Bayesian inference: Interacting particle approaches

Sebastian Reich

University of Potsdam & SFB 1294 Data Assimilation

June 9, 2022
Computational Bayesian inference
- Coupling of measures
- Invariance and ergodicity

Interacting particle systems for sampling
- Overdamped Langevin dynamics
- Gradient log density estimator

Kalman–Wasserstein gradient flow structure
- Gradient flow structures
- Numerical implementation
Part I

Computational Bayesian inference

Publications:
Daniel Huang, Jiaoyang Huang, SR & Andrew Stuart, *Efficient derivative-free Bayesian inference for large scale inverse problems*, arXiv:2204.04386
Prior (forecast):
\[ \Theta_f \sim \pi_f \]

negative log-likelihood:

**nonl. regression:** 
\[ l(y|\theta) = \frac{1}{2} (g(\theta) - y)^T R^{-1} (g(\theta) - y) \]

**logistic regression:** 
\[ l(y|\theta) = -y \log \sigma(\theta^T \phi_x) - (1 - y) \log(1 - \sigma(\theta^T \phi_x)) \]

\( g \) forward map, \( R \) error covariance matrix, \( y \) the data \((y \in \mathbb{R}/y \in \{0, 1\})\), \( \phi_x \) feature map, \( x \in \mathbb{R}^J \), \( \sigma(t) = 1/(1 + \exp(-t)) \).

**Bayesian posterior (analysis):**
\[ \pi_a(\theta|y) \propto e^{-l(y|\theta)} \pi_f(\theta). \]
Monte Carlo: Compute realisations $\theta^{(i)}_a$, $i = 1, \ldots, M$, from a random variable (RV)

$$\Theta_a \sim \pi_a$$

to approximate posterior expectation values

$$\mathbb{E}_a[f] \approx \frac{1}{M} \sum_{i=1}^{M} f(\theta^{(i)}_a).$$

Catch: The random variable $\Theta_a$ is not fully specified by Bayes’ theorem:

- approaches based on coupling of measures,
- and those based on invariance and ergodicity.
Find a pair of random variables

\[(\Theta_f, \Theta_a) \sim \pi_{fa}(\theta_f, \theta_a) = \pi_a(\theta_a | \theta_f) \pi_f(\theta_f)\]

such that

\[\Theta_f \sim \pi_f, \quad \Theta_a \sim \pi_a.\] (1)

**Catch:** Joint distribution \(\pi_{fa}(\theta_f, \theta_a)\) is not uniquely determined by its marginals (1):

- **Examples:** sequential Monte-Carlo, ensemble Kalman filter
- **optimal transportation** (minimise expected distance between \(\Theta_f\) and \(\Theta_a\); transport equation)
- **Schrödinger bridges** (minimise the Kullback–Leibler divergence to some reference measure; stochastic optimal control)
Do the opposite: \textbf{successively decouple}.

Define a sequence of random variables (\textit{stochastic process}) $\Theta_\tau$, $\tau \geq 0$, with $\Theta_0 = \Theta_f$ and

$$\Theta_a := \lim_{\tau \to \infty} \Theta_\tau \sim \pi_a.$$

\textbf{Catch:} Such stochastic processes typically satisfy:

- $\Theta_0 \sim \pi_a$ implies $\Theta_\tau \sim \pi_a$ for all $\tau > 0$ (\textit{invariance}) and
- $\Theta_a$ is \textit{independent} of $\Theta_0 = \Theta_f$ (\textit{ergodicity}), that is,

$$\pi_{fa}(\theta_f, \theta_a) = \pi_f(\theta_f) \pi_a(\theta_a | y).$$
Examples:

- **Langevin dynamics**

\[
\frac{d\Theta}{d\tau} = -\nabla_\theta V(\Theta)\,d\tau + \sqrt{2}\,dW_\tau. \tag{2}
\]

with

\[
V(\theta) = -\log \pi_a(\theta|y) .
\]

The SDE (2) is ergodic with unique invariant measure $\pi_a$ under appropriate conditions.

- In discrete time, we got **Markov chain Monte Carlo (MCMC)** methods.
Discrete-time Langevin:

stochastic process \( \{\Theta_n\}_{n \geq 0} \), \( \lim_{n \to \infty} \Theta_n \sim \pi_a \).

