Kernel Stein Discrepancy on Lie Groups: Theory and Applications

Xiaoda Qu, Xiran Fan, Baba C. Vemuri *Fellow, IEEE*

***Abstract*—Distributional approximation is a fundamental problem in machine learning with numerous applications across all fields of science and engineering and beyond. The key challenge in most approximation methods is the need to tackle the intractable normalization constant present in the candidate distributions used to model the data. This intractability is especially common in distributions of manifold-valued random variables such as rotation matrices, orthogonal matrices etc. In this paper, we focus on the distributional approximation problem in Lie groups since they are frequently encountered in many applications including but not limited to, computer vision, robotics, medical imaging and many more. We present a novel Stein's operator on Lie groups leading to a kernel Stein discrepancy (KSD), which is a normalization-free loss function. We present several theoretical results characterizing the properties of this new KSD on Lie groups and its minimizer namely, the minimum KSD estimator (MKSDE). Properties of MKSDE are presented and proved, including strong consistency, CLT and a closed form of the MKSDE for the von Mises-Fisher, the exponential family and the Riemannian normal distributions on $\mathrm{SO}(N)$. Finally, we present several experimental results depicting advantages of MKSDE over maximum likelihood estimation.**

***Index Terms*—Stein's Operator, Lie Groups, Riemannian Manifolds, Kernel Stein Discrepancy, Exponential Distribution, Riemannian Normal Distribution**

I. Introduction (Problem Motivation and Literature Review)

Distributional approximation is a ubiquitous problem in science and engineering with numerous applications. This fundamental problem can be formulated as follows: Suppose we wish to approximate a density q by a family of candidate probability densities $\{p_\alpha\}$, either parameterized by some finite dimensional parameters or non-parametrically represented via say, a neural network. The approximation is invariably accomplished via the optimization of a loss (cost) function D, which measures the dissimilarity between distributions, i.e., $\alpha^* := \operatorname{argmin} D(q, p_\alpha)$. A vanilla application of this fundamental technique known as maximum likelihood estimation (MLE), is an important constituent of a wide variety of algorithms, e.g., the well known expectation-maximization (EM) algorithm, the recursive stochastic filter namely the Kalman filter [1] and many others.

Most commonly used loss functions include likelihood or log-likelihood functions and the KL-divergence. However, in many cases, the family of candidate densities $\{p_\alpha\}$ are only known up to a highly intractable normalizing constant. In practice, one must approximate the normalizing constant and its derivative w.r.t α numerically in each step of say a gradient decent method employed in the optimization. *An intractable normalizing constant leads to a cost-accuracy trade-off i.e., higher the computational cost, better the accuracy of approximation. If one can circumvent this intractable constant altogether and yet achieve high accuracy in parameter estimation, it would be ideal and this is precisely what we will achieve in this paper.*

Such a situation arises even in high dimensional Euclidean space, not to mention the non-flat spaces. In many computer vision and machine learning applications such as, unmanned aerial vehicle tracking, aircraft trajectory analysis, human motion analysis etc., one encounters data in different Lie groups, as it is the most appropriate space to represent different types of object motions e.g., rotations, affine motions etc. Lie groups capture the underlying intrinsic geometric structure of these transformations that represent the object motions. In contrast, a vector space structure proves to be inadequate for most tasks in manipulating such data. For example, the arithmetic mean of several rotational motions – rotation (orthogonal) matrices – of an object is no longer a rotation. Therefore, to model the rotation of objects, it is common to use the geometric structure of the special orthogonal group $\mathrm{SO}(N) := \{R \in \mathbb{R}^{N\times N} : R^\top R = I, \det(R) = 1\}$ instead of $\mathbb{R}^{N\times N}$.

The issue of normalizing constant is significant for distributions in Lie groups. The integral on Lie groups w.r.t its volume measure, though well-defined, most often is computationally intractable. Even the most widely-used, well-behaved probability families on Lie group have an intractable normalizing constant. For example, von Mises-Fisher distribution (vMF) [2] and Bingham distribution [3], [4] on $\mathrm{SO}(N)$, whose normalizing constants are hypergeometric functions of the parameters, are widely used in pose estimation or rotation estimation with uncertainty [5]–[8] and Bayesian attitude estimation [9], [10] and many others. The Riemannian normal distribution [11], with a highly intractable normalizing constant, is encountered in probabilistic principal geodesic analysis (PPGA) [12], [13]. The one-axis model on special Euclidean group SE(3) [14, Ch. 6], for modeling the rigid body motion [15] contains a vMF component. These are just a few examples where intractability of the normalizing constant is encountered.

A normalization-free loss, called *kernel Stein discrepancy* (KSD), was first proposed in [16], [17], which combines Stein's method with the theory of reproducing kernel Hilbert space (RKHS). The KSD has been extensively studied recently in different aspects of a general framework [18]–[20], characterization scope [21], [22], asymptotic properties regarding

Xiaoda Qu is with Dept. of Statistics, Univ. of Florida; Xiran Fan is a Research Scientist at Visa, CA; Baba C. Vemuri is with Dept. of CISE, Univ. of Florida. This research was in part funded by the NIH NINDS and NIA grant RO1NS121099 to Vemuri. email: {quxiaoda, vemuri}@ufl.edu

its minimization [23], [24], and its relevant applications [17], [25], [26]. It involves an RKHS $\mathcal{H}_k$ of a kernel function k on $\mathbb{R}^d$, termed as *Stein's class* here, and a *Stein operator* S_{p_α}. The Stein operator $\mathcal{S}_{p_\alpha}$, depends on the candidate distribution p_α but is free from its normalizing constant, maps the elements in $\mathcal{H}_k^d$, the d-times product of $\mathcal{H}_k$, to real-valued functions. In addition, they must satisfy *Stein's identity*, that is, $\mathbb{E}_p[\mathcal{S}_p f] = 0$ for all $f \in \mathcal{H}_k^d$. Then, the KSD between p and another distribution q is defined as,

$$\mathrm{KSD}(p,q) := \sup\left\{\mathbb{E}_q[\mathcal{S}_p f] : f \in \mathcal{H}_k^d, \|f\|_{\mathcal{H}_k^d} \le 1\right\}. \quad (1)$$

In contrast to the KSD on $\mathbb{R}^d$, the existing generalizations of KSD to non-flat manifolds are somewhat lacking. In [27] Barp et al. developed a novel KSD limited to compact Riemannian manifolds and studied its convergence property using a higher order Sobolev space as the RKHS. Recently, in [28] Xu and Matsuda implemented the Euclidean KSD in a local coordinate chart, thus the KSD in their work is only applicable to distributions supported within the local coordinate chart. A more detailed discussion on these two works will be presented in section II-C.

Our Contributions and Paper Organization

In this work, we propose a novel Stein's operator (see equation 3) making use of the structure of Lie groups, leading to a KSD with a closed form (see equation 4), *a normalization-free loss function applicable to all Lie groups*. The asymptotic properties of the KSD and its minimizer are established in theorems 4.5, 4.6, 4.7 and 4.8 respectively. Furthermore, we present applications of the KSD to important problems on the widely encountered Lie group $\mathrm{SO}(N)$ in §V and example 1. Specifically, experiments VI-A will address the issue of the normalization constant that arises in MLE and how the estimation obtained using proposed normalization-free KSD yields far more accurate parameter estimates compared to MLE that uses approximations for the normalization constant.

a) Organization:: The rest of this paper is organized as follows: In Section II, we present the mathematical preliminaries required in the rest of the paper. Section III contains the key theoretical results involving the derivation of the KSD on Lie groups. In Section IV we present the minimum kernel Stein discrepancy estimator and its asymptotic properties, and analyze different practical situations that arise in applications. Proofs of all theorems are included in the appendix. Finally, we present experiments in Section VI and conclude in Section VII.

II. Background

In this section, we present a brief overview of relevant mathematics on Lie groups and the KSD on $\mathbb{R}^d$. For a more detailed discussion on Lie groups, we refer the reader to [29].

A. Manifolds

A manifold M is a second countable Hausdorff space that is locally homeomorphic to an open set in $\mathbb{R}^d$. The number d is the *dimension* of M. Such local homeomorphisms are called *charts*. A smooth structure on M is assigned via a collection of charts that covers the manifold such that the composition of every two charts is smooth on $\mathbb{R}^d$. A manifold with a smooth structure is called a *smooth manifold*.

B. Lie groups

A *Lie group* G is a group and a smooth d-dimensional manifold whose multiplication and group inverse are smooth. For any fixed $g \in G$, the *left (resp. right) multiplication action* $L_g : h \mapsto gh$, (resp. $R_g : h \mapsto hg$) is smooth, whose differential is denoted by dL_g (resp. dR_g) and its pullback is denoted by L_g^* (resp. R_g^*).

a) Left-invariant fields: A vector field is a smooth assignment of tangent vectors to the tangent space at each point of G. A *left-invariant field* is a vector field invariant w.r.t left-translations, i.e., dL_g or L_g^* for any $g \in G$. Clearly, if a group of left-invariant vector fields is linearly independent at some point, then they are linearly independent at every point. By the term *left-invariant vector (resp. covector or tensor) basis*, we mean a collection of left-invariant vector (resp. covector or tensor) fields that forms a basis at every point. Given a left-invariant basis, any other left-invariant field can be written as the linear combination of the given basis.

b) Volume measure and modular functions: Akin to the Lebesgue measure on $\mathbb{R}^d$, the canonical dominating measure on a Lie group is usually taken as the *left-invariant Haar measure* μ, defined via left invariant volume form Ω or the Haar theorem, see [30] and [31]. The left-invariant measure μ satisfies $\mu(gS) = \mu(S)$ for all $g \in G$ and Borel set S, which is unique up to a constant. For each fixed $g \in G$, note that $\mu_g(S) := \mu(Sg)$ is still a left-invariant measure, thus there exists a number $\Delta_g > 0$ such that $\mu_g = \Delta(g)\mu$. The function $\Delta : g \mapsto \Delta_g$ is the *modular function* of G. We say that G is *unimodular*, if $\Delta \equiv 1$, e.g., $\mathrm{SO}(N)$, $\mathrm{SE}(N)$, etc. All Abelian or compact Lie groups are unimodular. All probability densities used in this paper are with respect to the volume measure μ.

C. KSD on $\mathbb{R}^d$ and existing generalizations

Suppose $\mathcal{H}_k$ is an RKHS on $\mathbb{R}^d$ of a kernel k. Let $\mathcal{H}_k^d := \mathcal{H}_k \times \cdots \times \mathcal{H}_k$ be the d-times product of $\mathcal{H}_k$, endowed with the inner product $\langle f, g\rangle_{\mathcal{H}_k^d} = \sum_{l=1}^d \langle f_l, g_l\rangle_{\mathcal{H}_k}$ for $f = (f_1, \ldots, f_d)$ and $g = (g_1, \ldots, g_d)$ in $\mathcal{H}_k^d$. The most commonly used Stein operator $\mathcal{S}_p$ on $\mathbb{R}^d$ also adopted in [17], [22]–[26], denoted here by T_p, has the form

$$\mathcal{T}_p : f \mapsto \sum_{l=1}^d \left[\frac{\partial f_l}{\partial x^l} + f_l \frac{\partial}{\partial x^l}\log p\right], \quad f \in \mathcal{H}_k^d. \quad (2)$$

Then the KSD is defined by

$$\mathrm{KSD}(p,q) := \sup\left\{\mathbb{E}_q[\mathcal{T}_p f] : f \in \mathcal{H}_k^d, \|f\|_{\mathcal{H}_k^d} \le 1\right\}.$$

Clearly, one could easily see from (1) and Stein's identity that the KSD is always non-negative and satisfies $\mathrm{KSD}(p,p) = 0$. In fact, as discussed in [17], [22]–[26], if the kernel function k is C_0-universal [32], then KSD will uniquely characterize p, i.e., for p and q regular enough, we have $\mathrm{KSD}(p,q) =$

$0 \Leftrightarrow p = q$. Notably, the computation of T_p is free from the normalizing constant of p, so is the corresponding KSD in (1).

As one of the significant application of KSD, the *minimum kernel Stein discrepancy estimator* (MKSDE), first proposed by [23] and further researched in [24], [26], minimizes the KSD between a parametrized family p_θ and the empirical distribution of a group of samples to obtain an approximation p_{θ^*} of the underlying distribution of the samples. The MKSDE has good convergent properties and thus can serve as a normalization-free alternative to MLE.

Since the issue of normalization is even more severe on non-flat spaces, one would naturally wish to generalize such a method to manifolds. In [28], Xu and Matsuda use a local chart to construct a *differential form* ω, such that its differential $d\omega = p\mathcal{S}_p f\Omega$ (Ω is the volume form), and then apply Stokes's theorem to conclude that $\mathbb{E}_p[\mathcal{S}_p f] = \int p\mathcal{S}_p f\Omega = \int d\omega = 0$. However, as their construction relies on the local coordinates, $p\mathcal{S}_p f\Omega$ does not extend beyond the chart and is singular on the boundary of the chart, however, Stokes's theorem does require $p\mathcal{S}_p f\Omega$ to be globally smooth. Thus, the applicability of this method is limited to cases where p is supported inside the chart so that the $p\mathcal{S}_p f\Omega$ vanishes (smoothly) before reaching the boundary, or when a global chart exists. It is however well known that there is no global chart on any compact manifold, and most commonly used families are all globally supported, e.g., the vMF family on $\mathrm{SO}(N)$. Xu and Matsuda [28] present an example of the vMF family on the sphere and as stated earlier, the sphere has no global chart (spherical coordinates are not global as they are singular on the boundary) and it violates their Stein's operator construction that depends on the local charts. Further, the volume density J in their work is easy to compute on a sphere but not so on more complicated spaces, such as $\mathrm{SO}(N)$ and others.

In [27], Barp et al. adopted a second order Stein's operator free from the choice of coordinate charts in constructing an approximation to the posterior expectation of distributions supported only on compact manifolds. In addition, to successfully characterizing p uniquely with the KSD corresponding to this operator, one must consider a very large RKHS, i.e., the Sobolev space on M. Finding such a Sobolev space with an easy-to-derive closed form kernel can be challenging on a highly curved manifold. One possible way to address this challenge, not addressed in [27], is to use the restriction of a Sobolev-type kernel on $\mathbb{R}^d$ to the manifold [33]. However, this method does not guarantee a closed form expression for the kernel.

