A Riemann–Stein Kernel Method

ALESSANDRO BARPI\textsuperscript{1,2}, CHRIS. J. OATES\textsuperscript{3,2} EMILIO PORCU\textsuperscript{4,5} and MARK GIROLAMI\textsuperscript{1,2}

\textsuperscript{1}University of Cambridge, UK, E-mail: ab2286@eng.cam.ac.uk; mag92@eng.cam.ac.uk
\textsuperscript{2}Alan Turing Institute, UK
\textsuperscript{3}Newcastle University, UK, E-mail: chris.oates@ncl.ac.uk
\textsuperscript{4}Khalifa University, Abu Dhabi, E-mail: emilio.porcu@ku.ac.ae
\textsuperscript{5}Trinity College Dublin, IE

This paper proposes and studies a numerical method for approximation of posterior expectations based on interpolation with a Stein reproducing kernel. Finite-sample-size bounds on the approximation error are established for posterior distributions supported on a compact Riemannian manifold, and we relate these to a kernel Stein discrepancy (KSD). Moreover, we prove in our setting that the KSD is equivalent to Sobolev discrepancy and, in doing so, we completely characterise the convergence-determining properties of KSD. Our contribution is rooted in a novel combination of Stein’s method, the theory of reproducing kernels, and existence and regularity results for partial differential equations on a Riemannian manifold.

Keywords: Bayesian computation; kernel Stein discrepancy; reproducing kernel; Stein equation

1. Introduction

The focus of this paper is the numerical approximation of an integral

\[ \int_M f \, dP \]  

where \( P \) is a probability measure on a compact, connected, complete Riemannian manifold \( M \) and \( f \in C(M, \mathbb{R}) \). The notation \( C(A, B) \) indicates the set of continuous functions from a topological space \( A \) to a topological space \( B \). The distribution \( P \) is assumed to admit a density \( p \) with respect to the natural volume measure on \( M \), specified only up to an unknown normalisation constant. It is assumed that direct computation of the normalisation constant is difficult and therefore precluded. This final point demands special consideration and prevents standard numerical integration method from being used. This situation is of course regularly encountered in Bayesian statistics, where \( P \) is a posterior distribution whose density \( p \) is specified in un-normalised form as the product of a prior and a likelihood. In general the direct computation of the normalisation constant is difficult in the Bayesian context [33].

Several approaches to approximation of (1) have been developed. These range from heuristic approaches, such as variational inference [46] and the Laplace approximation [82], through to asymptotically exact approaches such as Markov chain Monte Carlo [MCMC; 34]. Among asymptotically exact approaches, MCMC is most widely-used and its convergence theory is well-developed [64]. However, the absolute error of the ergodic average is gated at \( O_p(n^{-1/2}) \), where \( n \) is the number of evaluations of the integrand. This rate is sub-optimal for an \( s \)-times weakly differentiable integrand when \( s > \frac{d}{2} \) and \( d \) is the dimension of the manifold; a consequence of the fact that the ergodic average does not exploit smoothness properties of the integrand [101]. In recent years, several alternatives to MCMC have been developed to address this convergence bottleneck, focussing on smooth integrals of low effective dimension for which sub-optimality of MCMC is most pronounced. These include transport maps [63], Riemann sums [77], quasi Monte Carlo ratio estimators [85, 24, 25], minimum energy designs [45],
support points [59] and estimators based on Stein’s method [58, 76, 57, 75, 18, 17, 43, 80, 100, 29]. The computational cost of some of these methods is higher than \( O(n) \) and, for the method that we study in this work, the cost is \( O(n^2) \). Thus any accelerated convergence being offered must be weighed against this increased computational overhead. Indeed, for a method with computational cost \( c = O(n^r) \) and error \( O_p(c^{-e/r}) \), MCMC has error \( O_p(c^{-1/2}) \) (\( e = \frac{1}{2}, r = 1 \)), while the method that we study has error \( O_p(c^{-s/3d}) \) (\( e = \frac{2}{3}, r = 3 \)), where \( s \) captures the regularity of the integrand; accelerated convergence therefore requires that \( s > \frac{3d}{2} \).

\[ \text{Problem 1.} \quad \text{Fix } P \in \mathcal{P}(M). \text{ Consider a set } \mathcal{H} \subset C(M, \mathbb{R}) \text{ and an operator } A_P : \mathcal{H} \to C(M, \mathbb{R}) \text{ with the property that, for all } Q \in \mathcal{P}(M), \]

\[ P = Q \iff \int_M A_P h \, dQ = 0 \quad \forall h \in \mathcal{H}. \]

Then \( (\mathcal{H}, A_P) \) is said to be a Stein characterisation of \( P \). In this case the set \( \mathcal{H} \) is called a Stein class and the operator \( A_P \) is called a Stein operator. If only the \( \Rightarrow \) implication holds, so that \( \mathcal{H} \) need not be rich enough to distinguish elements in \( \mathcal{P}(M) \), then we call \( (\mathcal{H}, A_P) \) a Stein pair for \( P \).

The reader is referred to [51] for probabilistic background on Stein’s method and to [1] for a recent overview of statistical applications of Stein’s method. For the moment, let us suppose that a Stein pair \( (\mathcal{H}, A_P) \) for \( P \) can be found. Then (1) can be directly approximated, as will now be explained: First, select a set of distinct locations \( X = \{x_i\}_{i=1}^n \subset M \) at which the integrand is to be evaluated. Then construct an estimator of the form

\[ P_X(f) := \inf_{\xi \in \mathbb{R}} \inf_{h \in \mathcal{H}} \sum_{i=1}^n \left( \xi + A_P h(x_i) - f(x_i) \right)^2 + R_1(\xi) + R_2(h) \quad (2) \]

where \( R_1 \) and \( R_2 \) are regularisation terms to be specified, whose purpose is to ensure that a (unique) minimum will exist. The form of (2) can be motivated as constructing an approximation \( f_X \) to the integrand \( f \), based on evaluation of \( f \) at the locations in \( X \), in the class of functions of the form \( f_X := \xi + A_P h \), where \( \xi \in \mathbb{R} \) and \( h \in \mathcal{H} \). The definition of a Stein pair ensures that \( \int_M f_X \, dP = \xi \), so (2) can be interpreted as the integral of an approximation \( f_X \) to the integrand. Numerical approximations of the form (2) are increasingly being adopted in applied and methodological work [55, 93, 15, 81, 47, 65]. The properties of (2) depend on the class \( \mathcal{H} \), the operator \( A_P \), the set of points \( X \) where \( f \) is evaluated, and the regularisation terms \( R_1, R_2 \). Thus the above formulation is quite general and some specific choices are discussed next.

**Previous Work for \( M = \mathbb{R}^d \):** All previous work on (2) has focussed on the Euclidean context with \( M = \mathbb{R}^d \). In particular, [4, 67] considered the case where \( \mathcal{H} \subset C(\mathbb{R}^d, \mathbb{R}) \) is a space of low-degree polynomials under no regularisation, i.e. \( R_1, R_2 \equiv 0 \). This was combined with the operator \( A_P : \mathcal{H} \to \)}
$C(\mathbb{R}^d, \mathbb{R})$, $\mathcal{A}_p h = \nabla \cdot (p \nabla h)/p$, a second-order differential operator that can be evaluated without access to $p$’s normalisation constant. This led to an over-constrained least-squares problem and (2) can be seen as a classical control variate method. Higher-degree polynomials were considered in [92] in conjunction with an appropriate regularisation to ensure the least-squares problem remains well-posed.

The innovation in [76] was to consider instead an infinite-dimensional normed space for $\mathcal{H}$. The operator in [76] was $\mathcal{A}_p : \mathcal{H} \rightarrow C(\mathbb{R}^d, \mathbb{R})$, where $\mathcal{H} \subset C(\mathbb{R}^d, \mathbb{R}^d)$ was a Cartesian product of reproducing kernel Hilbert spaces and $\mathcal{A}_p h = \nabla \cdot (ph)/p$ was a first-order differential operator that can again be evaluated without $p$’s normalisation constant. To complete the specification of the method, the following natural regularisation terms were proposed:

$$R_1(\xi) = \sigma^{-2} \xi^2$$

$$R_2(h) = \inf \left\{ \|h'\|_{\mathcal{H}}^2 : h' \in \mathcal{H}, \mathcal{A}_p(h' - h) = 0 \right\}$$

(3)

where $\sigma > 0$ was a parameter to be specified and $\| \cdot \|_{\mathcal{H}}$ denotes the norm associated to $\mathcal{H}$. That the term $R_2(h)$ should depend on $h$ through $\mathcal{A}_p h$ is natural, since $h$ enters into the approximation $f_X$ only through $\mathcal{A}_p h$. The solution to (2) can be computed in closed-form when reproducing kernels are employed and, since in this case $f_X$ interpolates the data $X$, we call $f_X$ a Stein kernel interpolant. In recent work [91] proposed a semi-parametric approach that combined elements from [76] and [4, 67].

In parallel work in the Euclidean context, [57] considered adding additional regularisation to (2), whilst alternatives to the squared error objective have also been considered, including an empirical variance estimator [9] and an estimator of asymptotic variance [8] for use in the MCMC context. The latter were shown, empirically, to improve estimator performance but at the cost of no longer having a closed-form expression for the estimator. In a different direction, [106] proposed to take $\mathcal{H}$ to be a finite-dimensional parametric neural network and empirically explored its potential, while [87] provided a theoretical analysis of stochastic gradient descent in that context.

**Related Work for General $M$:** Beyond the Euclidean case, in parallel to the present research, a Stein pair for distributions supported on a Riemannian manifold was provided in [54, 104] and this was extended to a Stein characterisation in [49]. The focus of [54] was the design of a gradient flow on a manifold, the focus of [104] was on goodness-of-fit testing and the focus of [49] was weak convergence control. Our analysis starts from the same Stein pair, but diverges thereafter, with our focus being on the numerical approximation of (1).

**1.2. Our Contributions**

Our contributions are as follows:

- **Generalisation to a Riemannian Manifold:** Integrals on manifolds arise in many important applications of Bayesian statistics, most notably directional statistics [60] and modelling of functional data on the sphere $\mathbb{S}^2$ [78]. In this context, MCMC methods have been developed to sample from distributions defined on a manifold [e.g. 23, 7, 16, 48, 44, 3]. In this paper we complement existing work by generalising the Stein kernel interpolant of [76] to integrals defined on a Riemannian manifold.

