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

Abstract

We develop Riemannian Stein Variational Gradient Descent (RSVGD), a Bayesian inference method that generalizes Stein Variational Gradient Descent (SVGD) to Riemann manifold. The benefits are two-folds: (i) for inference tasks in Euclidean spaces, RSVGD has the advantage over SVGD of utilizing information geometry, and (ii) for inference tasks on Riemann manifolds, RSVGD brings the unique advantages of SVGD to the Riemannian world. To appropriately transfer to Riemann manifolds, we conceive novel and non-trivial techniques for RSVGD, which are required by the intrinsically different characteristics of general Riemann manifolds from Euclidean spaces. We also discover Riemannian Stein's Identity and Riemannian Kernelized Stein Discrepancy. Experimental results show the advantages over SVGD of exploring distribution geometry and the advantages of particle-efficiency, iteration-effectiveness and approximation flexibility over other inference methods on Riemann manifolds.

Introduction

Bayesian inference is the central task for learning a Bayesian model to extract knowledge from data. The task is to estimate the posterior distribution of latent variables of the model given observed data. It has been in the focus of machine learning for decades, with quite a lot of methods emerging. Variational inference methods (VIs) aim to approximate the posterior by a tractable variational distribution. Traditional VIs typically use a statistical model, usually a parametric distribution family, as the variational distribution, and we call them model-based VIs (M-VIs). They cast the inference problem as an optimization problem, which can be efficiently solved by various techniques. However, due to the restricted coverage of the chosen distribution family (e.g. the mean-field form), there would be a gap blocking the approximation from getting any closer. Monte Carlo methods (MCs) estimate the posterior by directly drawing samples from it. Asymptotically accurate as they are, their performance for finite samples is not guaranteed and usually take effect slowly, especially for the widely applicable thus commonly used Markov Chain MCs (MCMCs), due to the positive autocorrelation of their samples.

*corresponding author.

Recently, a set of particle-based VIs (P-VIs) have been proposed. P-VIs use a certain number of samples, or particles, to represent the variational distribution, and update the particles by solving an optimization problem. Similar to MCs, this non-parametric particle form gives them great flexibility to reduce the gap of M-VIs, and beyond MCs, the optimization-based update rule makes them effective in iteration: every iteration is guaranteed to make progress. Although there are convergence analyses for some particular MCMCs from non-stationary to stationary, the principle of a general MCMC only guarantees that the sampler will remain in the stationary distribution. Moreover, MCs usually require a large sample size to take effect, while P-VIs achieve similar performance with much fewer particles, since their principle aims at finite sample performance. This particle-efficiency would save the storage of the inference result, and reduce the time for tasks afterwards such as test and prediction. Table 1 presents a comparison of the three kinds of inference methods.

Stein variational gradient descent (SVGD) (Liu and Wang 2016) is an outstanding example of P-VIs. SVGD updates particles by imposing a continuous-time dynamics on them that leads the variational distribution to evolve towards the posterior. Although the dynamics is restricted in a kernel-related space for tractability, the theory of SVGD makes no assumption on the variational distribution, indicating the best flexibility. SVGD has been applied to develop advanced inference methods (Wang and Liu 2016; Pu et al. 2017) as well as reinforcement learning (Liu et al. 2017; Haarnoja et al. 2017). Other instances of P-VIs include normalizing flows (NF) (Rezende and Mohamed 2015) and particle mirror descent (PMD) (Dai et al. 2016). NF uses a series of invertible transformations to adjust particles from a simple tractable distribution to fit the posterior. However, the invertibility requirement restricts its flexibility. PMD adopts the mirror descent method to formulate an optimization problem and uses a weighted kernel density estimator for the variational distribution, which is still a restricting assumption.

Another issue of Bayesian inference is the collaboration with Riemann manifold. This consideration bears its importance in two ways: (i) the posterior of some models itself is a distribution on a given Riemann manifold, e.g. the spherical admixture model (SAM) (Reisinger et al. 2010) has its posterior on hyperspheres; (ii) latent variables in

Table 1: A comparison of three kinds of inference methods.

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

a Bayesian model is a natural coordinate system (c.s.) of the Riemann manifold of likelihood distribution, so we can also conduct Bayesian inference on the distribution manifold with improved efficiency with the help of information geometry (Amari and Nagaoka 2007; Amari 2016). Much progress has been made recently for both Riemannian considerations. In the spirit of (i), Bonnabel (2013) and Zhang, Reddi, and Sra (2016) develop scalable and stable optimization methods on Riemann manifold to enhance inference for M-VIs, and Brubaker, Salzmann, and Urtasun (2012), Byrne and Girolami (2013) and Liu, Zhu, and Song (2016) develop efficient and scalable MCMCs on Riemann manifold. In the spirit of (ii), Hoffman et al. (2013) use natural gradient for M-VI, Girolami and Calderhead (2011) and Ma, Chen, and Fox (2015) develop efficient MCMCs with Riemann structure, and Li et al. (2016) apply the Riemannian MCMCs to the inference of Bayesian neural network. Little work has been done to enhance P-VIs with Riemann structure. Gemici, Rezende, and Mohamed (2016) attempt to generalize NF to Riemann manifold, but their method cannot be used for manifolds with no global c.s., such as hypersphere. Although the method can be implemented in the almost-global c.s., unbounded distortion near the boundary of the c.s. would unavoidably occur, which would cause numerical instability.