III. KSD on Lie groups

In this section, we present a novel Stein's operator on Lie groups denoted here by $\mathcal{A}_p$ and derive its corresponding KSD, making use of the group structure, specifically, left-invariant basis, akin to the basis $\frac{\partial}{\partial x^l}$ on $\mathbb{R}^d$.

A. Stein's operator on Lie groups and associated KSD

Suppose G is a connected Lie group with a modular function Δ and $\{D^l\}_{l=1}^d$ is a left-invariant basis on G. Suppose p is a locally Lipschitz probability density function on G, only known up to a constant. Suppose $\mathcal{H}_k$ is an RKHS with a kernel function k on G. Then, the *Stein's operator* $\mathcal{A}_p$ is defined as,

$$\mathcal{A}_p : f \mapsto \sum_{l=1}^{d} \left[D^l f_l + f_l D^l \log p + f_l D^l \Delta \right], \quad f \in \mathcal{H}_k^d. \tag{3}$$

Here $D^l \log p$ is set to 0 whenever $p = 0$. In particular, we specify each component $\mathcal{A}_p^l$ of the operator $\mathcal{A}_p$ as $\mathcal{A}_p^l : h \mapsto D^l h + hD^l \log p + hD^l \Delta$ for $h \in \mathcal{H}_k$. Thus, $\mathcal{A}_p f = \sum_{l=1}^d \mathcal{A}_p^l f_l$. In addition, the vector $(\mathcal{A}_p^1 f_1, \ldots, \mathcal{A}_p^d f_d)^\top$ is denoted by $\vec{\mathcal{A}}_p f$.

On unimodular groups, e.g, compact Lie groups or abelian Lie groups, the last term $f_l D^l$ disappears since the modular function Δ is a constant. Specifically, note that $(\mathbb{R}^d, +)$ is an abelian Lie group, thus $\mathcal{A}_p$ degenerates to T_p in (2) on $\mathbb{R}^d$. In contrast to the generalization by [28], the computation of $\mathcal{A}_p$, as well as its corresponding KSD defined through (1), is independent of the choice of a coordinate chart. Although $\mathcal{A}_p$ seemingly depends on the choice of left-invariant basis D^l, we will show later that different choices of D^l will actually lead to equivalent KSDs related to each other via the inequality (5).

Following result ensures that all $f \in \mathcal{H}_k$ are differentiable so that $\mathcal{A}_p f$ is meaningful.

Theorem 3.1. *If $k \in C^2(G \times G)$, i.e., twice continuously differentiable, then all $f \in \mathcal{H}_k$ is C^1. Furthermore, given a tangent vector $D \in T_{x_0}M$, $(Dk)_{x_0} \in \mathcal{H}_k$ and $Df(x_0) = \langle f, (Dk)_{x_0} \rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$. Here $(Dk)_{x_0}$ represents the function obtained by letting D act on the first argument of k and fix the first argument at x_0.*

Proof. See [34, Lem. 4.34]. □

From here on, we always assume that our kernel function is at least C^2. The *kernel Stein discrepancy* (KSD) on Lie groups defined via (1) and (3), share most of properties with the vanilla KSD on $\mathbb{R}^d$ [17], [25], defined via (1) and (2) – specifically, replacing S_p in (1) with T_p from (2) in the $\mathbb{R}^d$ case and $\mathcal{A}_p$ from (3) in the Lie group case respectively. For example, the shared properties include but are not limited to (i) non-negativeness, and most significantly, (ii) an integral closed form (4). This remarkable closed form is based on the structure of the RKHS. We denote by $\mathcal{A}_p^l k_x := \mathcal{A}_p^l k(x, \cdot)$ the function obtained by letting $\mathcal{A}_p^l$ act on the first argument of k and then fix the first argument of $\mathcal{A}_p^l k(\cdot, \cdot)$ at x. Under the condition in theorem 3.1, it is clear that $\mathcal{A}_p^l k_x \in \mathcal{H}_k$ for all x, and $\mathcal{A}_p^l h(x) = \langle h, \mathcal{A}_p^l k_x \rangle_{\mathcal{H}_k}$ for all $h \in \mathcal{H}_k$. Furthermore, the expectation $\mathbb{E}_q[\mathcal{A}_p^l h]$ can be written as $\mathbb{E}_q \langle h, \mathcal{A}_p^l k_{(\cdot)} \rangle_{\mathcal{H}_k} = \langle h, \mathbb{E}_q[\mathcal{A}_p^l k_{(\cdot)}] \rangle_{\mathcal{H}_k}$. The expectation $\mathbb{E}_q[\mathcal{A}_p^l k_{(\cdot)}]$ is the Bochner-integral [34, §A.5.4] of the map $x \mapsto \mathcal{A}_p^l k_x$ from G to $\mathcal{H}_k$ w.r.t to q, which is well-defined whenever the map $x \mapsto \mathcal{A}_p^l k_x$ is Bochner q-integrable [34, §A.5.4]. In such a case, the KSD can be represented by

$$\begin{aligned} \mathrm{KSD}(p, q) &= \sup_{\|f\|_{\mathcal{H}_k^d} \le 1} \mathbb{E}_q[\mathcal{A}_p f] = \sup_{\|f\|_{\mathcal{H}_k^d} \le 1} \langle f, \mathbb{E}_q[\vec{\mathcal{A}}_p k_{(\cdot)}] \rangle_{\mathcal{H}_k^d} \\ &= \|\mathbb{E}_q[\vec{\mathcal{A}}_p k_{(\cdot)}]\|_{\mathcal{H}_k^d}. \end{aligned}$$

Substituting the term $\mathcal{A}_p^l k_y$ for h in the property $\mathcal{A}_p^l h(x) = \langle h, \mathcal{A}_p^l k_x \rangle_{\mathcal{H}_k}$, we have,

$$k_p(x,y) := \sum_{l=1}^d k_p^l(x,y) := \sum_{l=1}^d \tilde{\mathcal{A}}_p^l \mathcal{A}_p^l k(x,y)$$
$$= \sum_{l=1}^d \langle \mathcal{A}_p^l k_x, \mathcal{A}_p^l k_y \rangle_{\mathcal{H}_k}.$$

Here, the tilde on $\tilde{\mathcal{A}}_p^l$ indicates that it acts on the second argument. Interchanging the inner product and the expectation again, we have the closed form of KSD, summarized in the following theorem, which is a *generalization to the classical version presented in* [17], [25] on $\mathbb{R}^d$.

Theorem 3.2 (Closed form KSD). *Suppose $k \in C^2(G \times G)$ and $\sqrt{k_p(x,x)}$ is p and q-integrable, then the map $x \mapsto \mathcal{A}_p^l k_x$ is Bochner q-integrable and*

$$\mathrm{KSD}^2(p,q) = \int_G \int_G k_p(x,y) q(x) q(y) \mu(dx) \mu(dy). \tag{4}$$

Proof. The proof is along the same lines of the theroem in the $\mathbb{R}^d$ case, see e.g. [17], [25], and hence omitted here. □

With theorem 3.2 in hand, we can establish the invariance of KSD with respect to different choice of basis. In particular, orthogonal transformation between basis will preserve the KSD. The following theorem formalizes this invariance.

Theorem 3.3 (Invariance of basis). *Suppose $A := (a_k^l)$ is the transformation between two bases D^l and E^k such that $D^l = \sum_k a_k^l E^k$, then the* KSD *with basis D^l and the* $\overline{\mathrm{KSD}}$ *with basis E^k satisfies*

$$\sqrt{\lambda_{\min}} \cdot \overline{\mathrm{KSD}} \leq \mathrm{KSD} \leq \sqrt{\lambda_{\max}} \cdot \overline{\mathrm{KSD}}, \tag{5}$$

where, $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $A^\top A$.

Proof. Let $\mathcal{T}_p$ be the Stein's operator with basis E^k. Note that we have $\mathcal{A}_p^l h = \sum_k a_k^l \mathcal{T}_p^k h$ for $h \in \mathcal{H}_k$. This linear transformation also commutes with Bochner integral, that is, $\mathbb{E}_q[\vec{\mathcal{A}}_p f] = A \cdot \mathbb{E}_q[\vec{\mathcal{T}}_p f]$. Recall that $\mathrm{KSD} = \|\mathbb{E}_q[\vec{\mathcal{A}}_p f]\|_{\mathcal{H}_k^d}$ and $\overline{\mathrm{KSD}} = \|\mathbb{E}_q[\vec{\mathcal{T}}_p f]\|_{\mathcal{H}_k^d}$. The rest of the proof is standard. □

In practice, almost all of the commonly used Lie groups are matrix Lie groups, whose Lie algebra (tangent space at identity) is a subspace of $\mathbb{R}^{N \times N}$. Thus, the Lie algebra will inherit the canonical inner product of $\mathbb{R}^{N \times N}$, i.e., $\langle A, B \rangle = \mathrm{tr}(A^\top B)$, and the orthonormal basis w.r.t this inner product is a canonical choice of the basis.

Under mild regularity conditions, the KSD will uniquely characterize p, as stated formally in the following theorem.

Theorem 3.4 (Characterization). *Suppose $k \in C^2(G \times G)$ is C_0-universal. Suppose further $\sqrt{k_p(x,x)}$ and $D^l \log(p/q)$ are integrable with respect to locally Lipschitz densities p and q for all l. Then $p = q \iff \mathrm{KSD}(p,q) = 0$.*

Proof. The proof is included in §A. □

Choice of kernel: In contrast to [27], where the k must reproduce the Sobolev space, a C_0-universal kernel is much easier to obtain on a Riemannian manifold. Specifically, [32, Ex. 6.12] showed that the bivariate function $\exp(-\frac{\tau}{2}\|x-y\|^2)$, $\tau > 0$ restricted on any closed set X in $\mathbb{R}^m$ is C_0-universal on X and we will use it in the next section. It is possible to consider other C_0-universal kernels such as the Laplace and the Matérn but we will focus on the above mentioned bivaraite function and derive the MKSDE formulas for commonly encountered distributions namely, the exponential family and the Riemannian normal. The derivations will be similar for the other C_0-universal kernels.

IV. Minimum kernel Stein discrepancy estimator

Given a locally Lipschitz density q and a family of locally Lipschitz densities $\{p_\alpha\}$, we define the *minimum kernel Stein discrepancy estimator* (MKSDE) as any of the global minimizers of the KSD, that is,

$$\widehat{\alpha}^{\mathrm{KSD}} := \operatorname{argmin} \mathrm{KSD}(\alpha) := \operatorname{argmin} \mathrm{KSD}(p_\alpha, q). \tag{6}$$

However, the $\mathrm{KSD}(\alpha)$ is typically not available in practice, as the integral in (4) may be intractable or in some cases we only have samples from q. In such cases, we choose to minimize a discrete KSD from a suite of different KSDs based on the situation. In this section, we will introduce these minimization schemes and study their asymptotic properties.

A. Minimization schemes

The KL-divergence $D(q||p_\alpha) = \mathbb{E}_q \log(p_\alpha) - \mathbb{E}_q \log(q)$ is commonly used for distribution approximation. If accessibility is limited to the samples from q instead of its closed form, then, we just minimize the discretized KL-divergence $\sum_{i=1}^n \log(p_\alpha(x_i))$, i.e., the log-likelihood function.

Similarly, the KSD can be discretized as a U-statistic $U_n = \frac{1}{n(n-1)} \sum_{i \neq j} k_p(x_i, x_j)$, based on which the kernel Stein goodness of fit test was developed in several prior research works [17], [25], [28]. We could minimize this U-statistic to obtain MKSDE from samples, or minimize the V-statistic

$$\mathrm{KSD}_n^2(\alpha) := \mathrm{KSD}_n^2(p_\alpha, q) := \frac{1}{n^2} \sum_{i,j} k_{p_\alpha}(x_i, x_j), \tag{7}$$

which has the analogous asymptotic property but is more stable for optimization since it results in a convex loss function, for the case of $\{p_\alpha\}$ being the von-Mises Fisher family or a mixture of von-Mises Fisher family.

Another situation that commonly occurs in practice is when we do have a closed form of p_α and q, but the integral in (4) is intractable. For example, in the rotation tracking problem encountered in robotics [35], one must approximate the posterior distribution of Q_k with a von Mises-Fisher distribution so that the tracking algorithm (Kalman filter) updates consistently lie in the same space. In a Bayesian fusion problem [9], one must go through a similar process to ensure that the result of the fusion stays in the family. In such cases, we can draw samples from q with any of the various sampling algorithms, e.g., Hamiltonion Monte Carlo or Metropolis-Hastings algorithms, and minimize (7). Alternatively, if these sampling methods are hard to implement, we could use the importance sampling scheme to sample from another distribution w and minimize the *weighted (discrete) KSD*:

$$\begin{aligned}\mathrm{wKSD}_n^2(\alpha) &:= \mathrm{wKSD}_n^2(p_\alpha, q)\\ &= \frac{1}{n^2}\sum_{i,j} k_{p_\alpha}(x_i,x_j)q(x_i)q(x_j)w^{-1}(x_i)w^{-1}(x_j).\end{aligned} \tag{8}$$

For example, on any compact Lie groups G, e.g., $\mathrm{SO}(N)$, we could directly sample from the uniform distribution on G, as the volume is finite. Then the KSD in (8) will degenerate to

$$\mathrm{wKSD}_n^2(\alpha) = \frac{1}{n^2}\sum_{i,j} k_{p_\alpha}(x_i,x_j)q(x_i)q(x_j). \tag{9}$$

Note that the normalizing constants of q and w are not necessary during the minimization.

B. Asymptotic properties of KSD and MKSDE

In this section, we study the asymptotic properties of KSD and MKSDE obtained from (7) and (8). Note that the weighted KSD in (8) will degenerate to KSD in (7) if $w = q$, thus it suffices to study the MKSDE obtained in (8). We denote the function

$$V_\alpha(x,y) := k_{p_\alpha}(x,y)q(x)q(y)w^{-1}(x)w^{-1}(y).$$

In this section, we assume that all the conditions in theorem 3.4 hold. We would like to emphasize that all the results in this section pertain to the weighted KSD in (8) on Lie groups, and are distinct from the classical results in [23] that involve the unweighted KSD on $\mathbb{R}^d$.

