# Appendix for: Riemannian Stein Variational Gradient Descent for Bayesian Inference

**Chang Liu, Jun Zhu***
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.tsinghua.edu.cn; dcszj@tsinghua.edu.cn

## A1. Proof of Lemma 1 (Continuity Equation on Riemann Manifold)

Let $F_{(\cdot)}(\cdot)$ be the flow of $X$. $\forall U \subset \mathcal{M}$ compact, consider the integral $\int_{F_t(U)} p_t \mu_g$. Since a particle in $U$ at time 0 will always in $F_t(U)$ at time $t$ and vice versa, the integral, i.e. the portion of particles in $F_t(U)$ at time $t$, is equal to the portion of particles in $U$ at time 0 for any time $t$. So it is a constant. Reynolds transport theorem gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} \int_{F_t(U)} p_t \mu_g = \int_{F_t(U)} \left( \frac{\partial p_t}{\partial t} + \mathrm{div}(p_t X) \right) \mu_g$$

for any $U$ and $t$, so the integrand must be zero and we derived the conclusion.

## A2. Well-definedness of KL-divergence on Riemann Manifold

We define the KL-divergence between two distributions on $\mathcal{M}$ by their p.d.f. $q^\mu$ and $p^\mu$ w.r.t. volume form $\mu$ as:

$$\mathrm{KL}(q||p) := \int_{\mathcal{M}} q^\mu \log(q^\mu / p^\mu) \mu.$$

To make this notion well-defined, we need to show that the right hand side of the definition is invariant of $\mu$. Let $\omega$ be another volume form. Since $\forall A \in \mathcal{M}$, $\mu(A)$ and $\omega(A)$ lie on the same 1-dimensional linear space (the space of $m$-forms at $A$), we have $\alpha(A) \in \mathbb{R}^+$ s.t. $\omega(A) = \alpha(A)\mu(A)$. Such a construction gives a smooth function $\alpha : \mathcal{M} \to \mathbb{R}^+$. By the definition of p.d.f., $q^\omega = q^\mu / \alpha$. So $\int_{\mathcal{M}} q^\omega \log(q^\omega / p^\omega)\omega = \int_{\mathcal{M}} q^\mu \log(q^\mu / p^\mu)\mu$, which indicates that the integral is independent of the chosen volume form.

## A3. Proof of Theorem 2

To formally prove Theorem 2, we first deduce a lemma, which gives the p.d.f. of the distribution transformed by a diffeomorphism on $\mathcal{M}$ (an invertible smooth transformation on $\mathcal{M}$).

**Lemma 8** (Transformed p.d.f.)**.** *Let $\phi$ be an orientation-preserving diffeomorphism on $\mathcal{M}$, and $p$ the p.d.f. of a distribution on $\mathcal{M}$. Denote $p_{[\phi]}$ as the p.d.f. of the distribution of the $\phi$-transformed random variable from the one obeying $p$, i.e. the transformed p.d.f. Then in any local coordinate system (c.s.) $(U, \Phi)$,*

$$p_{[\phi]} = \frac{\left(p\sqrt{|G|}\right) \circ \phi^{-1}}{\sqrt{|G|}} \left|\mathrm{Jac}\, \phi^{-1}\right|, \tag{9}$$

*where $G$ is the Riemann metric tensor in $(U, \Phi)$ and $|G|$ is its determinant, and $\mathrm{Jac}\, \phi^{-1}$ is the Jacobian determinant of $\Phi \circ \phi^{-1} \circ \Phi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$. The right hand side is coordinate invariant.*

*Proof.* Let $U$ be a compact subset of $\mathcal{M}$, and $(V, \Phi), V \subset U$ be a local c.s. of $U$ with coordinate chart $\{x^i\}_{i=1}^m$. On one hand, due to the definition of $p_{[\phi]}$, we have $\mathrm{Prob}_p(U) = \mathrm{Prob}_{p_{[\phi]}}(\phi(U))$. On the other hand, we can invoke the theorem of global change of variables on manifold (Abraham, Marsden, and Ratiu, 2012, Theorem 8.1.7), which gives $\mathrm{Prob}_p(U) =$

$$\int_U p\mu_g = \int_{\phi(U)} {\phi^{-1}}^*(p\mu_g) = \int_{\phi(U)} (p \circ \phi^{-1}) {\phi^{-1}}^*(\mu_g) \tag{10}$$

$$= \int_{\phi(U)} (p \circ \phi^{-1})(\sqrt{|G|} \circ \phi^{-1}) |\mathrm{Jac}\, \phi^{-1}| \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^m$$

$$= \int_{\phi(U)} \frac{\left(p\sqrt{|G|}\right) \circ \phi^{-1}}{\sqrt{|G|}} |\mathrm{Jac}\, \phi^{-1}| \mu_g \tag{11}$$

