Consensus-based sampling

Rencontres Inria-LJLL

Urbain Vaes
urbain.vaes@inria.fr

MATHERIALS – Inria Paris & CERMICS – École des Ponts ParisTech

18 October 2021

Collaborators

José Carrillo

UNIVERSITY OF OXFORD

Mathematical Institute

Franca Hoffmann

UNIVERSITÄT BONN

Hausdorff Center for Mathematics

Andrew Stuart

Department of Computing + Mathematical Sciences

Reference: J. A. Carrillo, F. Hoffmann, A. M. Stuart, and UV. Consensus Based Sampling. arXiv e-prints, 2021

Outline

Motivating example: Bayesian inverse problems

Paradigmatic inverse problem

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

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

- $\mathcal{G}$ 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,

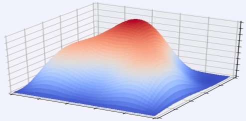
- $\mathcal{G}$ is expensive to evaluate;
- The derivatives of $\mathcal{G}$ are difficult to calculate;
- θ is a function $\rightarrow$ infinite dimension.

[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.

Mathematical model:

$$-\nabla \cdot \big(\theta(x)\nabla T(x)\big) = f(x), \qquad x \in \Omega,$$
$$T(x) = 0, \qquad x \in \partial\Omega.$$

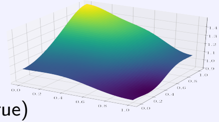
Solution:



Temperature field $T(x)$

Unknown parameter:

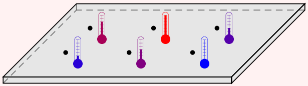
Thermal conductivity $\theta(x)$



(true)

Forward problem

Data:

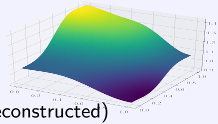


Noisy temperature measurements:

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

Inverse problem

MAP estimator:



(reconstructed)

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

Bayesian approach to inverse problems

Modeling step:

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

An application of Bayes' theorem gives the posterior distribution:

$$\rho^y(\theta) \propto \pi(\theta)\, \nu\big(y - \mathcal{G}(\theta)\big) = \text{prior} \times \text{likelihood}.$$

(In infinite dimension, use Radon–Nikodym derivative.)

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

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

Two approaches for extracting information:

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

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

Brief review of the recent literature on interacting particle methods

- 2006: Sequential Monte Carlo[3];
- 2010: Affine-invariant many-particle MCMC[4];
- 2013: Ensemble Kalman inversion[5];
- 2016: Stein variational gradient descent[6];
- 2017: Consensus-based optimization[7];
- 2020: Ensemble Kalman sampling[8];

Often parallelizable, and some can be studied through mean-field equations.

---

[3] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. J. R. Stat. Soc. Ser. B Stat. Methodol., 2006.

[4] J. Goodman and J. Weare. Ensemble samplers with affine invariance. Commun. Appl. Math. Comput. Sci., 2010.

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

[6] Q. Liu and D. Wang. Stein variational gradient descent: a general purpose Bayesian inference algorithm. In Advances In Neural Information Processing Systems, 2016.

[7] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. Math. Models Methods Appl. Sci., 2017.

[8] 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.

Our starting point: consensus-based optimization (CBO)[9]

CBO is an Optimization method based on the interacting particle system

$$\mathrm{d}\theta_t^{(j)} = -\Big(\theta_t^{(j)} - \mathcal{M}_\beta(\mu_t^J)\Big)\,\mathrm{d}t + \sqrt{2}\sigma\Big|\theta_t^{(j)} - \mathcal{M}_\beta(\mu_t^J)\Big|\mathrm{d}W_t^{(j)}. \qquad j = 1, \dots, J,$$

where $\mathcal{M}_\beta(\mu_t^J)$ is given by

$$\mathcal{M}_\beta(\mu_t^J) = \frac{\int \theta\, \mathrm{e}^{-\beta f(\theta)}\, \mu_t^J(\mathrm{d}\theta)}{\int \mathrm{e}^{-\beta f(\theta)}\, \mu_t^J(\mathrm{d}\theta)} = \frac{\sum_{j=1}^J \theta_t^{(j)} \exp\big(-\beta f(\theta_t^{(j)})\big)}{\sum_{j=1}^J \exp\big(-\beta f(\theta_t^{(j)})\big)}, \qquad \mu_t^J = \frac{1}{J}\sum_{j=1}^J \delta_{\theta_t^{(j)}}.$$

