On Wasserstein Gradient Flows and Particle-Based Variational Inference

Ruiyi Zhang

Duke University

Joint work with Chang Liu, Changyou Chen and Lawrence Carin

Workshop on Stein's Method, ICML 2019

Introduction

We are in an era of abundant data:

- Text, images, videos from the Internet; raw medical notes from doctors, *etc*.

We need tools for modeling, searching, visualizing, and understanding large-scale data sets.

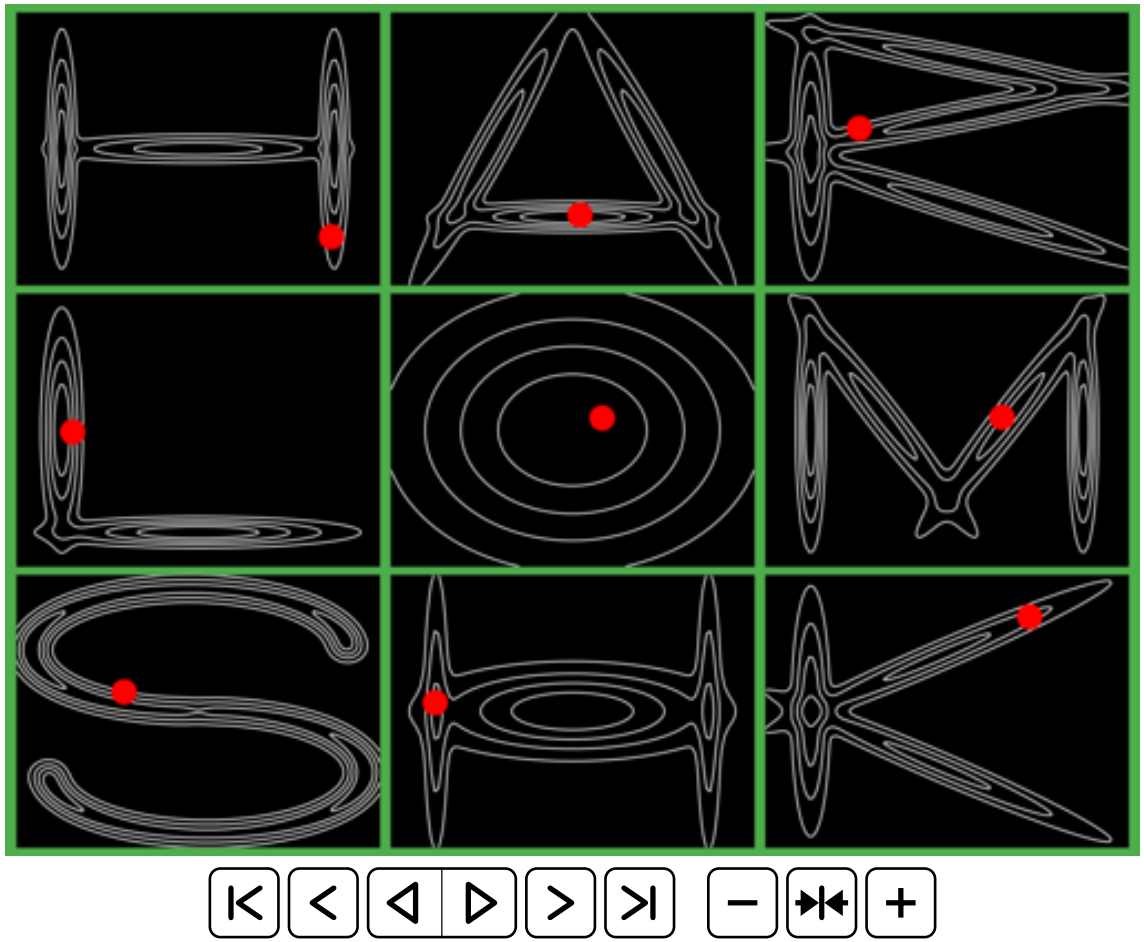
We want our modeling tools:

- Faithfully represent uncertainty in our model structure and parameters.
- Automatically deal with noise in our data.
- Exhibit robustness.

Modeling from a Bayesian perspective!

Demo: Markov-Chain-based Bayesian Sampling

- Nine mixtures of Gaussians[1].
- Sequential of samples connected by yellow lines.



[1] Demo by T. Broderick and D. Duvenaud.

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 MCMC.

A few natural questions:

- How do ParVIs work (unifying and understanding)?
- Can we accelerate ParVIs?

Related Work:

- Stein Variational Gradient Descent (SVGD) [12] simulates the gradient flow (steepest descending curves) of KL_p on $\mathcal{P}_{\mathcal{H}}(\mathcal{X})$ [11].
- Stochastic Gradient Langevin Dynamics (SGLD) [2] simulate the gradient flow of KL_p on the Wasserstein space $\mathcal{P}_2(\mathcal{X})$.

Stochastic Gradient Langevin Dynamic

- Given data $\mathcal{D} = \{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_N\}$, model prior $p(x)$, model (likelihood) $p(\mathcal{D}|x) = \prod_{i=1}^N p(\boldsymbol{b}_i|x)$ on *i.i.d.* assumption.
- Want to sample from the posterior distribution:

$$p(x|\mathcal{D}) \propto p(x)p(\mathcal{D}|x) = p(x)\prod_{i=1}^N p(\boldsymbol{b}_i|x) \ .$$

- SGLD are numerical solutions of continuous-time diffusion processes with stationary distribution equal to $p(x|\mathcal{D})$:

$$\mathrm{d}x_t = F(x_t)\mathrm{d}t + \mathrm{d}\mathcal{W}_t \ .$$

- Define the potential energy (negative unnormalized posterior):

$$U(x) \triangleq -\sum_{i=1}^N \log p(\boldsymbol{b}_i|x) - \log p(x) - \cancel{\log p(\mathcal{D})}$$

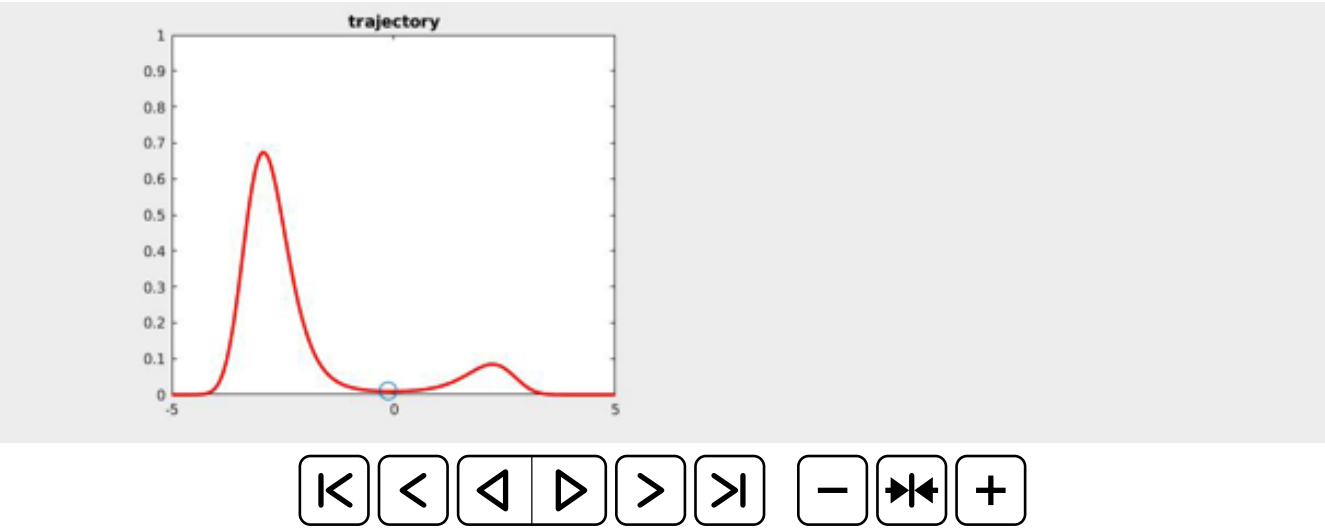
Stochastic Gradient Langevin Dynamic

- In case of large data, define a stochastic version of $U(x)$ with a minibatch of size n:

$$\tilde{U}(x) \triangleq -\frac{N}{n}\sum_{i=1}^{n}\log p(\boldsymbol{b}_i|x) - \log p(x)$$

