Riemannian Stein Variational Gradient Descent for Bayesian Inference

Chang Liu, Jun Zhu[1]

Dept. of Comp. Sci. & Tech., TNList Lab; Center for Bio-Inspired Computing Research
State Key Lab for Intell. Tech. & Systems, Tsinghua University, Beijing, China

{*chang-li14@mails., dcszj@*}*tsinghua.edu.cn*

AAAI'18 @ New Orleans

[1]Corresponding author.

Introduction

- Bayesian inference: given a dataset $\mathcal{D}$ and a Bayesian model $p(x, \mathcal{D})$, estimate the posterior of the latent variable $p(x|\mathcal{D})$.
- Comparison of current inference methods: model-based variational inference methods (M-VIs), Monte Carlo methods (MCs) and particle-based variational inference methods (P-VIs)

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

Stein Variational Gradient Descent (SVGD) [7]: a P-VI with minimal assumption and impressive performance.

Introduction

In this work:

Generalize SVGD to the Riemann manifold settings, so that we can:

Purpose 1

Adapt SVGD to tasks on Riemann manifold and introduce the first P-VI to the Riemannian world.

Purpose 2

Improve SVGD efficiency for usual tasks (ones on Euclidean space) by exploring information geometry.

Stein Variational Gradient Descent (SVGD)

The idea of SVGD:

- A deterministic continuous-time dynamics $\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \phi(x(t))$ on $\mathcal{M} = \mathbb{R}^m$ (where $\phi : \mathbb{R}^m \to \mathbb{R}^m$) will induce a continuously evolving distribution q_t on $\mathcal{M}$.
- At some instant t, for a fixed dynamics ϕ, find the decreasing rate of $\mathrm{KL}(q_t||p)$, i.e. the *Directional Derivative* $-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ in the "direction" of ϕ.
- Find ϕ that maximizes the directional derivative , i.e. the *Functional Gradient* ϕ^* (the steepest ascending "direction"). For close-form solution, ϕ^* is chosen from $\mathcal{H}^m$, where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) of some kernel.
- Apply the dynamics ϕ^* to samples $\{x^{(s)}\}_{s=1}^S$ of q_t: $\{x^{(s)} + \varepsilon\phi^*(x^{(s)})\}_{s=1}^S$ forms a set of samples of $q_{t+\varepsilon}$.

Roadmap

For a general Riemann manifold $\mathcal{M}$,

- Any deterministic continuous-time dynamics on $\mathcal{M}$ is described by a vector field X on $\mathcal{M}$. It induces a continuously evolving distribution on $\mathcal{M}$ with density q_t (w.r.t. Riemann volume form).
- Derive the *Directional Derivative* $-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ under dynamics X.
- Derive the *Functional Gradient* $X^* := (\max \cdot \arg\max)_{\|X\|_{\mathfrak{X}}=1} - \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$.
- Moreover, for *Purpose 1*, express X^* in the *Embedded Space* of $\mathcal{M}$ when $\mathcal{M}$ has no global coordinate systems (c.s.), e.g. hyperspheres.
- Finally, simulate the dynamics X^* for a small time step ε to update samples.

Derivation of the Directional Derivative

Let q_t be the evolving density under dynamics X.

Lemma (Continuity Equation on Riemann Manifold)

$$\frac{\partial q_t}{\partial t} = -\mathrm{div}(q_t X) = -X[q_t] - q_t \mathrm{div}(X).$$

- $X[q_t]$: the action of the vector field X on the smooth function q_t. In any c.s., $X[q_t] = X^i \partial_i q_t$.
- $\mathrm{div}(X)$: the divergence of vector field X. In any c.s., $\mathrm{div}(X) = \partial_i(\sqrt{|G|}X^i)/\sqrt{|G|}$, where G is the matrix expression under the c.s. of the Riemann metric of $\mathcal{M}$.

Theorem (Directional Derivative)

Let p be a fixed distribution. Then the directional derivative is

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) = \mathbb{E}_{q_t}[\mathrm{div}(pX)/p] = \mathbb{E}_{q_t}\big[X[\log p] + \mathrm{div}(X)\big].$$

Derivation of the Functional Gradient

The task now:

$$X^* := (\max \cdot \arg\max)_{X \in \mathfrak{X}, \|X\|_{\mathfrak{X}}=1} \mathcal{J}(X) := \mathbb{E}_q\big[X[\log p] + \mathrm{div}(X)\big],$$

where $\mathfrak{X}$ is some subspace of the space of vector fields on $\mathcal{M}$, such that the requirements are met:

Requirements on X^*, thus on $\mathfrak{X}$

- 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 expectation.

