Derivative-free Bayesian Inversion Using Multiscale Dynamics

ENUMATH minisymposium “Inference methodologies for high-dimensional and multiscale stochastic systems”

Urbain Vaes
urbain.vaes@inria.fr

MATHERIALS – Inria Paris & CERMICS – École des Ponts ParisTech

4 September 2025

Collaborators and references

Antonin Della Noce

Inria & École des Ponts

Grigorios Pavliotis

Imperial College London

Department of Mathematics

Andrew Stuart

Department of Computing + Mathematical Sciences

References:

- G. A. Pavliotis, A. M. Stuart, and UV. SIAM J. Appl. Dyn. Syst., 2022
- Ongoing work with A. Della Noce

Paradigmatic inverse problem and its difficulties

Paradigmatic inverse problem

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

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

- $\mathcal{G}$ 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 $\to$ infinite dimension.

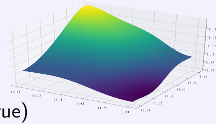
[1] M. Dashti and A. M. Stuart. In Handbook of uncertainty quantification. Vol. 1, 2, 3. Springer, Cham, 2017.

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:

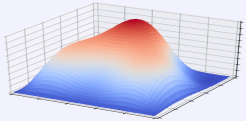
Thermal conductivity $u(x)$



(true)

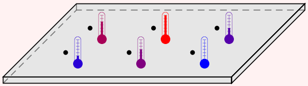
Forward problem

Solution:



Temperature field $T(x)$

Data:

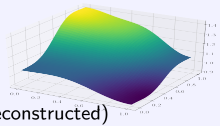


Noisy temperature measurements:

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

Inverse problem

MAP estimator:



(reconstructed)

Probabilistic approach for solving "$y = \mathcal{G}u + \eta$" [2],[3]

Bayesian approach to inverse problems

Modeling step:

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

An application of Bayes' theorem gives the posterior distribution $\rho^y(u) = \mathbf{P}[u|y]$ as

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

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

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

Two approaches for extracting information:

- Find the maximizer of ρ^y (maximum a posteriori estimation);
- Sample the posterior distribution ρ^y **(our focus in this presentation)**.

[2] Jari Kaipio and Erkki Somersalo. Springer-Verlag, New York, 2005.
[3] A. M. Stuart. Acta Numer., 2010.

The Ensemble Kalman Sampler (EKS)[4]

From now on, we drop the tildes so that $\rho^y(u) = \exp\big(-\Phi(u)\big) = \exp\left(\frac{1}{2}\left|y - \mathcal{G}(u)\right|_\Gamma^2\right)$

Ensemble Langevin Sampler $(j = 1, \dots, J)$

Let $\bar{u} := \frac{1}{J}\sum_{j=1}^J u^{(j)}$ and $C(U) = \frac{1}{J}\sum_{j=1}^J \big(u^{(j)} - \bar{u}\big) \otimes \big(u^{(j)} - \bar{u}\big)$

$$\mathrm{d}u^{(j)} = -C(U)\nabla\Phi(u^{(j)})\,\mathrm{d}t + \sqrt{2C(U)}\,\mathrm{d}W^{(j)}$$

Gradient-free approximation: with $\bar{\mathcal{G}} := \frac{1}{J}\sum_{j=1}^J \mathcal{G}(u^{(j)})$,

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

Ensemble Kalman Sampler $(j = 1, \dots, J)$

$$\mathrm{d}u^{(j)} = -\frac{1}{J}\sum_{k=1}^J \left\langle \mathcal{G}\big(u^{(k)}\big) - \bar{\mathcal{G}}, \mathcal{G}\big(u^{(j)}\big) - y \right\rangle_\Gamma \big(u^{(k)} - \bar{u}\big)\,\mathrm{d}t + \sqrt{2C(U)}\,\mathrm{d}W^{(j)}$$

---

[4] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. SIAM J. Appl. Dyn. Syst., 2020.

