Understanding and Accelerating Particle-Based Variational Inference

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu, Lawrence Carin
chang-li14@mails.tsinghua.edu.cn

Contributions

In theory:

- Particle-based Variational Inference methods (ParVIs), *e.g.*, Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016), approximate the Wasserstein gradient flow by a compulsory smoothing assumption.
- ParVIs either smooth the density or smooth functions, and they are equivalent.

In practice:

- The smoothing theory inspires two new ParVIs and a bandwidth selection method.
- The gradient flow perspective inspires an acceleration framework for all ParVIs.

New ParVIs

- GF with smoothed density (GFSD):

$v^{\text{GF}} = \nabla \log p - \nabla \log q$

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

- GF with smoothed function (GFSF):

$v^{\text{GF}} = \nabla \log p + \operatorname{argmin}_{u \in \mathcal{L}^2} \max_{\phi \in \mathcal{C}_c^\infty, \|\phi\|_{\mathcal{L}_q^2}=1} \left(\mathbb{E}_q[\phi \cdot u - \nabla \cdot \phi]\right)^2$

$\Longrightarrow$

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

$\hat{v}^{\text{GFSF}} = \hat{g} + \hat{K}' \hat{K}^{-1}$.

Bandwidth Selection

Dynamics $\mathrm{d}x = -\nabla \log q_t(x)\,\mathrm{d}t$ produces q_t obeying $\partial_t q_t(x) = \Delta q_t(x)$ (Heat Equation).

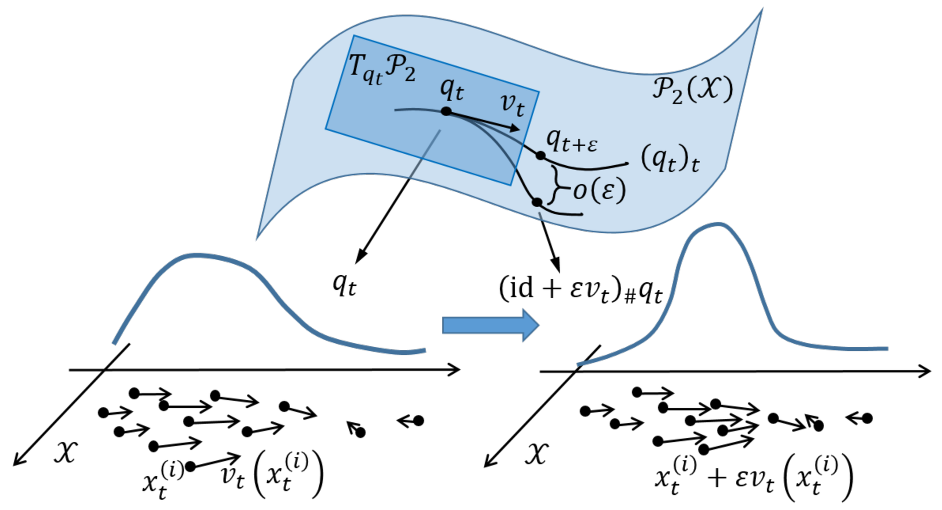
Approximate $q_t(x)$ by kernel smoothed density $\tilde{q}_h(x; \{x^{(i)}\}_{i=1}^N)$ with particles (h: bandwidth). Then:

- $q_{t+\varepsilon}(x) \approx \tilde{q}(x) + \varepsilon \Delta \tilde{q}(x)$ (HE).
- $q_{t+\varepsilon}(x) \approx \tilde{q}(x; \{x^{(i)} - \varepsilon \nabla \log \tilde{q}(x^{(i)})\}_{i=1}^N)$ (Dynamics on the particles).

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

(dimensionless)

Preliminary



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

Tangent vector v on $\mathcal{P}_2(\mathcal{X}) \Leftrightarrow$ vector field X on $\mathcal{X}$.

Gradient Flow (GF) of $\text{KL}_p(q)$: $v^{\text{GF}} = \nabla \log p - \nabla \log q$.

- ParVIs
  - $v^{\text{SVGD}}(\cdot) := \mathbb{E}_{q(x)}[K(x, \cdot) \nabla \log p(x) + \nabla_x K(x, \cdot)]$, $\mathcal{H}$: the RKHS of a kernel.
  - Blob (Chen *et al.*, 2018) $v^{\text{Blob}} := \nabla \log p - \nabla \log \tilde{q} - \nabla\big((q/\tilde{q}) * K\big)$, $\tilde{q} := q * K$ (convolution).

ParVIs Approximate Wass. Grad. Flow by Smoothing

$\mathcal{L}_q^2$ / $\mathcal{C}_c^\infty$: integr. / comp.-supp smth vec-val func.

