Report to Consensus Based Sampling

Zhenyu Huang

January 20, 2022

1 Introduction

The paper [1] consider the inverse problem of finding θ from y where

$$y = G(\theta) + \eta. \tag{1}$$

Here $y \in \mathbf{R}^K$ is the observation, $\theta \in \mathbf{R}^d$ is the unknown parameter, $G : \mathbf{R}^d \to \mathbf{R}^K$ is the forward model and η is the observational noise. Assume that the parameter and the noise are independent and normally distributed: $\theta \sim \mathrm{N}(0, \Sigma)$ and $\eta \sim \mathrm{N}(0, \Gamma)$. Applying the Bayes's formula, one can obtain the posterior density equals

$$\rho(\theta) = \frac{\exp(-f(\theta))}{\int_{\mathbf{R}^d} \exp(-f(\theta))\mathrm{d}\theta}, \tag{2}$$

where

$$f(\theta) := \frac{1}{2}|y - G(\theta)|_\Gamma^2 + \frac{1}{2}|\theta|_\Sigma^2.$$

There are two approaches for extracting information from (2). One approach is simply to seek the point of maximum posterior probability, the MAP point, defined by

$$\theta^* = \mathrm{argmin}_\theta \, f(\theta).$$

Another approach is to seek a Gaussian approximation of the measure, i.e., sample the posterior distribution $\rho(\theta)$.

2 The Paper's Work

The paper proposes a novel method for sampling and optimization tasks based on a stochastic interacting particle system which is called Consensus Based Sampling(CBS). It is based on ideas from Consensus Based Optimization(CBO) [3]. The development of the Ensemble Kalman Inversion(EKI) into the Ensemble Kalman Sampling(EKS) suggests a parallel development of CBO into a sampling methodology.

2.1 Starting Point: CBO

CBO is an optimization method based on the interacting particle system

$$\mathrm{d}\theta_t^{(j)} = -\left(\theta_t^{(j)} - \mathcal{M}_\beta\left(\mu_t^J\right)\right)\mathrm{d}t + \sqrt{2}\sigma\left|\theta_t^{(j)} - \mathcal{M}_\beta\left(\mu_t^J\right)\right|\mathrm{d}W_t^{(j)}, \quad j = 1, ..., J. \tag{3}$$

Here

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

$$\mu_t^J = \frac{1}{J}\sum_{j=1}^J \delta_{\theta_t^{(j)}}.$$

The idea behind the CBO method is to think about realizations of θ as explorers, in the landscape of the function $f(\theta)$, which can continuously exchange the evaluation of the function f at their position θ, through $\mathcal{M}_\beta(\rho_t)$. Then, the explorers compute a weighted average of their position in parameter space and direct their relaxation movement towards this average; this explains the first term on the right hand side of (3). The role of the second term is to impose the property of noise strength decreasing proportionally to the distance of the explorer to the weighted average.
[3] has shown that the mean field limit of (3) is

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

and the convergence result of the mean field limit solution is

$$\mathcal{M}_0\left(\mu_t\right) \underset{t\to\infty}{\longrightarrow} \widehat{\theta}(\beta), \quad \widehat{\theta}(\beta) \underset{\beta\to\infty}{\longrightarrow} \underset{\theta\in\mathbf{R}^d}{\arg\min} f(\theta).$$

2.2 Presentation of CBS

There are some notation: $\mathcal{M}_\beta$ is the weighted mean, $\mathcal{C}_\beta(\mu)$ is the weighted covariance, $\mathcal{L}_\beta(\mu)$ is the reweighting, defined by

$$\begin{array}{lll} \mathcal{M}_\beta(\mu) = \mathcal{M}\left(\mathcal{L}_\beta\mu\right), & \mathcal{C}_\beta(\mu) = \mathcal{C}\left(\mathcal{L}_\beta\mu\right), & \mathcal{L}_\beta : \mu \mapsto \frac{\mu \mathrm{e}^{-\beta f}}{\int \mu \mathrm{e}^{-\beta f}}, \\ \mathcal{M}(\mu) = \int \theta\mu(\mathrm{d}\theta), & \mathcal{C}(\mu) = \int (\theta - \mathcal{M}(\mu)) \otimes (\theta - \mathcal{M}(\mu))\mu(\mathrm{d}\theta). & \end{array}$$