Idea.\(^1\) Let \( \Theta_n \sim \bar{\pi}_n \); for any \( \gamma > 0 \):

- diffusion: \( \bar{\pi}_{n+1/2} \propto \bar{\pi}_n^{1/(1+\gamma)} \),
- Bayes/drift: \( \bar{\pi}_{n+1} \propto \pi_a^{\gamma/(1+\gamma)} \bar{\pi}_n + \bar{\pi}_n + 1/2 \)

Theorem. If \( \bar{\pi}_n = \pi_a \), then \( \bar{\pi}_{n+1} = \pi_a \). Convergence is exponential for all \( \gamma > 0 \).

\(^1\)Huang et al, arXiv:2204.04386
Diffusion step:

\[ \Theta_n \sim \tilde{\pi}_n = \mathcal{N}(\mu_n, \Sigma_n) \]

implies

\[ \tilde{\pi}_{n+1/2} = \mathcal{N}(\mu_n, (1 + \gamma)\Sigma_n). \]

Update step:

deterministic: \[ \Theta_{n+1/2} = \mu_n + (1 + \gamma)^{1/2}(\Theta_n - \mu_n) \]

stochastic: \[ \Theta_{n+1/2} = \Theta_n + \gamma^{1/2}\Sigma_n^{1/2}\Xi_n, \quad \Xi_n \sim \mathcal{N}(0, I). \]
Bayes/drift step:

i) Extended observations

\[ \tilde{y}_{\text{obs}} = \begin{pmatrix} y_{\text{obs}} \\ \mu_0 \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} G \\ I \end{pmatrix}, \quad \tilde{R} = \begin{pmatrix} R & 0 \\ 0 & \Sigma_0 \end{pmatrix}. \]

ii) negative log ”likelihood” function

\[ \tilde{l}(\theta | \tilde{y}_{\text{obs}}) := -\log \pi_a(\theta) = \frac{1}{2} (\tilde{G} \theta - \tilde{y}_{\text{obs}})^T \tilde{R}^{-1} (\tilde{G} \theta - \tilde{y}_{\text{obs}}). \]

iii) Kalman filter step with \( \tau = \gamma/(1 + \gamma) \), likelihood \( \tilde{l}(\theta | \tilde{y}_{\text{obs}}) \) and prior \( \mathcal{N}(\mu_n^{1/2}, \Sigma_n^{1/2}) \).
Continuous-time limit:\(^2\) \((\gamma \to 0)\)

$$\dot{\Theta}_\tau = -\frac{1}{2} \Sigma_{\Theta} \left\{ G^T R^{-1} (G \Theta_\tau + G \mu \Theta_\tau - 2 y_{obs}) + \Sigma_0^{-1} (\Theta_\tau + \mu \Theta_\tau - 2 \mu_0) \right\} + \Sigma_{\Theta_\tau}^{1/2} \dot{W}_\tau.$$ 

Alternatively:

$$\dot{\Theta}_\tau = -\frac{1}{2} \Sigma_{\Theta} \left\{ G^T R^{-1} (G \Theta_\tau + G \mu \Theta_\tau - 2 y_{obs}) + \Sigma_0^{-1} (\Theta_\tau + \mu \Theta_\tau - 2 \mu_0) \right\} + 2(\Theta_\tau - \mu \Theta_\tau).$$

\(^2\)Pidstrigach & SR, FoCM, 2022, Huang et al, arXiv:2204.04386
Part II

Interacting particle systems for sampling

Publications:
Nonlinear diffusion I

Nonlinear SDE:

\[ d\Theta_\tau = f(\Theta_\tau) d\tau + \sqrt{2}\sigma dW_\tau, \quad \Theta_0 \sim \pi_0, \]

\(W_\tau\) standard Brownian motion and e.g. \(f(\theta) = \nabla_\theta \log \pi_\theta(\theta|y)\).

Fokker–Planck equation: \(\Theta_\tau \sim \pi_\tau\)

\[ \partial_\tau \pi_\tau = -\nabla \cdot (\pi_\tau f) + \sigma \Delta \pi_\tau, \]

\[ = -\nabla \cdot (\pi_\tau \{f - \sigma \nabla \log \pi_\tau\}) \]
Nonlinear SDE:

\[ d\Theta_\tau = f(\Theta_\tau) d\tau + \sqrt{2}\sigma dW_\tau, \quad \Theta_0 \sim \pi_0, \]

\( W_\tau \) standard Brownian motion and e.g. \( f(\theta) = \nabla_\theta \log \pi_a(\theta|y) \).

