MATHerials

Inria

École des Ponts ParisTech

Derivative-free Bayesian Inversion Using Multiscale Dynamics

EnKF workshop 2021

Urbain Vaes

Outline of the presentation:

Collaborators

Grigorios Pavliotis
Imperial College London
Department of Mathematics

Andrew Stuart
Department of Computing + Mathematical Sciences

G. A. Pavliotis, A. M. Stuart, and U. Vaes. Derivative-free Bayesian Inversion Using Multiscale Dynamics. arXiv e-prints, 2021

Introduction to inverse problems

Paradigmatic inverse problem

Find an unknown parameter $u \in \mathcal{U}$ from data $y \in \mathbf{R}^m$ where

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

- $\mathcal{G}$ is the forward operator;
- η is observational noise.

Two difficulties[1] associated with this problem are the following:

- Because of the noise, it might be that $y \notin \mathrm{Im}(\mathcal{G})$;
- The problem might be underdetermined.

Additionally, in many PDE applications,

- $\mathcal{G}$ is expensive to evaluate;
- The derivatives of $\mathcal{G}$ are difficult to calculate;
- u is a function $\rightarrow$ infinite dimension.

[1] M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In Handbook of uncertainty quantification. Vol. 1, 2, 3. Springer, Cham, 2017.

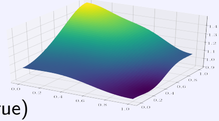
Example: inference of the thermal conductivity in a plate

Mathematical model:

$$
\begin{aligned}
-\nabla \cdot \big(u(x)\nabla T(x)\big) &= f(x), && x \in \Omega,\\
T(x) &= 0, && x \in \partial\Omega.
\end{aligned}
$$

Unknown parameter:

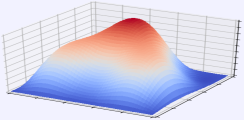
Thermal conductivity $u(x)$



(true)

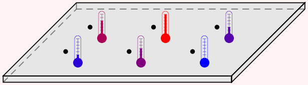
Forward problem

Solution:



Temperature field $T(x)$

Data:

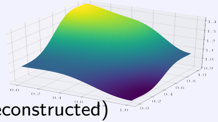


Noisy temperature measurements:

$$y = \big(T(x_1), \ldots, T(x_m)\big) + \eta.$$

Inverse problem

MAP estimator:



(reconstructed)

Optimization approach for solving "$y = \mathcal{G}u + \eta$"

Optimization approach

Find a minimizer of the regularized least-squares functional

$$u^\dagger = \arg\min_{u\in\mathcal{U}} \left(\frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2 + R(u) \right),$$

where $|x|_A^2 := \langle x, x \rangle_A := \langle x, A^{-1}x \rangle$ and $R(u)$ is a regularization term.

- Example regularization (Tikhonov):

$$R(u) = \frac{1}{2}\,|u - m|_\Sigma^2\,.$$

- Modeling step: choice of Γ, m, Σ.

Probabilistic approach for solving "$y = \mathcal{G}u + \eta$"[1]

Bayesian approach to inverse problems

Modeling step:

- Probability distribution on parameter: $u \sim \pi$, encoding our prior knowledge;
- Probability distribution for noise: $\eta \sim \nu$.

An application of Bayes' theorem gives the posterior distribution

$$\rho^y(u) \propto \pi(u)\,\nu\big(y - \mathcal{G}(u)\big) \qquad \text{(valid in finite dimension).}$$

In the Gaussian case where $\pi = \mathcal{N}(m, \Sigma)$ and $\nu = \mathcal{N}(0, \Gamma)$,

$$\rho^y(u) \propto \exp\left(-\left(\frac{1}{2}\,|y - \mathcal{G}(u)|_\Gamma^2 + \frac{1}{2}\,|u - m|_\Sigma^2\right)\right) =: \exp\big(-\Phi_R(u)\big).$$

Two approaches for extracting information:

- Find the maximizer of $\rho^y(u)$ (maximum a posteriori estimation);
- Sample the posterior distribution $\rho^y(u)$.

[1] A. M. Stuart. Inverse problems: a Bayesian perspective. Acta Numer., 2010.