- SVGD approximates Wass. grad. flow.

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

Note $v^{\text{GF}} = \max \cdot \operatorname{argmax}_{v \in \mathcal{L}_q^2, \|v\|_{\mathcal{L}_q^2}=1} \langle v^{\text{GF}}, v \rangle_{\mathcal{L}_q^2}$.

- Smoothing functions

Theorem 3. For Gaussian kernel K and abs. cont. q, $\mathcal{H}^D$ is isometrically isomorphic to $\mathcal{G} := \overline{\{\phi * K : \phi \in \mathcal{C}_c^\infty\}}^{\mathcal{L}_q^2}$.

Note $\overline{\mathcal{C}_c^\infty}^{\mathcal{L}_q^2} = \mathcal{L}_q^2$: $\mathcal{H}^D$ smooths $\mathcal{L}_q^2$.

- Smoothing the density

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

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

- Equivalence: for obj. in smth. fun. $\mathbb{E}_q[L(v)]$,

$\mathbb{E}_{\tilde{q}}[L(v)] = \mathbb{E}_{q*K}[L(v)]$ (smth. dens.)

$= \mathbb{E}_q[L(v) * K] = \mathbb{E}_q[L(v * K)]$ (smth. func.)

- Necessity: well-definedness of v^{GF}.

Theorem 4. For $q = \hat{q}$ and $v \in \mathcal{L}_p^2$, opt. problem for SVGD has no opt. solution. SVGD transfers the assmp. on q to func.

Accelerated First-Order Methods on Wass. Space

Applying Riemannian Accelerated Gradient (Liu *et al.*, 2017) and Riemannian Nesterov's method (Zhang & Sra, 2018) to $\mathcal{P}_2(\mathcal{X})$: requires exponential map and parallel transport.

- Exp. map (Villani, 2008): $\text{Exp}_q(v) = (\text{id} + v)_\# q$.
- Inverse exp. map: For pairwise close samples $\{x^{(i)}\}_i$ of q and $\{y^{(i)}\}_i$ of r, $\big(\text{Exp}_q^{-1}(r)\big)(x^{(i)}) \approx y^{(i)} - x^{(i)}$.
- Parallel transport: For pairwise close samples, $\big(\Gamma_q^r(v)\big)(y^{(i)}) \approx v(x^{(i)})$.

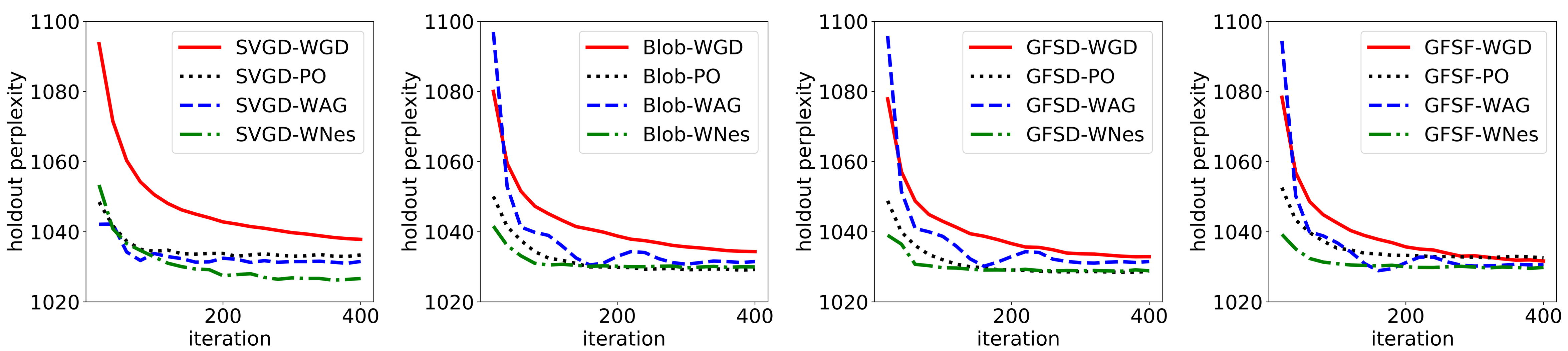
Alg. 1: Wasserstein Accelerate Gradient / Wasserstein Nesterov's method.

