Variational inference via Wasserstein gradient flows

Presentation of the paper *Lambert et al., NeurIPS 2022*

Gianluca Covini

Candidate for the PhD position in *Optimal Transport and Machine Learning* at WIAS

January 7, 2025

Outline

Problem

Variational inference

Given a target distribution $\pi \propto \exp(-V)$, determine an approximation:

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

where $\mathcal{P}$ is an ambiguity set.

The Kullback–Leibler (KL) divergence is defined as:

$$\mathrm{KL}(p\|\pi) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{dp}{d\pi}(x)\right) d\pi(x), & \text{if } p \ll \pi, \\ \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{\pi(x)}\right) dx, & \text{if } p, \pi \ll \text{Leb}. \end{cases}$$

Bayesian inference

Bayesian inference framework

We want to make inference (e.g., compute expectation, covariance...) on a posterior distribution π on a space Θ:

$$\pi(\theta) := p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}.$$

Defining the *potential* $V(\theta) = -\log(p(x \mid \theta)) - \log(p(\theta))$, then $\pi \propto \exp(-V)$.

Bayesian inference

Bayesian inference framework

We want to make inference (e.g., compute expectation, covariance...) on a posterior distribution π on a space Θ:

$$\pi(\theta) := p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}.$$

Defining the *potential* $V(\theta) = -\log(p(x \mid \theta)) - \log(p(\theta))$, then $\pi \propto \exp(-V)$.

Two approaches:

- Monte Carlo Markov chains (**MCMC**);
- Variational inference (**VI**).

MCMC

Idea: build a Markov chain $(X_t)_{t\geq 0}$ with π as stationary distribution. Then take $(X_t)_{t\geq \bar{t}}$ as samples from π.

MCMC

Idea: build a Markov chain $(X_t)_{t\geq 0}$ with π as stationary distribution. Then take $(X_t)_{t\geq \bar{t}}$ as samples from π.

Langevin MC

$(X_t)_{t\geq 0}$ are solution of the *Langevin SDE*

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \tag{1}$$

where $(B_t)_{t\geq 0}$ is the standard Brownian motion.

MCMC

Idea: build a Markov chain $(X_t)_{t\geq 0}$ with π as stationary distribution. Then take $(X_t)_{t\geq \bar{t}}$ as samples from π.

Langevin MC

$(X_t)_{t\geq 0}$ are solution of the *Langevin SDE*

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \tag{1}$$

where $(B_t)_{t\geq 0}$ is the standard Brownian motion.

The marginal laws μ_t of the solution of $(X_t)_{t\geq 0}$, are the solution of the *Fokker-Planck PDE*

$$\partial_t \mu_t = \text{div}(\mu_t \nabla \log \frac{\mu_t}{\pi})$$

and it converges to the stationary distribution π

MCMC

Idea: build a Markov chain $(X_t)_{t\geq 0}$ with π as stationary distribution. Then take $(X_t)_{t\geq \bar{t}}$ as samples from π.

Langevin MC

$(X_t)_{t\geq 0}$ are solution of the *Langevin SDE*

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \tag{2}$$

where $(B_t)_{t\geq 0}$ is the standard brownian motion.

Advantages:

- samples from real distribution;
- non-asymptotic guarantees when π is strongly log-concave.

MCMC

Idea: build a Markov chain $(X_t)_{t\geq 0}$ with π as stationary distribution. Then take $(X_t)_{t\geq \bar{t}}$ as samples from π.

Langevin MC

$(X_t)_{t\geq 0}$ are solution of the *Langevin SDE*

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \tag{2}$$

where $(B_t)_{t\geq 0}$ is the standard brownian motion.

Advantages:

- samples from real distribution;
- non-asymptotic guarantees when π is strongly log-concave.

Disadvantages:

- generally slow, in particular in high-dimensional settings.

VI

Variational inference

The *VI* setting consists in finding a tractable approximation $\hat{\pi}$ of π solving

$$\hat{\pi} \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{KL}(p||\pi)$$

and then computing the quantity of interests on $\hat{\pi}$.

VI

Variational inference

The *VI* setting consists in finding a tractable approximation $\hat{\pi}$ of π solving

$$\hat{\pi} \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{KL}(p||\pi)$$

and then computing the quantity of interests on $\hat{\pi}$.

Advantages:

- fast computation of statistics of $\hat{\pi}$.

VI

Variational inference

The *VI* setting consists in finding a tractable approximation $\hat{\pi}$ of π solving

$$\hat{\pi} \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{KL}(p||\pi)$$

and then computing the quantity of interests on $\hat{\pi}$.

Advantages:

- fast computation of statistics of $\hat{\pi}$.

Disadvantages:

- approximation of the target distribution;

VI

Variational inference

The *VI* setting consists in finding a tractable approximation $\hat{\pi}$ of π solving

$$\hat{\pi} \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{KL}(p||\pi)$$

and then computing the quantity of interests on $\hat{\pi}$.

Advantages:

- fast computation of statistics of $\hat{\pi}$.

Disadvantages:

- approximation of the target distribution;
- theoretical guarantees?

VI

Variational inference

The *VI* setting consists in finding a tractable approximation $\hat{\pi}$ of π solving

$$\hat{\pi} \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{KL}(p||\pi)$$

and then computing the quantities of interest on $\hat{\pi}$.

In the paper, the problem is addressed in the cases of

- $\mathcal{P} = \mathrm{BW}(\mathbb{R}^d) := \{\text{non-degenerate } d\text{-dimensional Gaussians}\}$;
- $\mathcal{P} = \{\text{mixtures of } d\text{-dimensional Gaussians}\}$.

Particle system

Theorem

Given $(X_t)_{t\geq 0}$ solutions of the Langevin diffusion, with $X_t \sim \pi_t$, $m_t = \mathbb{E}[X_t]$ and $\Sigma_t = \mathrm{cov}(X_t)$, then

$$\dot{m}_t = -\mathbb{E}_{\pi_t}[\nabla V(X_t)]$$
$$\dot{\Sigma}_t = 2I - \mathbb{E}_{\pi_t}[\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)]$$

Particle system

Theorem

Given $(X_t)_{t\geq 0}$ solutions of the Langevin diffusion, with $X_t \sim \pi_t$, $m_t = \mathbb{E}[X_t]$ and $\Sigma_t = \mathrm{cov}(X_t)$, then

$$\dot{m}_t = -\mathbb{E}_{\pi_t}[\nabla V(X_t)]$$
$$\dot{\Sigma}_t = 2I - \mathbb{E}_{\pi_t}[\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)]$$

Särkkä's heuristic

Taking $Y_t \sim p_t = \mathcal{N}(m_t, \Sigma_t)$, the system of ODE

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

yields to an evolution $(p_t)_{t\geq 0}$ of gaussians.

Problem geometry

Optimization problem

$$\hat{\pi} \in \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{KL}(p||\pi)$$

where $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$

The choice of $\mathcal{P}$ determines the geometry of the searching problem.

- In $\mathbb{R}^d$: $\dot{x}_t = -\nabla F(x_t)$
  ("continuous-time" gradient descent)
- In $\mathcal{P} = \mathrm{BW}(\mathbb{R}^d)$: $\partial_t \mu_t = -\nabla_{\mathrm{BW}} \mathcal{F}(\mu_t)$
  (gradient flow)

∇_{BW}? Discretization?

Results

Claim: The gradient flow of $\mathrm{KL}(\cdot||\pi)$ in the space $\mathrm{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance solves Särkkä's ODEs.

Results

Claim: The gradient flow of $\mathrm{KL}(\cdot||\pi)$ in the space $\mathrm{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance solves Särkkä's ODEs.

Consequences:

- $(p_t)_{t\geq 0}$ converges (rapidly) to the Gaussian VI solution $\hat{\pi}$ when π is strongly log-concave;

Results

Claim: The gradient flow of $\mathrm{KL}(\cdot||\pi)$ in the space $\mathrm{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance solves Särkkä's ODEs.

Consequences:

- $(p_t)_{t\geq 0}$ converges (rapidly) to the Gaussian VI solution $\hat{\pi}$ when π is strongly log-concave;
- the discretized algorithm has non-asymptotic sharp guarantees when π is strongly log-concave and log-smooth.