Outline

Inverse problems: optimization and sampling approaches

Solving inverse problems using EnKF

A multiscale derivative-free methodology

EnKF for inverse problems[1,2,3]

Artificial state-estimation problem amenable to EnKF

Dynamical system: $z_{n+1} = \Xi(z_n), \qquad \Xi(z) = \begin{pmatrix} u \\ \mathcal{G}(u) \end{pmatrix}.$

Data model: $y_{n+1} = (0 \;\; I\,) z_{n+1} + \eta_{n+1} = \mathcal{G}(u_{n+1}) + \eta_{n+1}, \qquad \eta_{n+1} \sim \mathcal{N}(0, h^{-1}\Gamma)$

Key idea: reuse the data from the inverse problem: $y_n = y$ for all $n \in \mathbf{N}$.

$\downarrow$ Continuous-time limit $h \to 0$ (viewing h as algorithmic time)

Interacting particle system for optimization (Ensemble Kalman Inversion):

$$\dot{u}^{(j)} = -\frac{1}{J} \sum_{k=1}^{J} \langle \mathcal{G}(u^{(k)}) - \bar{\mathcal{G}}, \mathcal{G}(u^{(j)}) - y \rangle_{\Gamma} (u^{(k)} - \bar{u}), \qquad j = 1, \dots, J,$$

with $\bar{u} = \frac{1}{J} \sum_{j=1}^{J} u^{(j)}$ and $\bar{\mathcal{G}} = \frac{1}{J} \sum_{j=1}^{J} \mathcal{G}(u^{(j)})$.

[1] Y. Chen and D. S. Oliver. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. Math. Geosci., January 2012.

[2] A. A. Emerick and A. C. Reynolds. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. Comput. Geosci., 2013.

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

Ensemble Kalman Inversion in the linear setting[1]

When $\mathcal{G}$ is linear,

$$\begin{aligned}
&\frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}(u^{(k)}) - \bar{\mathcal{G}}, \mathcal{G}(u^{(j)}) - y\rangle_{\Gamma}(u^{(k)} - \bar{u}) \\
&\quad = \frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}(u^{(k)} - \bar{u}), \mathcal{G}(u^{(j)}) - y\rangle_{\Gamma}(u^{(k)} - \bar{u}) \\
&\quad = \frac{1}{J}\sum_{k=1}^{J}\left(\nabla\Phi(u^{(j)})\cdot(u^{(k)} - \bar{u})\right)(u^{(k)} - \bar{u}) = C(U)\nabla\Phi(u^{(j)}),
\end{aligned}$$

with

$$C(U) = \frac{1}{J}\sum_{j=1}^{J}(u^{(j)} - \bar{u})\otimes(u^{(j)} - \bar{u}), \qquad \Phi(u) = \frac{1}{2}\left|\mathcal{G}(u) - y\right|_{\Gamma}^{2}.$$

$\rightarrow$ EKI is a preconditioned gradient descent:

$$\dot{u}^{(j)} = -C(U)\nabla\Phi(u^{(j)}), \qquad j = 1, \dots, J.$$

[1] C. Schillings and A. M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. SIAM J. Numer. Anal., 2017.

Ensemble Kalman Sampling (EKS)[1]

$$\dot{u}^{(j)} = -\frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}(u^{(k)}) - \bar{\mathcal{G}}, \mathcal{G}(u^{(j)}) - y\rangle_{\Gamma}(u^{(k)} - \bar{u}) - C(U)\Sigma^{-1}(u^{(j)} - m) + \sqrt{2C(U)}\,\dot{W}^{(j)}, \qquad j = 1, \dots, J.$$

In the linear setting:

$$\dot{u}^{(j)} = -C(U)\nabla\Phi_R(u^{(j)}) + \sqrt{2C(U)}\,\dot{W}^{(j)}, \qquad j = 1, \dots, J.$$

$\rightarrow$ preconditioned overdamped Langevin dynamics.

Mean field limit $J \to \infty$:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \Big(\mathrm{Cov}(\rho)\,(\nabla\Phi_R\,\rho + \nabla\rho)\Big).$$

[1] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. SIAM J. Appl. Dyn. Syst., 2020.

Main advantages of EKI and EKS:

- They are derivative-free;
- They are based on interacting particle systems;
- They are affine invariant[1] $\rightarrow$ self-preconditioning;
- They have good convergence properties in the linear setting:

Exponential convergence for the EKS mean field equation[2,3]

$$W_2\left(\rho_t, \rho_\infty\right) \leq C\, \mathrm{e}^{-t}\, W_2\left(\rho_0, \rho_\infty\right), \qquad \rho_\infty : \text{Bayesian posterior.}$$

Main limitation

Uncontrolled gradient approximation in the nonlinear case $\rightarrow$ sampling error!