$= \mathrm{Prob}_{\frac{\left(p\sqrt{|G|}\right)\circ\phi^{-1}}{\sqrt{|G|}} |\mathrm{Jac}\, \phi^{-1}|}(\phi(U))$, where ${\phi^{-1}}^*(\cdot)$ is the pull-back of $\phi^{-1}$ on the $m$-forms on $\mathcal{M}$. Combining both hands and noting the arbitrariness of $U$, we get the desired conclusion. $\square$

Let $F_{(\cdot)}(\cdot)$ be the flow of $X$. For any evolving distribution $p_t$ under dynamics $X$, by its definition, we have $p_t = {p_0}_{[F_t]}$. Due to the property of flow that for any $s, t \in \mathbb{R}$, $F_{s+t} = F_s \circ F_t = F_t \circ F_s$, we have $p_{s+t} = {p_0}_{[F_{s+t}]} = {p_0}_{[F_s \circ F_t]} = ({p_0}_{[F_s]})_{[F_t]} = (p_s)_{[F_t]}$.

Now, for a fixed $t_0 \in \mathbb{R}$, we let $p_t$ be the evolving distribution under $X$ that satisfies $p_{t_0} = p$, the target distribution. For sufficiently small $t > 0$, $F_t(\cdot)$ is a diffeomorphism on

*corresponding author.

$\mathcal{M}$. Equipped with all these knowledge, we begin the final deduction:

$$-\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_0} \mathrm{KL}(q_t||p) = -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathcal{M}} q_{t_0+t} \log \frac{q_{t_0+t}}{p_{t_0}} \mu_g$$

(Treat $q_{t_0+t}$ as $(q_{t_0})_{[F_t]}$ and apply Eqn. (9))

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathcal{M}} \frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} \left|\mathrm{Jac}\, F_t^{-1}\right| \cdot \left(\log \frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} + \log\left|\mathrm{Jac}\, F_t^{-1}\right| - \log p_{t_0}\right) \mu_g$$

(Apply $F_t^{-1}$ to the entire integral and invoke the theorem of global change of variables Eqn. (10))

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{F_t^{-1}(\mathcal{M})} \left(\left[\frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} \left|\mathrm{Jac}\, F_t^{-1}\right| \cdot \left(\log \frac{\left(q_{t_0}\sqrt{|G|}\right) \circ F_t^{-1}}{\sqrt{|G|}} + \log\left|\mathrm{Jac}\, F_t^{-1}\right| - \log p_{t_0}\right)\right] \circ F_t\right) F_t^*(\mu_g)$$

($F_t^{-1}(\mathcal{M}) = \mathcal{M}$ since $F_t^{-1}$ is a diffeomorphism on $\mathcal{M}$. $|\mathrm{Jac}\, F_t^{-1}| \circ F_t = |\mathrm{Jac}\, F_t|^{-1}$. See Eqn. (11) for the expression of $F_t^*(\mu_g)$, the pull-back of $F_t$ on $\mu_g$)

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathcal{M}} \frac{q_{t_0}\sqrt{|G|}}{\sqrt{|G|} \circ F_t} |\mathrm{Jac}\, F_t|^{-1} \cdot \left(\log \frac{q_{t_0}\sqrt{|G|}}{\sqrt{|G|} \circ F_t} - \log|\mathrm{Jac}\, F_t| - \log(p_{t_0} \circ F_t)\right) \cdot \frac{\sqrt{|G|} \circ F_t}{\sqrt{|G|}} |\mathrm{Jac}\, F_t|\, \mu_g$$

(Rearange terms)

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathcal{M}} q_{t_0} \left[\log q_{t_0} - \log\left(\frac{\left(p_{t_0}\sqrt{|G|}\right) \circ F_t}{\sqrt{|G|}} |\mathrm{Jac}\, F_t|\right)\right] \mu_g$$