Results

Claim: The gradient flow of $\mathrm{KL}(\cdot||\pi)$ in the space $\mathrm{BW}(\mathbb{R}^d)$ endowed with the Wasserstein distance solves Särkkä's ODEs.

Consequences:

- $(p_t)_{t\geq 0}$ converges (rapidly) to the Gaussian VI solution $\hat{\pi}$ when π is strongly log-concave;
- the discretized algorithm has non-asymptotic sharp guarantees when π is strongly log-concave and log-smooth.

An extension with $\mathcal{P}$ as space of mixtures of Gaussian is then derived.

Wasserstein space

Riemannian structure

We consider $\mathcal{P}_2(\mathbb{R}^d)$ with the following Riemannian structure (*Otto*):

$$T_\mu \mathcal{P}_2(\mathbb{R}^d) = \{\nabla\psi \mid \psi : \mathbb{R}^d \to \mathbb{R}\}, \qquad \mu \in \mathcal{P}_2(\mathbb{R}^d)$$

$$\langle v, w\rangle_\mu = \int_{\mathbb{R}^d} \langle v(\theta), w(\theta)\rangle_{\mathbb{R}^d} d\mu(\theta), \qquad v, w \in T_\mu \mathcal{P}_2(\mathbb{R}^d)$$

Wasserstein space

Riemannian structure

We consider $\mathcal{P}_2(\mathbb{R}^d)$ with the following Riemannian structure (*Otto*):

$$T_\mu \mathcal{P}_2(\mathbb{R}^d) = \{\nabla\psi \mid \psi : \mathbb{R}^d \to \mathbb{R}\}, \qquad \mu \in \mathcal{P}_2(\mathbb{R}^d)$$

$$\langle v, w\rangle_\mu = \int_{\mathbb{R}^d} \langle v(\theta), w(\theta)\rangle_{\mathbb{R}^d} d\mu(\theta), \qquad v, w \in T_\mu \mathcal{P}_2(\mathbb{R}^d)$$

Metric structure

$\mathcal{P}_2(\mathbb{R}^d)$ is a metric space with the *Wasserstein metric* (*Benamou - Brenier*):

$$W_2^2(\mu_0, \mu_1) = \inf\left\{\int_0^1 \| \nabla\xi \|_{\mu_t}^2 \, dt \mid \partial_t \mu_t = -\operatorname{div}(\mu_t \nabla\xi)\right\}$$

Bures-Wasserstein space

$\mathrm{BW}(\mathbb{R}^d)$ is identified with $\mathbb{R}^d \times \mathbf{S}^d_{++}$, where $\mathbf{S}^d_{++} = \{\Sigma \in \mathbb{R}^{d\times d} \mid \Sigma \succ 0, \Sigma^t = \Sigma\}$.

Bures-Wasserstein space

$\mathrm{BW}(\mathbb{R}^d)$ is identified with $\mathbb{R}^d \times \mathbf{S}^d_{++}$, where $\mathbf{S}^d_{++} = \{\Sigma \in \mathbb{R}^{d\times d} \mid \Sigma \succ 0, \Sigma^t = \Sigma\}$.

Properties

- It inherits the Riemannian (and metric) structure of $\mathcal{P}_2(\mathbb{R}^d)$;
- known closed-form for the Wasserstein distance and the optimal transport map.

Bures-Wasserstein space

$\mathrm{BW}(\mathbb{R}^d)$ is identified with $\mathbb{R}^d \times \mathbf{S}^d_{++}$, where $\mathbf{S}^d_{++} = \{\Sigma \in \mathbb{R}^{d\times d} \mid \Sigma \succ 0, \Sigma^t = \Sigma\}$.

Properties

- It inherits the Riemannian (and metric) structure of $\mathcal{P}_2(\mathbb{R}^d)$;
- known closed-form for the Wasserstein distance and the optimal transport map.

Riemannian structure

In particular, the tangent space is identified with

$$T_p \,\mathrm{BW}(\mathbb{R}^d) = \{x \mapsto a + S(x - m_p) \mid a \in \mathbb{R}^d, S \in \mathbf{S}^d\} \cong \mathbb{R}^d \times \mathbf{S}^d$$

where $\mathbf{S}^d = \{S \in \mathbb{R}^{d\times d} \mid S^t = S\}$.

Gradient flows in $\mathrm{BW}(\mathbb{R}^d)$

Goal: derive gradient flows $(p_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot \mid \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$.

Gradient flows in $\mathrm{BW}(\mathbb{R}^d)$

Goal: derive gradient flows $(p_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot \mid \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$.

- *Bures-JKO scheme* (proximal point algorithm):

Bures-JKO scheme

Given $h > 0$,

$$p_{k+1,h} := \operatorname*{argmin}_{p \in \mathrm{BW}(\mathbb{R}^d)} \{\mathrm{KL}(p \parallel \pi) + \frac{1}{2h} W_2^2(p, p_{k,h})\}$$

Gradient flows in $\mathrm{BW}(\mathbb{R}^d)$

Goal: derive gradient flows $(p_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot \mid \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$.

- *Bures-JKO scheme* (proximal point algorithm):

Bures-JKO scheme

Given $h > 0$,

$$p_{k+1,h} := \operatorname*{argmin}_{p \in \mathrm{BW}(\mathbb{R}^d)} \{\mathrm{KL}(p \parallel \pi) + \frac{1}{2h} W_2^2(p, p_{k,h})\}$$

We then define $p_t := \lim_{h\to 0} p_{\lfloor t/h \rfloor, h}$

Gradient flows in $\mathrm{BW}(\mathbb{R}^d)$

Goal: derive gradient flows $(p_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot \mid \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$.

- *Projection of the Wasserstein gradient on $T\,\mathrm{BW}(\mathbb{R}^d)$;*
- *Direct computation of Bures-Wasserstein gradient.*

Gradient flows in $\mathrm{BW}(\mathbb{R}^d)$

Goal: derive gradient flows $(p_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot \mid \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$.

- *Projection of the Wasserstein gradient on $T\,\mathrm{BW}(\mathbb{R}^d)$;*
- *Direct computation of Bures-Wasserstein gradient.*

Bures-Wasserstein gradient

$$\nabla_{\mathrm{BW}} f(m,\Sigma) = (\nabla_m f(m,\Sigma), 2\nabla_\Sigma f(m,\Sigma))$$

$$\nabla_{\mathrm{BW}} \mathrm{KL}(p \parallel \pi) = (\mathbb{E}_p[\nabla V], \mathbb{E}_p[\nabla^2 V] - \Sigma_p^{-1})$$

Gradient flows in $\mathrm{BW}(\mathbb{R}^d)$

Goal: derive gradient flows $(p_t)_{t\geq 0}$ of $\mathrm{KL}(\cdot \mid \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$.

- *Projection of the Wasserstein gradient on $T\,\mathrm{BW}(\mathbb{R}^d)$;*
- *Direct computation of Bures-Wasserstein gradient.*

Bures-Wasserstein gradient

$$\nabla_{\mathrm{BW}} f(m,\Sigma) = (\nabla_m f(m,\Sigma), 2\nabla_\Sigma f(m,\Sigma))$$

$$\nabla_{\mathrm{BW}} \mathrm{KL}(p \parallel \pi) = (\mathbb{E}_p[\nabla V], \mathbb{E}_p[\nabla^2 V] - \Sigma_p^{-1})$$

Gradient flow

In the previous ways, we obtain $\nabla_{\mathrm{BW}} KL(\mu \parallel \pi)$ in μ. The gradient flow $(p_t)_{t\geq 0}$ in $\mathrm{BW}(\mathbb{R}^d)$ is a solution of

$$\partial_t p_t = -\nabla_{\mathrm{BW}} KL(p_t \parallel \pi)$$

Särkkä equivalence

Theorem

$(p_t = p_{(m_t,\Sigma_t)})_{t\geq 0}$, gradient flow of $\mathrm{KL}(\mu \parallel \pi)$ in $\mathrm{BW}(\mathbb{R}^d)$, satisfies Särkkä's system of ODEs

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

Continuous-time convergence

α-convexity

We say $\mathcal{F}$, functional on $\mathrm{BW}(\mathbb{R}^d)$, is α-convex for $\alpha \in \mathbb{R}$ if, on the constant-speed geodesic $(p_t)_{t\in[0,1]}$,

$$\mathcal{F}(p_t) \leq (1-t)\mathcal{F}(p_0) + t\mathcal{F}(p_1) - \alpha \frac{t(1-t)}{2} W_2^2(p_0, p_1)$$

Continuous-time convergence

α-convexity

We say $\mathcal{F}$, functional on $\mathrm{BW}(\mathbb{R}^d)$, is α-convex for $\alpha \in \mathbb{R}$ if, on the constant-speed geodesic $(p_t)_{t\in[0,1]}$,

$$\mathcal{F}(p_t) \leq (1-t)\mathcal{F}(p_0) + t\mathcal{F}(p_1) - \alpha \frac{t(1-t)}{2} W_2^2(p_0, p_1)$$

Lemma

For any $\alpha \in \mathbb{R}$, $\nabla^2 V \succeq \alpha I$ (π strongly log-concave), then $\mathrm{KL}(\cdot||\pi)$ is α-convex on $\mathrm{BW}(\mathbb{R}^d)$.