Properties of the Ensemble Kalman Sampler (EKS)

- Good behavior for $\mathcal{G}(\theta) = \mathcal{G}_0(\theta) + \mathcal{G}_1(\theta/\varepsilon)$ with $\varepsilon \ll 1$[5]
- Mean field limit in the linear case[6],[7]

$$\begin{cases} \mathrm{d}\overline{u}_t = -\mathcal{C}(\rho_t)\nabla\Phi(\overline{u}_t)\,\mathrm{d}t + \sqrt{2\mathcal{C}(\rho_t)}\,\mathrm{d}W_t, \\ \rho_t = \mathrm{Law}(\overline{u}_t)\,. \end{cases} \qquad \text{(McKean SDE)}$$

- Exponential convergence of mean field solution to the posterior in the linear case[8]

$$W_2(\rho_t, \rho^y) \leqslant C\, W_2(\rho_0, \rho^y)\, \mathrm{e}^{-t}$$

- The method is affine invariant[9]

Main limitation

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

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

---

[5] O. Dunbar, A. B. Duncan, A. M. Stuart, and M.-T. Wolfram. SIAM J. Appl. Dyn. Syst., 2022.
[6] A. Garbuno-Inigo, N. Nüsken, and S. Reich. SIAM Journal on Applied Dynamical Systems, 2020.
[7] Z. Ding and Q. Li. SIAM J. Math. Anal., 2021.
[8] J. A. Carrillo and UV. Nonlinearity, 2021.
[9] A. Garbuno-Inigo, N. Nüsken, and S. Reich. SIAM Journal on Applied Dynamical Systems, 2020.

Towards a multiscale derivative-free methodology

Setting: we consider the task of generating samples from

$$\rho^y(u) := \exp\big(-\Phi(u)\big) \quad \text{where} \quad \Phi(u) = \big|y - \mathcal{G}(u)\big|_\Gamma^2$$

without employing the derivatives of $\mathcal{G}$

Objective: devise and mathematically analyze a gradient-free sampling method which

- exploits the sum-of-squares structure of Φ
- can be systematically refined to produce accurate samples from ρ^y
- is robust to scale separation in the multiscale setting $\mathcal{G}(x) = \mathcal{G}_0(x) + \mathcal{G}_1(x/\varepsilon)$.

Ideas from the optimization literature (1/2)

The Zero-th order Perturbed Stochastic Gradient Descent[10] to minimize Φ:

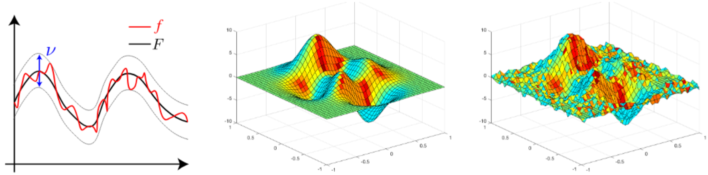
- **Iteration:** Sample $\xi_n^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_{d_u})$ and set

$$\frac{u_{n+1} - u_n}{\eta} = -\frac{1}{J\sigma^2} \sum_{i=1}^{m} \xi_n^{(i)} \Big(\Phi(u_n + \xi_n^{(i)}) - \Phi(u_n) \Big) + \text{ noise}$$

- **Idea:** For fixed x and $z \sim \mathcal{N}(0, \sigma^2 I_{d_u})$, it holds that

$$\mathbf{E}\big[\nabla \Phi(x+z)\big] = \frac{1}{\sigma^2} \mathbf{E}\Big[z \big(\Phi(x+z) - \Phi(x) \big) \Big]$$

$\rightsquigarrow$ when $J \gg 1$, method is a gradient descent for loss $\Phi \star \mathcal{N}(0, \sigma^2 I_{d_u})$



[10] C. Jin, L. T. Liu, R. Ge, and M. I. Jordan. Adv. Neural Inf. Process., 2018.

In[11], a similar method is proposed exploiting the sum-of-squares structure of Φ:

$$\frac{u_{n+1} - u_n}{\eta} = -\frac{1}{J\sigma^2} \sum_{j=1}^{J} \left\langle \mathcal{G}(u_n^{(j)}) - \mathcal{G}(u), \mathcal{G}(u) - y \right\rangle_\Gamma (u_n^{(j)} - u)$$

$$u_n^{(j)} = u_n + \sigma \xi_n^{(j)}, \qquad \xi_n^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_u}).$$

Motivation:

- Gradient approximation when $\mathcal{G}$ is linear

$$\frac{1}{J\sigma^2} \sum_{j=1}^{J} \left\langle \mathcal{G}\big(u_n^{(j)}\big) - \mathcal{G}(u), \mathcal{G}(u) - y \right\rangle_\Gamma = \left(\frac{1}{J} \sum_{j=1}^{J} \xi_n^{(j)} \otimes \xi_n^{(j)} \right) \nabla \Phi(u)$$

- Good gradient approximation when σ is small
- Works more generally for $\Phi(u) = \ell\big(\mathcal{G}(u)\big)$

[11] E. Haber, F. Lucka, and L. Ruthotto. arXiv e-prints, May 2018.