(Note the property of flow: $F_t = F_{-t}^{-1}$. Treat $p_{t_0-t}$ as $(p_{t_0})_{[F_{-t}]}$ and apply Eqn. (9) inversely)

$$= -\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\mathcal{M}} q_{t_0} \left[\log q_{t_0} - \log p_{t_0-t}\right] \mu_g$$

($\mathcal{M}$ is unchanged over time $t$ (otherwise an integral over the boundary would appear))

$$= \int_{\mathcal{M}} q_{t_0} \frac{\partial}{\partial t} (\log p_{t_0-t}) \Big|_{t=0} \mu_g = -\int_{\mathcal{M}} q_{t_0} \frac{\partial}{\partial t} (\log p_{t_0+t}) \Big|_{t=0} \mu_g$$

(Refer to Eqn. (3))

$$= \int_{\mathcal{M}} (q_{t_0}/p_{t_0}) \mathrm{div}(p_{t_0} X) \mu_g = \mathbb{E}_{q_{t_0}}[\mathrm{div}(p_{t_0} X)/p_{t_0}]$$

(Property of divergence)

$$= \mathbb{E}_{q_{t_0}}\left[X[\log p_{t_0}] + \mathrm{div}(X)\right].$$

Due to the arbitrariness of $t_0$, we get the desired conclusion and complete the proof.

## A4. Condition for Stein's Identity (Stein Class)

Now we derive the condition for Stein's identity to hold. We require $\mathbb{E}_p[\mathrm{div}(pX)/p] = 0$, which is

$$\int_{\mathcal{M}} \mathrm{div}(pX) \mu_g = \int_{\partial\mathcal{M}} \mathrm{i}_{(pX)} \mu_g = \sum_{i=1}^{m} \int_{\partial\mathcal{M}} p\sqrt{|G|}(-1)^{i+1} X^i \bigwedge \mathrm{d}x^{\neg i},$$

where the first equality holds due to Gauss' theorem (Abraham, Marsden, and Ratiu, 2012, Theorem 8.2.9), $\partial\mathcal{M}$ is the boundary of $\mathcal{M}$, $\mathrm{i}_X : A^k(\mathcal{M}) \to A^{k-1}(\mathcal{M})$ is the interior product or contraction, $(\mathrm{i}_X\omega)(A)[v_1, \dots, v_{k-1}] = \omega(A)[X(A), v_1, \dots, v_{k-1}]$, $X^i$ is the $i$-th component of $X$ under the natural basis of some local c.s., $\bigwedge \mathrm{d}x^{\neg i} := \mathrm{d}x^1 \wedge \dots \wedge \mathrm{d}x^{i-1} \wedge \mathrm{d}x^{i+1} \wedge \dots \wedge \mathrm{d}x^m$ with "$\wedge$" the wedge product (exterior product).

For manifolds like spheres, $\partial\mathcal{M}$ is empty and the above integral is always zero, so the Stein class is $\mathcal{T}(\mathcal{M})$. If $\partial\mathcal{M}$ is not empty, by its definition, around any point on $\partial\mathcal{M}$ there exists a c.s. $(V, \Psi)$ with coordinate chart $(y^1, \dots, y^m)$ such that $\forall A \in \partial\mathcal{M} \cap V, y^m(A) = 0$. Thus $\mathrm{d}y^m = 0$ and $(\partial\mathcal{M} \cap V, \tilde{\Psi} = (\Psi^1, \dots, \Psi^{m-1}))$ is a local c.s. of $\partial\mathcal{M}$. Then the condition for Stein's identity to hold becomes

$$\int_{\partial\mathcal{M}} p\tilde{X}^m \sqrt{|\tilde{G}|} \mathrm{d}y^1 \wedge \dots \wedge \mathrm{d}y^{m-1} = 0,$$

where $\tilde{G}$ is the Riemann metric tensor in $(\partial\mathcal{M} \cap V, \tilde{\Psi})$, and $\tilde{X}^m$ is the $m$-th component of $X$ in $(\partial\mathcal{M} \cap V, \tilde{\Psi})$.