- Stochastic gradient Langevin dynamic (SGLD) generates samples via

$$x_{\ell+1} = x_\ell + h_{\ell+1}\nabla_x \tilde{U}(\theta_\ell) + \sqrt{2h_{\ell+1}}\zeta_{\ell+1}, \qquad \zeta_{\ell+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$



Stein Variational Gradient Descent

- SVGD iteratively updates an interactive particle system $\{x_\ell^{(i)}\}_{i=1}^M$ via:

$$x_{\ell+1}^{(i)} = x_\ell^{(i)} + h\phi(x_\ell^{(i)}), \quad \phi = \arg\max_{\phi\in\mathcal{F}}\{\frac{\partial}{\partial h}\mathsf{KL}(q_{[h\phi]}||p(x|\mathcal{D}))\}$$

  - $q_{[h\phi]}$: density formed by the particles.
- When $\mathcal{F}$ is an RKHS induced by kernel $K(x, x')$, SVGD endows close-form updates:

$$x_{\ell+1}^{(i)} = x_\ell^{(i)} + \frac{h}{M}\sum_{j=1}^{M}\Big[\underbrace{K(x_\ell^{(j)}, x_\ell^{(i)})\nabla_{x_\ell^{(j)}}\tilde{U}(x_\ell^{(j)})}_{\text{move to high prob. region}} + \underbrace{\nabla_{x_\ell^{(j)}}K(x_\ell^{(j)}, x_\ell^{(i)})}_{\text{repulsive force}}\Big]$$

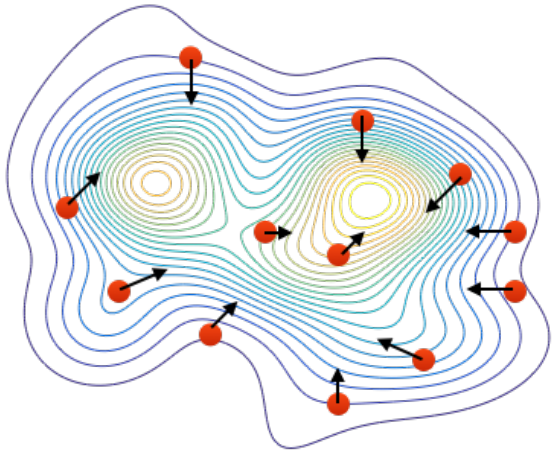


Image credit: Qiang Liu.

Wasserstein Gradient Flows

- $\mathcal{P}_2(\mathcal{X}) := \left\{ q\text{: distribution on } \mathcal{X} \;\middle|\; \exists x_0 \in \mathcal{X} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \right\}$
- WGFs are partial differential equations (PDEs) to describe evolutions of probability distributions over time.
- It has the following general form:

$$\partial_t q_t = \nabla \cdot \left(q_t \nabla \left(\frac{\delta F}{\delta q_t}(q_t) \right) \right) ,$$

  - $F : \mathcal{P}(\Omega) \to \mathbb{R}$ defines the landscape in the space of probability measures, called **energy functional**.
  - Consider $F = \mathrm{KL}_p(q)$, $v^{\mathrm{GF}} := -\nabla \frac{\delta \mathrm{KL}_p(q_t)}{\delta q_t} = \nabla \log p - \nabla \log q$.
  - Typically the stationary distribution q_∞ is our target distribution.

Wasserstein Gradient Flows

- $\mathcal{P}_2(\mathcal{X}) := \left\{ q\text{: distribution on } \mathcal{X} \;\middle|\; \exists x_0 \in \mathcal{X} \text{ s.t. } \mathbb{E}_q[d(x_0, x)^2] < +\infty \right\}$
- WGFs are partial differential equations (PDEs) to describe evolutions of probability distributions over time.
- It has the following general form:

$$\partial_t q_t = \nabla \cdot \left(q_t \nabla \left(\frac{\delta F}{\delta q_t}(q_t) \right) \right) ,$$

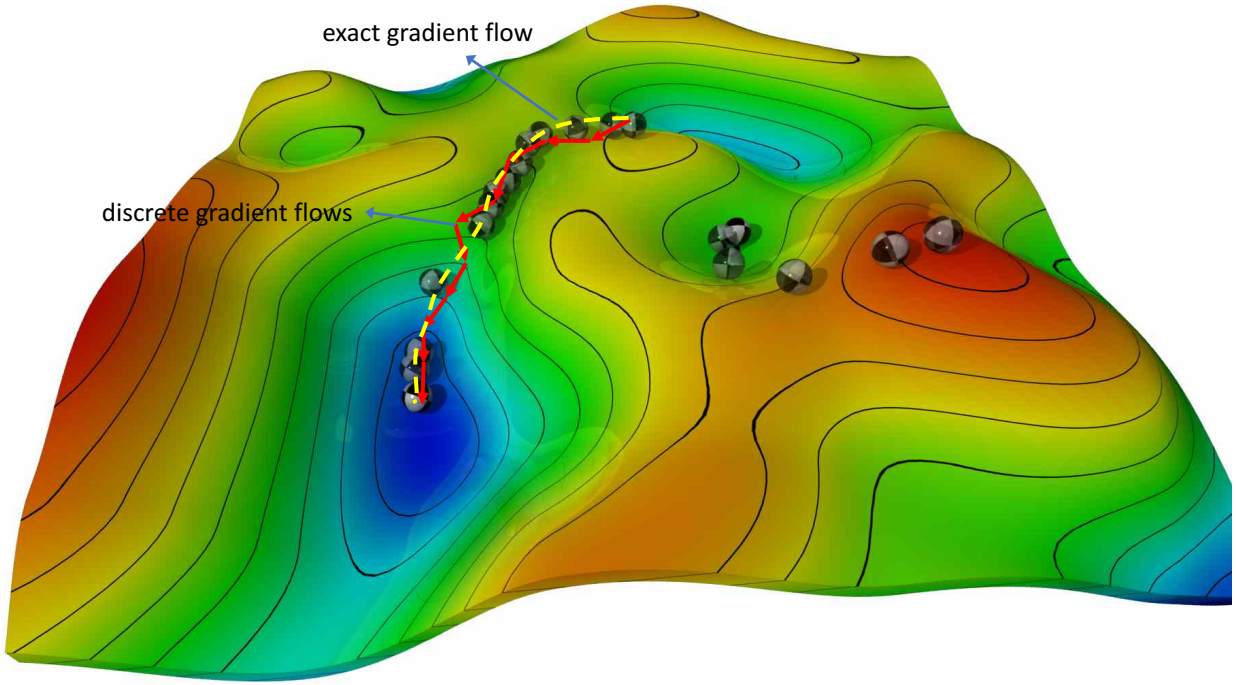
  - $F : \mathcal{P}(\Omega) \to \mathbb{R}$ defines the landscape in the space of probability measures, called **energy functional**.
  - Consider $F = \mathrm{KL}_p(q)$, $v^{\mathrm{GF}} := -\nabla \frac{\delta \mathrm{KL}_p(q_t)}{\delta q_t} = \nabla \log p - \nabla \log q$.
  - Typically the stationary distribution q_∞ is our target distribution.