Continuous-time convergence

Corollary

Let $\nabla^2 V \succeq \alpha I$ for a certain $\alpha \in \mathbb{R}$. Then, for any $p_0 \in \mathrm{BW}(\mathbb{R}^d)$, there exists a unique solution for the gradient flow in $\mathrm{BW}(\mathbb{R}^d)$ of $\mathrm{KL}(\cdot||\pi)$ started at p_0. Moreover,

1. If $\alpha > 0$, then for all $t \geq 0$,

$$W_2^2(p_t, \hat{\pi}) \leq \exp(-2\alpha t) W_2^2(p_0, \hat{\pi})$$

2. If $\alpha > 0$, then for all $t \geq 0$,

$$\mathrm{KL}(p_t \parallel \pi) - \mathrm{KL}(\hat{\pi} \parallel \pi) \leq \exp(-2\alpha t) \{\mathrm{KL}(p_0 \parallel \pi) - \mathrm{KL}(\hat{\pi} \parallel \pi)\}$$

3. If $\alpha = 0$, then for all $t > 0$,

$$\mathrm{KL}(p_t \parallel \pi) - \mathrm{KL}(\hat{\pi} \parallel \pi) \leq \frac{1}{2t} W_2^2(p_0, \hat{\pi})$$

Discretization

To discretize in t the gradient flow $(p_t)_{t\geq 0}$, two approaches are possible:

Discretization

To discretize in t the gradient flow $(p_t)_{t\geq 0}$, two approaches are possible:

- Numerical integration of Särkkä's system of ODE

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

with $Y_t \sim p_t = \mathcal{N}(m_t, \Sigma_t)$

Drawback: theoretical guarantees?

Discretization

- Bures-Wasserstein SGD algorithm.

Algorithm Bures–Wasserstein SGD

Require: $\alpha > 0$, step size $h > 0$, m_0 and Σ_0

1: **for** $k = 1, \dots, N$ **do**
2: Draw a sample $\hat{X}_k \sim p_k$
3: $m_{k+1} \leftarrow m_k - h\nabla V(\hat{X}_k)$
4: $M_k \leftarrow I - h(\nabla^2 V(\hat{X}_k) - \Sigma_k^{-1})$
5: $\Sigma_k^+ \leftarrow M_k \Sigma_k M_k$
6: $\Sigma_{k+1} \leftarrow \mathrm{clip}_{1/\alpha} \Sigma_k^+$
7: **end for**

It can be shown that

$$p_k^+ := p_{m_{k+1}, \Sigma_k^+} = \exp_{p_k}(-h\nabla_{\mathrm{BW}} \mathrm{KL}(p_k \parallel \pi)(\hat{X}_k)) \quad \text{(SG step of } h\text{)}.$$

Non-asymptotic results

Theorem

Assume that $0 \prec \alpha I \preceq \nabla^2 V \preceq I$ (π strongly log-concave and log-smooth). Also, assume that $h \leq \alpha^2/60$ and that we initialize Algorithm 1 at a matrix satisfying $\frac{\alpha}{9} I \preceq \Sigma_{\mu_0} \preceq \frac{1}{\alpha} I$. Then, for all $k \in \mathbb{N}$,

$$\mathbb{E}[W_2^2(p_k, \hat{\pi})] \leq \exp(-\alpha k h) W_2^2(p_0, \hat{\pi}) + \frac{36 d h}{\alpha^2}.$$

In particular, we obtain

$$\mathbb{E}[W_2^2(p_k, \hat{\pi})] \leq \varepsilon^2$$

provided we set $h \asymp \frac{\alpha^2 \varepsilon^2}{d}$ and the number of iterations to be $k \gtrsim \frac{d}{\alpha^3 \varepsilon^2} \log\left(\frac{W_2(p_0, \hat{\pi})}{\varepsilon}\right)$.

Summary (so far)

Results:

- Theoretical guarantees that close the gap between Gaussian VI and Langevin MC.

Summary (so far)

Results:

- Theoretical guarantees that close the gap between Gaussian VI and Langevin MC.

Drawbacks:

- Gaussians are not always a good approximation of the target distribution π.

Summary (so far)

Results:

- Theoretical guarantees that close the gap between Gaussian VI and Langevin MC.

Drawbacks:

- Gaussians are not always a good approximation of the target distribution π.

The drawback can be addressed by extending the previous model to mixtures of Gaussians.

Theorem

The set of the d-dimensional Gaussian mixtures is dense in $\mathcal{P}_2(\mathbb{R}^d)$ for the metric W_2.

Geometry of the problem

The space of mixtures of Gaussians can be identified with $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$ through this relation

$$\mu \in \mathcal{P}_2(\Theta) \quad \longleftrightarrow \quad p_\mu := \int_\Theta p_\theta \, d\mu(\theta)$$

where $\Theta = \mathbb{R}^d \times \mathbf{S}^{\mathbf{d}}_{++} \cong \mathrm{BW}(\mathbb{R}^d)$.

Geometry of the problem

The space of mixtures of Gaussians can be identified with $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$ through this relation

$$\mu \in \mathcal{P}_2(\Theta) \quad \longleftrightarrow \quad p_\mu := \int_\Theta p_\theta \, d\mu(\theta)$$

where $\Theta = \mathbb{R}^d \times \mathbf{S}^{\mathbf{d}}_{++} \cong \mathrm{BW}(\mathbb{R}^d)$.

Remark

The theory of optimal transport can be derived again in the space $\mathcal{P}_2(\mathcal{M})$ where $\mathcal{M}$ is a Riemannian manifold, in particular it is well-defined the gradient flow of $\mu \mapsto \mathrm{KL}(p_\mu \,\|\, \pi)$ in $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$.

Interacting particles algorithm

Let $\mu \in \mathcal{P}(\mathrm{BW}(\mathbb{R}^d)) \mapsto \mathcal{F}(\mu) = \mathrm{KL}(p_\mu \parallel \pi) \in \mathbb{R} \cup \{\infty\}$.

Theorem

The gradient flow $(\mu_t)_{t \geq 0}$ of the functional $\mathcal{F}$ over $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$ can be described as follows. Let $\theta_0 = (m_0, \Sigma_0) \sim \mu_0$, and let $\theta_t = (m_t, \Sigma_t)$ evolve according to the system of ODEs:

$$\dot{m}_t = -\mathbb{E}[\nabla \log \frac{p_{\mu_t}}{\pi}(Y_t)],$$
$$\dot{\Sigma}_t = -\mathbb{E}\left[\nabla^2 \log \frac{p_{\mu_t}}{\pi}(Y_t)\right]\Sigma_t - \Sigma_t \mathbb{E}\left[\nabla^2 \log \frac{p_{\mu_t}}{\pi}(Y_t)\right],$$

where $Y_t \sim \mathcal{N}(m_t, \Sigma_t)$. Then $\theta_t \sim \mu_t$.