Properties:

- Mean-field limit:

$$\partial_t \mu = \nabla \cdot \Big(\big(\theta - \mathcal{M}_\beta(\mu)\big)\mu\Big) + \sigma^2 \triangle\Big(\big|\theta - \mathcal{M}_\beta(\mu)\big|^2 \mu\Big).$$

- Convergence of the mean field solution: if f has a unique global minimizer,

$$\mathcal{M}_0(\mu_t) \xrightarrow[t\to\infty]{} \widehat{\theta}(\beta), \qquad \widehat{\theta}(\beta) \xrightarrow[\beta\to\infty]{} \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d} f(\theta).$$

[9] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. Math. Models Methods Appl. Sci., 2017.

Key tool for the analysis of CBO: Laplace's method

Laplace's method can be employed for studying the limit as $\beta \to \infty$ of the integral

$$I_\beta(\varphi) = \frac{\displaystyle\int_{\mathbf{R}^d} \varphi(\theta)\, \mathrm{e}^{-\beta f(\theta)}\, \mu(\mathrm{d}\theta)}{\displaystyle\int_{\mathbf{R}^d} \mathrm{e}^{-\beta f(\theta)}\, \mu(\mathrm{d}\theta)} =: \int_{\mathbf{R}^d} \varphi\, \mathrm{d}(\mathcal{R}_\beta \mu), \qquad \mathcal{R}_\beta : \mu \mapsto \frac{\mu\, \mathrm{e}^{-\beta f}}{\int \mu\, \mathrm{e}^{-\beta f}}.$$

Let $\theta_* = \arg\min f$. Under appropriate assumptions, it holds[10],[11]

$$I_\beta(\varphi) = \int_{\mathbf{R}^d} \varphi\, \mathrm{d}g_\beta + \mathcal{O}\left(\frac{1}{\beta^2}\right) \qquad \text{as } \beta \to \infty.$$

where $g_\beta = \mathcal{N}\Big(\theta_*, \beta^{-1}\big(\mathrm{Hess}\, f(\theta_*)\big)^{-1}\Big)$. In other words $\mathcal{R}_\beta \mu \approx g_\beta$ for large β.

Motivation:

$$\mathrm{e}^{-\beta f(\theta)} \approx \mathrm{e}^{-\beta\left(f(\theta_*) + \frac{1}{2}\, \mathrm{Hess}\, f(\theta_*) : \left((\theta - \theta_*) \otimes (\theta - \theta_*)\right)\right)}$$

[10] P. D. Miller. Applied asymptotic analysis. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2006.

[11] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. An analytical framework for consensus-based global optimization method. Mathematical Models and Methods in Applied Sciences, 2018.

Outline

Can we construct a sampling method using ideas from CBO?

Notation: $\mathcal{M}_\beta$ weighted mean, $\mathcal{C}_\beta$ weighted covariance, $\mathcal{R}_\beta$ reweighting:

$$\mathcal{M}_\beta(\mu) = \mathcal{M}(\mathcal{R}_\beta \mu)\,, \quad \mathcal{C}_\beta(\mu) = \mathcal{C}(\mathcal{R}_\beta \mu)\,, \quad \mathcal{R}_\beta : \mu \mapsto \frac{\mu\, \mathrm{e}^{-\beta f}}{\int \mu\, \mathrm{e}^{-\beta f}}\,,$$

$$\mathcal{M}(\mu) = \int \theta \mu(\mathrm{d}\theta)\,, \quad \mathcal{C}(\mu) = \int \bigl(\theta - \mathcal{M}(\mu)\bigr) \otimes \bigl(\theta - \mathcal{M}(\mu)\bigr) \mu(\mathrm{d}\theta)\,.$$

Discrete-time consensus-based sampling ($\beta \geq 0$)

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta(\mu_n) + \alpha\bigl(\theta_n - \mathcal{M}_\beta(\mu_n)\bigr) + \sqrt{\gamma \mathcal{C}_\beta(\mu_n)}\, \xi_n, & \xi_n \sim \mathcal{N}(0, I_d), \\ \mu_n = \mathrm{Law}(\theta_n). \end{cases}$$

We first assume $\mathrm{e}^{-f} = \mathcal{N}(a, A)$.

