Transport-Based Variational Bayesian Methods for Learning from Data

Daniele Bigoni[1] Joshua Chen[2] Peng Chen[2] Omar Ghattas[2] Youssef Marzouk[1] Tom O'Leary–Roseberry[2] Keyi Wu[2]

[1]Center for Computational Engineering, Massachusetts Institute of Technology

[2]Institute for Computational Engineering & Sciences, The University of Texas at Austin

AEOLUS

Advances in Experimental Design, Optimization and Learning for Uncertain Complex Systems

Abstract and Motivation

Inverse problems—in which we learn from data through the lens of models—arise across numerous field of science, engineering, technology, and medicine, and in particular our driving applications in advanced manufacturing and materials. Bayesian inference provides a systematic statistical framework for learning from data—e.g., fusing data with models and quantifying uncertainty in the results. Yet inference in large-scale settings—incorporating large multimodal data sets, complex physics-based models, and high-dimensional parameter spaces—remains an enormous computational challenge. Moreover, inference is often only an inner element of "outer loop" analyses such as optimization under uncertainty or optimal experimental design, and hence must be performed repeatedly and quickly. While advanced structure-exploiting sampling methods (e.g., Markov chain Monte Carlo or sequential Monte Carlo methods) that significantly accelerate sampling have been developed in recent years (e.g., [1, 2]), many large-scale complex problems remain out of reach. To overcome these barriers, we are developing new scalable inference strategies that replace *sampling* with *optimization*. In particular, we are advancing variational inference methodologies based on transportation of measures, which describe conditioning via the action of a nonlinear map. Transport maps offer a rich and flexible representation of complex posterior distributions in non-Gaussian settings, along with the ability to continuously trade off accuracy and computational cost. We propose (1) adaptive (semi-)parametric approaches and (2) completely nonparametric approaches for representing maps, each coupled with suitable optimization methods. In both cases, our methods exploit low-dimensional structure: low-dimensional data-informed subspaces, approximate independence, or approximate conditional independence. We demonstrate inference across a spectrum of problems, including inverse problems arising in PDEs, state-space models, and statistical models in machine learning.

Approach

We consider parametric and nonparametric **transport methods** for the solution of Bayesian inference problems. Given an intractable **target/posterior** distribution $\boldsymbol{\nu}_\pi$ on $\mathbb{R}^d$ with unnormalized density $\widetilde{\pi}$ and a tractable **reference** distribution $\boldsymbol{\nu}_\rho$, we seek a map $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushes forward $\boldsymbol{\nu}_\rho$ to $\boldsymbol{\nu}_\pi$, denoted $T_\sharp \boldsymbol{\nu}_\rho = \boldsymbol{\nu}_\pi$. This map renders challenging integration problems tractable:

$$\int f(\boldsymbol{x})\, \boldsymbol{\nu}_\pi(d\boldsymbol{x}) = \int f \circ T(\boldsymbol{x})\, \boldsymbol{\nu}_\rho(d\boldsymbol{x}) \,.$$

The map can be identified as the minimizer of the following variational problem

$$\arg\min_T \mathcal{D}_{\mathsf{KL}} \left(T_\sharp \boldsymbol{\nu}_\rho \| \boldsymbol{\nu}_\pi \right) = \arg\min_T \mathbb{E}_\rho \left[-\log T^\sharp \widetilde{\pi} \right] \tag{1}$$

- A **parametric** formulation [3] approximates the Knothe–Rosenblatt rearrangement within a space $\mathcal{T}_>$ of lower triangular monotone maps:
$$[T(\boldsymbol{x})]_i = T_i(x_1, \ldots, x_i) \,, \qquad \partial_{x_i} T_i > 0 \,.$$
- A **nonparametric** formulation can be obtained as the composition
$$T = S_1 \circ \cdots \circ S_n \qquad \text{of} \qquad S_i = \mathbb{I}_d + Q_i \,,$$
where Q_i belongs to a reproducing kernel Hilbert space.

