Journal Pre-proof

Learning on manifolds without manifold learning

H.N. Mhaskar, Ryan O'Dowd

PII: S0893-6080(24)00683-X
DOI: https://doi.org/10.1016/j.neunet.2024.106759
Reference: NN 106759

To appear in: *Neural Networks*

Received date : 19 February 2024
Revised date : 18 August 2024
Accepted date : 23 September 2024

Please cite this article as: H.N. Mhaskar and R. O'Dowd, Learning on manifolds without manifold learning. *Neural Networks* (2024), doi: https://doi.org/10.1016/j.neunet.2024.106759.

Learning on manifolds without manifold learning

H. N. Mhaskar* Ryan O'Dowd†

August 18, 2024

Abstract

Function approximation based on data drawn randomly from an unknown distribution is an important problem in machine learning. The manifold hypothesis assumes that the data is sampled from an unknown submanifold of a high dimensional Euclidean space. A great deal of research deals with obtaining information about this manifold, such as the eigendecomposition of the Laplace-Beltrami operator or coordinate charts, and using this information for function approximation. This two-step approach implies some extra errors in the approximation stemming from estimating the basic quantities of the data manifold in addition to the errors inherent in function approximation. In this paper, we project the unknown manifold as a submanifold of an ambient hypersphere and study the question of constructing a one-shot approximation using a specially designed sequence of localized spherical polynomial kernels on the hypersphere. Our approach does not require preprocessing of the data to obtain information about the manifold other than its dimension. We give optimal rates of approximation for relatively "rough" functions.

1 Introduction

In the past quarter-century, machine leaning has impacted our lives ubiquitously, from driving cars to military maneuvers. Shallow and deep neural networks have played a central role in these applications. In turn, a theoretical justification for the use of these networks is their universal approximation property: they can approximate arbitrarily well an arbitrary continuous function on an arbitrary compact subset of a Euclidean space of arbitrary dimension. In mathematical terms, the challenge can be formulated as follows. We are given a data of the form $\mathcal{D} = \{(y_j, z_j)\}_{j=1}^M$, drawn randomly from an unknown probability distribution τ, and we wish to find a parametrized model $G(\boldsymbol{\theta}; y)$ to minimize the *generalization error* $\mathbb{E}_\tau(\mathcal{L}(z, G(\boldsymbol{\theta}, y)))$ for a judiciously chosen loss functional $\mathcal{L}$. For example, in a shallow neural network of the form $\sum_{k=1}^m a_k \sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k)$, $\mathbf{x} \in \mathbb{R}^q$, the parameters are $\boldsymbol{\theta} = (a_k, w_{k,1}, \cdots, w_{k,q}, b_k)_{k=1}^m$. Since τ is not known, one minimizes instead the empirical risk obtained by discretizing the expected value in terms of the data. There is a huge amount of literature on the choice of the loss functional, usually involving the correct choice of one or more regularization terms, the difference between the minimal empirical risk and the generalization error in terms of the number of samples, strategies for the optimization involved, the geometry of the error surface, etc.

Writing $f(y) = \mathbb{E}_\tau(z|y)$, the fundamental problem is to approximate f given the data $\mathcal{D}$. The role of approximation theory is to estimate the relation between the minimum generalization error and number of parameters in $\boldsymbol{\theta}$ in terms of some properties of f and G. Naturally, there is a huge amount of literature in this direction as well, especially when $\mathcal{L}$ is a square loss or, since τ is unknown, the uniform or probabilistic loss. In the case of the square loss, the generalization error splits into the variance $\mathbb{E}_\tau(|z - f(y)|^2)$ and the bias $\mathbb{E}_\nu(|f(y) - G(\boldsymbol{\theta}, y)|^2)$, where ν is the marginal distribution of y.

We think that the whole paradigm of getting an insight into the number of parameters using approximation theory, and then using an optimization procedure to actually obtain the approximation in a decoupled manner needs to be revisited. We list some reasons.

1. The use of a global metric for measuring the generalization/approximation error is insensitive to local effects in the target function (see Example 2.1).

2. The use of the degree of approximation to get an insight on the model complexity may be misleading, as we will elaborate on shortly.

*Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711. email: hrushikesh.mhaskar@cgu.edu.
The research is supported in part by NSF grant DMS 2012355, and ONR grants N00014-23-1-2394, N00014-23-1-2790.
†Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711. email: ryan.o'dowd@cgu.edu.

3. There is no guarantee that the minimizer of the empirical risk would be the one which gives the best approximation error or generalization error. Moreover, absolute minima are hard to obtain, and perhaps not required in practice anyway.

4. The training process may be very sensitive to the initialization of the parameters. It is observed in [22] that with a wrong initialization of parameters, a deep network evaluating the ReLU activation function trained to approximate $|t| = t_+ + (-t)_+$ results in a constant output. This is a phenomenon which they have called "dead on arrival."

Since the problem is fundamentally one of function approximation, it is natural to question whether one could use a new paradigm where the approximation is constructed directly from the data, and the error on the data not yet seen can be estimated directly as well. So far, approximation theory has played only a marginal role in machine learning. There are several reasons for this.

1. Many papers on function approximation by shallow or deep networks ignore the fact that the approximation needs to be constructed from the data. For example, the dimension independent bounds are typically derived using probabilistic arguments resulting in estimates which could be misleading. Particularly, we have shown in [28, 31] that drastically different estimates are obtained for approximation by ReLU networks for the **same class of functions** depending on whether the networks are constructed from the data or not.

2. We do not typically know whether the assumptions on the target function involved in the approximation theory bounds are satisfied in practice, or whether the number of parameters is the right criterion to look at in the first place. For example, when one considers approximation by radial basis function (RBF) networks, it is observed in many papers (e.g., [27]) that the minimal separation among the centers is the right criterion rather than the number of parameters. It is shown that if one measures the degree of approximation in terms of the minimal separation, then one can determine the smoothness of the underlying target function by examining the rate at which the degrees of approximation converge to 0.

3. Most of the approximation theory literature focuses on the question of estimating the difference between f and $G(\boldsymbol{\theta}, \circ)$ in various norms and conditions on f, where the support of the marginal distribution ν is assumed to be a known domain, such as a torus, a cube, the whole Euclidean space, a hypersphere (simply referred to as a sphere in the remainder of this paper), etc.; equivalently, one assumes that the data points y_j are "dense" on such a domain. This creates a gap between theory, where the domain of ν is known, and practice, where it is not. One consequence of approximating, say on a cube, is the curse of dimensionality. That is, if the dimension of the input data is Q, then the number of components in the parameter vector $\boldsymbol{\theta}$ to achieve an accuracy of ϵ will be $\Omega(\epsilon^{-Q})$.

Rather than approximating on a known domain, a relatively recent idea is to assume that the support of the marginal distribution ν is an **unknown**, low-dimensional submanifold of the high-dimensional ambient space in which the data is located. This gives rise to a two-step procedure: *manifold learning*, where there is an effort to find information about the manifold itself, and then *function approximation* (which we have called *learning on the manifold* in the title of this paper), where we assume the necessary information about the manifold to be known, and study function approximation based on this information.

Works by Belkin, Niyogi, Singer, and others have shown that the so-called graph Laplacian (and the corresponding eigendecomposition) constructed from data points converges to the manifold Laplacian and its eigendecomposition. Some preliminary papers in this direction are: [2, 3, 40]. An introduction to the subject is given in [7]. Another approach is to estimate an atlas of the manifold, which thereby allows function approximation to be conducted via local coordinate charts. One such effort is to utilize the underlying parametric structure of the functions to determine the dimension of the manifold and the parameters involved [23]. Approximations utilizing estimated coordinate charts have been implemented, for example, via deep learning [10, 39], moving least-squares [42], local linear regression [6], or using Euclidean distances among the data points [8]. HNM and his collaborators carried out an extensive investigation of function approximation on manifolds, some of which is summarized in [30]. With the two-step procedure, the estimates obtained in function approximation need to be tempered by the errors accrued in the manifold learning step. In turn, the errors in the manifold learning step are very sensitive to the choice of different parameters used in the process.

The purpose of this paper is to introduce a direct method of approximation on *unknown* manifolds without trying to find out anything about the manifold other than its dimension. Toward this goal, we project the q-dimensional manifold $\mathbb{X}$ in question from the ambient space $\mathbb{R}^Q$ to a sphere $\mathbb{S}^Q$ of the same dimension. We can then use a

specially designed, localized, univariate kernel $\Phi_{n,q}$ (cf. (4.12)) which is a spherical polynomial of degree $< n$ on $\mathbb{S}^Q$, with n and q being tunable hyperparameters. Our construction is very simple; we define

$$F_n(\mathcal{D};x) := \frac{1}{M}\sum_{j=1}^{M} z_j \Phi_{n,q}(x\cdot y_j). \tag{1.1}$$

We note that $F_n(\mathcal{D};\circ)$ is a function defined on the ambient sphere $\mathbb{S}^Q$. The localization of the kernel allows us to adapt the approximation to the unknown manifold.

Our main theorem (cf. Theorem 5.1) has the following form:

Theorem 1.1. ***(Informal statement)*** *Let $\mathcal{D} = \{(y_j, z_j)\}_{j=1}^M$ be a set of random samples chosen from a distribution τ. Suppose f belongs to a smoothness class W_γ (detailed in Definition 5.1) with associated norm $||\circ||_{W_\gamma}$. Then under some additional conditions and with a judicious choice of n, we have with high probability:*

$$||F_n(\mathcal{D};\circ) - f||_{\mathbb{X}} \le c\left(||z|| + ||f||_{W_\gamma}\right)\left(\frac{\log M}{M}\right)^{\gamma/(q+2\gamma)}, \tag{1.2}$$

where c is a positive constant independent of f.

We note some mathematical features of our construction and theorem which we find interesting.

1. The usual approach in machine learning is to construct the approximation using an optimization procedure, usually involving a regularization term. The setting up of this optimization problem, especially the regularization term, requires one to assume that the function belongs to some special function class, such as a reproducing kernel Hilbert/Banach space. Thus, the constructions are not explicit nor universal. In contrast, our construction (1.1) does not require a prior on the function in order to use our model. Of course, the theorem and its high-probability convergence rates do require various assumptions on τ, the marginal distribution, the dimension of the manifold, the smoothness of the target function, etc. The point is that the construction itself does not require any assumptions.

2. A major problem in manifold learning is one of out of sample extension; i.e., extending the approximation to outside the manifold. A usual procedure for this in the context of approximation using the eigenstructure of the Laplace-Beltrami operator on the manifold is the Nyström extension [11]. However, this extension is no longer in terms of any orthogonal system on the ambient space, and hence there is no guarantee of the quality of approximation even if the function is known outside the manifold. In contrast, the point x in (1.1) is not restricted to the manifold, but rather freely chosen from $\mathbb{S}^Q$. That is, our construction defines an out of sample extension in terms of spherical polynomials on the ambient sphere, whose approximation capabilities are well studied.

3. In terms of M, the estimate in (1.2) depends asymptotically on the dimension q of the manifold rather than the dimension Q of the ambient space.

4. We do not need to know **anything** about the manifold (e.g., eigendecomposition or atlas estimate) itself apart from its dimension in order to prove our theorem. There are several papers in the literature for estimating the dimension from the data, for example [20, 21, 23]. However, the simplicity of our construction allows us to treat the dimension q as a tunable parameter to be determined by the usual division of the data into training, validation, and test data.

There are several other approaches superficially similar to our constructions. We will comment on some of these in Section 2. We describe the main idea behind our proofs in Section 3. The paper requires an understanding of the approximation properties of spherical polynomials. Accordingly, we describe some background on the spherical polynomials, our localized kernels, and their use in approximation theory on subspheres of the ambient sphere in Section 4. The main theorems for approximation on the unknown manifold are given in Section 5. The theorems are illustrated with three numerical examples in Section 6. One of these examples is closely related to an important problem in magnetic resonance relaxometry, in which one seeks to find the proportion of water molecules in the myelin covering in the brain based on a model that involves inversion of the Laplace transform. The proofs of the main theorems are given in Section 7. The appendix describes the encoding of the target function (A.1), gives some background about the theory of manifolds which is used in this paper (A.2), and describes in detail the Clenshaw algorithm used to evaluate our kernels and their implementation as a deep neural network (A.3).

We would like to thank Dr. Richard Spencer at the National Institute of Aging (NIH) for his helpful comments, especially on Section 6.2, verifying that our simulation is consistent with what is used in the discipline of magnetic resonance relaxometry.

2 Related ideas

Since our method is based on a highly localized kernel, it is expected to be comparable to the simple nearest neighbor algorithm. However, rather than specifying the number of neighbors to consider in advance, our method allows the selection of neighbors adaptively for each test point, controlled by the parameter n. Also, rather than taking a simple averaging, our method is more sophisticated, designed to give an optimal order of magnitude of the approximation error.

One of the oldest ideas for data based function approximation is the so-called Nadaraya-Watson estimator (NWE), given by

$$NW_h(x) = \frac{\sum_{j=1}^M z_j K(|x-y_j|/h)}{\sum_{j=1}^M K(|x-y_j|/h)},$$

where K is a kernel with an effectively small support—the Gaussian kernel $K(t) = \exp(-t^2)$, as a common example—and h is a scaling parameter. Another possible choice is a B-spline (including Bernstein polynomials) which has a compact support. This construction is designed to work on a Euclidean space by effectively shrinking the support of K using the scaling parameter $h \to 0$, analogously to spline approximation. The degree of approximation of such methods is measured in terms of h. It is well known (e.g., [13]) that the use of a positive kernel K suffers from the so-called saturation phenomenon: the degree of approximation cannot be smaller than $\mathcal{O}(h^2)$ unless the function is a trivial one in some sense.

Radial basis function (RBF) networks and neural networks are used widely for function approximation, using either interpolation or least square fit. Standard RBF networks, such as Gaussian networks or thin plate spline networks, use a fixed, scaled kernel. Typically, the matrices involved in either interpolation or least square approximation are very ill-conditioned, and the approximation is not highly localized.

Restricted to the sphere, both of the notions are represented by a zonal function (ZF) network. A *zonal function* on a sphere is a function of the form $x \mapsto g(x \cdot x_0)$. A ZF network is a linear combination of finitely many zonal functions. One may notice that

$$g(x \cdot x_0) = g\left(1 - \frac{|x-x_0|^2}{2}\right),$$

so we can see that a ZF network is also a neural/RBF network. Conversely, a neural/RBF network restricted to the sphere is a ZF network. The same observations about RBF networks hold for ZF networks as well. We note that all the papers we are aware of which deal with approximation by ZF networks actually end up approximating a spherical polynomial by the networks in question.

Rather than working with a fixed, scaled kernel, in this paper we deal with a sequence of highly localized polynomial kernels. We do not need to solve any system of equations or do any optimization to arrive at our construction. RBF networks and NWE were developed for approximation on Euclidean domains instead of unknown manifolds. Both have a single hyperparameter h and work analogously to the spline approximation. In contrast, our method is designed for approximation on unknown manifolds without having to learn anything about the manifold besides the dimension. It has two integer hyperparameters (n and q) and yields a polynomial approximation.

If one chooses h small enough relative to a fixed n then NWE may be able to outperform our method as measured in terms of a global error bound, such as the root mean square (RMS) error. If one instead chooses n large enough relative to a fixed h then our method may be able to outperform NWE. So in order to give a fair comparison in Example 2.1, we force the RMS error of both methods to be approximately equivalent and investigate the qualitative differences of the errors produced by each method. We additionally show that both methods in the example outperform an interpolatory RBF network.

Example 2.1. This example serves to illustrate two points. The first point is to compare the performance our method with NWE and an RBF interpolant. In doing so, we show that the error associated with our method is localized to singularities of the target function, whereas the other methods do not exhibit this behavior. The second point is that using a global error estimate such as RMS can be misleading. Even if the RMS error with a given method might be large, the percentage of test data points at which it is smaller than a threshold could be substantially higher due to the local effects in the target function.

