteorological Society **147** (2021), 2352–2374.
[3] M. Pulido, P. J. van Leeuwen, *Sequential monte carlo with kernel embedded mappings: The mapping particle filter*, Journal of Computational Physics **396** (2019), 400–415.

Likelihood-based methods for low frequency diffusion data & Statistical learning theory for neural operators

Sven Wang

We consider the problem of nonparametric inference in multi-dimensional diffusion models from low-frequency data. Due to the computational intractability of the likelihood, implementation of likelihood-based procedures in such settings is a notoriously difficult task. Exploiting the underlying (parabolic) PDE structure of the transition densities, we derive computable formulas for the likelihood function and its gradients. We then construct a Metropolis-Hastings Crank-Nicolson-type algorithm for Bayesian inference with Gaussian priors, as well as gradient-based methods for computing the MLE and Langevin-type MCMC. The performance of the algorithms is illustrated via numerical experiments.

We present statistical convergence results for the learning of mappings in infinite-dimensional spaces. Given a possibly nonlinear map between two separable Hilbert spaces, we analyze the problem of recovering the map from noisy input-output pairs corrupted by i.i.d. white noise processes or subgaussian random variables. We provide a general convergence results for least-squares-type empirical risk minimizers over compact regression classes, in terms of their approximation properties and metric entropy bounds, proved using empirical process theory. This extends classical results in finite-dimensional nonparametric regression to an infinite-dimensional setting. As a concrete application, we study an encoder-decoder based neural operator architecture. Assuming holomorphy of the operator, we prove algebraic (in the sample size) convergence rates in this setting, thereby overcoming the curse of dimensionality. To illustrate the wide applicability of our results, we discuss a parametric Darcy-flow problem on the torus.

Sequential measure transport for density estimation and its applications

Benjamin Zanger
(joint work with Tiangang Cui, Martin Schreiber, Olivier Zahm)

Transport-based methods are receiving growing interest because of their ability to sample easily from the approximated density. These methods aim at building a deterministic diffeomorphism $\mathcal{T}$, also called a transport map, which pushes forward an arbitrary reference probability density ρ_{ref} to a given target probability density π to be approximated. Typically, the transport map $\mathcal{T}$ is parameterized *e.g.* by invertible neural networks [1] and fitted using variational methods of the form

$$\min_{\mathcal{T}\in\mathcal{M}} \mathrm{D}(\pi||\mathcal{T}_\sharp\rho_{\text{ref}}) \tag{1}$$

with a statistical divergence $\mathrm{D}(\cdot\,||\cdot)$, typically the (reversed) KL-divergence. An emerging strategy for this problem is to first estimate π by $\widetilde{\pi}$ and then to compute a map $\mathcal{T}$ which exactly pushes forward ρ_{ref} to $\widetilde{\pi}$, known as the Knothe-Rosenblatt (KR) map, see [2, 3]. Among the infinitely many maps $\mathcal{T}$ which satisfy $\mathcal{T}_\sharp\rho_{\text{ref}} = \widetilde{\pi}$, the KR map is rather simple to evaluate since it requires only computing the cumulative distribution functions (CDFs) of the conditional marginals of $\widetilde{\pi}$. In general,

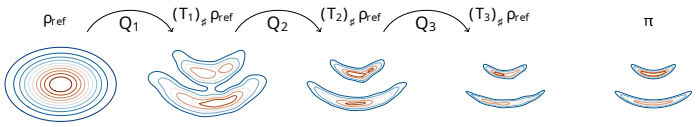


Figure 1. Visualization of the approximation of a bimodal density π (right) using $L = 3$ intermediate tempered densities estimated using SoS (4) and a Gaussian reference density ρ_{ref}.

problem (1) is difficult to solve when π is multimodal or when it concentrates on a low-dimensional manifold. A known solution to this problem is at the core of sequential Monte Carlo methods and has also been adopted *e.g.* in [3]. The idea is to introduce an arbitrary sequence of bridging densities

$$\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(L)} = \pi, \tag{2}$$

with increasing complexity. The sequential strategy consists in building L transport maps $\mathcal{Q}_1, \ldots, \mathcal{Q}_L$ one after the other. In general, there are two possible ways of combining these, in *forward* or *reverse* order,

$$\textit{forward: } \mathcal{T}_L = \mathcal{Q}_L \circ \mathcal{Q}_{L-1} \circ \cdots \circ \mathcal{Q}_1 \qquad \text{or} \qquad \textit{reverse: } \mathcal{T}_L = \mathcal{Q}_1 \circ \mathcal{Q}_2 \circ \cdots \circ \mathcal{Q}_L.$$