Fokker–Planck equation: \( \Theta_\tau \sim \pi_\tau \)

\[
\partial_\tau \pi_\tau = -\nabla \cdot (\pi_\tau f) + \sigma \Delta \pi_\tau, \\
= -\nabla \cdot (\pi_\tau \{f - \sigma \nabla \log \pi_\tau\})
\]

Mean-field ODE

\[ \dot{\Theta}_\tau = f(\Theta_\tau) - \sigma \nabla \log \pi_\tau. \]
Nonlinear diffusion

**Gaussian case:**

\[ \Theta_\tau \sim N(\mu_\tau, \Sigma_\tau) \implies -\nabla \log \pi_\tau(\theta) = \Sigma_\tau^{-1}(\theta - \mu_\tau). \]

---

\(^3\)Carrillo et al, Calc. Var. Part. Diff. Eqs, 2019
Nonlinear diffusion

**Gaussian case:**

\[ \Theta_\tau \sim N(\mu_\tau, \Sigma_\tau) \implies -\nabla \log \pi_\tau(\theta) = \Sigma_\tau^{-1}(\theta - \mu_\tau). \]

**Interacting particle dynamics:** \( \Theta_0^{(i)} \sim \pi_0, i = 1, \ldots, M, \)

\[ \dot{\Theta}_\tau^{(i)} = f(\Theta_\tau^{(i)}) - \sigma \nabla \log \tilde{\pi}_\tau(\Theta_\tau^{(i)}) \]

with approximative density \( \tilde{\pi}_\tau: \)

- Gaussian

\[ \tilde{\pi}_\tau(\theta) = n(\theta; \mu_M, \Sigma_M) \]

- Gaussian mixture \(^3\)

\[ \tilde{\pi}_\tau(\theta) = \frac{1}{M} \sum_{i=1}^{M} n(\theta; \Theta_\tau^{(i)}, \gamma I). \]

\(^3\)Carrillo et al, Calc. Var. Part. Diff. Eqs, 2019
Variational formulation:\(^4\)

\[
\partial_\alpha \log \pi := r^{(\alpha)} + \arg \min_\phi L_\alpha[\phi, \pi]
\]

\[
\partial_\alpha = \partial_{\theta^{(\alpha)}}, \quad r^{(\alpha)} \text{ the } \alpha \text{th component of an appropriate reference function}
\]

\[
r : \mathbb{R}^{N_\theta} \to \mathbb{R}^{N_\theta},
\]

\[
L_\alpha[\phi, \pi] := \int \pi(\theta) \left( \phi^2(\theta) + 2r^{(\alpha)}(\theta) \phi(\theta) + 2\partial_\alpha \phi(\theta) \right) d\theta
\]

\[
= \int \pi(\theta) \left( \phi(\theta) + r^{(\alpha)}(\theta) - \partial_\alpha \log \pi(\theta) \right)^2 +
\]

\[
\text{terms independent of } \phi
\]

---

Gradient log density estimator II

Estimator:

\[ \mathcal{L}_\alpha[\phi, \pi_\tau] \approx \mathcal{L}_\alpha[\phi, \pi_\tau^M] \]

\[ := \frac{1}{M} \sum_{i=1}^{M} \left( \phi^2(\Theta_T^{(i)}) + 2r^{(\alpha)}(\Theta_T^{(i)}) \phi(\Theta_T^{(i)}) + 2\partial_\alpha \phi(\Theta_T^{(i)}) \right) \]

and

\[ \partial_\alpha \log \pi_\tau(\theta) \approx r^{(\alpha)}(\theta) + \arg \min_{\phi \in \mathcal{F}} \mathcal{L}_\alpha[\phi, \pi_\tau^M](\theta). \]
Gradient log density estimator II

Estimator:

\[ \mathcal{L}_\alpha[\phi, \pi_\tau] \approx \mathcal{L}_\alpha[\phi, \pi^M_\tau] \]

\[ := \frac{1}{M} \sum_{i=1}^{M} \left( \phi^2(\Theta^{(i)}_\tau) + 2r^{(\alpha)}(\Theta^{(i)}_\tau) \phi(\Theta^{(i)}_\tau) + 2\partial_\alpha \phi(\Theta^{(i)}_\tau) \right) \]

and

\[ \partial_\alpha \log \pi_\tau(\theta) \approx r^{(\alpha)}(\theta) + \arg \min_{\phi \in \mathcal{F}} \mathcal{L}_\alpha[\phi, \pi^M_\tau](\theta). \]

Interacting particle ODE: \( i = 1, \ldots, M, \)

\[ \dot{\Theta}^{(i)}_\tau = f(\Theta^{(i)}_\tau) - \sigma \left( r^{(\alpha)}(\Theta^{(i)}_\tau) + \phi^{(\alpha)}_\tau(\Theta^{(i)}_\tau) \right). \]

with

\[ \phi^{(\alpha)}_\tau := \arg \min_{\phi \in \mathcal{F}} \mathcal{L}_\alpha[\phi, \pi^M_\tau] \]
Remarks