How to solve it?

1. Discrete gradient flows.
2. Blob methods.

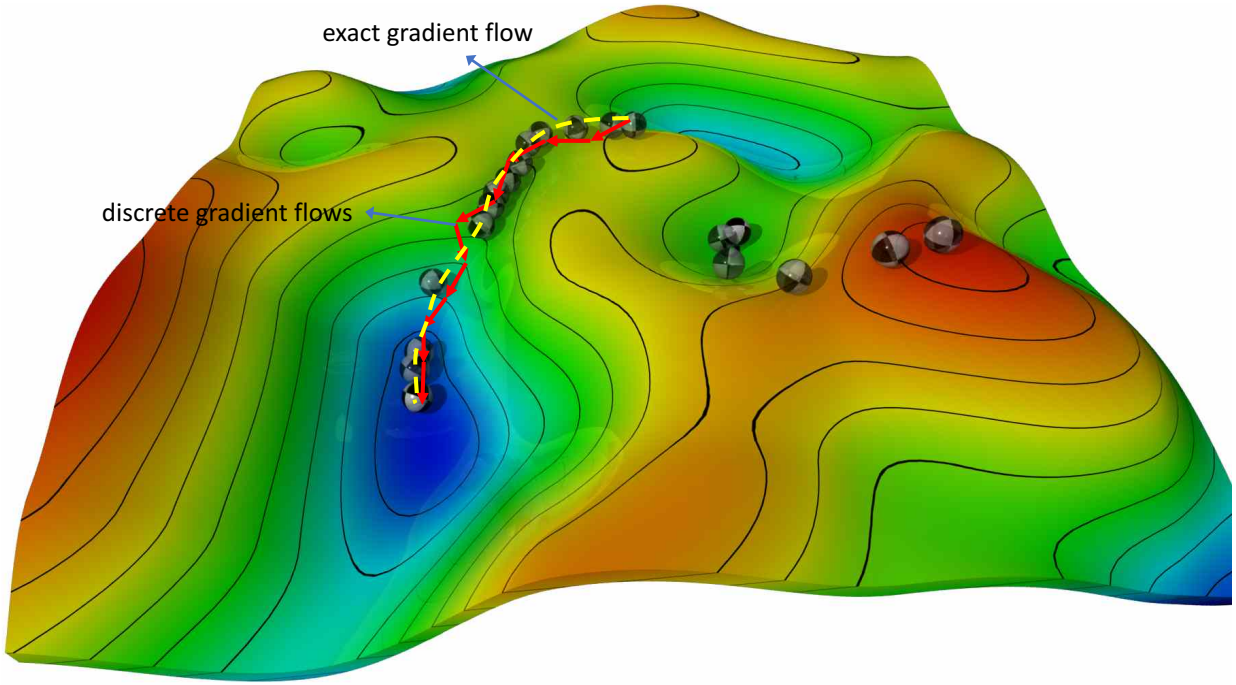
Discrete Gradient Flows

- Discretize the continuous-time PDE, *i.e.*, approximating q_t by $q_k^{(h)}$ obtained from an optimization problem, where h is the stepsize, and $t = kh$.



Discrete Gradient Flows

- Discretize the continuous-time PDE, *i.e.*, approximating q_t by $q_k^{(h)}$ obtained from an optimization problem, where h is the stepsize, and $t = kh$.



- Each intermediate solution is obtained via Minimizing Movement Scheme:

$$q_{k+1}^{(h)} = \arg\min_q F(q) + d_W^2(q, q_k^{(h)})/2h$$

Explanation of Discrete Gradient Flows

$$q_{k+1}^{(h)} = \arg\min_{q} F(q) + d_W^2(q, q_k^{(h)})/2h \tag{1}$$

Explanation of Discrete Gradient Flows

$$q_{k+1}^{(h)} = \arg\min_{q} F(q) + d_W^2(q, q_k^{(h)})/2h \tag{1}$$

- Consider the Euclidean case, where $q_k^{(h)}$ is replaced with a finite-dimension vector $x_k^{(h)}$,
  - d_W^2 corresponds to the Euclidean distance in Euclidean space.
- Iterative optimization in Eq. (1) becomes

$$x_{k+1}^{(h)} = \arg\min_{x} F(x) + \left\| x - x_k^{(h)} \right\|^2 /2h$$

$$\Rightarrow x_{k+1}^{(h)} = x_k^{(h)} - h\nabla_x F(x)$$

Gradient descent!

Explanation of Discrete Gradient Flows

$$q_{k+1}^{(h)} = \arg\min_{q} F(q) + d_W^2(q, q_k^{(h)})/2h \tag{1}$$

- Consider the Euclidean case, where $q_k^{(h)}$ is replaced with a finite-dimension vector $x_k^{(h)}$,
  - d_W^2 corresponds to the Euclidean distance in Euclidean space.
- Iterative optimization in Eq. (1) becomes

$$x_{k+1}^{(h)} = \arg\min_{x} F(x) + \left\| x - x_k^{(h)} \right\|^2 /2h$$

$$\Rightarrow x_{k+1}^{(h)} = x_k^{(h)} - h\nabla_x F(x)$$

Gradient descent!

Discrete gradient flows are gradient descent in the space of probability measures!

Numerical Solution for Discrete Gradient Flows

$$q_{k+1}^{(h)} = \arg\min_{q} \underbrace{F(q) + d_W^2(q, q_k^{(h)})/2h}_{E}$$

- Still infeasible to solve since $q_k^{(h)}$ are **infinite-dimensional**.

Numerical Solution for Discrete Gradient Flows

$$q_{k+1}^{(h)} = \arg\min_{q} \underbrace{F(q) + d_W^2(q, q_k^{(h)})/2h}_{E}$$

- Still infeasible to solve since $q_k^{(h)}$ are **infinite-dimensional**.

Particle approximation

- Approximate $q_k^{(h)}$ as $q_k^{(h)} \approx \frac{1}{M}\sum_{i=1}^{M} \delta_{x_k^{(i)}}$.
- Solving $q_k^{(h)}$ is equivalent to solving $x_k^{(i)}$'s.
- Update $x_k^{(i)}$ by gradient descent:

$$x_{k+1}^{(i)} = x_k^{(i)} - h \frac{\partial E}{\partial x}\Big|_{x_k^{(i)}}$$

Blob Methods

$$\partial_t q_t = \nabla \cdot (q_t \underbrace{\nabla(\frac{\delta F}{\delta q_t}(q_t))}_{-v^{\mathsf{GF}}}) \tag{2}$$

- Directly solve the original WGF with particle approximation:

Theorem 1

When approximating q_t with particles, Eq. (2) is reduced to solving

$$\mathrm{d}x_t^{(i)} = -v^{\mathsf{Blob}}(\{x_t^{(j)}\}_j)\mathrm{d}t \tag{3}$$

- Directly use numerical method to solve Eq. (3):

$$x_{k+1}^{(i)} = x_k^{(i)} - hv^{\mathsf{Blob}}(\{x_k^{(j)}\}_j)\Big|_{x_k^{(i)}}$$

SVGD as Wasserstein Gradient Flow

Following some manifold argument [10], v^{GF} can be reformulated as:

$$v^{\mathsf{GF}} = \max \cdot \operatorname*{argmax}_{v \in \mathcal{L}_q^2, \|v\|_{\mathcal{L}_q^2}=1} \left\langle v^{\mathsf{GF}}, v \right\rangle_{\mathcal{L}_q^2}. \tag{4}$$

We find:

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

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

- $\mathcal{H}^D$ 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.