In this work, we develop Riemannian Stein Variational Gradient Descent (RSVGD), the extension of SVGD to Riemann manifold. Our method can be applied for both approximating the posterior on Riemann manifold (i), and efficient inference by exploring the Riemann structure of distributions (ii). RSVGD inherits the significant advantages of SVGD, bringing the benefits to the field of inference on Riemann manifold, such as particle-efficiency and zero variational assumption. Technically, it is highly non-trivial to extent the idea to Riemann manifold, as many subtle properties of Riemann manifold must be carefully considered, which may lead to completely different treatment. We first review SVGD as an evolution under a dynamics of a flow and generalize the deduction to Riemann manifold. Then we solve for the optimal dynamics by a novel method, where the treatment of SVGD fails due to the intrinsically different properties of general Riemann manifold from Euclidean space. The expression in the embedded space is also derived for application to manifolds with no global c.s. like hyperspheres, which does not require choosing a c.s. and introduce no numerical problems. As side products, we also develop Riemannian Stein's identity and Riemannian kernelized Stein discrepancy, as an extension of the corresponding concepts. Finally, we apply RSVGD to the troublesome inference task of SAM, with its unique advantages validated in experiments.

Preliminaries

Riemann Manifolds

We briefly introduce basic concepts of Riemann manifold. For more details please refer to common textbooks e.g. Do Carmo (1992); Abraham, Marsden, and Ratiu (2012).

Basics Denote $\mathcal{M}$ an m-dimensional Riemann manifold. By definition at every point $A \in \mathcal{M}$ there exists a local coordinate system (c.s.) (U, Φ), where $U \subset \mathcal{M}$ is open and contains A, and $\Phi : U \to \mathbb{R}^m$ a homeomorphism between U and $\Phi(U)$. Denote $\mathcal{C}^\infty_A$ and $\mathcal{C}^\infty(\mathcal{M})$ as the set of functions $\mathcal{M} \to \mathbb{R}$ that are smooth around A and all over $\mathcal{M}$, respectively. A tangent vector v at A is a linear functional $\mathcal{C}^\infty_A \to \mathbb{R}$ that satisfies $v[(fg)(\cdot)] = f(A)v[g(\cdot)] + g(A)v[f(\cdot)], \forall f, g \in \mathcal{C}^\infty_A$. Intuitively $v[f]$ is the directional derivative of f at A along the direction of v. All such v forms an m-dimensional linear space $T_A\mathcal{M}$, called the tangent space at A. Its natural basis $\{\partial_i\}_{i=1}^m$ under (U, Φ) is defined as $\partial_i(A)[f] := \frac{\partial(f\circ\Phi^{-1})}{\partial x^i}(x^1, \dots, x^m)|_{\Phi(A)}$ (also denoted as $\partial_i f|_A$). We can then express v in component: $v = v^i\partial_i$, where we adopt Einstein's convention that duplicated subscript and superscript are summed out. A vector field X on $\mathcal{M}$ specifies at every $A \in \mathcal{M}$ a tangent vector $X(A) \in T_A\mathcal{M}$ smoothly with respect to (w.r.t.) A. Denote $\mathcal{T}(\mathcal{M})$ as the set of all such X. A Riemann structure is equipped to $\mathcal{M}$ if $\forall A \in \mathcal{M}$, $T_A\mathcal{M}$ is endowed with an inner product $g_A(\cdot, \cdot)$ (and g_A is smooth in A). In (U, Φ), for $u = u^i\partial_i, v = v^j\partial_j$, $g_A(u, v) = g_{ij}(A)u^iv^j$, where $g_{ij}(A) = g_A(\partial_i, \partial_j)$.

An object called Riemann volume form μ_g can be used to define a measure on $\mathcal{M}$ via integral: the measure of a compact region $U \subset \mathcal{M}$ is defined as $\int_U \mu_g$. Thus for any probability distribution absolutely continuous w.r.t. μ_g, we can define its probability density function (p.d.f.) p w.r.t. μ_g: $\mathrm{Prob}(U) = \int_U p\mu_g$. In the sequel, we require distributions such that their p.d.f. are smooth functions on $\mathcal{M}$, and we would say "distribution with p.d.f. p" as "distribution p".

Flows, Dynamics and Evolving Distributions These concepts constitute the fundamental idea of SVGD and our RSVGD. The notion of flow arises from the following fact: for a vector field X and a fixed point $A \in \mathcal{M}$, there exist a subset $U \subset \mathcal{M}$ containing A, and a one-parameter transformation $F_{(\cdot)}(\cdot) : (-\delta, \delta) \times U \to \mathcal{M}$ where $\delta \in \mathbb{R}$, such that $F_0(\cdot)$ is the identity map on U, and for $B \in U$ and $t_0 \in (-\delta, \delta)$, $\frac{\mathrm{d}}{\mathrm{d}t} f(F_t(B))|_{t=t_0} = X(F_{t_0}(B))[f], \forall f \in \mathcal{C}^\infty_{F_{t_0}(B)}$ (Do Carmo, 1992, Page 28). We call $F_{(\cdot)}(\cdot)$ the local flow of X around A. Under some condition, e.g. X has compact

support, there exists a local flow that is global (i.e. $U = \mathcal{M}$), which is called the **flow** of X.

We refer a **dynamics** here as a rule governing the motion on $\mathcal{M}$ over time t. Specifically, a dynamics gives the position $A(t)$ at any future time for any given initial position $A(0)$. If $t \in \mathbb{R}$ and $A(t)$ is smooth for all $A(0) \in \mathcal{M}$, we call it a continuous-time dynamics. Obviously, a flow can be used to define a continuous-time dynamics: $A(t) = F_t(A_0), A(0) = A_0$. Due to the correspondence between a vector field and a flow, we can also define a dynamics by a vector field X, and denote it as $\frac{\mathrm{d}A}{\mathrm{d}t} = X$. Thus we would say "the dynamics defined by the flow of X" as "dynamics X" in the following.

Let a random variable (r.v.) obeying some distribution p move under dynamics X from time 0 to t, which acts as a transformation on the r.v. Then the distribution of the transformed r.v. also evolves along time t, and we denote it as p_t, and call it an **evolving distribution** under dynamics X. Suppose there is a set of particles $\{A^{(i)}\}_{i=1}^N$ that distributes as p. Let each particle move individually under dynamics X for time t, then the new set of particles distributes as p_t.

