Variational inference via Wasserstein gradient flows

Sinho Chewi

NeurIPS 2022

Massachusetts Institute of Technology

Collaborators

Francis Bach (INRIA)

Silvère Bonnabel (UNC/ENSMP)

Marc Lambert (INRIA)

Philippe Rigollet (MIT)

Massachusetts Institute of Technology

Motivation from Bayesian Inference

Motivation: Large-scale Bayesian applications require computation of *summary statistics* of the posterior $\pi \propto \exp(-V)$.

Two main computational paradigms:

- Markov chain Monte Carlo (MCMC)
- variational inference (VI)

Markov Chain Monte Carlo (MCMC)

The most basic MCMC algorithm discretizes the Langevin diffusion

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

which has $X_\infty \sim \pi$.

Non-asymptotic guarantees: if V is *strongly convex* + *smooth*, we approximately sample from π after $O(d)$ queries to ∇V.

Variational Inference (VI)

Approximate π via:

$$\boxed{\hat{\pi} \in \arg\min_{p \in \mathcal{P}} \mathrm{KL}(p \,\|\, \pi)}$$

Common choices for $\mathcal{P}$:

- $\mathcal{P} = \{\text{product measures}\}$ (mean-field)
- $\mathcal{P} = \{\text{Gaussians}\}$ or $\{\text{mixtures of Gaussians}\}$ (**this talk**)

What is the computational complexity?

Särkkä's Heuristic

Let $(\pi_t)_{t\geq 0}$ be the law of the Langevin diffusion

$$\pi_t = \mathsf{law}(X_t)\,, \qquad \mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

Can we build a Gaussian approximation?

Särkkä's Heuristic

Let $(\pi_t)_{t\geq 0}$ be the law of the Langevin diffusion

$$\pi_t = \mathsf{law}(X_t)\,, \qquad \mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

Can we build a Gaussian approximation?

The mean $m_t = \mathbb{E}\,X_t$ and covariance $\Sigma_t = \operatorname{cov} X_t$ evolve via

$$\begin{aligned}
\dot m_t &= -\,\mathbb{E}\,\nabla V(X_t)\,,\\
\dot\Sigma_t &= 2I - \mathbb{E}[\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)]\,.
\end{aligned}$$

Massachusetts Institute of Technology

Särkkä's Heuristic

Let $(\pi_t)_{t\geq 0}$ be the law of the Langevin diffusion

$$\pi_t = \mathsf{law}(X_t)\,, \qquad \mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

Can we build a Gaussian approximation?

The mean $m_t = \mathbb{E}\,X_t$ and covariance $\Sigma_t = \operatorname{cov} X_t$ evolve via

$$\begin{aligned}
\dot m_t &= -\,\mathbb{E}\,\nabla V(X_t)\,, \\
\dot\Sigma_t &= 2I - \mathbb{E}[\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)]\,.
\end{aligned}$$

We cannot compute the expectations.

Särkkä's Heuristic

Heuristic from Kalman filtering [Särkkä '07]: replace X_t via $Y_t \sim p_t = \mathcal{N}(m_t, \Sigma_t)$.

$$\dot{m}_t = -\mathbb{E}\,\nabla V(Y_t)\,,$$
$$\dot{\Sigma}_t = 2I - \mathbb{E}[\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes \nabla V(Y_t)]\,.$$

This yields a Gaussian approximation $(p_t)_{t \geq 0}$.

Särkkä's Heuristic

Heuristic from Kalman filtering [Särkkä '07]: replace X_t via $Y_t \sim p_t = \mathcal{N}(m_t, \Sigma_t)$.

$$\begin{aligned}\dot m_t &= -\,\mathbb{E}\,\nabla V(Y_t)\,,\\ \dot\Sigma_t &= 2I - \mathbb{E}[\nabla V(Y_t)\otimes(Y_t - m_t) + (Y_t - m_t)\otimes\nabla V(Y_t)]\,.\end{aligned}$$

This yields a Gaussian approximation $(p_t)_{t\geq 0}$.

What is its interpretation? Convergence as $t \to \infty$? At what rate?

Massachusetts Institute of Technology

Wasserstein Gradient Flows

Theorem (Jordan, Kinderlehrer, Otto '98): The law $(\pi_t)_{t\geq 0}$ of the Langevin diffusion is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

Wasserstein Gradient Flows

Theorem (Jordan, Kinderlehrer, Otto '98): The law $(\pi_t)_{t \geq 0}$ of the Langevin diffusion is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

Theorem (Lambert, C., Bach, Bonnabel, Rigollet '22): The law $(p_t)_{t \geq 0}$ of Särkkä's process is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ which is constrained to lie in the space of Gaussians.

Massachusetts Institute of Technology

Wasserstein Gradient Flows

Theorem (Jordan, Kinderlehrer, Otto '98): The law $(\pi_t)_{t\geq 0}$ of the Langevin diffusion is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

Theorem (Lambert, C., Bach, Bonnabel, Rigollet '22): The law $(p_t)_{t\geq 0}$ of Särkkä's process is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ which is constrained to lie in the space of Gaussians.

We call this the Bures–Wasserstein space, $(\mathsf{BW}(\mathbb{R}^d), W_2)$.

Särkkä's Process as a Gradient Flow

Theorem (Lambert, C., Bach, Bonnabel, Rigollet '22): The law $(p_t)_{t\geq 0}$ of Särkkä's process is a gradient flow of $\mathrm{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ which is constrained to lie in the space of Gaussians.