For the case where $\mathcal{M}$ is a compact subset of Euclidean space $\mathbb{R}^m$, around any point $A$ on the boundary $\partial\mathcal{M}$, we take $(V, \Psi)$ such that $y^m = 0$ and the natural basis $\{\partial_i | \partial_i := \frac{\partial}{\partial y^i}, i = 1, \dots, m\}$ is orthonormal. Then $|\tilde{G}(A)| = 1$ and $\partial_m$ is perpendicular to the span of $\{\partial_1, \dots, \partial_{m-1}\}$, which is the tangent space of $\partial\mathcal{M}$ at $A$. So $\partial_m$ is the unit normal $\vec{n}$ to $\partial\mathcal{M}$, and $\tilde{X}^m$ is the component of $X$ along the normal direction, i.e. $\tilde{X}^m = X \cdot \vec{n}$. Denote the volume form $\mathrm{d}y^1 \wedge \dots \wedge \mathrm{d}y^{m-1}$ on $\partial\mathcal{M}$ as $\mathrm{d}S$, then the condition for Stein's identity is $\int_{\partial\mathcal{M}} pX \cdot \vec{n} \mathrm{d}S$, which meets the conclusion in (Liu and Wang 2016). We provide a generalization of the conclusion to general Riemann manifold.

## A5. Proof of Theorem 4

For any $X \in \mathfrak{X}$, let $f = \iota^{-1}(X)$ ($\iota$ is defined in the proof of Lemma 3), i.e. the only element in $\mathcal{H}_K$ such that $X =$

$\operatorname{grad} f$. Then in any c.s., $X = g^{ij}\partial_i f \partial_j$, and we have

$$
\begin{aligned}
&\mathcal{J}(X) := \mathbb{E}_q \left[ X[\log p] + \operatorname{div}(X) \right] \\
=&\mathbb{E}_q \left[ X^j \partial_j \log(p\sqrt{|G|}) + \partial_j X^j \right] \\
=&\mathbb{E}_q \left[ g^{ij}\partial_i f \partial_j \log(p\sqrt{|G|}) + \partial_j (g^{ij}\partial_i f) \right] \\
=&\mathbb{E}_q \left[ \left( g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij} \right) \partial_i f + g^{ij}\partial_i\partial_i f \right].
\end{aligned}
$$

Now we invoke the conclusions of Zhou (2008) that $\partial_i K(A,\cdot), \partial_i\partial_j K(A,\cdot) \in \mathcal{H}_K$, and for any $f \in \mathcal{H}_K$, $\langle f(\cdot), \partial_i K(A,\cdot)\rangle_{\mathcal{H}_K} = \partial_i f(A)$, $\langle f(\cdot), \partial_i\partial_j K(A,\cdot)\rangle_{\mathcal{H}_K} = \partial_i\partial_j f(A)$:

$$
\begin{aligned}
\mathcal{J}(X) =&\mathbb{E}_q \Big[ \left( g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij} \right) \langle f(\cdot), \partial_i K(A,\cdot)\rangle_{\mathcal{H}_K} \\
&+ g^{ij} \langle f(\cdot), \partial_i\partial_j K(A,\cdot)\rangle_{\mathcal{H}_K} \Big] \\
=&\mathbb{E}_q \Big[ \Big\langle f(\cdot), \left( g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij} \right) \partial_i K(A,\cdot) \\
&+ g^{ij}\partial_i\partial_j K(A,\cdot) \Big\rangle_{\mathcal{H}_K} \Big] \\
=&\Big\langle f(\cdot), \mathbb{E}_q \Big[ \left( g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij} \right) \partial_i K(A,\cdot) \\
&+ g^{ij}\partial_i\partial_j K(A,\cdot) \Big] \Big\rangle_{\mathcal{H}_K},
\end{aligned}
$$

where all the functions, differentiations and expectations are with argument $A$, if not specified. Define

$$
\begin{aligned}
\hat{f}(\cdot) =&\mathbb{E}_q \Big[ \left( g^{ij}\partial_j \log(p\sqrt{|G|}) + \partial_j g^{ij} \right) \partial_i K(A,\cdot) \\
&+ g^{ij}\partial_i\partial_j K(A,\cdot) \Big] \\
=&\mathbb{E}_q \Big[ g^{ij}\partial_j \log(p\sqrt{|G|}) \partial_i K(A,\cdot) \\
&+ \partial_j \big( \sqrt{|G|} g^{ij} \partial_i K(A,\cdot) \big) / \sqrt{|G|} \Big] \\
=&\mathbb{E}_q \Big[ g^{ij}\partial_j \log(p\sqrt{|G|}) \partial_i K(A,\cdot) + \Delta K(A,\cdot) \Big],
\end{aligned}
$$

we have $\mathcal{J}(X) = \langle f(\cdot), \hat{f}(\cdot)\rangle_{\mathcal{H}_K}$, and by the isometric isomorphism between $\mathcal{H}_K$ and $\mathfrak{X}$, we have $\mathcal{J}(X) = \langle \operatorname{grad} f, \operatorname{grad} \hat{f}\rangle_{\mathfrak{X}} = \langle X, \hat{X}\rangle_{\mathfrak{X}}$.