- **New Insight into Discrepancy:** The kernel Stein discrepancy [KSD; 19, 56, 37] is an upper bound on the worst-case error of the Stein kernel interpolant. For our setting, we prove that this discrepancy is equivalent to a classical Sobolev discrepancy on the manifold. An explicit statement
is presented in Theorem 1. Thus we completely characterise the convergence-determining properties of KSD in this context. This is our main result and our proof combines Stein’s method, the theory of reproducing kernels and existence and regularity results for partial differential equations on a Riemannian manifold.

• Finite-Sample-Size Error Bound: Bounds on the worst-case approximation error are established as consequences of the Sobolev connection in Theorem 1. The optimal convergence rate is established, under appropriate regularity assumptions on the distribution \( P \) and the point set \( X \). An explicit statement of the rate is presented in Theorem 2.

The paper now proceeds to Section 2, where we provide a brief mathematical background. In Section 3 we state our main results, with proofs contained in Section 4. A discussion is provided in Section 5.

2. Background

The purpose of this section is to introduce the mathematical tools needed for our development. The experienced reader may prefer to continue directly to Section 3 and return later to Section 2 if required. In Section 2.1 we recall the definition of a Riemannian manifold, in Section 2.2 we review geometric measure theory, in Section 2.3 we describe calculus on a Riemannian manifold, in Section 2.4 we define reproducing kernel Hilbert spaces and in Section 2.5 we construct a Sobolev norm on a Riemannian manifold.

2.1. Riemannian Manifolds

Let \( C^l(A, B) \) denote the set of all \( l \)-times continuously differentiable functions from \( A \) to \( B \), where \( A \) and \( B \) possess suitable structure for the notion of a continuous derivative to be well-defined.

A \( d \)-dimensional manifold \( M \), \( d \in \mathbb{N} \), is a Hausdorff topological space for which every point \( x \in M \) has an open neighbourhood \( U_x \) homeomorphic to an open subset of \( \mathbb{R}^d \) or an open subset of the closed \( d \)-dimensional upper half-space if \( M \) has a boundary \( \partial M \) [see p25 of 50]. If \( \phi: U \to \mathbb{R}^d \) is a homeomorphism (onto its image) with \( x \in U \subset M \), we say \( (U, \phi) \) is a coordinate patch around \( x \). This defines coordinate functions \( q_j := \pi_j \circ \phi \) over \( U \), where \( \pi_j : \mathbb{R}^d \to \mathbb{R} \) are the canonical projections and \( \circ \) denotes composition of functions. A \( C^\infty \) atlas is a collection of coordinate patches \( (U_i, \phi_i) \) that cover \( M \) such that the transition functions \( \phi_j \circ \phi_i^{-1} \) are \( C^\infty(\mathbb{R}^d, \mathbb{R}^d) \) whenever they are defined. A manifold is smooth if it possesses a \( C^\infty \) atlas. The tangent space \( T_x M \) at \( x \) is the vector space of \( C^\infty \) functions over \( C^\infty(M) \) satisfying Leibniz rule; i.e. the elements of \( T_x M \) are linear functionals \( t : C^\infty(M) \to \mathbb{R} \) such that \( t(ab) = t(a)b(x) + a(x)t(b) \) for all \( a, b \in C^\infty(M) \). If \( q_j \) are coordinates on a patch \( (U, \phi) \) containing \( x \), the coordinate vectors \( \partial q_j | x : f \mapsto \partial(f \circ \phi_j^{-1})/\partial q_j | \phi(x) \) define a basis of \( T_x M \). We say \( (M, g) \) is a Riemannian manifold if \( M \) has a metric tensor \( g \), i.e., a smooth map \( x \mapsto g_x \) such that \( g_x \) is an inner product on \( T_x M \). It will be convenient to represent the metric tensor \( g_x \) as a matrix \( G(x) \) with coordinates \( G_{ij}(x) := g_x \left( \partial q_i | x, \partial q_j | x \right) \).

The non-Euclidean manifolds of interest in most applications are closed; for example the group of rotations, Grassmannian manifolds, and Stiefel manifolds such as hyperspheres, though we will also allow manifolds with boundary for completeness.

**Standing Assumption 1.** \( M \) is a smooth, compact, complete, and connected manifold, that is either closed or is a manifold with boundary.
In the case of a manifold with boundary, the outward-pointing unit normal \( n \) to the boundary \( \partial M \) of the manifold can be defined via the fact that, since \((\partial M, g|_{\partial M})\) is a Riemannian submanifold of \((M, g)\), then for each \( x \in \partial M \), the metric \( g_x \) of \( M \) splits the tangent space \( T_x M \) into \( T_x \partial M \) and its orthogonal complement \( N_x \); i.e. \( T_x M = T_x \partial M \oplus N_x \). Here \( N_x \) is the one-dimensional space spanned by the normal vector to \( \partial M \). See e.g. [5].

The geodesic distance \( d_M(x, y) \) on a complete Riemannian manifold is defined as the infimum of the length \( \int_0^1 \sqrt{g(\dot{\gamma}(t), \dot{\gamma}(t))} dt \) over all continuously differentiable curves \( \gamma : [0, 1] \to M \) with \( \gamma(0) = x \) and \( \gamma(1) = y \), where \( \dot{\gamma} \) is the tangent vector to \( \gamma \).

**Example 1.** The sphere \( S^2 \) is a 2-dimensional Riemannian manifold. The coordinate patch \( \phi \) with \( \phi^{-1}(q) = (\cos q_1 \sin q_2, \sin q_1 \sin q_2, \cos q_2) \), with local coordinates \( q_1 \in (0, 2\pi) \), \( q_2 \in (0, \pi) \), covers almost all \( x \in S^2 \) (it does not cover the half great circle that passes through both poles and the point \((1, 0, 0))\). The tangent space \( T_{\phi^{-1}(q_1, q_2)} M \) is spanned by \( \partial_{q_1} = (-\sin q_1 \sin q_2, \cos q_1 \sin q_2, 0) \) and \( \partial_{q_2} = (\cos q_1 \cos q_2, \sin q_1 \cos q_2, -\sin q_2) \). Let \( s = s_1 \partial_{q_1} + s_2 \partial_{q_2} \in T_x M \) be associated with the coefficient vector \( s^+ = [s_1, s_2] \) and similarly for \( t^+ = [t_1, t_2] \). Taking the Euclidean inner product of these vectors shows that \( g_x(s, t) := \langle s, t \rangle_G = s^+ G t \) where \( G_{1,1} = \sin^2 q_2 \), \( G_{2,2} = 1 \), \( G_{1,2} = G_{2,1} = 0 \).

### 2.2. Geometric Measure Theory

Any Riemannian manifold has a natural measure \( V \) over its Borel algebra, called the *Riemannian volume measure*, with infinitesimal volume element denoted \( dV \). In a coordinate patch \( U_i \subset \mathbb{R}^d \), this measure can be expressed in terms of the Lebesgue measure: \( dV = \sqrt{\det(G(x))} \lambda^d(\mathrm{d}q) \) where \( G(f) = \int_M f dV \). In particular, when \( M \) is the Euclidean space this is just the Lebesgue measure, and when \( M \) is an embedded manifold in \( \mathbb{R}^m \), \( V \) is the Hausdorff measure [28]. When \( M \) has a non-empty boundary \( \partial M \), the latter inherits a measure by noting that \( \partial M \) is a submanifold of \( M \) and the restriction \( g|_{\partial M} \) of the metric \( g \) turns it into a Riemannian manifold \((\partial M, g|_{\partial M})\). Its associated Riemannian volume measure is usually denoted \( i_n dV \) in the literature [30].

**Example 2.** For the sphere \( M = S^2 \), \( dV = \sin q_2 dq_1 dq_2 \), where \( \sin q_2 \) is the area of the parallelogram spanned by \( \partial_{q_1}, \partial_{q_2} \).

### 2.3. Calculus on a Riemannian Manifold

To present a geometric, coordinate-independent construction of differential operators on manifolds would require either exterior calculus or the concept of a covariant derivative. To limit scope, we present two important differential operators in local coordinates and merely comment that the associated operators are in fact coordinate-independent; full details can be found in [5]. To this end, denote the gradient of a function \( f : M \to \mathbb{R} \) as

\[
\nabla f = \sum_{i,j=1}^{d} [G^{-1}]_{ij} \frac{\partial f \circ \phi^{-1}}{\partial q_j} \partial_{q_i},
\]

whenever it is assumed to exist. Likewise, define the divergence of a vector field \( s = s_1 \partial_{q_1} + \cdots + s_d \partial_{q_d} \) with \( s_i = s_i(x) \), assumed to exist, as

\[
\nabla \cdot s = \sum_{i=1}^{d} \frac{\partial}{\partial q_i} s_i + s_i \frac{\partial}{\partial q_i} \log \sqrt{\det(G)}.
\]
These two differential operators are sufficient for our work; for instance, they can be combined to obtain the Laplace–Beltrami operator $\Delta f := \nabla \cdot \nabla f$.

## 2.4. Reproducing Kernel Hilbert Spaces

A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions on $M$ is a Hilbert space for which the evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}, \delta_x(f) := f(x)$, are continuous for each $x \in M$. The dual space of bounded linear operators $L : \mathcal{H} \to \mathbb{R}$ is denoted $\mathcal{H}^*$. The Riesz-representation theorem implies that $\mathcal{H}$ is isomorphic to $\mathcal{H}^*$ and we can thus associate a vector $k_x \in \mathcal{H}$ to $\delta_x$ which satisfies $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$.

The symmetric function $k(x, y) := k_x(y)$ is called the reproducing kernel for $\mathcal{H}$ and we denote this as $\mathcal{H}(k)$ in the sequel. It can be checked $k$ is a semi-positive definite function on $M \times M$. Moore’s theorem states the converse is also true; any semi-positive function on $M$ defines an RKHS $\mathcal{H}(k)$ of functions on $M$ with $k$ as its reproducing kernel. See e.g. [10].

### Example 3.

Consider a finite collection of linearly independent real-valued functions $\{e_i\}_{i=1}^m$ on $M$. Let $\mathcal{H}$ denote the real vector space of functions $h : M \to \mathbb{R}$ of the form $h = \sum_{i=1}^m h_i e_i$, where $h_1, \ldots, h_m \in \mathbb{R}$. Equip $\mathcal{H}$ with an inner product $\langle h, g \rangle_{\mathcal{H}} = \sum_{i=1}^m h_i g_i$, where here $g = \sum_{i=1}^m g_i e_i$. Then $\mathcal{H}$ is a (finite-dimensional) RKHS with reproducing kernel $k(x, y) = \sum_{i=1}^m e_i(x) e_i(y)$.