Consequences:

- as $t \to \infty$, $p_t \to \hat{\pi} := \arg\min_{\mathrm{BW}(\mathbb{R}^d)} \mathrm{KL}(\cdot \,\|\, \pi)$
  $\implies$ solution to Gaussian VI

Massachusetts Institute of Technology

Särkkä's Process as a Gradient Flow

Theorem (Lambert, C., Bach, Bonnabel, Rigollet '22): The law $(p_t)_{t \geq 0}$ of Särkkä's process is a gradient flow of $\mathsf{KL}(\cdot \,\|\, \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ which is constrained to lie in the space of Gaussians.

Consequences:

- as $t \to \infty$, $p_t \to \hat{\pi} := \arg\min_{\mathsf{BW}(\mathbb{R}^d)} \mathsf{KL}(\cdot \,\|\, \pi)$
  $\implies$ solution to Gaussian VI
- use theory of gradient flows to obtain convergence rates

Consequences: Continuous-Time Convergence

Theorem (Lambert, C., Bach, Bonnabel, Rigollet '22):
If V is α-strongly convex and $\mathsf{KL}_\star := \mathsf{KL}(\hat\pi \,\|\, \pi)$:

1. ($\alpha > 0$)

$$W_2^2(p_t, \hat\pi) \le \exp(-2\alpha t)\, W_2^2(p_0, \hat\pi)\,,$$
$$\mathsf{KL}(p_t \,\|\, \pi) - \mathsf{KL}_\star \le \exp(-2\alpha t)\, \{\mathsf{KL}(p_0 \,\|\, \pi) - \mathsf{KL}_\star\}\,.$$

3. ($\alpha = 0$)

$$\mathsf{KL}(p_t \,\|\, \pi) - \mathsf{KL}_\star \le \frac{1}{2t}\, W_2^2(p_0, \hat\pi)\,.$$

Consequences: Discretization

Theorem (Lambert, C., Bach, Bonnabel, Rigollet '22): Assume $0 \prec \alpha I \preceq \nabla^2 V \preceq I$. For the iterates $(p_k)_{k\in\mathbb{N}}$ of Bures–Wasserstein SGD with step size $0 < h \le \frac{\alpha}{6}$,

$$\mathbb{E}\, W_2^2(p_k, \hat{\pi}) \le \exp(-\alpha k h)\, W_2^2(p_0, \hat{\pi}) + \frac{21 d h}{\alpha^2}\,.$$

$\Longrightarrow$ $\widetilde{O}(d)$ query complexity, akin to MCMC

Mixtures of Gaussians

There is a correspondence between measures over $\mathrm{BW}(\mathbb{R}^d)$ and mixtures of Gaussians:

$$\underbrace{\mu}_{\text{mixing measure}} \qquad \leftrightarrow \qquad \mathsf{p}_\mu := \int p \, \mathrm{d}\mu(p) \,.$$

Mixtures of Gaussians

There is a correspondence between measures over $\mathrm{BW}(\mathbb{R}^d)$ and mixtures of Gaussians:

$$\underbrace{\mu}_{\text{mixing measure}} \qquad \leftrightarrow \qquad \mathsf{p}_\mu := \int p \, \mathrm{d}\mu(p)\,.$$

Consequently, $\{\text{mixtures of Gaussians}\} \cong \mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$.

Mixtures of Gaussians

There is a correspondence between measures over $\mathrm{BW}(\mathbb{R}^d)$ and mixtures of Gaussians:

$$\underbrace{\mu}_{\text{mixing measure}} \qquad \leftrightarrow \qquad \mathsf{p}_\mu := \int p \, \mathrm{d}\mu(p)\,.$$

Consequently, $\{\text{mixtures of Gaussians}\} \cong \mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$.

What is the **gradient flow** of $\mu \mapsto \mathsf{KL}(\mathsf{p}_\mu \,\|\, \pi)$ over this space?

Gradient Flow for Mixtures of Gaussians

Theorem (Lambert, C., Bach, Bonnabel, Rigollet): The gradient flow of $\mu \mapsto \mathrm{KL}(\mathsf{p}_\mu \,\|\, \pi)$ over $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$ can be implemented as an interacting particle system: for $i \in [N]$,

$$\dot{m}_t^{(i)} = -\,\mathbb{E}\,\nabla \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\,,$$

$$\dot{\Sigma}_t^{(i)} = -\,\mathbb{E}\,\nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\,\Sigma_t^{(i)} - \Sigma_t^{(i)}\,\mathbb{E}\,\nabla^2 \ln \frac{\mathsf{p}_{\mu_t}}{\pi}(Y_t^{(i)})\,,$$

where $Y_t^{(i)} \sim \mathcal{N}(m_t^{(i)}, \Sigma_t^{(i)})$ and $\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{(m_t^{(i)}, \Sigma_t^{(i)})}$.

Massachusetts Institute of Technology

Mixture of Gaussians VI

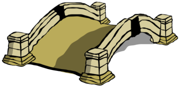
See our paper for an algorithm with changing weights based on Wasserstein–Fisher–Rao geometry.

Massachusetts Institute of Technology

Conclusion

Wasserstein gradient flows

variational inference (VI)



Kalman filtering

- We obtain an algorithm for Gaussian VI with quantitative computational guarantees.
- We propose algorithms for mixture of Gaussians VI based on Wasserstein gradient flows.