- Approximation space $\mathcal{F}$: (i) $L$-dimensional (random feature) space

$$\phi_{\tau}(\theta) = \sum_{l=1}^{L} \alpha_{\tau}^{(l)} \phi_{l}(\theta)$$

(ii) RKHS with kernel $k(\theta, \theta')$

$$\phi_{\tau}(\theta) = \sum_{i=1}^{M} \alpha_{\tau}^{(i)} k(\theta, \Theta_{\tau}^{(i)})$$

- (ii) with $\gamma = 1$, $f(\theta) = \nabla_{\theta} \log \pi_{a}(\theta|y)$, $r = f$, leads equations related to **Stein variational gradient descent**.\(^5\)

\(^5\)Q. Liu & D. Wang, NEURIPS, 2016
Example: Chaotic attractor Lorenz 63
Kalman–Wasserstein gradient flow structure

Publications:


Overdamped Langevin dynamics

\[ d\Theta_\tau = -\nabla_\theta V(\Theta_\tau) \, d\tau + \sqrt{2} \, dW_\tau. \]

is **not** invariant under **affine transformations**

\[ \theta' = A\theta + b; \]

that is

\[ d\Theta'_\tau = -AA^T \nabla_{\theta'} V(\Theta'_\tau) \, d\tau + \sqrt{2}A \, dW_\tau. \]
Affine invariant sampling methods

---

7 SR, BIT, 2011
8 Garbuno-Inigo et al, SIADS, 2020a
9 Garbuno-Inigo et al, SIADS, 2020b
Affine invariant sampling methods

Inspired by ensemble Kalman–Bucy filter:

\[ d\Theta_\tau = -\Sigma_\tau \nabla_\theta G(\Theta_\tau) R^{-1} \left( G(\Theta_\tau) d\tau + R^{1/2} dW_\tau - y d\tau \right) \]

---

7SR, BIT, 2011
8Garbuno-Inigo et al, SIADS, 2020a
9Garbuno-Inigo et al, SIADS, 2020b
Affine invariant sampling methods\textsuperscript{6}

Inspired by ensemble Kalman–Bucy filter:\textsuperscript{7}

\[ d\Theta_\tau = -\Sigma_\tau \nabla_\theta G(\Theta_\tau) R^{-1} \left( G(\Theta_\tau) d\tau + R^{1/2} dW_\tau - y d\tau \right) \]

**Ensemble Kalman sampler (EKS)\textsuperscript{8} / affine invariant Langevin dynamics (ALDI)\textsuperscript{9}**

\[ d\Theta_\tau = -\Sigma_\tau \nabla_\theta V(\Theta_\tau) d\tau + \sqrt{2} \Sigma_\tau^{1/2} dW_\tau. \]

\textsuperscript{7}SR, BIT, 2011
\textsuperscript{8}Garbuno-Inigo et al, SIADS, 2020a
\textsuperscript{9}Garbuno-Inigo et al, SIADS, 2020b
Kalman–Wasserstein gradient flow structure

**Nonlinear (affine invariant) Fokker–Planck equation**

\[
\partial_\tau \pi_\tau = -\nabla_\theta \cdot (\pi_\tau \Sigma_\tau \{ \nabla_\theta \log \pi_\tau - \nabla_\theta \log \pi_a \}) = -\text{grad}_{\pi}^{\text{AI}} \text{KL}(\pi_\tau || \pi_a)
\]

\(\Sigma_t\) the covariance matrix of \(\Theta_\tau \sim \pi_\tau\).

Metric \(g_\pi(\rho_1, \rho_2)\) on space of densities induced by **Mahalanobis distance** on \(\mathbb{R}^D\):

\[
\|a\|_{\Sigma^{-1}}^2 := a^T \Sigma^{-1} a
\]

---

Kalman–Wasserstein gradient flow structure

**Nonlinear (affine invariant) Fokker–Planck equation**\(^{10}\)

\[
\partial_{\tau} \pi_{\tau} = -\nabla_{\theta} \cdot (\pi_{\tau} \Sigma_{\tau} \{ \nabla_{\theta} \log \pi_{\tau} - \nabla_{\theta} \log \pi_{a} \}) \\
= -\text{grad}^{\text{AI}}_{\pi} \, \text{KL} (\pi_{\tau} \| \pi_{a})
\]

\(\Sigma_{t}\) the covariance matrix of \(\Theta_{\tau} \sim \pi_{\tau}\).