Question: Are there choices of (α, β, γ) such that $\mu_n = \mathrm{e}^{-f}$ is a steady state?

Determining the parameters

Discrete-time consensus-based sampling ($\beta \geq 0$)

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta(\mu_n) + \alpha\big(\theta_n - \mathcal{M}_\beta(\mu_n)\big) + \sqrt{\gamma \mathcal{C}_\beta(\mu_n)}\, \xi_n, & \xi_n \sim \mathcal{N}(0, I_d), \\ \mu_n = \mathrm{Law}(\theta_n). \end{cases}$$

A simple explicit calculation shows that

$$\mathcal{M}_\beta(\mathrm{e}^{-f}) = a,$$
$$\mathcal{C}_\beta(\mathrm{e}^{-f}) = (1+\beta)^{-1} A.$$

If $\theta_n \sim \mathcal{N}(a, A)$, then

$$\theta_{n+1} \sim \mathcal{N}(a, \alpha^2 A + \gamma (1+\beta)^{-1} A).$$

Therefore $\mathrm{e}^{-f} = \mathcal{N}(a, A)$ is a steady state if

$$\alpha \in [-1, 1], \qquad \gamma = (1 - \alpha^2)(1+\beta).$$

For what parameters is the target $\mathcal{N}(a, A)$ an attractor?

If $\theta_n \sim \mathcal{N}(m_n, C_n)$, then a calculation shows $\theta_{n+1} \sim \mathcal{N}(m_{n+1}, C_{n+1})$ with

$$m_{n+1} = \alpha m_n + (1-\alpha)\left(C_n^{-1} + \beta A^{-1}\right)^{-1}\left(\beta A^{-1} a + C_n^{-1} m_n\right),$$
$$C_{n+1} = \alpha^2 C_n + \gamma\left(C_n^{-1} + \beta A^{-1}\right)^{-1},$$

For e^{-f} to be an attractor for Gaussian initial conditions, we need in fact $\alpha \in (-1, 1)$.

Convergence result for target $\mathcal{N}(a, A)$ and Gaussian initial condition

If $\alpha \in (-1, 1)$ and $\gamma = (1-\alpha^2)(1+\beta)$, then

$$|m_n - a| + \|C_n - A\| \leq C\left(\frac{1-|\alpha|}{1+\beta} + |\alpha|\right)^n$$

Questions:

- Is $\mathcal{N}(a, A)$ an attractor for non-Gaussian initial conditions?
- What if the target e^{-f} is not Gaussian?

We will (partially) answer the second question.

Particle approximation of the mean-field dynamics

In practice, we approximate the mean-field equation by a particle system:

$$\theta_{n+1}^{(j)} = \mathcal{M}_\beta(\mu_n^J) + \alpha\big(\theta_n^{(j)} - \mathcal{M}_\beta(\mu_n^J)\big) + \sqrt{\gamma \mathcal{C}_\beta(\mu_n^J)}\, \xi_n^{(j)}, \qquad j = 1, \dots, J.$$

Here $\Theta_n = \{\theta_n^{(j)}\}_{j=1}^J$ is a set of particles and

$$\mu_n^J := \frac{1}{J} \sum_{j=1}^J \delta_{\theta_n^{(j)}}$$

is the associated empirical measure.

Motivation: if $\Theta_0 \sim \mu_0^{\otimes J}$ and $J \gg 1$, then it holds approximately $\Theta_n \sim \mu_n^{\otimes J}$, so

$$\mathcal{M}_\beta(\mu_n^J) \approx \mathcal{M}_\beta(\mu_n), \qquad \mathcal{C}_\beta(\mu_n^J) \approx \mathcal{C}_\beta(\mu_n),$$

by the law of large numbers.

Invariant subspace property[12]: $\mathrm{Span}\{\theta_n^{(j)}\}_{j=1}^J \subset \mathrm{Span}\{\theta_0^{(j)}\}_{j=1}^J$.