Theorem 4.5 (Asymptotic Distribution of KSD)**.** *Suppose $V_\alpha(x,x)$ is w-integrable. For fixed α, as $n \to \infty$*

1) $\mathrm{wKSD}_n(\alpha) \to \mathrm{KSD}(\alpha)$ *almost surely;*
2) *If* $\mathrm{KSD}(\alpha) \neq 0$,
$$\sqrt{n}[\mathrm{wKSD}_n^2(\alpha) - \mathrm{KSD}^2(\alpha)] \xrightarrow{d.} N\big(0, 4\tilde{\sigma}^2\big),$$
where $\tilde{\sigma}^2 := \mathit{var}_{y\sim w}[\mathbb{E}_{x\sim w}V_\alpha(x,y)]$.
3) *If* $\mathrm{KSD}(\alpha) = 0$, *then*
$$n\,\mathrm{wKSD}_n^2(\alpha) \xrightarrow{d.} \sum_{l=1}^{\infty}\lambda_l Z_l^2,$$
where $Z_l \overset{i.i.d}{\sim} N(0,1)$, *$\lambda_l$ are the eigenvalues of the operator* $K : L^2(\omega) \to L^2(\omega), Kg := \mathbb{E}_{y\sim w}[V_\alpha(\cdot,y)g(y)]$ *s.t.* $\sum_{l=1}^{\infty}\lambda_l = \mathbb{E}_w[V_\alpha(x,x)]$.

Moreover, we have

$$\begin{aligned}0 &\le \varliminf_{n\to\infty} \inf_\alpha \mathrm{wKSD}_n^2(\alpha)\\ &\le \varlimsup_{n\to\infty} \inf_\alpha \mathrm{wKSD}_n^2(\alpha) \le \inf_\alpha \mathrm{KSD}^2(\alpha)\end{aligned}$$

In particular, if $\inf_\alpha \mathrm{KSD}(\alpha) = 0$, *then* $\inf_\alpha \mathrm{wKSD}_n(\alpha) \to \inf_\alpha \mathrm{KSD}(\alpha)$ *almost surely;*

Proof. Note that $V_\alpha(\cdot,\cdot)$ is also positive definite, thus we have $V_\alpha(x,y)^2 \le V_\alpha(x,x)V_\alpha(y,y)$, thus $\mathbb{E}_q[V_\alpha(x,x)] < +\infty$ implies that $\mathbb{E}_{x,y\sim q}[V_\alpha(x,y)^2] < +\infty$. Therefore, (1), (2) and (3) are straightforward from the classical asymptotic results of V-statistics [36, §6.4.1]. To show the last statement, note that for fixed α, $\mathrm{wKSD}_n^2(\alpha) = \frac{n-1}{n}U_n + n^{-2}\sum_{i=1}^n V_\alpha(x_i,x_i)$, then $\mathrm{wKSD}_n^2(\alpha) \to \mathrm{KSD}^2(\alpha)$ almost surely by SLLN and the strong consistency of U-statistic [36, Thm. 5.4.A], with convergence rate of $O_p(n^{-1})$ by the convergence rate of U-statistic [36, Thm 5.4.B]. Take a sequence of α_m such that $\mathrm{KSD}(\alpha_m) \le \inf_\alpha \mathrm{KSD}(\alpha) + m^{-1}$, then

$$\begin{aligned}\inf_\alpha \mathrm{wKSD}_n(\alpha) \le \mathrm{wKSD}_n(\alpha_m) &\to \mathrm{KSD}(\alpha_m)\\ &\le \inf_\alpha \mathrm{KSD}(\alpha) + m^{-1}.\end{aligned}$$

We conclude the result due to the arbitrariness of m. □

Up until now, we did not assume any additional structure of the index α. To obtain the asymptotic behavior of the parameter, we re-tag the index of the density family by θ from some connected metric parameter space (Θ, d). Let $\Theta_0 \subset \Theta$ be the set of best approximators θ_0, i.e., $\mathrm{wKSD}(\theta_0) = \inf_\theta \mathrm{wKSD}(\theta)$. Let $\widehat{\Theta}_n$ be the set of MKSDE, i.e., minimizers $\hat{\theta}_n$ of $\mathrm{wKSD}_n(\theta)$ in (8), which is a random set. With the additional metric structure on the parameter space, we can establish a stronger asymptotic result on KSD.

Theorem 4.6. *Suppose $V_{(\cdot)}(\cdot,\cdot)$ is jointly continuous and $\sup_{\theta\in K} V_\theta(x,x)$ is w-integrable for any compact $K \subset \Theta$, then $\mathrm{wKSD}_n^2(\theta) \to \mathrm{KSD}^2(\theta)$ compactly almost surely, i.e., for any compact K, as $n \to \infty$,*

$$\mathrm{wKSD}_n^2(\theta) \to \mathrm{KSD}^2(\theta) \textit{ uniformly on } K, \quad \textit{almost surely}.$$

As a corollary, if Θ is locally compact, wKSD_n and KSD are all continuous on Θ.

Proof. The proof is included in §B. □

The condition in theorem 4.6 namely, the joint continuity of $V_{(\cdot)}(\cdot,\cdot)$, is much easier to test in practice than prior results in [37], [38] on the uniform strong law for U-statistics. With theorem 4.6 in hand, we can establish the strong consistency of MKSDE.

Theorem 4.7 (Strong consistency)**.** *Suppose the conditions in theorem 4.6 hold and $\Theta = \Theta_1 \times \Theta_2$ such that Θ_1 is compact, Θ_2 is convex, and for each fixed $\theta_1 \in \Theta_1$, $\mathrm{wKSD}_n(\theta_1,\cdot)$ is convex on Θ_2 and $\mathrm{KSD}(\theta_1,\cdot)$ attains minimum value on a non-empty and compact set $\tilde{\Theta}_0(\theta_1) \subset int(\Theta_2)$. Then Θ_0, $\widehat{\Theta}_n$ are non-empty for large n and $\sup_{\theta\in\widehat{\Theta}_n} d(\theta,\Theta_0) \to 0$ almost surely as $n \to \infty$.*

Proof. The proof is included in §C. □

It is worth noting that if the family p_θ is identifiable and q is a member of the family p_θ, then the set Θ_0 of global minimizers is a singleton $\{\theta_0\}$. In such a situation, the MKSDE always converges to the unique θ_0.

In [23, Thm. 3.3] Barp et al., showed the consistency of MKSDE, assuming that the parameter space Θ is either compact or satisfies the conditions of convexity. However, this is not applicable to the Riemannian normal distribution, as it has a compact $\mu \in \mathrm{SO}(N)$ and a convex $\tau > 0$ simultaneously. In contrast, theorem 4.6, as a stronger version, only requires checking for the existence of global minimizers for one component, instead of for both components, and thus applicable to this situation.

In addition, note that the MKSDE of vMF is a quadratic form of F. Therefore, we have:

Corollary 4.7.1. *The MKSDE for von Mises-Fisher distribution computed via (14) and (8), and the MKSDE for Riemannian normal distribution computed via (15) and (8) are strongly consistent.*

Proof. The proof is straightforward and hence omitted. □

To establish the asymptotic normality of MKSDE, we assume that Θ is a connected Riemannian manifold with a Levi-Civita connection ∇ and the Riemannian logarithm map, Log. We assume that Θ_0 and $\widehat{\Theta}_n$ are non-empty for n large enough, and $\hat{\theta}_n$ is a sequence of MKSDE that converges to one of the global minimizer θ_0 of $\mathrm{KSD}(\theta)$. Additionally, we assume the following conditions:

- **(A1)** $V_\theta(x,y)$ is jointly continuous, and twice continuously differentiable in θ.
- **(A2)** there exists a compact neighborhood K of θ_0 such that $\sup_{\theta\in K}\|\nabla V_{\theta_0}(x,y)\|$ is $w\times w$-integrable.
- **(A3)** $\|\nabla V_{\theta_0}(x,y)\|^2$ and $\|\mathcal{I}_{\theta_0}(x,y)\|$ are $w\times w$-integrable, $\|\mathcal{I}_{\theta_0}(x,x)\|$ is w-integrable.
- **(A4)** $\mathcal{I}_{\theta_0}(x,y)$ is equi-continuous at θ_0.
- **(A5)** $\Gamma := \frac{1}{2}\mathbb{E}_{x,y\sim w}[\mathcal{I}_{\theta_0}(x,y)]$ is invertible.

Here $\nabla V_\theta(x,y)$ represents the gradient of $V_\theta(x,y)$ w.r.t θ, and $\mathcal{I}_\theta(x,y)$ represents the Hessian of $V_\theta(x,y)$ w.r.t θ. In addition, let Σ be the covariance matrix of the random vector $\mathbb{E}_{Y\sim w}[V_{\theta_0}(x,Y)]$.

Theorem 4.8 (CLT for MKSDE)**.** *Under assumption **A1**, **A2**, **A3**, **A4** and **A5**, we have* $\sqrt{n}\,\mathrm{Log}_{\theta_0}(\hat{\theta}_n)\xrightarrow{d.}\mathcal{N}(0,\Gamma^{-1}\Sigma\Gamma^{-1})$.

Proof. In this work, the parameter space is not assumed to be a vector space and thus is not necessarily flat. However, as the MKSDE $\hat{\theta}_n$ converges to the ground truth θ_0, the sequence will finally enter a neighborhood of θ_0. Since a manifold is locally Euclidean, the parameter space Θ can be considered as a flat space without loss of generality. It is convenient to take the normal coordinates given by the exponential map Exp_{θ_0} at θ_0, as the Jacobian of the Exp_{θ_0} is I at 0 and the geodesic connecting $\hat{\theta}_n$ with θ_0 coincides with the line segments connecting $\mathrm{Log}_{\theta_0}(\theta)$ with 0 on the tangent space at θ_0. The rest of the proof is the same as the one given in [24, Thm. 12]. □

C. Applications

1) MKSDE goodness of fit test for a distribution family: The goodness of fit test is a hypothesis test that tests whether a group of samples can be well-modeled by a given distribution p, that is, if we denote by q the unknown underlying distribution of samples, we aim to test $H_0 : p = q$ versus $H_1 : p \neq q$. Several works [17], [25] utilized the KSD to develop normalization-free goodness of fit tests on $\mathbb{R}^d$. However, their methods only apply to a specific candidate p, and requires testing individually for each member of a parameterized family p_θ. In many applications, the candidate distribution for the given samples is usually not a specific distribution but a parameterized family. In this section, *we develop a one-shot method to test whether a group of samples x_i matches a family p_θ, using the MKSDE obtained by minimizing (7).*

The *MKSDE goodness of fit* performs the test $H_0 : \exists\theta_0, p_{\theta_0} = q$ versus $H_1 : \forall\theta, p_\theta \neq q$. Under the null hypothesis H_0, $n\,\mathrm{wKSD}_n^2(\theta_0) \sim \sum_{j=1}^\infty \lambda_i Z_j^2$ asymptotically by (3) in theorem 4.5. Let $\gamma_{1-\beta}$ be the $(1-\beta)$-quantile of $\sum_{j=1}^\infty \lambda_j Z_j^2$ with significance level β. We reject H_0 if $n\,\mathrm{wKSD}_n(\hat{\theta}_n) \geq \gamma_{1-\beta}$, as it implies $n\,\mathrm{wKSD}_n^2(\theta_0) \geq n\,\mathrm{wKSD}_n^2(\hat{\theta}_n) \geq \gamma_{1-\beta}$, since $\hat{\theta}_n$ minimizes wKSD_n^2.

The evaluation of $\gamma_{1-\beta}$, relies on following result:

Proposition 1. *Let $\hat{\lambda}_l'$ be the eigenvalues of the Gram matrix $n^{-1}(V_{\theta_0}(x_i,x_j))_{ij}$ (set $\hat{\lambda}_l' := 0$ for $l > n$). Suppose $V_{\theta_0}(x,x)$ is p_{θ_0}-integrable. Then $\sum_{l=1}^\infty(\hat{\lambda}_l' - \lambda_l)Z_l^2 \to 0$ in probability as $n\to\infty$.*

Proof. See [39, Theorem 1]. □

Therefore, the $\sum_{l=1}^n \hat{\lambda}_l Z_l^2$ can serve as an empirical estimate of the asymptotic distribution $\sum_{l=1}^\infty \lambda_l Z_j^2$.

Although the ground truth θ_0 is unknown in our setting, the minimizers $\hat{\theta}_n$ of wKSD_n^2 converge to θ_0 almost surely by theorem 4.7 under specific conditions. Moreover, if the function V_θ is continuous in θ, then we can use the eigenvalues $\hat{\lambda}_l$ of $n^{-1}(V_{\hat{\theta}_n}(x_i,x_j))_{ij}$ as an alternative. To sum up, we have

Theorem 4.9. *Suppose the conditions in theorem 4.7 hold and $\{p_\theta\}$ is identifiable, then $\sum_{l=1}^\infty(\hat{\lambda}_l - \lambda_l)Z_l^2 \to 0$ in probability as $n\to\infty$.*

Proof. The proof is included in §D. □

The MKSDE goodness of fit test algorithm is summarized in the algorithm block 1. To implement the test, we assume the conditions in theorem 4.7 hold and $\{p_\theta\}$ is identifiable.

