frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 20 October 2020
doi: 10.3389/fams.2020.00030

Kernel-Based Analysis of Massive Data

*Hrushikesh N. Mhaskar**

Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA, United States

Dealing with massive data is a challenging task for machine learning. An important aspect of machine learning is function approximation. In the context of massive data, some of the commonly used tools for this purpose are sparsity, divide-and-conquer, and distributed learning. In this paper, we develop a very general theory of approximation by networks, which we have called eignets, to achieve local, stratified approximation. The very massive nature of the data allows us to use these eignets to solve inverse problems, such as finding a good approximation to the probability law that governs the data and finding the local smoothness of the target function near different points in the domain. In fact, we develop a wavelet-like representation using our eignets. Our theory is applicable to approximation on a general locally compact metric measure space. Special examples include approximation by periodic basis functions on the torus, zonal function networks on a Euclidean sphere (including smooth ReLU networks), Gaussian networks, and approximation on manifolds. We construct pre-fabricated networks so that no data-based training is required for the approximation.

Keywords: Kernel based approximation, distributed learning, machine learning, inverse problems, probability estimation

Edited by:
Ke Shi,
Old Dominion University, United States

Reviewed by:
Jianjun Wang,
Southwest University, China
Alex Cloninger,
University of California, San Diego,
United States

Correspondence:
Hrushikesh N. Mhaskar
hrushikesh.mhaskar@cgu.edu

Specialty section:
This article was submitted to
Mathematics of Computation and
Data Science,
a section of the journal
Frontiers in Applied Mathematics and
Statistics

Received: *29 March 2020*
Accepted: *03 July 2020*
Published: *20 October 2020*

Citation:
Mhaskar HN (2020) Kernel-Based
Analysis of Massive Data.
Front. Appl. Math. Stat. 6:30.
doi: 10.3389/fams.2020.00030

1. INTRODUCTION

Rapid advances in technology have led to the availability and need to analyze a massive data. The problem arises in almost every area of life from medical science to homeland security to finance. An immediate problem in dealing with a massive data set is that it is not possible to store it in a computer memory; we therefore have to deal with the data piecemeal to keep access to an external memory to a minimum. The other challenge is to devise efficient numerical algorithms to overcome difficulties, for example, in using the customary optimization problems in machine learning. On the other hand, the very availability of a massive data set should lead also to opportunities to solve some problems heretofore considered unmanageable. For example, deep learning often requires a large amount of training data, which, in turn, helps us to figure out the granularity in the data. Apart from deep learning, distributed learning is also a popular way of dealing with big data. A good survey with the taxonomy for dealing with massive data was recently conducted by Zhou et al. [1].

As pointed out in Cucker and Smale [2], Cucker and Zhou [3], and Girosi and Poggio [4], the main task in machine learning can be viewed as one of approximation of functions based on noisy values of the target function, sampled at points that are themselves sampled from an unknown distribution. It is therefore natural to seek approximation theory techniques to solve the problem. However, most of the classical approximation theory results are either not constructive or study function approximation only on known domains. In this century, there is a new paradigm to consider function approximation on data-defined manifolds; a good introduction to the subject is in the special issue [5] of Applied and Computational Harmonic Analysis, edited by Chui and Donoho. In this theory, one assumes the *manifold hypothesis*, i.e., that the data is sampled from a probability distribution μ^* supported on a smooth, compact, and connected Riemannian manifold; for simplicity, even that μ^* is the Riemannian volume measure for the manifold, normalized to be a probability measure. Following (e.g., [6–10]), one constructs first a "graph Laplacian" from the data and finds its eigen decomposition. It is proved in the abovementioned papers that as the size of the data tends to infinity, the graph Laplacian converges to the Laplace-Beltrami operator on the manifold, and the eigenvalues (eigenvectors) converge to the corresponding quantities on the manifold. A great deal of work is devoted to studying the geometry of this unknown manifold (e.g., [11, 12]) based on the so-called heat kernel. The theory of function approximation on such manifolds is also well-developed (e.g., [13–17]).

A bottleneck in this theory is the computation of the eigendecomposition of a matrix, which is necessarily huge in the case of big data. Kernel-based methods have been used also in connection with approximation on manifolds (e.g., [18–22]). The kernels used in this method are constructed typically as a radial basis function (RBF) in the ambient space, and the methods are traditional machine learning methods involving optimization. As mentioned earlier, massive data poses a big challenge for the solution of these optimization problems. The theoretical results in this connection assume a Mercer's expansion in terms of the Laplacian eigenfunctions for the kernel, satisfying certain conditions. In this paper, we develop a general theory including several RBF kernels in use in different contexts (examples are discussed in section 2). Rather than using optimization-based techniques, we will provide a direct construction of the approximation based on what we have called eignets. An eignet is defined directly using the eigendecomposition on the manifold. We thus focus directly on the properties of Mercer expansion in an abstract and unified manner that enables us to construct local approximations suitable for working with massive data without using optimization.

It is also possible that the manifold hypothesis does not hold, and there is a recent work [23] by Fefferman et al. proposing an algorithm to test this hypothesis. On the other hand, our theory for function approximation does not necessarily use the full strength of Riemannian geometry. In this paper, we have therefore decided to work with a general locally compact metric measure space, isolating those properties which are needed for our analysis and substituting some that are not applicable in the current setting.

Our motivation comes from some recent works on distributed learning by Zhou et al. [24–26] as well as our own work on deep learning [27, 28]. For example, in Lin et al. [26], the approximation is done on the Euclidean sphere using a localized kernel introduced in Mhaskar [29], where the massive data is divided into smaller parts, each dense on the sphere, and the resulting polynomial approximations are added to get the final result. In Chui et al. [24], the approximation takes place on a cube, and exploits any known sparsity in the representation of the target function in terms of spline functions. In Mhaskar and Poggio [28] and Mhaskar [27], we have argued that from a function approximation point of view, the observed superiority of deep networks over shallow ones results from the ability of deep networks to exploit any compositional structure in the target function. For example, in image analysis, one may divide the image into smaller patches, which are then combined in a hierarchical manner, resulting in a tree structure [30]. By putting a shallow network at each node to learn those aspects of the target function that depend upon the pixels seen up to that level, one can avoid the curse of dimensionality. In some sense, this is a divide-and-conquer strategy, not so much on the data set itself but on the dimension of the input space.

The highlights of this paper are the following.

- In order to avoid an explicit, data-dependent eigendecomposition, we introduce the notion of an eignet, which generalizes several radial basis function and zonal function networks. We construct pre-fabricated eignets, whose linear combinations can be constructed just by using the noisy values of the target function as the coefficients, to yield the desired approximation.
- Our theory generalizes the results in a number of examples used commonly in machine learning, some of which we will describe in section 2.
- The use of optimization methods, such as empirical risk minimization has an intrinsic difficulty, namely, the minimizer of this risk may have no connection with the approximation error. There are also other problems, such as local minima, saddle points, speed of convergence, etc. that need to be taken into account, and the massive nature of the data makes this an even more challenging task. Our results do not depend upon any kind of optimization in order to determine the necessary approximation.
- We developed a theory for local approximation using eignets so that only a relatively small amount of data is used in order to approximate the target function in any ball of the space, the data being sub-sampled using a distribution supported on a neighborhood of that ball. The accuracy of approximation adjusts itself automatically depending upon the local smoothness of the target function on the ball.
- In normal machine learning algorithms, it is customary to assume a prior on the target function called smoothness class in approximation theory parlance. Our theory demonstrates clearly how a massive data can actually help to solve the inverse problem to determine the local smoothness of the target function using a wavelet-like representation based solely on the data.
- Our results allow one to solve the inverse problem of estimating the probability density from which the data is chosen. In contrast to the statistical approaches that we are aware of, there is no limitation on how accurate the approximation can be asymptotically in terms of the number of samples; the accuracy is determined entirely by the smoothness of the density function.
- All our estimates are given in terms of probability of the error being small rather

than the expected value of some loss function being small.

This paper is abstract, theoretical, and technical. In section 2, we present a number of examples that are generalized by our set-up. The abstract set-up, together with the necessary definitions and assumptions, are discussed in section 3. The main results are stated in section 4 and proved in section 8. The proofs require a great deal of preparation, which is presented in sections 5–7. The results in these sections are not all new. Many of them are new only in some nuance. For example, we have proven in section 7 the quadrature formulas required in the construction of our pre-fabricated networks in a probabilistic setting, and we have also substituted an estimate on the gradients by certain Lipschitz condition, which makes sense without the differentiability structure on the manifold as we had done in our previous works. Our Theorem 7.1 generalizes most of our previous results in this direction with the exception of [31, Theorem 2.3]. We have striven to give as many proofs as possible, partly for the sake of completion and partly because the results were not stated earlier in exactly the same form as needed here. In **Appendix A**, we give a short proof of the fact that the Gaussian upper bound for the heat kernel holds for arbitrary smooth, compact, connected manifolds. We could not find a reference for this fact. In **Appendix B**, we state the main probability theory estimates that are used ubiquitously in the paper.

2. MOTIVATING EXAMPLES

In this paper, we aim to develop a unifying theory applicable to a variety of kernels and domains. In this section, we describe some examples which have motivated the abstract theory to be presented in the rest of the paper. In the following examples, $q \geq 1$ is a fixed integer.

Example 2.1. Let $\mathbb{T}^q = \mathbb{R}^q/(2\pi\mathbb{Z}^q)$ be the q-dimensional torus. The distance between points $\mathbf{x} = (x_1, \cdots, x_q)$ and $\mathbf{y} = (y_1, \cdots, y_q)$ is defined by $\max_{1\leq k\leq q} |(x_k - y_k) \bmod 2\pi|$. The trigonometric monomial system $\{\exp(i\mathbf{k}\cdot\circ) : \mathbf{k} \in \mathbb{Z}^q\}$ is orthonormal with respect to the Lebesgue measure normalized to be a probability measure on $\mathbb{T}^q$. We recall that the periodization of a function $f : \mathbb{R}^q \to \mathbb{R}$ is defined formally by $f^\circ(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^q} f(\mathbf{x} + 2\mathbf{k}\pi)$. When f is integrable then the Fourier transform of f at $\mathbf{k} \in \mathbb{Z}^q$ is the same as the $\mathbf{k}$-th Fourier coefficient of f°. This Fourier coefficient will be denoted by $\widehat{f^\circ}(\mathbf{k}) = \hat{f}(\mathbf{k})$. A periodic basis function network has the form $\mathbf{x} \mapsto \sum_{k=1}^n a_k G(\mathbf{x} - \mathbf{x}_k)$, where G is a periodic function called the activation function. The examples of the activation functions in which we are interested in this paper include:

1. Periodization of the Gaussian.

$$G(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^q} \exp(-|\mathbf{x} - 2\pi\mathbf{k}|_2^2/2)\,,$$
$$\hat{G}(\mathbf{k}) = (2\pi)^{q/2} \exp(-|\mathbf{k}|_2^2/2).$$

2. Periodization of the Hardy multiquadric[1].

$$G(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^q} (\alpha^2 + |\mathbf{x} - 2\pi\mathbf{k}|_2^2)^{-1}\,,$$
$$\hat{G}(\mathbf{k}) = \frac{\pi^{(q+1)/2}}{\Gamma\left(\frac{q+1}{2}\right)\alpha} \exp(-\alpha|\mathbf{k}|_2), \qquad \alpha > 0. \qquad \square$$

Example 2.2. If $\mathbf{x} = (x_1, \cdots, x_q) \in [-1,1]^q$, there exists a unique $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q) \in [0,\pi]^q$ such that $\mathbf{x} = \cos(\boldsymbol{\theta})$. Therefore, $[-1,1]^q$ can be thought of as a quotient space of $\mathbb{T}^q$ where all points of the form $\boldsymbol{\varepsilon} \odot \boldsymbol{\theta} = \{(\varepsilon_1\theta_1, \cdots, \varepsilon_q\theta_q)\}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_q) \in \{-1,1\}^q$, are identified. Any function on $[-1,1]^q$ can then by lifted to $\mathbb{T}^q$, and this lifting preserves all the smoothness properties of the function. Our set-up below includes $[-1,1]^q$, where the distance and the measure are defined via the mapping to the torus, and suitably weighted Jacobi polynomials are considered to be the orthonormalized family of functions. In particular, if G is a periodic activation function, $\mathbf{x} = \cos(\boldsymbol{\theta})$, $\mathbf{y} = \cos(\boldsymbol{\phi})$, then the function $G^{\square}(\mathbf{x},\mathbf{y}) = \sum_{\boldsymbol{\varepsilon}\in\{-1,1\}^q} G(\boldsymbol{\varepsilon} \odot (\boldsymbol{\theta} - \boldsymbol{\phi}))$ is an activation function on $[-1,1]^q$ with an expansion $\sum_{\mathbf{k}\in\mathbb{Z}_+^q} b_{\mathbf{k}} T_{\mathbf{k}}(\mathbf{x}) T_{\mathbf{k}}(\mathbf{y})$, where $T_{\mathbf{k}}$'s are tensor product, orthonormalized, Chebyshev polynomials. Furthermore, $b_{\mathbf{k}}$'s have the same asymptotic behavior as $\hat{G}(\mathbf{k})$'s. $\square$

Example 2.3. Let $\mathbb{S}^q = \{\mathbf{x} \in \mathbb{R}^{q+1} : |\mathbf{x}|_2 = 1\}$ be the unit sphere in $\mathbb{R}^{q+1}$. The dimension of $\mathbb{S}^q$ as a manifold is q. We assume the geodesic distance ρ on $\mathbb{S}^q$ and the volume measure μ^* are normalized to be a probability measure. We refer the reader to Müller [33] for details, describing here only the essentials to get a "what-it-is-all-about" introduction. The set of (equivalence classes) of restrictions of polynomials in $q+1$ variables with total degree $< n$ to $\mathbb{S}^q$ are called spherical polynomials of degree $< n$. The set of restrictions of homogeneous harmonic polynomials of degree ℓ to $\mathbb{S}^q$ is denoted by $\mathbb{H}_\ell$ with dimension d_ℓ. There is an orthonormal basis $\{Y_{\ell,k}\}_{k=1}^{d_\ell}$ for each $\mathbb{H}_\ell$ that satisfies an addition formula

$$\sum_{k=1}^{d_\ell} Y_{\ell,k}(\mathbf{x}) Y_{\ell,k}(\mathbf{y}) = \omega_{q-1}^{-1} p_\ell(1) p_\ell(\mathbf{x}\cdot\mathbf{y}),$$

where ω_{q-1} is the volume of $\mathbb{S}^{q-1}$, and p_ℓ is the degree ℓ ultraspherical polynomial so that the family $\{p_\ell\}$ is orthonormalized with respect to the weight $(1-x^2)^{(q-2)/2}$ on $(-1,1)$. A zonal function on the sphere has the form $\mathbf{x} \mapsto G(\mathbf{x}\cdot\mathbf{y})$, where the activation function $G : [-1,1] \to \mathbb{R}$ has a formal expansion of the form

$$G(t) = \omega_{q-1}^{-1} \sum_{\ell=0}^{\infty} \hat{G}(\ell) p_\ell(1) p_\ell(t).$$

In particular, formally, $G(\mathbf{x}\cdot\mathbf{y}) = \sum_{\ell=0}^{\infty} \hat{G}(\ell) \sum_{k=1}^{d_\ell} Y_{\ell,k}(\mathbf{x}) Y_{\ell,k}(\mathbf{y})$. The examples of the activation functions in which we are interested in this paper include

[1] A Hardy multiquadric is a function of the form $\mathbf{x} \to (\alpha^2 + |\mathbf{x}|_2^2)^{-1}$, $\mathbf{x} \in \mathbb{R}^q$. It is one of the oft-used function in theory and applications of radial basis function networks. For a survey, see the paper [32] of Hardy.