A multiscale approach to sample $\rho^y(u) := \exp\left(-|y - \mathcal{G}(u)|_\Gamma^2\right)$

Multiscale sampler with small parameters σ and δ ($j = 1, \dots, J$)

Continuous-time dynamics

$$\mathrm{d}u = -\frac{1}{J\sigma^2} \sum_{j=1}^{J} \left\langle \mathcal{G}(u^{(j)}) - \mathcal{G}(u), \mathcal{G}(u) - y \right\rangle_\Gamma (u^{(j)} - u)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W$$

$$u^{(j)} = u + \sigma\, \xi^{(j)}$$

$$\mathrm{d}\xi^{(j)} = -\frac{1}{\delta^2}\, \xi^{(j)}\, \mathrm{d}t + \sqrt{\frac{2}{\delta^2}}\, \mathrm{d}W^{(j)} \qquad \xi^{(j)}(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_u})$$

where

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

Desired output: for large t, it holds approximately that $u(t) \sim \rho^y$. Relatedly

$$\mathbf{E}_{\rho^y}[\varphi] \approx \frac{1}{T} \int_0^T \varphi\big(u(t)\big)\, \mathrm{d}t$$

A multiscale approach: motivation

When σ is small, it holds with good accuracy that

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

$\rightarrow$ the equation for u reduces to

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

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

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

- **Averaging limit**: if $\sigma \ll 1$ and $\delta \ll 1$, then $u(t)$ approximately satisfies

$$\mathrm{d}u = -\nabla \Phi(u) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W. \qquad \text{(Overdamped Langevin dynamics)}$$

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

Let $\mathfrak{u}$ denote the solution to

$$\mathrm{d}\mathfrak{u} = -\nabla\Phi(\mathfrak{u})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W, \qquad \mathfrak{u}_0 = u_0.$$

Using standard tools from multiscale analysis[12], it is possible to prove

Theorem (Pathwise convergence to an overdamped Langevin dynamics)

Let $p \geqslant 1$ and assume that $\mathcal{G} \in C^2(\mathbf{T}^{d_u}, \mathbf{R}^K)$. Then for all $T > 0$ there is $C = C(T)$ s.t.

$$\mathbf{E}\left[\sup_{0 \leqslant t \leqslant T} |u_t - \mathfrak{u}_t|^p\right] \leqslant C\left(\frac{\delta^p}{J^{\frac{p}{2}}} + \sigma^p\right).$$

Ideas of the proof:

- convergence w.r.t. to δ^2 classical averaging approach[13]
- convergence w.r.t. to σ: Taylor expansions
- convergence w.r.t. to J: Law of large numbers in L^p

[12] G. A. Pavliotis and A. M. Stuart. Springer, New York, 2008.
[13] G. A. Pavliotis and A. M. Stuart. Springer, New York, 2008.

Discretization in time

Consider a dsicretization of the multiscale system in time based on

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

$$\widehat{u}_{n+1} = \widehat{u}_n - \frac{1}{J\sigma} \sum_{j=1}^{J} \left\langle \mathcal{G}(\widehat{u}_n + \sigma \widehat{\xi}_n^{(j)}) - \mathcal{G}(\widehat{u}_n), \mathcal{G}(\widehat{u}_n) - y \right\rangle_\Gamma \widehat{\xi}_n^{(j)} \, \Delta + \sqrt{2\Delta}\, x_n$$

$$\widehat{\xi}_{n+1}^{(j)} = \mathrm{e}^{-\frac{\Delta}{\delta^2}} \, \widehat{\xi}_n^{(j)} + \sqrt{1 - \mathrm{e}^{-\frac{2\Delta}{\delta^2}}} \, x_n^{(j)}$$

where $x_n \sim \mathcal{N}(0,1)$ and $x_n^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Theorem (Discrete strong convergence, ongoing work with A. Della Noce)

Assume that $\mathcal{G} \in C^2(\mathbf{T}^{d_u}, \mathbf{R}^{d_y})$. Then for all T there exists $C = C(T)$ such that

$$\mathbf{E}\left[\sup_{0 \leqslant n \leqslant \lfloor T/\Delta \rfloor} |\widehat{u}_n - \mathfrak{u}_{n\Delta}|^p \right] \leqslant C \left(\frac{\delta^p}{J^{\frac{p}{2}}} + \Delta^{\frac{p}{2}} + \sigma^p \right).$$

Existence and convergence of invariant measure (continuous-time)

Proposition (Convergence of stationary measures, ongoing work with A. Della Noce)

