Understanding and Accelerating Particle-Based Variational Inference

Chang Liu[†], Jingwei Zhuo[†], Pengyu Cheng[‡], Ruiyi Zhang[‡], Jun Zhu[†§], Lawrence Carin[‡§]

†: Department of Computer Science and Technology, Tsinghua University
‡: Department of Electrical and Computer Engineering, Duke University
§: Corresponding authors

chang-li14@mails.tsinghua.edu.cn

ICML 2019

Introduction

Particle-based Variational Inference Methods (ParVIs):

- Represent the variational distribution q by particles; update the particles to minimize $\mathrm{KL}_p(q)$.
- More flexible than classical VIs; more particle-efficient than MCMCs.

Introduction

Particle-based Variational Inference Methods (ParVIs):

- Represent the variational distribution q by particles; update the particles to minimize $\mathrm{KL}_p(q)$.
- More flexible than classical VIs; more particle-efficient than MCMCs.

What is known:

- Stein Variational Gradient Descent (SVGD) [13] simulates the gradient flow (steepest descending curves) of KL_p on $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ [12].
- The Blob and w-SGLD methods [5] simulate the gradient flow of KL_p on the Wasserstein space $\mathcal{P}_2(\mathcal{X})$.

Introduction

Particle-based Variational Inference Methods (ParVIs):

- Represent the variational distribution q by particles; update the particles to minimize $\mathrm{KL}_p(q)$.
- More flexible than classical VIs; more particle-efficient than MCMCs.

What is known:

- Stein Variational Gradient Descent (SVGD) [13] simulates the gradient flow (steepest descending curves) of KL_p on $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ [12].
- The Blob and w-SGLD methods [5] simulate the gradient flow of KL_p on the Wasserstein space $\mathcal{P}_2(\mathcal{X})$.

What remains unknown:

- Do ParVIs make assumptions when simulating the gradient flow? Does one assume stronger than another?
- Is it possible to accelerate the gradient flow?
- Is there a principle for selecting the bandwidth parameter?

Contributions

Findings:

- SVGD approximates the gradient flow on $\mathcal{P}_2(\mathcal{X})$.
- ParVIs approximate the $\mathcal{P}_2(\mathcal{X})$ gradient flow by a compulsory smoothing treatment.
- Various ParVIs either smooth the density or smooth functions, and they are equivalent.

Methods:

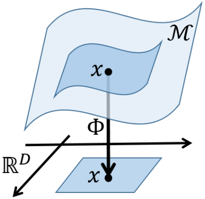
- Two novel ParVIs.
- An acceleration framework for general ParVIs.
- A principled bandwidth selection method for the smoothing kernel.

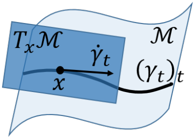
Basic Concepts

Spaces:

- Metric space: a set $\mathcal{M}$ with a distance function $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$.
- Riemannian manifold:
  A topological space $\mathcal{M}$ that locally behaves like an Euclidean space (manifold), and there is an inner product $\langle \cdot, \cdot \rangle_{T_x\mathcal{M}}$ in each of its tangent space $T_x\mathcal{M}$ (Riemannian).
  - Tangent space is the structure of manifolds.
  - Riemannian manifolds are metric spaces:

$$d(x, y) := \inf_{(\gamma_t)_t : \gamma_0 = x, \gamma_1 = y} \sqrt{\int_0^1 \langle \dot{\gamma}_t, \dot{\gamma}_t \rangle_{T_{\gamma_t}\mathcal{M}} \, \mathrm{d}t}.$$





Basic Concepts

Gradient flow $\{(x_t)_t\}$ of a function f: steepest descending curves.