Use aux. distr. r with ptcls. $\{y^{(i)}\}_i$. In iter k:

- Find $v(y_{k-1}^{(i)})$ by SVGD/Blob/GFSD/GFSF;
- $x_k^{(i)} = y_{k-1}^{(i)} + \varepsilon v(y_{k-1}^{(i)})$;
- $y_k^{(i)} = x_k^{(i)} + \begin{cases} \text{WAG: } \frac{k-1}{k}(y_{k-1}^{(i)} - x_{k-1}^{(i)}) + \frac{k+\alpha-2}{k}\varepsilon v(y_{k-1}^{(i)}); \\ \text{WNes: } c_1(c_2 - 1)(x_k^{(i)} - x_{k-1}^{(i)}); \end{cases}$

Pairwise-close condition holds.

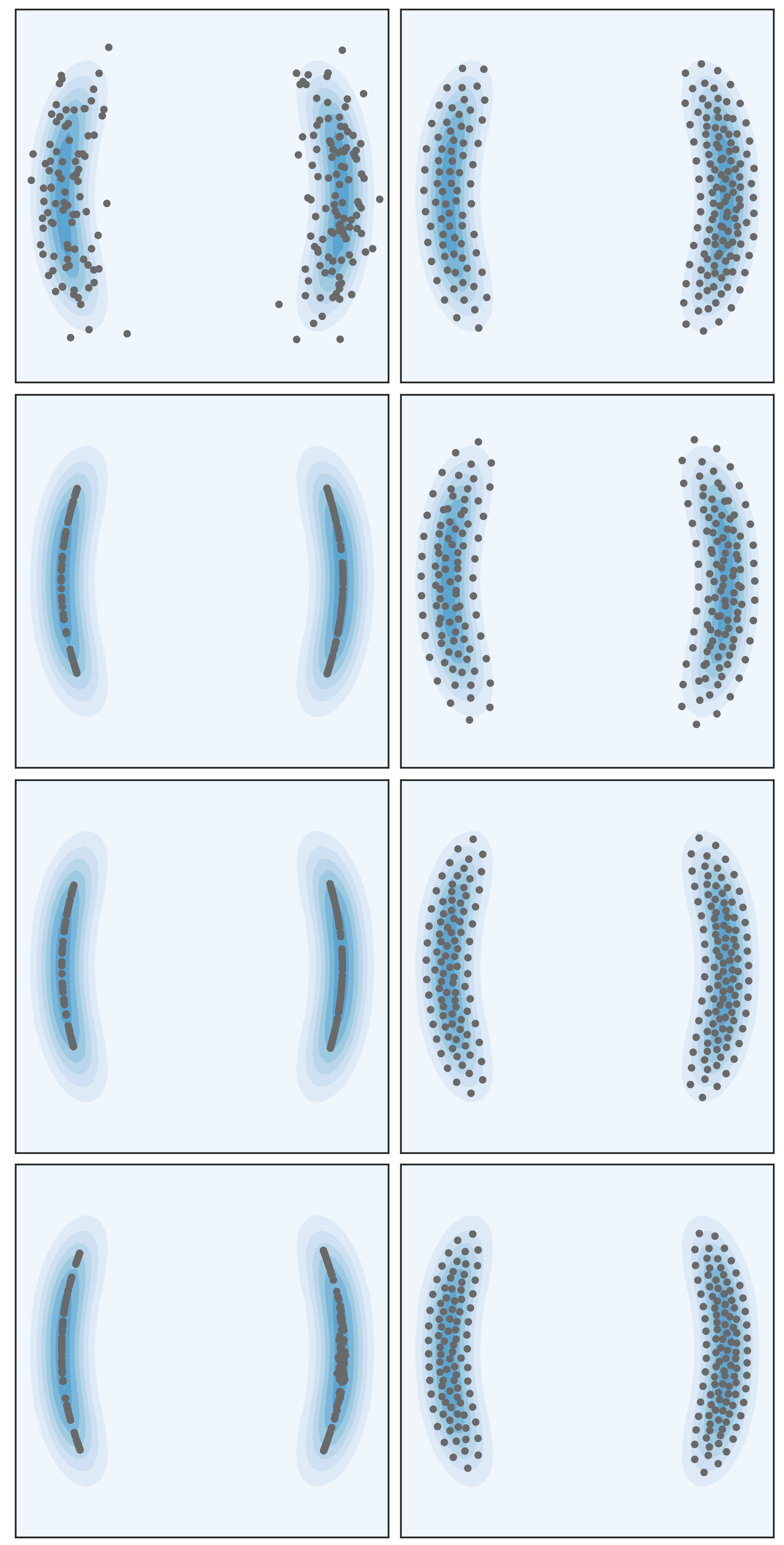
Experiments: Latent Dirichlet Allocation



Experiments

- Synthetic Experiment

Median method HE



SVGD

Blob

GFSD

GFSF

- Bayesian Logistic Regression