Algorithm 1 MKSDE goodness of fit test.

```
Input: population $x_1,\dots,x_n \sim q$; sample size $n$; number of generations $m$; significance level $\beta$.
Test: $H_0 : \exists\theta_0, p_{\theta_0} = q$ versus $H_1 : \forall\theta,\ p_\theta \neq q$.
Procedure:
    1. Find the minimizer $\hat{\theta}_n$ of $\mathrm{wKSD}_n^2(\theta)$ respectively either numerically using gradient descent or analytically as shown in the example 1.
    2. Obtain the eigenvalues $\hat{\lambda}_1,\dots,\hat{\lambda}_n$ of Gram matrix $n^{-1}(V_{\hat{\theta}_n}(x_i,x_j))_{ij}$.
    3. Sample $Z_j^i \sim N(0,1)$, $1 \leq j \leq n$, $1 \leq i \leq m$ independently.
    4. Compute $W^i = \sum_{j=1}^n \hat{\lambda}_j (Z_j^i)^2$.
    5. Determine estimation $\hat{\gamma}_{1-\beta}$ of $(1-\beta)$-quantile using $W^1,\dots,W^m$.
Output: Reject $H_0$ if $n \cdot \mathrm{wKSD}_n^2(\hat{\theta}_n) > \hat{\gamma}_{1-\beta}$.
```

2) KSD-EM algorithm: The EM algorithm is implemented to estimate the parameter θ of a probabilistic model $x \sim p(x|z,\theta)p(z|\theta)$ with an unobserved latent variable z, e.g., as in the probabilistic principal geodesic analysis (PPGA) technique [12], [13], [40] on Riemannian manifolds. In the classical EM algorithm, with samples $\boldsymbol{x}$, one generates an iterate θ^t

of θ by maximizing the log-likelihood $Q(\theta|\boldsymbol{x}, \theta^t)$ with z marginalized, i.e., $Q(\theta|\theta^t) := \mathbb{E}_{z\sim p(z|\boldsymbol{x},\theta^t)}[\log p(\theta|z, \boldsymbol{x}, \theta^t)]$. The expectation is estimated by $n^{-1}\sum_i \log p(z_i, \boldsymbol{x}|\theta^t)$, with samples z_i generated from $p(z|\boldsymbol{x}, \theta^t)$.

In applications, as the normalizing constant is most often intractable, one will encounter difficulty in either generating the samples z_i or maximizing the log-likelihood. However, one can now replace the log-likelihood loss with the weighted KSD from (8) and cope with the issue of the intractable normalizing constants in both $p(z|\boldsymbol{x}, \theta^t)$ and $p(\theta|z, \boldsymbol{x}, \theta^t)$. That is, sample $z_1, \dots, z_n$ from the distribution ω, and minimize

$$\begin{aligned} Q_{\mathrm{wKSD}}(\theta|\theta^t) := \frac{1}{n^2}\sum_{i,j=1}^{n} k_{p(z|\theta,\boldsymbol{x})}(z_i, z_j) \cdot p(z_i|\boldsymbol{x}, \theta^t) \\ \cdot\, p(z_j|\boldsymbol{x}, \theta^t)\omega(z_i)^{-1}\omega(z_j)^{-1} \end{aligned} \tag{10}$$

in one shot.

The KSD-EM algorithm is summarized in the algorithm block 2.

Algorithm 2 KSD-Expectation Maximization

```
Input: population $\boldsymbol{x} \sim q$; initial value $\theta^0$; iteration count $t = 0$; error tolerance $\epsilon$.
while $|Q_{\mathrm{wKSD}}(\theta^t|\theta^{t-1}) - Q_{\mathrm{wKSD}}(\theta^{t-1}|\theta^{t-2})| > \epsilon$ do
    1. Sample $z_1, \dots, z_n \sim \omega$.
    2. Compute $\theta^{t+1} := \mathrm{argmin}_\theta\, Q_{\mathrm{wKSD}}(\theta|\theta^t)$ numerically or analytically.
    3. $t \leftarrow t + 1$.
end while
Output: $\theta^t$.
```

V. KSD and MKSDE on $\mathrm{SO}(N)$

As $\mathrm{SO}(N)$ is one of the most widely encountered Lie group in applications, we demonstrate the mechanics of deriving the function $k_p(\cdot, \cdot)$ in (4) and closed form of MKSDE obtained by minimizing (8) for commonly used distribution families on $\mathrm{SO}(N)$ namely, the exponential family and the Riemannian normal distribution.

a) Notations: For notational convenience, we set up some notations that are summarized in table I.

TABLE I
Pertinent notation for $\mathrm{SO}(N)$ example

$\|A\|_F$	the Frobenius norm $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$ of A
$\mathscr{A}(A)$	the skew-symmetrization $\mathscr{A}(A) = (A - A^\top)/2$ of A
$\mathrm{vec}(A)$	vectorization of a matrix A by stacking the columns
$\otimes$	the Kronecker product, see [41] for example
Log	the matrix logarithm
$S_{N,N}$	the perfect shuffle matrix [41], defined as $(S_{N,N})_{ab} = 1$ whenever $a = Nk + l - N$ and $b = Nl + k - N$ for some $1 \leq k, l \leq N$, and equals 0 otherwise.
∇_X	the Euclidean gradient $\nabla_X f \in \mathbb{R}^{N\times N}$ of a differentiable function on $\mathrm{SO}(N)$ is defined as a $N \times N$ matrix such that $Df(X) = \mathrm{tr}[D^\top \nabla_X f]$ for all $D \in T_X\,\mathrm{SO}(N)$

b) Commonly-used distribution families: We now introduce the commonly-used distribution families in Lie Groups:

- *Exponential family*:
$$p(X|\theta) \propto \exp(\theta^\top \zeta(X) + \eta(X)), \quad \theta \in \mathbb{R}^m, m \in \mathbb{N},$$
where $\zeta : \mathrm{SO}(N) \to \mathbb{R}^m$ and $\eta : \mathrm{SO}(N) \to \mathbb{R}$ are continuously differentiable. Note that the widely used von Mises-Fisher family given by,
$$p(X|F) \propto \exp(\mathrm{tr}(F^\top X)), \quad F \in \mathbb{R}^{N\times N},$$
is a member of the exponential family, as $\mathrm{tr}(F^\top X) = \mathrm{vec}(F)^\top \mathrm{vec}(X)$.
- *Riemannian normal family*: For $\bar{X} \in \mathrm{SO}(N)$, $\sigma > 0$,
$$p(X|\bar{X}, \sigma) \propto \exp(-\frac{1}{2\sigma^2}\|\mathrm{Log}(\bar{X}^\top X)\|_F^2).$$
It should be noted that the Riemannian normal is *not* a member of the exponential family as it uses the intrinsic distance function on $\mathrm{SO}(N)$.

c) Choice of kernel on $\mathrm{SO}(N)$*:* Earlier, we chose as our C_0-universal kernel, the bivariate function $\exp(-\tau/2||x - y||^2), \tau > 0$ restricted to a closed subset $X \subset \mathbb{R}^m$. We now choose X to be $\mathrm{SO}(N)$, since $\mathrm{SO}(N)$ is closed in $\mathbb{R}^{N\times N}$, the Gaussian kernel $e^{-\frac{\tau}{2}\mathrm{tr}((X-Y)^\top(X-Y))}$ restricted to $\mathrm{SO}(N)$ is C_0-universal, which becomes

$$k(X, Y) = \exp(\tau\,\mathrm{tr}(X^\top Y)), \quad X, Y \in \mathrm{SO}(N),$$

where $\tau > 0$ is arbitrary. As mentioned earlier, the key to the choice of the kernel is the C_0-universality condition and one can choose other kernels such as the Laplace or Matérn, both of which can be shown to satisfy the universality condition based on the results in [42], [43]. The recipe for the derivation of closed form expressions for the restriction of these kernels to $\mathrm{SO}(N)$ and the MKSDE is rather complex and tedious but in principle similar to that presented here. Hence we will only focus on the above chosen restriction of the Gaussian kernel to $\mathrm{SO}(N)$.

d) Vector field basis on $\mathrm{SO}(N)$*:* Let E_{ij}, $1 \leq i < j \leq N$ be the matrix with all zeros except $\sqrt{2}/2$ at the $(i, j)^{th}$ entry and $-\sqrt{2}/2$ at $(j, i)^{th}$ entry of the matrix. Then XE_{ij}, $X \in \mathrm{SO}(N)$, is a standard orthonormal left-invariant basis on $\mathrm{SO}(N)$.

A. *k_p functions on* $\mathrm{SO}(N)$

Let $\mathcal{A}_p^{ij}$ represent the operator component corresponding to XE_{ij}. Clearly,

$$\begin{aligned} &\tilde{\mathcal{A}}_p^{ij}\mathcal{A}_p^{ij} k(X, Y) \\ &= e^{\tau\,\mathrm{tr}(X^\top Y)} \cdot \mathrm{tr}[(\nabla_X \log p + \tau Y)^\top X E_{ij}] \\ &\cdot \mathrm{tr}[(\nabla_Y \log p + \tau X)^\top Y E_{ij}] \\ &+ \tau\,\mathrm{tr}(E_{ij}^\top X^\top Y E_{ij}) \cdot e^{\tau\,\mathrm{tr}(X^\top Y)}. \end{aligned} \tag{11}$$

Recall that E_{ij} is an orthonormal basis of the space $\mathrm{Skew}(N)$ of all skew-symmetric matrices with the inner product $\langle\cdot, \cdot\rangle := \mathrm{tr}(\cdot^\top \cdot)$. Note that $\mathrm{tr}(A^\top E_{ij})$ is the inner product of A and E_{ij}, which actually equals $\mathrm{tr}(\mathscr{A}(A)^\top E_{ij})$, as $\mathscr{A}(A)$ is the orthogonal projection of A onto the space $\mathrm{Skew}(N)$. Furthermore, for any two matrices A and B,

$$\begin{aligned}&\sum_{i<j}\operatorname{tr}(A^\top E_{ij})\operatorname{tr}(B^\top E_{ij})\\&=\sum_{i<j}\operatorname{tr}(\mathscr{A}(A)^\top E_{ij})\operatorname{tr}(\mathscr{A}(B)^\top E_{ij})\\&=\sum_{i<j}\langle\mathscr{A}(A),E_{ij}\rangle\cdot\langle\mathscr{A}(B),E_{ij}\rangle\\&=\langle\mathscr{A}(A),\mathscr{A}(B)\rangle=\operatorname{tr}[\mathscr{A}(A)^\top\mathscr{A}(B)].\end{aligned}$$

Therefore, the first term on the RHS in (11) becomes

$$\begin{aligned}&\operatorname{tr}[\mathscr{A}(X^\top(\nabla_X\log p+\tau Y))^\top\mathscr{A}(Y^\top(\nabla_Y\log p+\tau X))]\\&\cdot e^{\tau\operatorname{tr}(X^\top Y)}.\end{aligned}$$

Moreover, note that $E_{ij}E_{ij}^\top$ equals the matrix which has all 0s except two $(1/2)$s at $(i,i)^{th}$ and $(j,j)^{th}$ entry respectively. The sum of $E_{ij}E_{ij}^\top$, $1\le i<j\le N$ actually equals $\frac{N-1}{2}I$. Recall that one can commute matrices inside $\operatorname{tr}(\cdot)$, thus the second term on the RHS of (11) becomes,

$$c(X,Y):=\frac{\tau}{2}(N-1)\operatorname{tr}(X^\top Y)e^{\tau\operatorname{tr}(X^\top Y)}.$$

Combining the two terms we obtain

$$\begin{aligned}k_p(X,Y)=\operatorname{tr}[&\mathscr{A}(X^\top(\nabla_X\log p+\tau Y))^\top\\&\mathscr{A}(Y^\top(\nabla_Y\log p+\tau X))]\cdot e^{\tau\operatorname{tr}(X^\top Y)}\\&+\frac{\tau}{2}(N-1)\operatorname{tr}(X^\top Y)e^{\tau\operatorname{tr}(X^\top Y)}\end{aligned}\tag{12}$$

Note that the last term, denoted as $c:=c(X,Y)$ in what follows is independent of the distribution p, and hence the parameters of p. We are now ready to substitute the expression of the gradient of $\log p$ for the exponential family and the von Mises-Fisher in particular, as well as the Riemannian normal into the above expression for the chosen kernel on $\mathrm{SO}(N)$ leading to closed form expressions of the kernel for each case.

a) Exponential family: Substituting the Euclidean gradient in to (12) we have

$$\begin{aligned}&k_\theta^{\mathrm{Exp}}(X,Y)\\&=\operatorname{tr}[\mathscr{A}(X^\top(\nabla_X(\theta^\top\zeta(X)+\eta(X))+\tau Y))^\top\\&\mathscr{A}(Y^\top(\nabla_Y(\theta^\top\zeta(Y)+\eta(Y))+\tau X))]\cdot e^{\tau\operatorname{tr}(X^\top Y)}\\&+c(X,Y).\end{aligned}\tag{13}$$

b) von Mises-Fisher: The Euclidean gradient of $\log p=\operatorname{tr}(F^\top X)$ is exactly F. Substituting this in to (12) we have,

$$\begin{aligned}&k_F^{\mathrm{vMF}}(X,Y)\\&=\operatorname{tr}[\mathscr{A}(X^\top(F+\tau Y))^\top\mathscr{A}(Y^\top(F+\tau X))]\\&\cdot e^{\tau\operatorname{tr}(X^\top Y)}+c(X,Y)\end{aligned}\tag{14}$$

c) Riemannian normal: Let $\varsigma=\sigma^{-2}$. The Euclidean gradient of $\log p=-\frac{\varsigma}{2}\operatorname{tr}[\operatorname{Log}(\bar{X}^\top X)\operatorname{Log}(\bar{X}^\top X)]$ at X is $\nabla_X\log p=\varsigma X\operatorname{Log}(X^\top\bar{X})$. Substituting this in to (12) leads to the following form:

$$\begin{aligned}k_{\bar{X},\varsigma}^{\mathrm{RN}}(X,Y)=c+\operatorname{tr}[&(\varsigma\operatorname{Log}(X^\top\bar{X})+\tau\mathscr{A}(X^\top Y))^\top\\&(\varsigma\operatorname{Log}(Y^\top\bar{X})+\tau\mathscr{A}(Y^\top X))]\cdot e^{\tau\operatorname{tr}(X^\top Y)}.\end{aligned}\tag{15}$$

B. Closed Form of MKSDE for the exponential family on $\mathrm{SO}(N)$

A closed form solution of MKSDE will evidently reduce the computational cost. Apparently, there is no general solution for any distribution family, since the parametrization $\alpha\mapsto p_\alpha$ of a family, as a map from the parameter space to the space of distributions, can be in any one of many different forms.