[1] A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. SIAM Journal on Applied Dynamical Systems, 2020.

[2] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. SIAM J. Appl. Dyn. Syst., 2020.

[3] J. A. Carrillo and U. Vaes. Wasserstein stability estimates for covariance-preconditioned Fokker–Planck equations. Nonlinearity, 2021.

Outline

Inverse problems: optimization and sampling approaches

Solving inverse problems using EnKF

A multiscale derivative-free methodology

Uncontrolled gradient approximation for the derivative-free EKS

The derivative-free ensemble Kalman sampler is based on the approximation

$$C(U)\nabla\Phi(u^{(j)}) \approx \frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}(u^{(k)}) - \bar{\mathcal{G}}, \mathcal{G}(u^{(j)}) - y\rangle_{\Gamma}(u^{(k)} - \bar{u}).$$

When the posterior is not Gaussian, this approximation can be inaccurate.

- The method produces approximate posterior samples;
- Can we correct the error?

Our contribution: a derivative free sampling method which

- can be systematically refined to produce accurate posterior samples and
- generalizes an existing derivative-free optimization method[1].

[1] E. Haber, F. Lucka, and L. Ruthotto. Never look back - A modified EnKF method and its application to the training of neural networks without back propagation. arXiv e-prints, May 2018.

A multiscale approach with small parameters σ and δ

$$\dot u = \overbrace{-\frac{1}{J\sigma^2}\sum_{j=1}^{J}\langle G(u^{(j)}) - G(u), G(u) - y\rangle_\Gamma \, (u^{(j)} - u)}^{\text{EnKF approximation of } C(\Xi)\nabla\Phi_R(u)} - C(\Xi)\Sigma^{-1}(u-m) + \sqrt{2}\,\dot W,$$

$$u^{(j)} = u + \sigma\,\xi^{(j)}, \qquad j = 1,\dots,J,$$

$$\dot\xi^{(j)} = -\frac{1}{\delta^2}\,\xi^{(j)} + \sqrt{\frac{2}{\delta^2}}\,\dot W^{(j)}, \qquad \xi^{(j)}(0) \sim \mathcal{N}(0, I_d), \qquad j = 1,\dots,J,$$

where

$$C(\Xi) = \frac{1}{J}\sum_{k=1}^{J}\xi^{(k)} \otimes \xi^{(k)},$$

- $u \in \mathbf{R}^d$: distinguished particle, provides useful information for sampling;
- $(u^{(1)},\dots,u^{(J)})$: collection of "explorers" useful for gradient approximation;
- σ: radius of exploration around the distinguished particle u;
- δ^2: correlation time of the Ornstein–Uhlenbeck processes $\xi^{(j)}$.

$$\mathbf{E}_{X\sim\rho^y}\varphi(X) \approx \frac{1}{T}\int_0^T \varphi\big(u(t)\big)\,\mathrm{d}t, \qquad \rho^y : \text{Bayesian posterior.}$$

A multiscale approach: motivation

When σ is small, it holds with good accuracy that

$$\mathcal{G}(u^{(k)}) - \mathcal{G}(u) \approx \nabla\mathcal{G}(u)(u^{(k)} - u).$$

$\to$ the equation for u reduces to

$$\begin{aligned}\dot{u} &= -\frac{1}{J}\sum_{k=1}^{J}\left(\xi^{(k)} \otimes \xi^{(k)}\right)\nabla\Phi(u) - C(\Xi)\Sigma^{-1}u + \sqrt{2}\dot{W} \\ &= -C(\Xi)\,\nabla\Phi_R(u) + \sqrt{2}\dot{W}.\end{aligned}$$

- $C(\Xi)\,\nabla\Phi_R(u)$ can be viewed as a projection of $\nabla\Phi_R(u)$ on $\mathrm{Span}\{\xi^{(1)}, \dots, \xi^{(J)}\}$.
- **Many-particle limit**: if $J \gg 1$, then

$$C(\Xi) = \frac{1}{J}\sum_{k=1}^{J}\xi^{(k)} \otimes \xi^{(k)} \approx I_d.$$

- **Averaging limit**: if $\delta \ll 1$, then $u(t)$ can be well approximated by the solution to

$$\dot{u} = -\nabla\Phi_R(u) + \sqrt{2}\dot{W}. \qquad \text{(Overdamped Langevin dynamics)}$$