For our implementation, we choose the reverse order, since this allows us to build every map $\mathcal{Q}_\ell$ by solving a variational problem of the type

$$\min_{\mathcal{Q}_\ell\in\mathcal{M}} \mathrm{D}(\mathcal{T}_{\ell-1}^\sharp \pi^{(\ell)}||(\mathcal{Q}_\ell)_\sharp\rho_{\text{ref}}), \quad \text{where} \quad \mathcal{T}_{\ell-1} = \mathcal{Q}_1 \circ \ldots \circ \mathcal{Q}_{\ell-1} \tag{3}$$

for statistical distances with the property that $\mathrm{D}(\pi||\mathcal{T}_\sharp\rho) = \mathrm{D}(\mathcal{T}^\sharp\pi||\rho)$. These problems are equivalent to estimating the pullback density $(\mathcal{T}_{\ell-1})^\sharp\pi^{(\ell)}$ with an intermediate approximation $\rho^{(\ell)} = (\mathcal{Q}_\ell)_\sharp\rho_{\mathrm{ref}}$.

In the presented work, we do the following contribution to the framework of sequential measure transport.

First, we employ Sum-of-Squares (SoS) densities to approximate the intermediate densities $\rho^{(\ell)}$ using α-divergences $D_\alpha(\cdot||\cdot)$. We sequentially solve the variational density approximation problem as in Equation (3) with D_α as the statistical divergence and where

$$\rho^{(\ell)}(\boldsymbol{x}) = \left(\Phi(\boldsymbol{x})^\top A_\ell \Phi(\boldsymbol{x})\right) \rho_{\mathrm{ref}}(\boldsymbol{x}), \tag{4}$$

for some arbitrary orthonormal basis function Φ in $L^2(\rho_{\mathrm{ref}})$. Here, the positivity of the matrix $A_\ell \succeq 0$ ensures the density $\rho^{(\ell)}$ to be positive. Since the α-divergence is defined for general *unnormalized* densities, it is not necessary to know the normalizing constant of π beforehand. α-divergences $D_\alpha(\cdot||\cdot)$ with parameter $\alpha \in \mathbb{R}$ include the Hellinger distance and KL-divergence, which have been used in previous works. The proposed SoS densities permit to efficiently normalize the estimated unnormalized density and to compute the KR map $\mathcal{Q}_\ell$ such that $(\mathcal{Q}_\ell)_\sharp\rho_{\mathrm{ref}} = \rho^{(\ell)}$. This combined use of α-divergence for performing SoS density estimation results in a *convex* optimization problems which can be efficiently solved using off-the-shelf toolboxes.

Second, we extend the methodology to the scenario where *only samples* $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}$ from π are available, as opposed to *point-evaluations* of the target density π. In this scenario, we propose to use diffusion-based bridging densities $\pi^{(\ell)}(\boldsymbol{x})$ where the distribution follows a time–inversed diffusion process such as the Ornstein-Uhlenbeck process with time parameters $t_{\ell-1} \leq t_\ell$ and $t_L = 0$. This idea is at the root of diffusion models [4].

Third, we present a novel convergence analysis using the geometric properties of α-divergences. The analysis is valid both for forward and reverse sequential methods and unifies and extends previous analyses proposed in [5, 3]. More interestingly, it guides the choice of bridging densities. In particular, we show that a smart choice of β_ℓ for tempered densities or of t_ℓ for diffusion-based densities yield a convergence rate of $\mathcal{O}(1/L^2)$ with respect to the number of layer L. While our convergence analysis is valid for any tool to build diffeomorphic maps $\mathcal{Q}$, it makes the assumption that these maps satisfy that

$$D_\alpha\left(\pi^{(\ell)}||(\mathcal{T}_\ell)_\sharp\rho_{\mathrm{ref}}\right) \leq \omega D_\alpha\left(\pi^{(\ell)}||(\mathcal{T}_{\ell-1})_\sharp\rho_{\mathrm{ref}}\right)$$

with $\omega < 1$. It is known that this is in practice hard to achieve. Combining our result, which is mainly focused on the design of schedulers of bridging densities so that $D_\alpha\left(\pi^{(\ell)}||\pi^{(\ell-1)}\right)$ is minimized, with a class of bridging densities and parametrization of $\mathcal{T}$ which has guarantees to achieve $\omega < 1$ for the given bridging density, is an open problem.

Last, we give an outlook for using sequential measure transport to solve optimal transport problems, where we mitigate the difficulty of estimating the optimal coupling by a sequence of entropic regularized problems.