## A6 Expressions in the Isometrically Embedded Space

In this part of appendix we express the functional gradient in the isometrically embedded space, for general Riemann manifolds and two specific Riemann manifolds.

### A6.1 For General Riemann Manifolds (Proposition 6)

Let $\Xi$ be an isometric embedding of $\mathcal{M}$ into $(\mathbb{R}^n, \{y^\alpha\}_{\alpha=1}^n)$. For a coordinate system (c.s.) $(U, \Phi)$ of $\mathcal{M}$ with coordinate chart $\{x^i\}_{i=1}^m$, define $\xi := \Xi \circ \Phi^{-1}$. We first develop a key tool. Let $h : \Xi(\mathcal{M}) \to \mathbb{R}$ be a smooth function on the embedded manifold. In $(U, \Phi)$ we define $f := h \circ \xi : U \to \mathbb{R}$ as a smooth function on an open subset of $\mathbb{R}^m$. By the chain rule of derivative, we have

$$
\partial_i f = \partial_\alpha h \frac{\partial y^\alpha}{\partial x^i} = M^\top \nabla h,
$$

where $M \in \mathbb{R}^{n\times m} : M_{\alpha i} = \frac{\partial y^\alpha}{\partial x^i}$, and $\nabla h$ is the usual gradient of $h$ as a function on $\mathbb{R}^n$. For isometric embedding, we have $g_{ij} = \sum_{\alpha=1}^n \frac{\partial y^\alpha}{\partial x^i}\frac{\partial y^\alpha}{\partial x^j}$, or in matrix form $G = M^\top M$.

From Eqn. (5), we know that $\hat{f}' = \mathbb{E}_q\big[f_1 + f_2\big]$ where $f_1 = (\operatorname{grad} K)[\log p]$ and $f_2 = \Delta K$. Then in any c.s. of $\mathcal{M}$,

$$
\begin{aligned}
f_1 =& g^{ij} (\partial_i \log p)(\partial_j K) \\
=& g^{ij} \frac{\partial y^\alpha}{\partial x^i} (\partial_\alpha \log p) \frac{\partial y^\beta}{\partial x^j} (\partial_\beta K) \\
=& (\nabla \log p)^\top (M G^{-1} M^\top) \nabla K, \\
f_2 =& g^{ij} (\partial_i K)(\partial_j \log\sqrt{|G|}) + \partial_i (g^{ij}\partial_j K) \\
=& (\nabla \log\sqrt{|G|})^\top (M G^{-1} M^\top) \nabla K + \frac{\partial y^\alpha}{\partial x^i} \partial_\alpha \Big( g^{ij} \frac{\partial y^\beta}{\partial x^j} \partial_\beta K \Big) \\
=& (\nabla \log\sqrt{|G|})^\top (M G^{-1} M^\top) \nabla K \\
&+ (M^\top \nabla)^\top (G^{-1} M^\top \nabla K) \\
=& (\nabla \log\sqrt{|G|})^\top (M G^{-1} M^\top) \nabla K \\
&+ \Big( (M^\top \nabla)^\top (G^{-1} M^\top) \Big) \nabla K \\
&+ \operatorname{tr}\Big( (\nabla\nabla^\top K)(M G^{-1} M^\top) \Big).
\end{aligned}
$$