However, if $\log p_\theta$ is linearly parametrized by the parameter θ, i.e., $\log p_\theta=\zeta(X)^\top\theta+\eta(X)+C$, for some ζ and η (which gives us exactly the exponential family), then the k_p function in (12) on $\mathrm{SO}(N)$ is a quadratic form of θ, so that the MKSDE obtained by minimizing the weighted KSD in (8) will have a closed form.

In this section, we derive the closed form of MKSDE for the exponential family on $\mathrm{SO}(N)$. For notational convenience, we denote $\Pi_X=\nabla_X\log p(X)$. First, we split k_p in (12) into several terms of different orders of Π_X and Π_Y:

$$\begin{aligned}&k_\theta^{\mathrm{Exp}}(X,Y)e^{-\tau\operatorname{tr}(X^\top Y)}\\&=C'+\operatorname{tr}[\mathscr{A}(X^\top\Pi_X)^\top\mathscr{A}(Y^\top\Pi_Y)]\\&+\tau\operatorname{tr}[\mathscr{A}(Y^\top X)^\top X^\top\Pi_X]+\tau\operatorname{tr}[\mathscr{A}(X^\top Y)^\top Y^\top\Pi_Y]\\&=C'+\underbrace{\frac{1}{2}\operatorname{tr}[\Pi_X^\top XY^\top\Pi_Y]}_{(I)}-\underbrace{\frac{1}{2}\operatorname{tr}[X^\top\Pi_XY^\top\Pi_Y]}_{(II)}\\&+\tau\underbrace{\operatorname{tr}[\mathscr{A}(Y^\top X)^\top X^\top\Pi_X]}_{(III)}+\tau\underbrace{\operatorname{tr}[\mathscr{A}(X^\top Y)^\top Y^\top\Pi_Y]}_{(IV)}\end{aligned}$$

The constant C' is independent of the Π_X and Π_Y, thus also independent of the parameters of p. The second-order terms will equal

$$\begin{aligned}(I)&=\frac{1}{2}[(I\otimes X^\top)\operatorname{vec}(\Pi_X)]^\top[(I\otimes Y^\top)\operatorname{vec}(\Pi_Y)]\\&=\frac{1}{2}\operatorname{vec}(\Pi_X)^\top(I\otimes XY^\top)\operatorname{vec}(\Pi_Y),\\(II)&=\frac{1}{2}\operatorname{vec}(\Pi_X)^\top(Y^\top\otimes X)\operatorname{vec}(\Pi_Y^\top)\\&=\frac{1}{2}\operatorname{vec}(\Pi_X)^\top(Y^\top\otimes X)S_{N,N}\operatorname{vec}(\Pi_Y).\end{aligned}$$

The first-order terms will equal

$$\begin{aligned}(III)&=\operatorname{vec}(X\mathscr{A}(Y^\top X))^\top\operatorname{vec}(\Pi_X),\\(IV)&=\operatorname{vec}(Y\mathscr{A}(X^\top Y))^\top\operatorname{vec}(\Pi_Y).\end{aligned}$$

Combining all the terms, we have

$$\begin{aligned}&k_\theta^{\mathrm{Exp}}(X,Y)e^{-\tau\operatorname{tr}(X^\top Y)}\\&=\frac{1}{2}\operatorname{vec}(\Pi_X)^\top(I\otimes XY^\top-Y^\top\otimes X\cdot S_{N,N})\operatorname{vec}(\Pi_Y)\\&+\tau\operatorname{vec}(X\mathscr{A}(Y^\top X))^\top\operatorname{vec}(\Pi_X)\\&+\tau\operatorname{vec}(Y\mathscr{A}(X^\top Y))^\top\operatorname{vec}(\Pi_Y)+C'.\end{aligned}\tag{16}$$

Next, we represent Π_X and Π_Y by θ. As $\theta^\top\zeta(X)$ is a scalar functions on $\mathrm{SO}(N)$, its Euclidean gradient $\nabla_X(\theta^\top\zeta(X))$ is a $N\times N$ matrix. However, if we extract θ from the Euclidean gradient, i.e., $\nabla_X(\theta^\top\zeta(X))=\theta^\top\nabla_X\zeta(X)$, then the Euclidean gradient $\nabla_X\zeta(X)$ will be a *3D matrix* of dimension $m\times N\times N$, since $\zeta(X)$ is a vector-valued function of X. To appropriately tackle this issue, we vectorize the $N \times N$ dimensional component of $\nabla_X \zeta(X)$, so that it is represented by a $m \times N^2$-dimensional 2D matrix, as elaborated upon next.

We denote $\zeta(X) := (\zeta_1(X), \dots, \zeta_m(X))^\top$, then each of $\nabla_X \zeta_i(X)$ is a $N \times N$ matrix. We vectorize each $\nabla_X \zeta_i(X)$ and stack them by rows to get

$$\mathscr{Z}(X) := \begin{bmatrix} \operatorname{vec}(\nabla_X \zeta_1(X))^\top \\ \vdots \\ \operatorname{vec}(\nabla_X \zeta_m(X))^\top \end{bmatrix} \in \mathbb{R}^{m \times N^2}.$$

Note that $Z(X)^\top \theta = \operatorname{vec}(\nabla_X[\theta^\top \zeta(X)])$. Therefore, we have $\operatorname{vec}(\Pi_X) = \mathscr{Z}(X)^\top \theta + \operatorname{vec}(\nabla_X \eta(X))$. Substituting this into (16) we get

$$\begin{aligned} & k_\theta^{\mathrm{Exp}}(X,Y) e^{-\tau \operatorname{tr}(X^\top Y)} \\ & = \frac{1}{2} \theta^\top \mathscr{Z}(X)(I \otimes XY^\top - Y^\top \otimes X \cdot S_{N,N}) \mathscr{Z}(Y)^\top \theta \\ & + \frac{1}{2} \operatorname{vec}(\nabla_X \eta(X))^\top (I \otimes XY^\top - Y^\top \otimes X \cdot S_{N,N}) \mathscr{Z}(Y)^\top \theta \\ & + \frac{1}{2} \operatorname{vec}(\nabla_Y \eta(Y))^\top (I \otimes YX^\top - X^\top \otimes Y \cdot S_{N,N}) \mathscr{Z}(X)^\top \theta \\ & + \tau \operatorname{vec}(X \mathscr{A}(Y^\top X))^\top \mathscr{Z}(X)^\top \theta \\ & + \tau \operatorname{vec}(Y \mathscr{A}(X^\top Y))^\top \mathscr{Z}(Y)^\top \theta. \end{aligned} \tag{17}$$

Although above formula is rather lengthy and monstrous, it is noteworthy that the first-order terms can be combined together by the summation in the weighted KSD in (8), as they are symmetric with respect to X and Y. We summarize the results for the exponential family in the following theorem:

Theorem 5.10. *Suppose $X_i \in \mathrm{SO}(N)$ are samples from w and p_θ is the exponential family on $\mathrm{SO}(N)$. Given kernel $k(X,Y) = \exp(\tau \operatorname{tr}(X^\top Y))$ on $\mathrm{SO}(N)$, the global minimizer of the weighted KSD in (8) has a closed form given below:*

$$\begin{aligned} \textit{Let } b = & \frac{1}{2n^2} \sum_{i,j} \operatorname{vec}(\nabla_{X_i} \eta(X_i))^\top (I \otimes X_i X_j^\top \\ & - X_j^\top \otimes X_i \cdot S_{N,N}) \mathscr{Z}(X_j)^\top \\ & \cdot e^{\tau \operatorname{tr}(X_i^\top X_j)} \cdot \frac{q(X_i)}{w(X_i)} \cdot \frac{q(X_j)}{w(X_j)} \\ & + \frac{\tau}{n^2} \sum_{i,j} \operatorname{vec}(X_i \mathscr{A}(X_j^\top X_i))^\top \mathscr{Z}(X_i)^\top \\ & \cdot e^{\tau \operatorname{tr}(X_i^\top X_j)} \frac{q(X_i)}{w(X_i)} \cdot \frac{q(X_j)}{w(X_j)} \\ \textit{and } A = & \frac{1}{2n^2} \sum_{i,j} \mathscr{Z}(X_i)(I \otimes X_i X_j^\top \\ & - X_j^\top \otimes X_i \cdot S_{N,N}) \mathscr{Z}(X_j)^\top \\ & \cdot e^{\tau \operatorname{tr}(X_i^\top X_j)} \frac{q(X_i)}{w(X_i)} \cdot \frac{q(X_j)}{w(X_j)}. \end{aligned} \tag{18}$$

Then, $\hat{\theta}_n = -A^{-1} b$. For the unweighted case, we just ignore the weighted ratio $\frac{q(X_i) q(X_j)}{w(X_i) w(X_j)}$.

Proof. Since (17) is a quadratic function of θ, the weighted KSD in (8), which is a weighted sum of k_p function in (17), will remain a quadratic function of θ. Therefore, the global minimizer is $\hat{\theta}_n = -A^{-1} b$. □

Example 1 (MKSDE of vMF)**.** *We now consider the MKSDE of the von Mises-Fisher family on $\mathrm{SO}(N)$. Note that $\log p = \operatorname{tr}(F^\top X) + C = \operatorname{vec}(F)^\top \operatorname{vec}(X)$, so that von Mises-Fisher family is a member of the exponential family if we set $\theta := \operatorname{vec}(F)$. Since $\zeta(X) = \operatorname{vec}(X)$ and $\eta(X) = 0$ in this case, we have $\mathscr{Z}(X) = I_{N^2 \times N^2}$ and $\nabla_X \eta(X) = 0$. Substituting this into (5.10) yields the closed form of MKSDE given below,*

$$\begin{aligned} \textit{Let } b = & \frac{\tau}{n^2} \sum_{i,j} \operatorname{vec}[X_i \mathscr{A}(X_i^\top X_j)] \\ & e^{\tau \operatorname{tr}(X_i^\top X_j)} \cdot \frac{q(X_i)}{w(X_i)} \cdot \frac{q(X_j)}{w(X_j)}, \\ \textit{and } A = & \frac{1}{2n^2} \sum_{i,j} [I \otimes X_i X_j^\top - (X_i^\top \otimes X_j) S_{N,N}] \\ & \cdot e^{\tau \operatorname{tr}(X_i^\top X_j)} \cdot \frac{q(X_i)}{w(X_i)} \cdot \frac{q(X_j)}{w(X_j)}. \end{aligned} \tag{19}$$

Then, $\widehat{F}_{\mathrm{wKSD}} = \operatorname{vec}^{-1}(A^{-1} b)$. For the unweighted case, we just ignore the weighted ratio $\frac{q(X_i) q(X_j)}{w(X_i) w(X_j)}$.

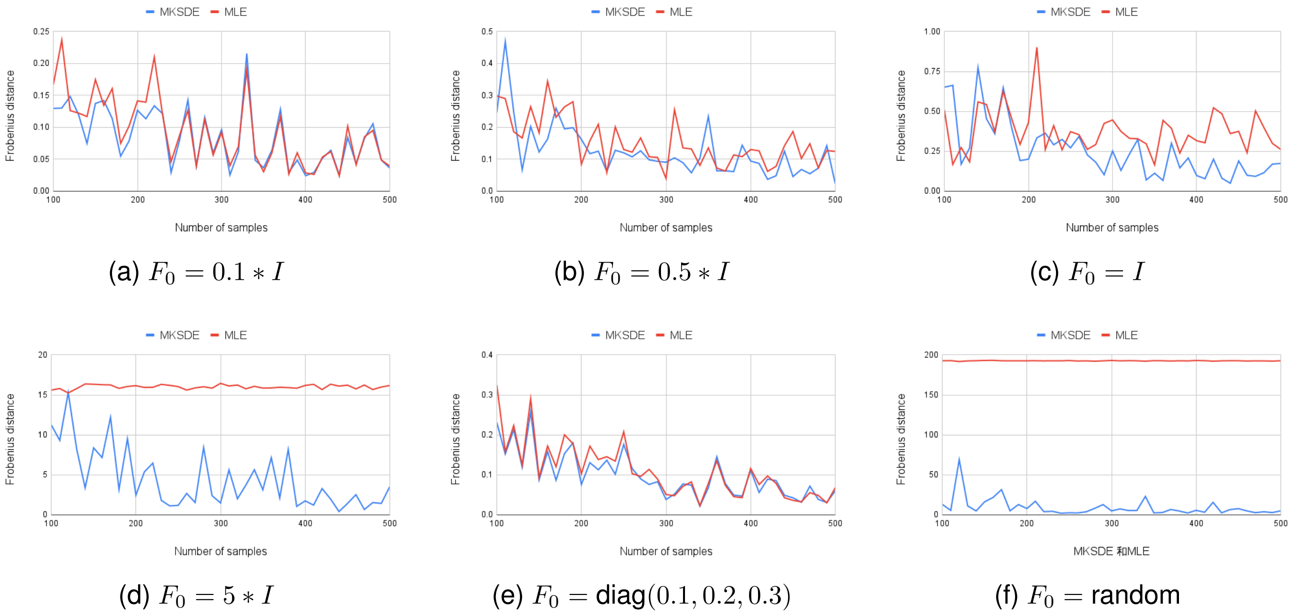
VI. Experiments

In this section we present two experiments to validate the advantage of our KSD over MLE. In the first experiment, we compare the MKSDE of F, a parameter of the vMF presented in example 1, to its MLE, illustrating that the issue of normalizing constant impairs the accuracy of MLE but has no affect on the MKSDE. In this experiment, based on estimated parameter F, we also estimate the orientation of a 3D object from a publicly available database, ModelNet10 [44]. In the second experiment, we compare the Cayley distribution [45] to the vMF via a goodness of fit test introduced in §IV-C1 to illustrate the power of our test. In these experiments, we set $\tau = 1$. Code for all the experiments in this paper is provided on GitHub at https://github.com/cvgmi/KSD-on-Lie-Groups.