Both parametric and nonparametric formulations encounter difficulties as the dimension of the problem increases:

- The expectation above is approximated with a quadrature (deterministic or random) whose accuracy deteriorates with dimension.
- Parametric maps can involve bases of exponentially increasing cardinality; this can be mitigated with sparse approximations [4].
- Nonparametric maps can exhibit "mode collapse" in the pushforward distribution, which can be avoided by projecting the map to low-dimensional subspaces [5].

Preliminary results: nonparametric maps

Stein variational Newton (SVN)[6]

For an ansatz representation

$$S_i(\boldsymbol{x}) = \sum_{n=1}^{N} c_n k_n(\boldsymbol{x}),$$

we have the Newton system

$$\mathbb{H}^{\boldsymbol{x}} \boldsymbol{c} = -\boldsymbol{g}^{\boldsymbol{x}},$$

with the gradient and Hessian

$$\boldsymbol{g}^{\boldsymbol{x}}_m = \mathbb{E}^{\mu_i}[-\nabla_{\boldsymbol{x}} \log \widetilde{\pi}\; k_m + \nabla_{\boldsymbol{x}} k_m]$$
$$\mathbb{H}^{\boldsymbol{x}}_{mn} = \mathbb{E}^{\mu_i}[-\nabla^2_{\boldsymbol{x}} \log \widetilde{\pi}\; k_m k_n + \nabla_{\boldsymbol{x}} k_n (\nabla_{\boldsymbol{x}} k_m)^\top].$$

Projection into subspaces

For a data-informed subspace

$$\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_r),$$

we project the parameter as

$$\boldsymbol{x} = \boldsymbol{x}^r + \boldsymbol{x}^\perp, \; \boldsymbol{x}^r = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{x}.$$

By $\boldsymbol{w} := \boldsymbol{\Psi}^\top \boldsymbol{x}$, we have

$$\nabla_{\boldsymbol{w}} \log \widetilde{\pi} = \boldsymbol{\Psi}^\top \nabla_{\boldsymbol{x}} \log \widetilde{\pi}$$
$$\nabla^2_{\boldsymbol{w}} \log \widetilde{\pi} = \boldsymbol{\Psi}^\top \nabla^2_{\boldsymbol{x}} \log \widetilde{\pi} \boldsymbol{\Psi}.$$

Projected SVN (pSVN)[5]

For an ansatz representation

$$S_i(\boldsymbol{w}) = \sum_{n=1}^{N} c_n k_n(\boldsymbol{w}),$$

we have the Newton system

$$\mathbb{H}^{\boldsymbol{w}} \boldsymbol{c} = -\boldsymbol{g}^{\boldsymbol{w}},$$

with the gradient and Hessian

$$\boldsymbol{g}^{\boldsymbol{w}}_m = \mathbb{E}^{\mu_i}[-\nabla_{\boldsymbol{w}} \log \widetilde{\pi}\; k_m + \nabla_{\boldsymbol{w}} k_m]$$
$$\mathbb{H}^{\boldsymbol{w}}_{mn} = \mathbb{E}^{\mu_i}[-\nabla^2_{\boldsymbol{w}} \log \widetilde{\pi}\; k_m k_n + \nabla_{\boldsymbol{w}} k_n (\nabla_{\boldsymbol{w}} k_m)^\top].$$