To ensure fair comparison, we use each of the three methods for approximation on $\mathbb{S}^1 = \{(\cos\theta, \sin\theta) : \theta \in (-\pi, \pi]\}$, where the Gaussian kernel can be expressed in the form of a zonal function as explained above.

We consider the function

$$f(\theta) = 1 + |\cos\theta|^{7/2} \sin(\cos\theta + \sin\theta)/2, \qquad \theta \in (-\pi, \pi]. \tag{2.1}$$

We note that the function is analytic except at $\theta = \pm\pi/2$, where it has a discontinuity in the 4th order derivative. Our training data consists of 2^{13} equidistantly spaced y_j's along the circle. We set $z_j = f(y_j)$, and examine the resulting error on a test data consisting of 2^{11} points chosen randomly according to the uniform distribution on $\mathbb{S}^1$.

We consider three approximation processes : (1) Nadaraya-Watson estimator NW_h with $K_h(t) = \exp(-t^2/h^2)$, (2) interpolatory approximation by the RBF network of the form $\sum a_k \exp(-|\circ - y_j|^2/h^2)$, (3) our method with the kernel $\Phi_{50,1}$.

We experimentally determined the optimal h value in NWE to be $\approx 7.45e\text{-}4$ (effectively simulating the minimization of the actual generalization error on the test data), resulting in an RMS error of 1.8462e-7. The same value of h was used for interpolation with the Gaussian RBF network, yielding a RMS error of 2.2290e-4. We then chose n so as to yield a (comparable to NWE) RMS error of 1.8594e-7 (though we note that in this case our method continues to provide a better approximation if n is further increased).

The detailed results are summarized in Figure 1 below.

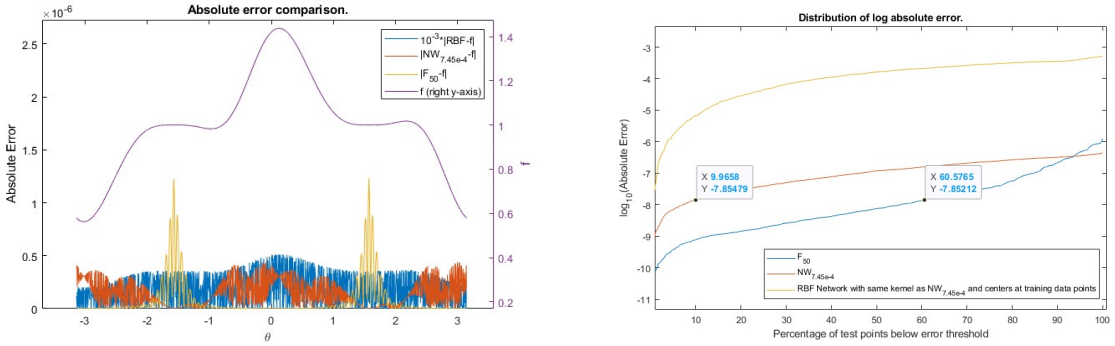

Figure 1: Error comparison between our method, the Nadaraya-Watson estimator, and an interpolatory RBF network. (Left) Comparison of absolute errors between the methods with the target function plotted on the right y-axis for benefit of the viewer. We note that the error from the RBF method is scaled by 10^{-3} so as to not dominate the figure. (Right) Percent point plot of the log absolute error for all three methods.

In the left plot in Figure 1, we can see a clear difference between the errors of the three methods. The (scaled by 10^{-3}) error from the RBF network jumps throughout the whole domain, signaling the ill-conditioned nature of the matrix. The error from the Nadaraya-Watson estimator exhibits some oscillation across the whole domain as well. The error with our method is localized to the two singularity points of the function. In other words, our method exhibits 1) sensitivity to the singularities of a function and 2) error adapting to the local smoothness of the function. In comparison, RBF networks and NWE do not always exhibit such behavior. On the right plot of Figure 1, we give a percent point plot of the log absolute error for all three methods. There are three curves corresponding to the three methods being compared. Each point (x, y) along a given curve indicates that the corresponding method approximated x% of test points with absolute error below 10^y. This plot can also be thought of as the inverse CDF for the random variable of the resulting log absolute error for a test point sampled uniformly at random. For example, whereas the Nadaraya-Watson estimator yields an error below $\approx 10^{-7.85}$ for only about 10% of the tested points, our method exhibits the same error or below for about 60% of the test points. Our method has the higher uniform error, but lower error for over 90% of the test points. Although the overall RMS error is roughly the same, our method exhibits a quicker decay from the uniform error. This illustrates, in particular, that measuring the performance using a global measure for the error, such as the uniform or RMS error can be misleading. The interpolatory RBF network performs the worst of the three methods as the right plot of Figure 1 shows clearly. ■

There are some efforts [15, 19] to do function approximation on manifolds using tensor product splines or RBF networks defined on an ambient space by first extending the target function to the ambient space. A locally adaptive polynomial approach is used in [41] for accomplishing function approximation on manifolds using the data. All these works require that the manifold be known.

In [29], we have suggested a direct approach to function approximation on an unknown submanifold of a Euclidean space using a localized kernel based on Hermite polynomials. This construction was used successfully in predicting diabetic sugar episodes [32] and recognition of hand gestures [24]. In particular, in [32], we constructed our approximation based on one clinical data set and used it to predict the episodes based on another clinical data set. In order to extend the applicability of such results to wearable devices, it is important that the approximation should be encoded by a hopefully small number of real numbers, which can then be hardwired or used for a

simpler approximation process [34]. However, the construction in [29] is a linear combination of kernels of the form $\Psi(|\circ - y_j|)$, where $\Psi(t) = P(t)\exp(-t^2/2)$ is a univariate kernel utilizing a judiciously chosen polynomial P. This means that we get a good approximation, but the space from which the approximation takes place changes with the point at which the approximation is desired. This does not allow us to encode the approximation using finitely many real numbers. In contrast, the method proposed in this paper allows us to encode the approximation using coefficients of the target function in the spherical harmonic expansion (defined in a distributional sense), computed empirically. This sequence can be reduced using connections between ultraspherical polynomials with different parameters and simple algorithms to detect and remove redundant coefficients. This is described in Appendix A.1. Moreover, the degree of the polynomials involved in [29] to obtain same the rate of convergence in terms of the number of samples is $\mathcal{O}(n^2)$, while the degree of the polynomials involved in this paper is $\mathcal{O}(n)$. We note that the construction in both the papers involve only univariate polynomials, so that the dimension of the input space enters only linearly in the complexity of the construction.

3 An overview of the proof

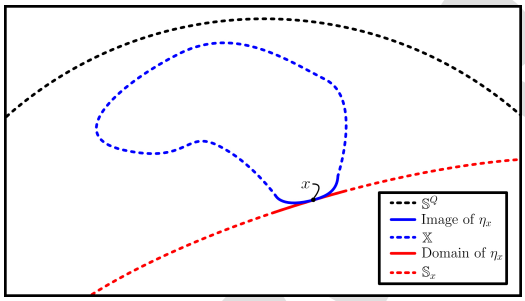

Figure 2: Visualization of our approximation approach. Here, $\mathbb{X}$ is a submanifold of the sphere $\mathbb{S}^Q$. The map η_x, analogous to the exponential map, allows us to relate the part of the integral in (3.3) near x with an integral on the tangent sphere at x via a change of variables (solid curves). The localization of the kernels in our method allow for the approximation to be extended over $\mathbb{X}$ and the tangent sphere $\mathbb{S}_x$ (dotted curves).

We can think of $F_n(\mathcal{D}; x)$ defined in (5.5) as an empirical approximation for an expected value with respect to the data distribution τ:

$$\mathbb{E}_\tau(F_n(\mathcal{D}; x)) = \int z\Phi_{n,q}(x \cdot y)d\tau(y). \tag{3.1}$$

Assuming that the marginal distribution of τ on $\mathbb{X}$ is absolutely continuous with respect to the Riemannian volume measure μ^* on $\mathbb{X}$; i.e., given by $f_0 d\mu^*$ for some smooth function f_0, we have

$$\mathbb{E}_\tau(F_n(\mathcal{D}; x)) = \int_{\mathbb{X}} f(y)f_0(y)\Phi_{n,q}(x \cdot y)d\mu^*(y). \tag{3.2}$$

Accordingly, we define an *integral reconstruction operator* by

$$\sigma_n(\mathbb{X}, f)(x) := \int_{\mathbb{X}} \Phi_{n,q}(x \cdot y)f(y)d\mu^*(y), \qquad f \in C(\mathbb{X}),\ x \in \mathbb{X}, \tag{3.3}$$

study the approximation properties of this operator, and use it with ff_0 in place of f. The approximation properties of the operator σ_n in the case of when $\mathbb{X}$ is the q-dimensional sphere $\mathbb{S}^q$ are well known (Proposition 4.2), and can be easily transferred to a q-dimensional equator of the ambient sphere $\mathbb{S}^Q$ (Section 4.4, Theorem 4.1). We introduce a local exponential map η_x at $x \in \mathbb{X}$ between $\mathbb{X}$ and the tangent equatorial sphere $\mathbb{S}_x$ (i.e., a rotated version of $\mathbb{S}^q$ that shares the tangent space with $\mathbb{X}$ at x). We give an illustration of this setup in Figure 2. Locally, a change of variable formula and the properties of this map allow us to compare the integral over a small manifold ball with that of its image on $\mathbb{S}_x$ (cf. (7.10)). We keep track of the errors using the Bernstein inequality for spherical polynomials (cf. (4.14)) and standard approximations between geodesic distances and volume elements on the manifold by those on $\mathbb{S}_x$. This constitutes the main part of the proof of the critical Lemma 7.1. We use the high localization property of our kernel $\Phi_{n,q}$ to lift the rest of the integral in (3.3) on $\mathbb{X}$ at any point $x \in \mathbb{X}$ to the rest of $\mathbb{S}_x$ with small error (cf. (7.11), (7.12)). After this, we can use known results from the theory of approximation on the sphere by spherical polynomials (cf. Proposition 4.2 and Theorem 4.1). A partition of unity argument is used often in the proof.

Having obtained the approximation result for the integral reconstruction operator, we then discretize the integral and keep track of the errors using concentration inequalities.

4 Background

In this section, we outline some important details about spherical harmonics (Section 4.1) which leads to the construction of the kernels of interest in this paper (Section 4.2). We then review some classical approximation results using these kernels on spheres (Section 4.3) and equators of spheres (Section 4.4).

4.1 Spherical harmonics

The material in this section is based on [33, 43]. Let $0 \le q \le Q$ be integers. We define the *q-dimensional sphere* embedded in $Q+1$-dimensional space as follows

$$\mathbb{S}^q := \{(x_1, \ldots, x_{q+1}, \underbrace{0, \ldots, 0}_{Q-q}) : x_1^2 + \cdots + x_{q+1}^2 = 1\}. \tag{4.1}$$

Observe that $\mathbb{S}^q$ is a q-dimensional compact manifold with geodesic defined by $\rho(x,y) = \arccos(x \cdot y)$.

Let μ_q^* denote the normalized volume measure on $\mathbb{S}^q$. By representing a point $x \in \mathbb{S}^q$ as $(x' \sin\theta, \cos\theta)$ for some $x' \in \mathbb{S}^{q-1}$, one has the recursive formula for measures

$$\frac{\omega_q}{\omega_{q-1}} d\mu_q^*(x) = \sin^{q-1}(\theta) d\theta d\mu_{q-1}^*(x'), \tag{4.2}$$

where ω_q denotes the surface volume of $\mathbb{S}^q$. One can write ω_q recursively by

$$\omega_q = \frac{2\pi^{(q+1)/2}}{\Gamma((q+1)/2)} = \begin{cases} 2\pi, & \text{if } q = 1, \\ \sqrt{\pi}\dfrac{\Gamma(q/2)}{\Gamma(q/2+1/2)}\omega_{q-1}, & \text{if } q \ge 2, \end{cases} \tag{4.3}$$

where Γ denotes the Gamma function.

The restriction of a homogenous harmonic polynomial in $q+1$ variables to the q-dimensional unit sphere $\mathbb{S}^q$ is called a spherical harmonic. The space of all spherical harmonics of degree ℓ in $q+1$ variables will be denoted by $\mathbb{H}_\ell^q$. The space of the restriction of all $q+1$ variable polynomials of degree $< n$ to $\mathbb{S}^q$ will be denoted by Π_n^q. We extend this notation for an arbitrary real value $x > 0$ by writing $\Pi_x^q := \Pi_{\lfloor x \rfloor}^q$. It is known that $\mathbb{H}_\ell^q$ is orthogonal to $\mathbb{H}_j^q$ in $L^2(\mu_q^*)$ whenever $j \ne \ell$, and $\Pi_n^q = \bigoplus_{\ell=0}^{n-1} \mathbb{H}_\ell^q$. In particular, $L^2(\mu_q^*) = \bigoplus_{\ell=0}^{\infty} \mathbb{H}_\ell^q$.

If we let $\{Y_{\ell,k}\}_{k=1}^{\dim(\mathbb{H}_\ell^q)}$ be an orthonormal basis for $\mathbb{H}_\ell^q$ with respect to μ_q^*, we can define

$$K_{q,\ell}(x,y) := \sum_{k=1}^{\dim(\mathbb{H}_\ell^q)} Y_{\ell,k}(x) Y_{\ell,k}(y). \tag{4.4}$$

In [33, 43], it is shown that

$$K_{q,\ell}(x,y) = \frac{\omega_q}{\omega_{q-1}} p_{q,\ell}(1) p_{q,\ell}(x \cdot y), \tag{4.5}$$

where $p_{q,\ell}$ denotes the orthonormalized ultraspherical polynomial of dimension q and degree ℓ. These ultraspherical polynomials satisfy the following orthogonality condition.

$$\int_{-1}^{1} (1-x^2)^{(q/2-1)} p_{q,m}(x) p_{q,n}(x) dx = \delta_{m,n}. \tag{4.6}$$

Computationally, it is customary to use the following recurrence relation:

$$\begin{aligned} \sqrt{\frac{(n+1)(n+q-1)}{(2n+q-1)(2n+q+1)}} p_{q,n+1}(x) &= x p_{q,n}(x) - \sqrt{\frac{n(n+q-2)}{(2n+q-1)(2n+q-3)}} p_{q,n-1}(x), \qquad n \ge 1, \\ p_{q,0}(x) = p_{q,0} &= 2^{1/2-q/2} \frac{\Gamma(q-1)}{\Gamma(q/2)}, \qquad p_{q,1}(x) = 2^{1/2-q/2} \frac{\sqrt{\Gamma(q)\Gamma(q+1)}}{\Gamma(q/2)} x. \end{aligned} \tag{4.7}$$

We note further that

$$p_{q,n}(1) = \frac{2^{1/2-q/2}}{\Gamma(q/2)}\sqrt{\frac{\Gamma(n+q-1)(2n+q-1)}{\Gamma(n+1)}}. \tag{4.8}$$

Remark 4.1. Many notations have been used for ultraspherical polynomials in the past. For example, [44] uses the notation of $P_n^{(\lambda)}$ for the Gegenbauer polynomials, also commonly denoted by $C_n^{(\lambda)}$. It is also usual to use a normalization, which we will denote by R_n^q in this remark, given by $R_n^q = p_{q,n}/p_{q,n}(1)$. Ultraspherical polynomials are also simply a special case of the Jacobi polynomials $P_n^{(\alpha,\beta)}$ where $\alpha = \beta$. Setting

$$h_{q,n} := 2^{q-1}\frac{\Gamma(n+q/2)^2}{n!\Gamma(n+q-1)(2n+q-1)}, \tag{4.9}$$

we have the following connection between these notations:

$$\begin{aligned} p_{q,n}(x) &= h_{q,n}^{-1/2}P_n^{(q/2-1,q/2-1)}(x) = \frac{\Gamma(q-1)}{\Gamma(q/2)}\sqrt{\frac{n!(2n+q-1)}{2^{q-1}\Gamma(n+q-1)}}C_n^{(q/2-1/2)}(x) \\ &= \frac{2^{1/2-q/2}}{\Gamma(q/2)}\sqrt{\frac{\Gamma(n+q-1)\Gamma(2n+q-1)}{\Gamma(n+1)}}R_n^q. \end{aligned} \tag{4.10}$$

■

Furthermore, the ultraspherical polynomials for the sphere of dimension d_1 can be represented by those for the sphere of dimension d_2 in the following manner

$$p_{d_1,n} = \sum_{\ell=0}^{n} C_{d_2,d_1}(\ell,n)p_{d_2,\ell}. \tag{4.11}$$

The coefficients C have been studied, and explicit formulas are given in [1, Equation 7.34] and [44, Equation 4.10.27].

