Riemannian Stein Variational Gradient Descent for Bayesian Inference

Chang Liu, Jun Zhu

chang-li14@mails.tsinghua.edu.cn, dcszj@tsinghua.edu.cn

Introduction

Task Bayesian inference: get access to the posterior of latent variable z given data x, $p(z|x) \propto p_0(z)p(x|z)$.

Proposal RSVGD: generalization of Stein Variational Gradient Descent (SVGD) [1] to Riemann manifold.

- SVGD: a P-VI with least assumption on the variational distribution (best flexibility).
- Importance to consider Riemann manifolds:
  Case (I): to do inference for posteriors defined on Riemann manifolds;
  Case (II): to improve efficiency by information geometry (to do inference on the distribution manifold) [2].

Table: A comparison of three kinds of inference methods.

Methods	M-VIs	MCs	P-VIs
Asymptotic Accuracy	No	Yes	Promising
Approximation Flexibility	Limited	Unlimited	Promisingly Unlimited
Iteration-Effectiveness	Yes	Weak	Strong
Particle-Efficiency	(does not apply)	Weak	Strong

M-VIs: model-based variational inference methods
MCs: Monte Carlo methods
P-VIs: particle-based variational inference methods

RSVGD: Directional Derivative

For $z \in \mathcal{M}$: m-dim Riemann manifold, a dynamics is defined by a vector field X.

Theorem 2 (Directional Derivative). $-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) = \mathbb{E}_{q_t}[\mathrm{div}(pX)/p] = \mathbb{E}_{q_t}\big[X[\log p] + \mathrm{div}(X)\big]$, *where in any c.s.,* $X[f] = X^i \partial_i f$ *(Einstein's convention),* $\mathrm{div}(X) = \partial_i(\sqrt{|G|}X^i)/\sqrt{|G|}$, $|G|$ *is the determinant of the Riemann metric tensor* G.

RSVGD: Functional Gradient

$$X^* = (\max \cdot \mathrm{argmax})_{X \in \mathfrak{X}, \|X\|_{\mathfrak{X}}=1} \mathbb{E}_{q_t}[\mathrm{div}(pX)/p].$$

Requirements for a reasonable and tractable $\mathfrak{X}$

- R1: X^* is a valid vector field on $\mathcal{M}$;
- R2: X^* is coordinate invariant;
- R3: X^* can be expressed in closed form, where q appears only in terms of $\mathbb{E}_q[\cdot]$.

SVGD's choice $\mathfrak{X} = \mathcal{H}^m$ does not meet R1 and R2!

Our Solution $\mathfrak{X} = \{\mathrm{grad}\, f | f \in \mathcal{H}\}$, where $\mathcal{H}$ is the RKHS of a Gaussian kernel on $\mathcal{M}$, and in any c.s., $(\mathrm{grad}\, f)^j = g^{ij}\partial_i f$ (g^{ij}: entries of G^{-1}). $f \to \mathrm{grad}\, f$ is a bijection, so $\langle \mathrm{grad}\, f, \mathrm{grad}\, h \rangle_{\mathfrak{X}} := \langle f, h \rangle_{\mathcal{H}}$.

Lemma 3. *$(\mathfrak{X}, \langle \cdot, \cdot \rangle_{\mathfrak{X}})$ is a Hilbert space.*

Theorem 4 (Functional Gradient). *For $X \in \mathfrak{X}$ as defined above, we have $\mathbb{E}_{q_t}[\mathrm{div}(pX)/p] = \langle X, \hat{X} \rangle_{\mathfrak{X}}$, where $\hat{X} = \mathrm{grad}\, \hat{f}$,*

$$\hat{f}(z') = \mathbb{E}_{q(z)}\Big[\big(\mathrm{grad}\, K(z,z')\big)[\log p(z)] + \Delta K(z,z')\Big],$$

and $\Delta f := \mathrm{div}(\mathrm{grad}\, f)$. Furthermore, $X^* = \hat{X}$.

Our solution satisfies all the requirements.

- In any c.s., $\hat{X}^{i\prime} =$
  $$g'^{ij}\partial'_j \mathbb{E}_q\Big[\big(g^{ab}\partial_a \log(p\sqrt{|G|}) + \partial_a g^{ab}\big)\partial_b K + g^{ab}\partial_a \partial_b K\Big],$$
  which is used for Case (II).
- Riemannian Kernelized Stein Discrepancy: $\max_{X \in \mathfrak{X}, \|X\|_{\mathfrak{X}}=1} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) =$
  $$\mathbb{E}_q \mathbb{E}_{q'}\Big[\big(\mathrm{grad}' \log p'\big)\big[(\mathrm{grad} \log p)[K]\big] + \Delta'\Delta K + (\mathrm{grad}' \log p')[\Delta K] + (\mathrm{grad} \log p)[\Delta' K]\Big].$$

Preliminaries: SVGD

SVGD in theory

For $z \in \mathbb{R}^m$, use a proper dynamics $X \in \mathbb{R}^m$: $\frac{\mathrm{d}}{\mathrm{d}t}z(t) = X(z(t))$ to evolve the variational distribution $q_t(z)$ towards the target $p(z)$.