To further simplify the expression, we mention it here that the operator $MG^{-1}M^\top = M(M^\top M)^{-1}M^\top$ is the orthogonal projection onto the column space of $M$, which is the tangent space of the embedded manifold. With $N \in \mathbb{R}^{n\times(n-m)}$ consisting of a set of orthonormal basis of the orthogonal complement of the tangent space, we can express the operator as $(I_n - NN^\top)$. Details are presented in Byrne and Girolami (2013) or Appendix A.2 of Liu, Zhu, and Song (2016). The advantage of using $N$ instead of $M$ is that it is independent of c.s. of $\mathcal{M}$, so we do not need to choose a set of c.s. covering $\mathcal{M}$ and conduct calculation in each c.s. Additionally, it is usually easier to find, and the expression with $N$ is more computationally economic. With this replacement, we have

$$
\begin{aligned}
f_1 + f_2 =& (\nabla \log p\sqrt{|G|})^\top (M G^{-1} M^\top) \nabla K \\
&+ \Big( (M^\top \nabla)^\top (G^{-1} M^\top) \Big) \nabla K \\
&+ \operatorname{tr}\Big( (\nabla\nabla^\top K)(M G^{-1} M^\top) \Big) \\
=& (\nabla \log p\sqrt{|G|})^\top (I_n - NN^\top) \nabla K \\
&+ \Big( (M^\top \nabla)^\top (G^{-1} M^\top) \Big) \nabla K \\
&+ \operatorname{tr}\Big( (\nabla\nabla^\top K) - (\nabla\nabla^\top K) NN^\top \Big) \\
=& (\nabla \log p\sqrt{|G|})^\top (I_n - NN^\top) \nabla K \\
&+ \Big( (M^\top \nabla)^\top (G^{-1} M^\top) \Big) \nabla K \\
&+ \nabla^\top \nabla K - \operatorname{tr}\Big( N^\top (\nabla\nabla^\top K) N \Big).
\end{aligned}
$$

Finally, $\hat{X} = \operatorname{grad} \hat{f} = g^{ij}\partial_i \hat{f} \partial_j = g^{ij} \frac{\partial y^\alpha}{\partial x^i} \partial_\alpha \hat{f} \frac{\partial y^\beta}{\partial x^j} \partial_\beta = MG^{-1}M\nabla\hat{f} = (I_n - NN^\top)\nabla\hat{f}$, which finishes the derivation.

Note that $M$ and $G$ depend on the choice of c.s. of $\mathcal{M}$. Note also that the parametric form of $\Xi^{-1}$ and $\xi^{-1}$ may not be unique (e.g. $\Xi^{-1}(y) = y$ and $\Xi^{-1}(y) = y + (1 - y^\top y)$ are both valid on $\Xi(\mathbb{S}^{n-1})$, but they give different gradients). Nevertheless, since $\hat{f}'$ is already a well-defined smooth function on $\mathcal{M}$ due to Eqn. (5), its expression in the embedded space w.r.t. any c.s. and any parametric form of $\Xi^{-1}$ and $\xi^{-1}$ should give the same result. We introduce $N$ in hope to explicitly express this independence, and we succeed for $\hat{X}'$ given $\hat{f}'$. For $\hat{f}'$, it is still a future work to make its expression explicitly independent of c.s. of $\mathcal{M}$ and parametric form of $\Xi^{-1}$ and $\xi^{-1}$.

**A6.2 For Hyperspheres (Proposition 7)** Let $\mathbb{S}^{n-1}$ be isometrically embedded in $\mathbb{R}^n$ via $\Xi : y \mapsto y$ the identity mapping. We select the c.s. $(U, \Phi)$ as the upper semi-hypersphere: $U := \{y \in \mathbb{R}^n | y^\top y = 1, y_n > 0\}$, $\Phi : y \mapsto (y_1, \ldots, y_{n-1})^\top \in \mathbb{R}^{n-1}$. Then we have $\Omega := \Phi(U) = \{x \in \mathbb{R}^{n-1} | x^\top x < 1\}$, and $\xi : \Omega \to \mathbb{R}^n, x \mapsto (x_1, \ldots, x_{n-1}, \sqrt{1 - x^\top x})^\top$. Furthermore,

$$M = \begin{pmatrix} I_{n-1} \\ -\frac{x^\top}{\sqrt{1-x^\top x}} \end{pmatrix},$$

and $G = I_{n-1} + \frac{xx^\top}{1-x^\top x}, G^{-1} = I_{n-1} - xx^\top, |G| = \frac{1}{1-x^\top x}$. The tangent space of $\Xi(\mathbb{S}^{n-1})$ at $y \in \mathbb{R}^n$ is a plane perpendicular to the direction of $y$, thus the orthogonal complement of the tangent space is the linear span of $y$, which indicates that $N = y$. Plugging in all these quantities in Eqn. (7), we can derive the result of Eqn. (8).