Metric \(g_{\pi}(\rho_{1}, \rho_{2})\) on space of densities induced by **Mahalanobis distance** on \(\mathbb{R}^{D}\):

\[
\|a\|^{2}_{\Sigma_{\tau}^{-1}} := a^{T} \Sigma_{\tau} a
\]

It holds that

\[
\frac{d}{d\tau} \text{KL}(\pi_{\tau} \| \pi_{a}) = -\int_{\mathbb{R}^{N}} \pi_{\tau} \left\| \nabla_{\theta} \frac{\delta \text{KL}(\pi_{\tau} \| \pi_{a})}{\delta \pi_{\tau}} \right\|^{2}_{\Sigma_{\tau}^{-1}} \leq 0 .
\]

Implementation of affine invariant Langevin dynamics (ALDI):

\[ d\Theta^{(i)}_{\tau} = -\sum_{\tau}^{M} \nabla_{\theta} V(\Theta^{(i)}_{\tau}) \, d\tau + \frac{D + 1}{M} (\Theta^{(i)}_{\tau} - \bar{\theta}_{\tau}^{M}) + \sqrt{2(\sum_{\tau}^{M})^{1/2}} \, dW^{(i)}_{\tau}, \]

\( i = 1, \ldots, M, \Theta_{\tau} \in \mathbb{R}^{D} \).
Implementation of affine invariant Langevin dynamics (ALDI):

\[
d\Theta^{(i)}_\tau = -\sum_{\tau}^{M} \nabla_\theta V(\Theta^{(i)}_\tau) \, d\tau + \frac{D + 1}{M} (\Theta^{(i)}_\tau - \bar{\theta}^M_\tau) + \sqrt{2} (\sum_{\tau}^{M})^{1/2} \, dW^{(i)}_\tau,
\]

\(i = 1, \ldots, M, \Theta_\tau \in \mathbb{R}^D.\)

Remarks.

- **correction term in orange** is needed for invariance of \(\pi_a\) (multiplicative noise),
- **invariance** and **ergodicity** holds provided \(M \geq D + 1\),
- ALDI is **affine invariant** for any \(M \geq 2\),
- **derivative-free formulation**.
Gradient-free implementations I

Can we avoid the computation of gradients?

**Idea:** Introduce localised covariance matrices

\[
\Sigma_\tau (\theta) := \frac{1}{C} \int (\theta' - \bar{\theta}_\tau)(\theta' - \bar{\theta}_\tau)^T e^{-\frac{1}{2\gamma} \|\theta' - \theta\|_2^2} \pi_\tau (\theta') \, d\theta', \tag{3}
\]

\(\bar{\theta}_\tau\) localised mean, \(\gamma > 0, \ C > 0\) a scaling constant.

**Localised** ALDI dynamics:

\[
d\Theta_\tau = -\Sigma_\tau (\Theta) \nabla_\theta V(\Theta_\tau) \, d\tau + \nabla_\theta \cdot \Sigma_\tau (\Theta_\tau) \, d\tau + \sqrt{2} \Sigma_\tau (\Theta_\tau)^{1/2} \, dW_\tau
\]
Let $\overline{V}_\tau(\theta)$ denote the expectation of $V(\Theta')$ w.r.t. density defined in (3), that is,

$$\tilde{\pi}_\tau(\theta'|\theta) = \frac{1}{C} e^{-\frac{1}{2\gamma} \|\theta'-\theta\|^2_{\Sigma_\tau} \pi_\tau(\theta')}.$$ 

**Catch:** $\tilde{\pi}_\tau(\theta'|\theta)$ is close to Gaussian with mean $\theta$ for $\gamma \ll 1$.

Allows for **derivative-free implementation** of ALDI/EnKBF with controllable errors as $\gamma \ll 1$ and $M \to \infty$:

$$\Sigma_\tau \nabla_\theta V \approx \Sigma_\tau \nabla_\theta \overline{V}_\tau \approx (\Theta' - \bar{\theta}_\tau)(V(\Theta') - \bar{V}_\tau)_\tau.$$
Nonlinear forward operator

\[ g(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2 \]

Figure: Bimodal posterior (a) and ALDI results (b)-(d) for two-dimensional multimodal posterior distribution.
Conclusions

- Increasing interest in interacting particle systems for sampling, inference, and optimisation
- Fruitful exchange between methods based on ergodicity and invariance and those based on coupling of measures
- Solid comparison is largely missing
- Affine invariance is highly desirable for applications in the natural sciences
- Gradient flow structures in the space of probability measures also appear as desirable; but in which metric and under which cost functional?