Consistent Interpolating Ensembles via the Manifold-Hilbert Kernel

Yutong Wang, Clayton Scott
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109
{yutongw, clayscot}@umich.edu

May 20, 2022

Abstract

Recent research in the theory of overparametrized learning has sought to establish generalization guarantees in the interpolating regime. Such results have been established for a few common classes of methods, but so far not for ensemble methods. We devise an ensemble classification method that simultaneously interpolates the training data, and is consistent for a broad class of data distributions. To this end, we define the *manifold-Hilbert kernel* for data distributed on a Riemannian manifold. We prove that kernel smoothing regression and classification using the manifold-Hilbert kernel are weakly consistent in the setting of Devroye et al. [19]. For the sphere, we show that the manifold-Hilbert kernel can be realized as a weighted random partition kernel, which arises as an infinite ensemble of partition-based classifiers.

1 Introduction

Ensemble methods are among the most often applied learning algorithms, yet their theoretical properties have not been fully understood [11]. Based on empirical evidence, Wyner et al. [40] conjectured that interpolation of the training data plays a key role in explaining the success of AdaBoost and random forests. However, while a few classes of learning methods have been analyzed in the interpolating regime [5, 3], ensembles have not.

Towards developing the theory of interpolating ensembles, we examine an ensemble classification method for data distributed on the sphere, and show that this classifier interpolates the training data and is consistent for a broad class of data distributions. To show this result, we develop two additional contributions that may be of independent interest. First, for data distributed on a Riemannian manifold M, we introduce the *manifold-Hilbert kernel* $K_M^{\mathcal{H}}$, a manifold extension of the *Hilbert kernel* [37]. Under the same setting as Devroye et al. [19], we prove that kernel smoothing regression with $K_M^{\mathcal{H}}$ is weakly consistent while interpolating the training data. Consequently, the classifier obtained by taking the sign of the kernel smoothing estimate has zero training error and is consistent.

Second, we introduce a class of kernels called weighted random partition kernels. These are kernels that can be realized as an infinite, weighted ensemble of partition-based histogram classifiers. Our main result is established by showing that when $M = \mathbb{S}^d$, the d-dimensional sphere, the manifold-Hilbert kernel is a weighted random partition kernel. In particular, we show that on the sphere, the manifold-Hilbert kernel is a weighted ensemble based on random hyperplane arrangements. This implies that the kernel smoothing classifier is a consistent, interpolating ensemble on $\mathbb{S}^d$. To our knowledge, this is the first demonstration of an interpolating ensemble method that is consistent for a broad class of distributions in arbitrary dimensions.

1.1 Problem statement

Consider the problem of binary classification on a Riemannian manifold M. Let (X, Y) be random variables jointly distributed on $M \times \{\pm 1\}$. Let $D^n := \{(X_i, Y_i)\}_{i=1}^n$ be the (random) training data consisting of n i.i.d copies of X, Y. A *classifier*, i.e., a mapping from D^n to a function $\widehat{f}(\bullet \| D^n) : M \to \{\pm 1\}$, has the **interpolating-consistent property** if, when X has a continuous distribution, both of the following hold: 1) $\widehat{f}(X_i \| D^n) = Y_i$, for all $i \in \{1, \dots, n\}$, and 2)

$$\Pr\{\widehat{f}(X \| D^n) \neq Y\} \to \inf_{f : M \to \{\pm 1\} \text{ measurable}} \Pr\{f(X) \neq Y\} \quad \text{in probability as } n \to \infty. \tag{1}$$

Our goal is to find an interpolating-consistent ensemble of *histogram classifiers*, to be defined below.

A *partition* on M, denoted by $\mathcal{P}$, is a set of subsets of M such that $P \cap P' = \emptyset$ for all $P, P' \in \mathcal{P}$ and $M = \bigcup_{P \in \mathcal{P}} P$. Given $x \in M$, let $\mathcal{P}[x]$ denote the unique element $P \in \mathcal{P}$ such that $x \in P$. The set of all partitions on a space M is denoted $\texttt{Part}(M)$. The *histogram classifier* with respect to D^n over $\mathcal{P}$ is the sign of the function $\widehat{h}(\bullet \| D^n, \mathcal{P}) : M \to \mathbb{R}$ given by

$$\widehat{h}(x \| D^n, \mathcal{P}) := \sum_{i=1}^n Y_i \cdot \mathbb{I}\{x \in \mathcal{P}[X_i]\}, \tag{2}$$

where $\mathbb{I}$ is the indicator function.

Definition 1.1. A *weighted random partition* (WRP) over M is a 3-tuple $(\Theta, \mathfrak{P}, \alpha)$ consisting of (i) *parameter space of partitions*: a set Θ where $\mathcal{P}_\theta \in \texttt{Part}(M)$ for each $\theta \in \Theta$, (ii) *random partitions*: a probability measure $\mathfrak{P}$ on Θ, and (iii) *weights*: a nonnegative function $\alpha : \Theta \to \mathbb{R}_{\geq 0}$.

Example 1.2 (Regular partition of the d-cube)**.** Let $M = [0, 1]^d$ and $\Theta = \{1, 2 \dots\} =: \mathbb{N}_+$. For each $n \in \mathbb{N}_+$, denote by $\mathcal{P}_n$ the regular partition of M into n^d d-cubes of side length $1/n$. For any probability mass function $\mathfrak{P}$ on $\mathbb{N}_+$ and weights $\alpha : \mathbb{N}_+ \to \mathbb{R}_{\geq 0}$, the 3-tuple $(\Theta, \mathfrak{P}, \alpha)$ is a WRP.

Below, WRPs will be denoted with 2-letter names in the sans-serif font, e.g., "rp" for a generic WRP, and "ha" for the weighted hyperplane arrangement random partition (Definition 5.1). The *weighted random partition kernel* associated to $\mathsf{rp} = (\Theta, \mathfrak{P}, \alpha)$ is defined as

$$K_M^{\mathsf{rp}} : M \times M \to \mathbb{R}_{\geq 0} \cup \{\infty\}, \quad K_M^{\mathsf{rp}}(x, z) := \mathbb{E}_{\theta \sim \mathfrak{P}}[\alpha(\theta)\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}]. \tag{3}$$

When $\alpha \equiv 1$, we recover the notion of unweighted random partition kernel introduced in [18]. Note that the kernel is symmetric since $\mathbb{I}\{x \in \mathcal{P}_\theta[z]\} = \mathbb{I}\{z \in \mathcal{P}_\theta[x]\}$. If $K_M^{\mathsf{rp}} < \infty$, then K_M^{rp} is a positive definite (PD) kernel. When K_M^{rp} can evaluate to ∞, the definition of a PD kernel is not applicable since the positive definite property is defined only for to kernels taking finite values [9].

Let $\texttt{sgn} : \mathbb{R} \cup \{\pm\infty\} \to \{\pm 1\}$ be the sign function. For a WRP, define the weighted infinite-ensemble

$$\widehat{u}(x \| D^n, K_M^{\mathsf{rp}}) := \sum_{i=1}^n Y_i \cdot K_M^{\mathsf{rp}}(x, X_i) = \mathbb{E}_{\theta \sim \mathfrak{P}}[\alpha(\theta)\widehat{h}(x \| D^n, \mathcal{P}_\theta)]. \tag{4}$$

Note that the equality on the right follows immediately from linearity of the expectation and the definition of $\widehat{h}(\bullet \| D^n, \mathcal{P}_\theta)]$ in Equation (2).

Main problem. Find a WRP such that $\texttt{sgn}(\widehat{u}(\bullet \| D^n, K_M^{\mathsf{rp}}))$ has the interpolating-consistent property.

1.2 Outline of approach and contributions

In the regression setting, we have (X, Y) jointly distributed on $M \times \mathbb{R}$. Let $m(x) := \mathbb{E}[Y | X = x]$. Recall from Belkin et al. [7, Equation (7)] the definition of the *kernel smoothing estimator* with a so-called *singular*[1] kernel $K : M \times M \to [0, +\infty]$:

$$\widehat{m}(x \| D^n, K) := \begin{cases} Y_i & : \exists i \in [n] \text{ such that } x = X_i \\ \frac{\sum_{i=1}^n Y_i K(x, X_i)}{\sum_{j=1}^n K(x, X_j)} & : \sum_{j=1}^n K(x, X_j) > 0 \\ 0 & : \text{otherwise.} \end{cases} \tag{5}$$

[1] The "singular" modifier refers to the fact that $K(x, x) = +\infty$ for all $x \in M$.

We note that Equation (5) is referred as the *Nadaraya-Watson* estimate in [7]. Now, we simply write $\widehat{m}_n(x)$ instead of $\widehat{m}(x\|D^n, K)$ when there is no ambiguity. Similarly, we write $\widehat{u}_n(x)$ instead of $\widehat{u}(x\|D^n, K)$ from earlier. Note that $\mathtt{sgn}(\widehat{m}_n(x)) = \mathtt{sgn}(\widehat{u}_n(x))$ if $\sum_{j=1}^n K(x, X_j) > 0$.

Observe that $\widehat{m}_n$ is interpolating by construction. Let μ_X denote the marginal distribution of X. The L_1-*error* of $\widehat{m}_n$ in approximating m is $J_n := \int_M |\widehat{m}_n(x) - m(x)|\mu_X(dx)$. For $M = \mathbb{R}^d$ and the *Hilbert kernel* defined by $K^{\mathcal{H}}_{\mathbb{R}^d}(x, z) := \|x - z\|^{-d}$, Devroye et al. [19] proved L_1-consistency for *regression*: $J_n \to 0$ in probability when Y is bounded and X is continuously distributed.

Our contributions. Our primary contribution is to demonstrate an ensemble method with the consistent-interpolating property. Toward this end, in Section 3, we introduce the manifold-Hilbert kernel $K^{\mathcal{H}}_M$ on a Riemannian manifold M. When show that when M is complete, connected, and smooth, kernel smoothing regression with $K^{\mathcal{H}}_M$ has the same consistency guarantee (Theorem 3.2) as $K^{\mathcal{H}}_{\mathbb{R}^d}$ mentioned in the preceding paragraph. In Section 5, we consider the case when $M = \mathbb{S}^d$, and show that the manifold-Hilbert kernel $K^{\mathcal{H}}_{\mathbb{S}^d}$ is a weighted random partition kernel (Proposition 5.2).