## 2.5. Sobolev Norm on a Riemannian Manifold

Let $S \subset \mathbb{R}^d$ be an open set and recall that the standard Sobolev space $W^s_2(S)$ is defined as the set of equivalence classes $f \in L^2(S)$ such that the weak derivatives $D^\alpha f := \partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d} f \in L^2(S)$ for all $|\alpha| := \alpha_1 + \cdots + \alpha_d \leq s$. The set $W^s_2(S)$, for $s > d/2$, becomes a RKHS when equipped with the norm\(^1\)

$$
\|f\|_{W^s_2(S)} := \left( \sum_{|\alpha| \leq s} \|D^\alpha f\|^2_{L^2(S)} \right)^{1/2}.
$$

In order to define a Sobolev space on a Riemannian manifold $(M, g)$, let $(U_i, \phi_i)$ be a $C^\infty$ atlas such that the $(U_i)$ form an open cover of $M$ and $(\rho_i)$ a partition of unity subordinate to $(U_i)$. Then let $W^s_2(M)$ be the set of real-valued functions on $M$ for which the norm

$$
\|f\|_{W^s_2(M)} := \left( \sum_i \| (\rho_i f) \circ \phi_i^{-1} \|^2_{L^2(\mathbb{R}^d)} \right)^{1/2}
$$

is finite. It can be shown that $W^s_2(M)$ is a RKHS [31]. Note that the norm depends on the choice of atlas and partition of unity. Different choices lead to different norms, however these are all equivalent [31]. To avoid confusion, we fix a specific atlas and partition of unity in the sequel. Further discussion can be found in [22]. The restriction to compact manifolds is fundamental to our analysis, as Sobolev norms on general manifolds are not equivalent, see section 3.2 [40]. This completes our mathematical background.

\(^1\)In a slight abuse of notation, for $f \in C(S, \mathbb{R})$ we interpret $\|f\|_{W^s_2(S)}$ as $\|\iota \circ f\|_{W^s_2(S)}$ where $\iota \circ f$ maps $f$ to its equivalence class in $W^s_2(S)$.\n
3. Results

In this section our novel theoretical results are stated. First in Section 3.1, we present and discuss the technical conditions that will be assumed. In Section 3.2 we explain how the choice of regularisers $R_1$, $R_2$ in (3) allows the estimator $P_X(f)$ in (2) to be computed. In Section 3.3 we explain how the error of the Stein kernel interpolant can be quantified. Then in Section 3.4 we establish an equivalence between the Stein kernel interpolant and approximation using Sobolev spaces on a Riemannian manifold. Finally, in Section 3.5, explicit sufficient conditions for the convergence of the estimator $P_X(f)$ are established.

3.1. Stein Pair on a Riemannian Manifold

The first task in this section is to present and discuss the technical conditions that will be assumed on $H$ and $A_P$ to ensure that $(H, A_P)$ is a Stein pair for $P$. Two normed spaces $(X, \| \cdot \|_X), (Y, \| \cdot \|_Y)$ are said to be equivalent if the sets $X$ and $Y$ are equal and if there exists $0 < C < \infty$ such that $C^{-1}\|x\|_X \leq \|x\|_Y \leq C\|x\|_X$ for all $x \in X$.

Standing Assumption 2. Let $k : M \times M \to \mathbb{R}$ be positive definite, with $H(k)$ equivalent to $W^{s+2}_2(M)$ for some $s > d/2$.

Kernels that reproduce Sobolev spaces are available in explicit forms for several manifolds of common interest; see e.g. [26] for the case of the sphere and the Grassmannian manifold. For general manifolds, however, numerical methods may be required to approximate a Sobolev kernel using an eigenbasis of the Laplace–Beltrami operator; see e.g. [22].

In this work a density $p$ for the distribution $P$ with respect to the Riemannian volume measure $V$ is required. It will be assumed that $p$ is positive on $M$ and that a number of derivatives of $p$ exist:

Standing Assumption 3. The function $\log p : M \to \mathbb{R}$ is $C^{s+1}(M, \mathbb{R})$, for the same exponent $s$ introduced in Assumption 2.

The Stein pair that we consider is based on the Riemannian manifold generalisation of the original differential operator considered in [4]; see also Section 3.4.3 of [66]:

Standing Assumption 4. The operator $A_P : H \to C^s(M, \mathbb{R})$ is the second order differential operator $A_P h = \nabla \cdot (p \nabla h)/p$.

Assumptions 2 and 3 ensure that $A_P$ in Assumption 4 is well-defined. Indeed, from the product rule

$$A_P h = \frac{p \nabla \cdot \nabla h + g(\nabla p, \nabla h)}{p} = \Delta h + g(\nabla \log p, \nabla h),$$

from Assumption 3 $\nabla \log p \in C^s(M, \mathbb{R})$, and from the Sobolev embedding theorem $\Delta h \in C^s(M, \mathbb{R})$ whenever $h \in W^{s+2}_2(M)$ and $s > d/2$ [42, Theorem 2.7], which is Assumption 2. In what follows, the operator in Assumption 4 will be called the Riemann–Stein operator due to its suitability for Stein’s method on a Riemannian manifold. The Riemann–Stein operator is an example of a weighted Laplacian operator; see [38]. This operator coincides with the operator studied in the concurrent work of [54, 104, 49].
Standing Assumption 5. If \((M, g)\) is a Riemannian manifold with boundary \(\partial M\), then
\[
\int_{\partial M} g(p \nabla h, n) \, i_n \, dV = 0 \quad \forall h \in \mathcal{H}.
\]

For a manifold with boundary, the boundary condition in Assumption 5 is either automatically satisfied if \(p\) vanishes on \(\partial M\) or must be enforced through a suitable restriction on \(\mathcal{H}\). It is unclear to us whether either of these sufficient conditions are compatible with Assumption 2 or Assumption 3. On the other hand, if \(M\) is a closed manifold, then no assumption is required.

Proposition 1. \((\mathcal{H}, A_P)\) is a Stein pair for \(P\).

Proof. From Assumptions 2 and 3 the operator \(A_P\) in Assumption 4 is well-defined. Thus for \(h \in \mathcal{H}\), using the above assumptions and the divergence theorem on a Riemannian manifold [see e.g. 99], we find
\[
\int_M A_P h \, dP = \int_M \nabla \cdot (p \nabla h) \, dV = \left\{ \begin{array}{ll}
\int_{\partial M} g(p \nabla h, n) \, i_n \, dV & \text{if } M \text{ is a manifold with boundary} \\
0 & \text{if } M \text{ is a closed manifold}. \end{array} \right.
\]

Finally, in the case of a manifold with boundary, Assumption 5 ensures the boundary integral in (4) is zero, as required. \(\square\)

3.2. Exact Computation of (2)

Our assumptions ensure that the estimator \(P_X(f)\) can be computed. The computations are analogous to those of traditional kriging [96], albeit based on a non-standard, non-radial kernel. It will be assumed that the point set \(X \subset M\) has already been generated. No specific requirements are needed on \(X\) in order for the estimator to be computed, however it will be assumed that its elements are distinct. First, define the function \(k_P : M \times M \to \mathbb{R}\) as \(k_P(x, x') := A_P^t A_P k(x, x')\), where the Riemann–Stein operators \(A_P\) and \(A_P^t\) act, respectively, on the first and second argument of the kernel. Then, for \(\sigma > 0\), define the function \(k_{P,\sigma} : M \times M \to \mathbb{R}\) as \(k_{P,\sigma}(x, x') = \sigma^2 + k_P(x, x')\). The representer theorem, together with basic properties of RKHS, implies that a function \(f_X = \xi + A_P h\) solving (2) is an element of \(\mathcal{H}(k_{P,\sigma})\), and identical calculations to those in [76] establish that
\[
P_X(f) = \sigma^2 1^\top \mathbf{K}_{P,\sigma}^{-1} f
\]
where \([\mathbf{K}_{P,\sigma}]_{ij} := k_{P,\sigma}(x_i, x_j)\) and \([f]_i := f(x_i)\). The requirement that elements of \(X\) are distinct, together with the fact that \(k_{P,\sigma}\) is equivalent to a known positive definite kernel (see Theorem 1), ensure that \(\mathbf{K}_{P,\sigma}\) is non-singular. The computation of (5) is associated with a \(O(n^3)\) cost, which can be contrasted with the \(O(n)\) cost of generating a point set \(X\) using MCMC. An increased cost (e.g. relative to MCMC) is typical for methods with accelerated convergence and can be justified when the convergence rate is sufficiently fast relative to the computational cost.

From (5), the operator \(P_X\) is seen to be linear. In particular, it is recognised as a weighted cubature rule \(P_X(f) = \sum_{i=1}^n w_i f(x_i)\) with weights \(w = [w_1, \ldots, w_n]^\top\) the solution to \(\mathbf{K}_{P,\sigma}^{-1} w = \sigma^2 1\). In practice, the parameter \(\sigma\) can either be set in a data-driven manner or eliminated altogether in such a way that constant functions are exactly integrated:
Proposition 2. Let $K_P$ denote the $n \times n$ matrix with entries $k_P(x_i, x_j)$. Then

$$
\lim_{\sigma \to \infty} P_X(f) = \left( \frac{K_P^{-1}1}{1^\top K_P^{-1}1} \right)^\top f. \quad (6)
$$

Proof. Note that $K_{P,\sigma} = \sigma^2 11^\top + K_P$. The proof is then an application of the Woodbury matrix inversion formula, which can be used to deduce that

$$
P_X(f) = \sigma^2 1^\top (\sigma^2 11^\top + K_P)^{-1}f = \frac{1^\top K_P^{-1}f}{\sigma^{-2} + 1^\top K_P^{-1}1}
$$

from which the result is immediately established. \qed

3.3. Connection to Kernel Stein Discrepancy

In addition to returning an estimate for the integral, the Riemann–Stein kernel method is accompanied by a parsimonious error assessment. Indeed, from the Moore–Aronszajn theorem, the kernel $k_{P,\sigma}$ induces an RKHS, denoted $H(k_{P,\sigma})$, and the worst case error of the cubature rule in (5) in the unit ball of $H(k_{P,\sigma})$ is given by

$$
\sup \left\{ |P_X(f) - \int_M f \, dP| : \|f\|_{H(k_{P,\sigma})} \leq 1 \right\} = (\sigma^{-2} + 1^\top K_P^{-1}1)^{-1/2}. \quad (7)
$$

The expression $(1^\top K_P^{-1}1)^{-1/2}$, which is recovered in the $\sigma \to \infty$ limit, can be interpreted as a kernel Stein discrepancy [KSD; 19, 56, 37]

$$
\text{KSD} \left( \sum_{i=1}^n w_i \delta(x_i), P \right) := \sqrt{\sum_{i,j=1}^n w_i w_j k_P(x_i, x_j)} \quad (8)
$$

for the specific choice of weights

$$
w = \frac{K_P^{-1}1}{1^\top K_P^{-1}1} \quad (9)
$$

that arise in Proposition 2. Indeed, minimisation of (8) over the weights $w$ subject to the non-degeneracy constraint $1^\top w = 1$ leads to (9), so that these weights are in a sense optimal. Under certain conditions on $P$ and $k$, KSD controls the standard notion of weak convergence of the weighted empirical measure $\sum_{i=1}^n w_i \delta(x_i)$ to the target $P$. Sufficient conditions for the case $M = \mathbb{R}^d$ and a first order differential operator were established in [37, 18].