A. MKSDE vs. MLE

a) vMF Parameter Estimation: The exact numerical solution of MLE for vMF requires the computation of the inverse of the derivative of the normalizing constant, a hypergeometric function of the parameter F. The commonly used MLE technique [46, §13.2.3] uses two direct approximate solutions, one is for the case when F is small while the other is for large F. Although the solution for small F is reasonably accurate, the solution method for large F is cumbersome and hard to implement, and both solutions poorly approximate for in-between values of F. Therefore, one must bear with either the inaccuracy of a direct approximate solution or the computational cost of achieving an convergent numerical solution.

Figure 1 depicts the Frobenius distance of $\hat{F}_{\mathrm{KSD}}$ and $\hat{F}_{\mathrm{MLE}}$ to the ground truth F_0, with varying values of F_0 and varying sample size n. For medium valued F_0, e.g., $F = 5I$, we used the approximation for small F_0. The last "random" F_0 is drawn randomly. It is evident that the accuracy of MLE decreases



(a) $F_0 = 0.1 * I$ (b) $F_0 = 0.5 * I$ (c) $F_0 = I$

(d) $F_0 = 5 * I$ (e) $F_0 = \text{diag}(0.1, 0.2, 0.3)$ (f) $F_0 = \text{random}$

Fig. 1. F-norms between estimators and ground truth

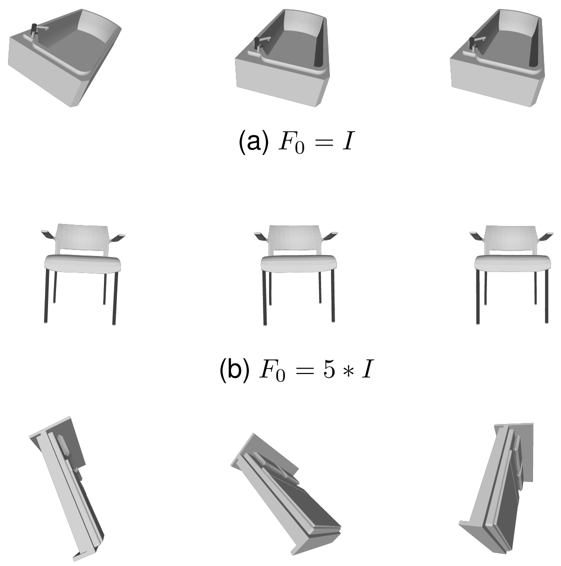
as F_0 becomes larger and the approximation worsens, while MKSDE remains accurate for all values of F_0. This result demonstrates the accuracy and stability of MKSDE over MLE.

b) Object Orientation Estimates: We now describe an experiment where samples are drawn from a vMF on $\text{SO}(3)$, with known parameter F_0. The vMF can then be sampled to generate distinct orientations, $X_i \in \text{SO}(3)$ which are applied to objects in the ModelNet10 [44] database of CAD (computer aided design) models and sample models are depicted in figure 2. Now, the goal of this experiment is to estimate the parameter F given the samples using MKSDE and MLE yielding $\hat{F}_{\text{KSD}}$ and $\hat{F}_{\text{MLE}}$ respectively, which can then be compared. Note that once the parameter F is estimated from the samples, we can then compute an estimate of the mode of the distribution (see [47]) which we will denote here by $\hat{X}$, given by $\hat{X} = U\text{diag}(1, 1, \det(UV))V^\top$, where U, V are orthogonal matrices obtained from the singular value decomposition (SVD) of $\hat{F}$. We can then easily compare the ground truth object orientation of the CAD models in the data base to estimated object orientations (that will correspond to the estimated mode of the vMF). Before presenting these object orientation estimates, we present comparisons between mKSDE and MLE estimates of the parameter F.

We now consider the estimated F ($\hat{F}_{\text{KSD}}$ and $\hat{F}_{\text{MLE}}$) obtained from a sample size of 500. In table II, we report the geodesic distance between the mode of ground truth orientation for the samples shown in Fig 2 and the mode of estimated F. From the table, it is evident that the gedesic distance of MKSDE estimates are consistently smaller than those for MLE indicating a superior performance of MKSDE.

B. MKSDE goodness of fit test

In this experiment, we measure the difference between a specific Cayley distribution to the vMF family. The Cayley



(a) $F_0 = I$

(b) $F_0 = 5 * I$

(c) $F_0 = \text{diag}(0.1, 0.2, 0.3)$

Fig. 2. Sampled rotation matrices from a matrix Fisher distribution with a varying parameter F_0, applied to example objects in the ModelNet10 database.

TABLE II
GEODESIC DISTANCE BETWEEN THE MODE OF ESTIMATED F AND THE GROUND TRUTH ORIENTATION OF THE SAMPLES SHOWN IN FIGURE 2

		left	middle	right
(a) $F_0 = I$	MKSDE	**1.53**	**2.82**	**3.01**
	MLE	1.54	3.02	3.07
(b) $F_0 = 5 * I$	MKSDE	**0.93**	**1.55**	**1.95**
	MLE	2.66	2.91	3.01
(c) $F_0 = \text{diag}(0.1, 0.2, 0.3)$	MKSDE	**1.62**	**1.65**	**1.64**
	MLE	1.76	1.79	1.78

distribution [45] has the density $p(X|M) \propto \det(I + XM^\top)^\kappa$ with parameters $M \in \mathrm{SO}(3)$ and $\kappa > 0$. As the vMF family is rotationally symmetric, the parameter M does not affect the dissimilarity between a specific Cayley distribution and the vMF family. We used the R package [48] to generate $n = 500$ samples from the Cayley distribution for varying κ, and perform the goodness of fit test in §III with different levels of significance β.

Table III depicts the $(1 - \beta)$-quantile and the statistic $n\,\mathrm{wKSD}_n^2(\hat{\theta}_n)$. As discussed in [45], the Cayley distribution resembles a uniform distribution for small κ, and a spiky local Gaussian distribution for large κ. It approximately belongs to the vMF class in both cases, but differs from vMF for κ in-between. This coincides with the results in table III.

TABLE III
MKSDE GOODNESS OF FIT

κ	0.2	0.5	1.0	1.5	2.0
$n \cdot \mathrm{wKSD}_n^2(\hat{\theta}_n)$	68.65	90.25	99.99	99.29	104.23
$\beta = 0.01$	83.98	94.23	108.18	116.88	139.76
$\beta = 0.05$	77.62	83.21	96.99	104.18	114.47
$\beta = 0.10$	73.22	77.79	90.71	96.46	107.12

VII. Conclusions

In this paper, we presented a novel Stein's operator defined on Lie Groups leading to a new kernel Stein discrepancy (KSD) which is a normalization-free loss function. We presented several theoretical results characterizing the properties of this new KSD on Lie groups and the MKSDE. We presented new theorems on MKSDE being strongly consistent and asymptotically normal, and a closed form expression of the MKSDE for the exponential family and in particular the vMF distribution on $\mathrm{SO}(N)$. Furthermore, we presented two algorithms, namely MKSDE goodness of fit and KSD-EM, for measuring the quality of model fitting and distributional parameter estimation with latent variables respectively. Finally, we presented several experiments demonstrating the advantages of MKSDE over MLE. Our future work will focus on exploring the practical implications of the choice of C_0-universal kernels in characterizing the Stein class.

Appendix

A. Proof of theorem 3.4

Theorem (Characterization)**.** *Suppose k is C_0-universal and $\sqrt{k_p(x,x)}$ is integrable with respect to locally Lipschitz densities p and q. Suppose further $D^l \log(p/q)$ is p and q-integrable for all l. Then $p = q \iff \mathrm{KSD}(p,q) = 0$.*

Proof. We begin with the forward implication, "$\Rightarrow$". By [34, Cor. 4.36], we have $|\mathcal{A}_p f(x)| \leq \|f\|_{\mathcal{H}_k} \cdot \sqrt{k_p(x,x)}$ for $f \in \mathcal{H}_k$. Moreover, the RKHS of a C_0-universal kernel contains only C_0, thus bounded functions, see [32, Prop. 2.3.2]. Therefore, for all $f \in \mathcal{H}_k$, $f, \mathcal{A}_p f$ are both q-integrable. Let div be the divergence operator with respect to the left-invariant Riemannian metric. Recall that $\mathrm{div}(D^l) = D^l \Delta$, thus $\mathrm{div}(p f_l D^l) = p \cdot f_l D^l \Delta + p \cdot D^l f_l + f_l \cdot D^l p = \mathcal{A}_p^l f_l \cdot p$. By generalized Stokes's theorem [49], if an integrable vector field on a complete manifold has integrable divergence, then its divergence integrates to 0. Therefore, $\mathbb{E}_p[\mathcal{A}_p f] = \int_G \mathcal{A}_p f \cdot p d\mu = \int_G \mathrm{div}(\sum_{l=1}^d p f_l D^l) d\mu = 0$. For readers unfamiliar with the divergence operator or Stokes's theorem, this proof maybe be interpreted as a general version of integration by parts on manifolds.

Now, we address the reverse implication, "$\Leftarrow$". Note that $\mathbb{E}_q[\mathcal{A}_p f] = \mathbb{E}_q[\mathcal{A}_p f - \mathcal{A}_q f] = \sum_l \mathbb{E}_q[f_l D^l \log(p/q)]$, since the operator $\mathcal{A}_q$ corresponding to q also has the Stein's identity. If k is C_0-universal, then $\mathcal{H}_k$ is dense in the space of continuous functions vanishing at infinity [32, Thm. 4.1.1]. Since $\mathbb{E}_q[f^l D^l \log(p/q)] = 0$ for all $f \in \mathcal{H}_k^d$, we have $D^l \log(p/q) = 0$ for all l. As we assume G is connected, $\log(p/q)$ is constant and thus $p = q$. □

B. Proof of theorem 4.6

Theorem. *Suppose $V_{(\cdot)}(\cdot,\cdot)$ is jointly continuous and satisfies that $\sup_{\theta \in K} V_\theta(x,x)$ is w-integrable for any compact $K \subset \Theta$, then $\mathrm{wKSD}_n^2(\theta) \to \mathrm{KSD}^2(\theta)$ compactly almost surely, i.e., for any compact K,*

$$\mathrm{wKSD}_n^2(\theta) \to \mathrm{KSD}^2(\theta) \textit{ uniformly on } K, \quad \textit{almost surely}.$$

As a corollary, if Θ is locally compact, wKSD_n and KSD are all continuous on Θ.

We prove a lemma first.

Lemma 1. *Suppose $f(\theta, x)$ is continuous on $\Theta \times G$ and $\sup_{\theta \in K} |f(\theta, x)|$ is P-integrable for all compact $K \subset \Theta$. Then*

$$n^{-1} \sum_{i=1}^n f(\theta, X_i) \to F(\theta) := \mathbb{E}[f(\theta, X_1)], \quad X_i \overset{i.i.d}{\sim} P.$$

compactly P-almost surely on Θ and thus $F(\theta)$ is continuous.

Proof. The lemma relies on a classical results [37, Thm. 1] regarding the uniform convergence of random functions. It suffices to check the condition of equi-continuity. Take a sequence of compact sets G_n such that $G = \bigcup G_n$, then the join continuity of $f(\cdot,\cdot)$ will implies the joint uniform continuity of $f(\theta, x)$ on $K \times G_n$, which further implies equi-continuity. □

Next we prove the theorem:

Proof. Note that,

$$\begin{aligned} \mathrm{wKSD}_n(\theta) :&= \frac{1}{n^2} \sum_{i=1}^n V_\theta(x_i, x_i) \\ &+ \frac{n-1}{n} \frac{1}{n(n-1)} \sum_{i \neq j} V_\theta(x_i, x_j) \end{aligned}$$

The first term tends to 0 compactly almost surely by lemma 1. For the second term, we apply a result [38, Thm. 1] regarding the uniform convergence of U-statistic to conclude that

$$\frac{1}{n(n-1)} \sum_{i \neq j} V_\theta(x_i, x_j) \to \mathbb{E}_{X,Y \sim p}[V_\theta(X, Y)]$$

compactly almost surely. These two statements together will imply the theorem. It suffices to test the conditions (i), (ii) and (iii) in [38, Thm. 1].

For condition (i), note that $V_\theta(\cdot,\cdot)$ is also a positive definite bivariate function, thus

$$\sup_{\theta\in K}|V_\theta(x,y)| \le \sup_{\theta\in K} V_\theta(x,x)^{1/2}\cdot \sup_{\theta\in K} V_\theta(y,y)^{1/2},$$

Recall that the w-integrability of $\sup_{\theta\in K} V_\theta(x,x)$ implies the w-integrability of $\sup_{\theta\in K} V_\theta(x,x)^{1/2}$ by Jensen's inequality. Therefore, $\sup_{\theta\in K}|V_\theta(x,y)|$ is $w\times w$-integrable. Therefore, the condition (i) holds.

For condition (ii), we take a sequence of compact subset G_n such that $G=\bigcup G_n$.

For condition (iii), note that $K\times G_n\times G_n$ is compact, thus $V_\theta(x,y)$ is uniformly continuous on $K\times G_n\times G_n$, thus equi-continuous in θ for $(x,y)\in G_n\times G_n$. Next we show the equi-continuity of $\mathbb{E}_{Y\sim P}[V_\theta(x,Y)]$ on $K\times G_n$. Given a fixed $(\theta_0,x_0)\in K\times G_n$, we choose a compact neighborhood $W\subset G$ of x_0. Since $V_\theta(x,y)$ is jointly continuous, $V_\theta(x,x)$ is bounded on the compact set $K\times W$ and we denote $C:=\sup_{\theta\in K,x\in W}[V_\theta(x,x)]^{1/2}<+\infty$. For any $\theta\in K$, $x\in W$ and $y\in G$, we have

$$\begin{aligned}|V_\theta(x,y)| &\le \sup_{\theta\in K,x\in W}[V_\theta(x,x)]^{1/2}\cdot\sup_{\theta\in K}[V_\theta(y,y)]^{1/2}\\ &= C\cdot\sup_{\theta\in K}[V_\theta(y,y)]^{1/2}\end{aligned}$$

For any $(\theta',x')\in K\times W$ that tends to (θ_0,x_0), we have $|V_{\theta'}(x',Y)-V_{\theta_0}(x_0,Y)|\le 2C\sup_{\theta\in K}[V_\theta(y,y)]^{1/2}$. Since the right hand side is a P-integrable function of y as the condition states, we have

$$\begin{aligned}&|\mathbb{E}_{Y\sim P}[V_{\theta'}(x',Y)]-\mathbb{E}_{Y\sim P}[V_{\theta_0}(x_0,Y)]|\\ &\qquad\le \mathbb{E}_{Y\sim P}|V_{\theta'}(x',Y)-V_{\theta_0}(x_0,Y)|\\ &\qquad\to 0,\quad \text{as } (\theta',x')\to(\theta_0,x_0).\end{aligned}$$

Therefore, $\mathbb{E}_{Y\sim w}[V_{(\cdot)}(\cdot,Y)]$ restricted on $K\times G$ is jointly continuous at (θ_0,x_0), and thus continuous on $K\times G$ by the arbitrariness of θ_0 and x_0. It is thus uniformly continuous on $K\times G_n$ and hence the condition of equi-continuity holds.