We demonstrate the capability of sequential measure transport methods with our proposed method in unsupervised learning and Bayesian inverse problems in moderate dimension.

References

[1] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, B. Lakshminarayanan, *Normalizing Flows for Probabilistic Modeling and Inference*, Journal of Machine Learning Research **22** (2021).

[2] S. Dolgovm, K. Anaya-Izquierdo, C. Fox, R. Scheichl, *Approximation and sampling of multivariate probability distributions in the tensor train decomposition*, Statistics and Computing **30**(3) (2020), 603–625.

[3] T. Cui, S. Dolgov, *Deep composition of tensor-trains using squared inverse Rosenblatt transports*, Foundations of Computational Mathematics (2022).

[4] A. Nichol, D. Prafulla, *Improved Denoising Diffusion Probabilistic Models*, Journal of Machine Learning Research (2021).

[5] J. Westermann, J. Zech, *Measure transport via polynomial density surrogates*, (arXiv:2311.04172).

[6] T. Cui, S. Dolgov, O. Zahm, *Scalable conditional deep inverse Rosenblatt transports using tensor trains and gradient-based dimension reduction*, Journal of Computational Physics **485** (2023).

Reporter: Elliot Addy

Participants

Elliot Addy
School of Mathematics
University of Edinburgh
James Clerk Maxwell Building
EH9 3FD Edinburgh
UNITED KINGDOM

Prof. Dr. Marc Bocquet
CEREA, École des Ponts ParisTech
and EdF R&D
6-8 Avenue Blaise Pascal
77455 Marne-la-Vallée
FRANCE

Prof. Dr. Alberto Carrassi
Dipartimento di Fisica
Universita degli Studi di Bologna
40127 Bologna
ITALY

Dr. Nisha Chandramoorthy
Dept. of Mathematics, Statistics
and Computer Science, M/C 249
University of Illinois at Chicago
Chicago, IL 60607-7045
UNITED STATES

Prof. Dr. Jana de Wiljes
Institut für Mathematik
Technische Universität Ilmenau
Postfach 100565
98684 Ilmenau
GERMANY

Dr. Svetlana Dubinkina
VU Amsterdam
1081 HV Amsterdam
NETHERLANDS

Dr. Gottfried Hastermann
Institut für Mathematik
Technische Universität Ilmenau
Postfach 100565
98684 Ilmenau
GERMANY

Dr. Oana Lang
Babes-Bolyai University
Faculty of Mathematics and Computer
Science, Department of Mathematics
400157 Cluj-Napoca
ROMANIA

Prof. Dr. Youssef Marzouk
Center for Computational Science
and Engineering
Laboratory for Information
and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge MA 02139
UNITED STATES

Aimee Maurais
Center for Computational Science
and Engineering
Massachusetts Institute of Technology
77 Massachusetts Avenue, 45-421
Cambridge, MA 02139-4307
UNITED STATES

Prof. Dr. Prashant Mehta
Coordinated Science Laboratory
Dept. of Mechanical Science
and Engineering
University of Illinois
at Urbana Champaign
1308 W. Main Street
Urbana, IL 61801
UNITED STATES

Prof. Joaquín Miguez
Departamento de Teoría de la Señal
y Comunicaciones
Universidad Carlos III de Madrid
Avenida de la Universidad, 30
28911 Leganes Madrid
SPAIN

Prof. Dr. Olga Mula
TU Eindhoven
P.O. Box 513
5600 Eindhoven
NETHERLANDS

Dr. Sahani Pathiraja
UNSW Sydney
Sydney NSW 2052
AUSTRALIA

Hans Reimann
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Daniel Sanz-Alonso
University of Chicago
Department of Statistics
5747 S. Ellis Avenue
Chicago IL 60637
UNITED STATES

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

Prof. Dr. Xin Tong
Department of Mathematics
National University of Singapore
10 Lower Kent Ridge Road
Singapore 119076
SINGAPORE

Prof. Dr. Elisabeth Ullmann
Department Mathematik
Technische Universität München
Boltzmannstraße 3
85748 Garching bei München
GERMANY

Prof. Dr. Peter Jan van Leeuwen
Department of Atmospheric Science
Colorado State University
3915 Laporte Ave
Fort Collins CO 80523-1371
UNITED STATES

Prof. Dr. Sven Wang
Institut für Mathematik
Fachbereich Mathematik
Humboldt-Universität Berlin
10099 Berlin
GERMANY

Benjamin Zanger
Laboratoire Jean Kuntzmann
Université Grenoble Alpes
INRIA Grenoble
150 Place du Torrent
Saint-Martin-d'Hères Cedex 9
FRANCE