3.4. First Main Result: Stein and Sobolev Kernels are Equivalent

Our analysis of the Stein kernel interpolant is rooted in a more fundamental result that establishes that the reproducing kernel Hilbert space associated to $k_{P,\sigma}$ is equivalent to a standard Sobolev space on $M$:
Theorem 1 (First Main Result). The Hilbert spaces $\mathcal{H}(k_{P,\sigma})$ and $W^s_2(M)$ are equivalent, meaning that they are equal as sets and that there exists a constant $0 < C < \infty$ such that $C^{-1} \|h\|_{\mathcal{H}(k_{P,\sigma})} \leq \|h\|_{W^s_2(M)} \leq C \|h\|_{\mathcal{H}(k_{P,\sigma})}$ for all $h \in W^s_2(M)$.

All proofs are provided in Section 4. Only one direction of this equivalence had been shown in earlier work; the embedding of $\mathcal{H}(k_{P,\sigma})$ into $W^s_2(M)$ was shown in the Euclidean setting in [75]. The main effort in our proof is to establish the converse direction; i.e. that the Stein equation $f = \xi + A_fh$ has a solution $h \in \mathcal{H}(k)$ and $\xi \in \mathcal{H}(\sigma^2)$ for each $f \in W^s_2(M)$. The argument makes use of the theory of partial differential equations on a Riemannian manifold.

Theorem 1 implies that the worst case error in (7) is equivalent to

$$\sup \left\{ \left| P_X(f) - \int_M f dP \right| : \|f\|_{W^s_2(M)} \leq 1 \right\},$$

(10)

the worst case error over the unit ball of $W^s_2(M)$. In particular, (7) converges to 0 only when $\sum_{i=1}^n w_i \delta(x_i)$ converges weakly to $P$, since equivalent normed spaces have equivalent dual spaces and the Sobolev discrepancy in (10) controls weak convergence to $P$ (see the proof of Corollary 1). Likewise, since $\mathcal{H}(k_P)$ is equivalent to the quotient space $\mathcal{H}(k_{P,\sigma})/\mathcal{H}(\sigma^2)$ and $\mathcal{H}(\sigma^2)$ is just a space of constant functions on $M$, it follows that $\mathcal{H}(k_P)$ is equivalent to the quotient space $W^s_2(M)/\mathcal{H}(\sigma^2)$. This provides new insight into KSD; we have shown that the KSD in (8) is equivalent to the worst case integration error over the unit ball in $W^s_2(M)/\mathcal{H}(\sigma^2)$.

Corollary 1. For any sequence $(P_n)_{n \in \mathbb{N}} \subset \mathcal{P}(M)$ of discrete measures $P_n := \sum_{i=1}^n w_i^{(n)} \delta(x_i^{(n)})$ with $\sum_{i=1}^n w_i^{(n)} = 1$ and $x_i^{(n)} \in M$, we have KSD($P_n, P$) → 0 if and only if $P_n$ converges weakly to $P$.

3.5. Second Main Result: Error Bounds for the Riemann–Stein Kernel Method

In this section we demonstrate how Theorem 1 leads to explicit finite-sample-size error bounds for the Riemann–Stein kernel method. It will be required that $X = \{x_i\}_{i=1}^n$ covers the manifold $M$ in the following sense [see e.g. 83]:

Definition 2. The fill distance of a set $X = \{x_i\}_{i=1}^n \subset M$ of points in $M$ is defined as $h_X := \max_{x \in M} \min_{i=1, \ldots, n} d_M(x, x_i)$, where $d_M$ is the geodesic distance on the Riemannian manifold.

The distribution $P$ defines a linear operator on $C(M, \mathbb{R})$ that we denote, in a small abuse of notation, by $P(f) := \int_M f dP$. A general error bound can now be stated:

Theorem 2. For all $f \in W^s_2(M)$,

$$|P_X(f) - P(f)| \leq C_s h_X^s \|f\|_{W^s_2(M)},$$

(11)

where $C_s$ is a constant depending on $s$ but independent of $f$ and of the point set $X$.

Thus when the points in $X$ cover the manifold $M$, in the sense that the fill distance $h_X$ is small, the estimator $P_X(f)$ is an accurate approximation to the true integral. For $P$ equivalent (in the sense of measures) to the natural volume measure $V$, an information-theoretic lower bound on the worst case error for any estimator $P_X$, based on a size $n$ point set $X$, is $Cn^{-s/d}$, for some $C > 0$ [13]. This
shows that the Riemann–Stein estimator \( P_X \) is rate-optimal whenever the point set \( X \) is selected such that the fill-distance is asymptotically minimised, i.e. \( h_X = O(n^{-1/d}) \). This is in principle a weak requirement, as under suitable conditions even independently sampled \( x_i \sim V \), where \( V \) denotes the normalised Riemannian measure on the manifold, achieve this rate in probability up to a logarithmic factor; see [79, Thm. 3.2, Cor. 3.3] and [26].

The most popular approaches to Bayesian computation are based on sampling, in particular MCMC. Next we present the consequences of Theorem 2 in the case where the point set \( X \) arises as MCMC output. For measurable \( A \subseteq M \) we let \( P(A) := P(1_A) \) where \( 1_A(x) = 1 \) if \( x \in M \), otherwise \( 1_A(x) = 0 \). Recall that notions of geometric and uniform ergodicity coincide on a compact state space when the invariant measure \( P \) is equivalent (in the sense of measures) to \( V \); we therefore describe a Markov chain \( (x_i)_{i \in \mathbb{N}} \) with \( n \)th step transition kernel \( P^n \) and invariant distribution \( P \) simply as ergodic if there exists a finite constant \( C \) and a number \( 0 \leq \rho < 1 \) such that \( |P^n(x_0, A) - P(A)| \leq C \rho^n \) for all \( x_0 \in M \) and all measurable \( A \subseteq M \). The reader is referred to [64] for background.

**Corollary 2** (Second Main Result). Suppose that \( (x_i)_{i \in \mathbb{N}} \) is an ergodic Markov chain with invariant distribution \( P \), initialised at an arbitrary point \( x_0 \in M \). Let \( X = \{x_i\}_{i=1}^n \) and denote expectation with respect to the sampling distribution of \( X \) as \( \mathbb{E} \). Then
\[
\mathbb{E} |P_X(f) - P(f)| \leq C'_s n^{-\frac{d}{2}} \log(n)^{\frac{d}{2}} \|f\|_{W_s^2(M)}^2,
\]
where \( C'_s \) is a constant depending on \( s \) but independent of \( f \).

This result shows that the Riemann–Stein kernel method provides an asymptotically more accurate approximation of \( P(f) \) compared to using MCMC whenever \( s > \frac{d}{2} \) (which was also assumed in Assumption 2). Note that the constant \( C'_s \) is dependent on \( P \) through the infimum of \( p(x) \) on \( x \in M \) and could be large when \( P \) is highly concentrated. If the \( O(n^3) \) computational cost is also taken into account, the Riemann–Stein method asymptotically out-performs MCMC on an error-per-cost basis only when \( s > \frac{3d}{2} \), following the argument in Section 1. As usual, the case of high-dimensional manifolds (i.e. \( d \) large) is likely to challenge any interpolation-based method unless stronger assumptions can be made on the integrand. An empirical validation is contained in Appendix A.

4. **Proofs**

This section contains proofs of the theoretical results just presented. To this end, there are two main theoretical challenges to be addressed: First, it is necessary to establish that \( \kappa_{P, \sigma} \) is a valid kernel so that, from the Moore–Aronszajn theorem, \( \kappa_{P, \sigma} \) defines an RKHS. This is addressed first in Section 4.1. The RKHS \( \mathcal{H}(\kappa_{P, \sigma}) \) will be called the Stein RKHS. The remainder of Section 4.1 is devoted to proving Theorem 1, which establishes the equivalence of the Stein RKHS with a standard Sobolev space on the Riemannian manifold. From this point onward, our interpolation error bounds are standard in the Sobolev space context and the proof for Theorem 2 is contained in Section 4.3. The corollary for MCMC are established in Section 4.4.

4.1. **Proof of Theorem 1**

It is required to establish that \( \mathcal{H}(\kappa_{P, \sigma}) \) is norm-equivalent to the Sobolev space \( W^s(M) \). This is performed in two parts, with a Sobolev embedding of the Stein RKHS and a Stein embedding of the Sobolev RKHS. For both parts we leverage results from the analysis of partial differential equations on
a Riemannian manifold; our main reference is [39]. The first result establishes how the kernel $k_P$ can be computed:

**Lemma 1.** Let $\langle \cdot, \cdot \rangle_{\mathcal{H}(k)}$ denote the inner product in $\mathcal{H}(k)$. Then $k_P(x,x') := \mathcal{A}'_P \mathcal{A}_P k(x,x') = \langle \mathcal{A}_P k(x,\cdot), \mathcal{A}'_P k(x',\cdot) \rangle_{\mathcal{H}(k)}$. In particular, $k_P$ is symmetric and semi-positive definite; i.e. $k_P$ is a kernel.

**Proof.** Since $k$ reproduces $W^2_s(M)$ for $s = 2 > (d/2) + 2$, it follows from the Sobolev embedding theorem that $k(x,\cdot) \in C^2(M,\mathbb{R})$. Thus second order differential operators, such as $\mathcal{A}_P$, can be applied to this function [Lem. 4.34 of 97].

Let $g$ be the metric tensor on differential forms associated to $g$ by the musical isomorphism [32]. Here $g(\nabla f,\nabla h) = g^{-1}(df, dh)$ where $dh$ is the linear functional on $T_x M$ such that $dh(v) = v(h)$.