Derivation of the Functional Gradient

R1: X^* is a valid vector field on $\mathcal{M}$.

- Why needed: deductions are based on valid vector fields.
- Note: non-trivial to guarantee!

Example (Vector fields on hyperspheres)

Vector fields on an even-dimensional hypersphere must have one zero-vector-valued point (critical point) due to the hairy ball theorem ([1], Theorem 8.5.13). The choice in SVGD $\mathfrak{X} = \mathcal{H}^m$ cannot guarantee R1.

Derivation of the Functional Gradient

R2: X^* is coordinate invariant.

- Concept: the expression of an object on $\mathcal{M}$ in any c.s. is the same. E.g. vector field, gradient and divergence.
- Why needed: necessary to avoid ambiguity or arbitrariness of the solution. The vector field X^* should be independent of the choice of c.s. in which it is expressed.
- Note: the choice in SVGD $\mathfrak{X} = \mathcal{H}^m$ cannot guarantee R2.

R3: X^* can be expressed in closed form, where q appears only in terms of expectation.

- Why needed: for tractable implementation, and for avoiding making restrictive assumptions on q.

Derivation of the Functional Gradient

Our Solution

$\mathfrak{X} = \{\mathrm{grad}\, f | f \in \mathcal{H}\}$, where $\mathcal{H}$ is the RKHS of some kernel.

- $\mathrm{grad}\, f$ is the gradient of the smooth function f. In any c.s., $(\mathrm{grad}\, f)^j = g^{ij}\partial_i f$, where g^{ij} is the entry of G^{-1} under the c.s.

Theorem

For Gaussian RKHS, $\mathfrak{X}$ is isometrically isomorphic to $\mathcal{H}$, thus it is a Hilbert space.

Our solution guarantees all the requirements:

- The gradient is a well-defined object on $\mathcal{M}$ and it is guaranteed to be a valid vector field and coordinate invariant (see paper for detailed interpretation).
- Close-form solution can be derived (see next).

Derivation of the Functional Gradient

The close-form solution:

Theorem (Functional Gradient)

$$X^{*\prime} = \mathrm{grad}' f^{*\prime},\ f^{*\prime} = \mathbb{E}_q\big[(\mathrm{grad}\, K)[\log p] + \Delta K\big],$$

where notations with prime "$'$" take x' as argument while others take x and K takes both, and $\Delta f := \mathrm{div}(\mathrm{grad}\, f)$. In any c.s.,

$$X^{*\prime i} = g'^{ij}\partial'_j \mathbb{E}_q\Big[\big(g^{ab}\partial_a \log(p\sqrt{|G|}) + \partial_a g^{ab}\big)\partial_b K + g^{ab}\partial_a\partial_b K\Big].$$

Derivation of the Functional Gradient

Purpose 2

Improve efficiency for the usual inference tasks on Euclidean space $\mathbb{R}^m$.

- Apply the idea of *information geometry* [3, 2]: for a Bayesian model with prior $p(x)$ and likelihood $p(\mathcal{D}|x)$, take $\mathcal{M} = \{p(\cdot|x) : x \in \mathbb{R}^m\}$ and treat x as the coordinate of $p(\cdot|x)$. In this global c.s., $G(x)$ is the Fisher information matrix of $p(\cdot|x)$ (and typically subtract by the Hessian of $\log p(x)$).
- Calculate the tangent vector at each sample using the c.s. expression
$$X^{*\prime i} = g'^{ij} \partial'_j \mathbb{E}_q \Big[\big(g^{ab} \partial_a \log(p\sqrt{|G|}) + \partial_a g^{ab} \big) \partial_b K + g^{ab} \partial_a \partial_b K \Big],$$
where the target distribution $p = p(x|\mathcal{D}) \propto p(x)p(\mathcal{D}|x)$ and the expectation is estimated by averaging over samples.
- Simulate the dynamics for a small time step ε to update samples:
$$x^{(s)} \leftarrow x^{(s)} + \varepsilon X^*(x^{(s)}).$$

Expression in the Embedded Space

Purpose 1

Enable applicability to inference tasks on non-linear Riemann manifolds.

In the coordinate space of $\mathcal{M}$:

- Some manifolds have no global c.s., e.g. hypersphere $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$ and Stiefel manifold [5]. Cumbersome switch among local c.s.
- G would be singular near the edge of coordinate space.