- Find the **Directional Derivative** of the objective $-\mathrm{KL}(q_t||p)$ wrt X:
  $$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) = \mathbb{E}_{q_t}[X^\top \nabla \log p + \nabla^\top X].$$
- The proper dynamics is the **Functional Gradient**:
  $$X^* = (\max \cdot \mathrm{argmax})_{X \in \mathfrak{X}, \|X\|_{\mathfrak{X}}=1} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p).$$

SVGD in practice

- For tractable X^*, restrict $\mathfrak{X}$ from $\mathbb{R}^m$ to $\mathcal{H}^m$, where $\mathcal{H}$ is the Reproducing Kernel Hilbert Space (RKHS) of some kernel K on $\mathbb{R}^m$, so that $X^*(z')$
  $$= \mathbb{E}_{q_t(z)}[K(z,z')\nabla_z \log p(z) + \nabla_z K(z,z')].$$
- q_t appears only in terms of expectation, so samples $\{z^{(s)}\}_{s=1}^S$ suffices and no need to restrict q_t to a specific parametric family.
- Update samples by discretizing the dynamics: $z^{(s)} \leftarrow z^{(s)} + \varepsilon X^*(z^{(s)})$.

RSVGD: Embedded Space

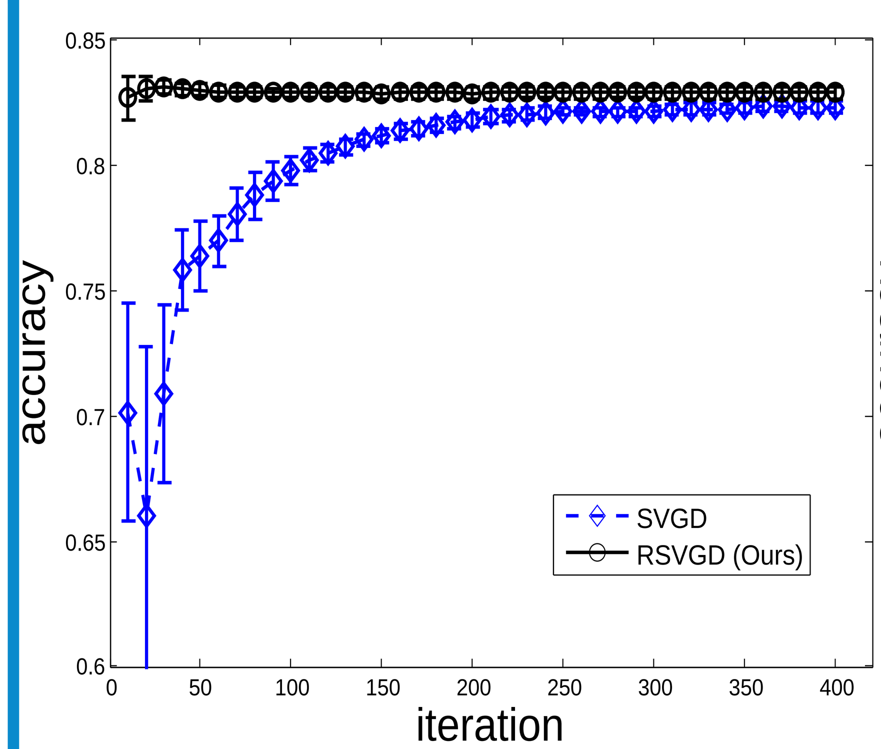
Why need an expression in the embedded space For Case (I), many manifolds have no global c.s., but are natural to express in the embedded space, e.g. hyperspheres $\mathbb{S}^{n-1} := \{y \in \mathbb{R}^n | \|y\|_2 = 1\}$.