Reynolds Transport Theorem

Reynolds transport theorem helps us to relate an evolving distribution to the corresponding dynamics. It is a generalization of the rule of differentiation under integral, and is the foundation of fluid mechanics. Let $X \in \mathcal{T}(\mathcal{M})$ and $F_{(\cdot)}(\cdot)$ be its flow. For smooth $f_{(\cdot)}(\cdot) : \mathbb{R} \times \mathcal{M} \to \mathbb{R}$ and any open subset $U \subset \mathcal{M}$, the theorem states that

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

where $\mathrm{div} : \mathcal{T}(\mathcal{M}) \to \mathcal{C}^\infty(\mathcal{M})$ is the divergence of a vector field. In any local c.s., $\mathrm{div}(X) = \partial_i(\sqrt{|G|}X^i)/\sqrt{|G|}$, where G, commonly called Riemann metric tensor, is the $m \times m$ matrix comprised of g_{ij} in that c.s., and $|G|$ is its determinant. More details can be found in e.g. (Romano, 2007, Page 164); (Frankel, 2011, Page 142); (Abraham, Marsden, and Ratiu, 2012, Page 469).

Stein Variational Gradient Descent (SVGD)

We review SVGD (Liu and Wang 2016) from a perspective inspiring our generalization work to Riemann manifold. SVGD is a particle-based variational inference method. It updates particles by applying an appropriate dynamics on them so that the distribution of the particles, an evolving distribution, approaches (in the KL-divergence sense) the target distribution. Note that in this case the manifold is the most common Euclidean space $\mathcal{M} = \mathbb{R}^m$. Denote the evolving distribution of the particles under dynamics X as q_t, and let p be the target distribution. The first key result of SVGD is

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p) = \mathbb{E}_{q_t}[X^\top \nabla \log p + \nabla^\top X], \qquad (1)$$

which measures the rate of q_t to approach p. As our desiderata is to maximize $-\mathrm{KL}(q_t||p)$, a desirable dynamics should maximize this approaching rate. This is analogous to the process of finding gradient when we want to maximize a function f: $\nabla f(x) = \alpha_0 v_0, (\alpha_0, v_0) = \max_{\|v\|=1} f'_v(x)$, where $f'_v(x)$ is the directional derivative along direction v. Similarly, we call $-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)$ the **directional derivative** along vector field X, and (Liu and Wang 2016) claims that the direction and magnitude of **functional gradient** of $-\mathrm{KL}(q_t||p)$ coincides with the maximizer and maximum of the directional derivative, and the dynamics of the functional gradient (itself a vector field) is the desired one. Simulating this dynamics and repeating the procedure updates the particles to be more and more representative for the target distribution.

The above result reveals the relation between the directional derivative and dynamics X. To find the functional gradient, we get the next task of solving the maximization problem. Note that in Euclidean case, the tangent space at any point is isometrically isomorphic to $\mathbb{R}^m$, so X can be described as m smooth functions on $\mathbb{R}^m$. Liu and Wang (2016) take X from the product space $\mathcal{H}_K^m$ of the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ of some smooth kernel K on $\mathbb{R}^m$. $\mathcal{H}_K$ is a Hilbert space of some smooth functions on $\mathbb{R}^m$.Its key property is that for any $A \in \mathbb{R}^m$, $K(A, \cdot) \in \mathcal{H}_K$, and $\langle f(\cdot), K(A, \cdot)\rangle_{\mathcal{H}_K} = f(A), \forall f \in \mathcal{H}_K$. In $\mathcal{H}_K^m$ the maximization problem can be solved in closed form, which gives the functional gradient:

$$X^*(\cdot) = \mathbb{E}_{q_t(A)}[K(A, \cdot)\nabla \log p(A) + \nabla K(A, \cdot)]. \qquad (2)$$

Note that the sample distribution q_t appears only in the form of expectation, which can be estimated by merely samples from q_t. This releases the assumption on the explicit form of q_t thus gives it great flexibility. It is a benefit of using KL-divergence to measure the difference between two distributions.

Riemannian SVGD

We now present Riemannian Stein Variational Gradient Descent (RSVGD), which has many non-trivial considerations beyond SVGD and requires novel treatments. We first derive the Riemannian counterpart of the directional derivative, then conceive a novel technique to find the functional gradient, in which case the SVGD technique fails. Finally we express RSVGD in the embedded space of the manifold and give an instance for hyperspheres, which is directly used for our application.

Derivation of the Directional Derivative

Now $\mathcal{M}$ is a general Riemann manifold. We first derive a useful tool for deriving the directional derivative.

Lemma 1 (Continuity Equation on Riemann Manifold)**.** *Let p_t be the evolving distribution under dynamics X, where $X \in \mathcal{T}(\mathcal{M})$. Then*

$$\frac{\partial p_t}{\partial t} = -\mathrm{div}(p_t X) = -X[p_t] - p_t \mathrm{div}(X). \qquad (3)$$

See Appendix[1] A1 for derivation, which involves Reynolds transport theorem.

[1]Appendix available at `http://ml.cs.tsinghua.edu.cn/~changliu/rsvgd/Liu-Zhu-appendix.pdf`

We now present our first key result, the directional derivative of $-\mathrm{KL}(q_t||p)$[2] w.r.t. vector field X.

Theorem 2 (Directional Derivative)**.** *Let q_t be the evolving distribution under dynamics X, and p a fixed distribution. Then the directional derivative is*

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

See Appendix A3 for proof. Theorem 2 is an extension of the key result of SVGD Eqn. (1) to general Riemann manifold case. In SVGD terms, we call $\mathcal{A}_p X = X[\log p] + \mathrm{div}(X)$ the generalized Stein's operator.