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

Affine invariance[13],[14],[15]

The CBS dynamics is affine invariant. We denote by

$$\mathrm{CBS}_n(\mu_0; \rho)$$

the law of θ_n when CBS is used to sample from ρ with initial condition $\theta_0 \sim \mu_0$.

It holds for any invertible affine transformations $T : \mathbf{R}^d \to \mathbf{R}^d$ that

$$\mathrm{CBS}_n\big(T_\sharp(\mu_0); T_\sharp(\rho)\big) = T_\sharp\big(\mathrm{CBS}_n(\mu_0; \rho)\big).$$

- Good performance for ill-conditioned targets;
- If $\mathrm{e}^{-f} = \mathcal{N}(a, A)$, then the convergence rate is independent of a and A.

[13] J. Goodman and J. Weare. Ensemble samplers with affine invariance. Commun. Appl. Math. Comput. Sci., 2010.

[14] B. Leimkuhler, C. Matthews, and J. Weare. Ensemble preconditioning for Markov chain Monte Carlo simulation. Stat. Comput., 2018.

[15] A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. SIAM J. Appl. Dyn. Syst., 2020.

Interpretation as discretization of McKean SDE

When $\alpha = \mathrm{e}^{-\Delta t}$ with $\Delta t \ll 1$, the CBS dynamics

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta(\mu_n) + \alpha\big(\theta_n - \mathcal{M}_\beta(\mu_n)\big) + \sqrt{\gamma \mathcal{C}_\beta(\mu_n)}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0, I_d), \\ \mu_n = \mathrm{Law}(\theta_n). \end{cases}$$

may be viewed as a discretization with time step Δt of the McKean SDE

$$\begin{cases} \mathrm{d}\theta_t = -\big(\theta_t - \mathcal{M}_\beta(\mu_t)\big)\,\mathrm{d}t + \sqrt{2(1+\beta)\mathcal{C}_\beta(\mu_t)}\,\mathrm{d}W_t, \\ \mu_t = \mathrm{Law}(\theta_t) \end{cases}$$

$\rightarrow$ Continuous-time sampling method with similar properties:

- Steady state is e^{-f} in the Gaussian setting;
- Exponential convergence in the Gaussian target/Gaussian initial condition setting:

$$|m_t - a| + \|C_t - A\| \leq C \exp\left(-\frac{\beta}{1+\beta} t\right)$$

Analysis beyond the Gaussian setting

We consider for simplicity the continuous-time dynamics:

$$\begin{cases} \mathrm{d}\theta_t = -\big(\theta_t - \mathcal{M}_\beta(\mu_t)\big)\,\mathrm{d}t + \sqrt{2(1+\beta)\mathcal{C}_\beta(\mu_t)}\,\mathrm{d}W_t, \\ \mu_t = \mathrm{Law}(\theta_t). \end{cases}$$

The law μ of θ_t evolves according to

$$\partial_t \mu = \nabla \cdot \Big(\big(\theta - \mathcal{M}_\beta(\mu)\big)\mu + (1+\beta)\mathcal{C}_\beta(\mu)\,\nabla\mu\Big).$$

- This dynamics propagates Gaussians even when e^{-f} is non-Gaussian;
- Any steady state must satisfy

$$\mu_\infty = \mathcal{N}\big(\mathcal{M}_\beta(\mu_\infty), (1+\beta)\mathcal{C}_\beta(\mu_\infty)\big).$$

$\rightarrow$ No convergence to e^{-f} in the case of a non-Gaussian target.

Let us introduce

$$\widehat{f}(\theta) = f(\theta_*) + \frac{1}{2} \operatorname{Hess} f(\theta_*) : \big((\theta - \theta_*) \otimes (\theta - \theta_*)\big).$$

The distribution $\mathrm{e}^{-\widehat{f}} \propto \mathcal{N}(\theta_*, C_*)$ is the Laplace approximation of e^{-f}.

Convergence result

Under appropriate assumptions (one-dimensional, convex),

- There exists a unique steady-state $\mathcal{N}\big(m_\infty(\beta), C_\infty(\beta)\big)$ satisfying

$$\big|m_\infty(\beta) - \theta_*\big| + \big\|C_\infty(\beta) - C_*\big\| = \mathcal{O}(\beta^{-1}).$$

- If the initial condition is Gaussian, then

$$\big|m(t) - m_\infty\big| + \big\|C(t) - C_\infty\big\| \leq C \exp\left(-\left(1 - \frac{k}{\beta}\right) t\right).$$

Idea of the proof: Laplace's method, then contraction argument.