Assume that $\mathcal{G} \in C^3(\mathbf{T}^{d_u}, \mathbf{R}^{d_y})$. Then

- *There exists a unique invariant measure $\mu_{\delta,\sigma,J}$ for the multiscale dynamics*
- *Let $\nu_{\delta,\sigma}$ denote the u-marginal of $\mu_{\delta,\sigma,J}$. Then for all $f \in W^{2,\infty}(\mathbf{T}^{d_u})$*

$$\left| \int_{\mathbf{T}^{d_u}} f(u)\, \nu_{\delta,\sigma,J}(\mathrm{d}u) - \int_{\mathbf{T}^{d_u}} f(u)\, \rho^y(\mathrm{d}u) \right| \leqslant C \|f\|_{W^{2,\infty}(\mathbf{T}^{d_u})} \left(\frac{\delta^2}{J} + \sigma \right)$$

Ideas of proof:

- Existence of $\mu_{\delta,\sigma,J}$ by classical "Lyapunov + minorization" technique[14]
- Weak convergence of $\nu_{\delta,\sigma,J}$ by technique from[15]: define $\varphi \in L^2_0(\rho^y)$ as the solution to $-\mathcal{L}_{\mathrm{ovd}}\varphi = f - \mathbf{E}_{\rho^y} f$ and notice that

$$\mathbf{E}_{\mu_{\delta,\sigma,J}}[f] = \mathbf{E}_{\mu_{\delta,\sigma,J}}\Big[(\mathcal{L}_{\mathrm{ovd}} - \mathcal{L}_{\delta,\sigma,J})\varphi\Big]$$

where $\mathcal{L}_{\delta,\sigma,J}$ and $\mathcal{L}_{\mathrm{ovd}}$ are the generators of the multiscale and limiting dynamics

[14] M. Hairer and J. C. Mattingly. In Seminar on Stochastic Analysis, Random Fields and Applications VI. Birkhäuser/Springer Basel AG, Basel, 2011.

[15] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. SIAM J. Numer. Anal., 2010.

Uniform-in-time weak convergence

Proposition (Convergence of stationary measures, ongoing work with A. Della Noce)

Assume that $\mathcal{G} \in C^{4}(\mathbf{T}^{d_u}, \mathbf{R}^{d_y})$*. Then for all* $f \in W^{3,\infty}(\mathbf{T}^{d_u})$

$$\forall t \geqslant 0, \qquad \Big|\mathbf{E}\big[f(u_t)\big] - \mathbf{E}\big[f(\mathfrak{u}_t)\big]\Big| \leqslant C \|f\|_{W^{3,\infty}(\mathbf{T}^{d_u})} \left(\frac{\delta^2}{J} + \sigma\right)$$

- Let $\mathcal{L}_{\delta,\sigma,J}$ denote the generator of the multiscale dynamics
- Let $\mathcal{L}_{\mathrm{ovd}} = -\nabla\Phi \cdot \nabla + \Delta$ denote the generator of the limiting Langevin

Idea of the proof: classical approach in weak error analysis, with first step

$$\begin{aligned} \mathrm{e}^{t\mathcal{L}_{\delta,\sigma,J}} f - \mathrm{e}^{t\mathcal{L}_{\mathrm{ovd}}} f &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}s} \left(\mathrm{e}^{s\mathcal{L}_{\delta,\sigma,J}} \, \mathrm{e}^{(t-s)\mathcal{L}_{\mathrm{ovd}}} f\right) \mathrm{d}s \\ &= \int_0^t \mathrm{e}^{s\mathcal{L}_{\delta,\sigma,J}} (\mathcal{L}_{\delta,\sigma,J} - \mathcal{L}_{\mathrm{ovd}}) \, \mathrm{e}^{(t-s)\mathcal{L}_{\mathrm{ovd}}} f \, \mathrm{d}s \end{aligned}$$

$+$ estimates on the solution of parabolic equation $\partial_t v = \mathcal{L}_{\mathrm{ovd}} v$

Improving convergence of the multiscale method with preconditioning

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

$$\mathrm{d}u = -\frac{1}{J\sigma^2} \sum_{j=1}^{J} \left\langle \mathcal{G}(u^{(j)}) - \mathcal{G}(u), \mathcal{G}(u) - y \right\rangle_\Gamma (u^{(j)} - u)\,\mathrm{d}t + \sqrt{2K}\,\mathrm{d}W$$

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

$$\mathrm{d}\xi^{(j)} = -\frac{1}{\delta^2}\,\xi^{(j)}\,\mathrm{d}t + \sqrt{\frac{2}{\delta^2}}\,\mathrm{d}W^{(j)}, \qquad \xi^{(j)}(0) \sim \mathcal{N}(0, I_{d_u}), \qquad j = 1, \dots, J,$$

