Consensus Based Sampling

Dynamics and Discretization: PDEs, Sampling, and Optimization

Franca Hoffmann

Hausdorff Center for Mathematics,
Rheinische Friedrich-Wilhelms-Universität Bonn

28 October 2021

Collaborators

José Carrillo

Mathematical Institute

Urbain Vaes

MATHERIALS – Inria Paris

Andrew Stuart

Department of Computing + Mathematical Sciences

Reference: J. A. Carrillo, F. Hoffmann, A.M. Stuart, and U. Vaes. **Consensus Based Sampling.** Studies in Applied mathematics, 2021.

Numerics: `https://figshare.com/s/8b1a068a63999a6c7e45`

Outline

Motivating example: Bayesian inverse problems

Paradigmatic inverse problem[1]

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, $\mathcal{G} : \mathbb{R}^d \mapsto \mathbb{R}^K$.
- η is observational noise, $\eta \sim \mathsf{N}(0, \gamma^2 I)$.

In many PDE applications,

- Calibration & Uncertainty Quantification;
- $\mathcal{G}$ is expensive to evaluate;
- The derivatives of $\mathcal{G}$ are not available.

[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.

Bayesian approach to inverse problems

Modeling step:

- Probability distribution on parameter: $\theta \sim \pi_0$, encoding our prior knowledge;
- Probability distribution for noise" $\mathbb{P}(y|\theta)$ with $y - \mathsf{G}(\theta) \sim \mathsf{N}(0, \gamma^2 I)$ likelihood

An application of Bayes' theorem gives the posterior distribution:

$$\rho^y(\theta) \propto \mathbb{P}(y|\theta)\,\pi_0(\theta) = \text{prior} \times \text{likelihood}.$$

In the Gaussian case where $\pi_0 = \mathcal{N}(m, \Sigma_0)$ and Gaussian noise,

$$\rho^y(\theta) \propto \exp\left(-\left(\frac{1}{2\gamma^2}\,|y - \mathcal{G}(\theta)|^2 + \frac{1}{2}\,|\theta - m|^2_{\Sigma_0}\right)\right) =: \exp\bigl(-f(\theta)\bigr).$$

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

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

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

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

Properties:

- Mean-field limit:

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

- 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.

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:

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

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

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

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

- Evolve particle ensemble: derivative-free algorithm

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

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

Determining the parameters

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

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

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

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

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

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

Questions:

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

Interpretation as discretization of McKean SDE

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

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

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

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

$\to$ 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:

$$|m_t - a|_A + \|C_t - A\|_A^{1/2} \leq C \exp\left(-\left(\frac{\beta}{1+\beta}\right)t\right)$$

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.

Convergence of the solution

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

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

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:

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

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),

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

Outline

Example 1: one-dimensional elliptic BVP – Sampling

Find $(\theta_1, \theta_2) \in \mathbf{R}^2$ from noisy observations of $(p(.25), p(.75)) \in \mathbf{R}^2$, 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 $p(1) = \theta_2$.

- Explicit solution $p(x, \theta)$ is available
- We define

$$G(\theta) = \begin{pmatrix} p(x_1, \theta) \\ p(x_2, \theta) \end{pmatrix}.$$

- Contour plots: $f(\theta) = \frac{1}{2\gamma^2}|y - G(\theta)|^2 + \frac{1}{2\sigma^2}|\theta|^2$

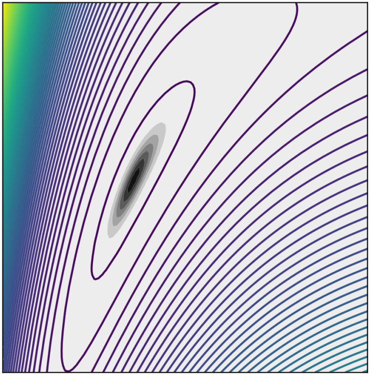


Figure: Contour plots.