Application to optimization

With the parameter choice $\gamma = (1-\alpha^2)$, we obtain an optimization method.

Discrete-time optimization variant:

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta(\mu_n) + \alpha\big(\theta_n - \mathcal{M}_\beta(\mu_n)\big) + \sqrt{(1-\alpha^2)\mathcal{C}_\beta(\mu_n)}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0, I_d), \\ \mu_n = \mathrm{Law}(\theta_n). \end{cases}$$

Continuous-time optimization variant:

$$\begin{cases} \mathrm{d}\theta_t = -\big(\theta_t - \mathcal{M}_\beta(\mu_t)\big)\,\mathrm{d}t + \sqrt{2\mathcal{C}_\beta(\mu_t)}\,\mathrm{d}W_t, \\ \mu_t = \mathrm{Law}(\theta_t) \end{cases}$$

Convergence result for the optimization method

If $\theta_0 \sim \mathcal{N}(m_0, C_0)$ and under appropriate assumptions (one-dimenisonal, convex),

$$W_2(\mu_n, \delta_{\theta_*}) \leq C n^{-p}, \qquad W_2(\mu_t, \delta_{\theta_*}) \leq C t^{-p}, \qquad p \in (0,1).$$

The convergence is slow but exact, which is an advantage compared to CBO as in[16].

[16] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. Math. Models Methods Appl. Sci., 2017.

Accelerating the optimization method by adapting β dynamically

Consider the case $\alpha = 0$ for simplicity:

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta(\mu_n) + \sqrt{\mathcal{C}_\beta(\mu_n)}\,\xi_n, & \xi_n \sim \mathcal{N}(0, I_d), \\ \mu_n = \mathrm{Law}(\theta_n). \end{cases}$$

We define the effective sample size for an ensemble $\Theta = \{\theta^{(j)}\}_{j=1}^J$ as

$$J_{\mathrm{eff}}(\Theta) := \frac{\left(\sum_{j=1}^J \omega_j\right)^2}{\sum_{j=1}^J |\omega_j|^2}, \qquad \omega_j := \mathrm{e}^{-\beta f(\theta^{(j)})}.$$

- If β is too large, the ensemble collapses to a point in 1 iteration;
- If β is small, the convergence is slow;
- If β is constant, $J_{\mathrm{eff}}(\Theta_n) \xrightarrow[n\to\infty]{} J$ and the weights become very close.

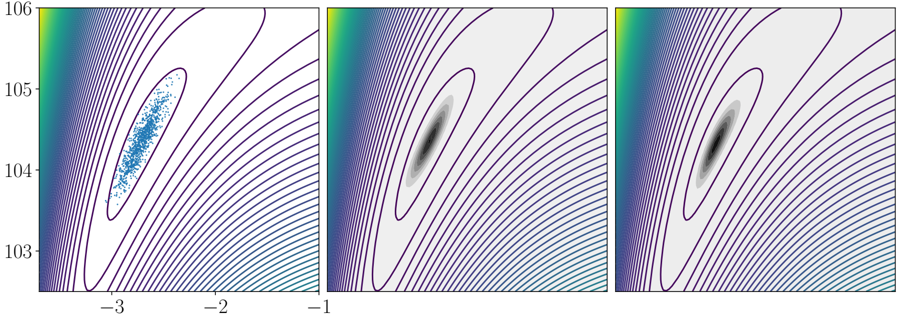
Idea: Take $\beta = \beta(n)$ such that $J_{\mathrm{eff}}/J = \eta \in (0,1)$ for all n.

Outline

Find $(\theta_1, \theta_2) \in \mathbf{R}^2$ from noisy observations of $\big(p(.25), p(.75)\big) \in \mathbf{R}^2$, where $p(x)$ solves

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{\theta_1}\,\frac{\mathrm{d}p}{\mathrm{d}x}\right) = 1, \qquad x \in [0,1],$$

with boundary conditions $p(0) = 0$ and $p(1) = \theta_2$.



Left: Particles at iteration $n = 100$ for fixed $\alpha = \beta = \frac{1}{2}$. **Middle:** Corresponding Gaussian density. **Right:** Bayesian posterior.