1.

$$G_r(x) := (1-2rx+r^2)^{-(q-1)/2}, \qquad x \in [-1,1],\ 0<r<1.$$

It is shown in Müller [33, Lemma 18] that

$$\widehat{G_r}(\ell) = \frac{(q-1)\omega_q}{2\ell+q-1} r^\ell, \qquad \ell = 1,2,\cdots.$$

2.

$$G_r^E(x) := \exp(rx), \qquad x \in [-1,1],\ r>0.$$

It is shown in Mhaskar et al. [34, Lemma 5.1] that

$$\hat{G_r^E}(\ell) = \frac{\omega_q r^\ell}{2^\ell\, \Gamma(\ell+\frac{q+1}{2})}\left(1+O(1/\ell)\right).$$

3. The smooth ReLU function $G(t) = \log(1+e^t) = t_+ + O(e^{-|t|})$. The function G has an analytic extension to the strip $\mathbb{R} + (-\pi,\pi)i$ of the complex plane. So, Bernstein approximation theorem [35, Theorem 5.4.2] can be used to show that

$$\limsup_{\ell\to\infty} |\hat{G}(\ell)|^{1/\ell} = 1/\pi. \qquad \square$$

Example 2.4. Let $\mathbb{X}$ be a smooth, compact, connected Riemannian manifold (without boundary), ρ be the geodesic distance on $\mathbb{X}$, μ^* be the Riemannian volume measure normalized to be a probability measure, $\{\lambda_k\}$ be the sequence of eigenvalues of the (negative) Laplace-Beltrami operator on $\mathbb{X}$, and ϕ_k be the eigenfunction corresponding to the eigenvalue λ_k; in particular, $\phi_0 \equiv 1$. This example, of course, includes Examples 2.1–2.3. An eignet in this context has the form $x \mapsto \sum_{k=1}^n a_k G(x, x_k)$, where the activation function G has a formal expansion of the form $G(x,y) = \sum_k b(\lambda_k)\phi_k(x)\phi_k(y)$. One interesting example is the heat kernel:

$$\sum_{k=0}^{\infty} \exp(-\lambda_k^2 t)\phi_k(x)\phi_k(y).$$

□

Example 2.5. Let $\mathbb{X} = \mathbb{R}^q$, ρ be the ℓ^∞ norm on $\mathbb{X}$, μ^* be the Lebesgue measure. For any multi-integer $\mathbf{k} \in \mathbb{Z}_+^q$, the (multivariate) Hermite function $\phi_{\mathbf{k}}$ is defined via the generating function

$$\sum_{\mathbf{k}\in\mathbb{Z}_+^q} \frac{\phi_{\mathbf{k}}(\mathbf{x})}{\sqrt{2^{|\mathbf{k}|_1}\mathbf{k}!}} \mathbf{w}^{\mathbf{k}} = \pi^{-1/4} \exp\left(-\frac{1}{2}|\mathbf{x}-\mathbf{w}|_2^2 + |\mathbf{w}|_2^2/4\right), \mathbf{w} \in \mathbb{C}^q. \tag{2.1}$$

The system $\{\phi_{\mathbf{k}}\}$ is orthonormal with respect to μ^*, and satisfies

$$\Delta\phi_{\mathbf{k}}(\mathbf{x}) - |\mathbf{x}|_2^2\phi_{\mathbf{k}}(\mathbf{x}) = -(2|\mathbf{k}|_1+1)\phi_k(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^q,$$

where Δ is the Laplacian operator. As a consequence of the so called Mehler identity, one obtains [36] that

$$\exp\left(-|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{y}|_2^2\right) \exp(-|\mathbf{y}|_2^2/4) = \left(\frac{3}{2\pi}\right)^{-q/2} \sum_{\mathbf{k}\in\mathbb{Z}_+^d} \phi_{\mathbf{k}}(\mathbf{x})\phi_{\mathbf{k}}(\mathbf{y})3^{-|\mathbf{k}|_1/2}. \tag{2.2}$$

A Gaussian network is a network of the form $\mathbf{x} \mapsto \sum_{k=1}^n a_k\left(-|\mathbf{x}-\mathbf{z}_k|_2^2\right)$, where it is convenient to think of $\mathbf{z}_k = \frac{\sqrt{3}}{2}\mathbf{y}_k$. □

3. THE SET-UP AND DEFINITIONS

3.1. Data Spaces

Let $\mathbb{X}$ be a connected, locally compact metric space with metric ρ. For $r>0$, $x \in \mathbb{X}$, we denote

$$\mathbb{B}(x,r) = \{y \in \mathbb{X} : \rho(x,y) \le r\}, \ \Delta(x,r) = \mathsf{closure}(\mathbb{X} \setminus \mathbb{B}(x,r)).$$

If $K \subseteq \mathbb{X}$ and $x \in \mathbb{X}$, we write as usual $\rho(K,x) = \inf_{y\in K}\rho(y,x)$. It is convenient to denote the set $\{x \in \mathbb{X};\ \rho(K,x) \le r\}$ by $\mathbb{B}(K,r)$. The diameter of K is defined by $\mathsf{diam}(K) = \sup_{x,y\in K}\rho(x,y)$.

For a Borel measure ν on $\mathbb{X}$ (signed or positive), we denote by $|\nu|$ its total variation measure defined for Borel subsets $K \subset \mathbb{X}$ by

$$|\nu|(K) = \sup_{\mathcal{U}} \sum_{U\in\mathcal{U}} |\nu(U)|,$$

where the supremum is over all countable measurable partitions $\mathcal{U}$ of K. In the sequel, the term measure will mean a signed or positive, complete, sigma-finite, Borel measure. Terms, such as measurable will mean Borel measurable. If $f : \mathbb{X} \to \mathbb{R}$ is measurable, $K \subset \mathbb{X}$ is measurable, and ν is a measure, we define[2]

$$\|f\|_{p,\nu,K} = \begin{cases} \left\{\left\{\int_K |f(x)|^p d|\nu|(x)\right\}^{1/p}, & \text{if } 1 \le p < \infty, \\ |\nu|\text{–ess}\sup\limits_{x\in K} |f(x)|, & \text{if } p = \infty. \end{cases}$$

The symbol $L^p(\nu, K)$ denotes the set of all measurable functions f for which $\|f\|_{p,\nu,K} < \infty$, with the usual convention that two functions are considered equal if they are equal $|\nu|$-almost everywhere on K. The set $C_0(K)$ denotes the set of all uniformly continuous functions on K vanishing at ∞. In the case when $K = \mathbb{X}$, we will omit the mention of K, unless it is necessary to mention it to avoid confusion.

We fix a non-decreasing sequence $\{\lambda_k\}_{k=0}^\infty$, with $\lambda_0 = 0$ and $\lambda_k \uparrow \infty$ as $k \to \infty$. We also fix a positive sigma-finite Borel measure μ^* on $\mathbb{X}$, and a system of orthonormal functions

[2] $|\nu|\text{–ess}\sup_{x\in K} |f(x)| = \inf\left\{t : |\nu|\left(\{x \in K : |f(x)| > t\}\right) = 0\right\}$

$\{\phi_k\}_{k=0}^\infty \subset L^1(\mu^*, \mathbb{X}) \cap C_0(\mathbb{X})$, such that $\phi_0(x) > 0$ for all $x \in \mathbb{X}$. We define

$$\Pi_n = span\,\{\phi_k : \lambda_k < n\}, \qquad n > 0. \tag{3.1}$$

It is convenient to write $\Pi_n = \{0\}$ if $n \le 0$ and $\Pi_\infty = \bigcup_{n>0} \Pi_n$. It will be assumed in the sequel that Π_∞ is dense in C_0 (and, thus, in every L^p, $1 \le p < \infty$). We will often refer to the elements of Π_∞ as **diffusion polynomials** in keeping with [13].

Definition 3.1. *We will say that a sequence $\{a_n\}$ (or a function $F:[0,\infty) \to \mathbb{R}$) is **fast decreasing** if $\lim_{n\to\infty} n^S a_n = 0$ (respectively, $\lim_{x\to\infty} x^S f(x) = 0$) for every $S > 0$. A sequence $\{a_n\}$ has **polynomial growth** if there exist $c_1, c_2 > 0$ such that $|a_n| \le c_1 n^{c_2}$ for all $n \ge 1$, and similarly for functions.*

Definition 3.2. *The space $\mathbb{X}$ (more precisely, the tuple $\Xi = (\mathbb{X}, \rho, \mu^*, \{\lambda_k\}_{k=0}^\infty, \{\phi_k\}_{k=0}^\infty)$) is called a **data space** if each of the following conditions is satisfied.*

1. *For each $x \in \mathbb{X}$, $r > 0$, $\mathbb{B}(x, r)$ is compact.*
2. *(**Ball measure condition**) There exist $q \ge 1$ and $\kappa > 0$ with the following property: for each $x \in \mathbb{X}$, $r > 0$,*

$$\mu^*(\mathbb{B}(x,r)) = \mu^*\left(\{y \in \mathbb{X} : \rho(x,y) < r\}\right) \le \kappa r^q. \tag{3.2}$$

   (In particular, $\mu^*\left(\{y \in \mathbb{X} : \rho(x,y) = r\}\right) = 0$.)

3. *(**Gaussian upper bound**) There exist $\kappa_1, \kappa_2 > 0$ such that for all $x, y \in \mathbb{X}$, $0 < t \le 1$,*

$$\left|\sum_{k=0}^\infty \exp(-\lambda_k^2 t)\phi_k(x)\phi_k(y)\right| \le \kappa_1 t^{-q/2} \exp\left(-\kappa_2 \frac{\rho(x,y)^2}{t}\right). \tag{3.3}$$

4. *(**Essential compactness**) For every $n \ge 1$, there exists a compact set $\mathbb{K}_n \subset \mathbb{X}$ such that the function $n \mapsto \mathrm{diam}(\mathbb{K}_n)$ has polynomial growth, while the functions*

$$n \mapsto \sup_{x \in \mathbb{X}\setminus\mathbb{K}_n} \sum_{\lambda_k < n} \phi_k(x)^2$$

   and

$$n \mapsto \int_{\mathbb{X}\setminus\mathbb{K}_n} \left(\sum_{\lambda_k < n} \phi_k(x)^2\right)^{1/2} d\mu^*(x)$$

   are both fast decreasing. (Necessarily, $n \mapsto \mu^*(\mathbb{K}_n)$ has polynomial growth as well.)

Remark 3.1. We assume without loss of generality that $\mathbb{K}_n \subseteq \mathbb{K}_m$ for all $n < m$ and that $\mu^*(\mathbb{K}_1) > 0$. □

Remark 3.2. If $\mathbb{X}$ is compact, then the first condition as well as the essential compactness condition are automatically satisfied. We may take $\mathbb{K}_n = \mathbb{X}$ for all n. In this case, we will assume tacitly that μ^* is a probability measure, and $\phi_0 \equiv 1$. □

Example 3.1. (**Manifold case**) This example points out that our notion of data space generalizes the set-ups in Examples 2.1–2.4. Let $\mathbb{X}$ be a smooth, compact, connected Riemannian manifold (without boundary), ρ be the geodesic distance on $\mathbb{X}$, μ^* be the Riemannian volume measure normalized to be a probability measure, $\{\lambda_k\}$ be the sequence of eigenvalues of the (negative) Laplace-Beltrami operator on $\mathbb{X}$, and ϕ_k be the eigenfunction corresponding to the eigenvalue λ_k; in particular, $\phi_0 \equiv 1$. If the condition (3.2) is satisfied, then $(\mathbb{X}, \rho, \mu^*, \{\lambda_k\}_{k=0}^\infty, \{\phi_k\}_{k=0}^\infty)$ is a data space. Of course, the assumption of essential compactness is satisfied trivially (see **Appendix B** for the Gaussian upper bound). □

Example 3.2. (**Hermite case**) We illustrate how Example 2.5 is included in our definition of a data space. Accordingly, we assume the set-up as in that example. For $a > 0$, let $\phi_{\mathbf{k},a}(x) = a^{-q/2}\phi_{\mathbf{k}}(a\mathbf{x})$. With $\lambda_{\mathbf{k}} = \sqrt{|\mathbf{k}|_1}$, the system $\Xi_a = (\mathbb{R}^q, \rho, \mu^*, \{\lambda_{\mathbf{k}}\}, \{\phi_{\mathbf{k},a}\})$ is a data space. When $a = 1$, we will omit its mention from the notation in this context. The first two conditions are obvious. The Gaussian upper bound follows by the multivariate Mehler identity [37, Equation 4.27]. The assumption of essential compactness is satisfied with $\mathbb{K}_n = \mathbb{B}(\mathbf{0}, cn)$ for a suitable constant c (cf. [38, Chapter 6]). □

In the rest of this paper, we assume $\mathbb{X}$ to be a data space. Different theorems will require some additional assumptions, two of which we now enumerate. Not every theorem will need all of these; we will state explicitly which theorem uses which assumptions, apart from $\mathbb{X}$ being a data space.

The first of these deals with the product of two diffusion polynomials. We do not know of any situation where it is not satisfied but are not able to prove it in general.

Definition 3.3. *(**Product assumption**) There exists $A^* \ge 1$ and a family $\{R_{j,k,n} \in \Pi_{A^*n}\}$ such that for every $S > 0$,*

$$\lim_{n\to\infty} n^S \left(\max_{\lambda_k, \lambda_j < n,\ p=1,\infty} \|\phi_k\phi_j - R_{j,k,n}\phi_0\|_p\right) = 0. \tag{3.4}$$

*We say that an **strong product assumption** is satisfied if, instead of (3.4), we have for every $n > 0$ and $P, Q \in \Pi_n$, $PQ \in \Pi_{A^*n}$.*

Example 3.3. In the setting of Example 3.2, if $P, Q \in \Pi_n$, then $PQ = R\phi_{\mathbf{0}}$ for some $R \in \Pi_{2n}$. So, the product assumption holds trivially. The strong product assumption does not hold. However, if $P, Q \in \Pi_n$, then $PQ \in \mathrm{span}\{\phi_{\mathbf{k},\sqrt{2}} : \lambda_k < n\sqrt{2}\}$. The manifold case is discussed below in Remark 3.3. □

Remark 3.3. One of the referees of our paper has pointed out three recent references [39–41], on the subject of the product assumption. The first two of these deal with the manifold case (Example 3.1). The paper [41] extends the results in Lu et al. [40] to the case when the functions ϕ_k are eigenfunctions of a more general elliptic operator. Since the results in these two papers are similar qualitatively, we will comment on Lu et al. [40] and Steinerberger [39].

In this remark only, let $K_t(x, y) = \sum_k \exp(-\lambda_k^2 t)\phi_k(x)\phi_k(y)$. Let $\lambda_k, \lambda_j < n$. In Steinerberger [39], Steinerberger relates $E_{An}(2, \phi_k\phi_j)$ [see (3.6) below for definition] with

$$\left\| \int_{\mathbb{X}} K_t(\circ, y)(\phi_k(y) - \phi_k(\circ))(\phi_j(y) - \phi_j(\circ))d\mu^*(y) \right\|_{2,\mu^*}.$$

While this gives some insight into the product assumption, the results are inconclusive about the product assumption as stated. Also, it is hard to verify whether the conditions mentioned in the paper are satisfied for a given manifold.

In Lu et al. [40], it is shown that for any $\epsilon, \delta > 0$, there exists a subspace V of dimension $O_\delta(\epsilon^{-\delta} n^{1+\delta})$ such that for all $\phi_k, \phi_j \in \Pi_n$, $\inf_{P\in V} \|\phi_k\phi_j - P\|_{2,\mu^*} \leq \epsilon$. The subspace V does not have to be Π_{An} for any A. Since the dimension of span$\{\phi_k\phi_j\}$ is $O(n^2)$, the result is meaningful only if $0 < \delta < 1$ and $\epsilon \geq n^{1-1/\delta}$.

In Geller and Pesenson [42, Theorem 6.1], it is shown that the strong product assumption (and, thus, also the product assumption) holds in the manifold case when the manifold is a compact homogeneous manifold. We have extended this theorem in Filbir and Mhaskar [17, Theorem A.1] for the case of eigenfunctions of general elliptic partial differential operators on arbitrary compact, smooth manifolds provided that the coefficient functions in the operator satisfy some technical conditions. □

In our results in section 4, we will need the following condition, which serves the purpose of gradient in many of our earlier theorems on manifolds.