The paper introduced the following discrete-time equation: given parameters $\lambda > 0$, $\beta > 0$ and $\alpha \in [0, 1)$,

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta\left(\rho_n\right) + \alpha\left(\theta_n - \mathcal{M}_\beta\left(\rho_n\right)\right) + \sqrt{\left(1-\alpha^2\right)\lambda^{-1}\mathcal{C}_\beta\left(\rho_n\right)}\boldsymbol{\xi}_n \\ \rho_n = \mathrm{Law}\left(\theta_n\right) \end{cases} \tag{4}$$

where $\boldsymbol{\xi}_n$ for $n \in \{0, 1, ...\}$ are i.i.d. r.v from $\mathrm{N}\left(\mathbf{0}, I_d\right)$.

Letting $\alpha = \exp(-\Delta t)$, the continuous-time limit associated with these dynamics is the following McKean SDE:

$$\begin{cases} \mathrm{d}\theta_t = -\left(\theta_t - \mathcal{M}_\beta\left(\rho_t\right)\right)\mathrm{d}t + \sqrt{2\lambda^{-1}\mathcal{C}_\beta\left(\rho_t\right)}\mathrm{d}\mathbf{W}_t \\ \rho_t = \mathrm{Law}\left(\theta_t\right) \end{cases} \tag{5}$$

In other words, CBS is proposed as alternative to the CBO method by using the weighted covariance $\mathcal{C}_\beta\left(\rho_t\right)$ as the modulation of the noise.
In order to determine the parameters, they denote by $g(\cdot; \mathbf{m}, C)$ the density of the Gaussian random variable $\mathrm{N}(\mathbf{m}, C)$:

$$g(\theta; \mathbf{m}, C) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\frac{1}{2}|\theta - \mathbf{m}|_C^2\right).$$

They also use the short-hand notation

$$\mathbf{m}_\beta(\mathbf{m}, C) := \mathcal{M}_\beta(g(\cdot; \mathbf{m}, C)), \quad C_\beta(\mathbf{m}, C) := \mathcal{C}_\beta(g(\cdot; \mathbf{m}, C)).$$

Let G be a linear map so that the posterior distribution given by (2) is Gaussian, and denote this Gaussian by $\mathrm{N}(\mathbf{a}, A)$. Computing the relationship between the mean and covariance of the Gaussian ρ and the mean and covariance of the Gaussian $\mathcal{L}_\beta\rho$ gives

$$\begin{aligned} \mathbf{m}_\beta(\mathbf{m}, C) &= \left(C^{-1} + \beta A^{-1}\right)^{-1}\left(\beta A^{-1}\mathbf{a} + C^{-1}\mathbf{m}\right), \\ C_\beta(\mathbf{m}, C) &= \left(C^{-1} + \beta A^{-1}\right)^{-1}. \end{aligned}$$

Therefore, the mean and covariance of a non-degenerate Gaussian steady state $g(\cdot; \mathbf{m}_\infty, C_\infty)$ for (4) satisfes

$$\begin{aligned} \mathbf{m}_\infty &= \left(C_\infty^{-1} + \beta A^{-1}\right)^{-1}\left(\beta A^{-1}\mathbf{a} + C_\infty^{-1}\mathbf{m}_\infty\right), \\ C_\infty &= \lambda^{-1}\left(C_\infty^{-1} + \beta A^{-1}\right)^{-1}. \end{aligned}$$

This has solution

$$\mathbf{m}_\infty = \mathbf{a}, \quad C_\infty = \frac{1-\lambda}{\lambda\beta}A. \tag{6}$$

Naturally, they focus on two choices of λ:
(i) $\lambda = 1$, when the method is used to minimize $f(\cdot)$, which will be referred to as CBS-O(α, β);
(ii) $\lambda = (1+\beta)^{-1}$ when the method is used for sampling the target distribution $\exp(-f(\cdot))$, which will be referred to as CBS(α, β).