Formal derivation of the averaging limit $\delta \to 0$

Simplified setting: $u \in \mathbf{R}$, quadratic $\Phi_R(u) = \frac{1}{2} k u^2$, one explorer ($J = 1$):

$$\dot{u} = -k\xi^2 u + \sqrt{2}\,\dot{W}^u,$$
$$\dot{\xi} = -\frac{1}{\delta^2}\,\xi + \sqrt{\frac{2}{\delta^2}}\,\dot{W}^\xi.$$

Over a small time interval $\delta^2 \ll \Delta t \ll 1$,

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} k\xi(s)^2 u(s)\,\mathrm{d}s \approx ku(t)\frac{1}{\Delta t}\int_t^{t+\Delta t} \xi(s)^2\,\mathrm{d}s$$
$$\approx ku(t)\int_{\mathbf{R}} \xi^2 \left(\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{\xi^2}{2}}\right)\mathrm{d}\xi = ku(t),$$

by ergodicity of the fast process ξ.

$\to$ When $\delta \ll 1$, the slow process $u(t)$ solves approximately

$$\dot{u} = -ku + \sqrt{2}\,\dot{W}^u.$$

Rigorous averaging result in joint limit $\sigma \to 0$, $\delta \to 0$

Let ϑ denote the solution to

$$\dot{\vartheta} = -\nabla \Phi_R(\vartheta) + \sqrt{2}\dot{W}.$$

Using standard tools from multiscale analysis[1], it is possible to prove

Theorem (Pathwise convergence to an overdamped Langevin dynamics)

Let $p > 1$ and assume that $\mathcal{G} \in C^2(\mathbf{T}^d, \mathbf{R}^K)$. Then for any $T > 0$, there exists a constant $C = C(T, J)$ such that

$$\mathbf{E}\left(\sup_{0 \leq t \leq T} |u_t - \vartheta_t|^p\right) \leq C(\delta^p + \sigma^p).$$

Future research directions:

- Generalization to unbounded domains;
- Convergence of the law in the longtime limit.

[1] G. A. Pavliotis and A. M. Stuart. Multiscale methods. Texts in Applied Mathematics. Springer, New York, 2008. Averaging and homogenization.

To discretize the multiscale system in time, we use

- the Euler–Maruyama method for u;
- the exact solution of the OU process for $\xi^{(j)}$;

$$
\begin{aligned}
\hat{u}_{n+1} = \hat{u}_n - \frac{1}{J\sigma} \sum_{j=1}^{J} \langle \mathcal{G}(\hat{u}_n + \sigma \hat{\xi}_n^{(j)}) - \mathcal{G}(\hat{u}_n), \mathcal{G}(\hat{u}_n) - y \rangle_{\Gamma} \, \hat{\xi}_n^{(j)} \, \Delta \\
- C(\hat{\Xi}_n) \Sigma^{-1} (\hat{u}_n - m) \Delta + \sqrt{2\Delta} \, x_n, \qquad x_n \sim \mathcal{N}(0,1), \\
\hat{\xi}_{n+1}^{(j)} = \mathrm{e}^{-\frac{\Delta}{\delta^2}} \, \hat{\xi}_n^{(j)} + \sqrt{1 - \mathrm{e}^{-\frac{2\Delta}{\delta^2}}} \, x_n^{(j)}, \qquad x_n^{(j)} \sim \mathcal{N}(0,1), \qquad j = 1, \dots, J.
\end{aligned}
$$

Theorem

Assume that $\mathcal{G} \in C^2(\mathbf{T}^d)$. Then there exists $C = C(T,J)$ such that

$$
\sup_{0 \le n \le \lfloor T/\Delta \rfloor} \mathbf{E} \left| \hat{u}_n - \vartheta_{n\Delta} \right|^2 \le C \left(\Delta + \sigma^2 + \log(1 + \delta^{-1}) \, \delta^2 \right).
$$

Importance of preconditioning

Simplified setting:

- Φ_R is quadratic:

$$\Phi_R = \frac{1}{2} |u|_C^2 \,, \qquad C \succ 0.$$

- Explicit Euler for $\dot{u} = -\nabla \Phi_R(u) = -C^{-1} u$:

$$u_{n+1} = (I - \Delta t \, C^{-1}) u_n$$

Stability requires $\Delta t < \lambda_{\min}(C)$! When $\Delta t = \frac{1}{2}\lambda_{\min}(C)$,

$$|u_n| \leq \left| 1 - \frac{1}{2} \frac{\lambda_{\min}(C)}{\lambda_{\max}(C)} \right|^n |u_0|.$$