The continuity of wKSD_n is straightforward since it is a finite summation of continuous functions. Since KSD is the uniform limit of wKSD_n on any compact set, KSD is continuous on any compact set. Therefore, KSD is continuous as continuity is a local property. This completes the proof. □

C. Proof of theorem 4.7

Theorem (Strong consistency)**.** *Suppose the conditions in theorem 4.2 hold and $\Theta=\Theta_1\times\Theta_2$ such that Θ_1 is compact, Θ_2 is convex, and for each fixed $\theta_1\in\Theta_1$, $\mathrm{wKSD}_n(\theta_1,\cdot)$ is convex on Θ_2 and $\mathrm{KSD}(\theta_1,\cdot)$ attains minimum value on a non-empty and compact set $\tilde\Theta_0(\theta_1)\subset int(\Theta_2)$. Then Θ_0, $\widehat\Theta_n$ are non-empty for large n and $\sup_{\theta\in\widehat\Theta_n} d(\theta,\Theta_0)\to 0$ almost surely.*

We now establish some lemmas that will be needed for the proof.

Lemma 2. *Suppose f is convex and C is a bounded open set. If there exists $x_0\in C$ such that $f(x_0)<\inf_{x\in\partial C}f(x)$, then $\inf_{x\notin C}f(x)=\inf_{x\in\partial C}f$ and the set of global minimizers of f is non-empty and contained in C.*

Proof. For all $x'\notin C$, there exists $0\le\beta<1$ such that $\beta x_0+(1-\beta)x'\in\partial C$, and thus

$$\begin{aligned}\beta f(x_0)+(1-\beta)f(x') &\ge f(\beta x_0+(1-\beta)x')\\ &> f(x_0)\Longrightarrow f(x')>f(x_0).\end{aligned}$$

Furthermore,

$$\begin{aligned}f(x') &\ge \beta f(x_0)+(1-\beta)f(x')\\ &\ge f(\beta x_0+(1-\beta)x')\ge \inf_{x\in\partial C}f(x).\end{aligned}$$

This concludes the proof. □

Note that Θ is locally compact, thus wKSD_n and KSD are continuous on Θ by theorem 4.3. Furthermore, as the pointwise limit preserves the convexity, $\mathrm{KSD}(\theta_1,\cdot)$ is also convex in θ_2 for all $\theta_1\in\Theta_1$. Let $m^*(\theta_1):=\inf_{\theta_2\in\Theta_2}\mathrm{KSD}(\theta_1,\theta_2)$, and denote by $\tilde\Theta_0(\theta_1)$ the set of minimizers of $\mathrm{KSD}(\theta_1,\cdot)$. Then we establish the next lemma.

Lemma 3. *For each $\bar\theta_1\in\Theta_1$, there exists a neighborhood V_1 of $\bar\theta_1$ and a bounded open set $V_2\subset\Theta_2$ such that for any $\theta_1\in V_1$, $\tilde\Theta_0(\theta_1)\subset V_2$.*

Proof. Take a bounded V_2 such that $\tilde\Theta_0(\bar\theta_1)\subset V_2$, then since $\mathrm{KSD}(\bar\theta_1,\cdot)-m^*(\bar\theta_1)$ is continuous and positive on ∂V_2, let $\epsilon:=\inf_{\theta_2\in\partial V_2}\mathrm{KSD}(\bar\theta_1,\theta_2)-m^*(\bar\theta_1)>0$. Note that $\{\bar\theta_1\}\times\partial V_2$ is a subset of the open set $\{\mathrm{KSD}(\cdot,\cdot)>\epsilon/2+m^*(\bar\theta_1)\}$, thus we apply tube lemma [50, Lem. 26.8] and get an open neighborhood V_1 of $\bar\theta_1$ such that $V_1'\times\partial V_2\subset\{\mathrm{KSD}(\cdot,\cdot)>\epsilon/2+m^*(\bar\theta_1)\}$, which implies that $\inf_{V_1'\times\partial V_2}\mathrm{KSD}>m^*(\bar\theta_1)+\epsilon/2$. However, we choose a $\tilde\theta_2\in\tilde\Theta_1(\bar\theta_1)\subset V_2$, there exists a neighborhood V_1'' of $\bar\theta_1$ such that $\sup_{\theta_1\in V_1''}\mathrm{KSD}(\theta_1,\tilde\theta_2)<m^*+\epsilon/2$, as KSD is continuous at $(\bar\theta_1,\tilde\theta_2)$. Therefore, for any $\theta_1\in V_1:=V_1'\cap V''$, we have $\mathrm{KSD}(\theta_1,\tilde\theta_2)<\inf_{\theta_2\in\partial V_2}\mathrm{KSD}(\theta_1,\theta)$. We now apply lemma 2, resulting in the minimizer set $\tilde\Theta_0(\theta_1)$ is contained in V_2. □

Now we are ready to prove the theorem. The proof is presented in several stages each of which is required to get to the desired result.

Proof. 1. First we show that Θ_0 is non-empty.

Let $m_0^*:=\inf_{\theta_1\in\Theta_1}m^*(\theta_1)=\inf_{\theta\in\Theta}\mathrm{KSD}(\theta)$. Take a sequence $\theta_{1,n}\in\Theta_1$ such that $m^*(\theta_{1,n})\downarrow m_0^*$. As Θ_1 is compact, $\theta_{1,n}$ has a convergent sub-sequence, still denoted as $\theta_{1,n}$, whose limit point is denoted as $\theta_{1,0}$. We apply lemma 3 to $\tilde\Theta_0(\theta_{1,0})$, thus there exists a compact neighborhood $V_1\times V_2$ of $\{\theta_{1,0}\}\times\tilde\Theta_0(\theta_{1,0})$ such that for all $\theta_1\in V_1$, $\tilde\Theta_0(\theta_1)\subset V_2$. Since KSD is continuous on the compact set $\{\theta_{1,0}\}\times\tilde\Theta_0(\theta_{1,0})$ and the size of $V_1\times V_2$ can be taken as small as needed, we can shrink $V_1\times V_2$ such that $\inf_{\theta\in V_1\times V_2}\mathrm{KSD}(\theta)>m^*(\theta_{1,n})-\epsilon$. However, note that $\theta_{1,n}$ will finally enter V_1, and from then on, $\{\theta_{1,n}\}\times\tilde\Theta_0(\theta_{1,n})\subset V_2$, which implies $\inf_{V_1\times V_2}\mathrm{KSD}(\theta)\le m^*(\theta_{1,n})\downarrow m_0^*$. Therefore, $m^*(\theta_{1,n})-\epsilon<\inf_{V_1\times V_2}\mathrm{KSD}(\theta)\le m^*(\theta_{1,n})\downarrow m_0^*$, and thus $m^*(\theta_{1,0})=m_0^*$ as ϵ is arbitrary. Therefore, Θ_0 is non-empty.

2. We now show that Θ_0 is compact.

Suppose $\{B_\alpha\} \subset \Theta$ is a collection of open balls such that $\Theta_0 \subset \bigcup B_\alpha$. For each θ_1, the θ_1-section $\Theta_0^{\theta_1} := \{(\theta, \theta') \in \Theta_0 | \theta = \theta_1\}$ of Θ_0, if non-empty, is exactly the set $\tilde{\Theta}_0(\theta_1)$. Since $\Theta_0^{\theta_1} = \tilde{\Theta}_0(\theta_1)$ is compact, there exists a finite subcollection of balls $B_i^{\theta_1}$, $1 \le i \le n_{\theta_1}$ such that $\{\theta_1\} \times \tilde{\Theta}_0(\theta_1) \subset \bigcup_i B_i^{\theta_1}$. We apply lemma 3 to each θ_1, and conclude there exists open neighborhoods $V_1^{\theta_1}$ and $V_2^{\theta_1}$ such that $\{\theta_1\} \times \tilde{\Theta}_0(\theta_1) \subset V_1^{\theta_1} \times V_2^{\theta_1} \subset \bigcup_i B_i^{\theta_1}$. If the section $\Theta_0^{\theta_1}$ is empty, we prescribe $V_1^{\theta_1}$ and $V_2^{\theta_2}$ with arbitrary neighborhood such that $\{\theta_1\} \times \tilde{\Theta}_0(\theta_1) \subset V_1^{\theta_1} \times V_2^{\theta_1}$. Note that $\{V_1^{\theta_1}\}_{\theta_1 \in \Theta_1}$ is a open cover of Θ_1, thus there exists an open cover $V_1^{\theta_{1,i}}$ for $\theta_{1,i}$, $1 \le i \le n$. Note that $B_j^{\theta_{1,i}}$, $1 \le i \le n$, $1 \le j \le n_{\theta_{1,i}}$ is a finite subcover of Θ_0. Therefore, Θ_0 is compact.

3. We now show that the minimizers $\hat{\theta}_n \in \widehat{\Theta}_n$ will eventually populate a compact set uniformly.

Since Θ_0 is compact, we choose an open set $K \subset \Theta_2$ with compact closure such that $\Theta_0 \subset \Theta_1 \times K$. Since $\mathrm{KSD} - m_0^*$ is continuous and positive on the compact set $\Theta_1 \times \partial K$, let $\epsilon := \inf_{\Theta_1 \times \partial K} \mathrm{KSD}(\theta) - m_0^* > 0$. By theorem 4.2, $\mathrm{wKSD}_n^2 \to \mathrm{KSD}$ uniformly on $\Theta_1 \times \bar{K}$, which implies for large enough n, $\mathrm{wKSD}_n > m_0^* + \epsilon/2$ uniformly on $\Theta_1 \times \partial K$ and $\mathrm{wKSD}_n < m_0^* + \epsilon/2$ uniformly on Θ_0. Therefore, we can apply lemma 2 to each $\mathrm{wKSD}_n(\theta_1, \cdot)$, and conclude that $\widehat{\Theta}_n \subset \Theta_1 \times K$ for large enough n.

4. We now show the theorem.

Since the global minimizers of wKSD_n will eventually populate the compact set $\Theta_1 \times \bar{K}$, WLOG, we assume that $\widehat{\Theta}_n$ is set of minimizers of wKSD_n over $\Theta_1 \times \bar{K}$.

For each $\delta > 0$, let $\Theta_\delta := \{\theta \in \Theta_1 \times \bar{K} : d(\theta, \Theta_0) < \delta\}$ be the closed δ-neighborhood of Θ_0 in $\Theta_1 \times \bar{K}$, and let $\Theta_\delta^c := \Theta_1 \times \bar{K} - \Theta_\delta$. Note that Θ_δ^c is compact, and $\mathrm{KSD} - m_0^*$ is positive on Θ_δ^c, thus we let $\epsilon_\delta := \inf_{\Theta_\delta^c} \mathrm{KSD}(\theta) - m_0^* > 0$. For n large enough such that $\sup |\mathrm{wKSD}_n - \mathrm{KSD}| < \epsilon_\delta/2$ and $\widehat{\Theta}_n \subset \Theta_1 \times \bar{K}$, we have

$$\begin{aligned} \mathrm{KSD}(\hat{\theta}_n) - m_0^* &< \mathrm{wKSD}_n(\hat{\theta}_n) + \epsilon_\delta/2 - m_0^* \\ &\le \mathrm{wKSD}_n(\theta_0) + \epsilon_\delta/2 - m_0^* \\ &< m_0^* + \epsilon_\delta - m_0^* = \epsilon_\delta, \quad \hat{\theta}_n \in \widehat{\Theta}_n, \end{aligned}$$

which implies $\hat{\theta}_n \in \Theta_\delta$, thus proving the theorem. □

D. Proof of theorem 4.9

Due to proposition 1, it suffices to show that $\sum_{l=1}^n (\hat{\lambda}_l' - \hat{\lambda}_l) Z_l^2 \to 0$ in probability. Note that $\hat{\lambda}_l'$, $\hat{\lambda}_l$, $1 \le l \le n$, are the eigenvalues of the matrices $H_0 := n^{-1}(V_{\theta_0}(x_i, x_j))_{ij}$ and $H_n := n^{-1}(V_{\hat{\theta}_n}(x_i, x_j))_{ij}$. The Weilandt-Hoffman inequality [51, Eq. (1.64)] states that

$$\sum_{l=1}^n |\hat{\lambda}_l' - \hat{\lambda}_l| \le \|H_0 - H_n\|_1,$$

where $\|H\|_1 := \operatorname{tr} \sqrt{H^\top H}$ represents the nuclear norm of a squared matrix H. We will show that $\|H_0 - H_n\|_1 \to 0$ almost surely, which will imply that $\sum_{l=1}^n (\hat{\lambda}_l' - \hat{\lambda}_l) Z_l^2$ converges to 0 in probability.