Definition 3.4. *We say that the system Ξ satisfies* ***Bernstein-Lipschitz condition*** *if for every $n > 0$, there exists $B_n > 0$ such that*

$$|P(x) - P(y)| \leq B_n \rho(x, y)\|P\|_\infty, \qquad x, y \in \mathbb{X},\ P \in \Pi_n. \quad (3.5)$$

Remark 3.4. Both in the manifold case and the Hermite case, $B_n = cn$ for some constant $c > 0$. A proof in the Hermite case can be found in Mhaskar [43] and in the manifold case in Filbir and Mhaskar [44]. □

3.2. Smoothness Classes

We define next the smoothness classes of interest here.

Definition 3.5. *A function $w : \mathbb{X} \to \mathbb{R}$ will be called a* ***weight function*** *if $w\phi_k \in C_0(\mathbb{X}) \cap L^1(\mathbb{X})$ for all k. If w is a weight function, we define*

$$E_n(w; p, f) = \min_{P\in\Pi_n} \|f - Pw\|_{p,\mu^*}, \quad n > 0, 1 \leq p \leq \infty, f \in L^p(\mathbb{X}). \quad (3.6)$$

We will omit the mention of w if $w \equiv 1$ on $\mathbb{X}$.

We find it convenient to denote by X^p the space $\{f \in L^p(\mathbb{X}) : \lim_{n\to\infty} E_n(p, f) = 0\}$; i.e., $X^p = L^p(\mathbb{X})$ if $1 \leq p < \infty$ and $X^\infty = C_0(\mathbb{X})$.

Definition 3.6. *Let $1 \leq p \leq \infty$, $\gamma > 0$, and w be a weight function.*
(a) For $f \in L^p(\mathbb{X})$, we define

$$\|f\|_{W_{\gamma,p,w}} = \|f\|_{p,\mu^*} + \sup_{n>0} n^\gamma E_n(w; p, f), \quad (3.7)$$

and note that

$$\|f\|_{W_{\gamma,p,w}} \sim \|f\|_{p,\mu^*} + \sup_{n\in\mathbb{Z}_+} 2^{n\gamma} E_{2^n}(w; p, f). \quad (3.8)$$

The space $W_{\gamma,p,w}$ comprises all f for which $\|f\|_{W_{\gamma,p,w}} < \infty$.
(b) We write $C_w^\infty = \bigcap_{\gamma>0} W_{\gamma,\infty,w}$. If B is a ball in $\mathbb{X}$, $C_w^\infty(B)$ comprises functions in $f \in C_w^\infty$, which are supported on B.
(c) If $x_0 \in \mathbb{X}$, the space $W_{\gamma,p,w}(x_0)$ comprises functions f such that there exists $r > 0$ with the property that, for every $\phi \in C_w^\infty(\mathbb{B}(x_0, r))$, $\phi f \in W_{\gamma,p,w}$.

Remark 3.5. In both the manifold case and the Hermite case, characterizations of the smoothness classes $W_{\gamma,p}$ are available in terms of constructive properties of the functions, such as the number of derivatives, estimates on certain moduli of smoothness or K-functionals, etc. In particular, the class C^∞ coincides with the class of infinitely differentiable functions vanishing at infinity. □

We can now state another assumption that will be needed in studying local approximation.

Definition 3.7. ***(Partition of unity)*** *For every $r > 0$, there exists a countable family $\mathcal{F}_r = \{\psi_{k,r}\}_{k=0}^\infty$ of functions in C^∞ with the following properties:*

1. *Each $\psi_{k,r} \in \mathcal{F}_r$ is supported on $\mathbb{B}(x_k, r)$ for some $x_k \in \mathbb{X}$.*
2. *For every $\psi_{k,r} \in \mathcal{F}_r$ and $x \in \mathbb{X}$, $0 \leq \psi_{k,r}(x) \leq 1$.*
3. *For every $x \in \mathbb{X}$, there exists a finite subset $\mathcal{F}_r(x) \subseteq \mathcal{F}_r$ such that*

$$\sum_{\psi_{k,r}\in\mathcal{F}_r(x)} \psi_{k,r}(y) = 1, \qquad y \in \mathbb{B}(x, r). \quad (3.9)$$

We note some obvious observations about the partition of unity without the simple proof.

Proposition 3.1. *Let $r > 0$, $\mathcal{F}_r$ be a partition of unity.*
(a) Necessarily, $\sum_{\psi_{k,r}\in\mathcal{F}_r(x)} \psi_{k,r}$ is supported on $\mathbb{B}(x, 3r)$.
(b) For $x \in \mathbb{X}$, $\sum_{\psi_{k,r}\in\mathcal{F}_r} \psi_{k,r}(x) = 1$.

The constant convention *In the sequel, $c, c_1, \cdots$ will denote generic positive constants depending only on the fixed quantities under discussion, such as Ξ, q, $\kappa, \kappa_1, \kappa_2$, the various smoothness parameters, and the filters to be introduced. Their value may be different at different occurrences, even within a single formula. The notation $A \sim B$ means $c_1 A \leq B \leq c_2 A$.* □

We end this section by defining a kernel that plays a central role in this theory.

Let $H : [0,\infty) \to \mathbb{R}$ be a compactly supported function. In the sequel, we define

$$\Phi_N(H; x, y) = \sum_{k=0}^{\infty} H(\lambda_k/N)\phi_k(x)\phi_k(y), \qquad N > 0,\ x, y \in \mathbb{X}. \tag{3.10}$$

If $S \geq 1$ is an integer, and H is S times continuously differentiable, we introduce the notation

$$|||H|||_S := \max_{0\leq k\leq S} \max_{x\in\mathbb{R}} |H^{(k)}(x)|.$$

The following proposition recalls an important property of these kernels. Proposition 3.2 is proven in Maggioni and Mhaskar [13] and more recently in much greater generality in Mhaskar [45, Theorem 4.3].

Proposition 3.2. *Let $S > q$ be an integer, $H : \mathbb{R} \to \mathbb{R}$ be an even, S times continuously differentiable, compactly supported function. Then, for every $x, y \in \mathbb{X}$, $N > 0$,*

$$|\Phi_N(H; x, y)| \leq \frac{cN^q |||H|||_S}{\max(1, (N\rho(x,y))^S)}. \tag{3.11}$$

In the sequel, let $h : \mathbb{R} \to [0, 1]$ be a fixed, infinitely differentiable, even function, non-increasing on $[0,\infty)$, with $h(t) = 1$ if $|t| \leq 1/2$ and $h(t) = 0$ if $t \geq 1$. If ν is any measure with a bounded total variation on $\mathbb{X}$, we define

$$\sigma_n(\nu, h; f)(x) = \int_{\mathbb{X}} \Phi_n(h; x, y) f(y) d\nu(y). \tag{3.12}$$

We will omit the mention of h in the notations; e.g., write $\Phi_n(x, y) = \Phi_n(h; x, y)$, and the mention of ν if $\nu = \mu^*$. In particular,

$$\sigma_n(f)(x) = \sum_{k=0}^{\infty} h\left(\frac{\lambda_k}{n}\right)\hat{f}(k)\phi_k(x),$$
$$n > 0,\ x \in \mathbb{X}, f \in L^1(\mathbb{X}) + C_0(\mathbb{X}), \tag{3.13}$$

where for $f \in L^1 + C_0$, we write

$$\hat{f}(k) = \int_{\mathbb{X}} f(y)\phi_k(y) d\mu^*(y) \tag{3.14}$$

.

3.3. Measures

In this section, we describe the terminology involving measures.

Definition 3.8. *Let $d \geq 0$. A measure $\nu \in \mathcal{M}$ will be called* ***d–regular*** *if*

$$|\nu|(\mathbb{B}(x, r)) \leq c(r + d)^q, \qquad x \in \mathbb{X}. \tag{3.15}$$

The infimum of all constants c that work in (3.15) will be denoted by $|||\nu|||_{R,d}$, and the class of all d-regular measures will be denoted by $\mathcal{R}_d$.

For example, μ^* itself is in R_0 with $|||\mu^*|||_{R,0} \leq \kappa$ [cf. (3.2)]. More generally, if $w \in C_0(\mathbb{X})$ then the measure $wd\mu^*$ is R_0 with $|||\mu^*|||_{R,0} \leq \kappa\|w\|_{\infty,\mu^*}$.

Definition 3.9. *(a) A sequence $\{\nu_n\}$ of measures on $\mathbb{X}$ is called an* ***admissible quadrature measure sequence*** *if the sequence $\{|\nu_n|(\mathbb{X})\}$ has polynomial growth and*

$$\int_{\mathbb{X}} P d\nu_n = \int_{\mathbb{X}} P d\mu^*, \qquad P \in \Pi_n,\ n \geq 1. \tag{3.16}$$

(b) A sequence $\{\nu_n\}$ of measures on $\mathbb{X}$ is called an ***admissible product quadrature measure sequence*** *if the sequence $\{|\nu_n|(\mathbb{X})\}$ has polynomial growth and*

$$\int_{\mathbb{X}} P_1P_2 d\nu_n = \int_{\mathbb{X}} P_1P_2 d\mu^*, \qquad P_1, P_2 \in \Pi_n,\ n \geq 1. \tag{3.17}$$

(c) By abuse of terminology, we will say that a measure ν_n is an ***admissible quadrature measure*** *(respectively, an* ***admissible product quadrature measure****) of order n if $|\nu_n| \leq c_1 n^c$ (with constants independent of n) and (3.16) [respectively, (3.17)] holds.*

In the case when $\mathbb{X}$ is compact, a well-known theorem called Tchakaloff's theorem [46, Exercise 2.5.8, p. 100] shows the existence of admissible product quadrature measures (even finitely supported probability measures). However, in order to construct such measures, it is much easier to prove the existence of admissible quadrature measures, as we will do in Theorem 7.1, and then use one of the product assumptions to derive admissible product quadrature measures.

Example 3.4. In the manifold case, let the strong product assumption hold as in Remark 3.3. If $n \geq 1$ and $\mathcal{C} \subset \mathbb{X}$ is a finite subset satisfying the assumptions of Theorem 7.1, then the theorem asserts the existence of an admissible quadrature measure supported on $\mathcal{C}$. If $\{\nu_n\}$ is an admissible quadrature measure sequence, then $\{\nu_{A^*n}\}$ is an admissible product quadrature measure sequence. In particular, there exist finitely supported admissible product quadrature measures of order n for every $n \geq 1$. □

Example 3.5. We consider the Hermite case as in Example 3.2. For every $a > 0$ and $n \geq 1$, Theorem 7.1 applied with the system Ξ_a yields admissible quadrature measures of order n supported on finite subsets of $\mathbb{R}^q$ (in fact, of $[-cn, cn]^q$ for an appropriate c). In particular, an admissible quadrature measure of order $n\sqrt{2}$ for $\Xi_{\sqrt{2}}$ is an admissible product quadrature measure of order n for $\Xi = \Xi_1$. □

3.4. Eignets

The notion of an eignet defined below is a generalization of the various kernels described in the examples in section 2.

Definition 3.10. *A function $b : [0,\infty) \to (0,\infty)$ is called a* ***smooth mask*** *if b is non-increasing, and there exists $B^* = B^*(b) \geq 1$ such that the mapping $t \mapsto b(B^*t)/b(t)$ is fast decreasing. A function $G : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is called a* ***smooth kernel*** *if there exists* *a measurable function* $W = W(G) : \mathbb{X} \to \mathbb{R}$ *such that we have a formal expansion (with a smooth mask b)*

$$W(y)G(x,y) = \sum_k b(\lambda_k)\phi_k(x)\phi_k(y), \qquad x, y \in \mathbb{X}. \tag{3.18}$$

If $m \geq 1$ *is an integer, an* ***eignet*** *with m neurons is a function of the form* $x \mapsto \sum_{k=1}^m a_k G(x, y_k)$ *for* $y_k \in \mathbb{X}$.

Example 3.6. In the manifold case, the notion of eignet includes all the examples stated in section 2 with $W \equiv 1$, except for the example of smooth ReLU function described in Example 2.3. In the Hermite case, (2.2) shows that the kernel $G(\mathbf{x}, \mathbf{y}) = \exp\left(-|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{y}|_2^2\right)$ defined on $\mathbb{R}^q \times \mathbb{R}^q$ is a smooth kernel, with $\lambda_\mathbf{k} = |\mathbf{k}|_1$, $\phi_\mathbf{k}$ as in Example 2.5, and $b(t) = \left(\frac{3}{2\pi}\right)^{-q/2} 3^{-t/2}$. The function W here is $W(\mathbf{y}) = \exp(-|\mathbf{y}|_2^2/4)$. □

Remark 3.6. It is possible to relax the conditions on the mask in Definition 3.10. Firstly, the condition that b should be non-increasing is made only to simplify our proofs. It is not difficult to modify them without this assumption. Secondly, let $b_0 : [0, \infty) \to \mathbb{R}$ satisfy $|b_0(t)| \leq b_1(t)$ for a smooth mask b_1 as stipulated in that definition. The function $b_2 = b + 2b_1$ is then a smooth mask and so is b_1. Let $G_j(x,y) = \sum_{k=0}^\infty b_j(\lambda_k)\phi_k(x)\phi_k(y)$, $j = 0, 1, 2$. Then $G_0(x,y) = G_2(x,y) - 2G_1(x,y)$. Therefore, all of the results in sections 4 and 8 can be applied once with G_2 and once with G_1 to obtain a corresponding result for G_0 with different constants. For this reason, we will simplify our presentation by assuming the apparently restrictive conditions stipulated in Definition 3.10. In particular, this includes the example of the smooth ReLU network described in Example 2.3. □

Definition 3.11. *Let* ν *be a measure on* $\mathbb{X}$ *(signed or having bounded variation), and* $G \in C_0(\mathbb{X} \times \mathbb{X})$. *We define*

$$\mathcal{D}_{G,n}(x,y) = \sum_{k=0}^\infty h\left(\lambda_k/n\right) b(\lambda_k)^{-1}\phi_k(x)\phi_k(y), n \geq 1, \; x, y \in \mathbb{X}, \tag{3.19}$$

and

$$\mathbb{G}_n(\nu; x, y) = \int_\mathbb{X} G(x,z)W(z)\mathcal{D}_{G,n}(z,y)d\nu(z). \tag{3.20}$$

Remark 3.7. Typically, we will use an approximate product quadrature measure sequence in place of the measure ν, where each of the measures in the sequence is finitely supported, to construct a sequence of networks. In the case when $\mathbb{X}$ is compact, Tchakaloff's theorem shows that there exists an approximate product quadrature measure of order m supported on $(\mathsf{dim}(\Pi_m) + 1)^2$ points. Using this measure in place of ν, one obtains a pre-fabricated eignet $\mathbb{G}_n(\nu)$ with $(\mathsf{dim}(\Pi_m) + 1)^2$ neurons. However, this is not an actual construction. In the presence of the product assumption, Theorem 7.1 leads to the pre-fabricated networks $\mathbb{G}_n$ in a constructive manner with the number of neurons as stipulated in that theorem. □

4. MAIN RESULTS

In this section, we assume the Bernstein-Lipschitz condition (Definition 3.4) in all the theorems. We note that the measure μ^* may not be a probability measure. Therefore, we take the help of an auxiliary function f_0 to define a probability measure as follows. Let $f_0 \in C_0(\mathbb{X})$, $f_0 \geq 0$ for all $x \in \mathbb{X}$, and $d\nu^* = f_0 d\mu^*$ be a probability measure. Necessarily, ν^* is 0-regular, and $k : \|\!|\nu^*|\!\|_{R,0} \leq k\|f_0\|_{\infty,\mu^*}$. We assume noisy data of the form (y, ϵ), with a joint probability distribution τ defined for Borel subsets of $\mathbb{X} \times \Omega$ for some measure space Ω, and with ν^* being the marginal distribution of y with respect to τ. Let $\mathcal{F}(y, \epsilon)$ be a random variable following the law τ, and denote

$$f(y) = \mathbb{E}_\tau(\mathcal{F}(y, \epsilon)|y). \tag{4.1}$$

It is easy to verify using Fubini's theorem that if $\mathcal{F}$ is integrable with respect to τ, then, for any $x \in \mathbb{X}$,

$$\mathbb{E}_\tau(\mathcal{F}(y,\epsilon)\Phi_n(x,y)) = \sigma_n(\nu^*; f)(x) := \int_\mathbb{X} f(y)\Phi_n(x,y)d\nu^*(y). \tag{4.2}$$

Let Y be a random sample from τ, and $\{\nu_n\}$ be an admissible product quadrature sequence in the sense of Definition 3.9. We define [cf. (3.20)]

$$\begin{aligned}\mathcal{G}_n(Y; \mathcal{F})(x) &= \mathcal{G}_n(\nu_{B^*n}, Y; \mathcal{F})(x) \\ &= \frac{1}{|Y|}\sum_{(y,\epsilon) \in Y} \mathcal{F}(y,\epsilon)\mathbb{G}_n(\nu_{B^*n}; x, y), \quad x \in \mathbb{X}, \; n = 1, 2, \cdots\end{aligned} \tag{4.3}$$

where B^* is as in Definition 3.10.