Proposition 7. *For $\mathbb{S}^{n-1}$ isometrically embedded in $\mathbb{R}^n$ with orthonormal basis $\{y^\alpha\}_{\alpha=1}^n$, we have $\hat{X}' = (I_n - y'y'^\top)\nabla' \hat{f}'$,*

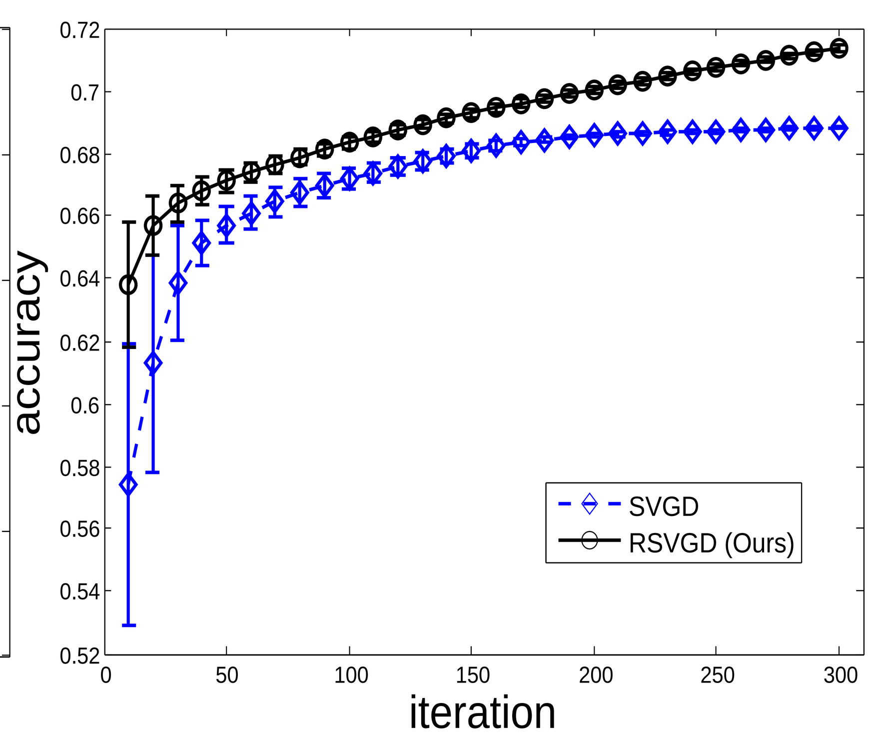
$$\hat{f}' = \mathbb{E}_q\Big[(\nabla \log p)^\top (\nabla K) + \nabla^\top \nabla K - y^\top(\nabla\nabla^\top K)y - (y^\top \nabla \log p + n - 1)y^\top \nabla K\Big].$$

Sample updating: $y^{(s)} \leftarrow \mathrm{Exp}_{y^{(s)}}(\varepsilon \hat{X}(y^{(s)}))$, where for $\mathbb{S}^{n-1}$, $\mathrm{Exp}_y(v) = y\cos(\|v\|) + (v/\|v\|)\sin(\|v\|)$.