Appendix A4 further discusses in detail the corresponding condition for Stein's identity to hold in this case. Stein's identity refers to the equality $-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t||p)\big|_{t=t_0} = 0$ for any t_0 such that $q_{t_0} = p$. Stein class is the set of X that makes Stein's identity hold. Stein's identity indicates that when q_t reaches its optimal configuration p, the directional derivative along any direction X in the Stein class is zero, in analogy to the well-known zero gradient condition in optimization. These concepts play an important rule when using the functional gradient to measure the distance between two distributions, i.e. the Stein discrepancy (Liu, Lee, and Jordan 2016).

Derivation of the Functional Gradient

Now that we have the directional derivative expressed in terms of X, we get the maximization problem to find the functional gradient:

$$\max_{X\in\mathfrak{X}, \|X\|_{\mathfrak{X}}=1} \mathcal{J}(X) := \mathbb{E}_q\big[X[\log p] + \mathrm{div}(X)\big], \qquad (4)$$

where we omit the subscript of q_t since it is fixed when optimizing w.r.t. X. Ideally $\mathfrak{X}$ should be $\mathcal{T}(\mathcal{M})$, but for a tractable solution we may restrict $\mathfrak{X}$ to some subset of $\mathcal{T}(\mathcal{M})$ which is at least a normed space. Once we get the maximizer of $\mathcal{J}(X)$, denoted by X^o, we have the functional gradient $X^* = \mathcal{J}(X^o)X^o \in \mathcal{T}(\mathcal{M})$.

Requirements Before we choose an appropriate $\mathfrak{X}$, we first list three requirements on a reasonable X^* (or equivalently on X^o since they differ only in scale). The first two requirements arise from special properties of general Riemann manifolds, which are so different from Euclidean spaces that make SVGD technique fail.

- 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.

We require R1 since all the above deductions are based on vector fields. In the Euclidean case, any set of m smooth functions $(f^1, \ldots, f^m)$ satisfies R1. But in general Riemann manifold, it is not enough that all the m components of a vector field is smooth in any coordinate system (c.s.). For example, due to the hairy ball theorem (Abraham, Marsden, and Ratiu, 2012, Theorem 8.5.13), vector fields on an even-dimensional hypersphere must have one zero-vector-valued point (critical point), which goes beyond the above condition. This disables the idea to use the SVGD technique in the coordinate space of a manifold.

R2 is required to avoid ambiguity or arbitrariness of the solution. Coordinate invariance is a key concept in the area of differential manifold. By definition the most basic way to access a general manifold is via c.s., so we can define an object on the manifold by its expression in c.s. If the form of the expression in any c.s. is the same, or equivalently the expression in a certain form refers to the same object in any c.s., we say that the form of the object is coordinate invariant. For intersecting c.s. (U, Φ) and (V, Ψ), if the expression in each c.s. with a same form gives different results, then the object is ambiguous on $U \cap V$. One may argue that it is possible to use one certain c.s. to uniquely define the object, whatever the form in that c.s., but firstly for manifolds without global c.s., e.g. hyperspheres, we cannot define an object globally on the manifold in this manner, and secondly for manifolds that have global c.s., the choice of the certain c.s. may be arbitrary, since all the c.s. are equivalent to each other: there is no objective reason for choosing one specific c.s. other than other c.s.

Expression $\mathrm{grad}\, f = g^{ij}\partial_i f \partial_j$ is coordinate invariant, while $\nabla f = \sum_i \partial_i f \partial_i$ is not. Suppose the above two expressions are written in c.s. $(U, \{x^i\}_{i=1}^m)$. Let $(V, \{y^a\}_{a=1}^m)$ be another c.s. with $\tilde{\partial}_a$ and $\tilde{g}^{ab}$ the corresponding objects on it. On $U \cap V$, $\mathrm{grad}\, f = \tilde{g}^{ab}\tilde{\partial}_a f \tilde{\partial}_b$, but $\nabla f = (\sum_i \frac{\partial y^a}{\partial x^i}\frac{\partial y^b}{\partial x^i})\tilde{\partial}_a f \tilde{\partial}_b$ while we expect $\sum_a \tilde{\partial}_a f \tilde{\partial}_a$, which do not match for general c.s. Consequently, the functional gradient of SVGD Eqn. (2) is not coordinate invariant. Again, the SVGD technique does not meet our demand here.

R3 cuts off the computational burden for optimizing w.r.t. X, and releases the assumption on the form of q so as to adopt the benefit of SVGD of flexibility.

Before presenting our solution, we would emphasize the difficulty by listing some possible attempts that actually fail. First note that $\mathfrak{X}$ is a subspace of $\mathcal{T}(\mathcal{M})$, which is a linear space on both $\mathbb{R}$ and $\mathcal{C}^\infty(\mathcal{M})$. But in either case $\mathcal{T}(\mathcal{M})$ is infinite dimensional and it is intractable to express vector fields in it. Secondly, as the treatment of SVGD, one may consider expressing vector fields component-wise. This idea specifies a tangent vector at each point, thus would easily violate R1 and R2, due to the intrinsically different characteristic of general Riemann manifolds from Euclidean space: tangent spaces at different points are not the same. R1 and R2 focus on global properties, one has to link tangent spaces at different points. Taking a c.s. could uniformly express the tangent vectors on the c.s., but this is not enough as stated above. One may also consider transforming tangent vectors at different points to one certain pivot point by parallel transport, but the arbitrariness of the pivot point would dissatisfy R2. The third possibility is to view a vector field as a map from $\mathcal{M}$ to $T\mathcal{M}$, the tangent bundle of $\mathcal{M}$, but $T\mathcal{M}$ is generally only a manifold but not a linear space, so we cannot apply theories of learning vector-valued functions (e.g. Micchelli and Pontil (2005)).

[2] Appendix A2 presents the well-definedness of the KL-divergence on Riemann manifold $\mathcal{M}$.