Remark 4.1. We note that the networks $\mathbb{G}_n$ are prefabricated independently of the data. The network $\mathcal{G}_n$ therefore has only $|Y|$ terms depending upon the data. □

Our first theorem describes local function recovery using local sampling. We may interpret it in the spirit of distributed learning as in Chui et al. [24] and Lin et al. [26], where we are taking a linear combination of pre-fabricated networks $\mathbb{G}_n$ using the function values themselves as the coefficients. The networks $\mathbb{G}_n$ have essentially the same localization property as the kernels Φ_n (cf. Theorem 8.2).

Theorem 4.1. *Let* $x_0 \in \mathbb{X}$ *and* $r > 0$. *We assume the partition of unity and find a function* $\psi \in C^\infty$ *supported on* $\mathbb{B}(x_0, 3r)$, *which is equal to 1 on* $\mathbb{B}(x_0, r)$, $\mathfrak{m} = \int_\mathbb{X} \psi d\mu^*$, *and let* $f_0 = \psi/\mathfrak{m}$, $d\nu^* = f_0 d\mu^*$. *We assume the rest of the set-up as described. If* $f_0 f \in W_{\gamma,\infty}$, *then for* $0 < \delta < 1$, *and* $|Y| \geq cn^{q+2\gamma}r^q \log(nB_n/\delta)$,

$$\begin{aligned}\mathsf{Prob}_\tau\Bigg(\Bigg\{\Bigg\|\frac{\mathfrak{m}}{|Y|}\sum_{(y,\epsilon)\in Y} \mathcal{F}(y,\epsilon)\mathbb{G}_n(\nu_{B^*n}; \circ, y) \\ -f\Bigg\|_{\infty,\mu^*,\mathbb{B}(x_0,r)} \geq c_3 n^{-\gamma}\Bigg\}\Bigg) \\ \leq \delta.\end{aligned} \tag{4.4}$$

Remark 4.2. If $\{y_1, \cdots, y_M\}$ is a random sample from some probability measure supported on $\mathbb{X}$, $s = \sum_{\ell=1}^{M} f_0(y_\ell)$, and we construct a sub-sample using the distribution that associates the mass $f_0(y_j)/s$ with each y_j, then the probability of selecting points outside of the support of f_0 is 0. This leads to a sub-sample Y. If $M \geq cn^{q+2\gamma} \log(nB_n/\delta)$, then the Chernoff bound, Proposition B.1(b), can be used to show that $|Y|$ is large, as stipulated in Theorem 4.1. □

Next, we state two inverse theorems. Our first theorem obtains accuracy on the estimation of the density f_0 using eignets instead of positive kernels.

Theorem 4.2. *With the set-up as in Theorem 8.3, let $\gamma > 0$, $f_0 \in W_{\gamma,\infty}$, and*

$$|Y| \geq \|f_0\|_{\infty,\mu^*} n^{q+2\gamma} \log\left(\frac{nB_n}{\delta}\right).$$

Then, with $\mathcal{F} \equiv 1$,

$$\mathsf{Prob}_\tau\left(\left\{\left\|\frac{1}{|Y|}\sum_{(y,\epsilon)\in Y} \mathbb{G}_n(\nu_{B^*n}; \circ, y) - f_0\right\|_\infty \geq c_3 n^{-\gamma}\right\}\right) \leq \delta. \tag{4.5}$$

Remark 4.3. Unlike density estimation using positive kernels, there is no inherent limit on the accuracy predicted by (4.5) on the estimation of f_0. □

The following theorem gives a complete characterization of the local smoothness classes using eignets. In particular, Part (b) of the following theorem gives a solution to the inverse problem of determining what smoothness class the target function belongs to near each point of $\mathbb{X}$. In theory, this leads to a **data-based detection** of singularities and sparsity analogous to what is assumed in Chui et al. [24] but in a much more general setting.

Theorem 4.3. *Let $f_0 \in C_0(\mathbb{X})$, $f_0(x) \geq 0$ for all $x \in \mathbb{X}$, and $d\nu^* = f_0 d\mu^*$ be a probability measure, τ, $\mathcal{F}$, and let f be as described above. We assume the partition of unity and the product assumption. Let $S \geq q+2$, $0 < \gamma \leq S$, $x_0 \in \mathbb{X}$, $0 < \delta < 1$. For each $j \geq 0$, suppose that Y_j is a random sample from τ with $|Y_j| \geq 2c_1 2^{j(q+2S)} |\!|\!|\nu^*|\!|\!|_{R,0} \log(c2^{2j}B_{2^j}/\delta)$. Then with τ-probability $\geq 1-\delta$,*
(a) If $f_0 f \in W_{\gamma,\infty}(x_0)$ then there exists a ball $\mathbb{B}$ centered at $\boldsymbol{x}_0$ such that

$$\sup_{j\geq 1} 2^{j\gamma} \|\mathcal{G}_{2^j}(Y_j; \mathcal{F}) - \mathcal{G}_{2^{j-1}}(Y_j; \mathcal{F})\|_{\infty,\mu^*,\mathbb{B}} < \infty. \tag{4.6}$$

(b) If there exists a ball $\mathbb{B}$ centered at x_0 for which (4.6) holds, then $f_0 f \in W_{\gamma,\infty,\phi_0}(x_0)$.

5. PREPARATORY RESULTS

We prove a lower bound on $\mu^*(\mathbb{B}(x,r))$ for $x \in \mathbb{X}$ and $0 < r \leq 1$ (cf. [47]).

Proposition 5.1. *We have*

$$\mu^*(\mathbb{B}(x,r)) \geq cr^q, \qquad 0 < r \leq 1,\ x \in \mathbb{X}. \tag{5.1}$$

In order to prove the proposition, we recall a lemma, proved in Mhaskar [14, Proposition 5.1].

Lemma 5.1. *Let $\nu \in R_d$, $N > 0$. If $g_1 : [0,\infty) \to [0,\infty)$ is a non-increasing function, then, for any $N > 0$, $r > 0$, $x \in \mathbb{X}$,*

$$N^q \int_{\Delta(x,r)} g_1(N\rho(x,y))d|\nu|(y) \leq$$
$$c\frac{2^q(1+(d/r)^q)q}{1-2^{-q}} |\!|\!|\nu|\!|\!|_{R,d} \int_{rN/2}^{\infty} g_1(u)u^{q-1}du. \tag{5.2}$$

Proof of Proposition 5.1.

Let $x \in \mathbb{X}$, $r > 0$ be fixed in this proof, although the constants will not depend upon these. In this proof, we write

$$K_t(x,y) = \sum_{k=0}^{\infty} \exp(-\lambda_k^2 t)\phi_k(x)\phi_k(y).$$

The Gaussian upper bound (3.3) shows that for $t > 0$,

$$\int_{\Delta(x,r)} |K_t(x,y)|d\mu^*(y) \leq \kappa_1 t^{-q/2} \int_{\Delta(x,r)} \exp(-\kappa_2\rho(x,y)^2/t)d\mu^*(y). \tag{5.3}$$

Using Lemma 5.1 with $d = 0$, $d\nu = d\mu^*$, $g_1(u) = \exp(-u^2)$, $N = \sqrt{\kappa_2/t}$, we obtain for $r^2/t \geq (q-2)/\kappa_2$:

$$\begin{aligned}
&\int_{\Delta(x,r)} |K_t(x,y)|d\mu^*(y)\\
&\leq c\int_{Nr/2}^{\infty} u^{q-1}\exp(-u^2)du = c_1 \int_{(Nr/2)^2}^{\infty} u^{q/2-1}e^{-u}du\\
&\leq c_2(r^2/t)^{(q-2)/2}\exp(-\kappa_2 r^2/(4t)).
\end{aligned} \tag{5.4}$$

Therefore, denoting in this proof only that $\kappa_0 = \|\phi_0\|_\infty$, we obtain that

$$\begin{aligned}
1 &= \int_{\mathbb{X}} K_t(x,y)\phi_0(y)d\mu^*(y) \leq \kappa_0 \int_{\mathbb{X}} |K_t(x,y)|d\mu^*(y)\\
&\leq \kappa_0\kappa_2 t^{-q/2}\mu^*(\mathbb{B}(x,r)) + c_3(r^2/t)^{(q-2)/2}\exp(-\kappa_2 r^2/(4t).
\end{aligned} \tag{5.5}$$

We now choose $t \sim r^2$ so that $c_3(r^2/t)^{(q-2)/2}\exp(-\kappa_3 r^2/(4t)) \leq 1/2$ to obtain (5.1) for $r \leq c_4$. The estimate is clear for $c_4 < r \leq 1$. □

Next, we prove some results about the system $\{\phi_k\}$.

Lemma 5.2. *For $n \geq 1$, we have*

$$\sum_{\lambda_k < n} \phi_k(x)^2 \leq cn^q, \qquad x \in \mathbb{X}. \tag{5.6}$$

and

$$\mathsf{dim}(\Pi_n) \leq cn^q\mu^*(\mathbb{K}_n). \tag{5.7}$$

In particular, the function $n \mapsto \mathsf{dim}(\Pi_n)$ has polynomial growth.

PROOF. The Gaussian upper bound with $x = y$ implies that

$$\sum_{k=0}^{\infty}\exp(-\lambda_k^2 t)\phi_k(x)^2 \le ct^{-q/2}, \qquad 0 < t \le 1,\ x \in \mathbb{X}.$$

The estimate (5.6) follows from a Tauberian theorem [44, Proposition 4.1]. The essential compactness now shows that for any $R > 0$,

$$\int_{\mathbb{X}\setminus\mathbb{K}_n}\sum_{\lambda_k<n}\phi_k(x)^2 d\mu^*(x) \le \left\{\sup_{x\in\mathbb{X}\setminus\mathbb{K}_n}\sum_{\lambda_k<n}\phi_k(x)^2\right\}^{1/2}$$

$$\int_{\mathbb{X}\setminus\mathbb{K}_n}\left(\sum_{\lambda_k<n}\phi_k(x)^2\right)^{1/2} d\mu^*(x) \le cn^{-R}.$$

In particular,

$$\begin{aligned}\mathrm{dim}(\Pi_n) &= \int_{\mathbb{X}}\sum_{\lambda_k<n}\phi_k(x)^2 d\mu^*(x)\\ &\le \int_{\mathbb{K}_n}\sum_{\lambda_k<n}\phi_k(x)^2 d\mu^*(x) + cn^{-R} \le cn^q\mu^*(\mathbb{K}_n).\end{aligned}$$

□

Next, we prove some properties of the operators σ_n and diffusion polynomials. The following proposition follows easily from Lemma 5.1 and Proposition 3.2. (cf. [14, 48]).

Proposition 5.2. *Let S, H be as in Proposition 3.2, $d > 0$, $\nu \in \mathcal{R}_d$, and $x \in \mathbb{X}$.*
(a) If $r \ge 1/N$, then

$$\int_{\Delta(x,r)}|\Phi_N(H;x,y)|d|\nu|(y) \le c(1+(dN)^q)(rN)^{-S+q}\|\!|\nu\|\!|_{R,d}|||H|||_S. \tag{5.8}$$

(b) We have

$$\int_{\mathbb{X}}|\Phi_N(H;x,y)|d|\nu|(y) \le c(1+(dN)^q)\|\!|\nu\|\!|_{R,d}|||H|||_S, \tag{5.9}$$

$$\|\Phi_N(H;x,\circ)\|_{\nu;\mathbb{X},p} \le cN^{q/p'}(1+(dN)^q)^{1/p}\|\!|\nu\|\!|_{R,d}^{1/p}|||H|||_S, \tag{5.10}$$

and

$$\left\|\int_{\mathbb{X}}|\Phi_N(H;\circ,y)|d|\nu|(y)\right\|_p \le c(1+(dN)^q)^{1/p'}\|\!|\nu\|\!|_{R,d}^{1/p'}(|\nu|(\mathbb{X}))^{1/p}|||H|||_S. \tag{5.11}$$

The following lemma is well-known; a proof is given in Mhaskar [15, Lemma 5.3].

Lemma 5.3. *Let (Ω_1,ν), (Ω_2,τ) be sigma–finite measure spaces, $\Psi : \Omega_1\times\Omega_2 \to \mathbb{R}$ be $\nu\times\tau$–integrable,*

$$\begin{aligned}M_\infty &:= \nu\text{–}ess\ sup_{x\in\Omega_1}\int_{\Omega_2}|\Psi(x,y)|d\tau(y) < \infty,\\ M_1 &:= \tau\text{–}ess\ sup_{y\in\Omega_2}\int_{\Omega_1}|\Psi(x,y)|d\nu(x) < \infty,\end{aligned} \tag{5.12}$$

and formally, for τ–measurable functions $f : \Omega_2 \to \mathbb{R}$,

$$T(f,x) := \int_{\Omega_2} f(y)\Psi(x,y)d\tau(y), \qquad x \in \Omega_1.$$

Let $1 \le p \le \infty$. If $f \in L^p(\tau;\Omega_2)$ then $T(f,x)$ is defined for ν–almost all $x \in \Omega_1$, and

$$\|Tf\|_{\nu;\Omega_1,p} \le M_1^{1/p}M_\infty^{1/p'}\|f\|_{\tau;\Omega_2,p}, \qquad f \in L^p(\Omega_2,\tau). \tag{5.13}$$

Theorem 5.1. *Let $n > 0$. If $P \in \Pi_{n/2}$, then $\sigma_n(P) = P$. Also, for any p with $1 \le p \le \infty$,*

$$\|\sigma_n(f)\|_p \le c\|f\|_p, \qquad f \in L^p. \tag{5.14}$$

If $1 \le p \le \infty$, and $f \in L^p(\mathbb{X})$, then

$$E_n(p,f) \le \|f - \sigma_n(f)\|_{p,\mu^*} \le cE_{n/2}(p,f). \tag{5.15}$$

PROOF. The fact that $\sigma_n(P) = P$ for all $P \in \Pi_{n/2}$ is verified easily using the fact that $h(t) = 1$ for $0 \le t \le 1/2$. Using (5.9) with μ^* in place of $|\nu|$ and 0 in place of d, we see that

$$\sup_{x\in\mathbb{X}}\int_{\mathbb{X}}|\Phi_n(x,y)|d\mu^*(y) \le c.$$

The estimate (5.14) follows using Lemma 5.3. The estimate (5.15) is now routine to prove. □

Proposition 5.3. *For $n \ge 1$, $P \in \Pi_n$, $1 \le p \le \infty$, and $S > 0$, we have*

$$\|P\|_{p,\mu^*,\mathbb{X}\setminus\mathbb{K}_{2n}} \le c(S)n^{-S}\|P\|_{p,\mu^*,\mathbb{X}}. \tag{5.16}$$