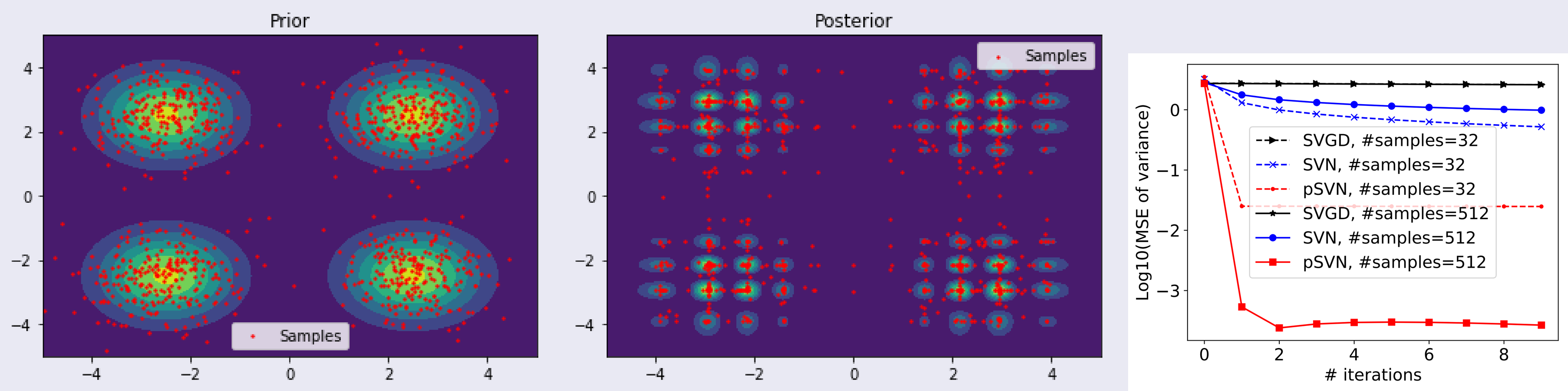


Figure 1: Prior samples (left) and posterior samples by SVN (middle) for a 2-D problem. Comparison of convergence/accuracy of SVGD, SVN, and pSVN for a 1025-D linear problem projected into 5-D.

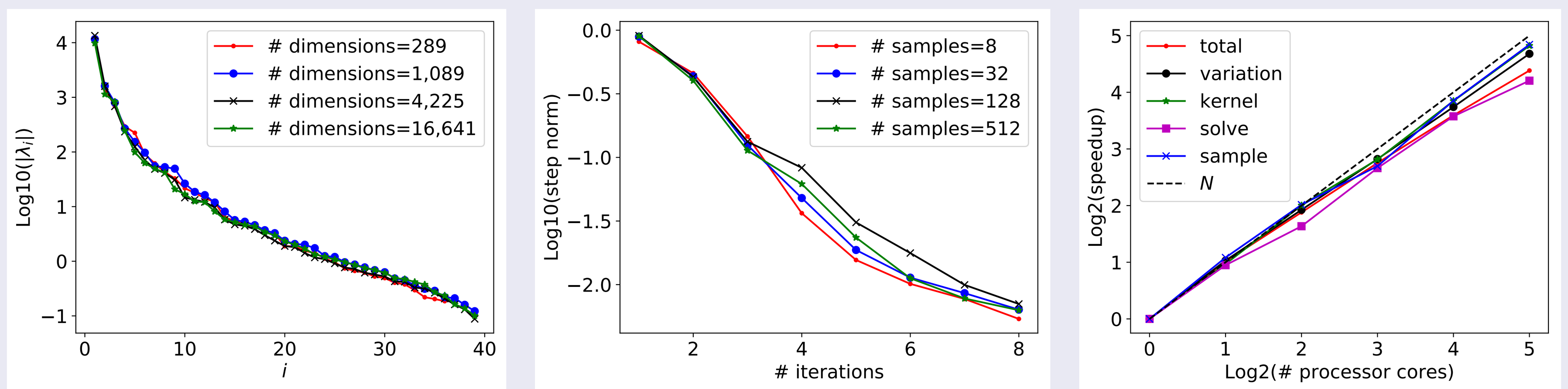
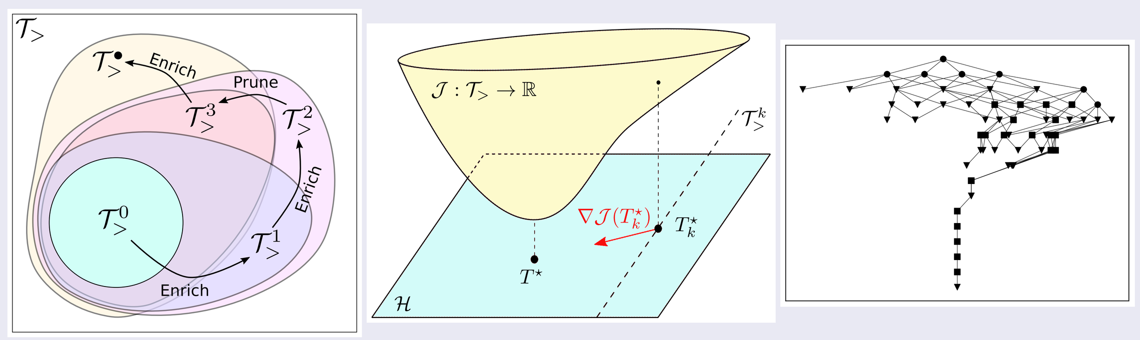