Recall that $\mathcal{A}_P'$ is used to denote the action of a differential operator $\mathcal{A}_P$ on $x'$. It follows that,

$$
g(\nabla' \log p, \nabla g(\nabla \log p, \nabla k)) = g^{-1}(d' \log p, d' g^{-1}(d \log p, dk))
$$

$$\quad= \sum_{i,j,k,l} g^{ij}(y)g'^{kl}(x) \partial_y \log p(y) \partial_{x'} \log p(x) \partial_y \partial_{x'} k(x, y)
$$

$$\quad= \nabla \log p \nabla' \log p(k).
$$

Here $\nabla \log p(k)$ is the function on $M$ which maps $x$ to $\nabla_x \log p(k) \in \mathbb{R}$, where $\nabla_x \log p \in T_x M$ is the gradient vector. Thus $\nabla \log p \nabla' \log p(k) = \nabla \log p(\nabla' \log p(k))$ is defined. Thus

$$\mathcal{A}'_P \mathcal{A}_P k(x,x') = \nabla \log p \nabla' \log p(k) + g^{-1}(d \log p, d' \Delta k) + g^{-1}(d' \log p, d' \Delta' k) + \Delta \Delta' k$$

$$\quad= \langle \mathcal{A}_P k(x,\cdot), \mathcal{A}_P'(x',\cdot) \rangle_{\mathcal{H}(k)}$$

where local coordinates verify that the last equality is established.

Finally, observe that the map $\phi : M \to \mathcal{H}(k)$, defined as $\phi(x) := \mathcal{A}_P k(x,\cdot)$, is a feature map for $k_P$, meaning that $k_P(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}(k)}$ for all $x,x' \in M$. It follows that $k_P$ is symmetric and semi-positive definite; indeed if $\{ w_i \}_{i=1}^n \subset \mathbb{R}$ and $\{ x_i \}_{i=1}^n \subset M$ then

$$\sum_{i=1}^n \sum_{j=1}^n w_i w_j k_P(x_i, x_j) = (\sum_{i=1}^n w_i \phi(x_i), \sum_{j=1}^n w_j \phi(x_j))_{\mathcal{H}(k)} = \| \sum_{i=1}^n w_i \phi(x_i) \|^2_{\mathcal{H}(k)} \geq 0$$

as required.

**Lemma 2.** $k_{P,\sigma}$ is symmetric and semi-positive definite; i.e. $k_{P,\sigma}$ is a kernel.

**Proof.** Let $\Phi := \mathbb{R} \times \mathcal{H}(k)$ denote the Hilbert space with inner product $\langle (c_1, h_1), (c_2, h_2) \rangle_{\Phi} := c_1 c_2 + \langle h_1, h_2 \rangle_{\mathcal{H}(k)}$ for all $c_1, c_2 \in \mathbb{R}$ and all $h_1, h_2 \in \mathcal{H}(k)$. Denote the associated norm $\| \|_{\Phi}$. Then the map $\phi : M \to \Phi$, defined as $\phi(x) := [\sigma, \mathcal{A}_P k(x,\cdot)]$, is a feature map for $k_{P,\sigma}$, meaning that $k_{P,\sigma}(x,x') = \langle \phi(x), \phi(x') \rangle_{\Phi}$ for all $x,x' \in M$. It follows that $k_{P,\sigma}$ is symmetric and semi-positive definite, as explained in the proof of Lemma 1.

The Stein RKHS $\mathcal{H}(k_{P,\sigma})$ has the form $\mathcal{H}(\mathcal{A}(\sigma^2)) \oplus \mathcal{H}(k_P)$ where $\mathcal{H}(\sigma^2)$ is the RKHS with constant kernel $\sigma^2$. Moreover, from Thm. 4.21 of [97], for $\zeta \in \mathcal{H}(k_P)$

$$\| \zeta \|^2_{\mathcal{H}(k_P)} = \inf \{ \| h \|^2_{\mathcal{H}(k)} : h \in \mathcal{H}, \mathcal{A}_P h = \zeta \}.$$ 

Thus, from Thm. 5 of [10],

$$\| f \|^2_{\mathcal{H}(k_{P,\sigma})} = \inf \{ \sigma^{-2} \xi^2 + \| \zeta \|^2_{\mathcal{H}(k_P)} : f = \xi + \zeta, \xi \in \mathbb{R}, \zeta \in \mathcal{H}(k_P) \}$$ (12)
\[
\text{infimum is sought in (12) in fact contains a single element. Let } f = \xi + A_P h, \xi \in \mathbb{R}, h \in \mathcal{H}(k) \text{.}
\]

Note that, since the only constant function in \( H \) is the zero function, the infimum is sought in (12) in fact contains a single element. Let \( f = \xi + A_P h \) where

\[
(\hat{\xi}, \hat{h}) = \arg \inf_{\xi \in \mathbb{R}, h \in \mathcal{H}(k)} \sum_{i=1}^{n} \left( \xi + A_P h(x_i) - f(x_i) \right)^2 + R_1(\xi) + R_2(h)
\]

and \( R_1, R_2 \) are defined in (3). The \( (\hat{\xi}, \hat{h}) \) exist and are unique from the representer theorem \([84]\). To understand the Riemann–Stein kernel method it suffices to study the convergence of \( \hat{f} \) to \( f \) in the Stein RKHS. To this end, note at this point that the proof of the representer theorem shows that \( \hat{f} \) is the orthogonal projection of \( f \) onto the span of \( \{k_{P,\sigma}(\cdot, x_1), \ldots, k_{P,\sigma}(\cdot, x_n)\} \) in \( \mathcal{H}(k_{P,\sigma}) \), and thus we have

\[
\|f - \hat{f}\|_{\mathcal{H}(k_{P,\sigma})} \leq \|f\|_{\mathcal{H}(k_{P,\sigma})},
\]

the so-called best approximation property.

Theorem 1 will now be proven in two parts; first we establish a Sobolev embedding of the Stein RKHS in Lemma 3 using a derivative counting argument, then a Stein embedding of the Sobolev RKHS is established in Lemma 4 using results from the theory of partial differential equations on a Riemannian manifold.

**Lemma 3 (Stein Embedding of Sobolev RKHS).** The space \( \mathcal{H}(k_{P,\sigma}) \) is continuously embedded in \( W^s_2(M) \), i.e. for some constant \( 0 < C < \infty \), we have \( \|f\|_{W^s_2(M)} \leq C\|f\|_{\mathcal{H}(k_{P,\sigma})} \).

**Proof.** The elements of \( \mathcal{H}(k_{P,\sigma}) \) are all of the form \( f = \xi + A_P h \) for \( \xi \in \mathbb{R}, h \in \mathcal{H}(k) \), and therefore it is sufficient to show that there is a finite constant \( C \) such that \( \|\xi + A_P h\|_{W^s_2(M)} \leq C\|f\|_{\mathcal{H}(k_{P,\sigma})} \). From the triangle inequality and the definition of Sobolev norm, \( \|\xi + A_P h\|_{W^s_2(M)} \leq C_1(\xi + \|A_P h\|_{W^{s+2}_2(M)}), \) where the constant \( C_1 \) depends on the volume of \( M \). Now, we also know that \( \|f\|^2_{\mathcal{H}(k_{P,\sigma})} = \sigma^{-2s^2} + \|h\|^2_{\mathcal{H}(k)} \) where \( h' \) is the element of \( \mathcal{H}(k) \) that minimises \( \|h'\|_{\mathcal{H}(k)} \) subject to \( A_P(h' - h) = 0 \). The result therefore follows if \( \|A_P h'\|_{W^s_2(M)} \leq C\|h'\|_{W^{s+2}_2(M)} \). This is a consequence of Lemma 2.4 in [39], and holds for the differential operator \( A_P = (1/p)D_a \) whenever the section \( a = pg(\cdot, \cdot) \) belongs to \( W^{s+2}_2(M, T^* M \otimes T^* M) \); see [6]. Here \( D_a := \nabla \cdot (p\nabla) \) is the differential operator associated to \( a \), and \( a \) is defined as the map that sends a point \( x \in M \) to the bilinear form \( p(x)g_x : T_x M \times T_x M \to \mathbb{R} \). Since \( M \) is compact and \( g \) is a smooth section, the condition holds for example whenever \( p \in C^{s+1}(M, \mathbb{R}) \). The final condition is implied by Assumption 3.

The converse of Lemma 3 is now established:

**Lemma 4 (Sobolev Embedding of Stein RKHS).** The space \( W^s_2(M) \) is continuously embedded in \( \mathcal{H}(k_{P,\sigma}) \); i.e. for some constant \( 0 < C < \infty \), we have \( \|f\|_{\mathcal{H}(k_{P,\sigma})} \leq C\|f\|_{W^s_2(M)} \).

**Proof.** Let \( f \in W^s_2(M) \) and consider solutions \( \xi \in \mathbb{R}, h \in W^{s+2}_2(M) \) to the partial differential equation

\[
A_P h + \xi = f \quad \text{in } M
\]

\[
g(\nabla h, n) = 0 \quad \text{on } \partial M,
\]

\[
= \inf \{ \sigma^{-2s^2} + \|h\|^2_{\mathcal{H}(k)} : f = \xi + A_P h, \xi \in \mathbb{R}, h \in \mathcal{H}(k) \}.
\]
sometimes called the Stein equation \[51\]. If the manifold has no boundary we simply drop the boundary condition and Stein equation reduces to \( A_P h + \xi = f \). Note that for the equation to be well-defined we must have \( \xi = P(f) \) due to Assumption 5. We re-write the first equation in (16) as \( D_a(h) = p(f - \xi) \) with the notation from the proof of Lemma 3. In order to obtain existence of solution we rely on the Fredholm alternative, according to which a solution exists if and only if \( p \) with the notation from the proof of Lemma 3. Observe that the adjoint of \( L \) is \( D \) defined by \( \langle L h, V \rangle \). In order to obtain existence of solution we rely on the Fredholm alternative, according to which a solution exists if and only if \( p(f - \xi) \) is orthogonal in \( L^2(M, dV) \) to the kernel of the adjoint in \( L^2(M, dV) \) of \( D_a \), see for example Corollary 10.4.10 \[73\]. Observe that the adjoint of \( D_a \) is \( D_a \) itself, since by the divergence theorem, for any \( h, u \in C^\infty(M) \)

\[
\int D_a(h)udV = \int u \nabla \cdot (p\nabla h)dV = -\int p\nabla h(u)dV = -\int p\nabla u(h)dV = \int h\nabla \cdot (p\nabla h)dV = \int hD_a(u)dV.
\]