The constant convention

In the sequel, $c, c_1, \cdots$ will denote generic positive constants depending upon the fixed quantities in the discussion, such as the manifold, the dimensions q, Q, and various parameters such as S to be introduced below. Their values may be different at different occurrences, even within a single formula. The notation $A \lesssim B$ means $A \le cB$, $A \gtrsim B$ means $B \lesssim A$, and $A \sim B$ means $A \lesssim B \lesssim A$.

4.2 Localized kernels

Let h be an infinitely differentiable function supported on $[0,1]$ where $h(x) = 1$ on $[0,1/2]$. This function will be fixed in the rest of this paper, and its mention will be omitted from the notation. Then we define the following univariate kernel for $t \in [-1,1]$:

$$\Phi_{n,q}(t) := \Phi_{n,q}(h;t) = \sum_{\ell=0}^{n} h\left(\frac{\ell}{n}\right)K_{\ell,q}(t) = \frac{\omega_q}{\omega_{q-1}}\sum_{\ell=0}^{n} h\left(\frac{\ell}{n}\right)p_{q,\ell}(1)p_{q,\ell}(t). \tag{4.12}$$

The following proposition lists some technical properties of these kernels which we will often use, sometimes without an explicit mention.

Proposition 4.1. *Let $x, y \in \mathbb{S}^Q$. For any $S > 0$, the kernel $\Phi_{n,q}(x,y)$ satisfies the **localization bound***

$$|\Phi_{n,q}(x\cdot y)| \lesssim \frac{n^q}{\max(1, n\arccos(x\cdot y))^S}, \tag{4.13}$$

where the constant involved may depend upon S. Further, we have the Lipschitz condition:

$$|\Phi_{n,q}(x\cdot y) - \Phi_{n,q}(x\cdot y')| \lesssim n^{q+1}|\arccos(x\cdot y) - \arccos(x\cdot y')|, \qquad y' \in \mathbb{S}^Q. \tag{4.14}$$

Proof. The estimate (4.13) is proved in [25]. Since $\theta \mapsto \Phi_{n,q}(\cos\theta)$ is a trigonometric polynomial of degree $< n$, the Bernstein inequality for the derivatives of trigonometric polynomials implies that

$$|\Phi_{n,q}(\cos\theta) - \Phi_{n,q}(\cos\phi)| \le n\|\Phi_{n,q}\|_\infty|\theta-\phi| \lesssim n^{q+1}|\theta-\phi|.$$

This leads easily to (4.14). ■

4.3 Approximation on spheres

Methods of approximating functions on $\mathbb{S}^q$ have been studied in, for example, [26, 38] and some details are summarized in Proposition 4.2.

For a compact set A, let $C(A)$ denote the space of continuous functions on A, equipped with the supremum norm $||f||_A = \max_{x\in A} |f(x)|$. We define the degree of approximation for a function $f \in C(\mathbb{S}^q)$ to be

$$E_n(f) := \inf_{P\in\Pi_n^q} ||f - P||_{\mathbb{S}^q}. \tag{4.15}$$

Let $W_\gamma(\mathbb{S}^q)$ be the class of all $f \in C(\mathbb{S}^q)$ such that

$$||f||_{W_\gamma(\mathbb{S}^q)} := ||f||_{\mathbb{S}^q} + \sup_{n\geq 0} 2^{n\gamma} E_{2^n}(f) < \infty. \tag{4.16}$$

We note that an alternative smoothness characterized in terms of constructive properties of f is explored by many authors; some examples are given in [12]. We define the approximation operator for $\mathbb{S}^q$ by

$$\sigma_n(f)(x) := \sigma_n(\mathbb{S}^q, f)(x) = \int_{\mathbb{S}^q} \Phi_{n,q}(x\cdot u) f(u) d\mu_q^*(u). \tag{4.17}$$

With this setup, we now review some bounds on how well $\sigma_n(f)$ approximates f (cf. [26]).

Proposition 4.2. *Let $n \geq 1$.*
(a) *For all $P \in \Pi_{n/2}^q$, we have $\sigma_n(P) = P$.*
(b) *For any $f \in C(\mathbb{S}^q)$, we have*

$$E_n(f) \leq ||f - \sigma_n(f)||_{\mathbb{S}^q} \lesssim E_{n/2}(f). \tag{4.18}$$

In particular, if $\gamma > 0$ then $f \in W_\gamma(\mathbb{S}^q)$ if and only if

$$||f - \sigma_n(f)||_{\mathbb{S}^q} \lesssim ||f||_{W_\gamma(\mathbb{S}^q)}\, n^{-\gamma}. \tag{4.19}$$

Remark 4.2. Part (a) is known as a *reproduction* property, which shows that polynomials up to degree $< n/2$ are unchanged when passed through the operator σ_n. Part (b) demonstrates that σ_n yields what we term a *good approximation*, where its approximation error is no more than some constant multiple of the degree of approximation. Part (c) not only gives the approximation bounds in terms of the smoothness parameter γ, but shows also that the rate of decrease of the approximation error obtained by $\sigma_n(f)$ **determines** the smoothness γ. ■

4.4 Approximation on equators

Let $SO(Q+1)$ denote group of all unitary $(Q+1)\times(Q+1)$ matrices with determinant equal to 1. A *q-dimensional equator* of $\mathbb{S}^Q$ is a set of the form $\mathbb{Y} = \{\mathcal{R}u : u \in \mathbb{S}^q\}$ for some $\mathcal{R} \in SO(Q+1)$. The goal in the remainder of this section is to give approximation results for equators.

Since there exist infinite options for $\mathcal{R} \in \mathrm{SO}(Q+1)$ to generate the set $\mathbb{Y}$, we first give a definition of degree of approximation in terms of spherical polynomials that is invariant to the choice of $\mathcal{R}$.

Fix $\mathbb{Y}$ to be a given q-dimensional equator of $\mathbb{S}^Q$ and let $\mathcal{R}, \mathcal{S} \in \mathrm{SO}(Q+1)$ mapping $\mathbb{S}^q$ to $\mathbb{Y}$. Observe that if $P \in \Pi_n^q$, then $P(\mathcal{R}^T\mathcal{S}\circ) \in \Pi_n^q$ and vice versa. As a result, the functions $F_\mathcal{R} = f(\mathcal{R}\circ)$ and $F_\mathcal{S} = f(\mathcal{S}\circ)$ defined on $\mathbb{S}^q$ satisfy

$$E_n(F_\mathcal{R}) = E_n(F_\mathcal{S}). \tag{4.20}$$

Since the degree of approximation in this context is invariant to the choice of $\mathcal{R} \in \mathrm{SO}(Q+1)$, we may simply choose any such matrix that maps $\mathbb{S}^q$ to $\mathbb{Y}$, drop the subscript $\mathcal{R}$ from $F_\mathcal{R}$, and define

$$E_n(\mathbb{Y}, f) := E_n(F). \tag{4.21}$$

This allows us to define the space $W_\gamma(\mathbb{Y})$ as the class of all $f \in C(\mathbb{Y})$ such that

$$||f||_{W_\gamma(\mathbb{Y})} := ||f||_{\mathbb{Y}} + \sup_{n\geq 0} 2^{n\gamma} E_{2^n}(\mathbb{Y}, f) < \infty. \tag{4.22}$$

We can also define the approximation operator on the set $\mathbb{Y}$ as

$$\sigma_n(\mathbb{Y}, f)(x) := \int_{\mathbb{Y}} \Phi_{n,q}(x\cdot y) f(y) d\mu_{\mathbb{Y}}^*(y), \tag{4.23}$$

where $\mu^*_{\mathbb{Y}}(y)$ is the probability volume measure on $\mathbb{Y}$. Let $F_{\mathcal{R}} \in C(\mathbb{S}^q)$ satisfy $F_{\mathcal{R}} = f \circ \mathcal{R}$. We observe that

$$\begin{aligned}\sigma_n(\mathbb{Y}, f)(x) &= \int_{\mathbb{S}^q} \Phi_{n,q}(x \cdot \mathcal{R}u) f(\mathcal{R}u) d\mu^*_q(u) \\ &= \int_{\mathbb{S}^q} \Phi_{n,q}(\mathcal{R}^T x \cdot u) F_{\mathcal{R}}(u) d\mu^*_q(u) \\ &= \sigma_n(\mathbb{S}^q, F_{\mathcal{R}})(\mathcal{R}^T x).\end{aligned} \tag{4.24}$$

We now give an analogue of Proposition 4.2 for approximation on equators.

Theorem 4.1. *Let $f \in C(\mathbb{Y})$.*
(a) *We have*

$$E_n(\mathbb{Y}, f) \le ||\sigma_n(\mathbb{Y}, f) - f||_{\mathbb{Y}} \lesssim E_{n/2}(\mathbb{Y}, f). \tag{4.25}$$

(b) *If $\gamma > 0$, then $f \in W_\gamma(\mathbb{Y})$ if and only if*

$$||\sigma_n(\mathbb{Y}, f) - f||_{\mathbb{Y}} \lesssim n^{-\gamma} ||f||_{W_\gamma(\mathbb{Y})}. \tag{4.26}$$

Proof. Let $F(\circ) = f(\mathcal{R}\circ)$ for some $\mathcal{R} \in \mathrm{SO}(Q+1)$ with $\mathbb{Y} = \{\mathcal{R}u : u \in \mathbb{S}^q\}$. To see (4.25), we check using Proposition 4.2 that

$$||\sigma_n(\mathbb{Y}, f) - f||_{\mathbb{Y}} = \left|\left|\sigma_n(\mathbb{S}^q, F)(\mathcal{R}^T\circ) - F(\mathcal{R}^T\circ)\right|\right|_{\mathbb{Y}} = ||\sigma_n(\mathbb{S}^q, F) - F||_{\mathbb{S}^q} \lesssim E_{n/2}(F) = E_{n/2}(\mathbb{Y}, f). \tag{4.27}$$

Additionally, $E_n(\mathbb{Y}, f) \le ||\sigma_n(\mathbb{Y}, f) - f||_{\mathbb{Y}}$ since $\sigma_n(\mathbb{Y}, f) = \sigma_n(\mathbb{S}^q, F)(\mathcal{R}^T x) \in \Pi^q_n$. Part (b) can be seen from part (a) and the definitions. ■

5 Function approximation on manifolds

In this section, we develop the notion of *smoothness* for the target function defined on a manifold, and state our main theorem: Theorem 5.1. For a brief introduction to manifolds and some results we will be using in this paper, see Appendix A.2.

Let $Q \ge q \ge 1$ be integers and $\mathbb{X}$ be a q-dimensional, compact, connected, submanifold of $\mathbb{S}^Q$ without boundary. Let ρ denote the geodesic distance and μ^* be the normalized volume measure (that is, $\mu^*(\mathbb{X}) = 1$). For any $x \in \mathbb{X}$, observe that the tangent space $\mathbb{T}_x(\mathbb{X})$ is a q-dimensional vector space tangent to $\mathbb{S}^Q$. We define $\mathbb{S}_x = \mathbb{S}_x(\mathbb{X})$ to be the q-dimensional equator of $\mathbb{S}^Q$ passing through x whose own tangent space at x is also $\mathbb{T}_x(\mathbb{X})$. As an important note, $\mathbb{S}_x$ is also a q-dimensional compact manifold.

In this paper we will consider many spaces, and need to define balls on each of these spaces, which we list in Table 1 below.

Space	Description	Definition
Ambient space	Euclidean ball	$B_{Q+1}(x,r) = \{y \in \mathbb{R}^{Q+1} : \|\|x-y\|\|_2 \le r\}$
Ambient sphere	Spherical cap	$S^Q(x,r) = \{y \in \mathbb{S}^Q : \arccos(x \cdot y) \le r\}$
Tangent space	Tangent ball	$B_{\mathbb{T}}(x,r) = \{y \in \mathbb{T}_x(\mathbb{X}) : \|\|x-y\|\|_2 \le r\}$
Tangent sphere	Tangent cap	$\mathbb{S}_x(r) = \{y \in \mathbb{S}_x : \arccos(x \cdot y) \le r\}$
Manifold	Geodesic ball	$\mathbb{B}(x,r) = \{y \in \mathbb{X} : \rho(x,y) \le r\}$

Table 1: Definition and description of balls in different spaces.

We also need to define the smoothness classes we will be considering for functions on $\mathbb{X}$. Let $C(\mathbb{X})$ denote the space of all continuous functions on $\mathbb{X}$, and $C^\infty(\mathbb{X}) \subset C(\mathbb{X})$ denote the space of all infinitely differentiable functions on $\mathbb{X}$. Let $\bar{\varepsilon}_x$ be the exponential map at x for $\mathbb{S}_x$ and ε_x be the exponential map at x for $\mathbb{X}$. Since both $\mathbb{X}$ and $\mathbb{S}_x$ are compact, we have some ι_1, ι_2 such that $\varepsilon_x, \bar{\varepsilon}_x$ are defined on $B_{\mathbb{T}}(x, \iota_1), B_{\mathbb{T}}(x, \iota_2)$ respectively for any x. We write $\iota^* = \min\{1, \iota_1, \iota_2\}$ and define $\eta_x : \mathbb{S}_x(\iota^*) \to \mathbb{X}$ by $\eta_x : \varepsilon_x \circ \bar{\varepsilon}_x^{-1}$. Thus,

$$\rho(x, \eta_x(y)) = \arccos(x \cdot y), \qquad x \in \mathbb{X},\ y \in \mathbb{S}_x(\iota^*). \tag{5.1}$$

Definition 5.1. We say that $f \in C(\mathbb{X})$ is **γ-smooth** for some $\gamma > 0$, or also that $f \in W_\gamma(\mathbb{X})$, if for every $x \in \mathbb{X}$ and $\phi \in C^\infty(\mathbb{X})$ supported on $\mathbb{B}(x, \iota^*)$, the function $F_{x,\phi} : \mathbb{S}_x \to \mathbb{R}$ defined by $F_{x,\phi} \coloneqq f(\eta_x(u))\phi(\eta_x(u))$ belongs to $W_\gamma(\mathbb{S}_x)$ as outlined in Section 4.3 (in particular, Equation (4.22)). We also define

$$||f||_{W_\gamma(\mathbb{X})} \coloneqq \sup_{x\in\mathbb{X}, ||\phi||_{W_\gamma(\mathbb{S}_x)}\leq 1} ||F_{x,\phi}||_{W_\gamma(\mathbb{S}_x)}. \tag{5.2}$$

Our main theorem, describing the approximation of ff_0 (the target function weighted by the density of data points) by the operator defined in (1.1), is the following. We note that approximation of ff_0 includes local approximation on $\mathbb{X}$ in the sense that when the training data is sampled only from a subset of $\mathbb{X}$, this fact can be encoded by f_0 being supported on this subset.