Recap of the Unifying ParVIs

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

- Gradient flow on $\mathcal{P}_2(\mathcal{X})$ for energy functional $\mathsf{KL}_p(q) := \mathbb{E}_q[\log(q/p)]$:
  - Approximation of vector-field v^{GF} ([17], Thm 23.18; [1], Example 11.1.2):

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

  - Minimizing Movement Scheme (MMS) ([1], Def. 2.0.6):

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

- 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 $\mathcal{P}_2(\mathcal{X})$ [7].

ParVIs are numerical solutions to Wasserstein gradient flows.

Particle-Based Variational Inference Methods (ParVIs)

- Stein Variational Gradient Descent (SVGD) [12]:

$$v^{\text{SVGD}}(\cdot) := \max_{v \in \mathcal{H}^D, \|v\|_{\mathcal{H}^D}=1} \cdot \operatorname{argmax} -\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathsf{KL}_p\big((\mathrm{id} + \varepsilon v)_{\#} q\big)\Big|_{\varepsilon=0}$$

$$= \mathbb{E}_{q(x)}[K(x,\cdot)\nabla \log p(x) + \nabla_x K(x,\cdot)],$$

  where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) of kernel K.
  - v^{SVGD} is the vector field of the gradient flow of KL_p on a kernel-related distribution manifold $\mathcal{P}_{\mathcal{H}}$ [11].
- Blob method (w-SGLD-B) [2]:

$$v^{\text{Blob}} := -\nabla\big(\frac{\delta}{\delta q}\mathbb{E}_q[\log(\tilde{q}/p)]\big)$$

$$= \nabla \log p - \nabla \log \tilde{q} - \nabla\big((q/\tilde{q}) * K\big), \quad \tilde{q} := q * K.$$

- GFSD method [10]:

$$v^{\text{GFSD}} := \nabla \log p - \nabla \log \tilde{q}, \quad \tilde{q} := q * K.$$

Particle-Based Variational Inference Methods (ParVIs)

- Stein Variational Gradient Descent (SVGD) [12]:

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

  where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) of kernel K.
  - v^{SVGD} is the vector field of the gradient flow of KL_p on a kernel-related distribution manifold $\mathcal{P}_{\mathcal{H}}$ [11].
- GFSF method [10]:

$$v^{\text{GFSF}} := \nabla \log p + \operatorname*{argmin}_{u\in\mathcal{L}^2} \max_{\substack{\phi\in\mathcal{H}^D,\\ \|\phi\|_{\mathcal{H}^D}=1}} \big(\mathbb{E}_q[\phi\cdot u - \nabla\cdot\phi]\big)^2.$$

  Solution: $\hat{v}^{\text{GFSF}} = \hat{g} + \hat{K}'\hat{K}^{-1}$. (Note $\hat{v}^{\text{SVGD}} = \hat{v}^{\text{GFSF}}\hat{K}$.)

  $\hat{g}_{:,i} = \nabla_{x^{(i)}} \log p(x^{(i)})$, $\hat{K}_{ij} = K(x^{(i)}, x^{(j)})$, $\hat{K}'_{:,i} = \sum_j \nabla_{x^{(j)}} K(x^{(j)}, x^{(i)})$.

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

Smoothing Functions

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

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

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 $\mathcal{G} := \overline{\{\phi * K : \phi \in \mathcal{C}^\infty_c\}}^{\mathcal{L}^2_q}$.

$\overline{\mathcal{C}^\infty_c}^{\mathcal{L}^2_q} = \mathcal{L}^2_q$ ([9], Thm. 2.11) $\Longrightarrow \mathcal{G}$ is roughly the kernel-smoothed $\mathcal{L}^2_q$.

- GFSF smoothed functions in a similar way as SVGD:

$$v^{\text{GFSF}} := \nabla \log p + \operatorname*{argmin}_{u \in \mathcal{L}^2} \max_{\substack{\phi \in \mathcal{H}^D, \\ \|\phi\|_{\mathcal{H}^D} = 1}} \left(\mathbb{E}_q[\phi \cdot u - \nabla \cdot \phi]\right)^2.$$

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

Smoothing the Density

- Blob [2] partially smooths the density.

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

- GFSD [10] fully smooths the density.

$$v^{\text{GF}} := \nabla \log p - \nabla \log q \Longrightarrow v^{\text{GFSD}} := \nabla \log p - \nabla \log \tilde{q}.$$

- DGF [21] adds an entropy regularizer in the primal objective function, encourage smoothing the density q.

Remark 4

Existing ParVI methods approximate Wasserstein Gradient flow by smoothing the **density** *or* **functions**.

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: $\operatorname{grad} \mathsf{KL}_p(q)$ undefined at $q = \hat{q} := \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}$.

Theorem 5 (Necessity of smoothing for SVGD)

For $q = \hat{q}$ and $v \in \mathcal{L}_p^2$, problem (4):

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

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!

Non-Asymptotic Convergence Theory [20]

- SVGD evolves following the ODE:

$$\mathrm{d}x_t^{(i)} = \frac{1}{M}\sum_{j=1}^{M}\left[K(x_t^{(j)}, x_t^{(i)})\log p(x_t^{(i)}) + \nabla_{x_t^{(j)}} K(x_t^{(j)}, x_t^{(i)})\right]\mathrm{d}t, \quad \text{for } \forall i$$

- Would the above ODE system converge? $\Rightarrow$ not really!

Theorem 6 (Pitfall of SVGD)

Define the expected particle distance (EPD) as: $\mathrm{EPD} \triangleq \sqrt{\sum_{i,j}^{M} \mathbb{E}\|x_t^{(i)} - x_t^{(j)}\|^2}$.
Under some assumptions, the EPD of SVGD is bounded as: $\mathrm{EPD} \leq C_0 e^{-2\lambda t}$,
where $C_0 = \sqrt{\sum_{i,j}^{M} \|x_0^{(i)} - x_0^{(j)}\|^2}$ *and some positive constant* λ.

Without considering numerical errors, the theorem implies particles in **SVGD would collapse under some circumstance!**

Non-Asymptotic Convergence Theory

- We propose a remedy variant by combining SGLD:

$$\mathrm{d}x_t^{(i)} = \left(\frac{1}{\beta} \log p(x_t^{(i)}) + \frac{1}{M} \sum_{j=1}^{M} K(x_t^{(i)} - x_t^{(j)}) \log p(x_t^{(j)}) + \frac{1}{M} \sum_{j=1}^{M} \nabla K(x_t^{(i)} - x_t^{(j)}) \right) \mathrm{d}t + \sqrt{2\beta^{-1}} \mathrm{d}\mathcal{W}_t^{(i)}$$

- Called stochastic particle optimization sampling (SPOS).

Theorem 7

Under certain assumptions, the EPD of SPOS is bounded as:
$\mathrm{EPD} \leq C_1 e^{-2\lambda t} + 4\sqrt{\frac{d}{\beta}\frac{M}{\lambda}}$, *for some positive constants* C_1 *and* λ.

The EPD of SPOS would not collapse to zero.

Non-asymptotic Convergence Bounds of SPOS

- Let q_T be the probability law of the particles at iteration T, we measure by $W_1(q_T, p)$

Theorem 8 (Fixed Stepsize (Informal))

Under certain assumptions, with a fixed stepsize h, $W_1(q_T, p)$ is bounded:

$$W_1(q_T, p) = O\left(\frac{1}{\sqrt{M}} + \exp\{-Th\} + Md^{\frac{3}{2}}T^{\frac{1}{2}}h^{\frac{1}{2}}\right) .$$

Theorem 9 (Decreasing Stepsize (Informal))

Denote $\tilde{h}_T \triangleq \sum_{k=0}^{T-1} h_k$. Under certain assumptions, if we set $h_k = h_0/(k+1)$ and $B_k = B_0 + [\log(k+1)]^{100/99}$, $W_1(q_T, p)$ is bounded

$$W_1(q_T, p) = O\left(\frac{1}{\sqrt{M}} + \exp\{-\tilde{h}_T\} + Md^{\frac{3}{2}}h_0\right) .$$

Larger particle number does NOT necessarily lead to **smaller errors**, due to limited computational budget and numerical errors!