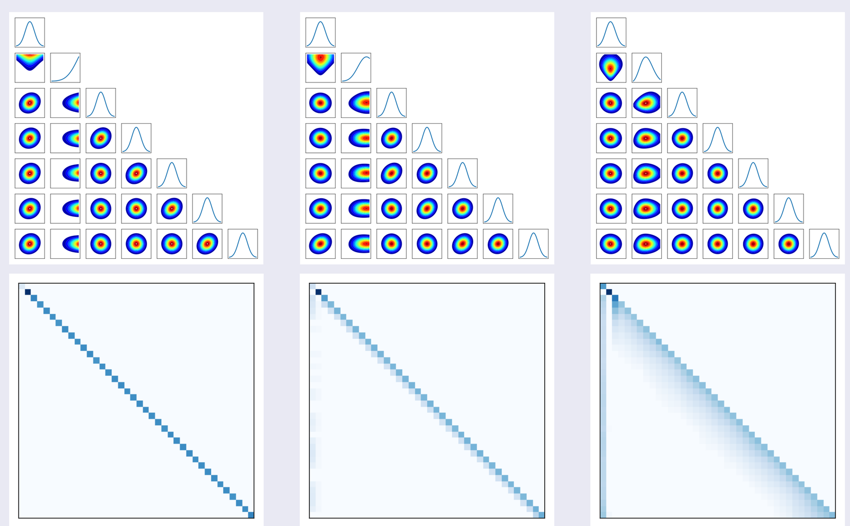
Figure 2: Scalability of pSVN w.r.t. # dimensions, samples, and cores for a nonlinear problem.

Preliminary results: adaptive semi-parametric maps

We seek $\mathcal{T}^\bullet_> \subset \mathcal{T}_>$ $(\dim \mathcal{T}^\bullet_> = n_\bullet)$ s.t.

$$\exists\, T[\boldsymbol{c}] \in \mathcal{T}^\bullet_>, \; \boldsymbol{c} \in \mathbb{R}^{n_\bullet} \quad \text{s.t.}$$
$$\mathbb{V}\left[\log \rho / T^\sharp \pi\right] < \varepsilon_\bullet$$



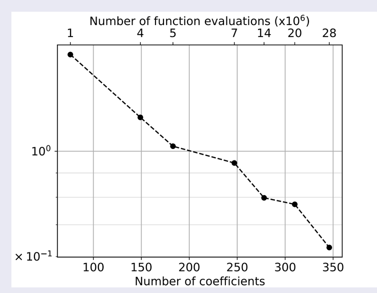


Slices of $T^\sharp \boldsymbol{\nu}_\pi$ (top) and sparsity of $\nabla_{\mathrm{x}} T$ (bottom) as the map improves through iterations 1, 2, and 10.