PROOF. In this proof, all constants will depend upon S. Using Schwarz inequality and essential compactness, it is easy to deduce that

$$\begin{aligned}\sup_{x\in\mathbb{X}\setminus\mathbb{K}_{2n}}\int_{\mathbb{X}}|\Phi_{2n}(x,y)|d\mu^*(y) &\le c_1 n^{-S},\\ \sup_{y\in\mathbb{X}}\int_{\mathbb{X}\setminus\mathbb{K}_{2n}}|\Phi_{2n}(x,y)|d\mu^*(x) &\le c_1 n^{-S}.\end{aligned} \tag{5.17}$$

Therefore, a use of Lemma 5.3 shows that

$$\|\sigma_{2n}(f)\|_{p,\mu^*,\mathbb{X}\setminus\mathbb{K}_{2n}} \le cn^{-S}\|f\|_p.$$

We use P in place of f to obtain (5.16). □

Proposition 5.4. *Let $n \ge 1$, $P \in \Pi_n$, $0 < p < r \le \infty$. Then*

$$\|P\|_r \le cn^{q(1/p-1/r)}\|P\|_p, \qquad \|P\|_p \le c\mu^*(\mathbb{K}_{2n})^{1/p-1/r}\|P\|_r. \tag{5.18}$$

PROOF. The first part of (5.18) is proved in Mhaskar [15, Lemma 5.4]. In that paper, the measure μ^* is assumed to be a probability measure, but this assumption was not used in this proof. The second estimate follows easily from Proposition 5.3. □

Lemma 5.4. *Let $R, n > 0$, $P_1, P_2 \in \Pi_n$, $1 \le p, r, s \le \infty$. If the product assumption holds, then*

$$E_{A^*n}(\phi_0; p, P_1P_2) \le cn^{-R}\|P_1\|_r\|P_2\|_s. \tag{5.19}$$

PROOF. In view of essential compactness, Proposition 5.4 implies that for any $P \in \Pi_n$, $1 \le r \le \infty$, $\|P\|_2 \le c_1 n^c\|P\|_r$. Therefore, using Schwarz inequality, Parseval identity, and Lemma 5.2, we conclude that

$$\sum_k |\hat{P}(k)| \le (\mathsf{dim}(\Pi_n))^{1/2}\|P\|_2 \le c_1 n^c\|P\|_r. \tag{5.20}$$

Now, the product assumption implies that for $p = 1, \infty$, and $\lambda_k, \lambda_j < n$, there exists $R_{j,k,n} \in \Pi_{A^*n}$ such that for any $R > 0$,

$$\|\phi_k\phi_j - R_{j,k,n}\phi_0\|_p \le cn^{-R-2c}, \tag{5.21}$$

where c is the constant appearing in (5.20). The convexity inequality

$$\|f\|_p \le \|f\|_\infty^{1/p'}\|f\|_1^{1/p}$$

shows that (5.21) is valid for all p, $1 \le p \le \infty$. So, using (5.20), we conclude that

$$\left\|P_1P_2 - \sum_{k,j}\widehat{P_1}(k)\widehat{P_2}(k)R_{j,k,n}\phi_0\right\|_p \le cn^{-R-2c}\left(\sum_k |\widehat{P_1}(k)|\right)$$
$$\left(\sum_k |\widehat{P_2}(k)|\right) \le cn^{-R}\|P_1\|_r\|P_2\|_s.$$

□

6. LOCAL APPROXIMATION BY DIFFUSION POLYNOMIALS

In the sequel, we write $g(t) = h(t) - h(2t)$, and

$$\tau_j(f) = \begin{cases}\sigma_1(f), & \text{if } j = 0,\\ \sigma_{2^j}(f) - \sigma_{2^{j-1}}(f), & \text{if } j = 1, 2, \cdots.\end{cases} \tag{6.1}$$

We note that

$$\tau_j(f)(x) = \sigma_{2^j}(\mu^*, g; f)(x) = \int_{\mathbb{X}} f(y)\Phi_{2^j}(g; x, y)d\mu^*(y), j = 1, 2, \cdots. \tag{6.2}$$

It is clear from Theorem 5.1 that for any p, $1 \le p \le \infty$,

$$f = \sum_{j=0}^{\infty}\tau_j(f), \qquad f \in X^p, \tag{6.3}$$

with convergence in the sense of L^p.

Theorem 6.1. *Let $1 \le p \le \infty$, $\gamma > 0$, $f \in X^p$, $x_0 \in \mathbb{X}$. We assume the partition of unity and the product assumption.*
(a) If $\mathbb{B}$ is a ball centered at x_0, then

$$\sup_{n\ge 0} 2^{n\gamma}\|f - \sigma_{2^n}(f)\|_{p,\mu^*,\mathbb{B}} \sim \sup_{j\ge 0} 2^{j\gamma}\|\tau_j(f)\|_{p,\mu^*,\mathbb{B}}. \tag{6.4}$$

(b) If there exists a ball $\mathbb{B}$ centered at x_0 such that

$$\sup_{n\ge 0} 2^{n\gamma}\|f - \sigma_{2^n}(f)\|_{p,\mu^*,\mathbb{B}} \sim \sup_{j\ge 0} 2^{j\gamma}\|\tau_j(f)\|_{p,\mu^*,\mathbb{B}} < \infty, \tag{6.5}$$

then $f \in W_{\gamma,p,\phi_0}(x_0)$.
(c) If $f \in W_{\gamma,p}(x_0)$, then there exists a ball $\mathbb{B}$ centered at x_0 such that (6.5) holds.

Remark 6.1. In the manifold case (Example 3.1), $\phi_0 \equiv 1$. So, the statements (b) and (c) in Theorem 6.1 provide necessary and sufficient conditions for $f \in W_{\gamma,p}(x_0)$ in terms of the local rate of convergence of the globally defined operator $\sigma_n(f)$ and the growth of the local norms of the operators τ_j, respectively In the Hermite case (Example 3.2), it is shown in Mhaskar [49] that $f \in W_{\gamma,p,\phi_0}$ if and only if $f \in W_{\gamma,p}$. Therefore, the statements (b) and (c) in Theorem 6.1 provide similar necessary and sufficient conditions for $f \in W_{\gamma,p}(x_0)$ in this case as well. □

The proof of Theorem 6.1 is routine, but we sketch a proof for the sake of completeness.

PROOF OF THEOREM 6.1

Part (a) is easy to prove using the definitions.
In the rest of this proof, we fix $S > \gamma + q + 2$. To prove part (b), let $\phi \in C^\infty$ be supported on $\mathbb{B}$. Then there exists $\{R_n \in \Pi_{2^n}\}_{n=0}^\infty$ such that

$$\|\phi - R_n\|_\infty \le c(\phi)2^{-nS}. \tag{6.6}$$

Further, Lemma 5.4 yields a sequence $\{Q_n \in \Pi_{A^*2^n}\}$ such that

$$\|R_n\sigma_{2^n}(f) - \phi_0 Q_n\|_p \le c2^{-nS}\|R_n\|_\infty\|\sigma_{2^n}(f)\|_p \le c(\phi)2^{-nS}\|f\|_p. \tag{6.7}$$

Hence,

$$\begin{aligned}
&E_{A^*2^n}(\phi_0; p, f\phi)\\
&\le \|f\phi - \phi_0 Q_n\|_p \le c(\phi)2^{-nS}\|f\|_p + \|f\phi - \sigma_{2^n}(f)R_n\|_p\\
&\le c(\phi)2^{-nS}\|f\|_p + \|(f - \sigma_{2^n}(f))\phi\|_p + \|\sigma_{2^n}(f)(\phi - R_n)\|_p\\
&\le c(\phi)\left\{2^{-nS}\|f\|_p + \|f - \sigma_{2^n}(f)\|_{p,\mu^*,\mathbb{B}} + \|\sigma_{2^n}(f)\|_p\|\phi - R_n\|_\infty\right\}\\
&\le c(\phi)2^{-nS}\|f\|_p + c(\phi, f)(A^*2^{-n})^\gamma.
\end{aligned}$$

Thus, $f\phi \in W_{\gamma,p,\phi_0}$ for every $\phi \in C^\infty$ supported on $\mathbb{B}$, and part (b) is proved.

To prove part (c), we observe that there exists $r > 0$ such that for any $\phi \in C^\infty(\mathbb{B}(x_0, 6r))$, $f\phi \in W_{\gamma,p}$. Using partition of unity [cf. Proposition 3.1(a)], we find $\psi \in C^\infty(\mathbb{B}(x_0, 6r))$ such that $\psi(x) = 1$ for all $x \in \mathbb{B}(x_0, 2r)$, and we let $\mathbb{B} = \mathbb{B}(x_0, r)$. In view of Proposition 3.2, $|\Phi_{2^n}(x,y)| \le c(r)2^{-n(S-q)}$ for all $x \in \mathbb{B}$ and $y \in \mathbb{X}\setminus\mathbb{B}(x_0,2r)$. Hence,

$$\begin{aligned}\|\sigma_{2^n}((1-\psi)f)\|_p &\le \left\|\int_{\mathbb{X}} |(1-\psi(y))f(y)\Phi_{2^n}(\circ,y)|d\mu^*(y)\right\|_p\\ &= \left\|\int_{\mathbb{X}\setminus\mathbb{B}(x_0,2r)} |(1-\psi(y))f(y)\Phi_{2^n}(\circ,y)|d\mu^*(y)\right\|_p\\ &\le c(\psi,r)2^{-n(S-q)}\|f\|_p. \qquad (6.8)\end{aligned}$$

Recalling that $\psi(x)=1$ for $x\in\mathbb{B}$ and $S-q\ge\gamma+2$, we deduce that

$$\begin{aligned}\|f-\sigma_{2^n}(f)\|_{p,\mu^*,\mathbb{B}} &= \|\psi f-\sigma_{2^n}(f)\|_{p,\mu^*,\mathbb{B}}\\ &\le \|\psi f-\sigma_{2^n}(\psi f)\|_{p,\mu^*,\mathbb{B}}+\|\sigma_{2^n}((1-\psi)f)\|_p\\ &\le cE_{2^n}(\psi f)+c(\psi,r)2^{-n(S-q)}\|f\|_p\\ &\le c(r,\psi,f)2^{-n\gamma}.\end{aligned}$$

This proves part (c). □

Let $\{\Psi_n : \mathbb{X}\times\mathbb{X}\to\mathbb{X}\}$ be a family of kernels (not necessarily symmetric). With a slight abuse of notation, we define when possible, for any measure ν with bounded total variation on $\mathbb{X}$,

$$\sigma(\nu,\Psi_n;f)(x)=\int_{\mathbb{X}} f(y)\Psi_n(x,y)d\nu(y),\quad x\in\mathbb{X},\ f\in L^1(\mathbb{X})+C_0(\mathbb{X}), \qquad (6.9)$$

and

$$\tau_j(\nu,\{\Psi_n\};f)=\begin{cases}\sigma(\nu,\Psi_1;f), & \text{if } j=0,\\ \sigma(\nu,\Psi_{2^j};f)-\sigma(\nu,\Psi_{2^{j-1}};f), & \text{if } j=1,2,\cdots.\end{cases} \qquad (6.10)$$

As usual, we will omit the mention of ν when $\nu=\mu^*$.

Corollary 6.1. *Let the assumptions of Theorem 6.1 hold, and* $\{\Psi_n : \mathbb{X}\times\mathbb{X}\to\mathbb{X}\}$ *be a sequence of kernels (not necessarily symmetric) with the property that both of the following functions of n are decreasing rapidly.*

$$\sup_{x\in\mathbb{X}}\int_{\mathbb{X}}|\Psi_n(x,y)-\Phi_n(x,y)|d\mu^*(y),\qquad \sup_{y\in\mathbb{X}}\int_{\mathbb{X}}|\Psi_n(x,y)-\Phi_n(x,y)|d\mu^*(x). \qquad (6.11)$$

(a) If $\mathbb{B}$ *is a ball centered at* x_0*, then*

$$\sup_{n\ge0}2^{n\gamma}\|f-\sigma(\Psi_{2^n};f)\|_{p,\mu^*,\mathbb{B}}\sim\sup_{j\ge0}2^{j\gamma}\|\tau_j(\{\Psi_n\};f)\|_{p,\mu^*,\mathbb{B}}. \qquad (6.12)$$

(b) If there exists a ball $\mathbb{B}$ *centered at* x_0 *such that*

$$\sup_{n\ge0}2^{n\gamma}\|f-\sigma(\Psi_{2^n};f)\|_{p,\mu^*,\mathbb{B}}\sim\sup_{j\ge0}2^{j\gamma}\|\tau_j(\{\Psi_n\};f)\|_{p,\mu^*,\mathbb{B}}<\infty, \qquad (6.13)$$

then $f\in W_{\gamma,p,\phi_0}(x_0)$.
(c) If $f\in W_{\gamma,p}(x_0)$*, then there exists a ball* $\mathbb{B}$ *centered at* x_0 *such that (6.13) holds.*

Proof. In view of Lemma 5.3, the assumption about the functions in (6.11) implies that $\|\sigma(\Psi_n;f)-\sigma_n(f)\|_p$ is decreasing rapidly. □

7. QUADRATURE FORMULA

The purpose of this section is to prove the existence of admissible quadrature measures in the general set-up as in this paper. The ideas are mostly developed already in our earlier works [17, 36, 43, 44, 50, 51] but always require an estimate on the gradient of diffusion polynomials. Here, we use the Bernstein-Lipschitz condition (Definition 3.4) instead.

If $\mathcal{C}\subset K\subset\mathbb{X}$, we denote

$$\delta(K,\mathcal{C})=\sup_{x\in K}\inf_{y\in\mathcal{C}}\rho(x,y),\qquad \eta(\mathcal{C})=\inf_{x,y\in\mathcal{C},x\ne y}\rho(x,y). \qquad (7.1)$$

If K is compact, $\epsilon>0$, a subset $\mathcal{C}\subset K$ is ϵ-distinguishable if $\rho(x,y)\ge\epsilon$ for every $x,y\in\mathcal{C}$, $x\ne y$. The cardinality the maximal ϵ-distinguishable subset of K will be denoted by $H_\epsilon(K)$.

Remark 7.1. If $\mathcal{C}_1\subset\mathcal{C}$ is a maximal $\delta(K,\mathcal{C})$-distinguishable subset of $\mathcal{C}$, $x\ne y$, then it is easy to deduce that

$$\delta(K,\mathcal{C})\le\eta(\mathcal{C}_1)\le2\delta(K,\mathcal{C}),\qquad \delta(K,\mathcal{C})\le\delta(K,\mathcal{C}_1)\le2\delta(K,\mathcal{C}).$$

In particular, by replacing $\mathcal{C}$ by $\mathcal{C}_1$, we can always assume that

$$(1/2)\delta(K,\mathcal{C})\le\eta(\mathcal{C})\le2\delta(K,\mathcal{C}). \qquad (7.2)$$

Theorem 7.1. *We assume the Bernstein-Lipschitz condition. Let* $n>0$, $\mathcal{C}_1=\{z_1,\cdots,z_M\}\subset\mathbb{K}_{2n}$ *be a finite subset,* $\epsilon>0$.
(a) There exists a constant $c(\epsilon)$ *with the following property: if* $\delta(\mathbb{K}_{2n},\mathcal{C}_1)\le c(\epsilon)\min(1/n,1/B_{2n})$*, then there exist non-negative numbers* W_k *satisfying*

$$0\le W_k\le c\delta(\mathbb{K}_{2n},\mathcal{C}_1)^q,\qquad \sum_{k=1}^M W_k\le c\mu^*(\mathbb{B}(\mathbb{K}_{2n},4\delta(\mathbb{K}_{2n},\mathcal{C}_1))), \qquad (7.3)$$

such that for every $P\in\Pi_n$,

$$\left|\sum_{k=1}^M W_k|P(z_k)|-\int_{\mathbb{X}}|P(x)|d\mu^*(x)\right|\le\epsilon\int_{\mathbb{X}}|P(x)|d\mu^*(x). \qquad (7.4)$$

(b) Let the assumptions of part (a) be satisfied with $\epsilon=1/2$*. There exist real numbers* $w_1,\cdots,w_M$ *such that* $|w_k|\le2W_k$, $k=1,\cdots,M$*, in particular,*

$$\sum_{k=1}^M|w_k|\le c\mu^*(\mathbb{B}(\mathbb{K}_{2n},4\delta(\mathbb{K}_{2n},\mathcal{C}_1))), \qquad (7.5)$$

and

$$\sum_{k=1}^M w_kP(z_k)=\int_{\mathbb{X}}P(x)d\mu^*(x),\qquad P\in\Pi_n. \qquad (7.6)$$

(c) Let $\delta>0$, $\mathcal{C}_1$ *be a random sample from the probability law* $\mu^*_{\mathbb{K}_{2n}}$ *given by*

$$\mu^*_{\mathbb{K}_{2n}}(B)=\frac{\mu^*(B\cap\mathbb{K}_{2n})}{\mu^*(\mathbb{K}_{2n})},$$

and $\epsilon_n = \min(1/n, 1/B_{2n})$*. If*

$$|\mathcal{C}_1| \ge c\epsilon_n^{-q}\mu^*(\mathbb{K}_{2n})\log\left(\frac{\mu^*(\mathbb{B}(\mathbb{K}_{2n},\epsilon_n))}{\delta\epsilon_n^q}\right),$$

then the statements (a) and (b) hold with $\mu^*_{\mathbb{K}_{2n}}$*-probability exceeding* $1-\delta$*.*

In order to prove Theorem 7.1, we first recall the following theorem [52, Theorem 5.1], applied to our context. The statement of Mhaskar [52, Theorem 5.1] seems to require that μ^* is a probability measure, but this fact is not required in the proof. It is required only that $\mu^*(\mathbb{B}(x,r)) \ge cr^q$ for $0 < r \le 1$.