- Slow convergence when $\lambda_{\min}(C) \ll \lambda_{\max}(C)$!
- Need for preconditioning:

$$\dot{u} = -K \nabla \Phi_R(u), \qquad \text{Optimal preconditioner: } K = C = \operatorname{Cov}\left(\frac{1}{2} \, \mathrm{e}^{-\Phi_R(u)} \right)$$

Improving convergence of the multiscale method with preconditioning

The method can be preconditioned with an appropriate matrix $K \succ 0$.

$$\dot{u} = -\frac{1}{J\sigma^2}\sum_{j=1}^{J}\langle \mathcal{G}(u^{(j)}) - \mathcal{G}(u), \mathcal{G}(u) - y\rangle_{\Gamma}\,(u^{(j)} - u) - C_K(\Xi)\Sigma^{-1}u + \sqrt{2K}\,\dot{W}$$

$$u^{(j)} = u + \sigma\sqrt{K}\xi^{(j)}, \qquad j = 1, \dots, J,$$

$$\dot{\xi}^{(j)} = -\frac{1}{\delta^2}\,\xi^{(j)} + \sqrt{\frac{2}{\delta^2}}\,\dot{W}^{(j)}, \qquad \xi^{(j)}(0) \sim \mathcal{N}(0, I_d), \qquad j = 1, \dots, J,$$

where $C_K(\Xi) := \sqrt{K}\,C(\Xi)\sqrt{K}$.

Formal justification: For small σ,

$$\dot{u} \approx -C_K(\Xi)\nabla\Phi_R + \sqrt{2K}\,\dot{W},$$

which, in the limit $\delta \to 0$, converges to

$$\dot{u} \approx -K\nabla\Phi_R + \sqrt{2K}\,\dot{W}.$$

In practice, we set $K \approx \mathrm{Cov}\left(\frac{1}{Z}\,\mathrm{e}^{-\Phi_R(u)}\right)$ approximated by ensemble Kalman sampling.

Example 1: effect of preconditioning

Here we use the multiscale method to find the minimizer of

$$\Phi(u) = \frac{1}{2}\left(|u_1 - 1|^2 + k^2|u_2 - 1|^2 + k^4|u_3 - 1|^2\right), \qquad k = 5.$$

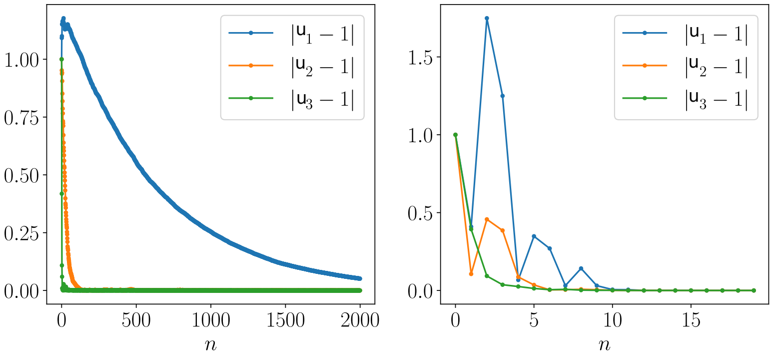


Figure: Error between the iterates and the MAP estimator, without (left) and with (right) preconditioning.

Example 2: two-dimensional elliptic BVP – MAP estimation

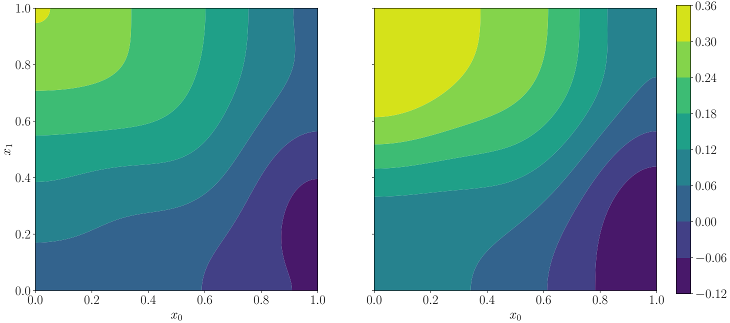
Inference of the conductivity in a plate

Find $u(x)$ from 100 noisy measurements of the temperature $T(x)$ where

$$-\nabla \cdot \left(\mathrm{e}^{u(x)} \nabla T(x)\right) = \mathrm{cst} \qquad x \in D = [0,1]^2, \quad + \text{ homogeneous Dirichlet BC.}$$