- On metric spaces: various defs ([1], Def. 11.1.1; [18], Def. 23.7), *e.g.*, the Mimimizing Movement Scheme (MMS) ([1], Def. 2.0.6):

$$x_{t+\varepsilon} = \operatorname*{argmin}_{x \in \mathcal{M}} f(x) + \frac{1}{2\varepsilon} d^2(x, x_t).$$

- On Riemannian manifolds: $\dot{x}_t = -\operatorname{grad} f(x_t)$, where:

$$\langle \operatorname{grad} f(x), v \rangle_{T_x\mathcal{M}} = v[f] := \sum_i v^i \partial_i f, \forall v \in T_x\mathcal{M},$$

  which is equivalent to:

$$\operatorname{grad} f(x) = \max \cdot \operatorname*{argmax}_{v \in T_x\mathcal{M}, \|v\|_{T_x\mathcal{M}}=1} \frac{\mathrm{d}}{\mathrm{d}t} f(x_t).$$

  It coincides with MMS on Riemannian manifolds.

The Wasserstein Space $\mathcal{P}_2(\mathcal{X})$

$$\mathcal{P}_2(\mathcal{X}) := \left\{ q\text{: distribution on } \mathcal{X} \;\middle|\; \exists x_0 \in \mathcal{X} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \right\}.$$

Consider Euclidean support space $\mathcal{X} = \mathbb{R}^D$ afterwards.

- $\mathcal{P}_2(\mathcal{X})$ as a metric space ([18], Def 6.4):

$$d_W(q, p) := \left(\inf_{\pi \in \Pi(q,p)} \mathbb{E}_{\pi(x,y)}[d(x, y)^2] \right)^{1/2},$$

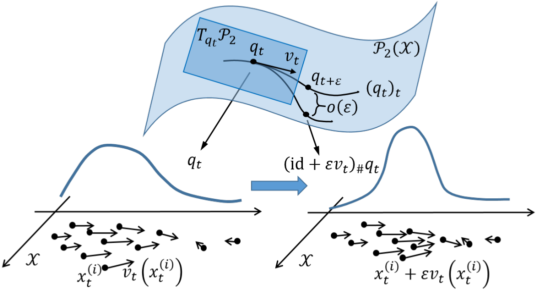
where

$$\Pi(q, p) := \left\{ \pi\text{: distribution on } \mathcal{X} \times \mathcal{X} \;\middle|\; \int_{\mathcal{X}} \pi(x, y)\, \mathrm{d}y = q(x), \int_{\mathcal{X}} \pi(x, y)\, \mathrm{d}x = p(y) \right\}.$$

The Wasserstein Space $\mathcal{P}_2(\mathcal{X})$

$$\mathcal{P}_2(\mathcal{X}) := \left\{ q\text{: distribution on } \mathcal{X} \;\middle|\; \exists x_0 \in \mathcal{X} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \right\}.$$

- $\mathcal{P}_2$ as a Riemannian manifold [17, 18, 1] ($\mathcal{X} = \mathbb{R}^D$):

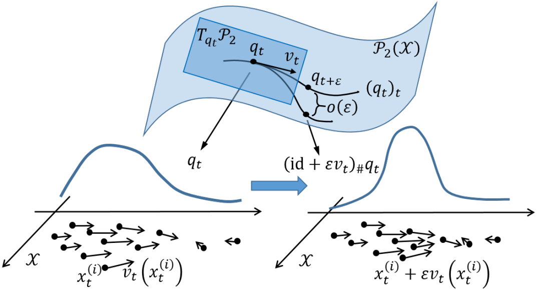


  - Tangent vector $\partial_t q_t$ on $\mathcal{P}_2(\mathcal{X}) \iff$ Vector field v_t on $\mathcal{X}$.
    $\{x^{(i)}\}_{i=1}^N \sim q_t \implies \{x^{(i)} + \varepsilon v_t(x^{(i)})\}_{i=1}^N \sim (\mathrm{id} + \varepsilon v_t)_\# q_t = q_{t+\varepsilon} + o(\varepsilon)$.
    ([1], Prop 8.1.8)

The Wasserstein Space $\mathcal{P}_2(\mathcal{X})$

$$\mathcal{P}_2(\mathcal{X}) := \left\{ q\text{: distribution on } \mathcal{X} \,\middle|\, \exists x_0 \in \mathcal{X} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \right\}.$$

- $\mathcal{P}_2$ as a Riemannian manifold [17, 18, 1] ($\mathcal{X} = \mathbb{R}^D$):