Optimization: objective functions

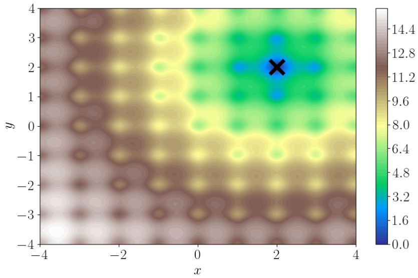
- the Ackley function, defined for $x \in \mathbf{R}^d$ by

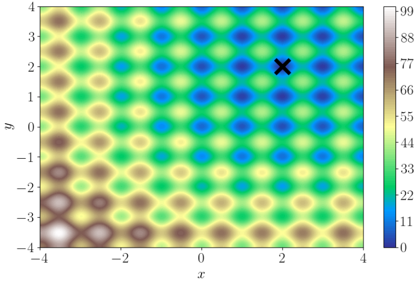
$$f_A(x) = -20 \exp\left(-\frac{1}{5}\sqrt{\frac{1}{d}\sum_{i=1}^{d}|x_i - b|^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos\big(2\pi(x_i - b)\big)\right) + \mathrm{e} + 20,$$

- the Rastrigin function, defined by

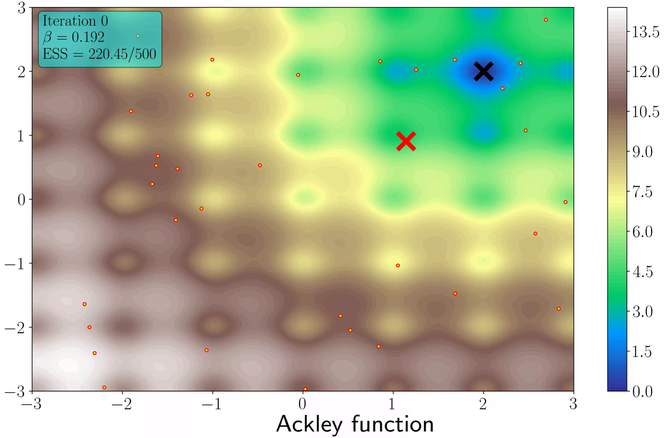
$$f_R(x) = \sum_{i=1}^{d}\Big((x_i - b)^2 - 10\cos\big(2\pi(x_i - b)\big) + 10\Big).$$

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

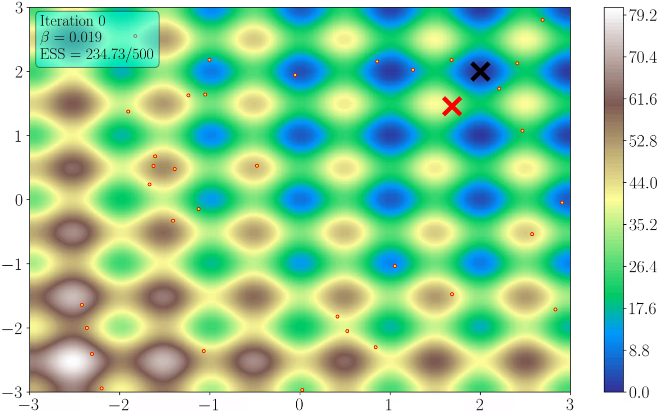




Convergence for $\alpha = .1$, adaptive β with $J_{\text{eff}}/J = .5$, and $J = 100$.



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



Rastrigin function

Outline

Overview: Gaussian Target $f(\theta) = \frac{1}{2}|\theta - a|_A^2$

- Denote $k_0 = \|A^{1/2}C_0^{-1}A^{1/2}\|_2$
- Gaussian initial condition with strictly positive definite covariance.

	Sampling		Optimization	
	Mean	Covariance	Mean	Covariance
$\alpha = 0$	$\left(\frac{1}{1+\beta}\right)^n$	$\left(\frac{1}{1+\beta}\right)^n$	$\frac{k_0}{k_0+\beta n}$	$\frac{k_0}{k_0+\beta n}$
$\alpha \in (0,1)$	$\left(\frac{1+\alpha\beta}{1+\beta}\right)^n$	$\left(\frac{1+\alpha^2\beta}{1+\beta}\right)^n$	$\left(\frac{k_0+\beta}{k_0+\beta+\beta(1-\alpha^2)n}\right)^{\frac{1}{1+\alpha}}$	$\frac{k_0+\beta}{k_0+\beta+\beta(1-\alpha^2)n}$
$\alpha = 1$	$\mathrm{e}^{-\left(\frac{\beta}{1+\beta}\right)t}$	$\mathrm{e}^{-\left(\frac{2\beta}{1+\beta}\right)t}$	$\left(\frac{k_0+\beta}{k_0+\beta+2\beta t}\right)^{\frac{1}{2}}$	$\frac{k_0+\beta}{k_0+\beta+2\beta t}$

Overview: Gaussian Target $f(\theta) = \frac{1}{2}|\theta - a|_A^2$

- Optimization mode: algebraic convergence.
- Sampling mode: exponential convergence.