Devroye et al. [19, Section 7] observed that the L_1-consistency of $\widehat{m}_n$ for regression implies the consistency for classification of $\mathtt{sgn} \circ \widehat{u}_n$. Furthermore, $\widehat{m}_n$ is interpolating for regression implies that $\mathtt{sgn} \circ \widehat{u}_n$ is interpolating for classification. These observations together with our results demonstrate the existence of a weighted infinite-ensemble classifier with the interpolating-consistent property.

1.3 Related work

Kernel regression. Kernel smoothing regression, or simply kernel regression, is an interpolator when the kernel used is singular, a fact known to Shepard [37] in 1968. Devroye et al. [19] showed that kernel regression with the Hilbert kernel is interpolating and weakly consistent for data with a density and bounded labels. Using singular kernels with compact support, Belkin et al. [7] showed that minimax optimality can be achieved under additional distributional assumptions.

Random forests. Wyner et al. [40] proposed that interpolation may be a key mechanism for the success of random forests and gave a compelling intuitive rationale. Belkin et al. [5] studied empirically the double descent phenomenon in random forests by considering the generalization performance past the interpolation threshold. The PERT variant of random forests, introduced by Cutler and Zhao [17], provably interpolates in 1-dimension. Belkin et al. [6] pose as an interesting question whether the result of Cutler and Zhao [17] extends to higher dimension. Many work have established consistency of random forest and its variants under different settings [13, 10, 36]. However, none of these work addressed interpolation.

Boosting. For classification under the noiseless setting (i.e., the Bayes error is zero), AdaBoost is interpolating and consistent (see Freund and Schapire [23, first paragraph of Chapter 12]). However, this setting is too restrictive and the result does not answer if consistency is possible when fitting the noise. Bartlett and Traskin [4] proved that AdaBoost with early stopping is universally consistent, however without the interpolation guarantee. To the best of our knowledge, whether AdaBoost or any other variant of boosting can be interpolating and consistent remains open.

Random partition kernels. Breiman [14] and Geurts et al. [25] studied infinite ensembles of simplified variants of random forest and connections to certain kernels. Davies and Ghahramani [18] formalized this connection and coined the term *random partition kernel*. Scornet [35] further developed the theory of random forest kernels and obtained upper bounds on the rate of convergence. However, it is not clear if these variants of random forests are interpolating.

Previously defined (unweighted) random partition kernels are bounded, and thus cannot be singular. On the other hand, the manifold-Hilbert kernel is always singular. To bridge between ensemble methods and theory on interpolating kernel smoothing regression, we propose *weighted* random partitions (Definition 1.1), whose associated kernel (Equation 3) can be singular.

Learning on Riemannian manifolds. Strong consistency of a kernel-based classification method on manifolds has been established by Loubes and Pelletier [29]. However, the result requires the kernel to be bounded and thus the method is not guaranteed to be interpolating. See Feragen and Hauberg [22] for a review of theoretical results regarding kernels on Riemannian manifolds.

Beyond kernel methods, other classical methods for Euclidean data have been extended to Riemannian manifolds, e.g., regression [38], classification [41], and dimensionality reduction and clustering [42][31]. To

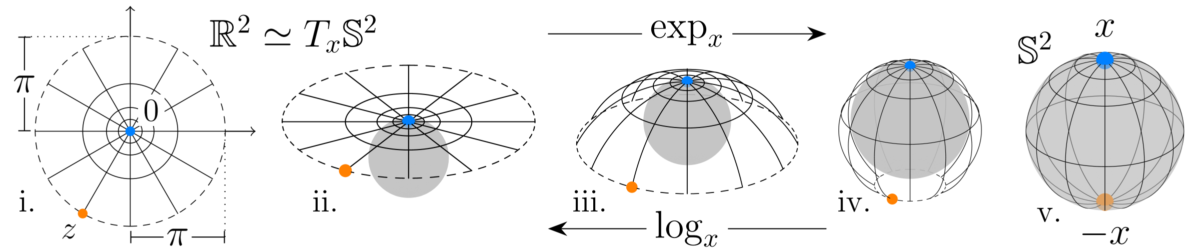

Figure 1: An illustration of the exponential map $\exp_x$ for the manifold $M = \mathbb{S}^2$, where x is the "northpole" (blue) and $-x$ the "southpole" (orange). The logarithm map $\log_x$, discussed in Section 4.1, is a right-inverse to $\exp_x$, i.e., $\exp_x \circ \log_x$ is the identity. *Panel i.* The tangent space $T_x\mathbb{S}^2$ visualized as $\mathbb{R}^2$. The dashed circle encloses a disc of radius π. *Panel ii.* The tangent space realized as the hyperplane tangent to sphere at x. *Panel iii-v.* Animation showing $\exp_x$ as a bijection from the open disc of radius π to $\mathbb{S}^2 \setminus \{-x\}$. The entire dashed circle in Panel i is mapped to $-x$ the southpole. Thus, $\log_x$ maps the southpole $-x$ to a point z on the dashed circle.

the best of our knowledge, no previous works have demonstrated an interpolating-consistent classifiers on manifolds other than $\mathbb{R}^d$.

In many applications, the data naturally belong to a Riemannian manifold. Spherical data arise from a range of disciplines in natural sciences. See the influential textbook by Mardia and Jupp [30, Ch.1§4]. For applications of the Grassmanian manifold in computer vision, see Jayasumana et al. [27] and the references therein. Topological data analysis [39] presents another interesting setting of manifold-valued data in the form of *persistence diagrams* [2, 28].

2 Background on Riemannian Manifolds

We give an intuitive overview of the necessary concepts and results on Riemannian manifolds. A longer, more precise version of this overview is in the Appendix Section A.1.

A smooth d-dimensional manifold M is a topological space that is locally diffeomorphic[2] to open subsets of $\mathbb{R}^d$. For simplicity, suppose that M is embedded in $\mathbb{R}^N$ for some $N \geq d$, e.g., $\mathbb{S}^d \subseteq \mathbb{R}^{d+1}$. Let $x \in M$ be a point. The *tangent space* at x, denoted T_xM, is the set of vectors that is tangent to M at x. Since linear combinations of tangent vectors are also tangent, the tangent space T_xM is a vector space. Tangent vectors can also be viewed as the time derivative of smooth curves. In particular, let $x \in M$. If $\epsilon > 0$ is an open set and $\gamma : (-\epsilon, \epsilon) \to M$ is a smooth curve such that $\gamma(0) = x$, then $\frac{d\gamma}{dt}(0) \in T_xM$.

A *Riemannian metric* on M is a choice of inner product $\langle \cdot, \cdot \rangle_x$ on T_xM for each x such that $\langle \cdot, \cdot \rangle_x$ varies smoothly with x. Naturally, $\|z\|_x := \sqrt{\langle z, z \rangle_x}$ defines a norm on T_xM. The length of a piecewise smooth curve $\gamma : [a, b] \to M$ is defined by $\mathtt{len}(\gamma) := \int_a^b \|\dot{\gamma}(t)\|_{\gamma(t)} dt$. Define $\mathtt{dist}_M(x, \xi) := \inf\{\mathtt{len}(\gamma) : \gamma$ is a piecewise smooth curve from x to $\xi\}$, which is a metric on M in the sense of metric spaces (see Sakai [34, Proposition 1.1]). For $x \in M$ and $r \in (0, \infty)$, the open metric ball centered at x of radius r is denoted $\mathtt{B}_x(r, M) := \{\xi \in M : \mathtt{dist}_M(x, \xi) < r\}$.

A curve $\gamma : [a, b] \to M$ is a geodesic if γ is *locally* distance minimizing and has constant speed, i.e., $\|\frac{d\gamma}{dt}(\tau)\|_{\gamma(\tau)}$ is constant. Now, suppose $x \in M$ and $v \in T_xM$ are such that there exists a geodesic $\gamma : [0, 1] \to M$ where $\gamma(0) = x$ and $\frac{d\gamma}{dt}(0) = v$. Define $\exp_x(v) := \gamma(1)$, the element reached by traveling along γ at time $= 1$. See Figure 1 for the case when $M = \mathbb{S}^2$.

For a fixed $x \in M$, the above function $\exp_x$, the *exponential map*, can be defined on an open subset of T_xM containing the origin. The Hopf-Rinow theorem ([20, Ch. 8, Theorem 2.8]) states that if M is connected and complete with respect to the metric $\mathtt{dist}_M$, then $\exp_x$ can be defined on all of T_xM.

[2]A diffeomorphism is a smooth bijection whose inverse is also smooth.

3 The Manifold-Hilbert kernel

Throughout the remainder of this work, we assume that M is a complete, connected, and smooth Riemannian manifold of dimension d.

Definition 3.1. We define the *manifold-Hilbert kernel* $K_M^{\mathcal{H}} : M \times M \to [0, \infty]$ for each $x, \xi \in M$ by $K_M^{\mathcal{H}}(x, \xi) := \mathtt{dist}_M(x, \xi)^{-d}$ if $x \neq \xi$ and $K_M^{\mathcal{H}}(x, x) := \infty$ otherwise.

Let λ_M be the *Riemann–Lebesgue volume measure* of M. Integration with respect to this measure is denoted $\int_M f d\lambda_M$ for a function $f : M \to \mathbb{R}$. For details of the construction of λ_M, see Amann and Escher [1, Proposition 1.5]. When $M = \mathbb{R}^d$, λ_M is the ordinary Lebesgue measure and $\int_{\mathbb{R}^d} f d\lambda_{\mathbb{R}^d}$ is the ordinary Lebesgue integral. For this case, we simply write λ instead of $\lambda_{\mathbb{R}^d}$.

We now state our first main result, a manifold theory extension of Devroye et al. [19, Theorem 1].