- Tangent space: $T_q\mathcal{P}_2 := \overline{\{\nabla\varphi \mid \varphi \in C_c^\infty\}}^{\mathcal{L}_q^2}$,
  $\mathcal{L}_q^2 := \{u : \mathbb{R}^D \to \mathbb{R}^D \mid \int_{\mathcal{X}} \|u(x)\|_2^2 \,\mathrm{d}q < \infty\}$, $\varphi : \mathbb{R}^D \to \mathbb{R}$.
  ([18], Thm 13.8; [1], Thm 8.3.1, Def 8.4.1, Prop 8.4.5)
- Riemannian metric: $\langle v, u\rangle_{T_q\mathcal{P}_2} := \int_{\mathcal{X}} v(x) \cdot u(x)\, q(x)\, \mathrm{d}x$.
  (consistent with the Wasserstein distance d_W [3])

The Wasserstein Space $\mathcal{P}_2(\mathcal{X})$

$$\mathcal{P}_2(\mathcal{X}) := \left\{ q\text{: distribution on } \mathcal{X} \;\middle|\; \exists x_0 \in \mathcal{X} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \right\}.$$

- Gradient flow on $\mathcal{P}_2(\mathcal{X})$ for $\mathrm{KL}_p(q) := \mathbb{E}_q[\log(q/p)]$:
  - $\mathcal{P}_2(\mathcal{X})$ as a Riemannian manifold:

$$v^{\mathsf{GF}} := -\operatorname{grad} \mathrm{KL}_p(q) = -\nabla\Big(\frac{\delta}{\delta q}\mathrm{KL}_p(q)\Big) = \nabla \log p - \nabla \log q.$$

    ([18], Thm 23.18; [1], Example 11.1.2)
  - Minimizing Movement Scheme (MMS) ([1], Def. 2.0.6):

$$q_{t+\varepsilon} = \operatorname*{argmin}_{q \in \mathcal{P}_2(\mathcal{X})} \mathrm{KL}_p(q) + \frac{1}{2\varepsilon} d_W^2(q, q_t).$$

  They coincide under the Riemannian structure.

  ([18], Prop. 23.1, Rem. 23.4; [1], Thm. 11.1.6; [8], Lem. 2.7)

Remark 1

The Langevin dynamics $\mathrm{d}x = \nabla \log p(x)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t(x)$ *(*B_t *is the Brownian motion) is also the gradient flow of* KL_p *on* $\mathcal{P}_2(\mathcal{X})$ *[9].*

Particle-Based Variational Inference Methods (ParVIs)

- Stein Variational Gradient Descent (SVGD) [13]:

$$v^{\mathsf{SVGD}}(\cdot) := \max_{v \in \mathcal{H}^D, \|v\|_{\mathcal{H}^D}=1} \cdot \operatorname{argmax} -\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathrm{KL}_p\big((\mathrm{id} + \varepsilon v)_{\#} q\big)\Big|_{\varepsilon=0}$$
$$= \mathbb{E}_{q(x)}[K(x,\cdot)\nabla \log p(x) + \nabla_x K(x,\cdot)],$$

  where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) of kernel K.
  - v^{SVGD} is the gradient flow of KL_p on a kernel-related distribution manifold $\mathcal{P}_{\mathcal{H}}$ [12].
- Blob (w-SGLD-B) [5]:

$$v^{\mathsf{Blob}} := -\nabla\Big(\frac{\delta}{\delta q}\mathbb{E}_q[\log(\tilde{q}/p)]\Big)$$
$$= \nabla \log p - \nabla \log \tilde{q} - \nabla\big((q/\tilde{q}) * K\big),$$
$$\tilde{q} := q * K.$$

Particle-Based Variational Inference Methods (ParVIs)

- Particle Optimization (PO) [4]: using MMS; estimate d_W by solving the dual optimal transport problem.

$$x_k^{(i)} = x_{k-1}^{(i)} + \varepsilon(v^{\mathsf{SVGD}}(x_{k-1}^{(i)}) + \mathcal{N}(0, \sigma^2 I)) + \mu(x_{k-1}^{(i)} - x_{k-2}^{(i)}).$$

- w-SGLD [5]: using MMS; estimate d_W by solving the primal problem. Similar update rule.