Moreover the kernel of \( D_a \) consists only of constant functions, since if \( h \) belongs to \( \ker D_a \), then multiplying by \( h \) and integrating yields

\[
D_a(h) = 0 \quad \implies \quad 0 = \int hD_a(h)dV = \int h\nabla \cdot (p\nabla h)dV = -\int g(\nabla h, \nabla h)dP,
\]

and the positive-definiteness of the Riemannian metric, \( g(\nabla h, \nabla h) \geq 0 \), implies that \( g(\nabla h, \nabla h) \) is the zero function, which is equivalent to \( \nabla h = 0 \), which is equivalent to \( h = \text{constant} \). Existence hence follows by noting that \( p(f - \xi) \) is orthogonal to the constant functions in \( L^2(M, dV) \), which is a consequence of the fact that its integral with respect to \( P \) vanishes,

\[
\int p(f - \xi)hdV = h\int (f - \xi)dP = 0 \quad \forall \text{ constant functions } h.
\]

The existence of a solution implies that \( W^2_2(M) \) is a subset of \( \mathcal{H}(k_{P,\sigma}) \), and it remains to prove that it is continuously embedded. To do so we will apply Lemma 10.4.9 \[73\], which states that there exists \( 0 < C < \infty \) such that, for all \( u \in W^{2,2+\varepsilon}_2(M) \) that are orthogonal to \( \ker D_a \) in \( L^2(M) \),

\[
\|u\|_{W^{2,2+\varepsilon}_2(M)} \leq C\|D_a(u)\|_{W^{2}_2(M)}.
\]

Let \( \overline{h} \) be a solution to the Stein equation, which exists from the above argument, and define \( h = \overline{h} - \int \overline{h}dV \). Then \( h \) is a solution with vanishing integral and is thus orthogonal to \( \ker D_a \). Hence, for some finite constants \( C_i \),

\[
\|h\|_{W^{2,2+\varepsilon}_2(M)} \leq C_1\|p(f - \xi)\|_{W^{2}_2(M)} \leq C_2\|f - \xi\|_{W^{2}_2(M)} \leq C_2(\|f\|_{W^{2}_2(M)} + \|\xi\|_{W^{2}_2(M)})
\]

where in a small abuse of notation \( \xi \) denotes also the constant function with value \( \xi \) and we have used the fact that the density \( p \) is bounded above on \( M \). Using Jensen’s inequality \( \|\xi\|_{W^{2}_2(M)} \leq C_3\|f\|_{L^1(M)} \leq C_4\|f\|_{L^2(M)} \leq C_5\|f\|_{W^{2}_2(M)} \), thus \( \|h\|_{W^{2,2+\varepsilon}_2(M)} \leq C_6\|f\|_{W^{2}_2(M)} \).

To complete the proof, we have from (13) that

\[
\|f\|_{\mathcal{H}(k_{P,\sigma})}^2 \leq \sigma^{-2} P(f)^2 + \|h\|_{\mathcal{H}(k)}^2
\]

(17)

where \( h \) is the unique element in \( \mathcal{H}(k) \) that satisfies (16). Now, from Jensen’s inequality \( P(f)^2 \leq P(f^2) \) and, from the definition of the Sobolev norm and the fact that \( p \) is bounded above on \( M \), we
have \( \|f\|_{L^2(\rho)} \leq C_7 \|f\|_{W_2^2(M)} \). Moreover, by hypothesis \( \mathcal{H}(k) \) is norm-equivalent to \( W_2^{s+2}(M) \); i.e. \( \|h\|_{\mathcal{H}(k)} \leq C_8 \|h\|_{W_2^{s+2}(M)} \) for some finite constant \( C_8 \). Thus

\[
\|f\|^2_{\mathcal{H}(k_{P,\alpha})} \leq (\sigma^{-2}C_7^2 + C_6^2 C_8^2)\|f\|^2_{W_2^s(M)}
\]

as required, with \( C = (\sigma^{-2}C_7^2 + C_6^2 C_8^2)^{1/2} \).

Together, Lemmas 3 and 4 establish that the Stein RKHS is equivalent to the Sobolev RKHS and thus Theorem 1 is established. This result, which is likely to be of independent interest, enables us to reduce the analysis of methods based on a Stein kernel to the analysis of methods based on a Sobolev kernel.

### 4.2. Proof of Corollary 1

The argument from the main text establishes that KSD\((P_n, P)\) → 0 if and only if MMD\(k(P_n, P)\) → 0, where

\[
\text{MMD}_k(Q, P) := \sup \{Q(f) - P(f) : \|f\|_{\mathcal{H}(k)} \leq 1\}
\]

is called the maximum mean discrepancy associated with the kernel \( k \) [90]. Recall that a kernel \( k \) is said to be characteristic if \( P \mapsto \int_M k(\cdot, x)dP(x) \) is an injective map into \( \mathcal{H}(k) \). Thm. 3.1 of [88] ensures that, for a continuous, bounded and characteristic kernel \( k \) on a compact Hausdorff space \( M \), MMD\(k\) metrises weak convergence in \( P(M) \). Thus we aim to establish the preconditions of this result.

In our setting, \( M \) is a compact Hausdorff space and, from Assumption 2 and the Sobolev embedding theorem, \( \mathcal{H}(k) \subset C(M, \mathbb{R}) \). The inclusion \( \mathcal{H}(k) \subset C(M, \mathbb{R}) \) implies \( k \) is continuous and bounded, due to Cor. 3 in [89], so it remains only to argue that \( k \) is a characteristic kernel.

Recall that a positive definite kernel \( k : M \times M \rightarrow \mathbb{R} \) on a compact Hausdorff space \( M \) is said to be \( c \)-universal if \( \mathcal{H}(k) \) is dense in \( C(M, \mathbb{R}) \) with respect to the uniform norm on \( M \) [94]. Moreover, recall that a \( c \)-universal kernel on a compact Hausdorff space is a characteristic kernel [Thm. 6 in [89]]. Thus we must argue that \( \mathcal{H}(k) \) in Assumption 2 is dense in \( C(M, \mathbb{R}) \) with respect to the uniform norm on \( M \). i.e. that \( W_2^{s+2}(M) \) is dense in \( C(M, \mathbb{R}) \). This is a standard fact, which follows from the fact that \( W_2^{s+2}(M) \) is a sub-algebra of \( C(M, \mathbb{R}) \) that separates points and vanishes nowhere, together with the Stone–Weierstrass theorem on a compact domain:

**Theorem 3 (Stone–Weierstrass on compact \( M \)).** Let \( M \) be a compact Hausdorff space and \( F \) be a sub-algebra of \( C(M, \mathbb{R}) \). Then \( F \) is dense in \( C(M, \mathbb{R}) \) (equipped with the uniform norm on \( M \)) if and only if \( F \) separates points (\( \forall x, y \in M, \exists f \in F \text{ s.t. } f(x) \neq f(y) \)) and vanishes nowhere (\( \forall x \in M, \exists f \in F \text{ s.t. } f(x) \neq 0 \)).

**Proof.** A special case of the more general result for locally compact spaces due to [98].

### 4.3. Proof of Theorem 2

In light of Theorem 1, the proof of Theorem 2 is relatively standard. Indeed, we have established that \( \mathcal{H}(k_{P,\alpha}) \) and \( W_2^s(M) \) are equivalent, so the convergence of \( \hat{f} \) to \( f \) can be studied in the standard Sobolev space context. The following Lemma 5 follows immediately from Prop. 7 and Thm. 8 in [31], together with the fact all Riemannian metrics (and their induced norms) are equivalent on a compact manifold, see page 22 of [42].
Lemma 5. Let $M$ be a smooth, compact, $d$-dimensional Riemannian manifold. Then there exists an atlas $(U_i, \phi_i)$ on $M$ and constants $C_1, C_2 > 0$ such that, if $x, y \in U_j$ for some $j$, then

$$C_1\|\phi_j(x) - \phi_j(y)\|_2 \leq d_M(x, y) \leq C_2\|\phi_j(x) - \phi_j(y)\|_2$$

Moreover, let $h_X \cap U_j := \sup_{x \in U_j} \min_{y \in X} d_M(x, y)$ denote the fill distance restricted to $U_j$. Then, if we choose an atlas as above, there exists finite constants $h_0, C$ such that, if $X \subset M$ is a finite point set with $h_X < h_0$, then for all $U_j$

$$h_X \cap U_j \leq C h_X.$$

The output of the Stein kernel method satisfies $P_X(f) = P(\hat{f})$ and in particular this means that

$$|P(f) - P_X(f)| = |P(f) - \hat{f}| = \left| \int_M (f - \hat{f}) dP \right| \leq \|f - \hat{f}\|_{L_2(P)}.$$

Using Lemma 5, and following the argument used in Thm. 10 of [31], we have that

$$\|f - \hat{f}\|_{L_2(P)} \leq C h_X^s \|f - \hat{f}\|_{W_2^2(M)}$$

Hence, for some finite constants $C_i$,

$$\|f - \hat{f}\|_{W_2^2(M)} \leq C_1 \|f - \hat{f}\|_{H(\mathcal{K}_{P,\sigma})} \quad \text{(Lemma 3)}$$

$$\leq C_1 \|f\|_{H(\mathcal{K}_{P,\sigma})} \quad \text{(best approximation property; (15))}$$

$$\leq C_2 \|f\|_{W_2^2(M)} \quad \text{(Lemma 4)}.$$

Thus $|P(f) - P_X(f)| \leq C_3 h_X^s \|f\|_{W_2^2(M)}$, as required. This completes the proof of Theorem 2.

4.4. Proof of Corollary 2

The proof proceeds in two stages, and first we state and prove a result for independent samples that will serve as a stepping stone toward the main result:

Corollary 3. Suppose that $(x_i)_{i \in \mathbb{N}}$ is a sequence of independent samples from the distribution $P$. Let $X = \{x_i\}_{i=1}^n$ and denote expectation with respect to the sampling distribution of $X$ as $\mathbb{E}$. Then

$$\mathbb{E}[P_X(f) - P(f)] \leq C'_s n^{-\frac{s}{2}} \log(n)^{\frac{3}{2}} \|f\|_{W_2^2(M)},$$

where $C'_s$ is a constant depending on $s$ but independent of $f$.