2.3 Particle Approximations

The particle approximations of the mean field dynamics (4)and (5) is

$$\theta_{n+1}^{(j)} = \mathcal{M}_\beta\left(\rho_n^J\right) + \alpha\left(\theta_n^{(j)} - \mathcal{M}_\beta\left(\rho_n^J\right)\right) + \sqrt{\left(1-\alpha^2\right)\lambda^{-1}\mathcal{C}_\beta\left(\rho_n^J\right)}\xi_n^{(j)}, \quad j = 1, \ldots, J. \tag{7}$$

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

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

is the associated empirical measure. Note that

$$\mathcal{M}_\beta\left(\rho_n^J\right) = \frac{\sum_{j=1}^J \mathrm{e}^{-\beta f\left(\theta_n^{(j)}\right)}\theta_n^{(j)}}{\sum_{j=1}^J \mathrm{e}^{-\beta f\left(\theta_n^{(j)}\right)}},$$
$$\mathcal{C}_\beta\left(\rho_n^J\right) = \frac{\sum_{j=1}^J \left(\left(\theta_n^{(j)} - \mathcal{M}_\beta\left(\rho_n^J\right)\right)\otimes\left(\theta_n^{(j)} - \mathcal{M}_\beta\left(\rho_n^J\right)\right)\right)\mathrm{e}^{-\beta f\left(\theta_n^{(j)}\right)}}{\sum_{j=1}^J \mathrm{e}^{-\beta f\left(\theta_n^{(j)}\right)}}.$$

This leads to the implementable algorithms used in the paper.

2.4 Key Properties of the Mean Field Limits

- The time evolution of the law of the solution (4) is governed by the following discrete-time dynamics on probability densities:

$$\rho_{n+1}(\theta) = \int_{\mathbf{R}^d} g\left(\theta; \mathcal{M}_\beta\left(\rho_n\right) + \alpha\left(u - \mathcal{M}_\beta\left(\rho_n\right)\right), \left(1-\alpha^2\right)\lambda^{-1}\mathcal{C}_\beta\left(\rho_n\right)\right)\rho_n(u)\mathrm{d}u. \tag{8}$$

  The time evolution of the law of the solution to (5) is governed by the following nonlinear and nonlocal Fokker–Planck equation:

$$\frac{\partial\rho}{\partial t} = \nabla\cdot\left(\left(\theta - \mathcal{M}_\beta(\rho)\right)\rho + \lambda^{-1}C_\beta(\rho)\nabla\rho\right). \tag{9}$$

- The CBS dynamics is affine invariant. Denote by $\left(\mathrm{CBS}_n\left(\mu_0;\rho\right)\right)$ 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

$$\mathrm{CBS}_n\left(T_\sharp\left(\mu_0\right); T_\sharp(\rho)\right) = T_\sharp\left(\mathrm{CBS}_n\left(\mu_0;\rho\right)\right). \tag{10}$$

- The steady states of (4) and (5) coincide, if they exist, and they are necessarily Gaussian. Precisely,

$$\rho_\infty(\bullet) = g\left(\bullet; \mathcal{M}_\beta\left(\rho_\infty\right), \lambda^{-1}\mathcal{C}_\beta\left(\rho_\infty\right)\right).$$

- Gaussianity is preserved along the flow, both in discrete and continuous time.

2.5 Convergence for Gaussian Targets

The paper establish the convergence of the moments of the solutions to (4) and (5), respectively, in the case of Gaussian initial conditions.

- Consider $\alpha = 0$ and initial conditions $(\mathbf{m}_0, C_0) \in \mathbf{R}^d \times \mathcal{S}^d_{++}$. Then the following statements hold:

$$|\mathbf{m}_n - \mathbf{a}|_A \leqslant \max(1, k_0)\, \lambda^n\, |\mathbf{m}_0 - \mathbf{a}|_A\,,$$
$$\|C_n - A\|_A \leqslant \max(1, k_0)\, \lambda^n\, \|C_0 - A\|_A\,.$$

- Consider $\alpha \in (0, 1)$, then the following statements hold:

$$|\mathbf{m}_n - \mathbf{a}|_A \leqslant \max(1, k_0)^{\frac{1}{1+\alpha}}\, ((1-\alpha)\lambda + \alpha)^n\, |\mathbf{m}_0 - \mathbf{a}|_A\,,$$
$$\|C_n - A\|_A \leqslant \max(1, k_0)\, \left(\left(1-\alpha^2\right)\lambda + \alpha^2\right)^n\, \|C_0 - A\|_A\,.$$