SVGD Approximates $\mathcal{P}_2(\mathcal{X})$ Gradient Flow

Reformulate v^{GF} as:

$$v^{\mathsf{GF}} = \max_{v \in \mathcal{L}_q^2, \|v\|_{\mathcal{L}_q^2}=1} \cdot \operatorname{argmax} \langle v^{\mathsf{GF}}, v \rangle_{\mathcal{L}_q^2}. \tag{1}$$

We find:

Theorem 2 (v^{SVGD} approximates v^{GF})

$$v^{\mathsf{SVGD}} = \max_{v \in \mathcal{H}^D, \|v\|_{\mathcal{H}^D}=1} \cdot \operatorname{argmax} \langle v^{\mathsf{GF}}, v \rangle_{\mathcal{L}_q^2}.$$

- $\mathcal{H}^D$ is a subspace of $\mathcal{L}_q^2$, so v^{SVGD} is the projection of v^{GF} on $\mathcal{H}^D$.
- The $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$-gradient-flow interpretation of SVGD: $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ is not a very nice manifold.
- All ParVIs approximate the $\mathcal{P}_2(\mathcal{X})$ gradient flow.

ParVIs Approximate $\mathcal{P}_2(\mathcal{X})$ Gradient Flow by Smoothing

Smoothing Functions

- SVGD restricts the optimization domain $\mathcal{L}_q^2$ to $\mathcal{H}^D$.

Theorem 3 ($\mathcal{H}^D$ smooths $\mathcal{L}_q^2$)

For $\mathcal{X} = \mathbb{R}^D$, a Gaussian kernel K on $\mathcal{X}$ and an absolutely continuous q, the vector-valued RKHS $\mathcal{H}^D$ of K is isometrically isomorphic to the closure $\mathcal{G} := \overline{\{\phi * K : \phi \in \mathcal{C}_c^\infty\}}^{\mathcal{L}_q^2}$.

$\overline{\mathcal{C}_c^\infty}^{\mathcal{L}_q^2} = \mathcal{L}_q^2$ ([11], Thm. 2.11) $\Longrightarrow \mathcal{G}$ is roughly the kernel-smoothed $\mathcal{L}_q^2$.

- PO solves the dual problem by restricting the optimization domain of Lipschitz functions to quadratic functions.

ParVIs Approximate $\mathcal{P}_2(\mathcal{X})$ Gradient Flow by Smoothing

Smoothing the Density

- Blob partially smooths the density.

$$v^{\mathsf{GF}} = -\nabla\Big(\frac{\delta}{\delta q}\mathbb{E}_q[\log(q/p)]\Big) \Longrightarrow v^{\mathsf{Blob}} = -\nabla\Big(\frac{\delta}{\delta q}\mathbb{E}_q[\log(\tilde{q}/p)]\Big).$$

- w-SGLD adds an entropy regularizer in the primal objective function.

$$d_W^2(\{x^{(i)}\}_{i=1}^N, \{y^{(j)}\}_{j=1}^N) \approx \min_{\pi_{ij}} \sum_{i,j} \pi_{ij} d_{ij}^2 + \lambda \sum_{i,j} \pi_{ij} \log \pi_{ij},$$

$$\text{s.t. } \sum_i \pi_{ij} = 1/N, \sum_j \pi_{ij} = 1/N.$$

ParVIs Approximate $\mathcal{P}_2(\mathcal{X})$ Gradient Flow by Smoothing

- Equivalence:
  Smoothing-function objective $= \mathbb{E}_q[L(v)]$, $L : \mathcal{L}_q^2 \to L_q^2$ linear.
  $$\Longrightarrow \mathbb{E}_{\tilde{q}}[L(v)] = \mathbb{E}_{q*K}[L(v)] = \mathbb{E}_q[L(v) * K] = \mathbb{E}_q[L(v * K)].$$
- Necessity: $\operatorname{grad} \mathrm{KL}_p(q)$ undefined at $q = \hat{q} := \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}$.