Solution We first present our solution, then check the above requirements. Note that theories of kernel and reproducing kernel Hilbert space (RKHS) (see e.g. Steinwart and Christmann (2008), Chapter 4) also apply to manifold case. Let $K : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ be a smooth kernel on $\mathcal{M}$ and $\mathcal{H}_K$ its RKHS. We require K such that zero function is the only constant function in $\mathcal{H}_K$. Such kernels include the commonly used Gaussian kernel (Steinwart and Christmann, 2008, Corollary 4.44).

We choose $\mathfrak{X} = \{\operatorname{grad} f | f \in \mathcal{H}_K\}$, where $\operatorname{grad} : \mathcal{C}^\infty(\mathcal{M}) \to \mathcal{T}(\mathcal{M})$ is the gradient of a smooth function, which is a valid vector field. In any c.s., $\operatorname{grad} f = g^{ij}\partial_i f \partial_j$, where g^{ij} is the (i, j)-th entry of G^{-1}. The following result indicates that $\mathfrak{X}$ can be made into an inner product space.

Lemma 3. *With a proper inner product, $\mathfrak{X}$ is isometrically isomorphic to $\mathcal{H}_K$, thus a Hilbert space.*

Proof. Define $\iota : \mathcal{H}_K \to \mathfrak{X}, f \mapsto \operatorname{grad} f$, which is linear: $\forall \alpha \in \mathbb{R}, f, h \in \mathcal{H}_K, \iota(\alpha f + h) = \alpha \operatorname{grad} f + \operatorname{grad} h = \alpha\iota(f) + \iota(h)$. For any $f, h \in \mathcal{H}_K$ that satisfy $\iota(f) = \iota(h)$, we have $\operatorname{grad}(f - h) = 0$, $f - h = \operatorname{const}_{\mathcal{H}_K} = 0_{\mathcal{H}_K}$, where the last equality holds for our requirement on K. So $f = h$ thus ι is injective. By definition of $\mathfrak{X}$, ι is surjective, so it is an isomorphism between $\mathfrak{X}$ and $\mathcal{H}_K$.

Define $\langle \cdot, \cdot \rangle_{\mathfrak{X}} : \langle X, Y \rangle_{\mathfrak{X}} = \langle \iota^{-1}(X), \iota^{-1}(Y) \rangle_{\mathcal{H}_K}$, $\forall X, Y \in \mathfrak{X}$. By the linearity of ι and that $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ is an inner product, one can easily verify that $\langle \cdot, \cdot \rangle_{\mathfrak{X}}$ is an inner product on $\mathfrak{X}$, and that ι is an isometric isomorphism. □

Next we present our second key result, that the objective $\mathcal{J}(X)$ can be cast as an inner product in $\mathfrak{X}$.

Theorem 4. *For $(\mathfrak{X}, \langle \cdot, \cdot \rangle_{\mathfrak{X}})$ defined above and $\mathcal{J}$ the objective in Eqn. (4), we have $\mathcal{J}(X) = \langle X, \hat{X} \rangle_{\mathfrak{X}}$, where*

$$\hat{X} = \operatorname{grad} \hat{f},$$

$$\hat{f}(\cdot) = \mathbb{E}_{q(A)}\Big[\big(\operatorname{grad} K(A, \cdot)\big)[\log p(A)] + \Delta K(A, \cdot)\Big], \quad (5)$$

and $\Delta f := \operatorname{div}(\operatorname{grad} f)$ is the Beltrami-Laplace operator.

Proof (sketch). Use $f = \iota^{-1}(X) \in \mathcal{H}_K$ to express the vector field $X = \operatorname{grad} f$. Expand $\mathcal{J}$ in some c.s., then cast partial derivatives into the form of inner product in $\mathcal{H}_K$ based on the results of Zhou (2008). Rearranging terms by the linearity of inner product and expectation, $\mathcal{J}$ can be written as an inner product in $\mathcal{H}_K$, thus an inner product in $\mathfrak{X}$ due to their isomorphism. See Appendix A5 for details. □

In the form of an inner product, the maximization problem (4) solves as: $X^o = \hat{X}/\|\hat{X}\|_{\mathfrak{X}}$. Then the functional gradient is $X^* = \mathcal{J}(X^o)X^o = \frac{\langle \hat{X}, \hat{X} \rangle_{\mathfrak{X}}}{\|\hat{X}\|_{\mathfrak{X}}} \cdot \frac{\hat{X}}{\|\hat{X}\|_{\mathfrak{X}}} = \hat{X}$.

Corollary 5. *(Functional Gradient) For $(\mathfrak{X}, \langle \cdot, \cdot \rangle_{\mathfrak{X}})$ taken as above, the functional gradient is $X^* = \hat{X}$.*

Now we check our solution for the three requirements. Since grad, div, Δ, expectation, and the action of tangent vector on smooth function $v[f]$ are all coordinate invariant objects, $\hat{f}$ is coordinate invariant thus a valid smooth function on $\mathcal{M}$. So its gradient $\hat{X}$ is a valid vector field (R1) and also coordinate invariant (R2). R3 is obvious from Eqn. (5).

As a bonus, we present the optimal value of the objective:

$$\mathcal{J}(\hat{X}) = \mathbb{E}_q \mathbb{E}_{q'}\Big[\big(\operatorname{grad}' \log p'\big)\big[(\operatorname{grad} \log p)[K]\big] + \Delta' \Delta K + (\operatorname{grad}' \log p')[\Delta K] + (\operatorname{grad} \log p)[\Delta' K]\Big],$$

where $K = K(A, A')$, and notations with prime "$'$" take A' as argument and others take A. We call it Riemannian Kernelized Stein Discrepancy (RKSD) between distributions q and p, a generalization of Kernelized Stein Discrepancy (Liu, Lee, and Jordan 2016; Chwialkowski, Strathmann, and Gretton 2016).