Theorem 7.2. *Let* τ *be a positive measure supported on a compact subset of* $\mathbb{X}$*,* $\epsilon > 0$*,* $\mathcal{A}$ *be a maximal* ϵ*-distinguishable subset of* $\mathsf{supp}(\tau)$*, and* $K = \mathbb{B}(\mathcal{A}, 2\epsilon)$*. There then exists a subset* $\mathcal{C} \subseteq \mathcal{A} \subseteq \mathsf{supp}(\tau)$ *and a partition* $\{Y_y\}_{y\in\mathcal{C}}$ *of* K *with each of the following properties.*

1. *(**volume property**) For* $y \in \mathcal{C}$*,* $Y_y \subseteq \mathbb{B}(y, 18\epsilon)$*,* $(\kappa_1/\kappa_2)7^{-q}\epsilon^q \le \mu^*(Y_y) \le \kappa_2(18\epsilon)^q$*, and* $\tau(Y_y) \ge (\kappa_1/\kappa_2)19^{-q}\min_{y\in\mathcal{A}}\tau(\mathbb{B}(y,\epsilon)) > 0$*.*
2. *(**density property**)* $\eta(\mathcal{C}) \ge \epsilon$*,* $\delta(K,\mathcal{C}) \le 18\epsilon$*.*
3. *(**intersection property**) Let* $K_1 \subseteq K$ *be a compact subset. Then*

$$\left|\{y \in \mathcal{C}: Y_y \cap K_1 \ne \emptyset\}\right| \le (\kappa_2^2/\kappa_1)(133)^q H_\epsilon(K_1).$$

PROOF OF THEOREM 7.1 (a), (b).

We observe first that it is enough to prove this theorem for sufficiently large values of n. In view of Proposition 5.3, we may choose n large enough so that for any $P \in \Pi_n$,

$$\|P\|_{1,\mu^*,\mathbb{X}\setminus\mathbb{K}_{2n}} \le n^{-S}\|P\|_1 \le (\epsilon/3)\|P\|_1. \tag{7.7}$$

In this proof, we will write $\delta = \delta(\mathbb{K}_{2n}, \mathcal{C}_1)$ so that $\mathbb{K}_{2n} \subset \mathbb{B}(\mathcal{C}_1, \delta)$. We use Theorem 7.2 with τ to be the measure associating the mass 1 with each element of $\mathcal{C}_1$, and δ in place of ϵ. If $\mathcal{A}$ is a maximal δ-distinguished subset of $\mathcal{C}_1$, then we denote in this proof, $K = \mathbb{B}(\mathcal{A}, 2\delta)$ and observe that $\mathbb{K}_{2n} \subset \mathbb{B}(\mathcal{C}_1, \delta) \subset K \subset \mathbb{B}(\mathbb{K}_{2n}, 4\delta)$. We obtain a partition $\{Y_y\}$ of K as in Theorem 7.2. The volume property implies that each Y_y contains at least one element of $\mathcal{C}_1$. We construct a subset $\mathcal{C}$ of $\mathcal{C}_1$ by choosing exactly one element of $Y_y \cap \mathcal{C}_1$ for each y. We may then re-index $\mathcal{C}_1$ so that, without loss of generality, $\mathcal{C} = \{z_1, \cdots, z_N\}$ for some $N \le M$, and re-index $\{Y_y\}$ as $\{Y_k\}$, so that $z_k \in Y_k$, $k = 1, \cdots, N$. To summarize, we have a subset $\{z_1, \cdots, z_N\} \subseteq \mathcal{C}_1$, and a partition $\{Y_k\}_{k=1}^N$ of $K \supset \mathbb{K}_{2n}$ such that each $Y_k \subset \mathbb{B}(z_k, 36\delta)$ and $\mu^*(Y_k) \sim \delta^q$. In particular (cf. (7.7)), for any $P \in \Pi_n$,

$$\|P\|_1 - \|P\|_{1,\mu^*,K} \le (\epsilon/3)\|P\|_1. \tag{7.8}$$

We now let $W_k = \mu^*(Y_k)$, $k = 1, \cdots, N$, and $W_k = 0$, $k = N+1, \cdots, M$.

The next step is to prove that if $\delta \le c(\epsilon)\min(1/n, 1/B_{2n})$, then

$$\sup_{y\in\mathbb{X}}\sum_{k=1}^N\int_{Y_k}|\Phi_{2n}(z_k,y) - \Phi_{2n}(x,y)|d\mu^*(x) \le 2\epsilon/3. \tag{7.9}$$

In this part of the proof, the constants denoted by $c_1, c_2, \cdots$ will retain their value until (7.9) is proved. Let $y \in \mathbb{X}$. We let $r \ge \delta$ to be chosen later, and write in this proof, $\mathcal{N} = \{k: \mathsf{dist}(y, Y_k) < r\}$, $\mathcal{L} = \{k: \mathsf{dist}(y, Y_k) \ge r\}$ and for $j = 0, 1, \cdots$, $\mathcal{L}_j = \{k: 2^j r \le \mathsf{dist}(y, Y_k) < 2^{j+1}r\}$. Since $r \ge \delta$, and each $Y_k \subset \mathbb{B}(z_k, 36\delta)$, there are at most $c_1(r/\delta)^q$ elements in $\mathcal{N}$. Using the Bernstein-Lipschitz condition and the fact that $\|\Phi_{2n}(\circ, y)\|_\infty \le c_2 n^q$, we deduce that

$$\begin{aligned}\sum_{k\in\mathcal{N}}\int_{Y_k}|\Phi_{2n}(z_k,y) - \Phi_{2n}(x,y)|d\mu^*(x) &\le c_3\mu^*(Y_k)n^q B_{2n}\delta(r/\delta)^q\\ &\le c_3\mu^*(\mathcal{B}(z_k,36\delta))n^q B_{2n}\delta(r/\delta)^q \le c_4(nr)^q B_{2n}\delta.\end{aligned} \tag{7.10}$$

Next, since $\mu^*(Y_k) \sim \delta^q$, we see that the number of elements in each $\mathcal{L}_j$ is $\sim (2^j r/\delta)^q$. Using Proposition 3.2 and the fact that $S > q$, we deduce that if $r \ge 1/n$, then

$$\begin{aligned}&\sum_{k\in\mathcal{L}}\int_{Y_k}|\Phi_{2n}(z_k,y) - \Phi_{2n}(x,y)|d\mu^*(x)\\ &= \sum_{j=0}^\infty\sum_{k\in\mathcal{L}_j}\int_{Y_k}|\Phi_{2n}(z_k,y) - \Phi_{2n}(x,y)|d\mu^*(x)\\ &\le c_5 n^q(nr)^{-S}\sum_{j=0}^\infty 2^{-jS}\left\{\sum_{k\in\mathcal{L}_j}\mu^*(Y_k)\right\}\\ &\le c_6(nr)^{q-S}.\end{aligned} \tag{7.11}$$

Since $S > q$, we may choose $r \sim_\epsilon n$ such that $c_6(nr)^{q-S} \le \epsilon/3$, and we then require $\delta \le \min(r, c_7(\epsilon)/B_{2n})$ so that, in (7.10), $c_4(nr)^q B_{2n}\delta \le \epsilon/3$. Then (7.10) and (7.11) lead to (7.9). The proof of (7.9) being completed, we resume the constant convention as usual.

Next, we observe that for any $P \in \Pi_n$,

$$P(x) = \int_{\mathbb{X}} P(y)\Phi_{2n}(x,y)d\mu^*(y), \qquad x \in \mathbb{X}.$$

We therefore conclude, using (7.9), that

$$\begin{aligned}&\left|\sum_{k=1}^N\mu^*(Y_k)|P(z_k)| - \int_K|P(x)|d\mu^*(x)\right|\\ &= \left|\sum_{k=1}^N\int_{Y_k}\big(|P(z_k)| - |P(x)|\big)\,d\mu^*(x)\right| \le \sum_{k=1}^N\int_{Y_k}|P(z_k)\\ &-P(x)|d\mu^*(x) \le \sum_{k=1}^N\int_{Y_k}\left|\int_{\mathbb{X}}P(y)\left\{\Phi_{2n}(z_k,y)\right.\right.\\ &\left.\left.-\Phi_{2n}(x,y)\right\}d\mu^*(y)\right|d\mu^*(x)\\ &\le \int_{\mathbb{X}}|P(y)|\left\{\sum_{k=1}^N\int_{Y_k}|\Phi_{2n}(z_k,y) - \Phi_{2n}(x,y)|d\mu^*(x)\right\}d\mu^*(y)\\ &\le (2\epsilon/3)\int_{\mathbb{X}}|P(y)|d\mu^*(y).\end{aligned}$$

Together with (7.8), this leads to (7.4). From the definition of $W_k = \mu^*(Y_k)$, $k = 1, \cdots, N$, $W_k \le c\delta^q$, and $\sum_{k=1}^N W_k =$

$\mu^*(K) = \mu^*(\mathbb{B}(\mathbb{K}_{2n}, 4\delta))$. Since $W_k = 0$ if $k \ge N+1$, we have now proven (7.3), and we have thus completed the proof of part (a).

Having proved part (a), the proof of part (b) is by now a routine application of the Hahn-Banach theorem [cf. [17, 44, 50, 51]]. We apply part (a) with $\epsilon = 1/2$. Continuing the notation in the proof of part (a), we then have

$$(1/2)\|P\|_1 \le \sum_{k=1}^{N} W_k|P(z_k)| \le (3/2)\|P\|_1, \qquad P \in \Pi_n. \quad (7.12)$$

We now equip $\mathbb{R}^N$ with the norm $|||(a_1, \cdots, a_N)||| = \sum_{k=1}^N W_k|a_k|$ and consider the sampling operator $\mathcal{S} : \Pi_n \to \mathbb{R}^N$ given by $\mathcal{S}(P) = (P(z_1), \cdots, P(z_N))$, let V be the range of this operator, and define a linear functional x^* on V by $x^*(\mathcal{S}(P)) = \int_{\mathbb{X}} P d\mu^*$. The estimate (7.12) shows that the norm of this functional is ≤ 2. The Hahn-Banach theorem yields a norm-preserving extension X^* of x^* to $\mathbb{R}^N$, which, in turn, can be identified with a vector $(w_1, \cdots, w_N) \in \mathbb{R}^N$. We set $w_k = 0$ if $k \ge N+1$. Formula (7.6) then expresses the fact that X^* is an extension of x^*. The preservation of norms shows that $|w_k| \le 2W_k$ if $k = 1, \cdots, N$, and it is clear that for $k = N+1, \cdots, M$, $|w_k| = 0 = W_k$. This completes the proof of part (b). □

Part (c) of Theorem 7.1 follows immediately from the first two parts and the following lemma.

Lemma 7.1. *Let ν^* be a probability measure on $\mathbb{X}$, $K \subset \mathsf{supp}(\nu^*)$ be a compact set. Let $\epsilon, \delta \in (0, 1]$, $\mathcal{C}$ be a maximal $\epsilon/2$-distinguished subset of K, and $\nu_\epsilon = \min_{x\in\mathcal{C}} \nu^*(\mathbb{B}(x, \epsilon/2))$. If*

$$M \ge c\nu_\epsilon^{-1} \log\left(c_1\mu^*(\mathbb{B}(K,\epsilon))/(\delta\epsilon^q)\right),$$

and $\{z_1, \cdots, z_M\}$ be random samples from the probability law ν^* then

$$\mathsf{Prob}_{\nu^*}\left(\{\delta(K, \{z_1, \cdots, z_M\}) > \epsilon\}\right) \le \delta. \quad (7.13)$$

PROOF. If $\delta(K, \{z_1, \cdots, z_M\}) > \epsilon$, then there exists at least one $x \in \mathcal{C}$ such that $\mathbb{B}(x, \epsilon/2) \cap \{z_1, \cdots, z_M\} = \emptyset$. For every $x \in \mathcal{C}$, $p_x = \nu^*(\mathbb{B}(x, \epsilon/2)) \ge \nu_\epsilon$. We consider the random variable z_j to be equal to 1 if $z_j \in \mathbb{B}(x, \epsilon/2)$ and 0 otherwise. Using (B.2) with $t = 1$, we see that

$$\begin{aligned}&\mathsf{Prob}\left(\mathbb{B}(x,\epsilon/2) \cap \{z_1, \cdots, z_M\}\right.\\ &\left.= \emptyset\right) \le \exp(-Mp_x/2) \le \exp(-cM\nu_\epsilon).\end{aligned}$$

Since $|\mathcal{C}| \le c_1\mu^*(\mathbb{B}(K,\epsilon))/\epsilon^q$,

$$\mathsf{Prob}\left(\{\delta(K, \{z_1, \cdots, z_M\}) > \epsilon\}\right) \le c_1\frac{\mu^*(\mathbb{B}(K,\epsilon))}{\epsilon^q}\exp(-cM\nu_\epsilon).$$

We set the right-hand side above to δ and solve for M to prove the lemma. □

8. PROOFS OF THE RESULTS IN SECTION 4

We assume the set-up as in section 4. Our first goal is to prove the following theorem.

Theorem 8.1. *Let τ, ν^*, $\mathcal{F}$, f be as described section 4. We assume the Bernstein-Lipschitz condition. Let $0 < \delta < 1$. We assume further that $|\mathcal{F}(y, \epsilon)| \le 1$ for all $y \in \mathbb{X}$, $\epsilon \in \Omega$. There exist constants c_1, c_2, such that if $M \ge c_1 n^q |||\nu^*|||_{R,0} \log(cnB_n/\delta)$, and $\{(y_1, \epsilon_1), \cdots, (y_M, \epsilon_M)\}$ is a random sample from τ, then*

$$\begin{aligned}&\mathsf{Prob}_{\nu^*}\left(\left\{\left\|\frac{1}{M}\sum_{j=1}^M \mathcal{F}(y_j,\epsilon_j)\Phi_n(\circ, y_j) - \sigma_n(\nu^*; f)\right\|_\infty\right.\right.\\ &\left.\left.\ge c_3\sqrt{\frac{n^q|||\nu^*|||_{R,0}\log(cnB_n|||\nu^*|||_{R,0}/\delta)}{M}}\right\}\right) \le \frac{\delta}{|||\nu^*|||_{R,0}}.\end{aligned} \quad (8.1)$$

In order to prove this theorem, we record an observation. The following lemma is an immediate corollary of the Bernstein-Lipschitz condition and Proposition 5.3.