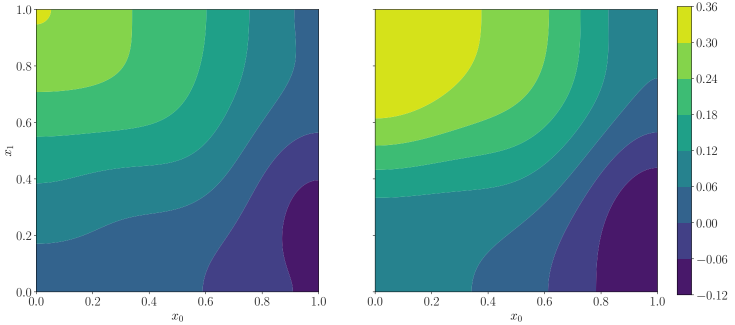
Example 2: Two-dimensional elliptic BVP – MAP estimation

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

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

Model: $\theta(x) \sim \mathcal{N}(0, \mathcal{C})$ in $L^2(D)$ where $\mathcal{C}^{-1} = (-\Delta + \tau^2 \mathcal{I})^{\upsilon}$ [17]

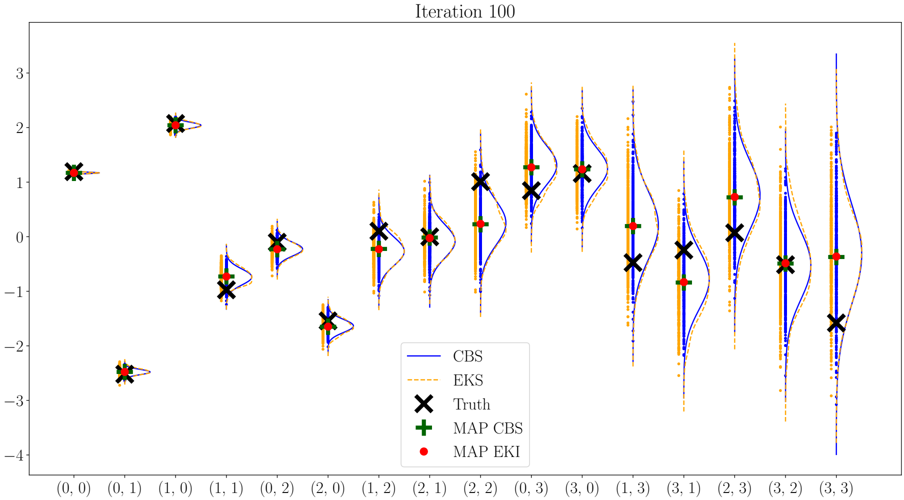
$$\text{KL expansion : } \quad \theta(x) = \sum \theta_i \sqrt{\lambda_i}\varphi_i(x), \qquad \theta_i \sim \mathcal{N}(0,1), \qquad \mathcal{C}\varphi_i = \lambda_i \varphi_i.$$



True (left) and reconstructed (right) log-conductivity ($\alpha = 0$, $J_{\mathrm{eff}}/J = .5$, $J = 512$)

[17] equipped with homogeneous Neumann boundary condition on the space of mean-zero functions.

Approximate posterior after 100 iterations of CBS with $\alpha = 0$, adaptive β, and $J = 512$.



Optimization: objective functions

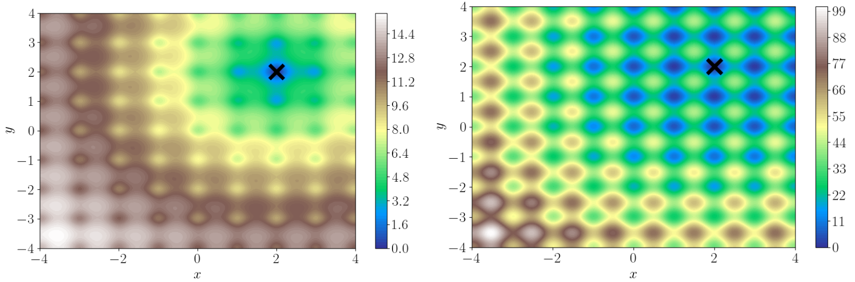
- the Ackley function, defined for $x \in \mathbf{R}^d$ by

$$f_A(x) = -20 \exp\left(-\frac{1}{5}\sqrt{\frac{1}{d}\sum_{i=1}^{d}|x_i - b|^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos\big(2\pi(x_i - b)\big)\right) + \mathrm{e} + 20,$$