Theorem 5.1. *We assume that*

$$\sup_{x\in\mathbb{X}, r>0} \frac{\mu^*(\mathbb{B}(x,r))}{r^q} \lesssim 1. \tag{5.3}$$

Let $\mathcal{D} = \{(y_j, z_j)\}_{j=1}^M$ be a random sample from a joint distribution τ. We assume that the marginal distribution of τ restricted to $\mathbb{X}$ is absolutely continuous with respect to μ^* with density f_0, and that the random variable z has a bounded range. We say $z \in [-||z||, ||z||]$. Let

$$f(y) \coloneqq \mathbb{E}_\tau(z|y), \tag{5.4}$$

and

$$F_n(\mathcal{D}; x) \coloneqq \frac{1}{M}\sum_{j=1}^M z_j \Phi_{n,q}(x \cdot y_j), \qquad x \in \mathbb{S}^Q, \tag{5.5}$$

where $\Phi_{n,q}$ is defined in (4.12).

Let $0 < \gamma < 2$ and $ff_0 \in W_\gamma(\mathbb{X})$. Then for every $n \geq 1$, $0 < \delta < 1/2$ and

$$M \gtrsim n^{q+2\gamma} \log(n/\delta), \tag{5.6}$$

we have with τ-probability $\geq 1 - \delta$:

$$||F_n(\mathcal{D}; \circ) - ff_0||_{\mathbb{X}} \lesssim \frac{\sqrt{||f_0||_{\mathbb{X}}}||z|| + ||ff_0||_{W_\gamma(\mathbb{X})}}{n^\gamma}. \tag{5.7}$$

Equivalently, for integer $M \geq 2$ and n satisfying (5.6), we have with τ-probability $\geq 1 - \delta$:

$$||F_n(\mathcal{D}; \circ) - ff_0||_{\mathbb{X}} \lesssim \left\{\sqrt{||f_0||_{\mathbb{X}}}||z|| + ||ff_0||_{W_\gamma(\mathbb{X})}\right\} \left(\frac{\log(M/\delta^{q+2\gamma})}{M}\right)^{\gamma/(q+2\gamma)}. \tag{5.8}$$

We discuss two corollaries of this theorem, which demonstrate how the theorem can be used for both estimation of the probability density f_0 and the approximation of the function f in the case when the training data is sampled from the volume measure on $\mathbb{X}$.

The first corollary is a result on function approximation in the case when the marginal distribution of y is μ^*; i.e., $f_0 \equiv 1$.

Corollary 5.1. *Assume the setup of Theorem 5.1. Suppose also that the marginal distribution of τ restricted to $\mathbb{X}$ is uniform. Then we have with τ-probability $\geq 1 - \delta$:*

$$||F_n(\mathcal{D}; \circ) - f||_{\mathbb{X}} \lesssim \frac{||z|| + ||f||_{W_\gamma(\mathbb{X})}}{n^\gamma}. \tag{5.9}$$

The second corollary is obtained by setting $f \equiv 1$, to point out that our theorem gives a method for density estimation. In practice, one may not have knowledge of f_0 (or even the manifold $\mathbb{X}$). So, the following corollary can be applied to estimate this critical quantity. We use this fact in our numerical examples in Section 6.

Typically, a positive kernel is used for the problem of density estimation in order to ensure that the approximation is also a positive measure. It is well known in approximation theory that this results in a saturation for the rate of convergence. Our method does not use positive kernels, and does not suffer from such saturation.

Corollary 5.2. *Assume the setup of Theorem 5.1. Then we have with τ-probability $\geq 1 - \delta$:*

$$\left|\left|\left|\frac{1}{M}\sum_{j=1}^M \Phi_{n,q}(\circ \cdot y_j)\right| - f_0\right|\right|_{\mathbb{X}} \lesssim \frac{||f_0||_{W_\gamma(\mathbb{X})}}{n^\gamma}. \tag{5.10}$$

6 Numerical examples

In this section, we illustrate our theory with some numerical experiments. In Section 6.1, we consider the approximation of a piecewise differentiable function, and demonstrate how the localization of the kernel leads to a determination of the locations of the singularities. The example in Section 6.2 is motivated by magnetic resonance relaxometry. Since it is relevant to our method for practical applications, we have included some discussion and results about how q effects the approximation in this example. The example in Section 6.3 illustrates how our method can be used for inverse problems in the realm of differential equations. In all the examples, we will examine how the accuracy of the approximation depends on the maximal degree n of the polynomial, the number M of samples, and the level of noise.

6.1 Piecewise differentiable function

In this example only we define the function to be approximated as

$$f(\theta) := 1 + |\cos\theta|^{1/2} \sin(\cos\theta + \sin\theta)/2, \tag{6.1}$$

defined on the ellipse

$$E = \{(3\cos\theta, 6\sin\theta) : \theta \in (-\pi, \pi]\}. \tag{6.2}$$

We project E to the sphere $\mathbb{S}^2$ using an inverse stereographic projection defined by

$$\mathbf{P}(\mathbf{x}) = \frac{(\mathbf{x}, 1)}{||(\mathbf{x}, 1)||_2}, \tag{6.3}$$

and call $\mathbb{X} = \mathbf{P}(E)$. Each $\mathbf{x} \in \mathbb{X}$ is associated with the value $\theta_{\mathbf{x}}$ satisfying $\mathbf{x} = \mathbf{P}\big((3\cos\theta_{\mathbf{x}}, 6\sin\theta_{\mathbf{x}})\big)$, so that $f(\mathbf{x}) := f(\theta_{\mathbf{x}})$ is a continuous function on $\mathbb{X}$.

We generate our data points by taking $\mathbf{y}_j = \mathbf{P}\big((3\cos\theta_j, 6\sin\theta_j)\big)$, where θ_j are each sampled uniformly at random from $(-\pi, \pi]$. We then define $z_j = f(\mathbf{y}_j) + \epsilon_j$ where ϵ_j are sampled from some mean-zero normal noise. Our data set is thus $\mathcal{D} := \{(\mathbf{y}_j, f(\mathbf{y}_j) + \epsilon_j)\}_{j=1}^M$. We will measure the magnitude of noise using the signal-to-noise ratio (SNR), defined by

$$20\log_{10}\Big(\big|\big|\big(f(\mathbf{y}_1), \ldots, f(\mathbf{y}_M)\big)\big|\big|_2 \Big/ \big|\big|\big(\epsilon_1, \ldots, \epsilon_M\big)\big|\big|_2\Big). \tag{6.4}$$

Since $f_0 \neq 1$ in this case, we could calculate f_0 from the projection, or we may estimate it using Corollary 5.2. That is,

$$f_0(\mathbf{x}) \approx \frac{1}{M}\sum_{j=1}^M \Phi_{n,1}(\mathbf{x}\cdot\mathbf{y}_j). \tag{6.5}$$

This option may be desirable in cases where f_0 is not feasible to compute (i.e. if the underlying domain of the data is unknown or irregularly shaped). Our approximation is thus:

$$F_n(\mathcal{D};\mathbf{x}) = \sum_{j=1}^M (f(\mathbf{y}_j) + \epsilon_j)\Phi_{n,1}(\mathbf{x}\cdot\mathbf{y}_j) \Bigg/ \left(\sum_{j=1}^M \Phi_{n,1}(\mathbf{x}\cdot\mathbf{y}_j)\right). \tag{6.6}$$

Figure 3 shows a plot of the true function and our approximation on the left y-axis and the absolute error on the right y-axis. The plot demonstrates that the approximation achieves much lower error than the uniform error bound at points where the function is relatively smooth, and only spikes locally at the singularities of the function ($\theta = \pm\pi/2$). Figure 4 displays three percent point plots illustrating how the distribution of $\log_{10}|F_n - f|$ behaves for various choices of n, M, ϵ. Each point (x, y) on a curve indicates that x% of test points were approximated by our method with absolute error below 10^y for the n, M, and ϵ value associated with the curve. The first graph shows the trend for various n values. As we increase n, we see consistent drop in log error. The second graph shows various values of M. We again see a decrease in the overall log error as M is increased. The third graph shows how the log error decreases as the noise decreases. We can see that the approximation is much worse for low SNR values, but nearly indistinguishable from the noiseless case when the SNR is above 60.

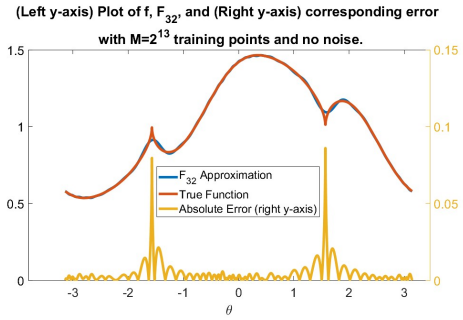

Figure 3: Left y-axis: Plot of the true function f compared with F_{32} constructed by 2^{13} noiseless training points. Right y-axis: Plot of $|f - F_{32}|$.

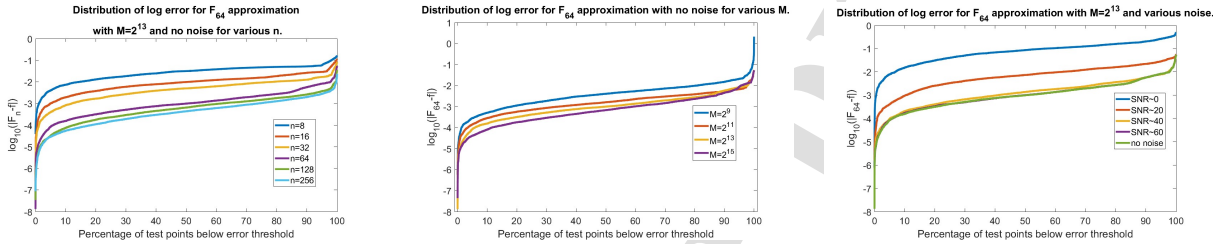

Figure 4: (Left) Percent point plot of log absolute error for various n with $M = 2^{13}$ training points and no noise. (Center) Percent point plot of log absolute error for various choices of M with no noise. (Right) Percent point plot of log absolute error for various noise levels with $M = 2^{13}$ training points.

6.2 Parameter estimation in bi-exponential sums

This example is motivated by magnetic resonance relaxometry, in which the proton nuclei of water are first excited with radio frequency pulses and then exhibit an exponentially decaying electromagnetic signal. When one may assume the presence of two water compartments undergoing slow exchange, with signal corrupted by additive Gaussian noise, the signal is modeled typically as a bi-exponential decay function (6.7) (cf. [37]):

$$F(t) = c_1 \exp(-t/T_{2,1}) + c_2 \exp(-t/T_{2,2}) + E(t),$$

where E is the noise, $T_{2,1}, T_{2,2} > 0$, and the time t is typically sampled at equal intervals. The problem is to determine $c_1, c_2, T_{2,1}, T_{2,2}$. The problem appears also in many other medical applications, such as intravoxel incoherent motion studies in magnetic resonance. An accessible survey of these applications is given in [18].

Writing $t = j\delta$, $\lambda_1 = \delta/T_{2,1}$, $\lambda_2 = \delta/T_{2,2}$, we may reformulate the data as

$$f(j) \coloneqq f(\lambda_1, \lambda_2, j) = c_1 e^{-\lambda_1 j} + c_2 e^{-\lambda_2 j} + \epsilon(j), \tag{6.7}$$

where $\epsilon(j)$ are samples of mean-zero normal noise.

In this example, suggested by Dr. Spencer at the National Institute of Aging (NIH), we consider the case where $c_1 = .7, c_2 = .3$ and use our method to determine the values λ_1, λ_2, given data of the form

$$\tilde{\mathbf{y}}(\lambda_1, \lambda_2) \coloneqq (f(1), f(2), \ldots, f(100)). \tag{6.8}$$

We "train" our approximation process with M samples of $(\lambda_1, \lambda_2) \in [.1, .7] \times [1.1, 1.7]$ chosen uniformly at random and then plugging those values into (6.7) to generate vectors of the form shown in (6.8). The dimension of the input data is $Q = 100$, however (in the noiseless case) the data lies on a $q = 2$ dimensional manifold, so we will use $\Phi_{n,2}$ to generate our approximations.

We note that our method is agnostic to the particular model (6.8) used to generate the data. We treat λ_1, λ_2 as functions of $\tilde{\mathbf{y}}$ without a prior knowledge of this function. In the noisy case, this problem does not perfectly fit the theory studied in this paper since the noise is applied to the input values $f(t)$ meaning we cannot assume they lie directly on an unknown manifold anymore. Nevertheless, we can see some success with our method. We define the operators

$$\mathbf{T}(\tilde{\mathbf{y}}) = 1000\tilde{\mathbf{y}} - (380, 189, 116, 0, \ldots, 0), \qquad \mathbf{P}(\circ) = \frac{(\circ, 100)}{||(\circ, 100)||_2} \tag{6.9}$$

and denote $\mathbf{y} = \mathbf{P}(\mathbf{T}(\tilde{\mathbf{y}}))$. In practice, the values used to define $\mathbf{T}$ and $\mathbf{P}$ need to be treated as hyperparameters of the model. In this example, we did not conduct a rigorous grid search. We use the same density estimation as in Section 6.1:

$$\mathrm{DE}\left(\mathbf{x}(\lambda_1, \lambda_2)\right) = \sum_{j=1}^{M} \Phi_{n,2}\big(\mathbf{x}(\lambda_1, \lambda_2) \cdot \mathbf{y}(\lambda_{1,j}, \lambda_{2,j})\big). \tag{6.10}$$

As a result, our approximation process looks like:

$$\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix} \approx \mathbf{F}_n\big(\mathbf{x}(\lambda_1, \lambda_2)\big) = \sum_{j=1}^{M} \begin{bmatrix}\lambda_{1,j} \\ \lambda_{2,j}\end{bmatrix} \Phi_{n,2}\big(\mathbf{x}(\lambda_1, \lambda_2) \cdot \mathbf{y}(\lambda_{1,j}, \lambda_{2,j})\big) \Big/ \mathrm{DE}\left(\mathbf{x}(\lambda_1, \lambda_2)\right). \tag{6.11}$$

Similar to Example 6.1, we will include figures showing how the results are effected as n, M, ϵ are adjusted. We measure noise using the signal-to-noise ratio (SNR) defined by

$$20 \log_{10}\left(||\tilde{\mathbf{y}}||_2 \Big/ \big|\big|\big(\epsilon(1), \ldots, \epsilon(100)\big)\big|\big|_2\right). \tag{6.12}$$

Unlike Example 6.1, we will now be considering percent approximation error instead of uniform error as it is more relevant in this problem. We define the *combined error* to be

$$\sum_{j=1}^{2} \frac{|\lambda_{j,\mathrm{true}} - \lambda_{j,\mathrm{approx}}|}{\lambda_{j,\mathrm{true}}}. \tag{6.13}$$

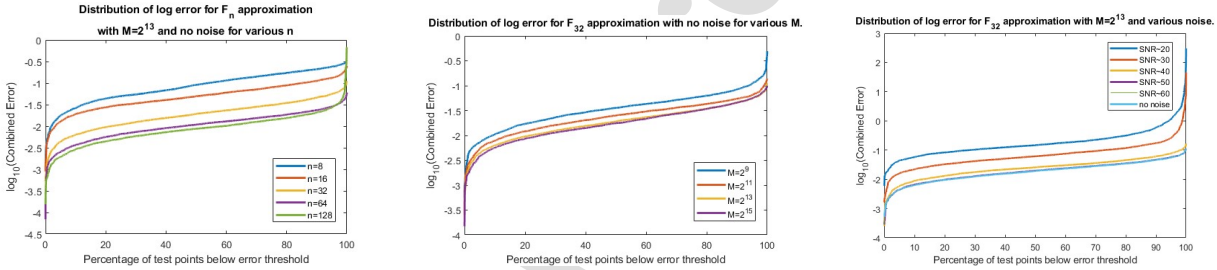

Figure 5: (Left) Percent point plot of log combined error for various n with $M = 2^{13}$ training points, and no noise. (Center) Percent point plot of log combined error for fixed $n = 32$, various choices of M, and no noise. (Right) Percent point plot of log combined error for fixed $n = 32$, fixed $M = 2^{13}$ training points, and various noise levels.