Lemma 8.1. *Let the Bernstein-Lipschitz condition be satisfied. Then for every $n > 0$ and $\epsilon > 0$, there exists a finite set $\mathcal{C}_{n,\epsilon} \subset \mathbb{K}_{2n}$ such that $|\mathcal{C}_{n,\epsilon}| \le cB_n^q\epsilon^{-q}\mu^*(\mathbb{B}(\mathbb{K}_{2n}, \epsilon))$ and for any $P \in \Pi_n$,*

$$\left|\max_{x\in\mathcal{C}_{n,\epsilon}} |P(x)| - \|P\|_\infty\right| \le \epsilon\|P\|_\infty. \quad (8.2)$$

PROOF OF THEOREM 8.1.

Let $x \in \mathbb{X}$. We consider the random variables

$$Z_j = \mathcal{F}(y_j, \epsilon_j)\Phi_n(x, y_j), \qquad j = 1, \cdots, M.$$

Then in view of (4.2), $\mathbb{E}_\tau(Z_j) = \sigma_n(\nu^*; f)(x)$ for every j. Further, Proposition 3.2 shows that for each j, $|Z_j| \le cn^q$. Using (5.10) with ν^* in place of ν, $N = n$, $d = 0$, we see that for each j,

$$\int_{\mathbb{X}\times\Omega} |Z_j|^2 d\tau \le \int_{\mathbb{X}} |\Phi_n(x, y)|^2 d\nu^*(y) \le cn^q|||\nu^*|||_{R,0}.$$

Therefore, Bernstein concentration inequality (B.1) implies that for any $t \in (0, 1)$,

$$\begin{aligned}&\mathsf{Prob}\left(\left\{\left|\frac{1}{M}\sum_{j=1}^M \mathcal{F}(y_j,\epsilon_j)\Phi_n(x,y_j) - \sigma_n(\nu^*; f)(x)\right| \ge t/2\right\}\right)\\ &\le 2\exp\left(-c\frac{t^2M}{n^q|||\nu^*|||_{R,0}}\right);\end{aligned} \quad (8.3)$$

We now note that Z_j, $\sigma_n(\nu^*; f)$ are all in Π_n. Taking a finite set $\mathcal{C}_{n,1/2}$ as in Lemma 8.1, so that $|\mathcal{C}_{n,1/2}| \le cB_n^q\mu^*(\mathbb{B}(\mathbb{K}_{2n}, 1/2)) \le c_1n^cB_n^q$, we deduce that

$$\begin{aligned}&\max_{x\in\mathcal{C}_{n,1/2}}\left|\frac{1}{M}\sum_{j=1}^M \mathcal{F}(y_j,\epsilon_j)\Phi_n(x,y_j) - \sigma_n(\nu^*; f)(x)\right|\\ &\ge (1/2)\left\|\frac{1}{M}\sum_{j=1}^M \mathcal{F}(y_j,\epsilon_j)\Phi_n(\circ,y_j) - \sigma_n(\nu^*; f)\right\|_\infty.\end{aligned}$$

Then (8.3) leads to

$$\mathrm{Prob}\left(\left\{\left\|\frac{1}{M}\sum_{j=1}^{M}\mathcal{F}(y_j,\epsilon_j)\Phi_n(x,y_j)-\sigma_n(\nu^*;f)(x)\right\|_\infty \ge t\right\}\right)$$
$$\le c_1 B_n^q n^c \exp\left(-c_2\frac{t^2M}{n^q|||\nu^*|||_{R,0}}\right). \tag{8.4}$$

We set the right-hand side above equal to $\delta/|||\nu^*|||_{R,0}$ and solve for t to obtain (8.1) (with different values of c, c_1, c_2). □

Before starting to prove results regarding eignets, we first record the continuity and smoothness of a "smooth kernel" G as defined in Definition 3.10.

Proposition 8.1. *If G is a smooth kernel, then $(x,y)\mapsto W(y)G(x,y)$ is in $C_0(\mathbb{X}\times\mathbb{X})\cap L^1(\mu^*\times\mu^*;\mathbb{X}\times\mathbb{X})$. Further, for any p, $1\le p\le\infty$, and $\Lambda\ge 1$,*

$$\sup_{x\in\mathbb{X}}\left\|W(\circ)G(x,\circ)-\sum_{k:\lambda_k<\Lambda}b(\lambda_k)\phi_k(x)\phi_k(\circ)\right\|_p \le c_1\Lambda^c b(\Lambda). \tag{8.5}$$

In particular, for every $x,y\in\mathbb{X}$, $W(\circ)G(x,\circ)$ and $W(y)G(\circ,y)$ are in C^∞.

Proof. Let b be the smooth mask corresponding to G. For any $S\ge 1$, $b(n)\le cn^{-S}b(n/B^*)\le cn^{-S}b(0)$. Thus, b itself is decreasing rapidly. Next, let $r>0$. Then remembering that $B^*\ge 1$ and b is non-increasing, we obtain that for $S>0$, $b(B^*\Lambda u)\le c(\Lambda u)^{-S-r-1}b(\Lambda u)$, and

$$\int_\Lambda^\infty t^r b(t)dt=(B^*\Lambda)^{r+1}\int_{1/B^*}^\infty u^r b(B^*\Lambda u)du$$
$$\le c\Lambda^{-S}\int_{1/B^*}^\infty u^{-S-1}b(\Lambda u)du$$
$$\le c\Lambda^{-S}\int_1^\infty u^{-S-1}b(\Lambda u)du\le c\Lambda^{-S}b(\Lambda). \tag{8.6}$$

In this proof, let $s(t)=\sum_{k:\lambda_k<t}\phi_k(x)^2$, so that $s(t)\le ct^q$, $t\ge 1$. If $\Lambda\ge 1$, then, integrating by parts, we deduce (remembering that b is non-increasing) that for any $x\in\mathbb{X}$,

$$\sum_{k:\lambda_k\ge\Lambda}b(\lambda_k)\phi_k(x)^2$$
$$=\int_\Lambda^\infty b(t)ds(t)=-b(\Lambda)s(\Lambda)-\int_\Lambda^\infty s(t)db(t)$$
$$\le c_1\left\{\Lambda^q b(\Lambda)-\int_\Lambda^\infty t^q db(t)\right\}\le c_2\left\{\Lambda^q b(\Lambda)\right.$$
$$\left.+\int_\Lambda^\infty t^{q-1}b(t)dt\right\}\le c_3\Lambda^q b(\Lambda). \tag{8.7}$$

Using Schwarz inequality, we conclude that

$$\sup_{x,y\in\mathbb{X}}\sum_{k:\lambda_k\ge\Lambda}b(\lambda_k)|\phi_k(x)\phi_k(y)|\le c_3\Lambda^q b(\Lambda). \tag{8.8}$$

In particular, since b is fast decreasing, $W(\circ)G(x,\circ)\in C_0(\mathbb{X})$ (and in fact, $W(y)G(x,y)\in C_0(\mathbb{X}\times\mathbb{X})$) and (8.5) holds with $p=\infty$. Next, for any $j\ge 0$, essential compactness implies that

$$\int_{\mathbb{X}\setminus\mathbb{K}_{2^{j+1}\Lambda}}\left(\sum_{k:\lambda_k\in[2^j\Lambda,2^{j+1}\Lambda)}b(\lambda_k)\phi_k(y)^2\right)^{1/2}$$
$$d\mu^*(y)\le c\Lambda^{-S-q}b(2^j\Lambda)^{1/2}.$$

So, there exists $r\ge q$ such that

$$\int_{\mathbb{X}}\left(\sum_{k:\lambda_k\in[2^j\Lambda,2^{j+1}\Lambda)}b(\lambda_k)\phi_k(y)^2\right)^{1/2}d\mu^*(y)$$
$$\le\int_{\mathbb{K}_{2^{j+1}\Lambda}}\left(\sum_{k:\lambda_k\in[2^j\Lambda,2^{j+1}\Lambda)}b(\lambda_k)\phi_k(y)^2\right)^{1/2}d\mu^*(y)$$
$$+c\Lambda^{-S-q}b(2^j\Lambda)^{1/2}$$
$$\le c\left((2^j\Lambda)^q b(2^j\Lambda)\right)^{1/2}\mu^*(\mathbb{K}_{2^{j+1}\Lambda})\le c\left((2^j\Lambda)^r b(2^j\Lambda)\right)^{1/2}.$$

Hence, for any $x\in\mathbb{X}$,

$$\int_{\mathbb{X}}\sum_{k:\lambda_k\ge\Lambda}b(\lambda_k)|\phi_k(x)\phi_k(y)|d\mu^*(y)$$
$$=\sum_{j=0}^\infty\int_{\mathbb{X}}\sum_{k:\lambda_k\in[2^j\Lambda,2^{j+1}\Lambda)}b(\lambda_k)|\phi_k(x)\phi_k(y)|d\mu^*(y)$$
$$\le\sum_{j=0}^\infty\left\{\sum_{k:\lambda_k\in[2^j\Lambda,2^{j+1}\Lambda)}b(\lambda_k)\phi_k(x)^2\right\}^{1/2}$$
$$\int_{\mathbb{X}}\left(\sum_{k:\lambda_k\in[2^j\Lambda,2^{j+1}\Lambda)}b(\lambda_k)\phi_k(y)^2\right)^{1/2}d\mu^*(y) \tag{8.9}$$
$$\le c\sum_{j=0}^\infty(2^j\Lambda)^r b(2^j\Lambda)\le c\sum_{j=0}^\infty\int_{2^{j-1}\Lambda}^{2^j\Lambda}t^{r-1}b(t)dt$$
$$=c\int_{\Lambda/2}^\infty t^{r-1}b(t)dt\le c\Lambda^{-S}b(\Lambda).$$

This shows that

$$\sup_{x\in\mathbb{X}}\left\|\sum_{k:\lambda_k\ge\Lambda}b(\lambda_k)|\phi_k(x)\phi_k(\circ)|\right\|_1\le c\Lambda^{-S}b(\Lambda). \tag{8.10}$$

In view of the convexity inequality,

$$\|f\|_p\le\|f\|_\infty^{1-1/p}\|f\|_1^{1/p},\qquad 1<p<\infty,$$

(8.8) and (8.10) lead to

$$\sup_{x\in\mathbb{X}}\left\|\sum_{k:\lambda_k\ge\Lambda}b(\lambda_k)|\phi_k(x)\phi_k(\circ)|\right\|_p\le c_1\Lambda^c b(\Lambda),\qquad 1\le p\le\infty.$$

In turn, this implies that $WG(x,\circ) \in L^p$ for all $x \in \mathbb{X}$, and (8.5) holds. □

A fundamental fact that relates the kernels Φ_n and the pre-fabricated eignets $\mathbb{G}_n$'s is the following theorem.

Theorem 8.2. *Let G be a smooth kernel and $\{\nu_n\}$ be an admissible product quadrature measure sequence. Then, for $1 \le p \le \infty$,*

$$\left\{\sup_{x\in\mathbb{X}} \|\mathbb{G}_n(\nu_{B^*n}; x, \circ) - \Phi_n(x,\circ)\|_p\right\}$$

is fast decreasing. In particular, for every $S > 0$

$$|\mathbb{G}_n(\nu_{B^*n}; x, y)| \le c(S)\left\{\frac{n^q}{\max(1, (N\rho(x,y))^S)} + n^{-2S}\right\}. \qquad (8.11)$$

PROOF. Let $x \in \mathbb{X}$. In this proof, we define $P_n = P_{n,x}$ by $P_n(z) = \sum_{k:\lambda_k<B^*n} b(\lambda_k)\phi_k(x)\phi_k(z)$, $z \in \mathbb{X}$, and note that $P_n \in \Pi_{B^*n}$. In view of Proposition 8.1, the expansion in (3.18) converges in $C_0(\mathbb{X}\times\mathbb{X})\cap L^1(\mu^*\times\mu^*; \mathbb{X}\times\mathbb{X})$, so that term-by-term integration can be made to deduce that for $y \in \mathbb{X}$,

$$\begin{aligned}\int_{\mathbb{X}} G(x,z)W(z)\mathcal{D}_{G,n}(z,y)d\mu^*(z) &= \int_{\mathbb{X}} P_n(z)\mathcal{D}_{G,n}(z,y)d\mu^*(z)\\ &\quad + \sum_{k:\lambda_k\ge B^*n} b(\lambda_k)\phi_k(x)\int_{\mathbb{X}}\phi_k(z)\mathcal{D}_{G,n}(z,y)d\mu^*(z).\end{aligned}$$

By definition, $\mathcal{D}_{G,n}(\circ,y) \in \Pi_n^q$, and, hence, each of the summands in the last expression above is equal to 0. Therefore, recalling that $h(\lambda_k/n) = 0$ if $\lambda_k > n$, we obtain

$$\begin{aligned}\int_{\mathbb{X}} G(x,z)W(z)\mathcal{D}_{G,n}(z,y)d\mu^*(z) &= \int_{\mathbb{X}} P_n(z)\mathcal{D}_{G,n}(z,y)d\mu^*(z)\\ &= \sum_{k:\lambda_k<B^*n} b(\lambda_k)\phi_k(x)\int_{\mathbb{X}}\phi_k(z)\mathcal{D}_{G,n}(z,y)d\mu^*(z)\\ &= \sum_{k:\lambda_k<B^*n} b(\lambda_k)\phi_k(x)h(\lambda_k/n)b(\lambda_k)^{-1}\phi_k(y)\\ &= \sum_k h(\lambda_k/n)\phi_k(x)\phi_k(y)\\ &= \Phi_n(x,y). \qquad (8.12)\end{aligned}$$

Since $\mathcal{D}_{G,n}(z,\circ) \in \Pi_n \subset \Pi_{B^*n}$, and ν_{B^*n} is an admissible product quadrature measure of order B^*n, this implies that

$$\Phi_n(x,y) = \int_{\mathbb{X}} P_n(z)\mathcal{D}_{G,n}(z,y)d\nu_{B^*n}(z), \qquad y \in \mathbb{X}. \qquad (8.13)$$

Therefore, for $y \in \mathbb{X}$,

$$\begin{aligned}&\mathbb{G}_n(\nu_{B^*n}; x, y) - \Phi_n(x,y)\\ &= \int_{\mathbb{X}}\left\{W(z)G(x,z) - P_n(z)\right\}\mathcal{D}_{G,n}(z,y)d\nu_{B^*n}(z).\end{aligned}$$

Using Proposition 8.1 (used with $\Lambda = B^*n$) and the fact that $\{|\nu_{B^*n}|(\mathbb{X})\}$ has polynomial growth, we deduce that

$$\begin{aligned}\|\mathbb{G}_n(\nu_{B^*n}; x,\circ) - \Phi_n(x,\circ)\|_p &\le |\nu_{B^*n}|(\mathbb{X})\\ &\quad\times\left\|W(\circ)G(x,\circ) - P_n\right\|_\infty \sup_{z\in\mathbb{X}}\|\mathcal{D}_{G,n}(z,\circ)\|_p\\ &\le c_1 n^c b(B^*n)\sup_{z\in\mathbb{X}}\|\mathcal{D}_{G,n}(z,\circ)\|_p. \qquad (8.14)\end{aligned}$$

In view of Proposition 5.4 and Proposition 5.2, we see that for any $z \in \mathbb{X}$,

$$\begin{aligned}\|\mathcal{D}_{G,n}(z,\circ)\|_p^2 &\le c_1 n^{2c}\|\mathcal{D}_{G,n}(z,\circ)\|_2^2\\ &= c_1 n^{2c}\sum_{k:\lambda_k<n}\left(h(\lambda_k/n)\, b(\lambda_k)^{-1}\phi_k(z)\right)^2\\ &\le c_1 n^{2c} b(n)^{-2}\|\Phi_n(z,\circ)\|_2^2 \le c_1 n^c b(n)^{-2}\|\Phi_n(z,\circ)\|_1^2\\ &\le c_1 n^c b(n)^{-2}.\end{aligned}$$

We now conclude from (8.14) that

$$\|\mathbb{G}_n(\nu_{B^*n}; x,\circ) - \Phi_n(x,\circ)\|_p \le c_1 n^c\frac{b(B^*n)}{b(n)}.$$

Since $\{b(B^*n)/b(n)\}$ is fast decreasing, this completes the proof. □

The theorems in section 4 all follow from the following basic theorem.