Theorem 3.2. *Suppose that X has a density f_X with respect to λ_M and that Y is bounded. Let $P_{Y|X}$ be a conditional distribution of Y given X and $m_{Y|X}$ be its conditional expectation. Let $\widehat{m}_n(x) := \widehat{m}(x \| D^n, K_M^{\mathcal{H}})$. Then*

1. *at almost all $x \in M$ with $f_X(x) > 0$, we have $\widehat{m}_n(x) \to m_{Y|X}(x)$ in probability,*
2. *$J_n := \int_M |\widehat{m}_n(x) - m_{Y|X}(x)| f_X(x) d\lambda_M(x) \to 0$ in probability.*

In words, the kernel smoothing regression estimate $\widehat{m}_n$ based on the manifold-Hilbert kernel is consistent and interpolates the training data, provided X has a density and Y is bounded. As a consequence, following the same logic as in Devroye et al. [19], the associated classifier $\mathtt{sgn} \circ \widehat{u}_n$ has the interpolating-consistent property. Before proving Theorem 3.2, we first review key concepts in probability theory on Riemannian manifolds.

3.1 Probability on Riemannian manifolds

Let $\mathcal{B}_M$ be the Borel σ-algebra of M, i.e., the smallest σ-algebra containing all open subsets of M. We recall the definition of M-valued random variables, following Pennec [32, Definition 2]:

Definition 3.3. Let $(\Omega, \mathbb{P}, \mathcal{A})$ be a probability space with measure $\mathbb{P}$ and σ-algebra $\mathcal{A}$. A *M-valued random variable* X is a Borel-measurable function $\Omega \to M$, i.e., $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}_M$.

Definition 3.4 (Density)**.** A random variable X taking values in M has a *density* if there exists a nonnegative Borel-measurable function $f : M \to [0, \infty]$ such that for all Borel sets B in M, we have $\Pr(X \in B) = \int_B f d\lambda_M$. The function f is said to be a *probability density function* (PDF) of X.

Next, we recall the definition of conditional distributions, following Dudley [21, Ch. 10 §2]:

Definition 3.5 (Conditional distribution[3])**.** Let (X, Y) be a random variable jointly distributed on $M \times \mathbb{R}$. Let $P_X(\cdot)$ be the probability measure corresponding to the marginal distribution of X. A *conditional distribution* for Y given X is a collection of probability measures $P_{Y|X}(\cdot|x)$ on $\mathbb{R}$ indexed by $x \in M$ satisfying the following:

1. For all Borel sets $A \subseteq \mathbb{R}$, the function $M \ni x \mapsto P_{Y|X}(A|x) \in [0, 1]$ is Borel-measurable.
2. For all $A \subseteq \mathbb{R}$ and $B \subseteq M$ Borel sets, $\Pr(Y \in A, X \in B) = \int_B P_{Y|X}(A|x) P_X(dx)$.

The *conditional expectation*[4] is defined as $m_{Y|X}(x) := \int_{\mathbb{R}} y P_{Y|X}(dy|x)$.

The existence of a conditional probability for a joint distribution (X, Y) is guaranteed by Dudley [21, Theorem 10.2.2]. When (X, Y) has a joint density f_{XY} and marginal density f_X, the above definition gives the classical formula $P_{Y|X}(A|x) = \int_A f_{XY}(x, y)/f_X(x) dy$ when $\infty > f_X(x) > 0$. See the first example in Dudley [21, Ch. 10 §2].

[3] also known as *disintegration measures* according to Chang and Pollard [16].

[4] More often, the conditional expectation is denoted $\mathbb{E}[Y|X = x]$. However, our notation is more convenient for function composition and compatible with that of [19].

3.2 Lebesgue points on manifolds

Devroye et al. [19] proved Theorem 3.2 when $M = \mathbb{R}^d$ and, moreover, that part 1 holds for the so-called *Lebesgue points*, whose definition we now recall.

Definition 3.6. Let $f : M \to \mathbb{R}$ be an absolutely integrable function and $x \in M$. We say that x is a *Lebesgue point* of f if $f(x) = \lim_{r\to 0} \frac{1}{\lambda_M(\mathtt{B}_x(r,M))} \int_{\mathtt{B}_x(r,M)} f d\lambda_M$.

For an integrable function, the following result states that almost all points are its Lebesgue points. For the proof, see Fukuoka [24, Remark 2.4].

Theorem 3.7 (Lebesgue differentation)**.** *Let $f : M \to \mathbb{R}$ be an absolutely integrable function. Then there exists a set $A \subseteq M$ such that $\lambda_M(A) = 0$ and every $x \in M \setminus A$ is a Lebesgue point of f.*

Next, for the reader's convenience, we restate Devroye et al. [19, Theorem 1], emphasizing the connection to Lebesgue points.

Theorem 3.8 (Devroye et al. [19])**.** *Let $M = \mathbb{R}^d$ be the flat Euclidean space. Then Theorem 3.2 holds. Moreover, Part 1 holds for all x that is a Lebesgue point to both f_X and $m_{Y|X} \cdot f_X$.*

The above result will be used in our proof of Theorem 3.2 below.

4 Proof of Theorem 3.2

The focal point of the first subsection is Lemma 4.1 which shows the Borel measurability of extensions of the so-called Riemannian logarithm. The second subsection contains two key results regarding densities of M-valued random variables transformed by the Riemannian logarithm. The final subsection proves Theorem 3.2 leveraging results from the preceding two subsections.

4.1 The Riemannian logarithm

Throughout, x is assumed to be an arbitrary point of M. Let $U_xM = \{v \in T_xM : \|v\|_x = 1\} \subseteq T_xM$ denote the set of unit tangent vectors. Define a function $\tau_x : U_xM \to (0,\infty]$ as follows[5]:

$$\tau_x(u) := \sup\{t > 0 : t = \mathtt{dist}_M(x, \exp_x(tu))\}.$$

The *tangent cut locus* is the set $\tilde{C}_x \subseteq T_xM$ defined by $\tilde{C}_x := \{\tau_x(u)u : u \in U_xM, \tau_x(u) < \infty\}$. Note that it is possible for $\tau_x(u) = \infty$ for all $u \in U_xM$ in which case $\tilde{C}_x$ is empty. The *cut locus* is the set $C_x := \exp_x(\tilde{C}_x) \subseteq M$.

The *tangent interior set* is $\tilde{I}_x := \{tu : 0 \le t < \tau_x(u), u \in U_xM\}$ and the *interior set* is the set $I_x := \exp_x(\tilde{I}_x)$. Finally, define $\tilde{D}_x := \tilde{I}_x \cup \tilde{C}_x$. Note that for each $z = tu \in \tilde{I}_x$, we have

$$\|z\|_x = t = \mathtt{dist}_M(x, \exp_x(tu)) = \mathtt{dist}_M(x, \exp_x(z)). \tag{6}$$

Consider the example where $M = \mathbb{S}^2$ as in Figure 1. Then $\tau_x(u) = \pi$ for all $u \in U_xM$. Thus, the tangent interior set $\tilde{I}_x = \mathtt{B}_0(\pi, \mathbb{R}^2)$, the open disc of radius π centered at the origin.

When restricted to $\tilde{I}_x$, the exponential map $\exp_x|_{\tilde{I}_x} : \tilde{I}_x \to I_x$ is a diffeomorphism. Its functional inverse, denoted by $\log_x|_{I_x}$, is called the *Riemannian Logarithm* [8, 43]. In previous works, $\log_x|_{I_x}$ is only defined from I_x to $\tilde{I}_x$. The next result shows that the domain of $\log_x|_{I_x} : I_x \to \tilde{I}_x$ can be extended to $\log_x : M \to \tilde{D}_x$ while remaining Borel-measurable.

Lemma 4.1. *For all $x \in M$, there exists a Borel measurable map $\log_x : M \to T_xM$ such that $\log_x(M) \subseteq \tilde{D}_x$ and $\exp_x \circ \log_x$ is the identity on M. Furthermore, for all $x, \xi \in M$, we have $\mathtt{dist}_M(x, \xi) = \|\log_x(\xi)\|_x$.*

[5] Positivity of τ_x is asserted at Sakai [34, eq. (4.1)]

Proof sketch. The full proof of the lemma is provided in Section A.2 of the Appendix. Below, we illustrate the idea of the proof using the example when $M = \mathbb{S}^2$ as in Figure 1.

Let $x \in \mathbb{S}^2$ be the "northpole" (the blue point). The tangent cut locus $\tilde{C}_x$ is the dashed circle in the left panel of Figure 1. The exponential map $\exp_x$ is one-to-one on $\tilde{D}_x$ except on the dashed circle, which all gets mapped to $-x$, the "southpole" (the orange point). A consequence of the measurable selection theorem[6] is that $\log_x$ can be extended to be a Borel-measurable right inverse of $\exp_x$ by selecting z point on $\tilde{C}_x$ such that $\log_x(-x) = z$. □

4.2 Random variable transforms

In the previous subsection, we showed that $\log_x : M \to T_xM$ is Borel-measurable. Now, recall that T_xM is equipped with the inner product $\langle \cdot, \cdot \rangle_x$, i.e., the Riemannian metric. Below, for each $x \in M$ choose an orthonormal basis on T_xM with respect to $\langle \cdot, \cdot \rangle$. Then T_xM is isomorphic as an inner product space to $\mathbb{R}^d$ with the usual dot product.

Our first result of this subsection is a "change-of-variables formula" for computing the densities of M-valued random variables after the $\log_x$ transform. Recall that λ_M is the Riemann-Lebesgue measure on M and λ is the ordinary Lebesgue measure on $\mathbb{R}^d = T_xM$.

Proposition 4.2. *Let $x \in M$ be fixed. There exists a Borel measurable function $\nu_x : M \to \mathbb{R}$ with the following properties:*

(i) Let X be a random variable on M with density f_X and let $Z := \log_x(X)$. Then Z is a random variable on T_xM with density $f_Z(z) := f_X(\exp_x(z)) \cdot \nu_x(\exp_x(z))$.

(ii) Let $f : M \to \mathbb{R}$ be an absolutely integrable function such that x is a Lebesgue point of f. Define $f : T_xM \to \mathbb{R}$ by $h(z) := f(\exp_x(z)) \cdot \nu_x(\exp_x(z))$. Then $0 \in T_xM$ is a Lebesgue point for h.

Proof sketch. The full proof of the proposition is in Appendix Section A.3. The function ν_x is the Jacobian of the change-of-variables formula for integrating $\int_{\tilde{B}} f_Z d\lambda$ where $\tilde{B} \subseteq T_xM$ is a Borel subset. See Appendix Lemma A.4 for the exact definition of ν_x. Part (i) is a simple consequence of this change-of-variables formula, which says that $\int_{\tilde{B}} f_Z d\lambda = \int_{\exp_x(\tilde{B})} h d\lambda_M$.