Interacting particles algorithm

To implement the *Gaussian particles* scheme we take $N < \infty$ Gaussian particles, i.e., mixtures of N Gaussians

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_0^{(i)}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(m_0^{(i)}, \Sigma_0^{(i)})} \leftrightarrow p_{\mu_0} = \frac{1}{N}\sum_{i=1}^{N} p_{(m_0^{(i)}, \Sigma_0^{(i)})}.$$

It follows

$$\mu_t = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_t^{(i)}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})} \leftrightarrow p_{\mu_t} = \frac{1}{N}\sum_{i=1}^{N} p_{(m_t^{(i)}, \Sigma_t^{(i)})}.$$

Interacting particles algorithm

To implement the *Gaussian particles* scheme we take $N < \infty$ Gaussian particles, i.e., mixtures of N Gaussians

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_0^{(i)}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(m_0^{(i)}, \Sigma_0^{(i)})} \leftrightarrow p_{\mu_0} = \frac{1}{N}\sum_{i=1}^{N} p_{(m_0^{(i)}, \Sigma_0^{(i)})}.$$

It follows

$$\mu_t = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_t^{(i)}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})} \leftrightarrow p_{\mu_t} = \frac{1}{N}\sum_{i=1}^{N} p_{(m_t^{(i)}, \Sigma_t^{(i)})}.$$

Or we can use a proximal point method analogous to the Bures-JKO scheme:

$$(\theta_t^{(1)} + h, \ldots, \theta_t^{(N)} + h) \approx$$

$$\approx \operatorname*{argmin}_{\theta^{(1)},\ldots,\theta^{(N)} \in \Theta} \left[\mathrm{KL}\left(\frac{1}{N}\sum_{i=1}^{N} p_{\theta^{(i)}} \middle\| \pi \right) + \frac{1}{2Nh}\sum_{i=1}^{N} W_2^2(p_{\theta^{(i)}}, p_{\theta_t^{(i)}}) \right].$$

Pros and cons

Advantages:

- (Hope of) asymptotic convergence to an arbitrary good approximation of the target distribution π given N sufficiently large.

Drawbacks:

- Lack of theoretical guarantees as in Gaussian VI and Langevin MC because

$$\mu \mapsto \mathrm{KL}(p_\mu \parallel \pi)$$

is in general not convex, even if π is strongly log-concave.

Extensions with changing weights

In the paper a possible extension is suggested to take into account mixtures of Gaussians with changing weights:

$$\mu_t = \sum_{i=1}^{N} w_t^{(i)} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})} \leftrightarrow p_{\mu_t} = \sum_{i=1}^{N} w_t^{(i)} p_{(m_t^{(i)}, \Sigma_t^{(i)})}.$$

Extensions with changing weights

In the paper a possible extension is suggested to take into account mixtures of Gaussians with changing weights:

$$\mu_t = \sum_{i=1}^{N} w_t^{(i)} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})} \leftrightarrow p_{\mu_t} = \sum_{i=1}^{N} w_t^{(i)} p_{(m_t^{(i)}, \Sigma_t^{(i)})}.$$

Idea: use the *Wasserstein-Fisher-Rao (Hellinger-Kantorovich) geometry*, which admits change of mass. Each Gaussian particle is provided with a mass which evolves with the other parameters.

Extensions with changing weights

Theorem

Let $Y_t^{(i)} \sim \mathcal{N}(m_t^{(i)}, \Sigma_t^{(i)})$, and let $r_t^{(i)} = \sqrt{w_t^{(i)}}$. Then, the system of ODEs is given by:

$$\dot{m}_t^{(i)} = -\mathbb{E}\left[\nabla \log \frac{p_{\mu_t}}{\pi}\left(Y_t^{(i)}\right)\right],$$

$$\dot{\Sigma}_t^{(i)} = -\mathbb{E}\left[\nabla^2 \log \frac{p_{\mu_t}}{\pi}\left(Y_t^{(i)}\right)\right]\Sigma_t^{(i)} - \Sigma_t^{(i)}\mathbb{E}\left[\nabla^2 \log \frac{p_{\mu_t}}{\pi}\left(Y_t^{(i)}\right)\right],$$

$$\dot{r}_t^{(i)} = -\left(\mathbb{E}\left[\log \frac{p_{\mu_t}}{\pi}\left(Y_t^{(i)}\right)\right] - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\log \frac{p_{\mu_t}}{\pi}\left(Y_t^{(j)}\right)\right]\right) r_t^{(i)}.$$

Summary

Results of the paper:

- Gaussian VI framework which close the gap with Langevin MC in terms of theoretical guarantees;
- mixtures of Gaussians VI framework with good empirical results.

Limits:

- Lack of theoretical guarantees for the mixture of Gaussians model;
- no in-depth analysis of the mixture of Gaussians VI with changing weights and the Wasserstein-Fisher-Rao geometry.

Thank you for your attention!

Proof of Särkkä's theorem

It is known (Ambrosio et al., 2008):

$$\nabla_{W_2} \mathrm{KL}(\mu \parallel \pi) = \nabla \log \frac{\mu}{\pi}$$

Let $(\pi_t)_{t \geq 0}$ evolve along the Fokker-Planck PDE.
According to Otto calculus, if $x_0 \sim \pi_0$ and

$$\dot{x_t} = v_t(x_t) = -\nabla \log \frac{\pi_t}{\pi}(x_t)$$

then $x_t \sim \pi_t$.
From explicit computation (integrating by parts) of

$$\dot{m_t} = \partial_t \mathbb{E}_{\pi_t}[x_t]$$
$$\dot{\Sigma}_t = \partial_t \mathbb{E}_{\pi_t}[x_t \otimes x_t] - \partial_t(\mathbb{E}_{\pi_t}[x_t] \otimes \mathbb{E}_{\pi_t}[x_t])$$

follows Särkkä's system of ODEs.

Bures-JKO scheme

Gaussian VI

$$p_{t+h} = \min_{p \in \mathrm{BW}(\mathbb{R}^d)} \mathcal{L}(m, \Sigma) = \mathrm{KL}(p_{m,\Sigma} \parallel \pi) + \\ + \frac{1}{2h} \parallel m_t - m \parallel^2 + \frac{1}{2h} \mathcal{B}^2(\Sigma_t, \Sigma)$$

Särkkä's system of ODEs can be derived by computing the critical points from

$$\nabla_m \mathcal{L}(m, \Sigma) = 0$$
$$\nabla_\Sigma \mathcal{L}(m, \Sigma) = 0$$

and then computing the limit for $h \downarrow 0$.
The process is analogous to the proximal point algorithm for mixture of Gaussians.

Otto calculus

In $\mathcal{P}(\mathbb{R}^d)$, with Otto-Wasserstein riemannian structure, for every *nice and regular* curve $(\mu_t)_{t\geq 0}$, there exist a tangent field $(v_t = \nabla\phi_t)_{t\geq 0}$ such that

$$\partial_t \mu_t + \mathrm{div}(\mu_t v_t) = 0$$

$$W_2^2(\mu_0, \mu_1) = \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \{\int_{\mathbb{R}^d\times\mathbb{R}^d} \| x - y \|^2 \, d\gamma(x, y)\}$$

$$= \int_{\mathbb{R}^d} \| x - T(x) \|^2 \, d\mu_0$$

$$\log_\mu \nu = \nabla\phi_{\mu\to\nu} - id$$

$$\exp_\mu v = (id + v)_\mu$$

Complements on Bures-Wasserstein space

If $(p_t)_{t\geq 0}$ is a curve in BW($\mathbb{R}^d$) with tangent vector in time 0 of (a, S), then

$$\dot{m}_t = a$$
$$\dot{\Sigma}_t = S\Sigma_0 + \Sigma_0 S$$

For any curve $(m_t, \Sigma_t)_{t\geq 0}$ with tangent vector (a, S) at time $t = 0$ is defined as $\nabla_{BW} f(m_0, \Sigma_0) = (\bar{a}, \bar{S})$ such that

$$\langle \nabla_{BW} f(m_0, \Sigma_0), (a, S)\rangle_{p_{m_0,\Sigma_0}} = \partial_t|_{t=0} f(m_t, \Sigma_t).$$

$$\langle \bar{a}, a\rangle + \langle \bar{S}, \Sigma_0 S\rangle = \langle \nabla_m f(m_0, \Sigma_0), a\rangle + 2\langle \nabla_\Sigma f(m_0, \Sigma_0), \Sigma_0 S\rangle.$$

It follows:

$$\nabla_{BW} f(m, \Sigma) = \left(\nabla_m f(m, \Sigma), 2\nabla_\Sigma f(m, \Sigma)\right).$$