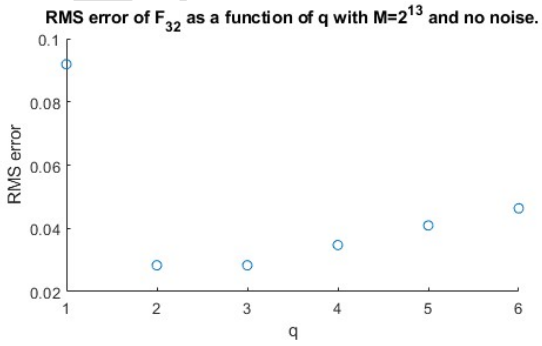

Figure 6: Plot of RMS error for approximation by F_{32} for various q values with $M = 2^{13}$ and no noise.

Figure 5 depicts three percent point plots showing the distribution of sorted $\log_{10}$(Combined Error) points for various n, M, ϵ. Each point (x, y) on a curve indicates that x% of test points were approximated by our method with combined error below 10^y for the n, M, and ϵ associated with the curve. In the first graph, we see the distribution of for various choices of n. As n increases, the overall log error decreases. An interesting phenomenon occurring in this figure is with the $n = 128$ case where the uniform error is actually higher than the $n = 64$ case. This is likely

due to the fact that overfitting can occur if n gets too large relative to a fixed M. The second graph illustrates how the approximation improves as M is increased. As expected, we see log error decay as we include more and more training points. In the third graph, we see that the approximation improves up to a limit as the noise decreases. There is very little noticeable difference between the noiseless case and any case where SNR>50.

Another question that may arise when utilizing our method on various data is what value of q to use. While the theory predicts that q should be associated with the intrinsic dimension of the manifold underlying the data, in practice this can only be estimated and so q should be treated as a hyperparameter. In Figure 6, we explore how changing q effects the approximation in this example. In this case, the intrinsic dimension is 2, and when $q = 2, 3$ the approximation does well. If q is chosen too high or two low, the approximation yields a greater error.

6.3 Darcy flow problem

In this section we will look at a numerical example from the realm of PDE inverse problems. Steady-state Darcy flow is given by the following PDE (see for example, [35, Eq. (4.7)]):

$$-\nabla \cdot (a \nabla y) = f, \tag{6.14}$$

defined on a domain D with the property that $y|_{\partial D} = 0$. The problem is to predict the *diffusion coefficient* a and *forcing term* f given some noisy samples of y on D. In this paper we consider a 1-dimensional version and suppose that $a = e^{-st}$ and $f = pe^{-st}$ for some p, s. We take noisy samples of $y(p, s; \circ) = y$ satisfying the following boundary value problem:

$$-\frac{d}{dt}(e^{-st}y'(t)) = pe^{-st}, \qquad y(1) = 0, y(0) = 1. \tag{6.15}$$

In this sample, we take a similar approach to that of Example 6.2 by "training" our model with a data set of the form $\{\mathbf{y}_j, (p_j, s_j)\}_{j=1}^M$, where $(p_j, s_j) \in [.1, .25] \times [1.5, 2.5]$ are sampled uniformly at random for each j. Letting y_j denote the y satisfying (6.15) with $p = p_j, s = s_j$, then $\mathbf{y}_j = \mathbf{P}(y_j(t_1), y_j(t_2), \ldots, y_j(t_{100}))$, where $t_1, t_2, \ldots, t_{100}$ are sampled uniformly from $[0, 1]$ and $\mathbf{P}$ is the projection to the sphere. In this example, the projection first consists of finding the center C and maximum spread over a single feature r of the data. That is,

$$\begin{aligned} C &= \left(\max_j y_j(t_1) + \min_j y_j(t_1), \ldots, \max_j y_j(t_{100}) + \min_j y_j(t_{100}) \right) \Big/ 2, \\ r &= \max_{t_i} \left(\max_j y_j(t_1) - \min_j y_j(t_1), \ldots, \max_j y_j(t_{100}) - \min_j y_j(t_{100}) \right). \end{aligned} \tag{6.16}$$

Then, we define

$$\mathbf{P}(\circ) = \frac{(\circ - C, r)}{||(\circ - C, r)||_2}. \tag{6.17}$$

Our approximation process then looks like:

$$\begin{bmatrix} p \\ s \end{bmatrix} \approx \mathbf{F}_n(\mathbf{y}) = \sum_{j=1}^{M} \begin{bmatrix} p_j \\ s_j \end{bmatrix} \Phi_{n,2}(\mathbf{y} \cdot \mathbf{y}_j) \Big/ \mathrm{DE}(\mathbf{y}), \tag{6.18}$$

where

$$\mathrm{DE}(\mathbf{y}) = \sum_{j=1}^{M} \Phi_{n,2}(\mathbf{y} \cdot \mathbf{y}_j). \tag{6.19}$$

Also similar to Example 6.2, we use the same notion of SNR and evaluate the success of our model using a *combined error*, now defined to be

$$\left(\left| \frac{p_{\text{true}} - p_{\text{approx}}}{p_{\text{true}}} \right| + \left| \frac{s_{\text{true}} - s_{\text{approx}}}{s_{\text{true}}} \right| \right). \tag{6.20}$$

In Figure 7, we provide some percent point plots from using our method on this data. Each point (x, y) on a curve indicates that $x\%$ of test points were approximated by our method with combined error below 10^y for the n, M, and ϵ associated with the curve. We see in the left-most plot that as we increase n, the error tends to decrease. In contrast to previous examples, the middle plot does not show much improvement by increasing M. This may be an indication of the fact that we have chosen a tight parameter space in this example (as compared to 6.2) and not many samples are needed to sufficiently cover the space. On the right-most plot, we see a decrease in error with the decrease of noise as expected, with convergence appearing to occur around the SNR=70 mark, as indicated by the green and light-blue lines being so close together.

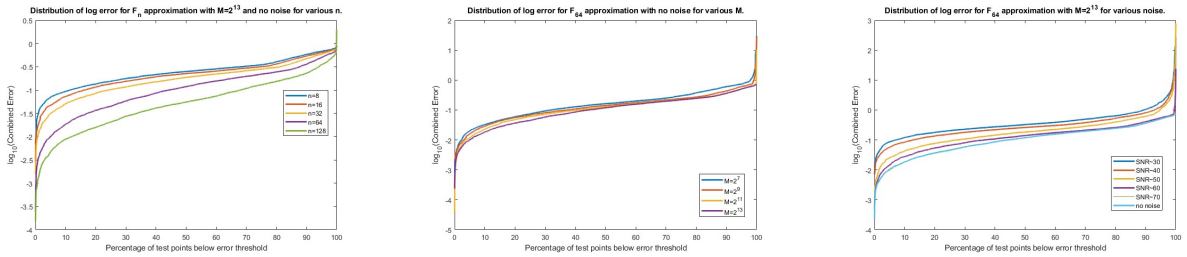

Figure 7: (Left) Percent point plot of log combined error for various n, fixed $M = 2^{13}$, and no noise. (Center) Percent point plot of log combined error for fixed $n = 64$, various M, and no noise. (Right) Percent point plot of log combined error for fixed $n = 64$, fixed $M = 2^{13}$, and various noise levels.

7 Proofs

The purpose of this section is to prove Theorem 5.1.

In Section 7.1, we study the approximation properties of the integral reconstruction operator defined in (3.3) (Theorem 7.1). In Section 7.2, we use this theorem with ff_0 in place of f, and use the Bernstein concentration inequality (Proposition 7.1) to discretize the integral expression in (3.3) and complete the proof of Theorem 5.1.

7.1 Integral reconstruction operator

In this section, we prove the following theorem which is an integral analogue of Theorem 5.1.

Theorem 7.1. *Let $0 < \gamma < 2$, $f \in W_\gamma(\mathbb{X})$, σ_n be as defined in (3.3). Then for $n \geq 1$, we have*

$$||f - \sigma_n(\mathbb{X}, f)||_{\mathbb{X}} \lesssim n^{-\gamma} ||f||_{W_\gamma(\mathbb{X})}. \tag{7.1}$$

In order to prove this theorem, we will use a covering of $\mathbb{X}$ using balls of radius ι^*, and a corresponding partition of unity. A key lemma to facilitate the details here is the following.

Lemma 7.1. *Let $x \in \mathbb{X}$. Let $g \in C(\mathbb{X})$ be supported on $\mathbb{B}(x, \iota^*)$. If $G(u) = g(\eta_x(u))$, $0 < \gamma < 2$. Then*

$$\left| \int_{\mathbb{X}} \Phi_{n,q}(x \cdot y) g(y) d\mu^*(y) - \int_{\mathbb{S}_x} \Phi_{n,q}(x \cdot u) G(u) d\mu^*_{\mathbb{S}_x}(u) \right| \lesssim n^{-\gamma} ||g||_{\mathbb{X}}. \tag{7.2}$$

If $\phi \in C^\infty(\mathbb{X})$ is supported on $\mathbb{B}(x, \iota^*)$, then we may apply this theorem with $g = f\phi$, thereby providing locally a lifting of the integral on $\mathbb{X}$ to the tangent equator $\mathbb{S}_x$ with the function corresponding to g on this tangent equator.

Naturally, the first step in this proof is to show that the Lebesgue constant for the kernel $\Phi_{n,q}$ is bounded independently of n (cf. (7.4)). Moreover, one can even leverage the localization of the kernel to improve on this bound when the integral is taken away from the point x (cf. (7.3)). These are both done in the following lemma.

Lemma 7.2. *Let $r > 0$ and $n \geq 1/r$. If $\Phi_{n,q}$ is given as in (4.12) with $S > q$, then*

$$\sup_{x \in \mathbb{X}} \int_{\mathbb{X} \setminus \mathbb{B}(x,r)} |\Phi_{n,q}(x \cdot y)| \, d\mu^*(y) \lesssim \max(1, nr)^{q-S}. \tag{7.3}$$

Additionally,

$$\sup_{x \in \mathbb{X}} \int_{\mathbb{X}} |\Phi_{n,q}(x \cdot y)| d\mu^*(y) \lesssim 1. \tag{7.4}$$

Proof. Recall from Proposition A.2 that $\rho(x, y) \sim \arccos(x \cdot y)$, so (4.13) implies

$$|\Phi_{n,q}(x \cdot y)| \lesssim \frac{n^q}{\max(1, n \arccos(x \cdot y))^S} \lesssim \frac{n^q}{\max(1, n\rho(x, y))^S}. \tag{7.5}$$

In this proof only, we fix $x \in \mathbb{X}$. Let $A_0 = \mathbb{B}(x, r)$ and $A_k = \mathbb{B}(x, 2^k r) \setminus \mathbb{B}(x, 2^{k-1} r)$, $k \geq 1$. Then $\mu^*(A_k) \lesssim 2^k q r^q$, and for any $y \in A_k$, $2^{k-1} r \leq \rho(x, y) \leq 2^k r$.

First, let $nr \geq 1$. In view of (5.3) and (7.5), it follows that

$$\begin{aligned}\int_{\mathbb{X}\setminus\mathbb{B}(x,r)} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) =& \sum_{k=1}^{\infty} \int_{A_k} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) \lesssim \sum_{k=1}^{\infty} \frac{\mu^*(A_k) n^q}{(n2^{k-1}r)^S} \\ \lesssim& (nr)^{q-S} \sum_{k=0}^{\infty} 2^{k(q-S)} \leq (nr)^{q-S}.\end{aligned} \tag{7.6}$$

Using this estimate with $r = 1/n$ and the condition (5.3) on the measures of balls we see that

$$\int_{\mathbb{X}} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) = \int_{A_0} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) + \int_{\mathbb{X}\setminus\mathbb{B}(x,r)} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) \lesssim 1 + (nr)^{q-S} \sim 1.$$

Since the choice of x was arbitrary, we have proven (7.4). Then (7.4) and (7.6) combined give the bounds for (7.3). ∎

Next, we prove Lemma 7.1.

Proof of Lemma 7.1.

Since $\gamma < 2$, we may choose (for sufficiently large n)

$$\delta = n^{-(\gamma+q+1)/(q+3)}, \quad n\delta = n^{(2-\gamma)/(q+3)} > 1, \quad S > \frac{2q+3\gamma}{2-\gamma}. \tag{7.7}$$

We may assume further that $\delta < \iota^*$. Then, by using (4.14) and Proposition A.2, we see that

$$\begin{aligned}\left|\Phi_{n,q}(x\cdot\eta_x(u)) - \Phi_{n,q}(x\cdot u)\right| \lesssim& n^{q+1} \left|\arccos(x\cdot\eta_x(u)) - \arccos(x\cdot u)\right| \\ =& n^{q+1} \left|\arccos(x\cdot\eta_x(u)) - \rho(x,\eta_x(u))\right| \lesssim n^{q+1}\rho(x,\eta_x(u))^3 \lesssim n^{q+1}\delta^3,\end{aligned} \tag{7.8}$$

for any $u \in \mathbb{S}_x(\delta)$. Let $\mathbf{g}_1, \mathbf{g}_2$ be the metric tensors associated with the exponential maps $\varepsilon_x : \mathbb{T}_x(\mathbb{X}) \to \mathbb{X}$ and $\bar{\varepsilon}_x : \mathbb{T}_x(\mathbb{X}) \to \mathbb{S}_x$, respectively. Then we have the following change of variables formulas (cf. Table 1):

$$\int_{\mathbb{B}(x,\delta)} d\mu^*(\varepsilon_x(v)) = \int_{B_{\mathbb{T}}(x,\delta)} \sqrt{|\mathbf{g}_1|}dv, \qquad \int_{\mathbb{S}_x(\delta)} d\mu_q^*(u) = \int_{B_{\mathbb{T}}(x,\delta)} \sqrt{|\mathbf{g}_2|}d\bar{\varepsilon}_x^{-1}(u). \tag{7.9}$$

We set $v = \bar{\varepsilon}_x^{-1}(u)$ and use the fact (cf. (A.10)) that on $B_{\mathbb{T}}(x,\delta)$, $|\sqrt{|\mathbf{g}_1|} - 1| \lesssim \delta^2$ and $|\sqrt{|\mathbf{g}_2|} - 1| \lesssim \delta^2$. Then by applying Equations (7.8), (7.9), (4.13), and (4.14), we can deduce

$$\begin{aligned}&\left|\int_{\mathbb{B}(x,\delta)} \Phi_{n,q}(x\cdot y) g(y) d\mu^*(y) - \int_{\mathbb{S}_x(\delta)} \Phi_{n,q}(x\cdot u) G(u) d\mu^*_{\mathbb{S}_x}(u)\right| \\ &\quad\leq \left|\int_{B_{\mathbb{T}}(x,\delta)} \Phi_{n,q}(x\cdot\varepsilon_x(v)) g(\varepsilon_x(v))(\sqrt{|\mathbf{g}_1|} - \sqrt{|\mathbf{g}_2|})dv\right| \\ &\qquad + \left|\int_{\mathbb{S}_x(\delta)} (\Phi_{n,q}(x\cdot\eta_x(u)) - \Phi_{n,q}(x\cdot u)) G(u) d\mu^*_{\mathbb{S}_x}(u)\right| \\ &\quad\lesssim ||g||_{\mathbb{X}} (\delta^{q+2} n^q + \delta^{q+3} n^{q+1}) \leq \delta^{q+3} n^{q+1} ||g||_{\mathbb{X}} (1/(n\delta) + 1) \lesssim ||g||_{\mathbb{X}} n^{-\gamma}.\end{aligned} \tag{7.10}$$

Now it only remains to examine the terms away from $\mathbb{S}_x(\delta), \mathbb{B}(x,\delta)$. Utilizing Lemma 7.2, and the fact that $S \geq \frac{2q+3\gamma}{2-\gamma}$, we have

$$\left|\int_{\mathbb{X}\setminus\mathbb{B}(x,\delta)} \Phi_{n,q}(x\cdot y) g(y) d\mu^*(y)\right| \lesssim ||g||_{\mathbb{X}} (n\delta)^{q-S} = ||g||_{\mathbb{X}} n^{(q-S)(2-\gamma)/(q+3)} \lesssim ||g||_{\mathbb{X}} n^{-\gamma}. \tag{7.11}$$

Similarly, again using Lemma 7.2 (with $\mathbb{S}_x$ as the manifold) and observing $||g||_{\mathbb{X}} = ||G||_{\mathbb{S}_x}$, we can conclude

$$\left|\int_{\mathbb{S}_x\setminus\mathbb{S}_x(\delta)} \Phi_{n,q}(x\cdot u) G(u) d\mu^*_{\mathbb{S}_x}(u)\right| \lesssim ||G||_{\mathbb{S}_x} (n\delta)^{q-S} \lesssim ||g||_{\mathbb{X}} n^{-\gamma}, \tag{7.12}$$

completing the proof. ∎

We are now in a position to complete the proof of Theorem 7.1.