Proof of Corollary 3. First, let $(Y_i : \Omega \to M)_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables distributed according to the normalised Riemannian measure $\tilde{V}$, i.e., $P \circ Y_i^{-1} \equiv \tilde{V}$, where $P$ is the probability measure on $\Omega$. Let $Y^n = \{Y_i\}_{i=1}^n$. From [79, Thm. 3.2, Cor. 3.3] we have that

$$\mathbb{E}[h_{Y^n}] \leq C''_s n^{-\frac{s}{2}} \log(n)^{\frac{3}{2}}$$

for some $C''_s$ a finite $s$-dependant constant. Since $M$ is compact and $p$ is continuous (Assumption 3), the Radon–Nikodym derivative $dP/d\tilde{V}$ can be bounded away from zero almost everywhere on $M$, say
by \( w > 0 \), and without loss of generality we can assume that \( w < 1 \). It follows that \( P \) can be represented as a bivariate mixture, one of whose components is \( \bar{V} \). Specifically, \( P = w\bar{V} + (1 - w)Q \), where \( Q \) has density

\[
\frac{dQ}{d\bar{V}} = \frac{1}{1-w} \left( \frac{dP}{d\bar{V}} - w \right).
\]

Let us introduce a Bernoulli random variable \( B \sim \text{Bernoulli}(w) \), and a random variable \( Z \) with conditional distributions

\[
Z|B \sim \begin{cases} \bar{V} & \text{if } B = 1 \\ Q & \text{if } B = 0 \end{cases}
\]

It follows that the law of \( Z \) is \( P \), since

\[
P \circ Z^{-1}(A) = \mathbb{P}(Z \in A|B = 1)\mathbb{P}(B = 1) + \mathbb{P}(Z \in A|B = 0)\mathbb{P}(B = 0)
= \bar{V}(A)w + Q(A)(1 - w)
= P(A),
\]

where \( w = \mathbb{P}(B = 1) \). Consider then such a sequence \( (Z_i)_{i \in \mathbb{N}} \) of i.i.d. random variables with the same law as \( Z \), and let \( Z := \{Z_i : 1 \leq i \leq n\} \). If \( m \) of the variables in \( Z \) were obtained by sampling from \( \bar{V} \) (which arise whenever the Bernoulli random variable takes the value \( B = 1 \)), then without loss of generality \( Z|m = (Y_1, \ldots, Y_m, W_{m+1}, \ldots, W_n) \), where \( W_i \sim Q \).

From the law of conditional expectation, the fact that \( X \) and \( Z \) are identically distributed, the fact that \( h_s Z|m \leq h_s Y_m \) and (19),

\[
\mathbb{E}[h_X] = \mathbb{E} \mathbb{E}[h_Z|m] \leq \mathbb{E} \mathbb{E}[h_{Y,m}|m] \\
= \mathbb{E} \mathbb{E} \left[ \min(\text{diam}(M), h_{Y,m}) | m \right] \\
\leq \mathbb{E} \left[ \min(\text{diam}(M), C''m^{-\frac{s}{d}} \log(m)^{\frac{s}{d}}) \right].
\]

Fix \( 0 < \epsilon < w \). From the law of total expectation, for \( n > 1 \):

\[
\begin{align*}
\mathbb{E} \left[ \min(\text{diam}(M), C''m^{-\frac{s}{d}} \log(m)^{\frac{s}{d}}) \right] & \leq \text{diam}(M) - \frac{\mathbb{P}(m = 0)}{n^{-\frac{s}{d}} \log(n)^{\frac{s}{d}}} \\
& \quad + \mathbb{E} \left[ \min(\text{diam}(M), C''m^{-\frac{s}{d}} \log(m)^{\frac{s}{d}}) | 0 < m \leq \epsilon n \right] \frac{\mathbb{P}(0 < m \leq \epsilon n)}{n^{-\frac{s}{d}} \log(n)^{\frac{s}{d}}} \\
& \quad + \sum_{m > \epsilon n} \mathbb{E} \left[ \min(\text{diam}(M), C''m^{-\frac{s}{d}} \log(m)^{\frac{s}{d}}) | m = \tilde{m} \right] \frac{\mathbb{P}(m = \tilde{m})}{n^{-\frac{s}{d}} \log(n)^{\frac{s}{d}}}.
\end{align*}
\]
The first term is seen to vanish as \( n \to \infty \):
\[
(*) = \text{diam}(M) n^{1/2} \log(n)^{1/2} (1 - w)^n \to 0
\]

Since \( m \sim \text{Binomial}(n, w) \), we have that \( \log(m) / \log(n) \leq 1 \) and moreover from Hoeffding’s inequality we have that
\[
\mathbb{P}(m \leq \epsilon n) \leq \exp(-2(w - \epsilon)^2 n).
\]

Thus, letting \( C''''_s = \max(\text{diam}(M), C''''_s) \), the second term is also seen to vanish as \( n \to \infty \):
\[
(**) \leq \mathbb{E} \left[ \min \left( \text{diam}(M), C''''_s \left( \frac{m}{n} \right)^{-1/2} \right) \right] \mathbb{P}(0 < m \leq \epsilon n)
\]
\[
\leq C''''_s n^{1/2} \mathbb{P}(0 < m \leq \epsilon n)
\]
\[
\leq C''''_s n^{1/2} \exp(-2(w - \epsilon)^2 n) \to 0
\]

For the final term, let \( g : [0, 1] \to \mathbb{R} \) be defined as
\[
g(x) = \begin{cases} x^{-1/2} & x \leq \epsilon \\ \epsilon^{-1/2} x & x > \epsilon \end{cases}
\]

which is observed to be a continuous and bounded. From the strong law of large numbers, \( m/n \) converges a.e. to \( w \), and thus in distribution to \( \delta(w) \). From the Portmanteau theorem we have that \( \mathbb{E}[g(m/n)] \to g(w) = w^{-s/d} \). Thus the third term is bounded as \( n \to \infty \):
\[
(***) \leq \sum_{\tilde{m} > \epsilon n} \mathbb{P}(m = \tilde{m})
\]
\[
\leq C''''_s \sum_{\tilde{m} > \epsilon n} \left( \frac{\tilde{m}}{n} \right)^{-1/2} \mathbb{P}(m = \tilde{m})
\]
\[
\leq C''''_s \mathbb{E} \left[ g \left( \frac{m}{n} \right) \right] \to C''''_s w^{-1/2} < \infty
\]

Together with Theorem 2, the result is now established. \( \square \)

Now we turn to the proof of Corollary 2:

**Proof of Corollary 2.** The proof focuses on \((x_{i\infty})_{i \in \mathbb{N}}\), which is sometimes referred to as the \( n \)-step jump chain and \( P^n \) is its transition kernel. Since \( M \) is compact and \( P \) is continuous (Assumption 3), the \( n \)th step transition distributions initialised from \( x_{i\infty} \in M \), denoted \( P^n_{i,\cdot}(\cdot) := P^n(x_{i\infty}, \cdot) \), are absolutely continuous with respect to \( P \) and therefore admit Radon-Nikodym derivatives \( dP^n_{i,\cdot} / dP \).

Since the Markov chain is ergodic, \( |dP^n_{i,\cdot} / dP - 1| \leq C \rho^n \) and we can select a value \( n_0 \in \mathbb{N} \) independent of the \( i \) index such that, for some \( w < 1 \) and all \( n \geq n_0 \), it holds almost everywhere that \( dP^n_{i,\cdot} / dP \geq w > 0 \). It follows that \( P^n_{i,\cdot} \) can be represented as a bivariate mixture, one of whose components is \( P \). Specifically, \( x_{i\infty} \) is equal in distribution to \( z_i \), where

\[
\begin{align*}
\text{if } B_i = 1 & \text{, then } B_i \overset{i.i.d.}{\sim} \text{Bernoulli}(w), \quad \frac{dQ_i}{dP} = \frac{1}{1-w} \left( \frac{dP^n_{i,\cdot-1}}{dP} - w \right), \\
\text{if } B_i = 0 &
\end{align*}
\]
Consider then such a sequence \((z_i)_{i \in \mathbb{N}}\) and let \(Z = \{z_i : 1 \leq i \leq [n/n_0]\}\) and \(Y = \{z_i : B_i = 1, 1 \leq i \leq [n/n_0]\}\). The elements of \(Y\) are independent samples from \(P\) and \(m := |Y| \sim \text{Binomial}([n/n_0])\).

The remainder of the proof is identical to that used for Corollary 3.

5. Discussion

This paper adds to the growing literature on computational uses for Stein’s method. Our contribution provides both a manifold generalisation and a refined theoretical analysis of the Stein kernel interpolant. This was achieved by characterising kernel Stein discrepancy as a Sobolev discrepancy in the context of a compact manifold, and this characterisation may be of independent interest. For example, [69] proposed to use Sobolev discrepancy to train generative adversarial networks (GAN); our result justifies the alternative use of KSD for training GANs in a manifold context. See also [68, 70, 2, 62] for uses of Sobolev discrepancy, which could instead be achieved using KSD. Our main contribution was theoretical, but for completeness our results are empirically verified on \(M = S^2\) in the electronic supplement.

Some extensions of this work could include; additional constructions to circumvent the need for gradient information on the target [41]; the simplification of computation when the density \(p\) can be factorised [107]; and extension to high- or infinite-dimensional spaces such as the Hilbert sphere \(S^\infty\) for functional data analysis on a manifold.

Appendix A: Empirical Validation

In this appendix we report experiments designed to validate our theoretical results for the Riemann–Stein kernel method. For this purpose we considered arguably the most important compact manifold; the sphere \(S^2\).

Differential Operators

The coordinate patch \(\phi\) from Example 1 in Section 2 can be used to compute the metric tensor

\[
G = \begin{pmatrix} \sin^2 q_2 & 0 \\ 0 & 1 \end{pmatrix}
\]

and a natural volume element \(dV = \sin q_2 \, dq_1 \, dq_2\). It follows that, for a function \(h : S^2 \to \mathbb{R}\), we have the gradient differential operator

\[
\nabla h = \frac{1}{\sin^2 q_2} \frac{\partial h}{\partial q_1} \frac{\partial}{\partial q_1} + \frac{\partial h}{\partial q_2} \frac{\partial}{\partial q_2}.
\]

Similarly, for a vector field \(s = s_1 \frac{\partial}{\partial q_1} + s_2 \frac{\partial}{\partial q_2}\), we have the divergence operator

\[
\nabla \cdot s = \frac{\partial s_1}{\partial q_1} + \frac{\partial s_2}{\partial q_2} + \frac{\cos q_2}{\sin q_2} s_2.
\]

Thus the Riemann–Stein operator \(A_P\) is:

\[
A_P h = \frac{\cos q_2}{\sin q_2} \frac{\partial h}{\partial q_2} + \frac{1}{\sin^2 q_2} \left\{ \frac{1}{p} \frac{\partial p}{\partial q_1} \frac{\partial h}{\partial q_1} + \frac{\partial^2 h}{\partial q_1^2} \right\} + \left\{ \frac{1}{p} \frac{\partial p}{\partial q_2} \frac{\partial h}{\partial q_2} + \frac{\partial^2 h}{\partial q_2^2} \right\}.
\] (20)
Turning this into expressions in terms of $x$ requires that we notice
\[
\frac{\cos q_2}{\sin q_2} = \frac{x_3}{\sqrt{1 - x_3^2}}, \quad \frac{1}{\sin^2 q_2} = \frac{1}{1 - x_3^2},
\]
and use chain rule for partial differentiation.