After deriving the functional gradient $X^* = \hat{X}$, we can simulate its dynamics in any c.s. (denoted as $(U, \{x^i\}_{i=1}^m)$) by $x^i(t + \varepsilon) = x^i(t) + \varepsilon \hat{X}^i(x(t))$ for each component i, which is a 1st-order approximation of the flow of $\hat{X}$, where

$$\hat{X}^i(A') = g'^{ij}\partial'_j \mathbb{E}_q\Big[\big(g^{ab}\partial_a \log(p\sqrt{|G|}) + \partial_a g^{ab}\big)\partial_b K + g^{ab}\partial_a \partial_b K\Big]. \quad (6)$$

This is the update rule of particles for inference tasks on Euclidean space, where the expectation is estimated by averaging over current particles, and the Riemann metric tensor g_{ij} is taken as the subtraction of the Fisher information of likelihood with the Hessian of prior p.d.f., as adopted by Girolami and Calderhead (2011). Note that g^{ij} is the entry of the inverse matrix.

Expression in the Embedded Space

From the above discussion, we can simulate the optimal dynamics in c.s. But it is not always the most convenient approach. For some manifolds, like hyperspheres $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n | \|x\|_2 = 1\}$ or Stiefel manifold $\mathbb{M}_{m,n} := \{M \in \mathbb{R}^{m \times n} | M^\top M = I_m\}$ (James 1976), on one hand, they have no global c.s. so we have to change c.s. constantly while simulating, and we have to compute g^{ij}, $|G|$ and $\partial_i K$ in each c.s. It is even hard to find a c.s. of a Stiefel manifold. On the other hand, such manifolds are defined as a subset of some Euclidean space, which is a natural embedded space. This motivates us to express the dynamics of the functional gradient in the embedded space and simulate in it.

Formally, an embedding of a manifold $\mathcal{M}$ is a smooth injection $\Xi : \mathcal{M} \to \mathbb{R}^n$ for some $n \geq m$. For a Riemann manifold, Ξ is isometric if $g_{ij} = \sum_{\alpha=1}^n \frac{\partial y^\alpha}{\partial x^i}\frac{\partial y^\alpha}{\partial x^j}$ in any c.s. (U, Φ), where $\frac{\partial y^\alpha}{\partial x^i}$ is for $y = \xi(x)$ with $\xi := \Xi \circ \Phi^{-1}$. The Hausdorff measure on $\Xi(\mathcal{M}) \subset \mathbb{R}^n$ induces a measure on $\mathcal{M}$, w.r.t. which we have p.d.f. p_H. We recognize that for isometric embedding, $p = p_H$.

Since Ξ is injective, $\Xi^{-1} : \Xi(\mathcal{M}) \to \mathcal{M}$ can be well-defined, so can ξ^{-1} for any c.s. The following proposition gives a general expression of the functional gradient in an isometrically embedded space.

Proposition 6. *Let all the symbols take argument in the isometrically embedded space $\mathbb{R}^n$ (with orthonormal basis $\{y^\alpha\}_{\alpha=1}^n$)) via composed with Ξ^{-1} or ξ^{-1}. We have*

$$\hat{X}' = (I_n - N'N'^\top)\nabla'\hat{f}',$$

$$\hat{f}' = \mathbb{E}_q\Big[\Big(\nabla \log\big(p\sqrt{|G|}\big)\Big)^\top \Big(I_n - NN^\top\Big)(\nabla K) + \nabla^\top\nabla K - \mathrm{tr}\Big(N^\top(\nabla\nabla^\top K)N\Big) + \Big((M^\top\nabla)^\top(G^{-1}M^\top)\Big)(\nabla K)\Big], \quad (7)$$

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

See Appendix A6.1 for derivation. Note that N does not depend on the choice of c.s. of $\mathcal{M}$ while M and G do, but the final result does not. Simulating the dynamics is quite different from the coordinate space case due to the constraint of $\Xi(\mathcal{M})$. A 1st-order approximation of the flow is

$$y(t+\varepsilon) = \mathrm{Exp}_{y(t)}\big(\varepsilon\hat{X}(y(t))\big),$$

where Exp_A is the exponential map at $A \in \mathcal{M}$, which maps $v \in T_A\mathcal{M}$ to the end point of moving A along the geodesic determined by v for unit time (parameter of the geodesic). In $\mathbb{R}^n$, A is moved along the straight line in the direction of v for length $\|v\|$, and in $\mathbb{S}^{n-1}$, A is moved along the great circle (orthodrome) tangent to v at A for length $\|v\|$.

Instantiation for Hyperspheres We demonstrate the above result with the instance of $\mathcal{M} = \mathbb{S}^{n-1}$.

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

$$\hat{f}' = \mathbb{E}_q\Big[(\nabla \log p)^\top(\nabla K) + \nabla^\top\nabla K - y^\top\big(\nabla\nabla^\top K\big)y - (y^\top\nabla \log p + n - 1)y^\top\nabla K\Big]. \quad (8)$$

Note that the form of the expression does not depend on any c.s. of $\mathbb{S}^{n-1}$. The exponential map on $\mathbb{S}^{n-1}$ is given by

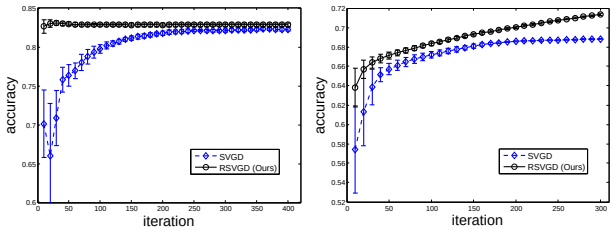
$$\mathrm{Exp}_y(v) = y\cos(\|v\|) + (v/\|v\|)\sin(\|v\|).$$

See Appendix A6.2 for more details. Appendix A6.3 further provides the expression for the product manifold of hyperspheres $(\mathbb{S}^{n-1})^P$, which is the specific manifold on which the inference task of Spherical Admixture Model is defined.

Experiments

We find the naïve mini-batch implementation of RSVGD does not perform well in experiments, which may require further investigation on the impact of gradient noise on the dynamics. So we only focus on the full-batch performance of RSVGD in experiments[3].