Proof of Theorem 7.1.

Let $x \in \mathbb{X}$. Choose $\phi \in C^\infty$ such that $0 \le \phi(y) \le 1$ for all $y \in \mathbb{X}$, $\phi(y) = 1$ on $\mathbb{B}(x, \iota^*/2)$, and $\phi(y) = 0$ on $\mathbb{X}\backslash\mathbb{B}(x, \iota^*)$. Then $f\phi$ is supported on $\mathbb{B}(x, \iota^*)$ and $F(u) := \phi(\eta_x(u))f(\eta_x(u))$ belongs to $W_\gamma(\mathbb{S}_x)$. We observe that $||f||_{\mathbb{X}} \lesssim ||f||_{W_\gamma(\mathbb{X})}$. By Lemma 7.2,

$$\left|\int_{\mathbb{X}} \Phi_{n,q}(x\cdot y)f(y)(1-\phi(y))d\mu^*(y)\right| \le ||f||_{\mathbb{X}} \int_{\mathbb{X}\backslash\mathbb{B}(x,\iota^*/2)} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) \lesssim n^{-\gamma}\, ||f||_{W_\gamma(\mathbb{X})}\,. \tag{7.13}$$

Note that the constant above is chosen to account for the case where $n < 2/\iota^*$. By Lemma 7.1,

$$\left|\int_{\mathbb{X}} \Phi_{n,q}(x\cdot y)f(y)\phi(y)d\mu^*(y) - \sigma_n(\mathbb{S}_x, F)(x)\right| \lesssim n^{-\gamma}\, ||f\phi||_{\mathbb{X}} \lesssim n^{-\gamma}\, ||f||_{W_\gamma(\mathbb{X})}\,. \tag{7.14}$$

Observe that since $f(x) = F(x)$ and $||F||_{W_\gamma(\mathbb{S}_x)} \le ||f||_{W_\gamma(\mathbb{X})}$,

$$\begin{aligned}
&|f(x) - \sigma_n(\mathbb{X}, f)(x)| \\
&\quad\le |f(x) - F(x)| + |F(x) - \sigma_n(\mathbb{S}_x, F)(x)| + |\sigma_n(\mathbb{S}_x, F)(x) - \sigma_n(\mathbb{X}, f)(x)| \\
&\quad\lesssim 0 + n^{-\gamma}\, ||F||_{W_\gamma(\mathbb{S}_x)} + \left|\sigma_n(\mathbb{S}_x, F)(x) - \int_{\mathbb{X}} \Phi_{n,q}(x\cdot y)f(y)\phi(y)d\mu^*(y)\right| + \left|\int_{\mathbb{X}} \Phi_{n,q}(x\cdot y)f(y)(1-\phi(y))d\mu^*(y)\right| \\
&\quad\le n^{-\gamma}\, ||f||_{W_\gamma(\mathbb{X})}\,.
\end{aligned} \tag{7.15}$$

Since this bound is independent of x, the proof is completed. ∎

7.2 Discretization

In order to complete the proof of Theorem 5.1, we need to discretize the integral operator in Theorem 7.1 while keeping track of the error. If the manifold were known and we could use the eigendecomposition of the Laplace-Beltrami operator, we could do this discretization without losing the accuracy using quadrature formulas (cf., e.g., [30]). In our current set up, it is more natural to use concentration inequalities. We will use the inequality summarized in Proposition 7.1 below (c.f. [5]).

Proposition 7.1 (Bernstein concentration inequality)**.** *Let $Z_1, \cdots, Z_M$ be independent real valued random variables such that for each $j = 1, \ldots, M$, $|Z_j| \le R$, and $\mathbb{E}(Z_j^2) \le V$. Then for any $t > 0$,*

$$\text{Prob}\left(\left|\frac{1}{M}\sum_{j=1}^{M}(Z_j - \mathbb{E}(Z_j))\right| \ge t\right) \le 2\exp\left(-\frac{Mt^2}{2(V + Rt/3)}\right). \tag{7.16}$$

In the following, we will set $Z_j(x) = z_j\Phi_{n,q}(x\cdot y_j)$, where (y_j, z_j) are sampled from τ. The following lemma estimates the variance of Z_j.

Lemma 7.3. *With the setup from Theorem 5.1, we have*

$$\sup_{x\in\mathbb{X}} \int |z\Phi_{n,q}(x\cdot y)|^2\, d\tau(y,z) \lesssim n^q\, ||z||^2\, ||f_0||_{\mathbb{X}}\,, \qquad x \in \mathbb{S}^Q. \tag{7.17}$$

Proof.

We observe that (4.13) and Lemma 7.2 imply that

$$\sup_{x\in\mathbb{X}} \int_{\mathbb{X}} \Phi_{n,q}(x\cdot y)^2 d\mu^*(y) \lesssim n^q \sup_{x\in\mathbb{X}} \int_{\mathbb{X}} |\Phi_{n,q}(x\cdot y)|\, d\mu^*(y) \lesssim n^q. \tag{7.18}$$

Hence,

$$\sup_{x\in\mathbb{X}} \int |z(y,\epsilon)\Phi_{n,q}(x\cdot y)|^2\, d\tau(y,z) \le ||z||^2\, ||f_0||_{\mathbb{X}} \sup_{x\in\mathbb{X}} \int_{\mathbb{X}} \Phi_{n,q}(x\cdot y)^2 d\mu^*(y) \lesssim n^q\, ||z||^2\, ||f_0||_{\mathbb{X}}\,. \tag{7.19}$$

∎

A limitation of the Bernstein concentration inequality is that it only considers a single x value. Since we are interested in supremum-norm bounds, we must first relate the supremum norm of Z_j over all $x \in \mathbb{S}^Q$ to a finite set of points. We set up the connection in the following lemma.

Lemma 7.4. *Let ν be any (bounded variation) measure on $\mathbb{X}$. Then there exists a finite set $\mathcal{C}$ of size $|\mathcal{C}| \sim n^Q$ such that*

$$\left|\left|\int_{\mathbb{X}} \Phi_{n,q}(\circ \cdot y) d\nu(y)\right|\right|_{\mathbb{S}^Q} \leq 2 \max_{x \in \mathcal{C}} \left|\int_{\mathbb{X}} \Phi_{n,q}(x \cdot y) d\nu(y)\right|. \tag{7.20}$$

Proof.

In view of the Bernstein inequality for the derivatives of spherical polynomials, we see that

$$|P(x) - P(y)| \leq cn \, ||x - y||_\infty \, ||P||_\infty, \qquad P \in \Pi_n^Q. \tag{7.21}$$

We can see by construction that $\int_{\mathbb{X}} \Phi_{n,q}(t \cdot y) d\nu(y)$ is a polynomial of degree $< n$ in the variable t. Since $\mathbb{S}^Q$ is a compact space and polynomials of degree $< n$ are continuous functions, there exists some $x^* \in \mathbb{S}^Q$ such that

$$\left|\left|\int_{\mathbb{X}} \Phi_{n,q}(\circ \cdot y) d\nu(y)\right|\right|_{\mathbb{S}^Q} = \left|\int_{\mathbb{X}} \Phi_{n,q}(x^* \cdot y) d\nu(y)\right|. \tag{7.22}$$

Let c be the same as in (7.21) and $\mathcal{C}$ be a finite set satisfying

$$\max_{x \in \mathbb{S}^Q} \min_{y \in \mathcal{C}} ||x - y||_\infty \leq \frac{1}{2cn}. \tag{7.23}$$

Since $\mathbb{S}^Q$ is a compact Q-dimensional space, the set $\mathcal{C}$ needs no more than $\sim n^Q$ points.

Then there exists some $z^* \in \mathcal{C}$ such that

$$\left|\int_{\mathbb{X}} (\Phi_{n,q}(x^* \cdot y) - \Phi_{n,q}(z^* \cdot y)) d\nu(y)\right| \lesssim n \, ||x^* - z^*||_\infty \left|\int_{\mathbb{X}} \Phi_{n,q}(x^* \cdot y) d\nu(y)\right|, \tag{7.24}$$

which implies (7.20).

■

With this preparation, we now state the following theorem which gives a bound on the difference between our discrete approximation F_n and continuous approximation σ_n with high probability.

Theorem 7.2. *Assume the setup of Theorem 5.1. Then for every $n \geq 1$ and $M \gtrsim n^{q+2\gamma} \log(n/\delta)$ we have*

$$\text{Prob}_\tau \left(||F_n(\mathcal{D}; \circ) - \sigma_n(\mathbb{X}, f f_0)||_{\mathbb{S}^Q} \geq c \, ||z|| \, n^{-\gamma} \sqrt{||f_0||_{\mathbb{X}}} \right) \leq \delta. \tag{7.25}$$

Proof. In this proof only, constants $c, c_1, c_2, \ldots$ will maintain their value once used. Let $Z_j(x) = z_j \Phi_{n,q}(x \cdot y_j)$. Since z is integrable with respect to τ, one has the following for any $x \in \mathbb{S}^Q$:

$$\mathbb{E}_\tau(Z_j(x)) = \int_{\mathbb{X}} \mathbb{E}_\tau(z|y) \Phi_{n,q}(x \cdot y) d\nu^*(y) = \int_{\mathbb{X}} f(y) \Phi_{n,q}(x \cdot y) f_0(y) d\mu^*(y) = \sigma_n(\mathbb{X}, f f_0)(x). \tag{7.26}$$

We have from (4.13) that $|Z_j| \lesssim n^q \, ||z||$. Lemma 7.3 informs us that $\mathbb{E}_\tau(Z_j^2) \lesssim n^q \, ||z||^2 \, ||f_0||_{\mathbb{X}}$. Assume $0 < r \leq 1$ and set $t = r \, ||z|| \, ||f_0||_{\mathbb{X}}$. From Proposition 7.1, we see

$$\begin{aligned} \text{Prob}_\tau \left(\left| \frac{1}{M} \sum_{j=1}^M Z_j(x) - \sigma_n(\mathbb{X}, f f_0)(x) \right| \geq t \right) &\leq 2 \exp\left(-c_1 \frac{M t^2}{(n^q \, ||z||^2 \, ||f_0||_{\mathbb{X}} + n^q \, ||z|| \, t/3)} \right) \\ &\leq 2 \exp\left(-c_2 \frac{M \, ||f_0||_{\mathbb{X}} \, r^2}{n^q} \right). \end{aligned} \tag{7.27}$$

Let $\delta \in (0, 1/2)$, $\mathcal{C}$ be a finite set satisfying (7.23) with $|\mathcal{C}| \leq c_3 n^Q$ (without loss of generality we assume $c_3 \geq 1$),

$$c_4 \geq \frac{\max\left(\log_2(c_3) + 1, Q\right)}{c_2}, \tag{7.28}$$

and

$$M \geq c_4 n^{q+2\gamma} \log(n/\delta). \tag{7.29}$$

We now fix

$$r \equiv \sqrt{c_4 \frac{n^q}{M \, ||f_0||_{\mathbb{X}}} \log(n/\delta)}. \tag{7.30}$$

Notice that since $||f_0||_{\mathbb{X}} \geq 1$, our assumption of M in (7.29) implies

$$r \leq n^{-\gamma} \Big/ \sqrt{||f_0||_{\mathbb{X}}} \leq 1, \tag{7.31}$$

so our choice of r may be substituted into (7.27). Further,

$$r \, ||z|| \, ||f_0||_{\mathbb{X}} \leq c \, ||z|| \, n^{-\gamma} \sqrt{||f_0||_{\mathbb{X}}}. \tag{7.32}$$

With this preparation, we can conclude

$$\begin{aligned}
&\text{Prob}_\tau \left(\left\| F_n(\mathcal{D}; \circ) - \sigma_n(\mathbb{X}, f f_0) \right\|_{\mathbb{S}^Q} \geq c \, ||z|| \, n^{-\gamma} \sqrt{||f_0||_{\mathbb{X}}} \right) \\
&\quad \leq \text{Prob}_\tau \left(\left\| \frac{1}{M} \sum_{j=1}^{M} Z_j - \sigma_n(\mathbb{X}, f f_0) \right\|_{\mathbb{S}^Q} \geq r \, ||z|| \, ||f_0||_{\mathbb{X}} \right) && \text{(from (7.32))} \\
&\quad \leq \text{Prob}_\tau \left(\max_{x_k \in \mathcal{C}} \left(\left| \frac{1}{M} \sum_{j=1}^{M} Z_j(x_k) - \sigma_n(\mathbb{X}, f f_0)(x_k) \right| \right) \geq t \right) && \text{(by Lemma 7.4)} \\
&\quad \leq \sum_{k=1}^{|\mathcal{C}|} \text{Prob}_\tau \left(\left| \frac{1}{M} \sum_{j=1}^{M} Z_j(x_k) - \sigma_n(\mathbb{X}, f f_0)(x_k) \right| \geq t \right) && (7.33) \\
&\quad \leq |\mathcal{C}| \exp\left(-c_2 \frac{M \, ||f_0||_{\mathbb{X}} \, r^2}{n^q} \right) && \text{(from (7.27))} \\
&\quad \leq c_3 n^{Q - c_2 c_4} \delta^{c_2 c_4} && \text{(from (7.30))} \\
&\quad \leq c_3 n^{Q-Q} \left(\frac{1}{2} \right)^{\log_{1/2}(1/c_3)} \delta && \text{(from (7.28) and } \delta < 1/2) \\
&\quad \leq \delta.
\end{aligned}$$

■

We are now ready for the proof of Theorem 5.1.

Proof of Theorem 5.1 (and Corollary 5.1 and 5.2). Since $f, f_0 \in W_\gamma(\mathbb{X})$, we can determine that $f f_0 \in W_\gamma(\mathbb{X})$ as well. Utilizing Theorem 7.1 with $f f_0$ and Theorem 7.2, we obtain with probability at least $1 - \delta$ that

$$\begin{aligned}
||F_n(\mathcal{D}; \circ) - f f_0||_{\mathbb{X}} &\leq ||F_n(\mathcal{D}; \circ) - \sigma_n(\mathbb{X}; f f_0)||_{\mathbb{X}} + ||\sigma_n(\mathbb{X}; f f_0) - f f_0||_{\mathbb{X}} \\
&\lesssim \frac{\sqrt{||f_0||_{\mathbb{X}}} \, ||z|| + ||f f_0||_{W_\gamma(\mathbb{X})}}{n^\gamma}.
\end{aligned} \tag{7.34}$$

Corollary 5.1 is seen immediately by setting $f_0 = 1$. Corollary 5.2 follows from setting $z = 1$ and then observing that $f = 1$ and $\sqrt{||f_0||_{\mathbb{X}}} \lesssim ||f_0||_{W_\gamma(\mathbb{X})}$. ■

8 Conclusions

We have discussed a central problem of machine learning, namely to approximate an unknown target function based only on the data drawn from an unknown probability distribution. While the prevalent paradigm to solve this problem in general is to minimize a loss functional, we have initiated a new paradigm where we can do the approximation directly from the data, under the so-called manifold assumption. This is a substantial theoretical

improvement over the classical manifold learning technology, which involves a two-step procedure: first to get some information about the manifold and then to do the approximation. Our method is a "one-shot" method that bypasses collecting any information about the manifold itself: it learns on the manifold without manifold learning. Our construction in itself does not require any assumptions on the probability distribution or the target function. We derive uniform error bounds with high probability regardless of the nature of the distribution, provided we know the dimension of the unknown manifold. The theorems are illustrated with some numerical examples. One of these is closely related to an important problem in magnetic resonance relaxometry, in which one seeks to find the proportion of water molecules in the myelin covering in the brain based on a model that involves the inversion of Laplace transform.