Proof of continuous-time convergence

$\mathcal{F}$ functional, $(p_t)_{t\geq 0}$ and $(q_t)_{t\geq 0}$ solutions of g.f. of $\mathcal{F}$.

$$\mathrm{F}(\mathrm{p}_t) \geq \mathcal{F}(q_t) + \langle \nabla(\mathcal{F}(q_t)), \log_{q_t}(p_t)\rangle_{q_t} + \tfrac{\alpha}{2} d^2(p_t, q_t)$$
$$\partial_t d^2(p_t, q_t) \leq -2\alpha d^2(p_t, q_t)$$
Grownwall's inequality: $\mathrm{d}^2(p_t, q_t) \leq \exp(-2\alpha t) d^2(p_0, q_0)$

Taking $q_t = p^*, \quad \forall t \geq 0$, if $\alpha > 0$

$$0 = \mathcal{F}(p_*) \overset{\text{Conv.}}{\geq} \mathcal{F}(p) + \langle \nabla \mathcal{F}(p), \log_p(p_*)\rangle_p + \frac{\alpha}{2} d^2(p, p_*)$$
$$\overset{\text{Young}}{\geq} \mathcal{F}(p) - \frac{1}{2\alpha} \|\nabla \mathcal{F}(p)\|_p^2 - \frac{\alpha}{2} \underbrace{\|\log_p(p_*)\|_p^2}_{=d^2(p,p_*)} + \frac{\alpha}{2} d^2(p, p_*)$$

Proof of continuous-time convergence

p_t gradient flow $\implies \partial_t \mathcal{F}(p_t) = - \parallel \nabla \mathcal{F}(p_t) \parallel^2_{p_t}$

$\partial_t \mathcal{F}(p_t) \leq -2\alpha \mathcal{F}(p_t) \overset{\text{Gronwall}}{\implies} \mathcal{F}(p_t) \leq \exp(-2\alpha t)\mathcal{F}(p_0)$

If $\alpha = 0$,

$$\text{Lyapunov functional: } \mathcal{L}_t := t\mathcal{F}(p_t) + \frac{1}{2} d^2(p_t, p^*)$$

$$\partial_t \mathcal{L}_t = \mathcal{F}(p_t) - t \parallel \nabla \mathcal{F}(p_t) \parallel^2_{p_t} + \langle \log_{p_t}(p^*), \nabla \mathcal{F}(p_t) \rangle_{p_t} \leq 0$$

$$\mathcal{L}_t \leq \mathcal{L}_0 \implies \mathcal{F}(p_t) \leq \frac{d^2(p_0, p^*)}{2t}$$

Complements on Bures-SGD

We have $g_p := \nabla_{BW}\mathcal{F}(p) = (\mathbb{E}_p[\nabla V], \mathbb{E}_p[\nabla^2 V] - \Sigma^{-1})$, and therefore the stochastic gradient $\hat{g}_p := (\nabla V(\hat{X}), \nabla^2 V(\hat{X}) - \Sigma^{-1})$.

Clip operator:

$$\text{clip}^\tau : \Sigma = \sum_{i=1}^d \lambda_i u_i u_i^t \mapsto \text{clip}^\tau \Sigma := \sum_{i=1}^d (\lambda_i \wedge \tau) u_i u_i^t$$

Lemma (Altschuler et al., 2023): For any $m \in \mathbb{R}^d, \tau > 0$, and $\Sigma, \Sigma' \in \mathbf{S}^d_{++}$,

$$W_2(p_{m,\text{clip}^\tau \Sigma}, p_{m,\text{clip}^\tau \Sigma'}) \leq W_2(p_{m,\Sigma}, p_{m,\Sigma'})$$

Proof of non-asymptotic guarantees

Lemma: If $0 \prec \alpha I \preceq \nabla^2 V \preceq I$ and $h \leq \alpha^2/60$. Then, if in Bures-SGD, $\Sigma_k \succeq \frac{\alpha}{9} I$, then $\Sigma_k^+ \succeq \frac{\alpha}{9} I$.

Sketch of proof: The idea is to write Σ_k^+ as a generalized Bures-Wasserstein barycenter at Σ_k for a certain distribution P, i.e.

$$\Sigma_k^+ = \exp_{\Sigma_k}\left(\int \log_{\Sigma_k}(\Sigma) dP(\Sigma)\right)$$

Then, some inequality results from Altschuler et al., 2023 are exploited.

Proof of non-asymptotic guarantees

$$\mathcal{F}_k := \sigma(\hat{X}_0, \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{k-1})$$

We bound

$$\mathbb{E}[W_2^2(p_{k+1}, \hat{\pi}) \mid \mathcal{F}_k] \leq \mathbb{E}[W_2^2(p_k^+, \hat{\pi}) \mid \mathcal{F}_k]$$

By writing the Wasserstein distance as expectation under the optimal coupling of $X_k \sim p_k$ and $Z \sim \hat{\pi}$ and then bounding it through strong convexity of $\mathrm{KL}(\cdot \parallel \pi)$, it holds

$$\mathbb{E}[W_2^2(p_{k+1}, \hat{\pi}) \mid \mathcal{F}_k] \leq (1 - 2\alpha h) W_2^2(p_k, \hat{\pi}) + h^2 \mathrm{err}$$

Proof of non-asymptotic guarantees

Using the previous lemma, the definition od the stochastic gradient and Poincaré inequality, it holds

$$\text{err} \leq \frac{36d}{\alpha} + 6W_2^2(p_k, \hat{\pi})$$

Assuming $h \leq \frac{\alpha^2}{60}$,

$$\mathbb{E}[W_2^2(p_{k+1}, \hat{\pi}) \mid \mathcal{F}_k] \leq (1 - \alpha h) W_2^2(p_k, \hat{\pi}) + \frac{36dh^2}{\alpha}$$

The theorem follows by iterating this result and by small step approximation.

Complements on mixtures of Gaussians

Given $(\mathcal{M}, g)$ Riemannian manifold,

$$\mathcal{P}_2(\mathcal{M}) := \{\mu \in \mathcal{P}(\mathcal{M}) \mid \int d^2(p_0, \cdot) d\mu < \infty \text{ for some } p_0 \in \mathcal{M}\}$$

$$W_2^2(\mu, \nu) := [\inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int d^2(x, y) d\gamma(x, y)]$$

$$T_\mu \mathcal{P}_2(\mathcal{M}) := \overline{\{\nabla\psi \mid \psi \in \mathcal{C}_c^\infty(\mathcal{M})\}}^{L^2(\mu)}$$

$$\langle v, w \rangle_\mu := \int g_p(v(p), w(p)) d\mu(p)$$

Given a functional $\mathcal{F}$, $\nabla_{W_2}\mathcal{F}(\mu)$ is the element of $T_\mu\mathcal{P}_2(\mathcal{M})$ such that, for any curve $(\mu_t)_{t\geq 0}$ which satisfies the transport equation $\partial_t\mu_t + \text{div}(\mu_t v_t) = 0$ with $\mu_0 = \mu$, it holds

$$\partial_t|_{t=0} \mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu), v_0 \rangle_\mu = \int g(\nabla_{W_2}\mathcal{F}(\mu), v_0)\, d\mu$$

Complements on mixtures of Gaussians

It follows, using the transport equation and integrating by parts, that

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla \delta \mathcal{F}(\mu)$$

where $\delta \mathcal{F}(\mu) : \mathcal{M} \to \mathbb{R}$ is the first variation of $\mathcal{F}$ at μ. It satisfies

$$\partial_t|_{t=0} \mathcal{F}(\mu_t) = \int \delta \mathcal{F}(\mu)\, \partial_t|_{t=0} \mu_t$$

To find the system of ODEs for the KL divergence functional when $\mathcal{M} = \mathrm{BW}(\mathbb{R}^d)$, the first variation of the functional is computed and its Bures-Wasserstein gradient. The ODEs for the evolution of the gradient flow then follow the relation

$$\dot{m}_0 = a$$

$$\dot{\Sigma}_0 = S\Sigma_0 + \Sigma_0 S$$

for every curve $(m_t, \Sigma_t)_{t \geq 0}$ in the $\mathrm{BW}(\mathbb{R}^d)$ space with tangent vector (a, S) in 0.