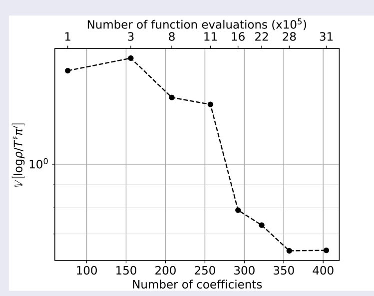
Stochastic volatility (SV) model $(d = 36)$:

$$Z_{n+1} = \mu + \phi(Z_n - \mu) + \eta_n \,,$$
$$Y_n = \varepsilon_n \exp(Z_n/2) \,, n = [1, \ldots, 34]$$

Monte Carlo



Adaptive Sparse Grid



Potential Impact

Inverse and inference problems arise across numerous scientific and technological areas, and address the foundational problem of how we learn from data through the lens of models. In particular, numerous problems of DOE interest fall in this category. For example, in the case of PDEs or ODEs, we can infer: subsurface permeability and contaminant concentrations from well measurements; ice basal boundary conditions from satellite observations of surface velocities; state estimation of oceans from altimetry and ocean probes; material microstructural properties from X-ray scattering data; as-built interior geometry of accelerators from measured EM fields; neutron star merger dynamics from measurements of gravitational fields; biomolecular potentials from dynamics; combustion reaction mechanisms from species concentrations; and so on. Beyond differential equation models, machine learning of graph-based, kernel-based, agent-based, Gaussian process-based, or neural network-based models is fundamentally an inference problem. In all of these, a Bayesian framework is attractive since it is capable of rigorously accosting for uncertainties, given uncertainties in observations, parameters, and the models themselves. The methods we are developing offer the hope of tackling large-scale instances of these, and many more, problems. To make our developments more accessible, they are being incorporated into our open source libraries for inverse problems and uncertainty quantification [7, 8].

Synergy

As mentioned above, the methods developed here are applicable to a broad spectrum of model-based inference-from-data problems. In many cases, we seek to infer infinite-dimensional fields, such as initial conditions, boundary conditions, sources, heterogeneous material properties, or geometry. Upon discretization, these lead to very high dimensional parameter spaces. The scalability of the methods we are developing relies of exploiting the underlying intrinsic low-dimensional structure. One technique we employ to achieve this is through low-rank approximations of Hessians of the log posterior. However some Hessians may not admit a global low rank structure. There is an opportunity to develop other compressed representations—requiring few forward model solves—using for example analytical representations, hierarchical matrices, or product-convolution approximations, which are currently active areas of research.

References

[1] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1525–A1555, 2014.

[2] Tiangang Cui, Kody J.H. Law, and Youssef M. Marzouk. Dimension-independent likelihood-informed MCMC. *Journal of Computational Physics*, 304:109–137, 2016.

[3] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. Sampling via measure transport: an introduction. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of Uncertainty Quantification*. Springer, 2016.

[4] D. Bigoni, J. Chen, P. Chen, O. Ghattas, and Y. Marzouk. Adaptive construction of transport maps via sparse quadrature and sparse polynomial approximation. *In preparation*, 2019.

[5] P. Chen, K. Wu, J. Chen, T. O'Leary-Roseberry, and O. Ghattas. Projected Stein variational Newton: A fast and scalable Bayesian inference method in high dimensions. *Submitted, arXiv:1901.08659*, 2019.

[6] G. Detommaso, T. Cui, A. Spantini, Y. Marzouk, and R. Scheichl. A Stein variational Newton method. In *Advances in Neural Information Processing Systems 31*, pages 9187–9197. 2018.

[7] MUQ: MIT Uncertainty Quantification Library. `http://muq.mit.edu/home`.

[8] U. Villa, N. Petra, and O. Ghattas. hIPPYlib: An extensible software framework for large-scale inverse problems. *Journal of Open Source Software*, 3, 2018.