[3]Codes and data available at `http://ml.cs.tsinghua.edu.cn/~changliu/rsvgd/`



(a) On Splice19 dataset (b) On Covertype dataset

Figure 1: 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 (Liu and Wang 2016).

Bayesian Logistic Regression

We test the empirical performance of RSVGD on Euclidean space (Eqn. 6) by the inference task of Bayesian logistic regression (BLR). BLR generates latent variable from prior $w \sim \mathcal{N}(0, \alpha I_m)$, and for each datum x_d, draws its label from Bernoulli distribution $y_d \sim \mathrm{Bern}(s(w^\top x_d))$, where $s(x) = 1/(1+e^{-x})$. The inference task is to estimate the posterior $p(w|\{x_d\}, \{y_d\})$. Note that for this model, as is shown in Appendix A7, all the quantities involving Riemann metric tensor G has a closed form, except G^{-1}, which can be solved by either direct numerical inversion, or an iterative method over the dataset.

Kernel As stated before, in the Euclidean case we essentially conduct RSVGD on the distribution manifold. To specify a kernel on it, we note that the Euclidean space of the latent variable is a global c.s. $(\Phi, \mathbb{R}^m)$ of the manifold. By the mapping Φ, any kernel on $\mathbb{R}^m$ is a kernel on the manifold (Steinwart and Christmann (2008), Lemma 4.3). In this sense we choose the Gaussian kernel on $\mathbb{R}^m$. Furthermore, we implement the kernel for both SVGD and RSVGD as the summation of several Gaussian kernels with different bandwidths, which is still a valid kernel (Steinwart and Christmann (2008), Lemma 4.5). We find it slightly better than the median trick of SVGD in our experiments.

Setups We compare RSVGD with SVGD (full-batch) for test accuracy along iteration. We fix $\alpha = 0.01$ and use 100 particles for both methods. We use the Splice19 dataset (1,000 training entries, 60 features), one of the benchmark datasets compiled by Mika et al. (1999), and the Covertype dataset (581,012 entries, 54 features) also used by Liu and Wang (2016). RSVGD updates particles by the aforementioned 1st-order flow approximation, which is effectively the vanilla gradient descent, while SVGD uses the recommended AdaGrad with momentum. We also tried gradient descent for SVGD, with no better outcomes.

Results We see from Fig. 1 that RSVGD makes more effective updates than SVGD on both datasets, indicating the benefit of RSVGD to utilize the distribution geometry for Bayesian inference. Although AdaGrad with momentum, which SVGD uses, counts for a method for estimating the geometry empirically (Duchi, Hazan, and Singer 2011), our method provides a more principled solution, with more precise results.

Spherical Admixture Model

We investigate the advantages of RSVGD on Riemann manifold (Eqn. 7) by the inference task of Spherical Admixture Model (SAM) (Reisinger et al. 2010), which is a topic model for data on hyperspheres, e.g. *tf-idf* feature of documents. The model first generates corpus mean $\mu \sim \mathrm{vMF}(m, \kappa_0)$ and topics $\{\beta_k\}_{k=1}^P : \beta_k \sim \mathrm{vMF}(\mu, \sigma)$, then for document d, generates its topic proportion $\theta_d \sim \mathrm{Dir}(\alpha)$ and content $v_d \sim \mathrm{vMF}(\beta\theta_d/\|\beta\theta_d\|, \kappa)$, where vMF is the von Mises-Fisher distribution (Mardia and Jupp 2000) for random variable on hyperspheres, and Dir is the Dirichlet distribution.

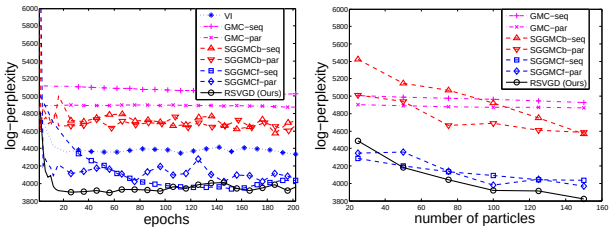
The inference task of SAM is to estimate the posterior of the topics $p(\beta|v)$. Note that the topics $\beta = (\beta_1, \ldots, \beta_P)$ lies in the product manifold of hyperspheres $(\mathbb{S}^{n-1})^P$.

Kernel Like the Gaussian kernel in the Euclidean case, we use the vMF kernel $K(y, y') = \exp(\kappa y^\top y')$ on hyperspheres. Note that the vMF kernel on $\mathbb{S}^{n-1}$ is the restriction of the Gaussian kernel on $\mathbb{R}^n$: $\exp(-\frac{\kappa}{2}\|y - y'\|^2) = \exp(-\kappa)\exp(\kappa y^\top y')$ for $y, y' \in \mathbb{S}^{n-1}$, so it is a valid kernel on $\mathbb{S}^{n-1}$ (Steinwart and Christmann (2008), Lemma 4.3). We also recognized that the arcsine kernel $K(y, y') = \arcsin(y^\top y')$ is also a kernel on $\mathbb{S}^{n-1}$, due to the non-negative Taylor coefficients of the arcsine function (Steinwart and Christmann (2008), Lemma 4.8). But unfortunately it does not work well in experiments.

For a kernel on $(\mathbb{S}^{n-1})^P$, we set $K(y, y') = \prod_{k=1}^P \exp(\kappa y_{(k)}^\top y'_{(k)})$. Again we use the summed kernel trick for RSVGD.