We view our paper as the beginning of a new direction. As such, there are plenty of future research projects, some of which we plan to undertake ourselves.

- Find alternative methods that improve upon the error estimates on **unknown** manifolds, and more general compact sets. The encoding described in Section A.1 gives a representation of a function on an unknown manifold. Such an encoding is useful in the emerging area of approximation of operators. It is clear that the encoding described in Section A.1 for functions on manifolds itself forms a submanifold of a Euclidean space, which in turn can be projected to a submanifold of a sphere. We plan to develop this theme further in the context of approximation of operators defined in different function spaces.
- Explore real-life applications other than the examples which we have discussed in this paper.
- We feel that our method will work best if we are working in the right feature space. One of the vexing problems in machine learning is to identify the right features in the data. Deep networks are supposed to be doing this task automatically. However, there is no clear explanation of whether they work in every problem or otherwise develop a theory of what "features" should mean and how deep networks can extract these automatically.

List of Symbols

η_x Composite map $\varepsilon_x \circ \overline{\varepsilon}_x^{-1}$ defined in Section 5.

$\iota_1, \iota_2, \iota^*$ Injectivity radii of $\varepsilon_x, \overline{\varepsilon}_x, \eta_x$, respectively.

$\mathbb{H}_\ell^q$ Space of homogenous, harmonic polynomials of degree ℓ in q dimensions.

$\mathbb{S}^q, \mu_q^*$ Sphere of q-dimensions with probability measure μ_q^* as defined in (4.1).

$\mathbb{S}_x$ The unique q-dimensional equator of $\mathbb{S}^Q$ that shares a tangent space $\mathbb{T}_x$ with the point $x \in \mathbb{X}$.

$\mathbb{X}, \rho, \mu^*$ Submanifold of $\mathbb{S}^Q$ with geodesic ρ and normalized volume measure μ^*.

$\mathbb{Y}$ Equator of $\mathbb{S}^Q$, as defined in Section 4.4.

$\mathbf{g}_1, \mathbf{g}_2$ Metric tensors associated with $\varepsilon_x, \overline{\varepsilon}_x$, respectively. See A.2 for details.

$\mathcal{D}, \tau$ Set of data $\mathcal{D} = \{y_j, z_j\}_{j=1}^M$ sampled from distribution τ. It is assumed the y_j's lie on a q-dimensional submanifold of $\mathbb{S}^Q$.

ω_q Volume of the q-dimensional sphere.

$\Phi_{n,q}$ Localized kernels as defined in (4.12).

Π_n^q Space of spherical polynomials of degree $< n$. See Section 4.1.

σ_n Continuous approximation (a.k.a. integral reconstruction) operator as defined on $\mathbb{S}^q$ in (4.17), $\mathbb{Y}$ in (4.23), and $\mathbb{X}$ in (3.3).

$\varepsilon_x, \overline{\varepsilon}_x$ Exponential map for $\mathbb{X}, \mathbb{S}_x$ respectively. See A.2 for details.

$B_{Q+1}(x,r), S^Q(x,r), B_{\mathbb{T}}, \mathbb{S}_x, \mathbb{B}$ Balls on various spaces. See Table 1 for reference.

E_n Degree of approximation as defined in Sections 4.3 and 4.4.

F_n Our proposed constructive approximation for finite data as defined in (5.5).

$p_{q,\ell}$ orthonormalized ultraspherical polynomial of degree ℓ and dimension q as defined in Section 4.1.

Q, q Ambient dimension of the data and dimension of the underlying manifold, respectively.

W_γ Smoothness class of functions as defined on $\mathbb{S}^q$ in Section 4.3, on $\mathbb{Y}$ in Section 4.4, and on $\mathbb{X}$ in Definition 5.1.

$Y_{\ell,k}$ Basis elements for the space of homogenous, harmonic polynomials of degree ℓ.

Appendix

A.1 Encoding

Our construction in (5.5) allows us to encode the target function in terms of finitely many real numbers. For each integer $\ell \geq 0$, let $\{Y_{Q,\ell,k}\}_{k=1}^{\dim(\mathbb{H}_\ell^Q)}$ be an orthonormal basis for $\mathbb{H}_\ell^Q$ on $\mathbb{S}^Q$. We define the encoding of f by

$$\hat{z}(\ell,k) := \frac{1}{M}\sum_{j=1}^{M} z_j Y_{Q,\ell,k}(y_j). \tag{A.1}$$

Given this encoding, the decoding algorithm is given in the following proposition.

Proposition A.1. *Assume $\Phi_{n,q}$ is given as in (4.12). Given the encoding of f as given in (A.1), one can rewrite the approximation in (5.5) as*

$$F_n(\mathcal{D};x) = \sum_{\ell=0}^{n} \Gamma_{\ell,n} \sum_{k=1}^{\dim(\mathbb{H}_\ell^Q)} \hat{z}(\ell,k) Y_{Q,\ell,k}(x) \qquad x \in \mathbb{S}^Q, \tag{A.2}$$

where

$$\Gamma_{\ell,n} := \frac{\omega_q \omega_{Q-1}}{\omega_Q \omega_{q-1}} \sum_{i=\ell}^{n} h\left(\frac{i}{n}\right) \frac{p_{q,i}(1)}{p_{Q,\ell}(1)} C_{Q,q}(\ell,i), \tag{A.3}$$

and $C_{Q,q}(\ell,i)$ is defined in (4.11).

Proof. The proof follows from writing out

$$F_n(\mathcal{D};x) = \frac{\omega_q}{M\omega_{q-1}} \sum_{j=1}^{M} z_j \sum_{i=1}^{n} h\left(\frac{i}{n}\right) p_{q,i}(1) p_{q,i}(x\cdot y_j), \tag{A.4}$$

making substitutions using (4.11), (4.4), and (4.5), then collecting terms. ∎

Remark A.1. The encoding (A.1) is not parsimonious. Since the basis functions $\{Y_{Q,\ell,k}\}_{\ell=0,k=1}^{n,\dim(\mathbb{H}_\ell^Q)}$ is not necessarily independent on $\mathbb{X}$, the encoding can be made parsimonious by exploiting linear relationships in this system. Given a reparametrization the functions as $\{Y_j\}_{j=1}^{\sum_{\ell=0}^{n}\dim(\mathbb{H}_\ell^Q)}$, we form the discrete Gram matrix G by the entries

$$G_{i,j} := \frac{1}{M}\sum_{k=1}^{M} Y_i(y_k)Y_j(y_k) \approx \int_{\mathbb{X}} Y_i(y)Y_j(y) f_0 d\mu^*(y). \tag{A.5}$$

In practice, one may formulate a QR decomposition by fixing some first basis vector and proceeding by the Gram-Schmidt process until a basis is formed, then setting some threshold on the eigenvalues to get the desired dependencies among the Y_j's. ∎

A.2 Background on manifolds

This introduction to manifolds covers the main ideas which we use in this paper without going into much detail. We mostly follow along with the notation and definitions in [14]. For details, we refer the reader to texts such as [4, 14, 17].

Definition A.1 (Differentiable Manifold)**.** A (boundary-less) *differentiable manifold* of dimension q is a set $\mathbb{X}$ together with a family of open subsets $\{U_\alpha\}$ of $\mathbb{R}^q$ and functions $\{\mathbf{x}_\alpha\}$ such that

$$\mathbf{x}_\alpha : U_\alpha \to \mathbb{X} \tag{A.6}$$

is injective, and the following 3 properties hold:

- $\bigcup_\alpha \mathbf{x}_\alpha(U_\alpha) = \mathbb{X}$,
- $\mathbf{x}_\alpha(U_\alpha) \cap \mathbf{x}_\beta(U_\beta) = W \neq \emptyset$ implies that $\mathbf{x}_\alpha^{-1}(W), \mathbf{x}_\beta^{-1}(W)$ are open sets and $\mathbf{x}_\beta^{-1} \circ \mathbf{x}_\alpha$ is an infinitely differentiable function.
- The family $\mathcal{A}_\mathbb{X} = \{(U_\alpha, \mathbf{x}_\alpha)\}$ is maximal regarding the above conditions.

Remark A.2. The pair $(U_\alpha, \mathbf{x}_\alpha)$ gives a *local coordinate chart* of the manifold, and the collection of all such charts $\mathcal{A}_\mathbb{X}$ is known as the *atlas*. ■

Definition A.2 (Differentiable Map)**.** Let $\mathbb{X}_1, \mathbb{X}_2$ be differentiable manifolds. We say a function $\phi : \mathbb{X}_1 \to \mathbb{X}_2$ is (infinitely) *differentiable*, denoted by $\phi \in C^\infty(\mathbb{X})$, at a point $x \in \mathbb{X}_1$ if given a chart $(V, \mathbf{y})$ of $\mathbb{X}_2$, there exists a chart $(U, \mathbf{x})$ of $\mathbb{X}_1$ such that $x \in \mathbf{x}(U)$, $\phi(\mathbf{x}(U)) \subseteq \mathbf{y}(V)$, and $\mathbf{y}^{-1} \circ \phi \circ \mathbf{x}$ is infinitely differentiable at $\mathbf{x}^{-1}(p)$ in the traditional sense.

For any interval I of $\mathbb{R}$, a differentiable function $\gamma : I \to \mathbb{X}$ is known as a *curve*. If $x \in \mathbb{X}$, $\epsilon > 0$, and $\gamma : (-\epsilon, \epsilon) \to \mathbb{X}$ is a curve with $x = \gamma(0)$, then we can define the *tangent vector* to γ at $\gamma(t_0)$ as a functional $\gamma'(t_0)$ acting on the class of differentiable functions $f : \mathbb{X} \to \mathbb{R}$ by

$$\gamma'(t_0) f \coloneqq \frac{d(f \circ \gamma)}{dt}(t_0). \tag{A.7}$$

The *tangent space* of $\mathbb{X}$ at a point $x \in \mathbb{X}$, denoted by $\mathbb{T}_x(\mathbb{X})$, is the set of all such functionals $\gamma'(0)$.

A *Riemannian* manifold is a differentiable manifold with a family of inner products $\{\langle \circ, \circ \rangle_x\}_{x \in \mathbb{X}}$ such that for any $X, Y \in \mathbb{T}_x(\mathbb{X})$, the function $\varphi : \mathbb{X} \to \mathbb{C}$ given by $x \mapsto \langle X(x), Y(x) \rangle_x$ is differentiable. We can define an associated norm $||X|| = \langle X(x), X(x) \rangle_x$. The length $L(\gamma)$ of a curve γ defined on $[a, b]$ is defined to be

$$L(\gamma) \coloneqq \int_a^b ||\gamma'(t)||\, dt. \tag{A.8}$$

We will call a curve $\gamma : [a, b] \to \mathbb{X}$ a *geodesic* if $L(\gamma) = \inf\{L(r) : r : [a, b] \to \mathbb{X}, r \text{ is a curve}\}$. It is well-known that if γ is a geodesic, then $\gamma'(t) \cdot \gamma''(t) = 0$ for any $t \in [a, b]$.

In the sequel, we assume that $\mathbb{X}$ is a compact, connected, Riemannian manifold. Then for every $x, y \in \mathbb{X}$ there exists a geodesic $\gamma : [a, b] \to \mathbb{X}$ such that $\gamma(a) = x, \gamma(b) = y$. The quantity $\rho(x, y) = L(\gamma)$ defines a metric on $\mathbb{X}$ such that the corresponding metric topology is consistent with the topology defined by any atlas on $\mathbb{X}$.

For any $x \in \mathbb{X}$, there exists a neighborhood $V \subset \mathbb{X}$ of x, a number $\delta = \delta(x) > 0$ and a mapping $\mathcal{E} : (-2, 2) \times U \to \mathbb{X}$, where $U = \{(y, v) : y \in V,\ v \in T_y\mathbb{X},\ \|v\|_2 < \delta\}$ such that $t \mapsto \mathcal{E}(t, y, v)$ is the unique geodesic of $\mathbb{X}$ which, at $t = 0$, passes through y and has the property that $\partial\mathcal{E}/\partial t = v$ for each $(y, v) \in U$. As a result, we can define the *exponential map* at x to be the function $\varepsilon_x : B_\mathbb{T}(x, \delta(x)) \subset \mathbb{T}_x(\mathbb{X}) \to \mathbb{X}$ by $\varepsilon_x(v) = \mathcal{E}(1, x, v)$. Intuitively, the line joining x and v in $\mathbb{T}_x(\mathbb{X})$ is mapped to the geodesic joining x with $\varepsilon_x(v)$. We call the supremum of all $\delta(x)$ for which the exponential map is so defined the *injectivity radius* at x, denoted by $\iota(x)$. We call $\iota^* = \inf_{x \in \mathbb{X}} \iota(x)$ the *global injectivity radius* of $\mathbb{X}$. Since $x \mapsto \iota(x)$ is a continuous function of x, and $\iota(x) > 0$ for each x, it follows that $\iota^* > 0$ when $\mathbb{X}$ is compact. Correspondingly, on compact manifolds, one can conclude that for $y \in B_\mathbb{T}(x, \iota^*)$, $\rho(x, \varepsilon_x(y)) = ||x - y||$.

Next, we discuss the metric tensor and volume element on $\mathbb{X}$. Let $(U, \mathbf{x})$ be a coordinate chart with $0 \in U$, $\mathbf{x}(0) = x \in \mathbb{X}$, and $\partial_j(x)$ be the tangent vector at x to the coordinate curve $t \mapsto \mathbf{x}((\underbrace{0, \ldots, 0}_{j-1}, t, 0, \ldots, 0))$. Then we can define the metric tensor $\mathbf{g}$ to be the matrix where $\mathbf{g}_{ij} = \langle \partial_i(x), \partial_j(x) \rangle_x$. When one expands the metric tensor

$\mathbf{g}$ as a Taylor series in local coordinates on $\mathbb{B}(x, \iota^*)$, it can be shown [36, pg. 21] that for any $\delta < \iota^*$, on the ball $\mathbb{B}(x, \delta)$ we have

$$|\mathbf{g}| = 1 + O(\delta^2). \tag{A.9}$$

In turn, this implies

$$\sqrt{|\mathbf{g}|} - 1 \lesssim \delta^2. \tag{A.10}$$

The following proposition lists some important properties relating the geodesic distance ρ on an unknown submanifold of $\mathbb{S}^Q$ with the geodesic distance on $\mathbb{S}^Q$ as well as the Euclidean distance on $\mathbb{R}^{Q+1}$.