- Consider the limiting case $\alpha \to 1$, then the following statements hold:

$$|\mathbf{m}(t) - \mathbf{a}|_A \leqslant \max\left(1, k_0^{\lambda/2}\right) \mathrm{e}^{-(1-\lambda)t}\, |\mathbf{m}_0 - \mathbf{a}|_A\,,$$
$$\|C(t) - A\|_A \leqslant \max\left(1, k_0^{\lambda}\right) \mathrm{e}^{-2(1-\lambda)t}\, \|C_0 - A\|_A\,.$$

The proof of above results is based on the following recurrence relation:

$$(\mathbf{m}_{n+1} - \mathbf{a}) = \left[\alpha I_d + (1-\alpha) A\, (A + \beta C_n)^{-1}\right] (\mathbf{m}_n - \mathbf{a})\,,$$
$$C_{n+1} = \left[\alpha^2 I_d + \left(1-\alpha^2\right) \lambda^{-1} A\, (A + \beta C_n)^{-1}\right] C_n,$$

And the equations for the moments:

$$\dot{\mathbf{m}} = -\beta C (A + \beta C)^{-1} (\mathbf{m} - \mathbf{a}),$$
$$\dot{C} = -2\beta C (A + \beta C)^{-1} \left(C - \left(\frac{1-\lambda}{\beta\lambda}\right) A\right).$$

2.6 Convergence Beyond The Gaussian Setting

Because of the last two properties in 2.4, the CBS is only exact for Gaussian problems like the ensemble Kalman sampler. However, the paper shows that under appropriate assumptions (convexity of the potential and one-dimenisonal), there exist a unique steady state which is Gaussian, close to the Laplace approximation of the posterior distribution.
Introduce

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

Then the distribution $\mathrm{e}^{-\widehat{f}} \propto \mathcal{N}(\theta_*, C_*)$ is the Laplace approximation of the target e^{-f}. Assume f satisfes some convexity and incremental condition, the paper obtain the following results for $d = 1$:

- For steady state,

$$\left|\begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix} - \begin{pmatrix} m_* \\ C_* \end{pmatrix}\right| \leqslant \frac{k}{\beta}.$$

- For $\alpha \in [0,1)$,

$$\left|\begin{pmatrix} m_n \\ C_n \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix}\right| \leqslant \left(\alpha + (1-\alpha^2)\frac{k}{\beta}\right)^n \left|\begin{pmatrix} m_0 \\ C_0 \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix}\right|.$$

- For $\alpha \to 1$,

$$\left|\begin{pmatrix} m(t) \\ C(t) \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix}\right| \leqslant \exp\left(-\left(1-\frac{2k}{\beta}\right)t\right) \left|\begin{pmatrix} m_0 \\ C_0 \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix}\right|.$$

The proof of the above results is based on Laplace method. Laplace method is aimed for determining the asymptotic behaviour as $\beta \to \infty$ of the integral

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

The paper shows

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

where $g_\beta = \mathcal{N}\left(\theta_*, \beta^{-1}\left(\text{ Hess } f\left(\theta_*\right)\right)^{-1}\right)$. In other words $\mathcal{L}_\beta \mu \approx g_\beta$ for large β. More precisely, the mean term satisfes

$$m_\beta(m,C) = \frac{\int_{\mathbf{R}} \theta \mathrm{e}^{-\beta f(\theta)} g(\theta; m, C) \mathrm{d}\theta}{\int_{\mathbf{R}} \mathrm{e}^{-\beta f(\theta)} g(\theta; m, C) \mathrm{d}\theta} = \theta_* + \mathcal{O}_R\left(\beta^{-1}\right).$$

And the covariance term satisfes

$$C_\beta(m,C) = \frac{\int_{\mathbf{R}} \left(\theta - \theta_*\right)^2 \mathrm{e}^{-\beta f(\theta)} g(\theta; m, C) \mathrm{d}\theta}{\int_{\mathbf{R}} \mathrm{e}^{-\beta f(\theta)} g(\theta; m, C) \mathrm{d}\theta} - \left(m_\beta(m,C) - \theta_*\right)^2 = \left(\text{ Hess } f\left(\theta_*\right)\right)^{-1} + \mathcal{O}_R\left(\beta^{-1}\right).$$