For part (ii), the key observations are that (a) $\nu_x(\exp_x(0)) = \nu_x(0) = 1$ and (b) the volumes of $\mathtt{B}_x(r, M)$ and $\mathtt{B}_0(r, T_xM)$ are equal as $r \to 0$. More precisely, $\lim_{r\to 0} \frac{\lambda_M(\mathtt{B}_x(r,M))}{\lambda(\mathtt{B}_0(r,T_xM))} = 1$. From these two observations, it is straightforward to directly verify Definition 3.6. □

Proposition 4.3. *Let (X, Y) have a joint distribution on $M \times \mathbb{R}$ such that the marginal of X has a density f_X on M. Let $P_{Y|X}(\cdot|\cdot)$ be a conditional distribution for Y given X. Let $x \in M$. Define $Z := \log_x(X)$ and consider the joint distribution (Z, Y) on $T_pM \times \mathbb{R}$. Then $P_{Y|Z}(\cdot|\cdot) := P_{Y|X}(\cdot|\exp_x(\cdot))$ is a conditional distribution for Y given Z. Consequently, $m_{Y|X} \circ \exp_x = m_{Y|Z}$.*

Proof sketch. The full proof of the Proposition is in Appendix Section A.4. The idea is the same as in the proof of Proposition 4.2, except that the probability density f_Z is replaced by an appropriate conditional probability density. □

4.3 Finishing up the Proof of Theorem 3.2

Fix $x \in M$ such that x is a Lebesgue point of f_X and $m_{Y|X} \cdot f_X$. Note that by Theorem 3.7, almost all $x \in M$ has this property. Next, let $Z = \log_x(X)$ and f_Z be as in Proposition 4.2-(i). Then

1. $f_Z = (f_X \circ \exp_x) \cdot (\nu_x \circ \exp_x)$, and
2. $(m_{Y|X} \circ \exp_x) \cdot f_Z = (m_{Y|X} \circ \exp_x) \cdot (f_X \circ \exp_x) \cdot (\nu_x \circ \exp_x)$.

Now, proposition 4.2-(ii) implies that 0 is a Lebesgue point of both f_Z and $(m_{Y|X} \circ \exp_x) \cdot f_Z$. Furthermore, by Proposition 4.3, we have $m_{Y|X} \circ \exp_x = m_{Y|Z}$. Thus, 0 is a Lebesgue point of f_Z and $m_{Y|Z} \cdot f_Z$.

[6]Kuratowski–Ryll-Nardzewski measurable selection theorem (see [12, Theorem 6.9.3])

Now, let $D_n := \{(X_i, Y_i)\}_{i\in[n]}$. Define $Z_i := \log_x(X_i)$, which are i.i.d copies of the random variable $Z := \log_x(X)$, and let $\tilde{D}_n := \{(Z_i, Y_i)\}_{i\in[n]}$. Then we have

$$\begin{aligned}\widehat{m}(x\|D^n, K_M^{\mathcal{H}}) &\overset{\text{(a)}}{=} \frac{\sum_{i=1}^n Y_i \cdot \mathtt{dist}_M(x, X_i)^{-d}}{\sum_{j=1}^n \mathtt{dist}_M(x, X_j)^{-d}} \overset{\text{(b)}}{=} \frac{\sum_{i=1}^n Y_i \cdot \|Z_i\|_x^{-d}}{\sum_{j=1}^n \|Z_j\|_x^{-d}} \\ &\overset{\text{(c)}}{=} \frac{\sum_{i=1}^n Y_i \cdot \mathtt{dist}_{\mathbb{R}^d}(0, Z_i)^{-d}}{\sum_{j=1}^n \mathtt{dist}_{\mathbb{R}^d}(0, Z_j)^{-d}} \overset{\text{(d)}}{=} \widehat{m}(0\|\tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}})\end{aligned}$$

where equations marked by (a) and (d) follow from Equation (5), (b) from Lemma 4.1, and (c) from the fact that the inner product space T_xM with $\langle\cdot,\cdot\rangle_x$ is isomorphic to $\mathbb{R}^d$ with the usual dot product. By Theorem 3.8, we have $\widehat{m}(0\|\tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}}) \to m_{Y|Z}(0)$ in probability. In other words, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\{|\widehat{m}(0\|\tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}}) - m_{Y|Z}(0)| > \epsilon\} = 0.$$

By Proposition 4.3, we have $m_{Y|Z}(0) = m_{Y|Z}(\exp_x(0)) = m_{Y|Z}(x)$. Therefore,

$$\left\{|\widehat{m}(0\|\tilde{D}^n, K_{\mathbb{R}^d}^{\mathcal{H}}) - m_{Y|Z}(0)| > \epsilon\right\} = \left\{|\widehat{m}(x\|D^n, K_M^{\mathcal{H}}) - m_{Y|X}(x)| > \epsilon\right\}$$

as events. Thus, $\widehat{m}(x\|D^n, K_M^{\mathcal{H}}) \to m_{Y|X}(x)$ converges in probability, proving Theorem 3.2 part 1. As noted in Devroye et al. [19, §2], part 2 of Theorem 3.2 is an immediate consequence of part 1.

5 Application to the d-Sphere

The d-dimensional round sphere is $\mathbb{S}^d := \{x \in \mathbb{R}^{d+1} : x_1^2 + \cdots + x_{d+1}^2 = 1\}$. Here, a *round* sphere assumes that $\mathbb{S}^d$ has the *arc-length metric*:

$$\mathtt{dist}_{\mathbb{S}^d}(x, z) = \angle(x, z) = \cos^{-1}(x^\top z) \in [0, \pi]. \tag{7}$$

Let $\mathcal{S}$ be a set and $\sigma : M \to \mathcal{S}$ be a function. The *partition induced* by σ is defined by $\{\sigma^{-1}(s) : s \in \mathtt{Range}(\sigma)\}$. For example, when $M = \mathbb{S}^d$ and $W \in \mathbb{R}^{(d+1)\times h}$, then the function $\sigma_W : \mathbb{S}^d \to \{\pm 1\}^h$ defined by $\sigma_W(x) = \mathtt{sgn}(W^\top x)$ induces a hyperplane arrangement partition.

Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denote the positive and non-negative integers.

Definition 5.1 (Random hyperplane arrangement partition)**.** Let $d \in \mathbb{N}$ and $M = \mathbb{S}^d$. Let $q < 0$ be a negative number, and let H be a random variable with probability mass function $p_H : \mathbb{N}_0 \to [0, 1]$ such that $p_H(h) > 0$ for all h. Define the following weighted random partition $\mathsf{ha} := (\Theta, \mathfrak{P}, \alpha)$:

1. The parameter space $\Theta = \bigsqcup_{h=0}^\infty \mathbb{R}^{(d+1)\times h}$ is the disjoint union of all $(d+1) \times h$ matrices. Element of Θ are matrices $\theta = W \in \mathbb{R}^{(d+1)\times h}$ where the number of columns $h \in \{0, 1, 2, \dots\}$ varies. By convention, if $h = 0$, the partition $\mathcal{P}_\theta = \mathcal{P}_W$ is the trivial partition $\{\mathbb{S}^d\}$. If $h > 0$, $\mathcal{P}_W$ is the partition induced by $x \mapsto \mathtt{sgn}(W^\top x)$.
2. The probability $\mathfrak{P}$ is constructed by the procedure where we first sample $h \sim p_H(h)$, then sample the entries of $W \in \mathbb{R}^{d\times h}$ i.i.d according to $\mathtt{Gaussian}(0, 1)$.
3. For $\theta \in \Theta$, define $\alpha(\theta) := \pi^q p_H(h)^{-1}(-1)^h \binom{q}{h}$, where $\binom{q}{h} := \frac{1}{h!}\prod_{j=0}^{h-1}(q - j)$.

Note that $(-1)^h\binom{q}{h} = \frac{1}{h!}\prod_{j=0}^{h-1}(q - j) > 0$ when $q < 0$.

Theorem 5.2. *Let* $\mathsf{ha} = (\Theta, \mathfrak{P}, \alpha)$ *be as in Definition 5.1. Then*

$$K_{\mathbb{S}^d}^{\mathsf{ha}}(x, z) = \begin{cases} \angle(x, z)^q & : \angle(x, z) \neq 0 \\ +\infty & : \textit{otherwise.} \end{cases}$$

When $q = -d$*, we have* $K_{\mathbb{S}^d}^{\mathsf{ha}} = K_{\mathbb{S}^d}^{\mathcal{H}}$ *where the right hand side is the manifold-Hilbert kernel.*

Proof of Theorem 5.2. Before proceeding, we have the following useful lemma:

Lemma 5.3. *Let* $\mathsf{rp} = (\Theta, \mathfrak{P}, \alpha)$ *be a WRP. Let* H *be a random variable. Let* $\theta \sim \mathfrak{P}$. *Suppose that for all* $x, z \in M$, *the random variables* $\alpha(\theta)$ *and* $\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}$ *are conditionally independent given* H. *Then we have* $K_M^{\mathsf{rp}}(x,z) = \mathbb{E}_H\Big[\overline{\alpha}(H) \cdot \mathbb{E}_{\theta\sim\mathfrak{P}}[\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}|H]\Big]$ *where* $\overline{\alpha}(h) := \mathbb{E}_{\theta\in\mathfrak{P}}[\alpha(\theta)|H = h]$ *for a realization* h *of* H.

The lemma follows immediately from the Definition of $K_M^{\mathsf{rp}}(x,z)$ in Equation 3 and the conditional independence assumption. Now, we proceed with the proof of Theorem 5.2.