**A6.3 For the Product Manifold of Hyperspheres** To fit the inference task of Spherical Admixture Model (Reisinger et al. 2010) (SAM), we need to further specify the manifold as the product manifold of hyperspheres, $(\mathbb{S}^{n-1})^P$. Let $(\mathcal{M})^P$ be a general product manifold. For any point $A = (A_{(1)}, \ldots, A_{(P)}) \in (\mathcal{M})^P$, $(\bigotimes_{k=1}^P U_{(k)}, \bigotimes_{k=1}^P \{x_{(k)}^{i_{(k)}}\}_{i_{(k)}=1}^{n-1})$ is a local c.s., where each $(U_{(k)}, \{x_{(k)}^{i_{(k)}}\}_{i_{(k)}=1}^{n-1}$ is a local c.s. of $\mathcal{M}_{(k)}$ around $A_{(k)}$. In this c.s., $\{\partial_{(k),i_{(k)}} | k = 1, \ldots, P, i_{(k)} = 1, \ldots, n-1\}$ is the natural basis, and the Riemann structure in the tangent space is defined by direct product of inner product space: $g_{(k,\ell),i_{(k)},j_{(\ell)}} = \delta_{k\ell} g_{i_{(k)},j_{(\ell)}}$. By this construction, one can derive the expressions for the gradient of a smooth function $f \in \mathcal{C}^\infty((\mathcal{M})^P)$ and the divergence of a vector field $X = \sum_{k=1}^P X_{(k)}^{i_{(k)}} \partial_{(k),i_{(k)}} \in \mathcal{T}((\mathcal{M})^P)$: $\operatorname{grad} f = \sum_{k=1}^P g_{(k)}^{i_{(k)} j_{(k)}} \partial_{(k),i_{(k)}} f \partial_{(k),j_{(k)}}$, $\operatorname{div}(X) = \sum_{k=1}^P \Big(\partial_{(k),i_{(k)}} X_{(k)}^{i_{(k)}} + X_{(k)}^{i_{(k)}} \partial_{(k),i_{(k)}} \log \sqrt{|G_{(k)}|}\Big)$, as well as the Beltrami-Laplacian $\Delta f$.

For $y = (y_{(1)}, \ldots, y_{(P)}) \in (\mathbb{S}^{n-1})^P$ with each $y_{(k)} \in \mathbb{S}^{n-1}$, and kernel $K(y, y') = \prod_{k=1}^P K_{(k)}(y_{(k)}, y'_{(k)})$, we have the following result:

**Proposition 9.** $\hat{X}'_{(\ell)} = (I_d - y_{(\ell)} {y'_{(\ell)}}^\top) \nabla'_{(\ell)} \hat{f}'$,

$$\begin{aligned} \hat{f}' =& \mathbb{E}_q \Big[ K \sum_{k=1}^P \Big[ (\nabla_{(k)} \log p)^\top (\nabla_{(k)} \log K_{(k)}) + \\ & \nabla_{(k)}^\top \nabla_{(k)} \log K_{(k)} - y_{(k)}^\top \big(\nabla_{(k)} \nabla_{(k)}^\top K_{(k)}\big) y_{(k)} \\ & + \big\| \nabla_{(k)} \log K_{(k)} \big\|^2 - (y_{(k)}^\top \nabla_{(k)} \log K_{(k)})^2 \\ & - (y_{(k)}^\top \nabla_{(k)} \log p + n - 1) y_{(k)}^\top \nabla_{(k)} \log K_{(k)} \Big] \Big]. \end{aligned} \tag{12}$$

This proposition directly constructs the algorithm of RSVGD for the inference task of SAM, where each $y_{(k)}$ is a topic lying on a hypersphere.