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 other smoothing functions.

Toy Experiments: Bandwidth Selection

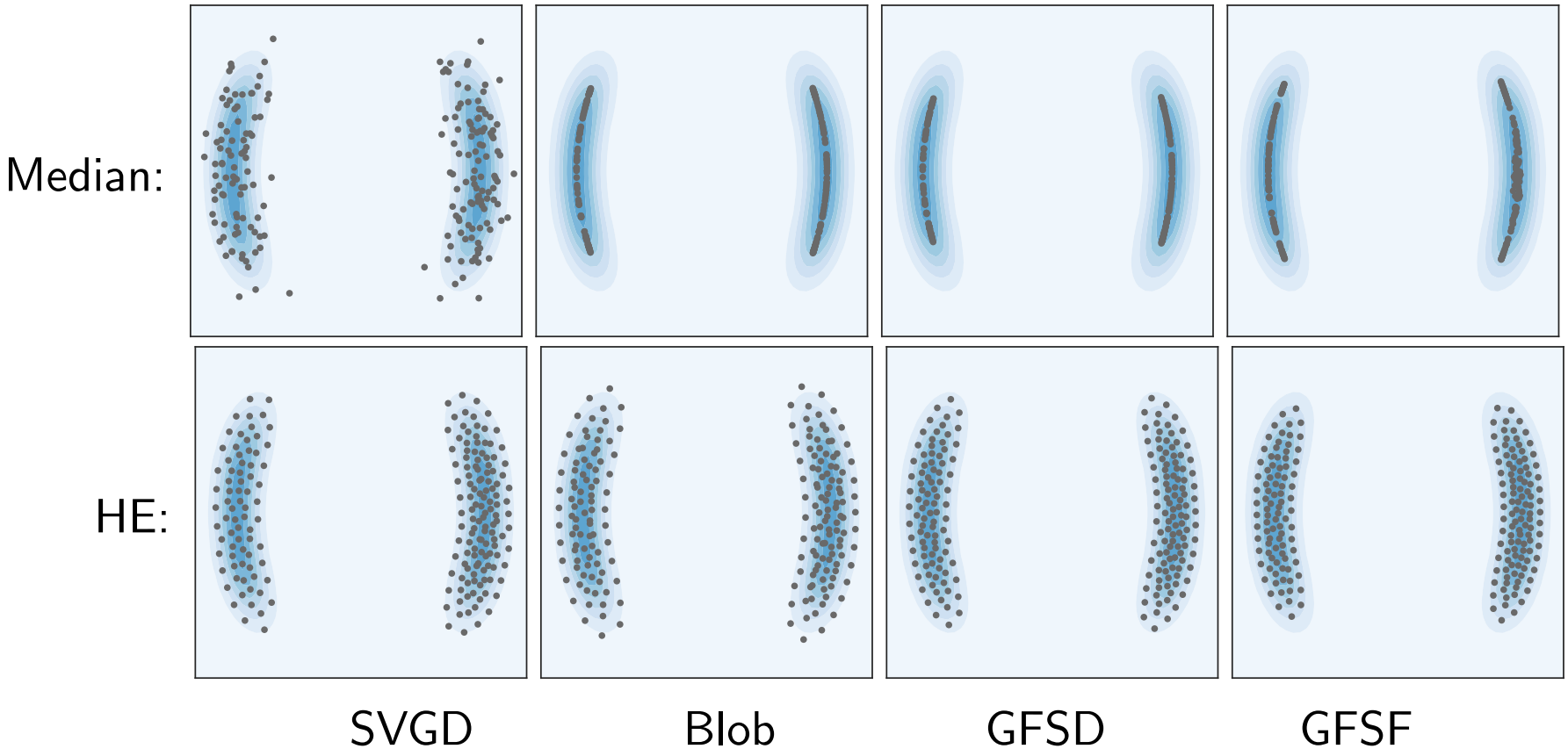
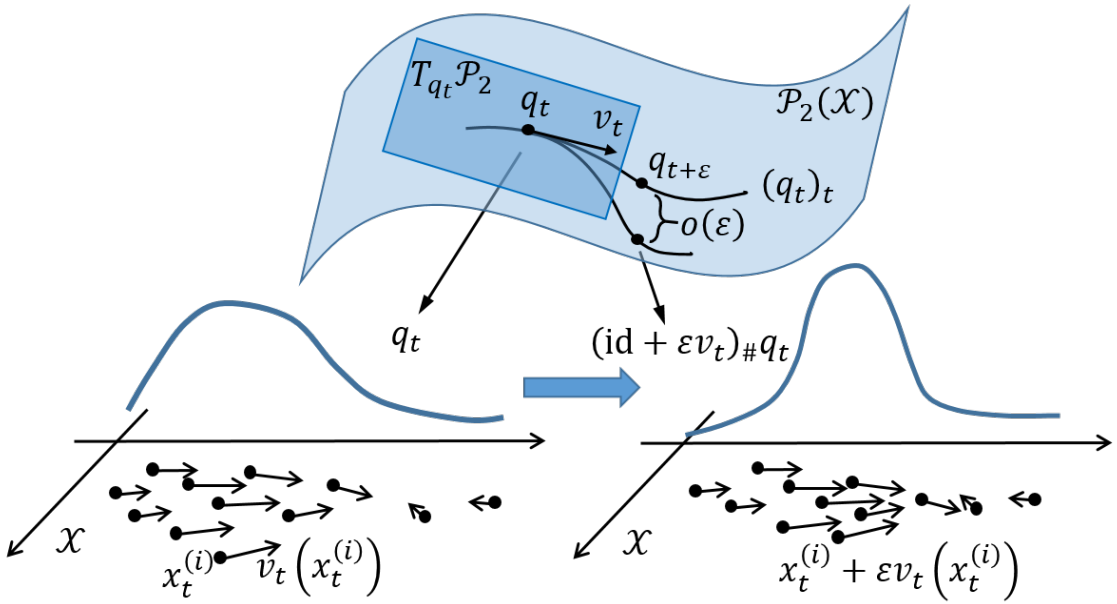


Figure: Comparison of HE (bottom row) with the median method (top row) for bandwidth selection.

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

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

- $\mathcal{P}_2$ as a Riemannian manifold [16, 17, 1] ($\mathcal{X} = \mathbb{R}^D$):



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

- Recap: ParVIs are numerical solution of the Wasserstein Gradient Flows.