Theorem 4 (Necessity of smoothing for SVGD)

For $q = \hat{q}$ and $v \in \mathcal{L}_p^2$, problem (1):

$$\max_{v\in\mathcal{L}_p^2, \|v\|_{\mathcal{L}_p^2}=1} \left\langle v^{\mathsf{GF}}, v \right\rangle_{\mathcal{L}_{\hat{q}}^2},$$

has no optimal solution. In fact the supremum of the objective is infinite, indicating that a maximizing sequence of v tends to be ill-posed.

ParVIs rely on the smoothing assumption! No free lunch!

New ParVIs with Smoothing

- Gradient Flow with Smoothed Density (GFSD):
  Fully smooth the density:
  $$v^{\mathsf{GFSD}} := \nabla \log p - \nabla \log \tilde{q}.$$
- Gradient Flow with Smoothed test Functions (GFSF):
  $$v^{\mathsf{GF}} = \nabla \log p - \nabla \log q$$
  $$\implies v^{\mathsf{GF}} = \nabla \log p + \operatorname*{argmin}_{u \in \mathcal{L}^2} \max_{\substack{\phi \in \mathcal{C}_c^\infty, \\ \|\phi\|_{\mathcal{L}_q^2} = 1}} \Big(\mathbb{E}_q[\phi \cdot u - \nabla \cdot \phi]\Big)^2.$$
  Smooth ϕ: take ϕ from $\mathcal{H}^D$:
  $$v^{\mathsf{GFSF}} := \nabla \log p + \operatorname*{argmin}_{u \in \mathcal{L}^2} \max_{\substack{\phi \in \mathcal{H}^D, \\ \|\phi\|_{\mathcal{H}^D} = 1}} \Big(\mathbb{E}_q[\phi \cdot u - \nabla \cdot \phi]\Big)^2.$$
  Solution: $\hat{v}^{\mathsf{GFSF}} = \hat{g} + \hat{K}' \hat{K}^{-1}$. (Note $\hat{v}^{\mathsf{SVGD}} = \hat{v}^{\mathsf{GFSF}} \hat{K}$.)
  $\hat{g}_{:,i} = \nabla_{x^{(i)}} \log p(x^{(i)})$, $\hat{K}_{ij} = K(x^{(i)}, x^{(j)})$, $\hat{K}'_{:,i} = \sum_j \nabla_{x^{(j)}} K(x^{(j)}, x^{(i)})$.

Nesterov's Acceleration Methods on Riemannian Manifolds

$r_k \in \mathcal{P}_2(\mathcal{X})$: auxiliary variable. $v_k := -\operatorname{grad} \mathrm{KL}(r_k)$.

- Riemannian Accelerated Gradient (RAG) [14] (with simplification):

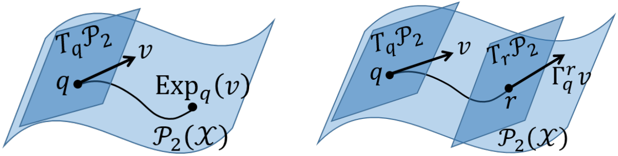
$$\begin{cases} q_k = \mathrm{Exp}_{r_{k-1}}(\varepsilon v_{k-1}), \\ r_k = \mathrm{Exp}_{q_k}\left[-\Gamma_{r_{k-1}}^{q_k}\left(\frac{k-1}{k}\mathrm{Exp}_{r_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon v_{k-1}\right)\right]. \end{cases}$$

- Riemannian Nesterov's method (RNes) [20] (with simplification):

$$\begin{cases} q_k = \mathrm{Exp}_{r_{k-1}}(\varepsilon v_{k-1}), \\ r_k = \mathrm{Exp}_{q_k}\left\{c_1 \mathrm{Exp}_{q_k}^{-1}\left[\mathrm{Exp}_{r_{k-1}}\left((1-c_2)\mathrm{Exp}_{r_{k-1}}^{-1}(q_{k-1}) + c_2 \mathrm{Exp}_{r_{k-1}}^{-1}(q_k)\right)\right]\right\}. \end{cases}$$

Required:

- Exponential map $\mathrm{Exp}_q : T_q\mathcal{P}_2(\mathcal{X}) \to \mathcal{P}_2(\mathcal{X})$ and its inverse.
- Parallel transport $\Gamma_q^r : T_q\mathcal{P}_2(\mathcal{X}) \to T_r\mathcal{P}_2(\mathcal{X})$.