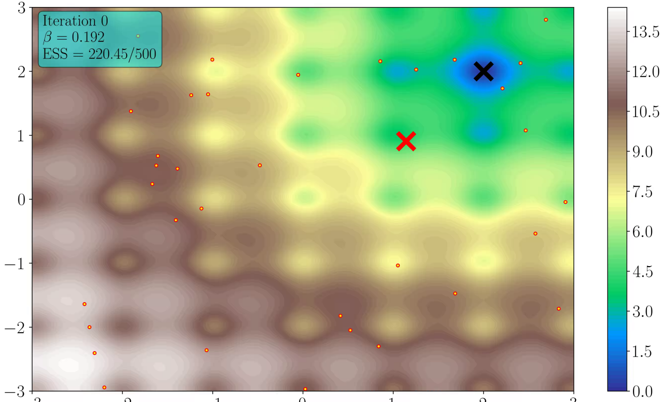
- the Rastrigin function, defined by

$$f_R(x) = \sum_{i=1}^{d}\Big((x_i - b)^2 - 10\cos\big(2\pi(x_i - b)\big) + 10\Big).$$

Minimizer: $x_* = (b, \dots, b)$, where $b \in \mathbf{R}$. Below $b = 2$.



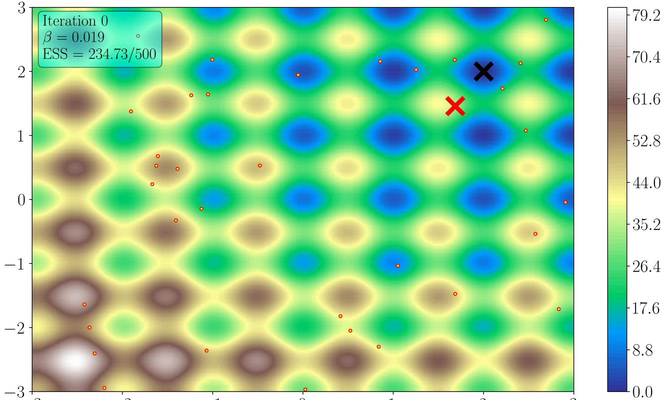
Convergence for $\alpha = .1$, adaptive β with $J_{\text{eff}}/J = .5$, and $J = 100$.



Ackley function

Optimization: illustration of the convergence

Convergence for $\alpha = .1$, adaptive β with $J_{\text{eff}}/J = .5$, and $J = 100$.



Rastrigin function

Outline

Towards exact sampling

Like ensemble Kalman-based methods, CBS is exact only for Gaussian targets.

How can we generate exact samples from e^{-f}?

- **Idea 1.** Metropolize CBS, i.e. construct a Markov chain such that:

$$\left\{\theta_n^{(j)}\right\}_{j=1}^{J} \xrightarrow[n\to\infty]{\text{Law}} \rho^{\otimes J}.$$

Example proposal: select $j_* \sim \mathcal{U}\{1, J\}$ and propose

$$\theta_{n+1}^{*,(j)} = \begin{cases} \mathcal{M}_\beta + \alpha\left(\theta_n^{(j)} - \mathcal{M}_\beta\right) + \sqrt{\gamma \mathcal{C}_\beta}\,\xi_n^{(j)} & \text{if } j = j_*, \\ \theta_n^{(j)} & \text{if } j \neq j_*, \end{cases}$$

with $\xi_n^{(j)} \sim \mathcal{N}(0, I_d)$ and

$$\mathcal{M}_\beta = \mathcal{M}_\beta(\mu_n^J), \qquad \mathcal{C}_\beta = \mathcal{C}_\beta(\mu_n^J), \qquad \mu_n^J = \frac{1}{J}\sum_{j=1}^{J} \delta_{\theta_n^{(j)}}.$$

Then accept or reject using Metropolis–Hastings' method.