Let $\phi := \angle(x,z)/\pi$. Let $H \sim p_H$ and $\theta \sim \mathfrak{P}$ be the random variables in Definition 5.1. Note that by construction, the following condition is satisfied: for all $x, z \in M$, the random variables $\alpha(\theta)$ and $\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}$ are conditionally independent given H. In fact, $\alpha(\theta) = \pi^q p_H(h)^{-1}(-1)^h\binom{q}{h}$ is constant given $H = h$. Hence, applying Lemma 5.3, we have

$$\begin{aligned}
K_{\mathbb{S}^d}^{\mathsf{ha}}(x,z) &= \mathbb{E}_H\Big[\overline{\alpha}(H) \cdot \mathbb{E}_{\theta\sim\mathfrak{P}}[\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}|H]\Big] \\
&= \sum_{h=0}^{\infty} \pi^q(-1)^h\binom{q}{h} \cdot \mathbb{E}_{\theta\sim\mathfrak{P}}[\mathbb{I}\{x \in \mathcal{P}_\theta[z]\}|H = h] = \sum_{h=0}^{\infty} \pi^q(-1)^h\binom{q}{h} \cdot \Pr\{x \in \mathcal{P}_\theta[z]|H = h\}.
\end{aligned}$$

Next, we claim that $\Pr\{x \in \mathcal{P}_\theta[z]|H = h\} = (1-\phi)^h$. When $h = 0$, $x \in \mathcal{P}_\theta[z]$ is always true since $\mathcal{P}_\theta = \{\mathbb{S}^d\}$ is the trivial partition. In this case, we have $\Pr\{x \in \mathcal{P}_\theta[z]|H = h\} = 1 = (1-\phi)^0$. When $h > 0$, we recall a result of Pinelis [33]:

Lemma 5.4. *Let* $x, z \in \mathbb{S}^d$. *Let* $w \in \mathbb{R}^{d+1}$ *be a random vector whose entries are sampled i.i.d according to* $\mathtt{Gaussian}(0,1)$. *Then* $\Pr\{\mathtt{sgn}(w^\top x) = \mathtt{sgn}(w^\top z)\} = 1 - (\angle(x,z)/\pi)$.

Let $W = [w_1, \ldots, w_h]$ be as in Definition 5.1 where w_j denotes the j-th column of W. Then by construction, w_j is distributed identically as w in Lemma 5.4. Furthermore, w_j and $w_{j'}$ are independent for $j, j' \in [h]$ where $j \neq j'$. Thus, the claim follows from

$$\begin{aligned}
\Pr\{x \in \mathcal{P}_\theta[z]|H = h\} &\overset{\text{(a)}}{=} \Pr\{\mathtt{sgn}(W^\top x) = \mathtt{sgn}(W^\top z)|H = h\} \\
&\overset{\text{(b)}}{=} \prod_{j=1}^{h} \Pr\{\mathtt{sgn}(w_j^\top x) = \mathtt{sgn}(w_j^\top z)\} \overset{\text{(c)}}{=} \prod_{j=1}^{h} (1-\phi) = (1-\phi)^h.
\end{aligned}$$

where equality (a) follows from Definition 5.1, (b) from $W \in \mathbb{R}^{(d+1)\times h}$ having i.i.d standard Gaussian entries given $H = h$, and (c) from Lemma 5.4. Putting it all together, we have

$$K_{\mathfrak{P},\alpha}^{\mathsf{part}}(x,z) = \sum_{h=0}^{\infty} \pi^q(-1)^h\binom{q}{h}(1-\phi)^h = \pi^q\sum_{h=0}^{\infty}\binom{q}{h}(\phi-1)^h = \angle(x,z)^q.$$

For the last step, we used the fact that for all $q \in \mathbb{R}$ the binomial series $(1+t)^q = \sum_{h=0}^{\infty}\binom{q}{h}t^h$ converges absolutely for $|t| < 1$ (when $\phi \in (0,1]$) and diverges to $+\infty$ for $t = -1$ (when $\phi = 0$). ☐

Corollary 5.5. *Let* $q := -d$ *and* $K_{\mathbb{S}^d}^{\mathsf{ha}}$ *be as in Theorem 5.2. The infinite-ensemble classifier* $\mathtt{sgn}(\widehat{u}(\bullet\|D^n, K_{\mathbb{S}^d}^{\mathsf{ha}}))$ *(see Equation 4 for definition) has the interpolating-consistent property.*

Proof. As observed in Devroye et al. [19, Section 7], for an arbitrary kernel K, the L_1-consistency of $\widehat{m}(\bullet\|D^n, K)$ for regression implies the consistency for classification of $\mathtt{sgn}(\widehat{u}(\bullet\|D^n, K))$. Furthermore, $\widehat{m}(\bullet\|D^n, K)$ is interpolating for regression implies that $\mathtt{sgn}(\widehat{u}(\bullet\|D^n, K))$ is interpolating for classification. While the argument there is presented in the $\mathbb{R}^d$ case, the argument holds in the more general manifold case *mutatis mutandis*.

Thus, by Theorem 3.2, we have $\mathtt{sgn}(\widehat{u}(\bullet\|D^n, K_{\mathbb{S}^d}^{\mathcal{H}}))$ is consistent for classification, i.e., Equation (1) holds. It is also interpolating since $\widehat{m}(\bullet\|D^n, K)$ is interpolating. By Proposition 5.2, we have $K_{\mathbb{S}^d}^{\mathsf{ha}} = K_{\mathbb{S}^d}^{\mathcal{H}}$. Thus $\mathtt{sgn}(\widehat{u}(\bullet\|D^n, K_{\mathbb{S}^d}^{\mathsf{ha}}))$ is an ensemble method having the interpolating-consistent property. ☐

6 Discussion

We have shown that using the manifold-Hilbert kernel in kernel smoothing regression, also known as Nadaraya-Watson regression, results in a consistent estimator that interpolates the training data on a Riemannian manifold M. Furthermore, when $M = \mathbb{S}^d$ is the sphere, we showed that the manifold-Hilbert kernel is a weighted random partition kernel, where the random partitions are induced by random hyperplane arrangements. This demonstrates an ensemble method that has the interpolating-consistent property.

A limitation of this work is that the random hyperplane arrangement partition is data-independent. Thus, the resulting ensemble method considered in this work are easier to analyze than popular ensemble methods used in practice. Nevertheless, we believe our work offers one theoretical basis towards understanding generalization in the interpolation regime of ensembles of histogram classifiers over data-dependent partitions, e.g., decision trees à la CART [15].

Acknowledgements

The authors were supported in part by the National Science Foundation under awards 1838179 and 2008074.

References

[1] Herbert Amann and Joachim Escher. *Analysis III.* Springer, 2009, pp. 389–455.

[2] Rushil Anirudh, Vinay Venkataraman, Karthikeyan Natesan Ramamurthy, and Pavan Turaga. "A Riemannian framework for statistical analysis of topological persistence diagrams". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.* 2016, pp. 68–76.

[3] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070.

[4] Peter L Bartlett and Mikhail Traskin. "AdaBoost is consistent". In: *Journal of Machine Learning Research* 8.Oct (2007), pp. 2347–2368.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[6] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 2300–2311.

[7] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. "Does data interpolation contradict statistical optimality?" In: *The 22nd International Conference on Artificial Intelligence and Statistics.* PMLR. 2019, pp. 1611–1619.

[8] Thomas Bendokat, Ralf Zimmermann, and P-A Absil. "A Grassmann manifold handbook: Basic geometry and computational aspects". In: *arXiv preprint arXiv:2011.13699* (2020).

[9] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Springer Science & Business Media, 2011.

[10] Gérard Biau, Luc Devroye, and Gäbor Lugosi. "Consistency of random forests and other averaging classifiers." In: *Journal of Machine Learning Research* 9.9 (2008).

[11] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* 25.2 (2016), pp. 197–227.

[12] Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory.* Vol. 1. Springer, 2007.

[13] Leo Breiman. "Consistency for a simple model of random forests". In: (2004).

[14] Leo Breiman. *Some infinity theory for predictor ensembles.* Technical Report. Department of Statistics, 2000.

[15] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press, 1984.

[16] Joseph T Chang and David Pollard. "Conditioning as disintegration". In: *Statistica Neerlandica* 51.3 (1997), pp. 287–317.

[17] Adele Cutler and Guohua Zhao. "PERT-perfect random tree ensembles". In: *Computing Science and Statistics* 33 (2001), pp. 490–497.

[18] Alex Davies and Zoubin Ghahramani. "The random forest kernel and other kernels for big data from random partitions". In: *arXiv preprint arXiv:1402.4293* (2014).

[19] Luc Devroye, László Györfi, and Adam Krzyżak. "The Hilbert kernel regression estimate". In: *Journal of Multivariate Analysis* 65.2 (1998), pp. 209–227.

[20] Manfredo Perdigao Do Carmo. *Riemannian Geometry*. Vol. 6. Springer, 1992.

[21] Richard M Dudley. *Real Analysis and Probability*. CRC Press, 2018.

[22] Aasa Feragen and Søren Hauberg. "Open Problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive definite geodesic exponential kernel?" In: *Conference on Learning Theory*. PMLR. 2016, pp. 1647–1650.

[23] Yoav Freund and Robert E Schapire. "Boosting: Foundations and Algorithms". In: *MIT Press* 1.6 (2012), p. 7.

[24] Ryuichi Fukuoka. "Mollifier smoothing of tensor fields on differentiable manifolds and applications to Riemannian Geometry". In: *arXiv preprint math/0608230* (2006).

[25] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine learning* 63.1 (2006), pp. 3–42.

[26] James J Hebda. "Parallel translation of curvature along geodesics". In: *Transactions of the American Mathematical Society* 299.2 (1987), pp. 559–572.

[27] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. "Kernel methods on Riemannian manifolds with Gaussian RBF kernels". In: *IEEE transactions on pattern analysis and machine intelligence* 37.12 (2015), pp. 2464–2477.

[28] Tam Le and Makoto Yamada. "Persistence Fisher kernel: A Riemannian manifold kernel for persistence diagrams". In: *Advances in Neural Information Processing Systems* 31 (2018).

[29] Jean-Michel Loubes and Bruno Pelletier. "A kernel-based classifier on a Riemannian manifold". In: *Statistics & Decisions* 26.1 (2008), pp. 35–51.

[30] Kanti V Mardia and Peter E Jupp. *Directional statistics*. Vol. 2. Wiley Online Library, 2000.

[31] Kanti V Mardia, Henrik Wiechers, Benjamin Eltzner, and Stephan F Huckemann. "Principal component analysis and clustering on manifolds". In: *Journal of Multivariate Analysis* 188 (2022), p. 104862.