These lead to the existence of a fixed point of the map

$$\Phi_\beta : \begin{pmatrix} m \\ C \end{pmatrix} \mapsto \begin{pmatrix} m_\beta(m,C) \\ \lambda^{-1} C_\beta(m,C) \end{pmatrix}, \quad \lambda = (1+\beta)^{-1},$$

which implies the existence of a steady state solution both for the iterative scheme (8) with any $\alpha \in [0,1)$ and for the nonlinear Fokker-Planck equation (9). Morever, the paper shows that the map Φ_β is a contraction for sufficiently large β, then they finally obtain the convergence results.

3 Conclusion

The purposed 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.
- is exact only for Gaussian targets like EKS.

4 Perspective

The paper [1] gives the convergence results from the mean field limit viewpoint. But in pratice, we use the particle form (7) to implement the algorithm. So, we may be more interested in proving the convergence and error estimates at the particle level. In the paper [2], they provide a simple and elementary convergence and error analysis for a general time-discrete CBO algorithm.
Consider time-discrete analogue of (3). For this, set

$$h := \Delta t, \quad X_n := X(nh), \quad n = 0, 1, \cdots .$$

Then

$$\begin{cases} X_{n+1}^i = X_n^i - \gamma \left(X_n^i - \bar{X}_n^* \right) - \sum_{l=1}^d \left(x_n^{i,l} - \bar{x}_n^{*,l} \right) \eta_n^l e_l, \quad n \geq 0, i = 1, \cdots, N, \\ \bar{X}_n^* = \left(x_n^{*,1}, \cdots, x_n^{*,d} \right) := \dfrac{\sum_{j=1}^N X_n^j e^{-\beta L\left(X_n^j\right)}}{\sum_{j=1}^N e^{-\beta L\left(X_n^j\right)}}. \end{cases} \tag{11}$$

where the random variables $\left\{ \eta_n^l \right\}_{n,l}$ are i.i.d. with

$$\mathbb{E}\left[\eta_n^l \right] = 0, \quad \mathbb{E}\left[\left| \eta_n^l \right|^2 \right] = \zeta^2, \quad n = 1, \cdots, \quad l = 1, \cdots, d.$$

The paper [2] shows that system parameters satisfy some restricted condition:

$$(1 - \gamma)^2 + \zeta^2 < 1, \tag{12}$$

then the N-state ensemble $\{X_n^i\}$ exhibit a global consensus, i.e:

$$\lim_{n\to\infty} \mathbb{E}\left|X_n^i - X_n^j\right|^2 = 0 \quad \text{and} \quad \mathbb{P}\left\{\lim_{n\to\infty}\left|X_n^i - X_n^j\right| = 0\right\} = 1, \quad \forall i,j = 1,\cdots,N.$$

Under the same assumption (12), one can obtain that there exists a common constant state $X_\infty = \left(x_\infty^1,\cdots,x_\infty^d\right)$ such that

$$\lim_{n\to\infty} X_n^i = X_\infty \quad \text{a.s., } 1 \le i \le N, \quad \text{such that} \quad L\left(X_\infty\right) \sim \min_X L(X).$$

Under the assumption (12) and for a well-prepared initial random variable X_{in} such that $X_n^i \sim X_{in}$, they derive a key estimate

$$\operatorname*{essinf}_{\omega\in\Omega} L\left(X_\infty\right) \le L\left(X_*\right) + \frac{d}{2}\frac{\log\beta}{\beta} + \mathcal{O}(\beta^{-1}).$$

For the first-order Euler type discrete model and predictor-corrector type discrete model, these assumptions can be satisfied.
The analytical method and assumptions purposed in [2] may be introduced in CBS, but the question is how to deal with the noise term which is not the same as CBO. I will investigate this more in future work.

References

[1] J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes. Consensus based sampling, 2021.

[2] Seung-Yeal Ha, Shi Jin, and Doheon Kim. Convergence and error estimates for time-discrete consensus-based optimization algorithms. *Numerische Mathematik*, 147(2):255–282, 2021.

[3] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 27(01):183–204, 2017.