The proof relies on following lemma:

Lemma 4. *Suppose $\mathcal{H}$ is a Hilbert space and $x_i, y_i \in \mathcal{H}$, $1 \le i \le n$ and $\Sigma := (\langle x_i, y_j \rangle)_{ij}$. Then $\|\Sigma\|_1 \le \|\vec{x}\| \cdot \|\vec{y}\|$. Here $\|\vec{x}\| := \sqrt{\sum_{i=1}^n \|x_i\|^2}$ and $\|\vec{y}\|$ likewise. Consequently,*

$$\|(\langle x_i, x_j \rangle)_{ij} - (\langle y_i, y_j \rangle)_{ij}\|_1 \le \|\vec{x} - \vec{y}\| \cdot (\|\vec{x}\| + \|\vec{y}\|)$$

Proof. Let $\mathcal{H}_0$:=span$\{x_1, \ldots, x_n, y_1, \ldots, y_n\}$ and $m := \dim \mathcal{H}_0$. Take an orthogonal basis $\{e\}_{i=1}^m$ of $\mathcal{H}_0$ and let $x_i = \sum_{j=1}^m a_i^j e_j$ and $y_i = \sum_{j=1}^m b_i^j e_j$ for $1 \le i \le n$, and let $A := (a_i^j)_{ij}$, $B := (b_i^j)_{ij}$. Then $\Sigma = AB^\top$. Let $\Sigma = PSQ^\top$ be the singular value decomposition of Σ, then apply the Cauchy-Schwartz inequality

$$\begin{aligned} \|\Sigma\|_1 = \mathrm{tr}(S) &= \mathrm{tr}(P^\top AB^\top Q) \\ &\le \|P^\top A\|_F \|B^\top Q\|_F = \|A\|_F \|B\|_F. \end{aligned}$$

The first part of the lemma follows from the fact that $\|A\|_F = \|x\|^2$ and $\|B\|_F = \|y\|^2$. For the second part, just note that

$$\begin{aligned} &\|(\langle x_i, x_j \rangle)_{ij} - (\langle y_i, y_j \rangle)_{ij}\|_1 \\ &\le \|(\langle x_i, x_j - y_j \rangle)_{ij}\|_1 + \|(\langle x_i - y_i, y_j \rangle)_{ij}\|_1 \\ &\le \|\vec{x} - \vec{y}\|(\|\vec{x}\| + \|\vec{y}\|). \end{aligned}$$

□

Next we prove the theorem.

Proof. Recall the construction in §III, the vector $\vec{\mathcal{A}}_p k_x := (\mathcal{A}_p^1 k_x, \ldots, \mathcal{A}_p^d k_x)^\top$ is an element in $\mathcal{H}_k^d$, and

$$V_p(x_i, x_j) = \left\langle \frac{p(x_i)}{\omega(x_i)} \vec{\mathcal{A}}_p k_{x_i}, \frac{p(x_j)}{\omega(x_j)} \vec{\mathcal{A}}_p k_{x_j} \right\rangle_{\mathcal{H}_k^d}.$$

For notational simplicity, let $\phi_\theta(x) := \frac{p(x)}{\omega(x)} \vec{\mathcal{A}}_{p_\theta} k_x \in \mathcal{H}_k^d$, then note that $H_0 = n^{-1}(\langle \phi_{\theta_0}(x_i), \phi_{\theta_0}(x_j) \rangle)_{ij}$ and $H_n = n^{-1}(\langle \phi_{\hat{\theta}_n}(x_i), \phi_{\hat{\theta}_n}(x_j) \rangle)_{ij}$. Apply the lemma 4 to get

$$\begin{aligned} n\|H_n - H_0\|_1 \le &\left(\sqrt{\sum_{i=1}^n \|\phi_{\hat{\theta}_n}(x_i)\|^2} + \sqrt{\sum_{i=1}^n \|\phi_{\theta_0}(x_i)\|^2} \right) \\ &\cdot \sqrt{\sum_{i=1}^n \|\phi_{\hat{\theta}_n}(x_i) - \phi_{\theta_0}(x_i)\|^2}. \end{aligned}$$

Let $f(\theta, x) := \|\phi_\theta(x) - \phi_{\theta_0}(x)\|^2$, then above inequality implies

$$\begin{aligned} \|H_n - H_0\|_1 \le &\sqrt{n^{-1} \sum_{i=1}^n f(\hat{\theta}_n, x_i)} \\ &\cdot \left(\sqrt{n^{-1} \sum_{i=1}^n f(\hat{\theta}_n, x_i)} + 2\sqrt{n^{-1} \sum_{i=1}^n \|\phi_{\theta_0}(x_i)\|^2} \right). \end{aligned}$$

Clearly, by law of large numbers,

$$\begin{aligned} n^{-1} \sum_{i=1}^n \|\phi_{\theta_0}(x_i)\|^2 &= n^{-1} \sum_{i=1}^n V_{\theta_0}(x_i, x_i) \\ &\xrightarrow{\omega-\text{a.s}} \int_G V_{\theta_0}(x, x) p_{\theta_0}(x) \mu(dx) < +\infty, \end{aligned}$$

thus $n^{-1} \sum_{i=1}^n \|\phi_{\theta_0}(x_i)\|^2 = O_p(1)$.

To show $\|H_n - H_1\| \to 0$ ω-a.s. (almost surely w.r.t. the distribution ω), it suffices to show $n^{-1}\sum_{i=1}^{n} f(\hat{\theta}_n, x_i) \to 0$ ω-a.s.. Note that

$$f(\theta, x) \leq 2\|\phi_\theta(x)\|^2 + 2\|\phi_{\theta_0}(x)\|^2 = 2V_\theta(x,x) + 2V_{\theta_0}(x,x),$$

thus $\sup_{\theta \in K} f(\theta, x)$ is ω-integrable for any compact set $K \subset \Theta$, as the conditions in theorem 4.7 state. Therefore, we can apply lemma 1 to conclude that

$$n^{-1}\sum_{i=1}^{n} f(\theta, x_i) \to F(\theta) := \int_G f(\theta, x)\omega(dx),$$

compactly ω-a.s. and $F(\theta)$ is continuous. Note that $f(\theta_0, x) = \|\phi_{\theta_0}(x) - \phi_{\theta_0}(x)\|^2 = 0$, thus $F(\theta_0) = 0$. Note that $\hat{\theta}_n \to \theta_0$ ω-a.s., thus $n^{-1}\sum_{i=1}^{n} f(\hat{\theta}_n, x_i) \to 0$ ω-a.s.. □

References

[1] H. W. Sorenson, Ed., *Kalman Filtering: Theory and Application*. IEEE Press, 1985.

[2] T. D. Downs, "Orientation statistics," *Biometrika*, vol. 59, no. 3, pp. 665–676, 1972.

[3] P. D. Hoff, "Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 438–456, 2009.

[4] C. Khatri and K. V. Mardia, "The von mises–fisher matrix distribution in orientation statistics," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 95–106, 1977.

[5] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, and T. Birdal, "Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1627–1654, 2022.

[6] I. Gilitschenski, R. Sahoo, W. Schwarting, A. Amini, S. Karaman, and D. Rus, "Deep orientation uncertainty learning based on a bingham loss," in *International conference on learning representations*, 2020.

[7] V. Peretroukhin, M. Giamou, D. M. Rosen, W. N. Greene, N. Roy, and J. Kelly, "A smooth representation of belief over so (3) for deep rotation learning with uncertainty," *arXiv preprint arXiv:2006.01031*, 2020.

[8] S. Prokudin, P. Gehler, and S. Nowozin, "Deep directional statistics: Pose estimation with uncertainty quantification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 534–551.

[9] T. Lee, "Bayesian attitude estimation with the matrix fisher distribution on so (3)," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3377–3392, 2018.

[10] W. Wang and T. Lee, "Matrix fisher–gaussian distribution on SO(3) × $\mathbb{R}^n$ and bayesian attitude estimation," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2175–2191, 2021.

[11] X. Pennec, "Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, pp. 127–154, 2006.

[12] M. Zhang and T. Fletcher, "Probabilistic principal geodesic analysis," *Advances in neural information processing systems*, vol. 26, 2013.

[13] Y. Zhang, J. Xing, and M. Zhang, "Mixture probabilistic principal geodesic analysis," in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy: 4th International Workshop, MBIA 2019, and 7th International Workshop, MFCA 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 4*. Springer, 2019, pp. 196–208.

[14] C. Ley and T. Verdebout, *Applied directional statistics: modern methods and case studies*. CRC Press, 2018.

[15] K. Oualkacha and L.-P. Rivest, "On the estimation of an average rigid body motion," *Biometrika*, vol. 99, no. 3, pp. 585–598, 2012.

[16] C. J. Oates, M. Girolami, and N. Chopin, "Control functionals for monte carlo integration," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 695–718, 2017.

[17] K. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *International conference on machine learning*. PMLR, 2016, pp. 2606–2615.

[18] C. Ley, G. Reinert, and Y. Swan, "Stein's method for comparison of univariate distributions," 2017.

[19] G. Mijoule, G. Reinert, and Y. Swan, "Stein operators, kernels and discrepancies for multivariate continuous distributions," *arXiv preprint arXiv:1806.03478*, 2018.

[20] G. Mijoule, M. Raič, G. Reinert, and Y. Swan, "Stein's density method for multivariate continuous distributions," *Electronic Journal of Probability*, vol. 28, pp. 1–40, 2023.

[21] J. Gorham and L. Mackey, "Measuring sample quality with stein's method," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[22] G. Jackson and M. Lester, "Measuring sample quality with kernels," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1292–1301.

[23] A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey, "Minimum stein discrepancy estimators," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] C. J. Oates, "Minimum kernel discrepancy estimators," in *Monte Carlo and Quasi-Monte Carlo Methods*, A. Hinrichs, P. Kritzer, and F. Pillichshammer, Eds. Springer Verlag, 2022.

[25] Q. Liu, J. Lee, and M. Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *International conference on machine learning*. PMLR, 2016, pp. 276–284.

[26] T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates, "Robust generalised bayesian inference for intractable likelihoods," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 3, pp. 997–1022, 2022.

[27] A. Barp, C. Oates, E. Porcu, M. Girolami *et al.*, "A riemann-stein kernel method," *arXiv preprint arXiv:1810.04946*, 2018.

[28] W. Xu and T. Matsuda, "Interpretable stein goodness-of-fit tests on riemannian manifold," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 502–11 513.

[29] S. Helgason, *Differential geometry, Lie groups, and symmetric spaces*. Academic press, 1979.

[30] J. M. Lee, *Introduction to Smooth Manifolds*. Springer, 2013.

[31] L. C. Evans and R. F. Garzepy, *Measure Theory and Fine Properties of Functions*. Routledge, 2018.

[32] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá, "Vector valued reproducing kernel hilbert spaces and universality," *Analysis and Applications*, vol. 8, no. 01, pp. 19–61, 2010.

[33] E. Fuselier and G. B. Wright, "Scattered data interpolation on embedded submanifolds with restricted positive definite kernels: Sobolev error estimates," *SIAM Journal on Numerical Analysis*, vol. 50, no. 3, pp. 1753–1776, 2012.

[34] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Science & Business Media, 2008.

[35] S. Suvorova, S. D. Howard, and B. Moran, "Tracking rotations using maximum entropy distributions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 2953–2968, 2021.

[36] R. J. Serfling, *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

[37] H. Rubin, "Uniform convergence of random functions with applications to statistics," *The Annals of Mathematical Statistics*, pp. 200–203, 1956.

[38] I.-K. Yeo and R. A. Johnson, "A uniform strong law of large numbers for u-statistics with application to transforming to near symmetry," *Statistics & probability letters*, vol. 51, no. 1, pp. 63–69, 2001.

[39] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, "A fast, consistent kernel two-sample test," *Advances in neural information processing systems*, vol. 22, 2009.

[40] M. Zhang and P. T. Fletcher, "Bayesian principal geodesic analysis in diffeomorphic image registration," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III 17*. Springer, 2014, pp. 121–128.

[41] C. F. Van Loan, "The ubiquitous kronecker product," *Journal of computational and applied mathematics*, vol. 123, no. 1-2, pp. 85–100, 2000.

[42] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective hilbert space embeddings of probability measures," 2008, pp. 111–122, 21st Annual Conference on Learning Theory, COLT 2008 ; Conference date: 09-07-2008 Through 12-07-2008.

[43] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet, "Universality, characteristic kernels and rkhs embedding of measures," *Journal of Machine Learning Research*, vol. 12, no. 70, pp. 2389–2410, 2011. [Online]. Available: http://jmlr.org/papers/v12/sriperumbudur11a.html

[44] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[45] C. A. León, J.-C. Massé, and L.-P. Rivest, “A statistical model for random rotations,” *Journal of Multivariate Analysis*, vol. 97, no. 2, pp. 412–430, 2006.

[46] K. V. Mardia, P. E. Jupp, and K. Mardia, *Directional statistics*. Wiley Online Library, 2000, vol. 2.

[47] D. W. Eggert, L. Adele, and R. B. Fisher, “Estimating 3-d rigid body transformations: a comparison of four major algorithms,” *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 272–290, 1997.

[48] B. Stanfill, H. Hofmann, and U. Genschel, “rotations: An r package for so(3) data,” *The R Journal*, vol. 6, pp. 68–78, 2014. [Online]. Available: https://journal.r-project.org/archive/2014-1/stanfill-hofmann-genschel.pdf

[49] M. P. Gaffney, “A special stokes’s theorem for complete riemannian manifolds,” *Annals of Mathematics*, pp. 140–145, 1954.

[50] J. R. Munkres, “Topology,” *(No Title)*, 2000.

[51] T. Tao, *Topics in random matrix theory*. American Mathematical Society, 2023, vol. 132.

Xiaoda Qu is a PhD student in the Department of Statistics, University of Florida. He received his BS in Mathematics and Applied Mathematics from Zhejiang University. His research interests are in geometric statistics and machine learning. His currently a PhD candidate in University of Florida.

Xiran Fan received her BS in Statistics from Nankai University and her PhD in Statistics from the University of Florida. Her research interests are in geometric machine Learning and machine learning. She is currently a research scientist at Visa, California, USA.

Baba C. Vemuri received his PhD in Electrical and Computer Engineering from the University of Texas at Austin. Currently, he is a Distinguished University Professor and holds the Wilson and Marie Collins professorship in Engineering at the University of Florida. He is a professor in the Department of Computer and Information Sciences and Engineering, and holds affiliate appointments in the Department of Statistics and Mathematics at the University of Florida. His research interests lie in Geometric Statistics, Computer Vision, Machine Learning and Medical Imaging. He received the IEEE Computer Society’s Edward McCluskey Technical Achievement Award (2017) and is a Fellow of the IEEE (2001) and the ACM (2009).