**A7 Implementation of RSVGD for Bayesian Logistic Regression** From the model description in the main context, we have

$$\begin{aligned} \text{log-prior:} \quad & \log p_0(w) = -\frac{w^\top w}{2\alpha} + \text{const}, \\ \text{log-likelihood:} \quad & \log p(\{y_d\} | w, \{x_d\}) \\ & = \sum_{d=1}^D \Big( y_d w^\top x_d - \log(1 + e^{w^\top x_d}) \Big) + \text{const}, \\ \text{log-posterior:} \quad & \log p(w | \{y_d\}, \{x_d\}) = -\frac{w^\top w}{2\alpha} \\ & + \sum_{d=1}^D \Big( y_d w^\top x_d - \log(1 + e^{w^\top x_d}) \Big) + \text{const}. \end{aligned}$$

So we have the gradient of the target density

$$\nabla \log p(w | \{y_d\}, \{x_d\}) = -\frac{1}{\alpha} w + \sum_{d=1}^D \big( y_d - s(w^\top x_d) \big) x_d,$$

and the Riemann metric tensor

$$\begin{aligned} G(w) =& \mathcal{I}\big(p(\{y_d\} | w, \{x_d\})\big) - \nabla\nabla^\top \log p_0(w) \\ =& \mathbb{E}_{p(\{y_d\}|w,\{x_d\})} \big[ \big(\nabla \log p(\{y_d\} | w, \{x_d\})\big) \\ & \big(\nabla \log p(\{y_d\} | w, \{x_d\})\big)^\top \big] \\ & - \nabla\nabla^\top \log p_0(w) \\ =& \sum_{d=1}^D c_d x_d x_d^\top + \frac{1}{\alpha} I_m, \end{aligned}$$

where $\mathcal{I}(\cdot)$ is the Fisher information of a distribution, and $c_d = s(w^\top x_d)(1 - s(w^\top x_d))$. For $G^{-1}$, direct numerical inversion is applicable, with time complexity $\mathcal{O}(m^3)$. Another method, with time complexity $\mathcal{O}(m^2 D)$, can be derived by iteratively applying the Sherman-Morrison formula (Sherman and Morrison 1950):

$$\begin{aligned} G_d^{-1} &= G_{d-1}^{-1} - \frac{c_d (G_{d-1}^{-1} x_d)(G_{d-1}^{-1} x_d)^\top}{1 + c_d x_d^\top G_{d-1}^{-1} x_d}, \\ G^{-1} &= G_D^{-1}, G_0^{-1} = \alpha I_m. \end{aligned}$$

For small datasets, or for mini-batch of data, this implementation would be advantageous. But in our experiments we found that direct inversion is still more efficient.

To continue, we first note $\partial_i G := \partial_{w_i} G = \sum_{d=1}^{D} f_d x_{di} x_d x_d^\top$, where $f_d = \frac{1-e^{w^\top x_d}}{1+e^{w^\top x_d}} c_d$. Note also that $\partial_i G_{jk} = \sum_{d=1}^{D} f_d x_{di} x_{dj} x_{dk}$, so the indices $i, j, k$ are completely permutable. Particularly, $\partial_i G_{jk} = \partial_j G_{ik}$. For the gradient of the log-determinant,

$$\partial_i \log |G(w)| = \mathrm{tr}(G^{-1} \partial_i G) = \sum_{d=1}^{D} f_d (x_d^\top G^{-1} x_d) x_{di},$$

and for the gradient of the inverse matrix,

$$\begin{aligned}
&\sum_{j=1}^{m} \partial_j G_{ij}^{-1}(w) = -G_{(i,:)}^{-1} \sum_{j=1}^{m} (\partial_j G) G_{(:,j)}^{-1} \\
&= - \sum_{k=1}^{m} G_{(i,k)}^{-1} \sum_{j=1}^{m} \sum_{\ell=1}^{m} (\partial_j G)_{(k,\ell)} G_{(\ell,j)}^{-1} \\
&= - \sum_{k=1}^{m} G_{(i,k)}^{-1} \sum_{j=1}^{m} \sum_{\ell=1}^{m} (\partial_k G)_{(j,\ell)} G_{(\ell,j)}^{-1} \\
&= - \sum_{k=1}^{m} G_{(i,k)}^{-1} \mathrm{tr}\big( (\partial_k G) G^{-1} \big) \\
&= - G_{(i,:)}^{-1} \nabla \log |G(w)|.
\end{aligned}$$

Now all the quantities needed for RSVGD (Eqn. (6)) are derived.

## References


Abraham, R.; Marsden, J. E.; and Ratiu, T. 2012. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media.

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

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.

Liu, C.; Zhu, J.; and Song, Y. 2016. Stochastic gradient geodesic mcmc methods. In *Advances In Neural Information Processing Systems*, 3009–3017.

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.

Sherman, J., and Morrison, W. J. 1950. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics* 21(1):124–127.

Zhou, D.-X. 2008. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics* 220(1):456–463.