[32] Xavier Pennec. "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements". In: *Journal of Mathematical Imaging and Vision* 25.1 (2006), pp. 127–154.

[33] Iosif Pinelis. *Probability of two points being divided by an high-dimensional hyperplane*. MathOverflow. URL:https://mathoverflow.net/q/323697 (version: 2019-02-21). 2019.

[34] Takashi Sakai. *Riemannian Geometry*. Vol. 149. American Mathematical Society, 1996.

[35] Erwan Scornet. "Random forests and kernel methods". In: *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1485–1500.

[36] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. "Consistency of random forests". In: *The Annals of Statistics* 43.4 (2015), pp. 1716–1741.

[37] Donald Shepard. "A two-dimensional interpolation function for irregularly-spaced data". In: *Proceedings of the 1968 23rd ACM national conference*. 1968, pp. 517–524.

[38] P Thomas Fletcher. "Geodesic regression and the theory of least squares on Riemannian manifolds". In: *International journal of computer vision* 105.2 (2013), pp. 171–185.

[39] Larry Wasserman. "Topological data analysis". In: *Annual Review of Statistics and Its Application* 5 (2018), pp. 501–532.

[40] Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. "Explaining the success of AdaBoost and random forests as interpolating classifiers". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1558–1590.

[41] Zhigang Yao and Zhenyue Zhang. "Principal boundary on Riemannian manifolds". In: *Journal of the American Statistical Association* 115.531 (2020), pp. 1435–1448.

[42] Zhenyue Zhang and Hongyuan Zha. "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment". In: *SIAM journal on scientific computing* 26.1 (2004), pp. 313–338.

[43] Ralf Zimmermann. "A matrix-algebraic algorithm for the Riemannian logarithm on the Stiefel manifold under the canonical metric". In: *SIAM Journal on Matrix Analysis and Applications* 38.2 (2017), pp. 322–342.

Appendix: Consistent Interpolating Ensembles via the Manifold-Hilbert Kernel

A.1 Basics of Riemannian Manifolds

In this section, we review the main concepts from Riemannian manifold theory essential to this work. Our main references are Sakai [34] and Do Carmo [20]. Throughout, $d \in \mathbb{N}$ denotes the dimension. We use the word *smooth* to mean infinitely differentiable.

Manifolds. A smooth *manifold* M of dimension d is a Hausdorff, second countable topological space together with an *atlas*: a set $\texttt{Atlas} := \{(U_\alpha, \varphi_\alpha)\}_{\alpha \in A}$ where 1). $\{U_\alpha\}_{\alpha \in A}$ is an open cover of M, 2). for each $\alpha \in A$, $\varphi_\alpha : U_\alpha \to \varphi_\alpha(U_\alpha) \subseteq \mathbb{R}^d$ is a homeomorphism onto its image, and 3). $\varphi_\alpha \circ \varphi_\beta^{-1} : \varphi_\beta(U_\alpha \cap U_\beta) \to \varphi_\alpha^{-1}(U_\alpha \cap U_\beta)$ is smooth for each pair $\alpha, \beta \in A$. An element (U, φ) of $\texttt{Atlas}$ is called a *chart*.

Smooth maps. A real-valued function $f : M \to \mathbb{R}$ is a *smooth function* if $f \circ \varphi^{-1}$ is smooth (in the elementary calculus sense) for all charts (U, φ). The set of all smooth functions is denoted $\texttt{Fn}(M)$, which forms an $\mathbb{R}$-vectorspace. Let N be another smooth manifold with atlas $\mathcal{B}$. A function $\Phi : M \to N$ is a *smooth map* if $g \circ \Phi \in \texttt{Fn}(M)$ for all $g \in \texttt{Fn}(N)$.

Tangent space. Let $x \in M$. A *derivation at* x is a linear function $v : \texttt{Fn}(M) \to \mathbb{R}$ satisfying the *product rule*: $v[fg] = f(x)v[f] + g(x)v[g]$ for all $f, g \in \texttt{Fn}(M)$. The *tangent space at* x, denoted T_xM, is the vector space of all derivations at x. Elements of T_xM are referred to as *tangent vectors* at x. For a given chart (U, φ) where $x \in U$, define a derivation at x, denoted $\partial_i|_x$, by $f \mapsto \frac{d(f \circ \varphi^{-1})}{dz_i}(\varphi(x))$ where $\frac{d}{dz_i}$ is the i-th partial derivative in ordinary calculus. It is a fact that $\{\partial_i|_x : i = 1, \ldots, d\}$ is a basis for T_xM.

Although the above definition of a tangent vector is abstract, it can be concretely interpreted in terms of derivative along a curve. Let $a < t_0 < b$ be real numbers. A *curve through* x is a smooth map $\gamma : (a, b) \to M$ such that $\gamma(t_0) = x$. Then $\texttt{Fn}(M) \ni f \mapsto \frac{d}{dt} f(\gamma(t))|_{t=t_0} \in \mathbb{R}$ defines a derivation at x. Oftentimes, this derivation is denoted $\dot{\gamma}(t_0) \in T_xM$

Riemannian metric. The *tangent bundle* is the set $TM := \bigcup_x T_xM$, which itself is a smooth manifold of dimension $2d$. A *vector field on* M is a smooth map $\mathsf{V} : M \to TM$ such that $\mathsf{V}(x) \in T_xM$ for all $x \in M$. The set of all vectors fields on M is denoted $\texttt{Vf}(M)$.

A *Riemannian metric* on M is a choice of an inner product $\langle \cdot, \cdot \rangle_x$ (and thus, a norm $\| \cdot \|_x$) on T_xM for each $x \in M$ such that the function $M \to \mathbb{R}$ given by $x \mapsto \langle \mathsf{V}(x), \mathsf{U}(x) \rangle_x$ is smooth for all $\mathsf{V}, \mathsf{U} \in \texttt{Vf}(M)$. As shorthands, when x is clear from context, we drop the subscripts and simply write $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ instead. Choosing an orthonormal basis for T_xM with respect to $\langle \cdot, \cdot \rangle_x$ for each x, we can identify T_xM with $\mathbb{R}^d$ with the ordinary dot inner product.

Let $x \in M$ and (U, φ) be a chart such that $x \in U$. Define $g_{ij}(x) = \langle \partial_i|_x, \partial_j|_x \rangle_x$. Denote by $G(x)$ the $d \times d$ positive definite matrix $[g_{ij}(x)]_{ij}$. Below, we will refer to the function $G : U \to \mathbb{R}^{d \times d}$ as the *coordinate representation* of the Riemannian metric. Define $g^{ij}(x) := [G(x)^{-1}]_{ij}$. The *Christoffel symbols* with respect to (U, φ) are defined by $\Gamma_{ij}^k := \frac{1}{2} \sum_{\ell=1}^d g^{k\ell}(\partial_i|_x g_{j\ell} + \partial_j|_x g_{i\ell} - \partial_\ell|_x g_{ij})$. Note that $g_{k\ell}$, $g^{k\ell}$, G, Γ_{ij}^k, and $\partial_i|_x g_{j\ell}$ are all functions with domain U.

Geodesics. Fix a chart (U, φ). Consider a smooth curve $\gamma : [a, b] \to U$. Let $\zeta_i(t) := [\varphi(\gamma(t))]_i$ be the i-th component functions. The curve γ is a *geodesic* if ζ is a solution to the following system of second order ordinary differential equations (ODEs): $\frac{d^2\zeta_i}{dt^2} + \sum_{j,\ell=1}^d \Gamma_{j\ell}^i \circ \gamma \frac{d\zeta_j}{dt} \frac{d\zeta_\ell}{dt} = 0$ for all $i = 1, \ldots, d$ at all time $t \in [a, b]$.

Geodesics are minimizers of the so-called *energy functional* $E(\gamma) = \frac{1}{2} \int_a^b \|\dot{\gamma}(t)\|_{\gamma(t)}^2 dt$. The above system of ODEs are the analog of the "first derivative test" for local minimizers of E. Thus, geodesics are defined independently of the choice of the chart.

Exponential map. For $x \in M$ and $v \in T_xM$, there exists $\epsilon > 0$ and a unique geodesic curve $\gamma_v : [-\epsilon, \epsilon] \to M$ such that $\gamma_v(0) = x$ and $\dot{\gamma}_v(0) = v$. This follows from the existence and uniqueness of the solution to an ODE given initial conditions where the ODE is as discussed above. Note that although geodesics are previously defined in U where (U, φ) is a chart, they can be extended outside of U using additional charts.

Let $x \in M$ and $v \in T_xM$ be fixed and let $\gamma_v : [-\epsilon, \epsilon] \to M$ be as in the preceding paragraph. If $\|v\|_x \leq \epsilon$, then define $\exp_x(v) := \gamma_v(1)$. A fundamental fact is that $\exp_x$, known as the *exponential map at* x, can be defined on an open set of T_xM containing the origin.

Distance function. Let $x, \xi \in M$ and $a < b$ be real numbers. A *piecewise smooth curve* from x to ξ is a piecewise smooth map $\gamma : [a, b] \to M$ such that $\gamma(a) = x$ and $\gamma(b) = \xi$. Assume that M is connected. Then for all $x, \xi \in M$, there exists a piecewise smooth curve from x to ξ. The *length of* γ is defined as $\mathtt{len}(\gamma) := \int_a^b \|\dot\gamma(t)\|_{\gamma(t)} dt$. Define $\mathtt{dist}_M(x, \xi) := \inf\{\mathtt{len}(\gamma) : \gamma \text{ is a piecewise smooth curve from } x \text{ to } \xi\}$, which is a metric on M in the sense of metric spaces (see [34, Proposition 1.1]). For $x \in M$ and $r \in (0, \infty)$, the open ball centered at x of radius r is denoted $\mathtt{B}_x(r, M) := \{z \in M : \mathtt{dist}_M(x, z) < r\}$.

Complete Riemannian manifolds. A Riemannian manifold is *complete* if it is a complete metric space under the metric $\mathtt{dist}_M$. The Hopf-Rinow theorem ([20, Ch. 8, Theorem 2.8]) states that if M is connected and complete, then the exponential $\exp_x$ can be defined on the entire T_xM.