Leveraging the Riemannian Structure of $\mathcal{P}_2(\mathcal{X})$

- Exponential map ([18], Coro. 7.22; [1], Prop. 8.4.6; [8], Prop. 2.1): $\mathrm{Exp}_q(v) = (\mathrm{id} + v)_{\#} q$, *i.e.*, $\{x^{(i)}\}_i \sim q \Rightarrow \{x^{(i)} + v(x^{(i)})\}_i \sim \mathrm{Exp}_q(v)$.
- Inverse exponential map: require the optimal transport map.
  - Sinkhorn methods [6, 19] appear costly and unstable.
  - Make approximations when $\{x^{(i)}\}_i$ and $\{y^{(i)}\}_i$ are pairwise close: $d(x^{(i)}, y^{(i)}) \ll \min\left\{ \min_{j \neq i} d(x^{(i)}, x^{(j)}), \min_{j \neq i} d(y^{(i)}, y^{(j)}) \right\}$.

Proposition 5 (Inverse exponential map)

For pairwise close samples $\{x^{(i)}\}_i$ of q and $\{y^{(i)}\}_i$ of r, we have $\left(\mathrm{Exp}_q^{-1}(r)\right)(x^{(i)}) \approx y^{(i)} - x^{(i)}$.

- Parallel transport
  - Hard to implement analytical results [15, 16].
  - Use Schild's ladder method [7, 10] for approximation.

Proposition 6 (Parallel transport)

For pairwise close samples $\{x^{(i)}\}_i$ of q and $\{y^{(i)}\}_i$ of r, we have $\left(\Gamma_q^r(v)\right)(y^{(i)}) \approx v(x^{(i)})$, $\forall v \in T_q \mathcal{P}_2$.

Acceleration Framework for ParVIs

Algorithm 1 The acceleration framework with Wasserstein Accelerated Gradient (WAG) and Wasserstein Nesterov's method (WNes)

WAG: select acceleration factor $\alpha > 3$;
WNes: select or calculate $c_1, c_2 \in \mathbb{R}^+$;
Initialize $\{x_0^{(i)}\}_{i=1}^N$ distinctly; let $y_0^{(i)} = x_0^{(i)}$;
for $k = 1, 2, \cdots, k_{\max}$, **do**
  for $i = 1, \cdots, N$, **do**
    Find $v(y_{k-1}^{(i)})$ by SVGD/Blob/GFSD/GFSF;
    $x_k^{(i)} = y_{k-1}^{(i)} + \varepsilon v(y_{k-1}^{(i)})$;
    $y_k^{(i)} = x_k^{(i)} + \begin{cases} \text{WAG: } \frac{k-1}{k}(y_{k-1}^{(i)} - x_{k-1}^{(i)}) + \frac{k+\alpha-2}{k}\varepsilon v(y_{k-1}^{(i)}); \\ \text{WNes: } c_1(c_2-1)(x_k^{(i)} - x_{k-1}^{(i)}); \end{cases}$
  end for
end for
Return $\{x_{k_{\max}}^{(i)}\}_{i=1}^N$.

Bandwidth Selection via the Heat Equation

Note

Under the dynamics $\mathrm{d}x = -\nabla \log q_t(x)\,\mathrm{d}t$, q_t evolves following the heat equation (HE): $\partial_t q_t(x) = \Delta q_t(x)$.

Smoothing the density: $q_t(x) \approx \tilde{q}(x) = \tilde{q}(x; \{x^{(i)}\}_{i=1}^N)$. Then for $q_{t+\varepsilon}(x)$,

- Due to HE, $q_{t+\varepsilon}(x) \approx \tilde{q}(x) + \varepsilon \Delta \tilde{q}(x)$.
- Due to the effect of the dynamics, updated particles $\{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N$ approximate $q_{t+\varepsilon}$, so $q_{t+\varepsilon}(x) \approx \tilde{q}(x; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N)$.

Objective: $\sum_k \left(\tilde{q}(x^{(k)}) + \varepsilon \Delta \tilde{q}(x^{(k)}) - \tilde{q}(x^{(k)}; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N) \right)^2$.