Formal justification: For small σ,

$$\mathrm{d}u \approx -C_K(\Xi)\nabla\Phi\,\mathrm{d}t + \sqrt{2K}\,\mathrm{d}W$$

where $C_K(\Xi) := \sqrt{K}\,C(\Xi)\sqrt{K}$. In the limit $\delta \to 0$, this converges to

$$\mathrm{d}u \approx -K\nabla\Phi + \sqrt{2K}\,\mathrm{d}W$$

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

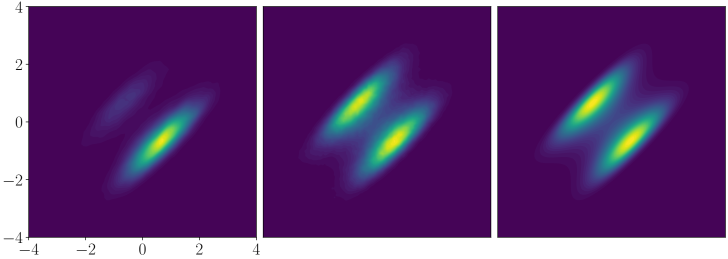
Example 1: Bimodal target distribution

Inverse problem with bimodal posterior

Find $u = (u_1, u_2) \in \mathbf{R}^2$ from

$$y = |u_1 - u_2|^2 + \eta, \qquad \eta \sim \mathcal{N}(0, 1).$$

Prior distribution $u \sim \mathcal{N}(0, I_2)$. Below $y = 2$.



Left: Approximate posterior using EKS.
Middle: Approximate posterior using multiscale method.
Right: True Bayesian posterior.

Example 2: two-dimensional elliptic BVP – MAP estimation

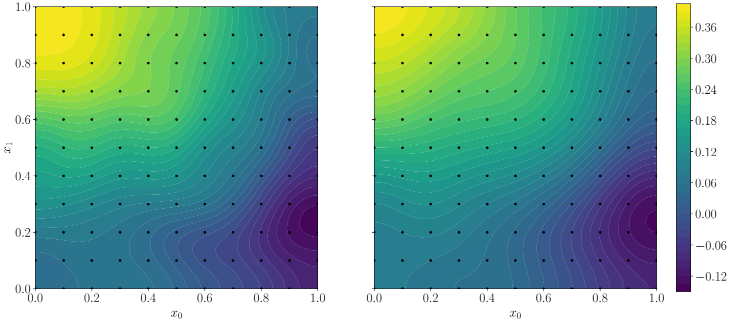
Inference of the conductivity in a plate

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

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

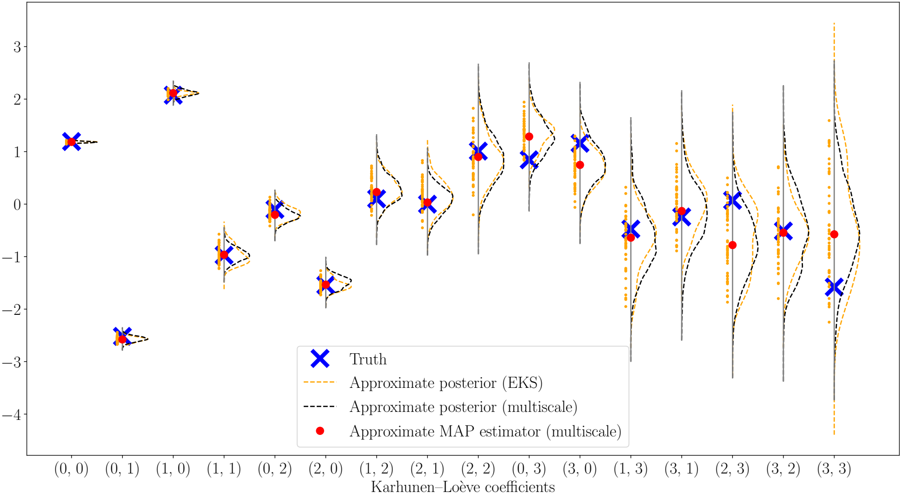
Model: $u(x) \sim \mathcal{N}(0, \mathcal{C})$ with $\mathcal{C} = (-\Delta + \tau^2 \mathcal{I})^{-\alpha}$:

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



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

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:

- Adaptive σ for computational efficiency;
- Unbounded state space
- Alternative (e.g. semi-implicit) time discretizations.
- Alternative derivative-free methodologies.

Thank you for your attention!