A.2 Proof of Lemma 4.1

This section uses definitions and notations introduced in Section 4.1. In particular, recall the cut locus C_x, the tangent cut locus $\tilde C_x$, the interior set I_x and the tangent interior set $\tilde I_x$. The proof of Lemma 4.1 is presented towards the end of the section. At this point, we compile some facts from various sources about the cut locus.

Lemma A.1. *For all $x \in M$, we have*

1. *C_x is a closed subset of M (Hebda [26, Proposition 1.2]).*
2. *$I_x \cap C_x = \emptyset$ and $I_x \cup C_x = M$ (Sakai [34, Ch II, Lemma 4.4 (1)])*
3. *I_x is an open subset of M (immediate from 1 and 2 above)*
4. *$\exp_x : \tilde I_x \to I_x$ is a diffeomorphism ([34, Ch II, Lemma 4.4 (2)])*
5. *$\lambda_M(C_x) = 0$, where λ_M is the Riemann-Lebesgue measure ([34, Lemma 4.4 (3)])*
6. *τ_x is continuous and $\inf_{u \in U_xM} \tau_x(u) > 0$ ([34, Ch II, Propositions 4.1 (2) and 4.13 (1)])*

While the following lemma is elementary, we provide a proof since we could not find one in the literature.

Lemma A.2. *For all $x \in M$, the (topological) closure of $\tilde I_x$ in T_xM is $\tilde D_x$. Furthermore, for all $x \in M$, we have $\exp_x(\tilde D_x) = M$.*

Proof of Lemma A.2. Take a convergent sequence $\{t_iu_i\}_{i\in\mathbb{N}} \subseteq \tilde I_x$ where $u_i \in U_xM$ and $0 \le t_i < \tau_x(u_i)$. Let $v^* = \lim_i t_iu_i$. Our goal is to show that $v^* \in \tilde D_x = \tilde I_x \cup \tilde C_x$.

Since U_xM is compact, we may assume that $u^* := \lim_i u_i$ exists after passing to a subsequence if necessary. Furthermore, $\|t_iu_i\|_x = t_i$ implies that $t^* := \lim_i t_i$ exists as well (i.e., $t^* < \infty$). Hence, $v^* = t^*u^*$.

Consider the case that $\tau_x(u^*) = \infty$. Then $0 \le t^* < \tau_x(u^*)$ implies that $v^* = t^*u^* \in \tilde I_x$. For the other case that $t(u) < \infty$, we first note that $t_iu_i \in \tilde I_x$ implies that $t_i < \tau_x(u_i)$. Taking the limit of both sides, we have $t^* = \lim_i t_i \le \lim_i \tau_x(u_i) = \tau_x(u^*)$. Note that the last limit can be exchanged since τ_x is continuous (Lemma A.1 part 6). Thus, either $t^* < \tau_x(u^*)$ in which case $v^* \in \tilde I_x$, or $t^* = \tau_x(u^*)$ in which case $v^* = \tau_x(u^*)u^* \in \tilde C_x$.

For the "furthermore" part, note that

$$\exp_x(\tilde D_x) = \exp_x(\tilde I_x \cup \tilde C_x) = \exp_x(\tilde I_x) \cup \exp_x(\tilde C_x) = I_x \cup C_x = M$$

where the last equality is Lemma A.1 part 2. □

Proof of Lemma 4.1. Denote by $\mathrm{cl}(T_xM)$ the set of closed subsets of T_xM. Define $\psi : M \to \mathrm{cl}(T_xM)$ by $\psi(\xi) := \{x \in \tilde D_x : \exp_x(x) = \xi\} = \exp_x^{-1}(\xi) \cap \tilde D_x$. Note that $\psi(\xi)$ is a closed set by Lemma A.2.

We claim that ψ is *weakly-measurable*, i.e., for every open set $\tilde U \subseteq T_xM$, the subset of M defined by $\{\xi \in M : \psi(\xi) \cap \tilde U \ne \emptyset\}$ is Borel. To see this, note that

$$\begin{aligned}
&\{\xi \in M : \psi(\xi) \cap \tilde U \ne \emptyset\} \\
&= \{\xi \in M : \exp_x^{-1}(\xi) \cap \tilde D_x \cap \tilde U \ne \emptyset\} \\
&= \{\xi \in M : \exp_x(\tilde D_x \cap \tilde U) \ni \xi\} \\
&= \exp_x(\tilde D_x \cap \tilde U).
\end{aligned}$$

As inner product spaces, T_xM and $\mathbb{R}^d$ are isomorphic (see Section A.1-Riemannian metric). Since, T_xM and $\mathbb{R}^d$ are homeomorphic as topological spaces, $\mathbb{R}^d$ being locally compact implies T_xM is locally compact as well. Thus, we can write $\tilde{U} = \bigcup_{i\in\mathbb{N}} \tilde{K}_i$ as a countable union of compact sets $\tilde{K}_i \subseteq T_xM$. Furthermore, $\tilde{D}_x \cap \tilde{U} = \bigcup_{i\in\mathbb{N}} \tilde{D}_x \cap \tilde{K}_i$ and so $\exp_x(\tilde{D}_x \cap \tilde{U}) = \bigcup_{i\in\mathbb{N}} \exp_x(\tilde{D}_x \cap \tilde{K}_i)$.

Since $\exp_x$ is continuous, $\exp_x(\tilde{D}_x \cap \tilde{K}_i)$ is a compact subset of M, and hence closed and bounded by the Hopf-Rinow theorem ([20, Ch. 8, Theorem 2.8]). Thus, $\exp_x(\tilde{D}_x \cap \tilde{U}) = \bigcup_{i\in\mathbb{N}} \exp_x(\tilde{D}_x \cap \tilde{K}_i)$ is a countable union of closed sets, which is Borel. This proves the claim that ψ is weakly Borel measurable.

By the Kuratowski–Ryll-Nardzewski measurable selection theorem (see [12, Theorem 6.9.3]), there exists a Borel measurable function $M \to T_xM$, which we denote by $\log_x$, such that $\log_x(\xi) \in \psi(\xi) = \exp_x^{-1}(\xi)$ for all $\xi \in M$, as desired. By construction, $\log_x(\xi) \in \exp_x^{-1}(\xi)$ for all $\xi \in M$, and so $\exp_x(\log_x(\xi)) = \xi$ is immediate.

For the "furthermore" part, let $\xi \in M$ be arbitrary and let $z := \log_x(\xi) \in \tilde{D}_x$. Let $\{z_i\} \subseteq \tilde{I}_x$ be a sequence such that $\lim_i z_i = z$. By Equation (6), we have $\mathtt{dist}_M(x, \exp_x(z_i)) = \|z_i\|_x$. By continuity of $\mathtt{dist}_M$ and $\exp_x$, we have $\mathtt{dist}_M(x,\xi) = \mathtt{dist}_M(x, \exp_x(z)) = \lim_i \mathtt{dist}_M(x, \exp_x(z_i))$. To conclude, we have $\lim_i \mathtt{dist}_M(x, \exp_x(z_i)) = \lim_i \|z_i\|_x = \|z\|_x = \|\log_x(\xi)\|_x$, as desired. □

A.3 Proof of Proposition 4.2

Recall from Section A.1-Riemannian metric, given a chart (U, φ), one can define the matrix-valued function $G : U \to \mathbb{R}^{d\times d}$ referred to earlier as the coordinate representation of the Riemannian metric. Now, Lemma A.1 part 3 states that I_x is an open neighborhood of x. Furthermore, $\tilde{I}_x$ is an open subset of T_xM, which is identified with $\mathbb{R}^d$ using an orthonormal basis (see Section A.1-Riemannian metric). Hence, $\{(I_x, \log_x|_{I_x})\}_{x\in M}$ is an atlas of M (see Section A.1-Manifolds).

Definition A.3. The chart $(I_x, \log_x|_{I_x})$ is called a *normal coordinate system* at x. Let $G : I_x \to \mathbb{R}^{d\times d}$ be the coordinate representation of the Riemannian metric for this chart. To emphasize the dependency on x, we write $G_x := G$. Denote by $G_x^\perp : M \to \mathbb{R}^{d\times d}$ the zero extension of G_x to the rest of M, i.e., $G_x^\perp(\xi) = G_x(\xi)$ for $\xi \in I_x$ and $G_x^\perp(\xi)$ is the zero matrix for $\xi \notin I_x$.

The normal coordinate system has the property that $G_x(x) = G_x^\perp(x)$ is the identity matrix. This is the result of Sakai [34, Ch. II §2 Exercise 4].

Lemma A.4 (Change-of-Variables)**.** *Let $x \in M$ be fixed. Define the function $\nu_x : M \to \mathbb{R}$ by $\nu_x(\xi) = \sqrt{|\det G_x^\perp(\xi)|}$ where $G_x^\perp$ is as in Definition A.3. Then ν_x is Borel-measurable. Furthermore, ν_x satisfies the following property: Let $f : M \to \mathbb{R}$ be an absolutely integrable function. Define the function*

$$h : T_xM \to \mathbb{R} \quad \textit{by} \quad h(z) := f(\exp_x(z)) \cdot \nu_x(\exp_x(z)).$$

Then (i) *$h(0) = f(x)$ and* (ii) *for all Borel set $\tilde{B} \subseteq T_xM$ we have $\int_B f d\lambda_M = \int_{\tilde{B}} h d\lambda$ where $B := \exp_x(\tilde{B} \cap \tilde{I}_x)$.*

Proof of Lemma A.4. We first show that ν_x is Borel-measurable. Recall that $G_x^\perp : M \to \mathbb{R}^{d\times d}$ is the zero extension of $G_x : I_x \to \mathbb{R}$, which is by definition smooth (see Section A.1-Riemannian metric). In particular, $G_x : I_x \to \mathbb{R}$ is continuous and so $\sqrt{\det(G_x(\bullet))}$ is Borel-measurable. Now, note that $\sqrt{\det(G_x^\perp(\bullet))}$ is the zero extension of $\sqrt{\det(G_x(\bullet))}$ from I_x to M. Hence, $\sqrt{\det(G_x^\perp(\bullet))}$, which is ν_x by definition, is Borel-measurable.