In the embedded space of $\mathcal{M}$:

- $\mathcal{M}$ can be expressed globally, and is natural for $\mathbb{S}^{n-1}$ and Stiefel manifold.
- No singularity problems.
- Requires exponential map and density w.r.t. Hausdorff measure, which are available for $\mathbb{S}^{n-1}$ and Stiefel manifold.

Expression in the Embedded Space

Proposition (Functional Gradient in the Embedded Space)

Let m-dim Riemann manifold $\mathcal{M}$ isometrically embedded in $\mathbb{R}^n$ (with orthonormal basis $\{y^\alpha\}_{\alpha=1}^n$) via $\Xi : \mathcal{M} \to \mathbb{R}^n$. Let p be the density w.r.t. the Hausdorff measure on $\Xi(\mathcal{M})$. Then $X^{*\prime} = (I_n - N' N'^\top) \nabla' f^{*\prime}$,

$$f^{*\prime} = \mathbb{E}_q \Big[\Big(\nabla \log \big(p \sqrt{|G|} \big) \Big)^\top \Big(I_n - N N^\top \Big) (\nabla K) + \nabla^\top \nabla K \\ - \mathrm{tr}\Big(N^\top (\nabla \nabla^\top K) N \Big) + \Big((M^\top \nabla)^\top (G^{-1} M^\top) \Big) (\nabla K) \Big],$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix, $\nabla = (\partial_{y^1}, \ldots, \partial_{y^n})^\top$, $M \in \mathbb{R}^{n \times m} : M_{\alpha i} = \frac{\partial y^\alpha}{\partial x^i}$, $N \in \mathbb{R}^{n \times (n-m)}$ is the set of orthonormal basis of the orthogonal complement of $\Xi_*(T_x \mathcal{M})$, and $\mathrm{tr}(\cdot)$ is the trace.

- Simulating the dynamics requires the exponential map Exp of $\mathcal{M}$:

$$y^{(s)} \leftarrow \mathrm{Exp}_{y^{(s)}}(\varepsilon X^*(y^{(s)})).$$

$\mathrm{Exp}_y(v)$: moves y on $\Xi(\mathcal{M})$ "straightly" along the direction of v.

Expression in the Embedded Space

Proposition (Functional Gradient for Embedded Hyperspheres)

For $\mathbb{S}^{n-1}$ isometrically embedded in $\mathbb{R}^n$ with orthonormal basis $\{y^\alpha\}_{\alpha=1}^n$, we have $X^{*\prime} = (I_n - y' {y'}^\top)\nabla' f^{*\prime}$, where $f^{*\prime} =$
$\mathbb{E}_q\Big[(\nabla \log p)^\top (\nabla K) + \nabla^\top \nabla K - y^\top \big(\nabla\nabla^\top K\big) y - (y^\top \nabla \log p + n - 1) y^\top \nabla K\Big]$.

- Exponential map on $\mathbb{S}^{n-1}$:

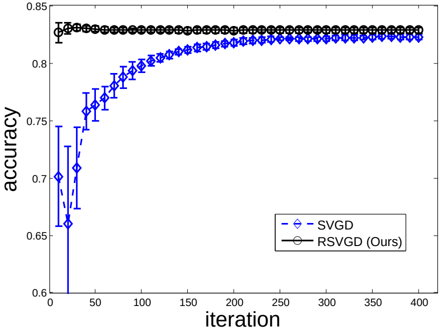
$$\mathrm{Exp}_y(v) = y\cos(\|v\|) + (v/\|v\|)\sin(\|v\|).$$

BLR: for *Purpose 2*

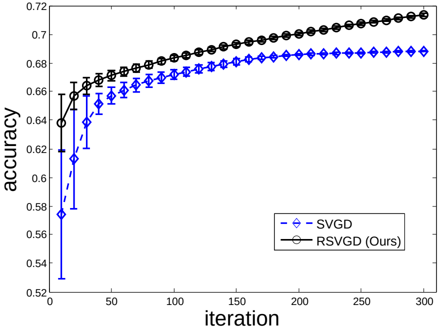
- Model: Bayesian Logistic Regression (BLR)
  $w \sim \mathcal{N}(0, \alpha I_m)$, $y_d \sim \mathrm{Bern}(\sigma(w^\top x_d))$, where $\sigma(x) = 1/(1 + e^{-x})$.
  - Euclidean task: $w \in \mathbb{R}^m$.
  - Posterior: $p(w|\{(x_d, y_d)\})$, log-density gradient known.
  - Riemann metric tensor G: $\mathrm{FisherInfo} - \mathrm{Hessian}$, known in close form.
- Kernel: Gaussian kernel in the coordinate space.
- Baselines: vanilla SVGD.
- Evaluation: averaged test accuracy.