Nesterov's Acceleration Method on Riemannian Manifolds

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

- Riemannian Accelerated Gradient (RAG) [13] (with simplification):

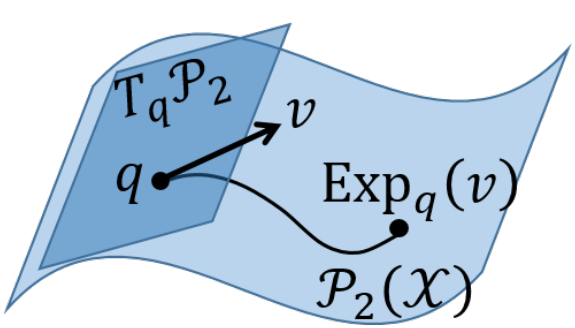
$$\begin{cases} q_k = \operatorname{Exp}_{r_{k-1}}(\varepsilon v_{k-1}), \\ r_k = \operatorname{Exp}_{q_k}\left[-\Gamma_{r_{k-1}}^{q_k}\left(\frac{k-1}{k}\operatorname{Exp}_{r_{k-1}}^{-1}(q_{k-1}) - \frac{k+\alpha-2}{k}\varepsilon v_{k-1}\right)\right]. \end{cases}$$

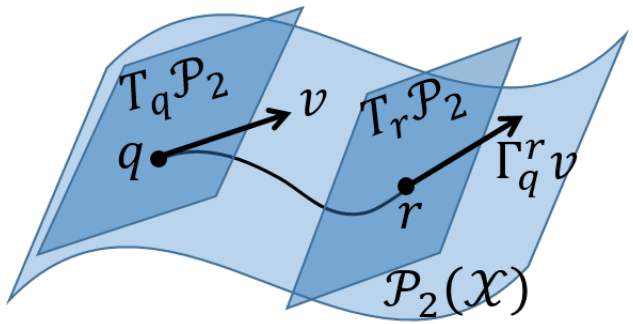
- Riemannian Nesterov's method (RNes) [19] (with simplification):

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

Required:

- Exponential map $\operatorname{Exp}_q : T_q\mathcal{P}_2(\mathcal{X}) \to \mathcal{P}_2(\mathcal{X})$ 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 ([17], Coro. 7.22; [1], Prop. 8.4.6; [5], 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 [3, 18] 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 10 (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 [14, 15].
  - Use Schild's ladder method [4, 8] for approximation.

Proposition 11 (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/DGF/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 $\{x_{k_{\max}}^{(i)}\}_{i=1}^N$.

Bayesian Logistic Regression (BLR)

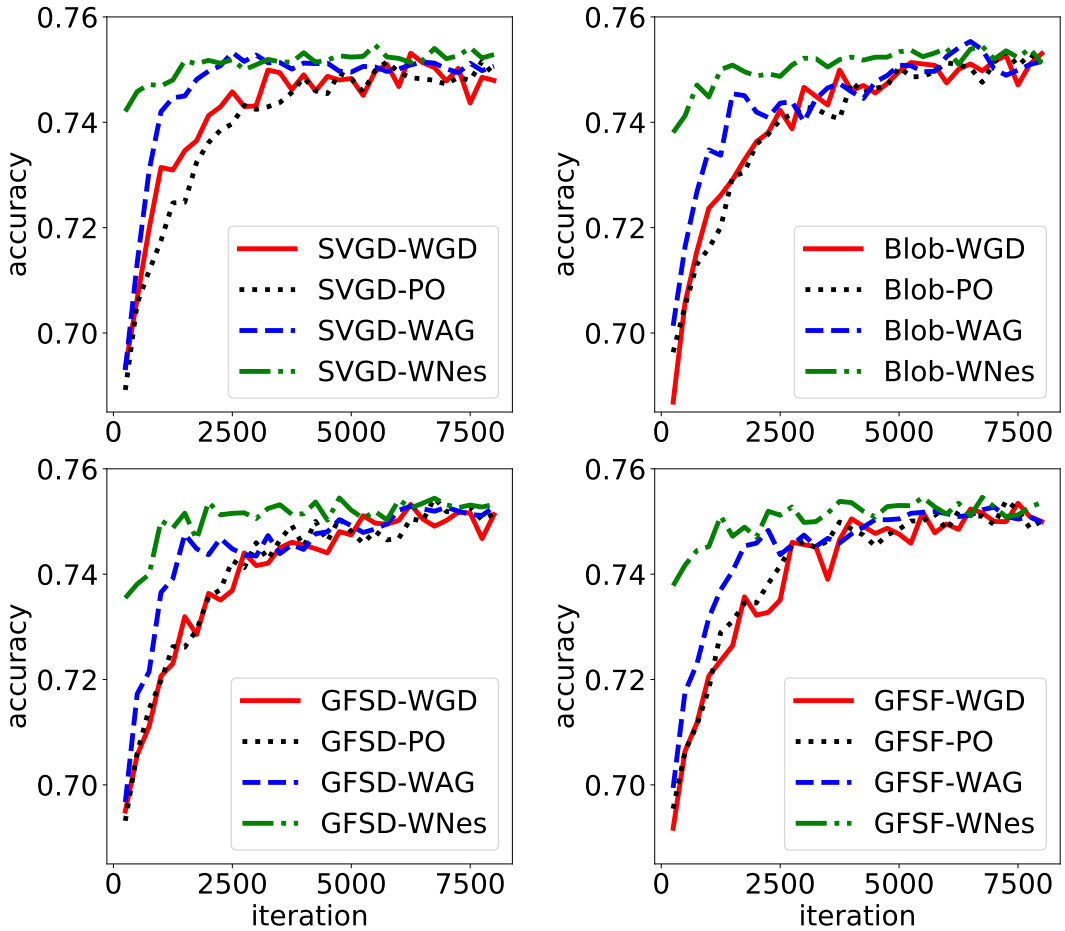


Figure: Acceleration effect of WAG and WNes on BLR on the Covertype dataset, measured by prediction accuracy on test dataset. Each curve is averaged over 10 runs.

Latent Dirichlet Allocation (LDA)

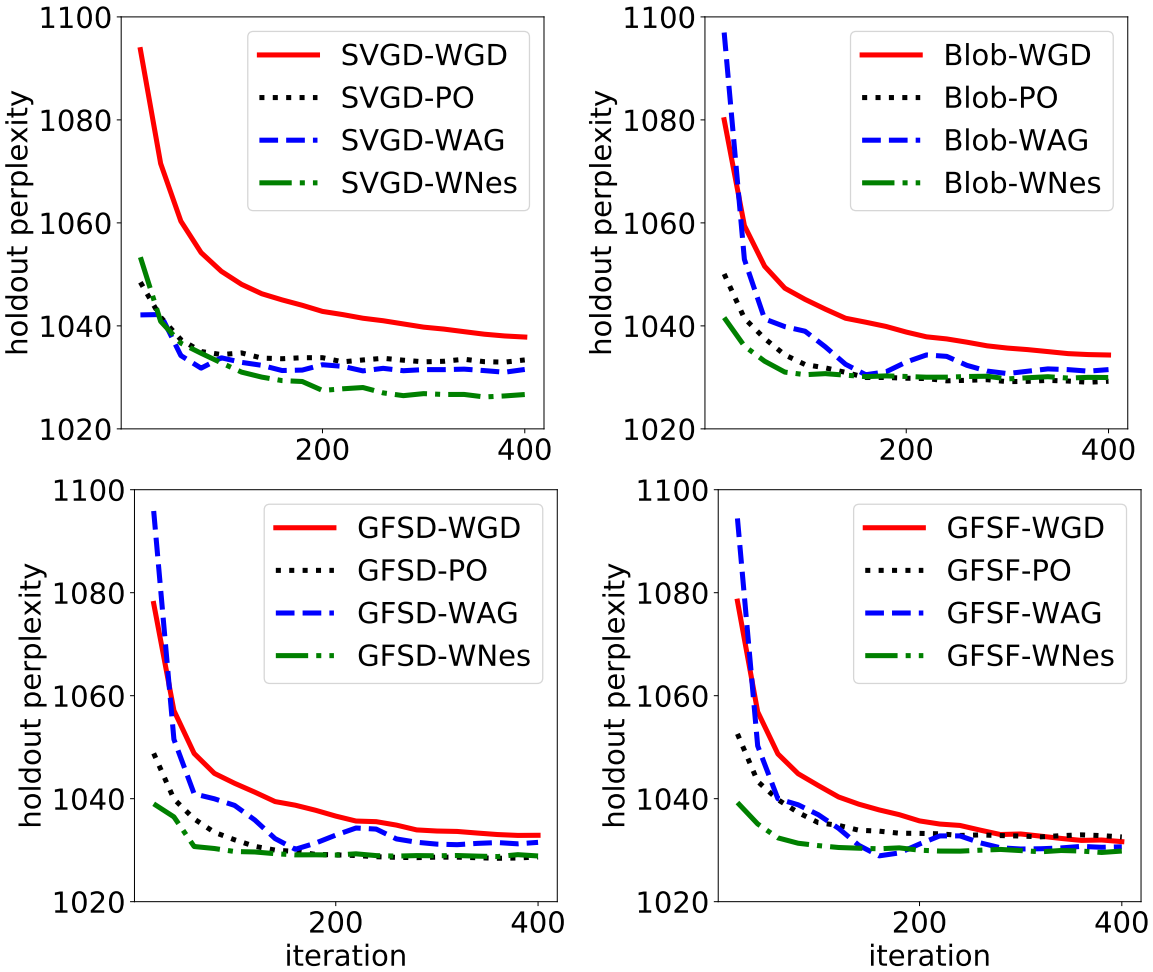


Figure: Acceleration effect of WAG and WNes on LDA. Inference results are measured by the hold-out perplexity. Curves are averaged over 10 runs.