Model: $u(x) \sim \mathcal{N}(0, \mathcal{C})$ with $\mathcal{C} = (-\Delta + \tau^2 \mathcal{I})^{-\alpha}$:

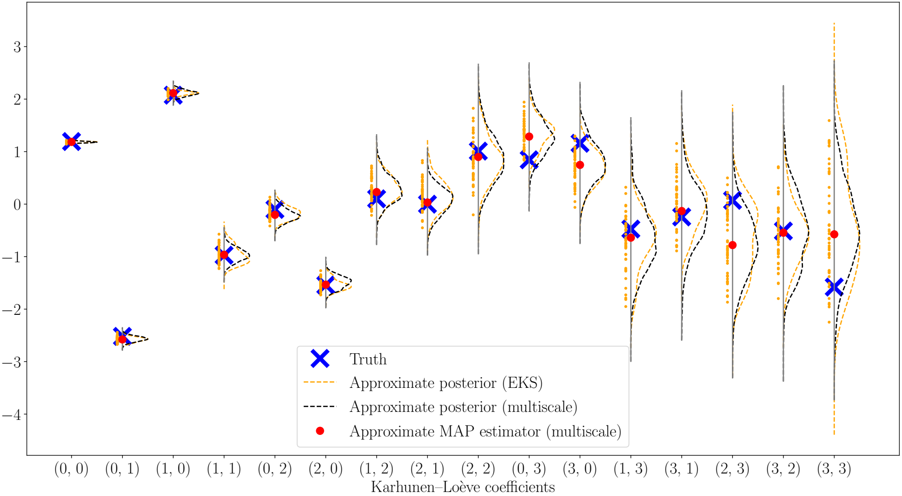
$$\text{KL expansion :} \quad u(x) = \sum u_i \sqrt{\lambda_i} \varphi_i(x), \qquad u_i \sim \mathcal{N}(0,1), \qquad \mathcal{C}\varphi_i = \lambda_i \varphi_i.$$



True (left) and reconstructed (right) log-conductivity ($\delta = \sigma = 10^{-5}$, $J = 8$)

Example 2: two-dimensional elliptic boundary value problem – Sampling

Approximate posterior from 10,000 iterations of the multiscale method:



Conclusions and perspectives for future work

In this presentation, we presented a novel method for sampling and optimization which

- is derivative-free and based on a system of interacting particles;
- is provably refineable over finite time intervals;
- can be preconditioned using information from EnKF methods for efficiency.

Many interesting questions remain open:

- Uniform-in-time weak error estimate;
- Estimate on invariant measure of multiscale system;
- Adaptive σ for computational efficiency;
- Alternative (e.g. semi-implicit) time discretizations.

Thank you for your attention!

References I

J. A. Carrillo and U. Vaes. Wasserstein stability estimates for covariance-preconditioned Fokker–Planck equations. Nonlinearity, 34(4):2275–2295, 2021.

Y. Chen and D. S. Oliver. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. Math. Geosci., 44(1):1–26, January 2012.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In Handbook of uncertainty quantification. Vol. 1, 2, 3, pages 311–428. Springer, Cham, 2017.

A. A. Emerick and A. C. Reynolds. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. Comput. Geosci., 17(2):325–350, 2013.

A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. SIAM J. Appl. Dyn. Syst., 19(1):412–441, 2020.

A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. SIAM Journal on Applied Dynamical Systems, 19(3):1633–1658, 2020.

- E. Haber, F. Lucka, and L. Ruthotto. Never look back - A modified EnKF method and its application to the training of neural networks without back propagation. arXiv e-prints, May 2018. arXiv: 1805.08034 [math.NA].
- M. A. Iglesias, K. J. H. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems. Inverse Problems, 29(4):045001, 20, 2013.
- G. A. Pavliotis and A. M. Stuart. Multiscale methods. Volume 53 of Texts in Applied Mathematics. Springer, New York, 2008, pages xviii+307. Averaging and homogenization.
- G. A. Pavliotis, A. M. Stuart, and U. Vaes. Derivative-free Bayesian Inversion Using Multiscale Dynamics. arXiv e-prints, 2102.00540, 2021.
- C. Schillings and A. M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. SIAM J. Numer. Anal., 55(3):1264–1290, 2017.
- A. M. Stuart. Inverse problems: a Bayesian perspective. Acta Numer., 19:451–559, 2010.