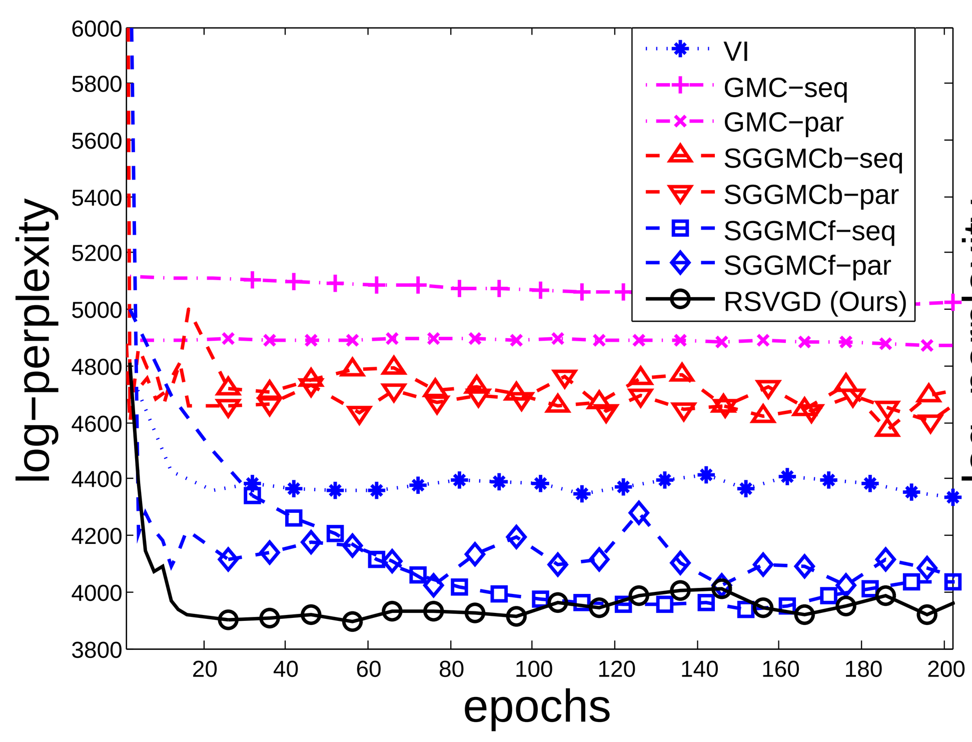
Experiments



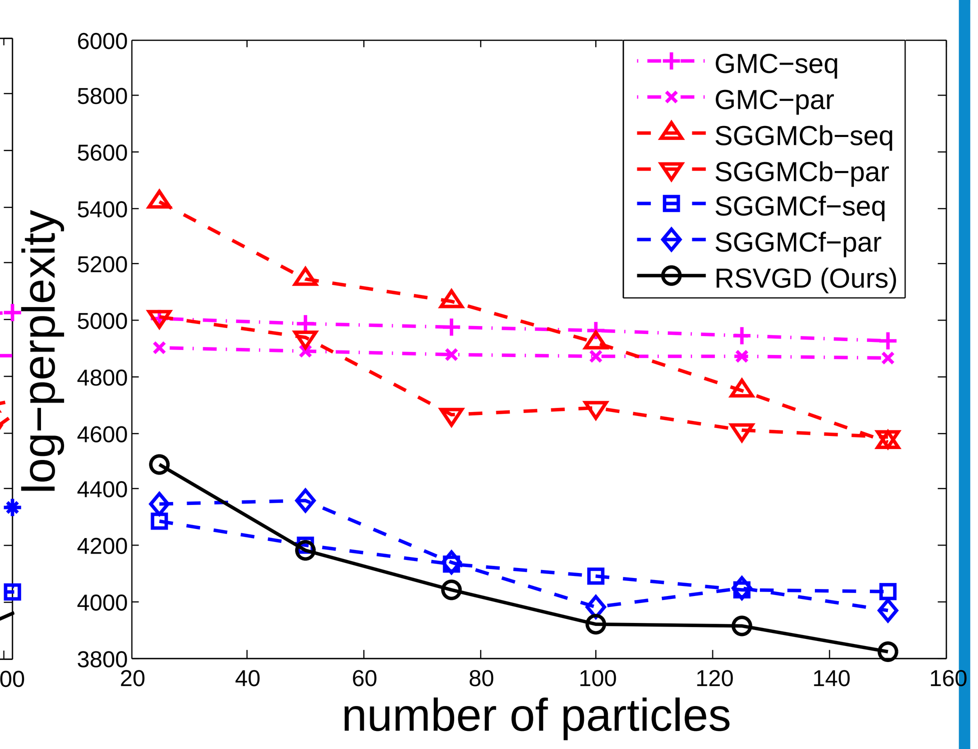
Case (II): On Splice19 dataset



Case (II): On Covertype dataset



Case (I): with 100 particles



Case (I): at 200 epoch

Case (II): Bayesian Logistic Regression

$w \sim \mathcal{N}(0, \alpha I_m)$, $y_d \sim \mathrm{Bern}(1/(1+e^{-w^\top x_d}))$.
Target: $p(w|\{y_d\},\{x_d\})$.
$G(w) = \mathcal{I}(p(\{y_d\}|w,\{x_d\})) - \nabla\nabla^\top \log p_0(w)$,
$\mathcal{I}(p(\cdot|w))$ is the Fisher information matrix.

Case (I): Spherical Admixture Model [3]

Topic model: corpus mean $\mu \sim \mathrm{vMF}(m, \kappa_0)$, topic $\beta_k \sim \mathrm{vMF}(\mu, \sigma)$, topic proportion $\theta_d \sim \mathrm{Dir}(\alpha)$ and content $v_d \sim \mathrm{vMF}(\beta\theta_d/\|\beta\theta_d\|, \kappa)$.
Target: $p(\beta|v)$. Note $\mu, \beta_k, v_d \in \mathbb{S}^{n-1}$!

We use vMF kernel $K(y,y') = \exp(\kappa y^\top y')$ on $\mathbb{S}^{n-1}$.
Baselines:
VI [3], and MCMCs: GMC [4], SGGMC [5].
Evaluation: log-perplexity$:= -\mathbb{E}_{\hat{p}(\beta|v)}[\log p(v_{\text{test}}|\beta)]$ (the lower the better).

[1] Liu, Q., and Wang, D. 2016. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, 2370-2378.
[2] Amari, S.-I. 2016. *Information geometry and its applications*. Springer.
[3] Reisinger, J.; Waters, A.; Silverthorn, B.; and Mooney, R. J. 2010. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 903-910.
[4] Byrne, S., and Girolami, M. 2013. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics* 40(4):825-845.
[5] Liu, C.; Zhu, J.; and Song, Y. 2016. Stochastic gradient geodesic mcmc methods. In *Advances In Neural Information Processing Systems*, 3009-3017.