- **Idea 2.** Use information from CBS to precondition a standard MCMC method:

$$\theta_{n+1}^* = m_\infty + \lambda\left(\theta_n - m_\infty\right) + \sqrt{1-\lambda^2}\xi_n, \qquad \xi_n \sim \mathcal{N}(0, xC_\infty).$$

where $x \geq 1$ and $\mathcal{N}(m_\infty, C_\infty)$ is the approximate posterior from CBS.

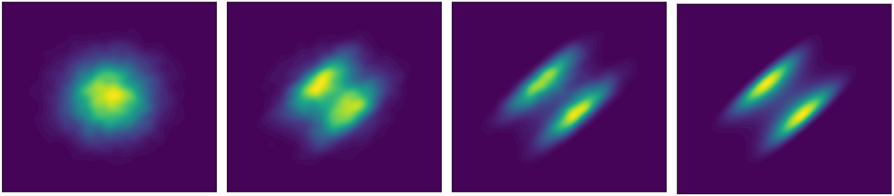
Metropolization: proof of concept

We consider the bimodal example from[18], arising from the Bayesian inverse problem with

$$G : \mathbf{R}^2 \ni \theta \mapsto |\theta_1 - \theta_2|^2 \in \mathbf{R},$$

noise $\eta \sim \mathcal{N}(0,\, I_2)$, prior distribution $\mathcal{N}(0, I_2)$, and data $y = 2$.

Empirical distribution of metropolized CBS with $J = 10^4$ particles and $\alpha = \frac{1}{2}$:



(a) $n = 10^3$ (b) $n = 10^4$ (c) $n = 10^5$ (d) Exact

In practice, expectations could be estimated using time and ensemble averages:

$$\mathbf{E}_{\theta \sim \rho} h(\theta) = \mathbf{E}_{\Theta \sim \rho^{\otimes J}} \left(\frac{1}{J} \sum_{j=1}^{J} h(\theta^{(j)}) \right) \approx \frac{1}{JN} \sum_{n=1}^{N} \sum_{j=1}^{J} h(\theta_n^{(j)})$$

[18] S. Reich and S. Weissmann. Fokker-Planck particle systems for Bayesian inference: computational approaches. SIAM/ASA J. Uncertain. Quantif., 2021.

Preconditioning: proof of concept

Assumption: target distribution has density with respect to $\pi = \mathcal{N}(m, C)$.

The MCMC method from[19] is based on the proposal

$$\theta^*_{n+1} = m + \lambda(\theta_n - m) + \sqrt{1-\lambda^2}\,\xi_n, \qquad \xi_n \sim \mathcal{N}(0, C),$$

- The method is robust with respect to refinement (e.g. more KL modes);
- The acceptance probability is close to 1 if the target is close to $\mathcal{N}(m, C)$ but
- it can be $\ll 1$ for very anisotropic targets.

A variation on this method using information from CBS:

- Calculate a Gaussian approximation $\mathcal{N}(m_\infty, C_\infty)$ of the posterior using CBS;
- Use a modification of the above method based on the proposal

$$\theta^*_{n+1} = m_\infty + \lambda\,(\theta_n - m_\infty) + \sqrt{1-\lambda^2}\xi_n, \qquad \xi_n \sim \mathcal{N}(0, xC_\infty), \qquad x \geq 1.$$

- Leads to better acceptance probability (≈ 1 if target is Gaussian and $\alpha = 1$).

[19] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. Statist. Sci., 2013.

Conclusions

The proposed method

- can be used for sampling or optimization;
- is based on ideas from consensus-based optimization;
- is based on a stochastic interacting particle system:
  - can be parallelized easily;
  - can be studied from a mean field viewpoint.
- is derivative-free, so well suited for PDE inverse problems;
- converges exponentially fast at the mean-field level (for sampling);
- is affine-invariant, so convergence rate is independent of target in Gaussian setting.

Perspectives:

- Can we study the method with adaptive β?
- Can we prove convergence at the particle level[20]?
- Can we correct the sampling error in the non-Gaussian setting[21]?

[20] A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. SIAM J. Appl. Dyn. Syst., 2020.

[21] E. Cleary, A. Garbuno-Inigo, S. Lan, T. Schneider, and A. M. Stuart. Calibrate, emulate, sample. J. Comp. Phys., 2021.

Thank you for your attention!