Lack of convexity

The functional $\mu \mapsto \mathrm{KL}(p_\mu \parallel \pi)$ is not convex even when π is strongly log-concave.
Let $d = 1, V = 0$, then $\mathrm{KL}(p_\mu \parallel \pi) = \mathcal{H}(p_\mu)$. Let $\mu_0 = \mathcal{N}(0,1) \otimes \delta_1, \mu_1 = \mathcal{N}(0,\tau^2) \otimes \delta_1$. For the optimal coupling, $\sigma_0^2 = \sigma_1^2 = 1$ and $m_1 = \tau m_0$. The geodesic is thus $\{\mu_t = \mathcal{N}(0,(1-t+t\tau)^2) \otimes \delta_1\}_{t\in[0,1]}$. Thus,

$$p_{\mu_t} = \int \mathcal{N}(m,\sigma^2) d\mu_t(m,\sigma^2) = \mathcal{N}(0, 1 + (1-t+t\tau)^2)$$

Hence,

$$\mathcal{H}(p_{\mu_t}) = \int p_{\mu_t} \log p_{\mu_t} = -\frac{1}{2}\log(2\pi e) - \frac{1}{2}\log(1+(1-t+t\tau)^2)$$

Concave in $[0,1]$ for $\tau = 1/2$.

Complements on HK (WFR) geometry

Fisher-Rao (Hellinger)

- Space $\mathcal{M}_+(\mathbb{R}^d)$ of positive measures.
- Fisher-Rao (Hellinger) distance:

$$d^2_{FR}(\mu_0, \mu_1) = \int (\sqrt{\mu_0} - \sqrt{\mu_1})^2$$

- Curves follow:

$$\partial_t \mu_t = \alpha_t \mu_t$$

$$\|\alpha\|^2_\mu = \frac{1}{4} \int \alpha^2 \, d\mu$$

Wasserstein

- Space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures.
- Wasserstein distance.
- Curves follow:

$$\partial_t \mu_t + \mathrm{div}(\mu_t v_t) = 0$$

$$\|v\|^2_\mu = \int \|v\|^2 \, d\mu$$

Complements on HK (WFR) geometry

Wasserstein-Fisher-Rao (Hellinger-Kantorovich) geometry

- Space $\mathcal{M}_+(\mathbb{R}^d)$
- $T_\mu \mathcal{M}_+(\mathbb{R}^d) = \{(\alpha, v) \mid \exists u : \mathbb{R}^d \to \mathbb{R}, \alpha = u, v = \nabla u\}$
- Curves follow:

$$\partial_t \mu_t + \mathrm{div}(\mu_t v_t) = \alpha_t \mu_t$$

- Norm:

$$\| (\alpha, v) \|_\mu^2 = \int (\alpha^2 + \| v \|^2) d\mu$$

- Wasserstein-Fisher-Rao distance:

$$WFR^2(\mu_0, \mu_1) := \inf\{ \int_0^1 \| (\alpha_t, v_t) \|_{\mu_t}^2 \, dt \mid (\mu_t, \alpha_t, v_t) \text{ solves reaction-diffusion PDE}\}$$

Complements on HK (WFR) geometry

The idea to develop the particle system algorithm is to define *particles with mass* (x_t, r_t) where the trajectory $t \mapsto x_t$ develops along a diffusion process and the mass along a reaction process, i.e.

$$\begin{aligned}\dot{x}_t &= v_t(x_t)\\ \dot{r}_t &= (\alpha_t - \int \alpha_t d\mu_t) r_t \quad \text{(normalized to evolve in } \mathcal{P}_2(\mathcal{M}))\end{aligned}$$

To derive the gradient flow, we take v_t and α_t as the WFR gradient of the KL functional.

Empirical results

Target: Bayesian logistic regression.

- synthetic dataset: $\mathcal{D} = \{(x_i, y_i) : i = 1, \dots, N\}$;
- binary label $y_i \in \{0, 1\}$, given by ($z \in \mathbb{R}^d$)

$$\pi(y_i \mid x_i, z) = \sigma(x_i^t z)^{y_i}(1 - \sigma(x_i^t z))^{1-y_i}$$

- target posterior with uninformative prior:

$$\pi(z \mid \mathcal{D}) = \frac{1}{Z} \prod_{i=1}^{N} \pi(y_i \mid x_i, z)$$

- Langevin dynamic associated to:

$$-\nabla V(z) = \nabla \log \pi(z \mid \mathcal{D}) = \sum_{i=1}^{N} (y_i - \sigma(x_i^t z)) x_i$$

- x_i generated from two Gaussian distributions $\mathcal{N}(m_{y_i}^*, \Sigma^*)$ with separation factor $\| m_1^* - m_0^* \| =: s$.

Empirical results

The $\mathrm{KL}(\hat{\pi} \parallel \pi)$ is computed up to the normalizing constant, i.e., $\mathrm{KL}(\hat{\pi} \parallel \pi) - \log Z$. As comparison it is taken the Laplace approximation, i.e. $\hat{\pi}^{Laplace} = \mathcal{N}(z_0, (\nabla^2 \log \frac{1}{\pi}(z_0))^{-1})$, obtained from the Taylor approximation of $\log \pi(z)$ around a proper mode (point where the gradient is 0) z_0 find with the L-BFGS algorithm.

Gaussian VI (through Särkkä's integration) - $d = 2$

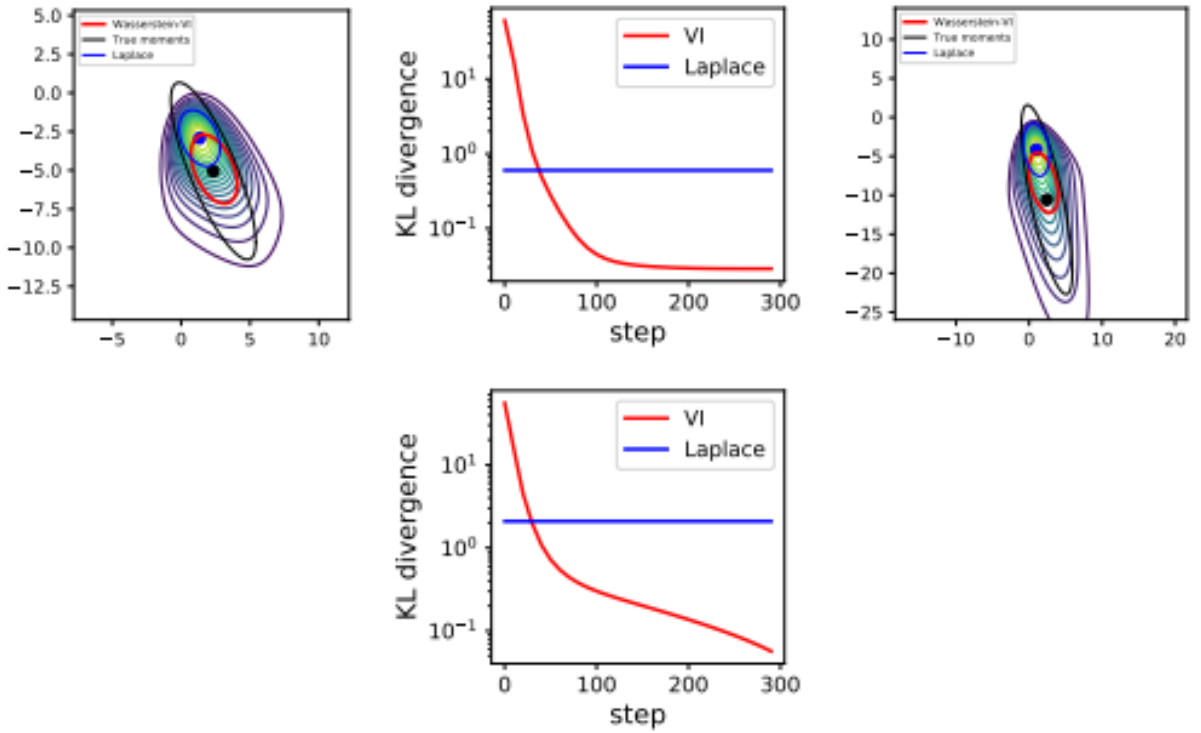


Figure 7: Results in dimension $d = 2$, $N = 10$ for a separation factor $s = 1.5$ (upper row) and $s = 2$ (lower row). The left column shows the true density via contour lines, the true mean (black dot) and covariance (black ellipsoid), and the results of the Laplace and Wasserstein VI approximations as blue and red ellipsoids respectively. The right column shows the evolution of the left KL divergence for Gaussian VI on a logarithmic scale. The corresponding KL divergence obtained with Laplace approximation is shown as a blue straight line.