BLR: for *Purpose 2*

- Results:



(a) On Splice19 dataset



(b) On Covertype dataset

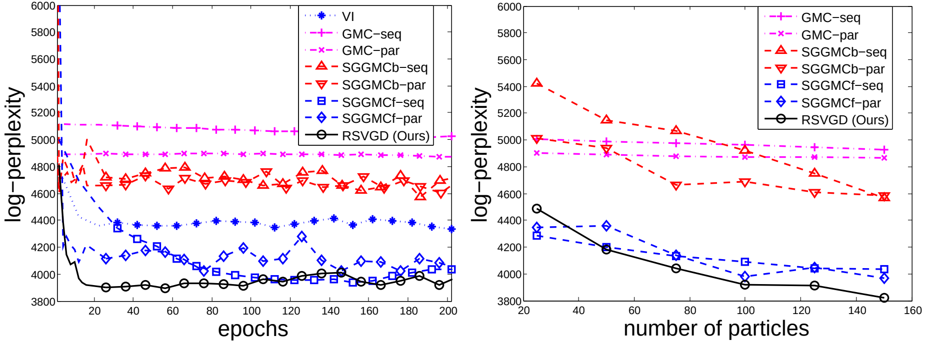
Figure: Test accuracy along iteration for BLR. Both methods are run 20 times on Splice19 and 10 times on Covertype. Each run on Covertype uses a random train(80%)-test(20%) split as in [7].

SAM: for *Purpose 1*

- Model: Spherical Admixture Model (SAM) [8]
  Observed var.: *tf-idf* representation of documents: $v_d \in \mathbb{S}^{V-1}$.
  Latent var.: spherical topics: $\beta_t \in \mathbb{S}^{V-1}$.
  - Non-linear Riemann manifold task: $\beta \in (\mathbb{S}^{V-1})^T$.
  - Posterior: $p(\beta|v)$ (w.r.t. the Hausdorff measure), log-density gradient can be estimated [6].
- Kernel: von-Mises Fisher (vMF) kernel $K(y, y') = \exp(\kappa y^\top y')$, the restriction of Gaussian kernel in $\mathbb{R}^n$ on $\mathbb{S}^{n-1}$.
- Baselines:
  - Variational Inference (VI) [8]: the vanilla inference method of SAM.
  - Geodesic Monte Carlo (GMC) [4]: MCMC for RM in the embed. sp.
  - Stochastic Gradient GMC (SGGMC) [6]: SG-MCMC for RM in the embeded space. (-b: mini-batch grad. est. -f: full-batch grad. est.)
  - For MCMCs, -seq: samples from one chain. -par: newest samples from multiple chains.
- Evaluation: log-perplexity (negative log-likelihood of test dataset under the trained model) [6].

SAM: for *Purpose 1*

- Results:



(a) Results with 100 particles

(b) Results at 200 epochs

Figure: Results on the SAM inference task on 20News-different dataset, in log-perplexity. We run SGGMCf for full batch and SGGMCb for a mini-batch size of 50.

Thank you!

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu.
Manifolds, tensor analysis, and applications, volume 75.
Springer Science & Business Media, 2012.

Shun-Ichi Amari.
Information geometry and its applications.
Springer, 2016.

Shun-Ichi Amari and Hiroshi Nagaoka.
Methods of information geometry, volume 191.
American Mathematical Soc., 2007.

Simon Byrne and Mark Girolami.
Geodesic monte carlo on embedded manifolds.
Scandinavian Journal of Statistics, 40(4):825–845, 2013.

I. M. James.
The topology of Stiefel manifolds, volume 24.
Cambridge University Press, 1976.

Chang Liu, Jun Zhu, and Yang Song.
Stochastic gradient geodesic mcmc methods.
In *Advances In Neural Information Processing Systems*, pages 3009–3017, 2016.

Qiang Liu and Dilin Wang.
Stein variational gradient descent: A general purpose bayesian inference algorithm.
In *Advances in Neural Information Processing Systems*, pages 2370–2378, 2016.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney.
Spherical topic models.
In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 903–910, 2010.