Setups The manifold constraint isolates the task from most prevalent inference methods, including SVGD. A mean-field variational inference method (VI) is proposed by the original work of SAM. Liu, Zhu, and Song (2016) present more methods based on advanced Markov chain Monte Carlo methods (MCMCs) on manifold, including Geodesic Monte Carlo (GMC) (Byrne and Girolami 2013), and their proposed Stochastic Gradient Geodesic Monte Carlo (SGGMC), a scalable mini-batch method. We implement the two MCMCs in both the standard sequential (-seq) way, and the parallel (-par) way: run multiple chains and collect the last sample on each chain. To apply RSVGD for inference, as with GMC and SGGMC cases, we adopt the framework of Liu, Zhu, and Song (2016), which directly provides an estimate of the all-we-need information $\nabla_\beta \log p(\beta|v)$. We run all the methods on the 20News-different dataset (1,666 training entries, 5,000 features) with default hyperparameters as the same as (Liu, Zhu, and Song 2016). We use epoch (the amount of visit to the dataset) instead of iteration since SGGMCb is run with mini-batch.

Results Fig. 2(a) shows that RSVGD makes the most effective progress along epoch, indicating its iteration-effectiveness. VI converges fast but to a less satisfying state due to its restrictive variational assumption, while RSVGD, by the advantage of high approximation flexibility, achieves a better result comparable to MCMC results. GMC methods make consistent but limited progress, where 200 epochs is too short for them. Although SGGMC methods perform outstandingly in (Liu, Zhu, and Song 2016), they are embarrassed here by limited particle size. Both fed on full-batch gradient, SGGMCf has a more active dynamics than GMC thus can explore a broader region, but still not as efficient as RSVGD. The -seq and -par versions of an MCMC behave similarly, although -par seems better at early stage since it has more particles.



(a) Results with 100 particles (b) Results at 200 epochs

Figure 2: 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.

Fig. 2(b) presents the results over various numbers of particles, where we find the particle-efficiency of RSVGD. For fewer particles, SGGMCf methods have the chance to converge well in 200 epochs thus can be better than RSVGD. For more particles MCMCs make less salient progress since the positive autocorrelation limits the representativeness of limited particles.

Conclusion

We develop Riemannian Stein Variational Gradient Descent (RSVGD), an extension of SVGD (Liu and Wang 2016) to Riemann manifold. We generalize the idea of SVGD and derive the directional derivative on Riemann manifold. To solve for a valid and close-formed functional gradient, we first analyze the requirements by Riemann manifold and illustrate the failure of SVGD techniques in our case, then propose our solution and validate it. Experiments show the benefit of utilizing distribution geometry on inference tasks on Euclidean space, and the advantages of particle-efficiency, iteration-effectiveness and approximation flexibility on Riemann manifold.

Possible future directions include exploiting Riemannian Kernelized Stein Discrepancy which would be more appropriate with a properly chosen manifold. Mini-batch version of RSVGD for scalability is also an interesting direction, since the naïve implementation does not work well as mentioned. Applying RSVGD to a broader stage is also promising, including Euclidean space tasks like deep generative models and Bayesian neural networks (with Riemann metric tensor estimated in the way of (Li et al. 2016)), and Riemann manifold tasks like Bayesian matrix completion (Song and Zhu 2016) on Stiefel manifold.

Acknowledgements This work is supported by the National NSF of China (Nos. 61620106010, 61621136008, 61332007) and Tiangong Institute for Intelligent Computing. We thank Jingwei Zhuo and Yong Ren for inspiring discussions.

References

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

Amari, S.-I., and Nagaoka, H. 2007. *Methods of information geometry*, volume 191. American Mathematical Soc.

Amari, S.-I. 2016. *Information geometry and its applications*. Springer.

Bonnabel, S. 2013. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control* 58(9):2217–2229.

Brubaker, M. A.; Salzmann, M.; and Urtasun, R. 2012. A family of mcmc methods on implicitly defined manifolds. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 161–172.

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

Chwialkowski, K.; Strathmann, H.; and Gretton, A. 2016. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2606–2615.

Dai, B.; He, N.; Dai, H.; and Song, L. 2016. Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, 985–994.

Do Carmo, M. P. 1992. *Riemannian Geometry*.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul):2121–2159.

Frankel, T. 2011. *The geometry of physics: an introduction*. Cambridge University Press.

Gemici, M. C.; Rezende, D.; and Mohamed, S. 2016. Normalizing flows on riemannian manifolds. *arXiv preprint arXiv:1611.02304*.

Girolami, M., and Calderhead, B. 2011. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(2):123–214.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *The Journal of Machine Learning Research* 14(1):1303–1347.

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

Li, C.; Chen, C.; Carlson, D. E.; and Carin, L. 2016. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *AAAI*, volume 2, 4.

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, Y.; Ramachandran, P.; Liu, Q.; and Peng, J. 2017. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*.

Liu, Q.; Lee, J. D.; and Jordan, M. I. 2016. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Ma, Y.-A.; Chen, T.; and Fox, E. 2015. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, 2917–2925.

Mardia, K. V., and Jupp, P. E. 2000. Distributions on spheres. *Directional Statistics* 159–192.

Micchelli, C. A., and Pontil, M. 2005. On learning vector-valued functions. *Neural computation* 17(1):177–204.

Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; and Mullers, K.-R. 1999. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, 41–48. IEEE.

Pu, Y.; Gan, Z.; Henao, R.; Li, C.; Han, S.; and Carin, L. 2017. Stein variational autoencoder. *arXiv preprint arXiv:1704.05155*.

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.

Rezende, D., and Mohamed, S. 2015. Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, 1530–1538.

Romano, G. 2007. Continuum mechanics on manifolds. *Lecture notes University of Naples Federico II, Naples, Italy* 1–695.

Song, Y., and Zhu, J. 2016. Bayesian matrix completion via adaptive relaxed spectral regularization. In *The 30th AAAI Conference on Artificial Intelligence (AAAI-16)*.

Steinwart, I., and Christmann, A. 2008. *Support vector machines*. Springer Science & Business Media.

Wang, D., and Liu, Q. 2016. Learning to draw samples: With application to amortized mle for generative adversarial learning. *arXiv preprint arXiv:1611.01722*.

Zhang, H.; Reddi, S. J.; and Sra, S. 2016. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, 4592–4600.

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