Gaussian VI (through Särkkä's integration) - $d = 10$

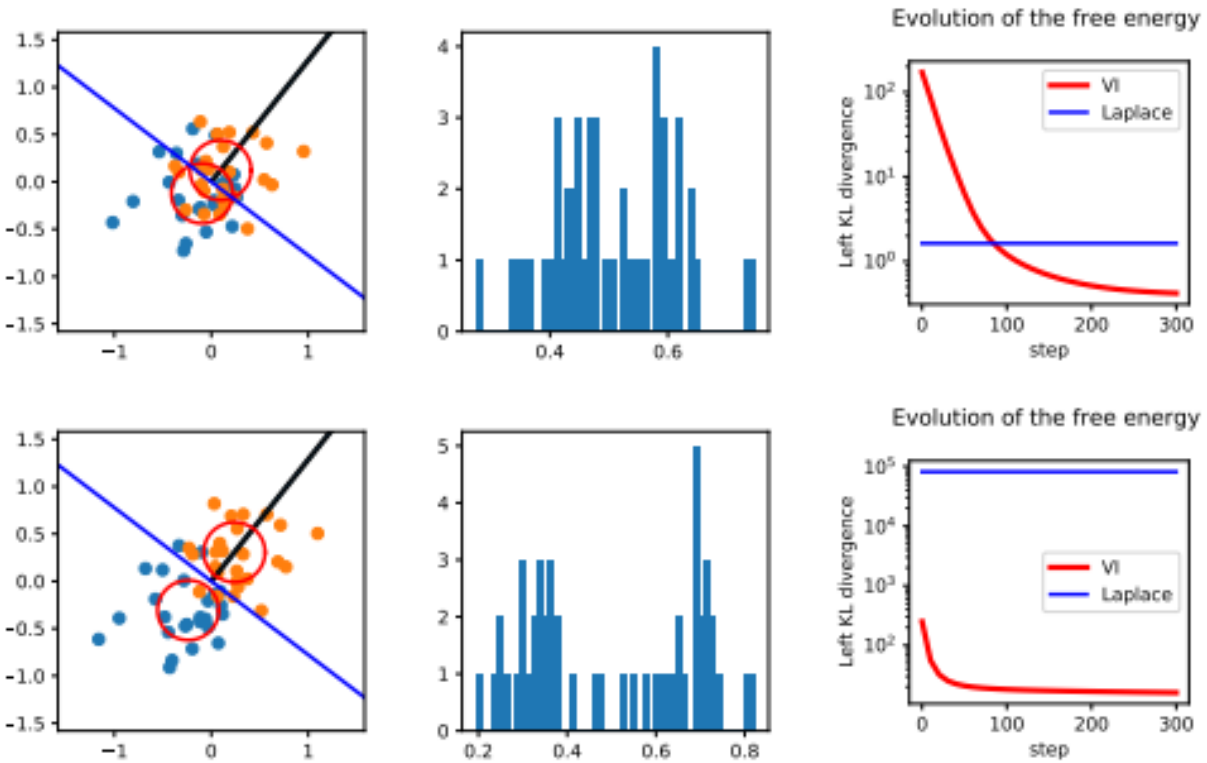


Figure 8: Results in dimension $d = 10$, $N = 50$ for a separation factor $s = 0.6$ (upper row) and $s = 1.5$ (lower row). Left column: synthetic dataset projected onto the two first coordinates. Middle column: histogram representing the number of examples predicted at a given probability by the obtained classifier. Right column: convergence in terms of unnormalized KL divergence. The unnormalized KL is computed via (55) letting $Z = 1$ (upper row) and $Z = 10^{20}$ (lower row). The Runge–Kutta step size is set to 0.1.

Gaussian VI (through Särkkä's integration) - $d = 100$

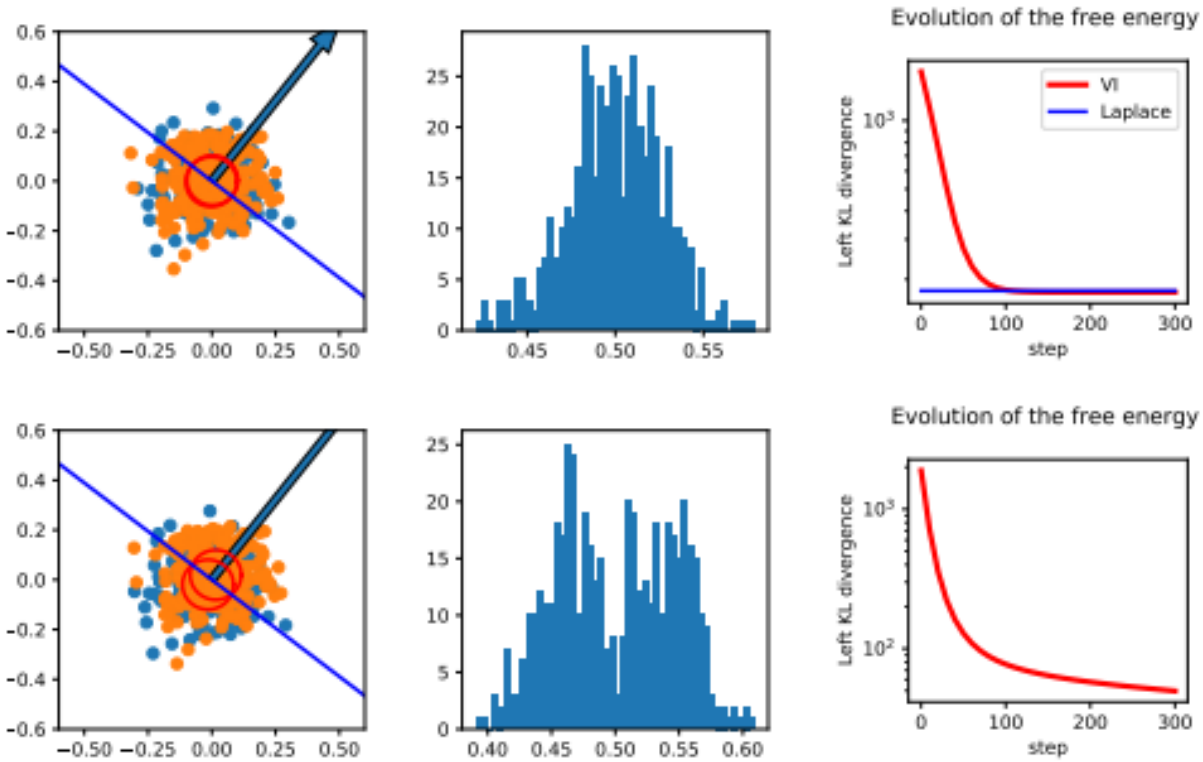


Figure 9: Same as Figure 8 but with dimension $d = 100$, $N = 500$, with separation factor $s = 0.05$ (upper row) and $s = 0.3$ (lower row). The unnormalized KL is computed letting $Z = 1$ (upper row) and $Z = 10^{100}$ (lower row). The unnormalized KL divergence for the Laplace method is not shown in the lower plot because it is too large to be visualized.