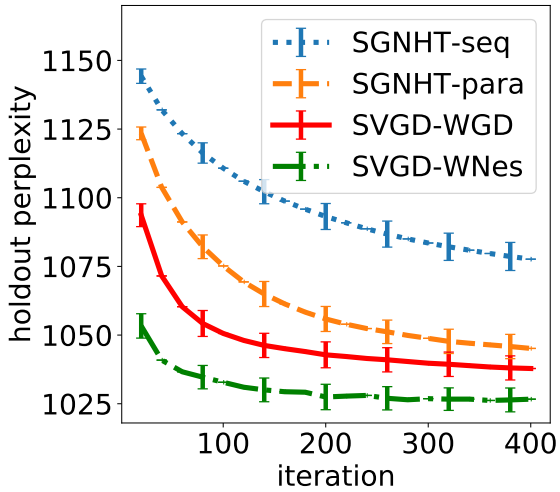


Figure: Comparison of SVGD and SGNHT on LDA, as representatives of ParVIs and MCMCs. Average over 10 runs.

Application I: Thompson Sampling

- Given past observations $\mathcal{D} = \{d_i\}_{i=1}^t \triangleq \{(\mathbf{x}_i, \boldsymbol{a}_i, r_i)\}_{i=1}^t$, model prior $p(x)$, model (likelihood) $p(\mathcal{D}|x) = \prod p(\boldsymbol{d}_i|x)$ on *i.i.d.* assumption.
- In Thompson sampling (TS), we want to sample from the posterior:

$$p(x|\mathcal{D}) \propto p(x)p(\mathcal{D}|x) = p(x)\prod_{i=1}^{t} p(\boldsymbol{d}_i|x) \ .$$

- We employ ParVIs to approximate the intractable posterior: particle-interactive Thompson sampling [22].

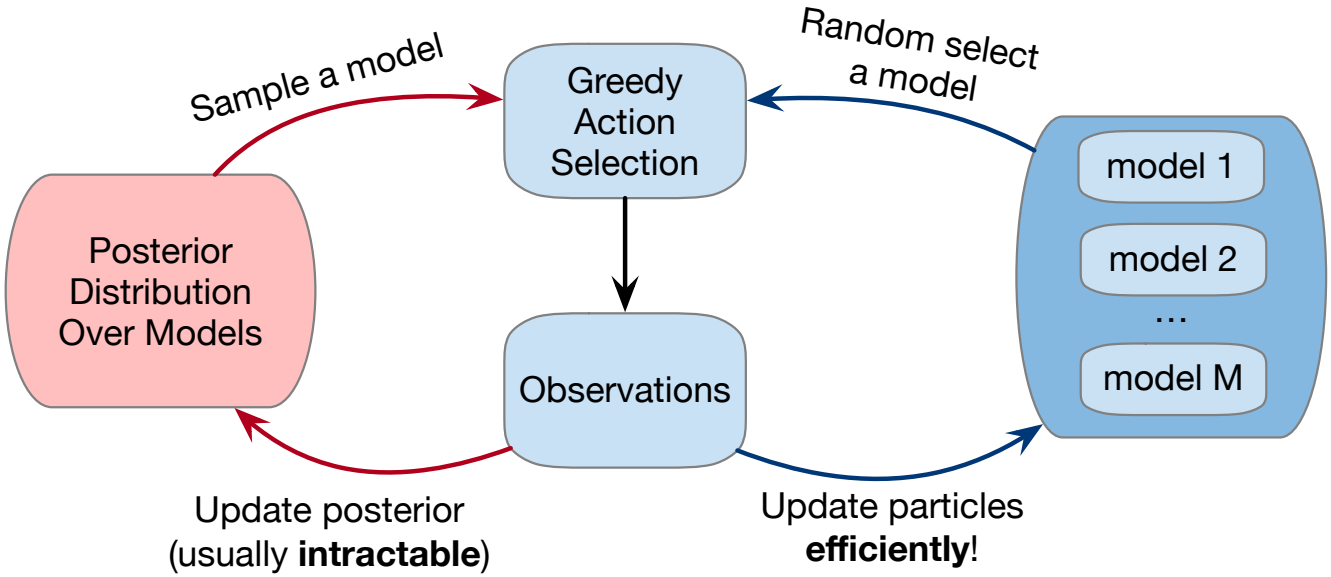


Figure: vanilla TS v.s. particle-interactive TS

Experiments on Thompson Sampling

Methods:

- Linear TS: not scalable and poor expressive power.
- Neural Linear: performs linear TS on extracted features.
- VI-TS (Gaussian): underestimate uncertainty leads to high variances.
- π-TS : particle-based variational inference.

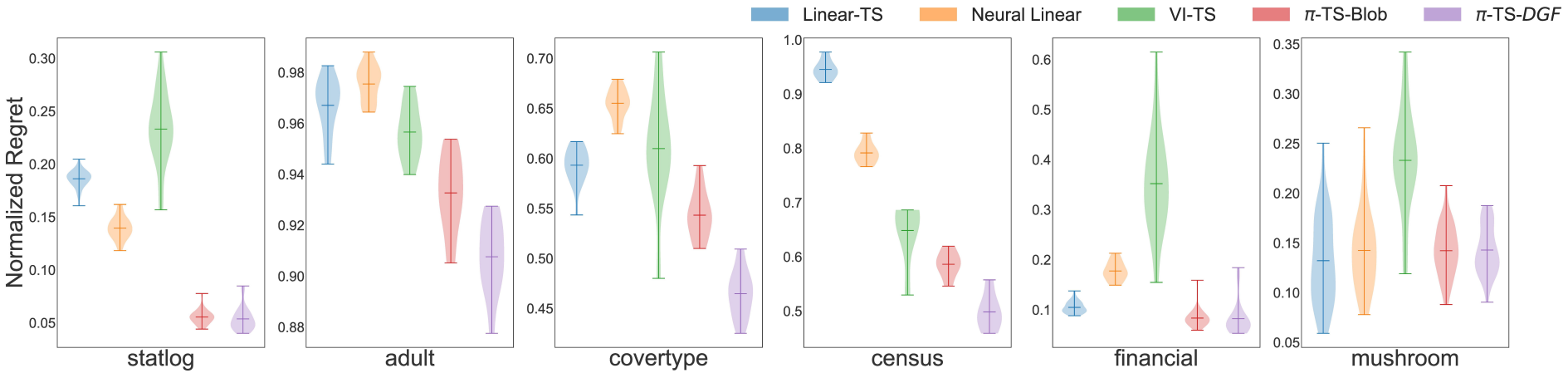


Figure: Normalized Regret comparison on real public datasets.