Next, we prove the "Furthermore" part (i). Note that $\exp_x(0) = x$. Moreover, $G_x^\perp(x) = G_x(x)$ is the identity matrix as asserted after Definition A.3 (see Sakai [34, Ch. II §2 Exercise 4]). Thus, $h(0) = f(\exp_x(0))\sqrt{|\det G_x^\perp(\exp_x(0))|} = f(x)\sqrt{1} = f(x)$, as desired.

For the "Furthermore" part (ii), we first note that $\tilde{B} = (\tilde{B} \cap \tilde{I}_x) \cup (\tilde{B} \cap \tilde{C}_x)$ expresses $\tilde{B}$ as a disjoint union. Thus, $B = \exp_x(\tilde{B}) = \exp_x(\tilde{B} \cap \tilde{I}_x) \cup \exp(\tilde{B} \cap \tilde{C}_x)$ expresses B as a disjoint union as well. Moreover, $\exp(\tilde{B} \cap \tilde{C}_x) \subseteq \exp(\tilde{C}_x) = C_x$, which has λ_M-measure zero (Lemma A.1 part 5).

Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition 3.1). Now, we directly compute to obtain the formula

$$\begin{aligned}
\int_{\tilde{B}} h d\lambda &= \int_{\tilde{B}\cap\tilde{I}_x} f \circ \exp_x \sqrt{|\det(G_x^{\perp} \circ \exp_x)|} d\lambda \\
&= \int_{\log_x(\exp_x(\tilde{B}\cap\tilde{I}_x))} f \circ \exp_x \sqrt{|\det(G_x^{\perp} \circ \exp_x)|} d\lambda \\
&= \int_{\exp_x(\tilde{B}\cap\tilde{I}_x)} f d\lambda_M \qquad \because \text{Amann and Escher [1, Ch XII, Thm 1.10]} \\
&= \int_{\exp_x(\tilde{B}\cap\tilde{I}_x)} f d\lambda_M + \int_{\exp_x(\tilde{B}\cap\tilde{C}_x)} f d\lambda_M \\
&= \int_B f d\lambda_M,
\end{aligned}$$

as desired. ∎

Proposition A.5. *Let $x \in M$ be fixed. Let X be a random variable on M with density f_X where the underlying probability space is $(\Omega, \mathbb{P}, \mathcal{A})$ (see Definition 3.3). Define $Z := \log_x(X)$. Then Z is a random variable on T_xM such that for all events $E \in \mathcal{A}$ and Borel sets $\tilde{B} \subseteq T_xM$ we have* $\Pr(E \cap \{Z \in \tilde{B}\}) = \Pr(E \cap \{X \in \exp_x(\tilde{B} \cap \tilde{I}_x)\})$,

Proof of Proposition A.5. To start with, we have

$$\begin{aligned}
&\Pr(E \cap \{Z \in \tilde{B}\}) \\
&= \Pr(E \cap \{Z \in \tilde{B} \cap \tilde{D}_x\}) \qquad \because \log_x(M) \subseteq \tilde{D}_x \\
&= \Pr(E \cap \{Z \in \tilde{B} \cap \tilde{I}_x\}) + \Pr(E \cap \{Z \in \tilde{B} \cap \tilde{C}_x\}) \qquad \because \tilde{D}_x = \tilde{I}_x \cup \tilde{C}_x,\ \emptyset = \tilde{I}_x \cap \tilde{C}_x \\
&= \Pr(E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{I}_x\}) + \Pr(E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{C}_x\}).
\end{aligned}$$

Since $\exp_x : \tilde{I}_x \to I_x$ is a diffeomorphism (Lemma A.1-part 4) with inverse $\log_x$, we have

$$E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{I}_x\} = E \cap \{X \in \exp_x(\tilde{B} \cap \tilde{I}_x)\}$$

as sets. On the other hand,

$$E \cap \{\log_x(X) \in \tilde{B} \cap \tilde{C}_x\} \subseteq \{X \in C_x\}.$$

Finally, $\Pr(X \in C_x) = \int_{C_x} f_X d\lambda_M = 0$ since C_x has λ_M-measure zero (Lemma A.1-part 5). ∎

Proof of Proposition 4.2 part (i). Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition 3.1). Let $E = \Omega$ in Proposition A.5. Then we have

$$\begin{aligned}
&\Pr(Z \in \tilde{B}) \\
&= \Pr(X \in \exp_x(\tilde{B} \cap \tilde{I}_x)) \qquad \because \text{Part (i)} \\
&= \int_{\exp_x(\tilde{B}\cap\tilde{I}_x)} f_X d\lambda_M \qquad \because f_X \text{ is the density of } X \\
&= \int_{\tilde{B}\cap\tilde{I}_x} (f_X \circ \exp_x) \cdot (\nu_x \circ \exp_x) d\lambda \qquad \because \text{Lemma A.4} \\
&= \int_{\tilde{B}} f_Z d\lambda \qquad \because \text{Definition of } f_Z
\end{aligned}$$

By assumption, f_X is Borel-measurable. By Lemma A.4, ν_x is Borel-measurable. Since $\exp_x$ is continuous, we have that both $f_X \circ \exp_x$ and $\nu_x \circ \exp_x$ are Borel-measurable. This proves that f_Z is Borel-measurable. Hence, the integrand is Borel-measurable and a density function for Z. ∎

Proof of Proposition 4.2 part (ii). Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition 3.1). By Lemma A.1 part 6, we have $\tau_x^* := \inf_{u \in U_x M} \tau_x(u) > 0$. Now, let $r \in (0, \tau_x^*)$. By the definition of r, we have $\mathtt{B}_x(r, M) \subseteq \tilde{I}_x$. Hence letting $z = \log_x(\xi)$ for $\xi \in \mathtt{B}_x(r, M)$, by Equation (6) we have

$$\mathtt{dist}_M(x, \xi) = \mathtt{dist}_M(x, \exp_x(z)) = \|z\|_x. \tag{8}$$

Thus,

$$\log_x(\mathtt{B}_x(r, M)) = \{z \in T_x M : \|z\|_x < r\} = \mathtt{B}_0(r, T_x M) \tag{9}$$

and

$$\mathtt{B}_x(r, M) = \exp_x(\mathtt{B}_0(r, T_x M)). \tag{10}$$

Thus, by Lemma A.4, we have

$$\int_{\mathtt{B}_x(r,M)} f d\lambda_M = \int_{\mathtt{B}_0(r,T_x M)} h d\lambda. \tag{11}$$

Before proceeding, we need the following lemma:

Lemma A.6. *For all $x \in M$, we have $\lim_{r\to 0} \frac{\lambda_M(\mathtt{B}_x(r,M))}{\lambda(\mathtt{B}_0(r,T_x M))} = 1$.*

Proof of Lemma A.6. Let $\omega_d := \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ be the volume of the unit ball in $\mathbb{R}^d$ where Γ is the gamma function. Then $\lambda(\mathtt{B}_0(r, T_x M)) = \omega_d r^d$. Next, [34, Ch II.5 Exercise 3] states that

$$\lim_{r\to 0} \frac{r^d \omega_d - \lambda_M(\mathtt{B}_x(r, M))}{r^{d+2}} = \frac{\omega_d}{6(d+2)} S_x$$

where $S_x \in \mathbb{R}$ is a constant that depends only on x (it is the scalar curvature of M at x). By simple algebra, the above yields

$$0 = \lim_{r\to 0} \frac{1}{r^2}\left(1 - \frac{\lambda_M(\mathtt{B}_x(r, M))}{\omega_d r^d} - \frac{S_x r^2}{6(d+2)}\right)$$

In particular, we have $\lim_{r\to 0} 1 - \frac{\lambda_M(\mathtt{B}_x(r,M))}{\omega_d r^d} = 0$, as desired. $\square$

Now we continue with the proof of Proof of Proposition 4.2 part (ii). We observe that

$$\begin{aligned}
f(x) &= \lim_{r\to 0} \frac{\int_{\mathtt{B}_x(r,M)} f d\lambda_M}{\lambda_M(\mathtt{B}_x(r, M))} && \because x \text{ is a Lebesgue point of } f \\
&= \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_x(M))} h d\lambda}{\lambda_M(\mathtt{B}_x(r, M))} && \because \text{definition of } h \text{ and equation (11)} \\
&= \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_x M)} h d\lambda}{\lambda_M(\mathtt{B}_x(r, M))} \frac{\lambda_M(\mathtt{B}_x(r, M))}{\lambda(\mathtt{B}_0(r, T_x M))} && \because \text{Lemma A.6} \\
&= \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_x M)} h d\lambda}{\lambda(\mathtt{B}_0(r, T_x M))}.
\end{aligned}$$

Since $f(x) = h(0)$ (Lemma A.4) , we've shown that

$$g(0) = \lim_{r\to 0} \frac{\int_{\mathtt{B}_0(r,T_x M)} h d\lambda}{\lambda(\mathtt{B}_0(r, T_x M))}.$$

Thus, 0 is a Lebesgue point of h, as desired. $\square$

A.4 Proof of Proposition 4.3

Recall that λ is the shorthand for the ordinary Lebesgue measure $\lambda_{\mathbb{R}^d}$ (see paragraph right after Definition 3.1). Let $A \subseteq \mathbb{R}$ and $\tilde{B} \subseteq T_x M$ be Borel subsets. Then

$$\begin{aligned}
&\int_{\tilde{B}} P_{Y|Z}(A|z) f_Z(z) d\lambda(z) \\
&= \int_{\tilde{B}} P_{Y|X}(A|\exp_x(z)) f_Z(z) d\lambda(z) \qquad \because \text{Definition of } P_{Y|Z=z} \\
&= \int_{\exp_x(\tilde{B}\cap\tilde{I}_p)} P_{Y|X}(A|x) f_X(x) d\lambda_M(x) \qquad \because \text{Lemma A.4 and Proposition 4.2 (ii)} \\
&= \Pr(Y \in A, X \in \exp_x(\tilde{B} \cap \tilde{I}_x)) \\
&= \Pr(Y \in A, Z \in \tilde{B}) \qquad \because \text{Proposition 4.2 (i) with } E := \{Y \in A\}
\end{aligned}$$

This proves that $P_{Y|Z}(\cdot|\cdot)$ is a conditional probability for Y given Z.