Theorem 8.3. *We assume the strong product assumption and the Bernstein-Lipschitz condition. With the set-up just described, we have*

$$\text{Prob}_{\nu^*}\left(\left\{\left\|\mathcal{G}_n(Y;\mathcal{F}) - \sigma_n(f_0f)\right\|_\infty \ge c_3\sqrt{\frac{n^q|||\nu^*|||_{R,0}\log(cnB_n|||\nu^*|||_{R,0}/\delta)}{|Y|}}\right\}\right) \le \frac{\delta}{|||\nu^*|||_{R,0}}. \qquad (8.15)$$

In particular, for $f \in X^\infty(\mathbb{X})$, Then

$$\begin{aligned}\text{Prob}_{\nu^*}&\left(\left\{\left\|\mathcal{G}_n(Y;\mathcal{F}) - f_0f\right\|_\infty\right.\right.\\ &\left.\left.\ge c_3\left(\sqrt{\frac{n^q|||\nu^*|||_{R,0}\log(cnB_n|||\nu^*|||_{R,0}/\delta)}{|Y|}} + E_{n/2}(\infty, f_0f)\right)\right\}\right)\\ &\le \frac{\delta}{|||\nu^*|||_{R,0}}. \qquad (8.16)\end{aligned}$$

PROOF. Theorems 8.1 and Theorem 8.2 together lead to (8.15). Since $\sigma_n(\nu^*; f) = \sigma_n(f_0f)$, the estimate 8.16 follows from Theorem 5.1 used with $p = \infty$. □

PROOF OF THEOREM 4.1.

We observe that with the choice of f_0 as in this theorem, $|||\nu^*|||_{R,0} \le \|f_0\|_\infty \le 1/\mathfrak{m}$. Using $\mathfrak{m}\delta$ in place of δ, we obtain Theorem 4.1 directly from Theorem 8.3 by some simple calculations. □

PROOF OF THEOREM 4.2.

This follows directly from Theorem 8.3 by choosing $\mathcal{F} \equiv 1$. □

PROOF OF THEOREM 4.3.

In view of Theorem 8.3, our assumptions imply that for each $j \geq 0$,

$$\mathsf{Prob}_{\nu^*}\left(\left\{\left\|\mathcal{G}_{2^j}(Y;\mathcal{F}) - \sigma_{2^j}(f_0 f)\right\|_\infty \leq c2^{-jS}\right\}\right) \leq \delta/2^{j+1}.$$

Consequently, with probability $\geq 1-\delta$, we have for each $j \geq 1$,

$$\left\|\mathcal{G}_{2^j}(Y;\mathcal{F}) - \mathcal{G}_{2^{j-1}}(Y_j;\mathcal{F}) - \tau_j(f_0 f)\right\|_\infty \leq c2^{-jS}.$$

Hence, the theorem follows from Theorem 6.1. □

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

1. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: opportunities and challenges. *Neurocomputing.* (2017) **237**:350–61. doi: 10.1016/j.neucom.2017.01.026
2. Cucker F, Smale S. On the mathematical foundations of learning. *Bull Am Math Soc.* (2002) **39**:1–49. doi: 10.1090/S0273-0979-01-00923-5
3. Cucker F, Zhou DX. *Learning Theory: An Approximation Theory Viewpoint*, Vol. **24**. Cambridge: Cambridge University Press (2007).
4. Girosi F, Poggio T. Networks and the best approximation property. *Biol Cybernet.* (1990) **63**:169–76. doi: 10.1007/BF00195855
5. Chui CK, Donoho DL. Special issue: diffusion maps and wavelets. *Appl Comput Harm Anal.* (2006) **21**:1–2. doi: 10.1016/j.acha.2006.05.005
6. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* (2003) **15**:1373–96. doi: 10.1162/089976603321780317
7. Belkin M, Niyogi P. Towards a theoretical foundation for Laplacian-based manifold methods. *J Comput Syst Sci.* (2008) **74**:1289–308. doi: 10.1016/j.jcss.2007.08.006
8. Belkin M, Niyogi P. Semi-supervised learning on Riemannian manifolds. *Mach Learn.* (2004) **56**:209–39. doi: 10.1023/B:MACH.0000033120.25363.1e
9. Lafon SS. *Diffusion maps and geometric harmonics* (Ph.D. thesis), Yale University, New Haven, CT, United States (2004).
10. Singer A. From graph to manifold Laplacian: the convergence rate. *Appl Comput Harm Anal.* (2006) **21**:128–34. doi: 10.1016/j.acha.2006.03.004
11. Jones PW, Maggioni M, Schul R. Universal local parametrizations via heat kernels and eigenfunctions of the Laplacian. *Ann Acad Sci Fenn Math.* (2010) **35**:131–74. doi: 10.5186/aasfm.2010.3508
12. Liao W, Maggioni M. Adaptive geometric multiscale approximations for intrinsically low-dimensional data. *arXiv.* (2016) 1611.01179.
13. Maggioni M, Mhaskar HN. Diffusion polynomial frames on metric measure spaces. *Appl Comput Harm Anal.* (2008) **24**:329–53. doi: 10.1016/j.acha.2007.07.001
14. Mhaskar HN. Eignets for function approximation on manifolds. *Appl Comput Harm Anal.* (2010) **29**:63–87. doi: 10.1016/j.acha.2009.08.006
15. Mhaskar HN. A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. *Neural Netw.* (2011) **24**:345–59. doi: 10.1016/j.neunet.2010.12.007
16. Ehler M, Filbir F, Mhaskar HN. Locally learning biomedical data using diffusion frames. *J Comput Biol.* (2012) **19**:1251–64. doi: 10.1089/cmb.2012.0187
17. Filbir F, Mhaskar HN. Marcinkiewicz-Zygmund measures on manifolds. *J Complexity.* (2011) **27**:568–96. doi: 10.1016/j.jco.2011.03.002
18. Rosasco L, Belkin M, Vito ED. On learning with integral operators. *J Mach Learn Res.* (2010) **11**:905–34.
19. Rudi A, Carratino L, Rosasco L. Falkon: an optimal large scale kernel method. *arXiv.* (2017) 1705.10958. Available online at: http://jmlr.org/papers/v11/rosasco10a.html.
20. Lu S, Pereverzev SV. *Regularization Theory for Ill-Posed Problems*. Berlin: de Gruyter (2013).
21. Mhaskar H, Pereverzyev SV, Semenov VY, Semenova EV. Data based construction of kernels for semi-supervised learning with less labels. *Front Appl Math Stat.* (2019) **5**:21. doi: 10.3389/fams.2019.00021
22. Pereverzyev SV, Tkachenko P. Regularization by the linear functional strategy with multiple kernels. *Front Appl Math Stat.* (2017) **3**:1. doi: 10.3389/fams.2017.00001
23. Fefferman C, Mitter S, Narayanan H. Testing the manifold hypothesis. *J Am Math Soc.* (2016) **29**:983–1049. doi: 10.1090/jams/852
24. Chui CK, Lin S-B, Zhang B, Zhou DX. Realization of spatial sparseness by deep relu nets with massive data. *arXiv.* (2019) 1912.07464.
25. Guo ZC, Lin SB, Zhou DX. Learning theory of distributed spectral algorithms. *Inverse Probl.* (2017) **33**:074009. doi: 10.1088/1361-6420/aa72b2
26. Lin SB, Wang YG, Zhou DX. Distributed filtered hyperinterpolation for noisy data on the sphere. *arXiv.* (2019) 1910.02434.
27. Mhaskar HN, Poggio T. Deep vs. shallow networks: an approximation theory perspective. *Anal Appl.* (2016) **14**:829–48. doi: 10.1142/S0219530516400042
28. Mhaskar H, Poggio T. Function approximation by deep networks. *arXiv.* (2019) 1905.12882.
29. Mhaskar HN. On the representation of smooth functions on the sphere using finitely many bits. *Appl Comput Harm Anal.* (2005) **18**:215–33. doi: 10.1016/j.acha.2004.11.004
30. Smale S, Rosasco L, Bouvrie J, Caponnetto A, Poggio T. Mathematics of the neural response. *Foundat Comput Math.* (2010) **10**:67–91. doi: 10.1007/s10208-009-9049-1
31. Mhaskar HN. On the representation of band limited functions using finitely many bits. *J Complexity.* (2002) **18**:449–78. doi: 10.1006/jcom.2001.0637
32. Hardy RL. Theory and applications of the multiquadric-biharmonic method 20 years of discovery 1968–1988. *Comput Math Appl.* (1990) **19**:163–208. doi: 10.1016/0898-1221(90)90272-L
33. Müller A. *Spherical Harmonics*, Vol. **17**. Berlin: Springer (2006).
34. Mhaskar HN, Narcowich FJ, Ward JD. Approximation properties of zonal function networks using scattered data on the sphere. *Adv Comput Math.* (1999) **11**:121–37. doi: 10.1023/A:1018967708053
35. Timan AF. *Theory of Approximation of Functions of a Real Variable: International Series of Monographs on Pure and Applied Mathematics*, Vol. **34**. New York, NY: Dover Publications (2014).
36. Chui CK, Mhaskar HN. A unified method for super-resolution recovery and real exponential-sum separation. *Appl Comput Harmon Anal.* (2019) **46**:431–51. doi: 10.1016/j.acha.2017.12.007
37. Chui CK, Mhaskar HN. A Fourier-invariant method for locating point-masses and computing their attributes. *Appl Comput Harmon Anal.* (2018) **45**:436–52. doi: 10.1016/j.acha.2017.08.010
38. Mhaskar HN. *Introduction to the Theory of Weighted Polynomial Approximation*, Vol. **56**. Singapore: World Scientific Singapore (1996).
39. Steinerberger S. On the spectral resolution of products of laplacian eigenfunctions. *arXiv.* (2017) 1711.09826.
40. Lu J, Sogge CD, Steinerberger S. Approximating pointwise products of laplacian eigenfunctions. *J Funct Anal.* (2019) **277**:3271–82. doi: 10.1016/j.jfa.2019.05.025
41. Lu J, Steinerberger S. On pointwise products of elliptic eigenfunctions. *arXiv.* (2018) 1810.01024.

42. Geller D, Pesenson IZ. Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds. *J Geometr Anal.* (2011) **21**:334–71. doi: 10.1007/s12220-010-9150-3
43. Mhaskar HN. Local approximation using Hermite functions. In: N. K. Govil, R. Mohapatra, M. A. Qazi, G. Schmeisser eds. *Progress in Approximation Theory and Applicable Complex Analysis*. Cham: Springer (2017). p. 341–62. doi: 10.1007/978-3-319-49242-1_16
44. Filbir F, Mhaskar HN. A quadrature formula for diffusion polynomials corresponding to a generalized heat kernel. *J Fourier Anal Appl.* (2010) **16**:629–57. doi: 10.1007/s00041-010-9119-4
45. Mhaskar HN. A unified framework for harmonic analysis of functions on directed graphs and changing data. *Appl Comput Harm Anal.* (2018) **44**:611–44. doi: 10.1016/j.acha.2016.06.007
46. Rivlin TJ. *The Chebyshev Polynomials*. New York, NY: John Wiley and Sons (1974).
47. Grigorlyan A. Heat kernels on metric measure spaces with regular volume growth. *Handb Geometr Anal.* (2010) **2**. Available online at: https://www.math.uni-bielefeld.de/~grigor/hga.pdf.
48. Mhaskar HN. Approximate quadrature measures on data-defined spaces. In: Dick J, Kuo FY, Wozniakowski H, editors. *Festschrift for the 80th Birthday of Ian Sloan*. Berlin: Springer (2017). p. 931–62. doi: 10.1007/978-3-319-72456-0_41
49. Mhaskar HN. On the degree of approximation in multivariate weighted approximation. In: M. D. Buhman, and D. H. Mache, eds. *Advanced Problems in Constructive Approximation*. Basel: Birkhäuser (2003). p. 129–41. doi: 10.1007/978-3-0348-7600-1_10
50. Mhaskar HN. Approximation theory and neural networks. In: *Proceedings of the International Workshop in Wavelet Analysis and Applications*. Delhi (1999). p. 247–89.
51. Mhaskar HN, Narcowich FJ, Ward JD. Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature. *Math Comput.* (2001) **70**:1113–30. doi: 10.1090/S0025-5718-00-01240-0
52. Mhaskar HN. Dimension independent bounds for general shallow networks. *Neural Netw.* (2020) **123**:142–52. doi: 10.1016/j.neunet.2019.11.006
53. Hörmander L. The spectral function of an elliptic operator. *Acta Math.* (1968) **121**:193–218. doi: 10.1007/BF02391913
54. Shubin MA. *Pseudodifferential Operators and Spectral Theory*. Berlin: Springer (1987).
55. Grigor'yan A. Gaussian upper bounds for the heat kernel on arbitrary manifolds. *J Diff Geom.* (1997) **45**:33–52. doi: 10.4310/jdg/1214459753
56. Boucheron S, Lugosi G, Massart P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford University Press (2013).
57. Hagerup T, Rüb C. A guided tour of Chernoff bounds. *Inform Process Lett.* (1990) **33**:305–8. doi: 10.1016/0020-0190(90)90214-I

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

A. GAUSSIAN UPPER BOUND ON MANIFOLDS

Let $\mathbb{X}$ be a compact and connected smooth q-dimensional manifold, $g(x) = (g_{i,j}(x))$ be its metric tensor, and $(g^{i,j}(x))$ be the inverse of $g(x)$. The Laplace-Beltrami operator on $\mathbb{X}$ is defined by

$$\Delta(f)(x) = \frac{1}{\sqrt{|g(x)|}} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_i \left(\sqrt{|g(x)|}\, g^{i,j}(x) \partial_j f \right),$$

where $|g| = \det(g)$. The symbol of Δ is given by

$$a(x,\xi) = \frac{1}{\sqrt{|g(x)|}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sqrt{|g(x)|}\, g^{i,j}(x) \right) \xi_i \xi_j.$$

Then $a(x,\xi) \geq c|\xi|^2$. Therefore, Hörmander's theorem [53, Theorem 4.4], [54, Theorem 16.1] shows that for $x \in \mathbb{X}$,

$$\sum_{\lambda_j < \lambda} \phi_k(x)^2 \leq c\lambda^q, \qquad \lambda \geq 1. \tag{A.1}$$

In turn, [44, Proposition 4.1] implies that

$$\sum_{k=0}^{\infty} \exp(-\lambda_k^2 t)\phi_k(x)^2 \leq c t^{-q/2}, \qquad t \in (0,1],\ x \in \mathbb{X}.$$

Then [55, Theorem 1.1] shows that (3.3) is satisfied.

B. PROBABILISTIC ESTIMATES

We need the following basic facts from probability theory. Proposition B.1(a) below is a reformulation of Boucheron et al. [56, section 2.1, 2.7]. A proof of Proposition B.1(b) below is given in Hagerup and Rüb [57, Equation (7)].

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

$$\text{Prob}\left(\left| \frac{1}{M} \sum_{j=1}^{M} (Z_j - \mathbb{E}(Z_j)) \right| \geq t \right) \leq 2\exp\left(-\frac{Mt^2}{2(V + Rt)} \right). \tag{B.1}$$

*(b) (**Chernoff bound**) Let $M \geq 1$, $0 \leq p \leq 1$, and $Z_1, \cdots, Z_M$ be random variables taking values in $\{0,1\}$, with $\text{Prob}(Z_k = 1) = p$. Then for $t \in (0,1]$,*

$$\text{Prob}\left(\sum_{k=1}^{M} Z_k \leq (1-t)Mp \right) \leq \exp(-t^2 Mp/2),$$
$$\text{Prob}\left(\left| \sum_{k=1}^{M} Z_k - Mp \right| \geq tMp \right) \leq 2\exp(-t^2 Mp/2). \tag{B.2}$$