Take $\varepsilon \to 0$, make the objective dimensionless (h/x^2 is dimensionless):

$\frac{1}{h^{D+2}} \sum_k \left[\Delta \tilde{q}(x^{(k)}; \{x^{(i)}\}_i) + \sum_j \nabla_{x^{(j)}} \tilde{q}(x^{(k)}; \{x^{(i)}\}_i) \cdot \nabla \log \tilde{q}(x^{(j)}; \{x^{(i)}\}_i) \right]^2$.

Also applicable to smoothing functions.

Toy Experiments

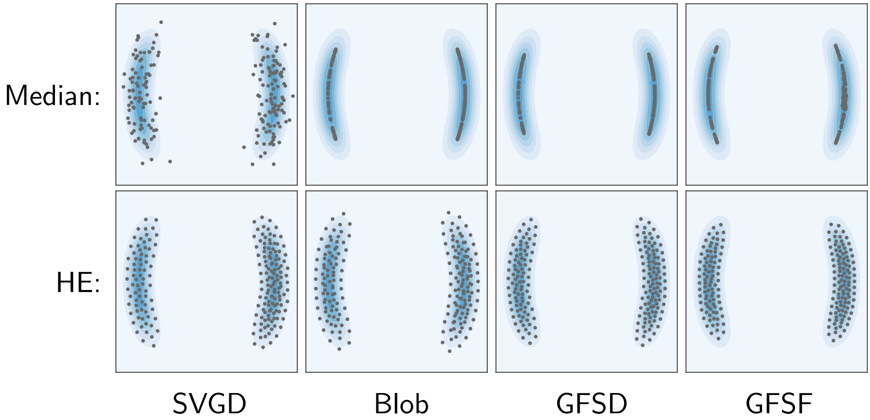


Figure: Comparison of HE (bottom row) with the median method (top row) for bandwidth selection.

Bayesian Logistic Regression (BLR)

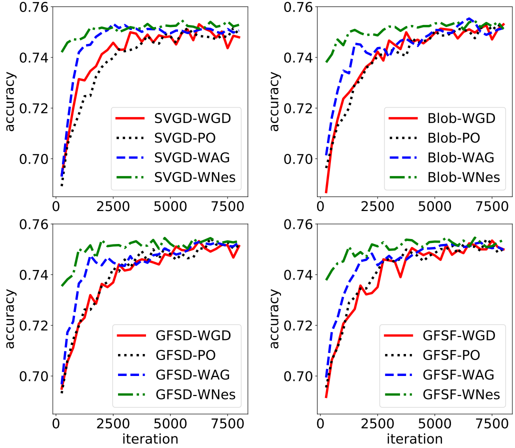


Figure: Acceleration effect of WAG and WNes on BLR on the Covertype dataset, measured by prediction accuracy on test dataset. Each curve is averaged over 10 runs.

Bayesian Neural Networks (BNNs)

Table: Results on BNN on the Kin8nm dataset (one of the UCI datasets [2]). Results are averaged over 20 runs.

Method	Avg. Test RMSE ($\times 10^{-2}$)			
	SVGD	Blob	GFSD	GFSF
WGD	8.4±0.2	8.2±0.2	8.0±0.3	8.3±0.2
PO	7.8±0.2	8.1±0.2	8.1±0.2	8.0±0.2
WAG	7.0±0.2	**7.0±0.2**	7.1±0.1	7.0±0.1
WNes	**6.9±0.1**	7.0±0.2	**6.9±0.1**	**6.8±0.1**

Method	Avg. Test LL			
	SVGD	Blob	GFSD	GFSF
WGD	1.042±0.016	1.079±0.021	1.087±0.029	1.044±0.016
PO	1.114±0.022	1.070±0.020	1.067±0.017	1.073±0.016
WAG	1.167±0.015	**1.169±0.015**	1.167±0.017	1.190±0.014
WNes	**1.171±0.014**	1.168±0.014	**1.173±0.016**	**1.193±0.014**

Latent Dirichlet Allocation (LDA)