Application II: Reinforcement Learning (Soft-Q Learning)

1. Q-function update [6]:

$$Q(\boldsymbol{a}_t, \boldsymbol{s}_t) = r(\boldsymbol{a}_t, \boldsymbol{s}_t) + \gamma \mathbb{E}_{\boldsymbol{s}_{t+1} \sim \rho_\pi}[V_\pi(\boldsymbol{s}_{t+1}) - \alpha \mathcal{H}(\pi(\cdot|\boldsymbol{s}_{t+1}))]$$

where $V_\pi(\boldsymbol{s}_{t+1}) \triangleq \log \int_{\mathcal{A}} \exp(Q(\boldsymbol{a}, \boldsymbol{s}_{t+1}))\mathrm{d}\boldsymbol{a}$.

$$\pi^*(\boldsymbol{a}_t|\boldsymbol{s}_t) = \arg\max_\pi \sum_t \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_t) \sim \rho_\pi} [r(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\boldsymbol{s}_t))]$$

2. Policy Optimization:
   - Approximate the policy $\pi^*(\cdot|\boldsymbol{s}) \triangleq p_{s,\pi} \propto \exp(Q(\cdot, \boldsymbol{s}))$ via ParVIs
3. Interact with the environment, collect more data (repeat 1-3).

- The policy distribution π is a sampling network: $\boldsymbol{a}_t \sim \pi^\phi(\cdot|\boldsymbol{s}_t)$, and the policy optimization as WGFs (DP-WGF) is:

$$\pi^{\boldsymbol{\phi}}_{k+1} = \arg\min_{\pi^{\boldsymbol{\phi}}} \left\{ \mathrm{KL}\left(\pi^{\boldsymbol{\phi}} \| p_{s,\pi}\right) + \frac{d_w^2(\pi^{\boldsymbol{\phi}}, \pi^{\boldsymbol{\phi}}_k)}{2\varepsilon} \right\} .$$

DP-WGF can be regarded as policy gradient with **Wasserstein trust-region** [21].

Experiments on Reinforcement Learning

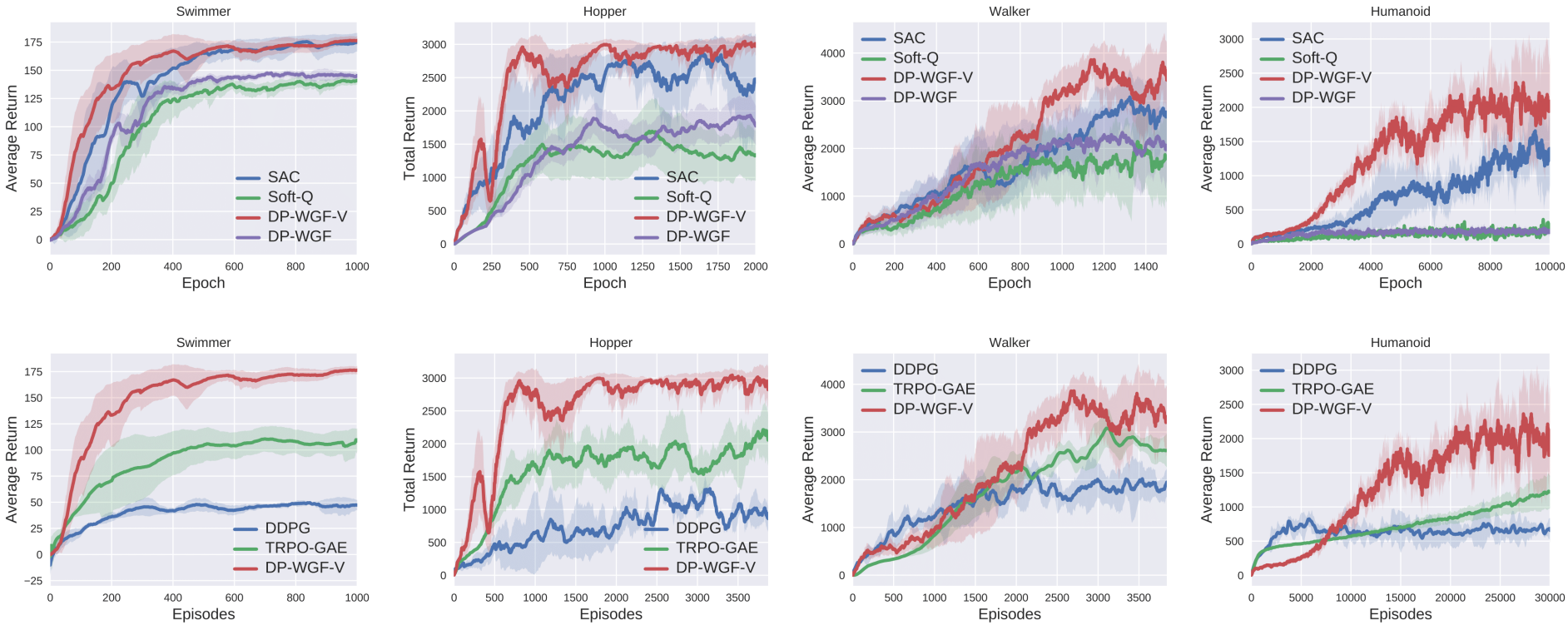


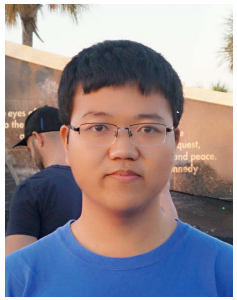
Figure: Average return in MuJoCo tasks by Soft-Q, SAC and DP-WGF-V (first row), and by DDPG, TRPO-GAE and DP-WGF-V (second row).

		WGF-DP-V		SAC		TRPO-GAE		DDPG	
Domain	Threshold	MaxReturn.	Episodes	MaxReturn	Epsisodes	MaxReturn	Episodes	MaxReturn	Episodes
Swimmer	100	**181.60**	**76**	180.83	112	110.58	433	49.57	N/A
Walker	3000	**4978.59**	**2289**	4255.05	2388	3497.81	3020	2138.42	N/A
Hopper	2000	**3248.76**	**678**	3146.51	736	2604	1749	1317	N/A
Humanoid	2000	3077.84	**18740**	2212.51	26476	**5411.15**	32261	2230.60	34652

Table: Average return by TRPO-GAE, SAC, DDPG and DP-WGF-V

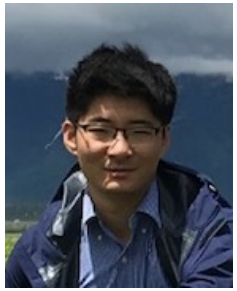
Summary

- ParVIs are numerical solutions to Wasserstein gradient flows (Unifying).
- ParVIs rely on smoothing: either the density or functions (Understanding).
- ParVIs can be accelerated via leveraging the Riemannian Structure.
- Variants of ParVIs: GFSF, GFSD [10], DGF [21], Blob [2], etc.
- Outperform existing methods on a number of applications.



Ruiyi Zhang

Changyou Chen



Chang Liu

Lawrence Carin

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.

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.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine.

Reinforcement learning with deep energy-based policies.

arXiv preprint arXiv:1702.08165, 2017.


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.


Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu, and Lawrence Carin.
Accelerated first-order methods on the wasserstein space for bayesian inference.
In *ICML*, 2019.


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.


Jianyi Zhang, Ruiyi Zhang, and Changyou Chen.
Stochastic particle-optimization sampling and the non-asymptotic convergence theory.
arXiv preprint arXiv:1809.01293, 2018.


Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin Duke.
Policy optimization as wasserstein gradient flows.
In *ICML*, 2018.


Ruiyi Zhang, Zheng Wen, Changyou Chen, and Lawrence Carin.
Scalable thompson sampling via optimal transport.
In *AISTATS*, 2019.