Proposition A.2. *Let η_x be defined as in Section 5.*

(a) *For every $\eta_x(u) \in \mathbb{B}(x, \iota^*)$,*

$$|\arccos(x \cdot \eta_x(u)) - \rho(x, \eta_x(u))| \lesssim \rho(x, \eta_x(u))^3. \tag{A.11}$$

(b) *For any $x, y \in \mathbb{X}$,*

$$\rho(x, y) \sim \arccos(x \cdot y). \tag{A.12}$$

Proof. First, we observe the fact that $||x - y||_2 \sim \arccos(x \cdot y)$ because $||x - y||_2 / 2 = \sin(\arccos(x \cdot y)/2)$ and $\theta/\pi \leq \sin(\theta/2) \leq \theta/2$ for all $\theta \in [0, \pi]$. Fix $x \in \mathbb{X}$ and let γ be a geodesic on $\mathbb{X}$ parametrized by length t from x. In particular we then have $||\gamma'(0)||_2 = 1$ and $\gamma'(0) \cdot \gamma''(0) = 0$. Taking a Taylor expansion for $\gamma(t)$ with $|t| < \iota^*$ (we recall that $\iota^* \leq 1$), we can see

$$\begin{aligned} \gamma'(0) \cdot (\gamma(t) - \gamma(0)) =& \gamma'(0) \cdot \left(\gamma'(0)t + \frac{1}{2}\gamma''(t)t^2 + O(t^3) \right) \\ =& ||\gamma'(0)||_2^2 t + \gamma'(0) \cdot \gamma''(0)t^2 + O(t^3) \\ =& t + O(t^3). \end{aligned} \tag{A.13}$$

For any $y \in \mathbb{B}(x, \iota^*)$, there exists a unique $u \in \mathbb{S}_x(\iota^*)$ such that $y = \eta_x(u)$. We can write $y = \gamma(t)$ for some geodesic γ. We know, $t = \rho(x, y) \geq \arccos(x \cdot y) \geq ||x - y||_2 = ||\gamma(t) - \gamma(0)||_2$. Using the Cauchy-Schwarz inequality, we see

$$0 \leq t - ||x - y||_2 \leq t - \gamma'(0) \cdot (\gamma(t) - \gamma(0)) \lesssim t^3. \tag{A.14}$$

As a result we can conclude

$$\rho(x, \eta_x(u)) - \arccos(x \cdot \eta_x(u)) \leq \rho(x, \eta_x(u)) - ||\eta_x(u) - x||_2 \lesssim \rho(x, \eta_x(u))^3, \tag{A.15}$$

showing (A.11). Letting c be the constant built into the notation of (A.11), then if we fix $x \in \mathbb{X}$ and let $y \in \mathbb{B}\big(x, \sqrt{1/(2c)}\big)$, we have

$$\frac{1}{2}\rho(x, y) \leq \rho(x, y) - c\rho(x, y)^3 \leq \arccos(x \cdot y). \tag{A.16}$$

Furthermore, since $A = \overline{\mathbb{X} \setminus \mathbb{B}\big(x, \sqrt{1/(2c)}\big)}$ is a compact set and $g_x(y) = \arccos(x \cdot y)/\rho(x, y)$ is a continuous function of y defined on A, we can conclude that g_x attains a minimum on A. Therefore,

$$\rho(x, y) \sim \arccos(x \cdot y) \tag{A.17}$$

for every $y \in \mathbb{X}$. We note that the constants involved in this proof vary continuously with respect to the choice of x, so in the theorem we may simply use the supremum over all such constants which must be finite since $\mathbb{X}$ is compact. ■

A.3 Network representation

Let $\{p_k\}$ be a system of orthonormal polynomials satisfying a recurrence relation

$$p_k(x) = a_k x p_{k-1}(x) + b_k p_{k-2}(x), \qquad k = 1, 2, \cdots, \quad b_1 = 0. \tag{A.18}$$

The Clenshaw algorithm is a modification of the classical Horner method to compute polynomials expressed in the monomial basis that evaluates a polynomial expressed in terms of the orthonormalized polynomials $\{p_k\}$ [9, 16].

Algorithm 1 Clenshaw algorithm to compute $\sum_{k=0}^{n-1} C_k p_k$, where $p_k(x) = a_k x p_{k-1}(x) + b_k p_{k-2}(x)$, $k = 1, 2, \cdots, n-1$, $b_1 = 0$.

a) **Input:** p_0, $C_0, \cdots, C_{n-1}$, x, $a_{n+1}, \cdots, a_1$, $b_{n+1}, \cdots, b_1$.
b) **Output:** The value of $\sum_{k=0}^{n-1} C_k p_k$.
1: $\mathsf{out1} \leftarrow 0, \mathsf{out2} \leftarrow 0, C_{-1} \leftarrow 0, C_n \leftarrow 0$.
2: **for** $k = n+1$ down to 1 **do**
3: $\quad \mathsf{temp} \leftarrow a_k * \mathsf{out}_1 * x + \mathsf{out2}$
4: $\quad \mathsf{out2} \leftarrow b_k * \mathsf{out1} + C_{k-2}$
5: $\quad \mathsf{out1} \leftarrow \mathsf{temp}$.
6: **end for**
7: **Return:** $\mathsf{out1} * p_0$.

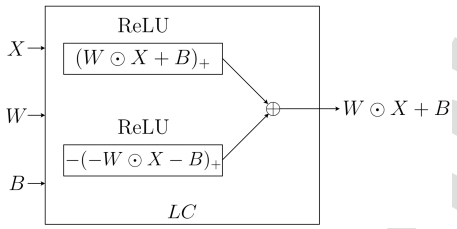

Figure 8: The implementation of a linear combination as a ReLU network. Here all operations are pointwise. The symbols $\odot$ represents Hadamard product of matrices, $\oplus$ is the sum of matrices.

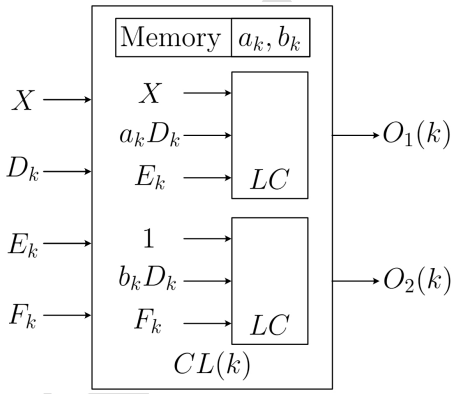

Figure 9: One step of the Clenshaw algorithm, using two circuits of the form LC (4 neurons) as in Figure 8. The circuit diagram is shown in general with four input pins and two output pins.

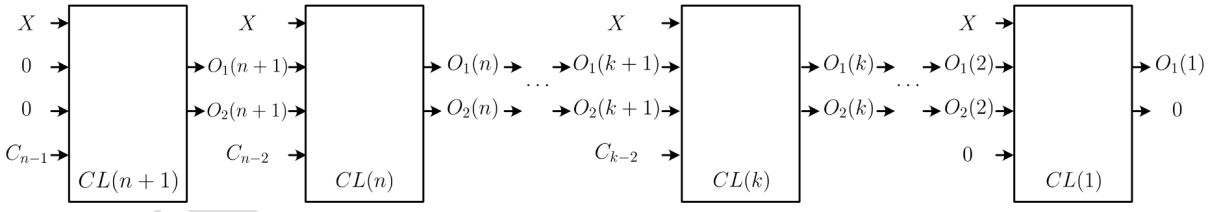

Figure 10: Unrolling the Clenshaw algorithm as a cascade of the circuits of the form CL(k) as in Figure 9.

To understand the method, let $P = \sum_{k=0}^{n-1} C_k p_k$. It is convenient to write $C_k = 0$ if $k \geq n$ or $k < 0$. The recurrence (A.18) shows that

$$C_k p_k(x) + C_{k-1} p_{k-1}(x) + C_{k-2} p_{k-2}(x) = (a_k C_k x + C_{k-1})\, p_{k-1}(x) + (b_k C_k + C_{k-2})\, p_{k-2}. \tag{A.19}$$

This leads to Algorithm 1.

By algorithm unrolling, we may express this algorithm in terms of a deep neural network evaluating a ReLU activation function. The network is a cascade of different circuits. The most fundamental is the implementation of a linear combination as a ReLU network (see Figure 8)

$$ax + b = (ax + b)_+ - (-ax - b)_+.$$

Using the circuits LC in Figure 8, we next construct a circuit to implement recursive reduction (A.19). This is illustrated in Figure 9. Finally, we unroll the Clenshaw algorithm by cascading the circuits CL(k) from Figure 9 for $k = n + 1$ down to $k = 1$ with different inputs and outputs as shown in Figure 10. We use this in order to compute $\Phi_{n,q}(x \cdot y_j)$ by using the recursive formula for ultraspherical polynomials (4.7) in the following way. We set

$$\begin{aligned} C_k &= \frac{\omega_q}{\omega_{q-1}} h(k/n) p_{q,k}(1), \\ a_k &= \begin{cases} \dfrac{\sqrt{\Gamma(q)\Gamma(q+1)}}{\Gamma(q-1)} & k = 1 \\ \sqrt{\dfrac{(2k+q-3)(2k+q-1)}{k(n+q-2)}} & k \geq 2 \end{cases}, \\ b_k &= \sqrt{\frac{(k-1)(k+q-3)(2k+q-1)}{k(k+q-2)(2k+q-5)}}. \end{aligned} \tag{A.20}$$

For the matrix X shown in Figure 10, we consider the $(Q+1) \times N$ test data matrix S where each column represents one test data x, and a $(Q+1) \times M$ train data matrix R where column j represents data point y_j. Then we set $X = S^T R$. In this way, we would return $\Phi_{n,q}(S^T R)$ from running Algorithm 1, with a time complexity of $O(NMn)$.

References

[1] R. Askey. *Orthogonal Polynomials and Special Functions*. Society for Industrial and Applied Mathematics, 1975.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

[4] W. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry: An Introduction to Differentiable Manifolds and Riemannian Geometry*. ISSN. Elsevier Science, 1975.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

[6] M.-Y. Cheng and H.-T. Wu. Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association*, 108(504):1421–1434, 2013.

[7] C. K. Chui and D. L. Donoho. Special issue: Diffusion maps and wavelets. *Appl. and Comput. Harm. Anal.*, 21(1), 2006.

[8] C. K. Chui and H. N. Mhaskar. Deep nets for local manifold learning. *Frontiers in Applied Mathematics and Statistics*, 4:12, 2018.

[9] C. W. Clenshaw. A note on the summation of chebyshev series. *Mathematics of Computation*, 9:118–120, 1955.

[10] A. Cloninger, R. R. Coifman, N. Downing, and H. M. Krumholz. Bigeometric organization of deep nets. *Applied and Computational Harmonic Analysis*, 44(3):774–785, 2018.

[11] R. R. Coifman and S. Lafon. Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21(1):31–52, 2006.

[12] F. Dai and Y. Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer New York, 2013.

[13] R. A. De Vore. *The approximation of continuous functions by positive linear operators*, volume 293. Springer, 2006.

[14] M. P. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

[15] E. Fuselier and G. B. Wright. Scattered data interpolation on embedded submanifolds with restricted positive definite kernels: Sobolev error estimates. *SIAM Journal on Numerical Analysis*, 50(3):1753–1776, 2012.

[16] W. Gautschi. *Orthogonal polynomials: computation and approximation*. Oxford University Press on Demand, 2004.

[17] V. Guillemin and A. Pollack. *Differential Topology*. AMS Chelsea Pub., 2010.

[18] A. A. Istratov and O. F. Vyvenko. Exponential analysis in physical phenomena. *Review of Scientific Instruments*, 70(2):1233–1257, 1999.

[19] N. Lehmann, L.-B. Maier, S. Odathuparambil, and U. Reif. Ambient approximation on hypersurfaces. *Constructive Approximation*, 49:175–190, 2019.

[20] W. Liao and M. Maggioni. Adaptive geometric multiscale approximations for intrinsically low-dimensional data. *arXiv preprint arXiv:1611.01179*, 2016.

[21] W. Liao, M. Maggioni, and S. Vigogna. Learning adaptive multiscale approximations to data and functions near low-dimensional sets. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 226–230. IEEE, 2016.

[22] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis. Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*, 2019.

[23] L. Manoni, C. Turchetti, and L. Falaschetti. An effective manifold learning approach to parametrize data for generative modeling of biosignals. *IEEE Access*, 8:207112–207133, 2020.

[24] E. Mason, H. N. Mhaskar, and A. Guo. A manifold learning approach for gesture identification from micro-doppler radar measurements. *Neural Networks*, 152:353–369, 2022. arXiv preprint arXiv:2110.01670, 2021.

[25] H. N. Mhaskar. Polynomial operators and local smoothness classes on the unit interval. *Journal of Approximation Theory*, 131(2):243–267, 2004.

[26] H. N. Mhaskar. On the representation of smooth functions on the sphere using finitely many bits. *Applied and Computational Harmonic Analysis*, 18(3):215–233, 2005.

[27] H. N. Mhaskar. Eignets for function approximation on manifolds. *Applied and Computational Harmonic Analysis*, 29(1):63–87, 2010.

[28] H. N. Mhaskar. Dimension independent bounds for general shallow networks. *Neural Networks*, 123:142–152, 2020.

[29] H. N. Mhaskar. A direct approach for function approximation on data defined manifolds. *Neural Networks*, 132:253–268, 2020.

[30] H. N. Mhaskar. Kernel-based analysis of massive data. *Frontiers in Applied Mathematics and Statistics*, 6:30, 10 2020.

[31] H. N. Mhaskar. Function approximation with zonal function networks with activation functions analogous to the rectified linear unit functions. *Journal of Complexity*, 51:1–19, April 2019.

[32] H. N. Mhaskar, S. V. Pereverzyev, and M. D. van der Walt. A function approximation approach to the prediction of blood glucose levels. *Frontiers in Applied Mathematics and Statistics*, 7:53, 2021.

[33] C. Müller. *Spherical harmonics*, volume 17. Springer, 2006.

[34] V. Naumova, L. Nita, J. U. Poulsen, and S. V. Pereverzyev. Meta-learning based blood glucose predictor for diabetic smartphone app, 2014.

[35] B. Raonic, R. Molinaro, T. De Ryck, T. Rohner, F. Bartolucci, R. Alaifari, S. Mishra, and E. de Bézenac. Convolutional neural operators for robust and accurate learning of pdes. *Advances in Neural Information Processing Systems*, 36, 2024.

[36] J. Roe. *Elliptic operators, topology and asymptotic methods*. Addison Wesley Longman Inc., 2 edition, 1998.

[37] M. Rozowski, J. Palumbo, J. Bisen, C. Bi, M. Bouhrara, W. Czaja, and R. G. Spencer. Input layer regularization for magnetic resonance relaxometry biexponential parameter estimation. *Magnetic Resonance in Chemistry*, 60(11):1076–1086, 2022.

[38] K. P. Rustamov. On approximation of functions on the sphere. *Izvestiya: Mathematics*, 43(2):311, apr 1994.

[39] J. Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

[40] A. Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006. Special Issue: Diffusion Maps and Wavelets.

[41] B. Sober, Y. Aizenbud, and D. Levin. Approximation of functions over manifolds: A moving least-squares approach. *arXiv preprint arXiv:1711.00765*, 2017.

[42] B. Sober, Y. Aizenbud, and D. Levin. Approximation of functions over manifolds: A moving least-squares approach. *Journal of Computational and Applied Mathematics*, 383:113140, 2021.

[43] E. M. Stein and G. Weiss. *Introduction to Fourier analysis on Euclidean spaces (PMS-32)*, volume 32. Princeton university press, 2016.

[44] G. Szegö. *Orthogonal Polynomials*. American Math. Soc: Colloquium publ. American Mathematical Society, 1975.

- Constructive approximation on an unknown manifold directly from noisy data
- No need to learn information about the manifold, such as eigenfunctions
- Universal construction, theoretical performance guarantees for rough functions
- Out-of-sample extension based on spherical harmonics included by design
- Encoding and decoding methods with theoretical performance guarantees

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Hrushikesh Mhaskar reports financial support was provided by National Science Foundation. Hrushikesh Mhaskar reports financial support was provided by Office of Naval Research. Ryan O'Dowd reports financial support was provided by Office of Naval Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.