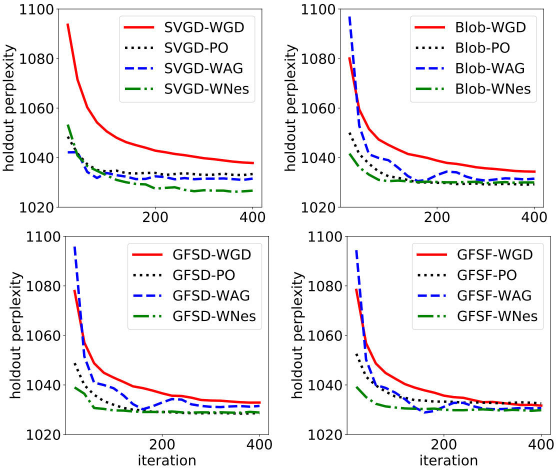


Figure: Acceleration effect of WAG and WNes on LDA. Inference results are measured by the hold-out perplexity. Curves are averaged over 10 runs.

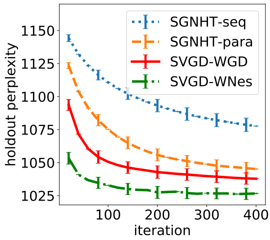


Figure: Comparison of SVGD and SGNHT on LDA, as representatives of ParVIs and MCMCs. Average over 10 runs.

Thank you!

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré.
Gradient flows: in metric spaces and in the space of probability measures.
Springer Science & Business Media, 2008.

Arthur Asuncion and David Newman.
Uci machine learning repository, 2007.

Jean-David Benamou and Yann Brenier.
A computational fluid mechanics solution to the monge-kantorovich mass transfer problem.
Numerische Mathematik, 84(3):375–393, 2000.

Changyou Chen and Ruiyi Zhang.
Particle optimization in stochastic gradient mcmc.
arXiv preprint arXiv:1711.10927, 2017.

Changyou Chen, Ruiyi Zhang, Wenlin Wang, Bai Li, and Liqun Chen.
A unified particle-optimization framework for scalable bayesian sampling.
arXiv preprint arXiv:1805.11659, 2018.

Marco Cuturi.
Sinkhorn distances: Lightspeed computation of optimal transport.
In *Advances in neural information processing systems*, pages 2292–2300, 2013.

J Ehlers, F Pirani, and A Schild.
The geometry of free fall and light propagation, in the book "general relativity" (papers in honour of jl synge), 63–84, 1972.

Matthias Erbar et al.
The heat equation on manifolds as a gradient flow in the wasserstein space.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 46, pages 1–23. Institut Henri Poincaré, 2010.

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.

Arkady Kheyfets, Warner A Miller, and Gregory A Newton.
Schild's ladder parallel transport procedure for an arbitrary connection.
International Journal of Theoretical Physics, 39(12):2891–2898, 2000.

Ondrej Kováčik and Jiří Rákosník.
On spaces $l\hat{p}(x)$ and $w\hat{k}$, $p(x)$.
Czechoslovak Mathematical Journal, 41(4):592–618, 1991.

Qiang Liu.

Stein variational gradient descent as gradient flow.
In *Advances in neural information processing systems*, pages 3118–3126, 2017.

Qiang Liu and Dilin Wang.
Stein variational gradient descent: A general purpose bayesian inference algorithm.
In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao.
Accelerated first-order methods for geodesically convex optimization on riemannian manifolds.
In *Advances in Neural Information Processing Systems*, pages 4875–4884, 2017.

John Lott.
Some geometric calculations on wasserstein space.
Communications in Mathematical Physics, 277(2):423–437, 2008.

John Lott.
An intrinsic parallel transport in wasserstein space.
Proceedings of the American Mathematical Society, 145(12):5329–5340, 2017.

Felix Otto.
The geometry of dissipative evolution equations: the porous medium equation.
2001.

Cédric Villani.
Optimal transport: old and new, volume 338.
Springer Science & Business Media, 2008.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha.
A fast proximal point method for computing wasserstein distance.
arXiv preprint arXiv:1802.04307, 2018.

Hongyi Zhang and Suvrit Sra.
An estimate sequence for geodesically convex optimization.
In *Conference On Learning Theory*, pages 1703–1723, 2018.