$\longrightarrow$ this is analogous to what is known about the EKI and EKS methods.

- Sharp convergence rates.
- Discrete time: smaller choices of α provide a faster rate of convergence.

$\longrightarrow$ choosing $\alpha = 0$ is therefore the most favorable choice in this regard.

- Larger β increases the speed of convergence, without limit as $\beta \to \infty$ for $\alpha = 0$;
- In the case $\alpha > 0$, increasing β is favourable but does not give rates which increase without limit.

Unique Attractor $N(a, A)$

- The mean-field dynamics admit infinitely many steady states given by all Dirac distributions and $\mathcal{N}(a, A)$,
- Convergence to $\mathcal{N}(a, A)$ starting from Gaussians.

Overview: Beyond Gaussians

- $f \in C^2(\mathbf{R}^d)$ and
$$\ell I_d \le L \le \mathrm{Hess}\, f(\theta) \le U \le u I_d\,,$$
for all $\theta \in \mathbf{R}^d$ and some $\ell, u > 0$.
- Let $k > 0$ independent of n, t, α and β, denote
$$\widetilde{k}_0 := \|L^{1/2} C_0^{-1} L^{1/2}\|_2 \qquad q > 2\max(2, u/\ell)\,.$$
- Gaussian initial condition with strictly positive definite covariance.

	Sampling		Optimization	
	Mean ($d = 1$)	Covariance ($d = 1$)	Mean ($d = 1$)	Covariance (any d)
$\alpha = 0$	$\left(\frac{k}{\beta}\right)^n$	$\left(\frac{k}{\beta}\right)^n$	$\lesssim \frac{\log(n)}{n}$	$\frac{\widetilde{k}_0}{\widetilde{k}_0 + \beta n}$
$\alpha \in (0,1)$	$\left(\alpha + (1-\alpha^2)\frac{k}{\beta}\right)^n$	$\left(\alpha + (1-\alpha^2)\frac{k}{\beta}\right)^n$	$\lesssim n^{-1/q}$ (not optimal)	$\frac{\widetilde{k}_0 + \beta}{\widetilde{k}_0 + \beta + \beta(1-\alpha^2)n}$
$\alpha = 1$	$\mathrm{e}^{-\left(1 - \frac{2k}{\beta}\right)t}$	$\mathrm{e}^{-\left(1 - \frac{2k}{\beta}\right)t}$	$\lesssim t^{-1/q}$ (not optimal)	$\frac{\widetilde{k}_0 + \beta}{\widetilde{k}_0 + \beta + 2\beta t}$

Overview: Beyond Gaussians

- Optimization mode: algebraic convergence.
- Sampling mode: exponential convergence.

Steady State

- Sampling: steady state whose mean is close to the minimizer of f for large β in any dimension.
- The steady state is unique and arbitrarily close to the Laplace approximation of the target distribution (for β sufficiently large) in one dimension.
- The density μ_∞ is a steady state of both the discrete-in-time scheme with any $\alpha \in [0, 1)$ and the nonlinear Fokker–Planck equation corresponding to $\alpha = 1$ in one dimension.

Take-Aways

- Laplace approximation: For $\beta \to \infty$ the measure $\mathcal{R}_\beta \rho$ concentrates on δ_{θ_*}.
  $\implies$ for $\beta \gg 1$, consider a Gaussian approximation around $\mathcal{M}_\beta(\rho)$ with covariance $\mathcal{C}_\beta(\rho)$.
- The rescaling of the covariance by $(1-\alpha^2)(1+\beta)$ enables recovery of a Gaussian approximation of the desired target measure $\exp\big(-f(\bullet)\big)$.
- Fixing the scale at $(1-\alpha^2)$ allows the covariance to remain small when optimization of $f(\bullet)$ is the desired goal.
- Laplace method allows us to provide convergence guarantees beyond the Gaussian setting.

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[12]?
- Can we correct the sampling error in the non-Gaussian setting[13]?

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

[13] 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!

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

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

Affine invariance[15],[16],[17]

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.

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

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

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

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.