Mixture of Gaussian VI

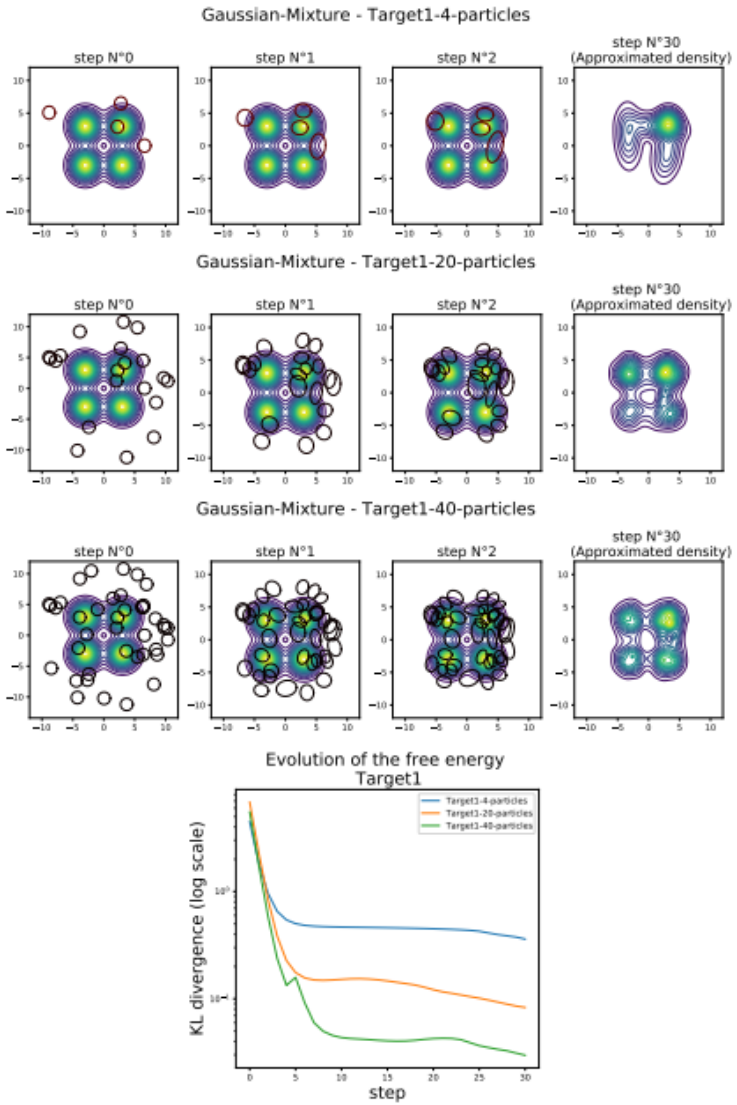


Figure 13: A target with 4 equally weighted modes and isotropic covariances.

Mixture of Gaussian VI

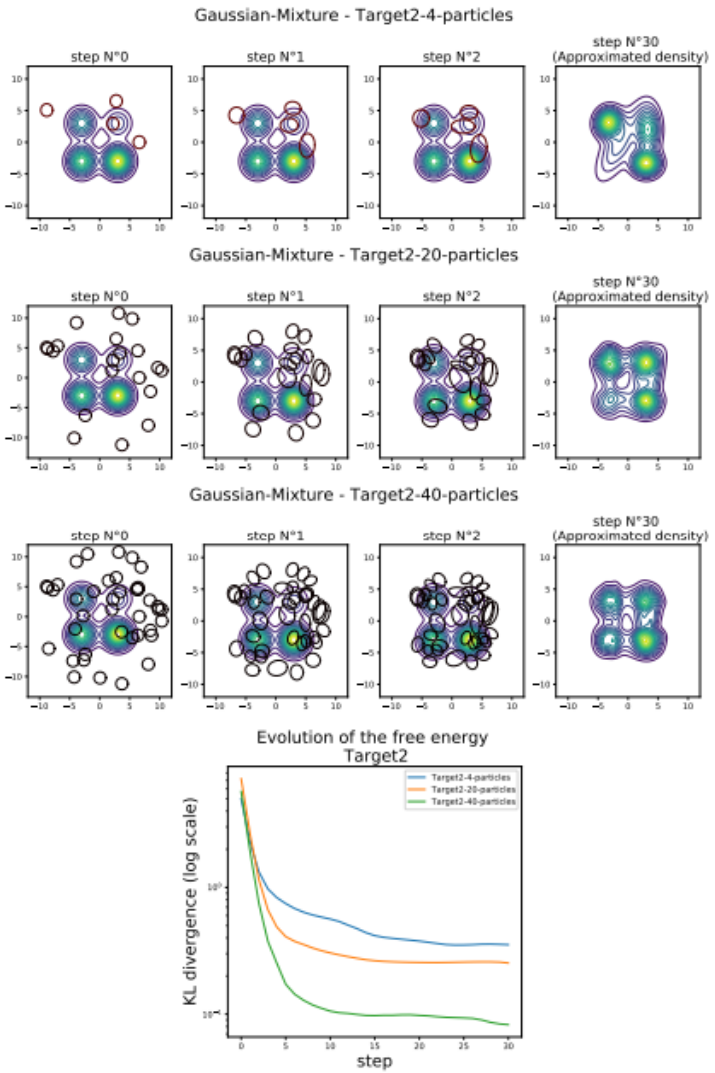
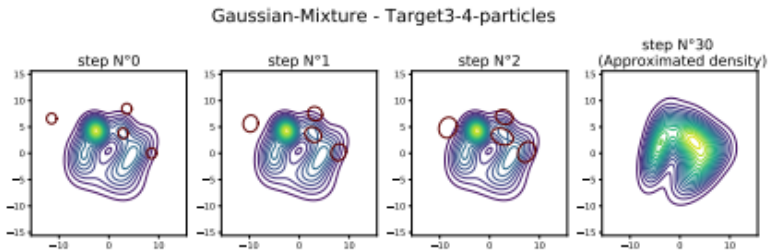
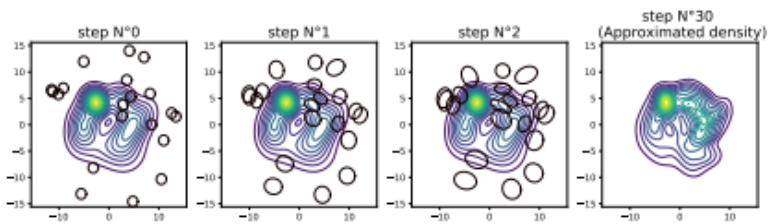
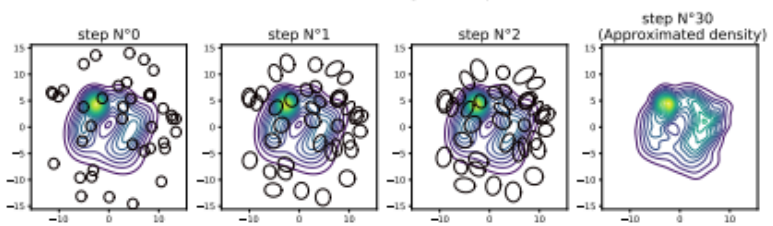


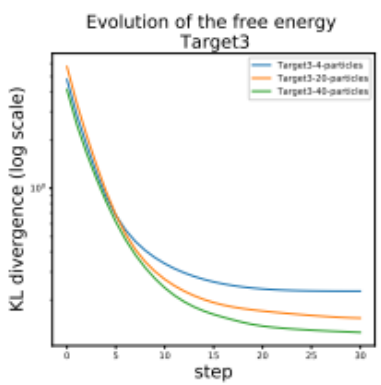
Figure 14: A target with 4 non-equally weighted modes and isotropic covariances.

Mixture of Gaussian VI









Bibliography (1/3)

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet.
Variational inference via Wasserstein gradient flows.
In *Advances in Neural Information Processing Systems (NeurIPS) 35*, 2022.

Jason Altschuler, Sinho Chewi, Patrik Gerber, and Austin J. Stromme.
Averaging on the Bures–Wasserstein manifold: dimension-free convergence of gradient descent.
arXiv preprint, art. arXiv:2106.08502, 2023.

Bibliography (2/3)


Felix Otto.
The geometry of dissipative evolution equations: the porous medium equation.
Comm. Partial Differential Equations, 26(1–2):101–174, 2001.


Richard Jordan, David Kinderlehrer, and Felix Otto.
The variational formulation of the Fokker–Planck equation.
SIAM Journal on Mathematical Analysis, 29(1):1–17, 1998.

Bibliography (3/3)


Cédric Villani.
Optimal Transport, volume 338 of Grundlehren der Mathematischen Wissenschaften.
Springer-Verlag, Berlin, 2009.


Jean-David Benamou and Yann Brenier.
A numerical method for the optimal time-continuous mass transport problem and related problems.
In *Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997)*, volume 226 of Contemp. Math., pages 1–11. Amer. Math. Soc., Providence, RI, 1999.