**Choice of Kernel**

The performance of the Riemann–Stein kernel method depends, of course, on the selection of a reproducing kernel $k$ to define the space $\mathcal{H}(k)$. For standard manifolds, such as the sphere $M = \mathbb{S}^2$, several function spaces and their reproducing kernels have been studied [e.g. 78]. For more general manifolds, an extrinsic kernel can be induced under embedding into an ambient space [52], or the stochastic partial differential approach [27, 53, 74] can be used to numerically approximate a suitable intrinsic kernel. Note that none of the theoretical development in this paper relies on an embedding of $M$ into an ambient space; all of our analysis is intrinsic to the manifold.

Although our focus is on $\mathbb{S}^2$, for generality the remainder of this section discusses kernels on $\mathbb{S}^d$. A kernel on $\mathbb{S}^d$ is characterised by a scalar $\sigma > 0$ and a sequence $(b_{n,d})_{n=0}^{\infty}$ of $d$-Schoenberg coefficients, such that $0 \leq b_{n,d}$ and $\sum_{n=0}^{\infty} b_{n,d} = 1$, in the sense that
\[
k(x, y) = \sigma^2 \sum_{n=0}^{\infty} b_{n,d} C_n^{(d-1)/2}(x \cdot y), \quad x, y \in \mathbb{S}^d
\]
where $C_n^\lambda$ are the Geigenbauer polynomials of degree $n$ and order $\lambda > 0$ [12, 61, 20, 36, 21]. It is known that $b_{n,d} \approx n^{-2\alpha}$ if and only if $\mathcal{H}(k)$ is norm-equivalent to the Sobolev space $W_2^{\alpha}(\mathbb{S}^d)$ [21]. Thus in principle one has much scope to design an appropriate kernel whilst also ensuring that Assumption 2 is satisfied. Therefore we need to elicit some additional desiderata (D) to constrain ourselves to those kernels which are most useful:

**Desiderata 1.** The kernel should have an explicit form that can be easily differentiated.

**Desiderata 2.** The kernel should have easily customisable smoothness.

For the Riemann–Stein kernel method to be practical, D1 must hold. For the Stein kernel method to be flexibly used, D2 must hold. Note that one cannot, for example, just restrict the Matérn kernel on $\mathbb{R}^3$ to $\mathbb{S}^2$; in order that the restriction is positive definite we require a strong condition $\nu \in (0, \frac{1}{2}]$ on the smoothness parameter [36]. This fact makes this approach not suitable for our work, where higher order derivatives of the kernel are needed. Moreover, the restriction of a Sobolev space to a manifold incurs a loss-of-smootheness that depends on the manifold. In what follows, two different kernels that each reproduce Sobolev spaces are presented:

**Kernel 1** (Brauchart; [14]). For $\alpha + \frac{1}{2} - \frac{d}{2} \in \mathbb{N}$ and $\alpha > \frac{d}{2}$, the kernel
\[
k_1(x, y) = C(1)^3 F_2 \left[ \begin{array}{c} \frac{d}{2} + \frac{1}{2} - \alpha, \frac{d}{2} - \alpha, \frac{d}{2} + 1 - \alpha, \frac{d}{2} - 2\alpha - \alpha \parallel x - y \parallel^2 / 2 \\ \frac{d}{2} + 1 - \alpha, 1 + \frac{d}{2} - 2\alpha \end{array} \right] + C(2) \parallel x - y \parallel^{2\alpha - 2}, \tag{22}
\]
defined for \( x, y \in \mathbb{S}^d \), reproduces the Sobolev space \( W^\alpha_2(\mathbb{S}^d) \). In this paper \( pFq \) is the generalised hypergeometric function. The constant terms in the kernel in (22) are as follows:

\[
C^{(1)} = \frac{2^{2\alpha-2} \left( \frac{d}{2} \right)_{2\alpha-2}}{2\alpha - d} \frac{\Gamma(d+1)}{\Gamma(d)} \frac{\Gamma(\alpha - \frac{d}{2} + \frac{1}{2}) \Gamma(\alpha - \frac{d}{2} + \frac{1}{2})}{\sqrt{\pi} \Gamma(d)(\frac{1}{2})_{\alpha - d + \frac{1}{2}}} \frac{1}{(\alpha - d + \frac{1}{2})_{\alpha - d + \frac{1}{2}}}
\]

where \((z)_n := \Gamma(z + n)/\Gamma(z)\) is the Pochhammer symbol. From properties of hypergeometric functions, this kernel has an explicit closed form when \( \alpha + \frac{1}{2} - \frac{d}{2} \in \mathbb{N} \), so that \( D1 \) is satisfied. Moreover, \( D2 \) holds for this kernel.

**Kernel 2 (Wendland; [102]).** Consider the compact support positive definite functions due to [102]

\[
\phi_{i,j}(r) = \begin{cases} 
  p_{i,j}(r) & \text{if } r \leq 1 \\
  0 & \text{if } r > 1 
\end{cases}
\]

defined for \( r \geq 0 \) where \( p_{i,j} : [0, \infty) \to \mathbb{R} \) is a particular polynomial selected such that \( \phi_{i,j} \) is positive definite and \( \phi_{i,j} \in C^2 \). The radial basis function

\[
k_2(x, y) = \phi_{i,j}\left( \frac{\|x - y\|_2}{\lambda} \right)
\]

for \( \lambda > 0 \) reproduces \( W^{j+\frac{1}{2} + \frac{1}{2}}j_2(\mathbb{R}^d) \) on \( x, y \in \mathbb{R}^d \) [see e.g. Thm. 2.1 in 103]. In the particular case where \( i = d + 1 \), the restriction of \( k_2 \) to \( x, y \in \mathbb{S}^d \) reproduces \( W^\alpha_2(\mathbb{S}^d) \) with \( \alpha = j + \frac{1}{2} \); see Thm. 4.1 of [71]. See also [35, 72, 105, 11]. Desiderata \( D2 \) is therefore satisfied. Moreover, this kernel has a closed form so that \( D1 \) is satisfied.

**Assessment**

To numerically assess the convergence of the Riemann–Stein kernel method, consider the von Mises-Fisher distribution \( P \) whose density with respect to \( V \) is

\[
p(x) = \frac{\|c\|_2}{4\pi \sinh(\|c\|_2)} \exp(c^\top x).
\]

For illustration, we suppose that the normalisation constant is unknown and we are told only that \( p(x) \propto \exp(c^\top x) \). This is sufficient to construct the differential operator \( A_P \) as previously described. Our aim in what follows is to validate our theoretical analysis; for this reason in all experiments we fixed \( \lambda = 2 \) for \( k_2 \) as a convenient default. Further theoretical work will be needed to understand the properties of the Riemann–Stein kernel method when kernel parameters are adaptively estimated [96].

In what follows we first considered point sets \( X = \{x_i\}_{i=1}^n \) whose elements were quasi-uniformly distributed on \( \mathbb{S}^2 \), being obtained by minimising a generalised electrostatic potential energy [Reisz’s energy; 86]. Note that these points, being uniform, do not arise as an approximation to \( P \); rather, they are intended to asymptotically minimise \( h_X \) as motivated by Theorem 2. First, explicit function approximations are presented based on \( k_1, k_2 \) respectively in Figures 1a and 1b. These function approximations are \( \hat{f} = \xi + A_P h \) where \( \xi, h \) solve (2). Here the integrand \( f(x) \) was based on a Rosenbrock function and represents a modest challenge to an interpolation-based integration method.
Figure 1: Function approximation with the Riemann–Stein kernel. Here $f$ represents the exact integrand. Each panel represents a kernel (a) $k_1$, (b) $k_2$, varying both the smoothness $\alpha$ of these kernel and the number $n$ of evaluations of the integrand. The point set was quasi-uniform over $S^2$.

It was observed that both kernels provided an accurate approximation when a large number of points were used. Next, for various values of $n$, we computed the worst case integration error in (7) (i.e. the generalised KSD). Results in Figures 2a and 2b show that, for both kernels $k_1$ and $k_2$, convergence of the generalised KSD occurred at the rate that was theoretically predicted. Finally, we re-evaluated the worst case integration error (i.e. the generalised KSD), based instead on a point set $X$ generated as the realisation of an MCMC sample path. Here we employed a Metropolis-Hastings Markov chain with
Figure 2: The worst case integration error (i.e. generalised kernel Stein discrepancy; KSD) of the Riemann–Stein kernel method was plotted for two different kernels, (a) $k_1$, (b) $k_2$, varying both the smoothness $\alpha$ of these kernels and the number $n$ of evaluations of the integrand. The point set was quasi-uniform over $S^2$. Dashed lines represent the slope of the convergence rates that we have theoretically established.

Figure 3: The worst case integration error (i.e. generalised kernel Stein discrepancy; KSD) of the Riemann–Stein kernel method was plotted for two different kernels, (a) $k_1$, (b) $k_2$, varying both the smoothness $\alpha$ of these kernels and the number $n$ of evaluations of the integrand. The point set was obtained as the realisation of a Markov chain whose invariant distribution was $P$. Here the arithmetic mean estimator is presented along with standard error bars, averaged over multiple realisations of the Markov chain. Dashed lines represent the slope of the convergence rates that we have theoretically established.
the normalised Riemannian measure $\bar{V}$ as the proposal. Results are presented in Figures 3a and 3b, where again the theoretically obtained convergence rate was validated.

Acknowledgements

CJO and MG were supported by the Lloyd’s Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. AB was supported by a Roth scholarship from the Department of Mathematics at Imperial College London, UK. EP was partially supported by FONDECYT Grant [1170290], Chile, and by Initiativa Científica Milenio - Minecon Nucleo Milenio MESC. MG was supported by the EPSRC grants [EP/K034154/1, EP/R018413/1, EP/P020720/1, EP/L014165/1], an EPSRC Established Career Fellowship [EP/J016934/1] and a Royal Academy of Engineering Research Chair in Data Centric Engineering. The authors are grateful for discussions with Andrew Duncan, Toni Karvonen, Chang Liu, Gustav Holzegel, Julio Delgado and Andrew Stuart, and to an Associate Editor and Reviewer for constructive feedback on an earlier version of the manuscript.

References


