Neural Networks 132 (2020) 253–268

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A direct approach for function approximation on data defined manifolds

H.N. Mhaskar[1]

Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, United States of America

ARTICLE INFO

Article history:
Received 23 March 2020
Received in revised form 29 July 2020
Accepted 17 August 2020
Available online 25 August 2020

Keywords:
Manifold learning
Deep networks
Gaussian networks
Weighted polynomial approximation

ABSTRACT

In much of the literature on function approximation by deep networks, the function is assumed to be defined on some known domain, such as a cube or a sphere. In practice, the data might not be dense on these domains, and therefore, the approximation theory results are observed to be too conservative. In manifold learning, one assumes instead that the data is sampled from an unknown manifold; i.e., the manifold is defined by the data itself. Function approximation on this unknown manifold is then a two stage procedure: first, one approximates the Laplace–Beltrami operator (and its eigen-decomposition) on this manifold using a graph Laplacian, and next, approximates the target function using the eigen-functions. Alternatively, one estimates first some atlas on the manifold and then uses local approximation techniques based on the local coordinate charts.

In this paper, we propose a more direct approach to function approximation on *unknown*, data defined manifolds without computing the eigen-decomposition of some operator or an atlas for the manifold, and without any kind of training in the classical sense. Our constructions are universal; i.e., do not require the knowledge of any prior on the target function other than continuity on the manifold. We estimate the degree of approximation. For smooth functions, the estimates do not suffer from the so-called saturation phenomenon. We demonstrate via a property called good propagation of errors how the results can be lifted for function approximation using deep networks where each channel evaluates a Gaussian network on a possibly unknown manifold.

1. Introduction

One of the main problems of machine learning is the following. Given data $\{(\mathbf{y}_j, f(\mathbf{y}_j)+\epsilon_j)\}_{j=1}^M$, where f is an unknown function, $\mathbf{y}_j$'s are sampled randomly from a probability distribution μ^* defined on a subset of $\mathbb{R}^Q$ for some typically high dimension Q, and ϵ_j's are realizations of a mean zero random variable, find an approximation P from a class V_n to f (Cucker & Smale, 2002; Cucker & Zhou, 2007; Girosi & Poggio, 1990), where $\{V_n\}$ is a nested sequence of subsets of $L^2(\mu^*)$. In practice, this approximation is found by empirical risk minimization, assuming some prior on f, such as that it belongs to some reproducing kernel Hilbert space with a known kernel, or that it has a certain number of derivatives, or that it satisfies some conditions on its Fourier transform. To set up the minimization problem, one needs to know in advance the complexity of the model P, typically, the number of parameters desired to be estimated. In theory, the usual way of estimating this number is to estimate the so called approximation error, $\inf_{P\in V_n} \mathbb{E}_{\mu^*}((f-P)^2)$. Necessarily, this results in a fundamental gap in the theory, namely, that the minimizer of the empirical risk may have no connection with the minimizer of the approximation error.

Since the fundamental problem is one of function approximation, it is natural to wonder if appropriate tools in approximation theory can be developed in order to close this gap. One of the difficulties in doing so is that most of the results in classical approximation theory assume that the approximation takes place on a known domain, such as the cube, or Euclidean space, or sphere or similar known manifold. In turn, this requires that the data should be dense on this domain; i.e., the domain should be the (exact) support of μ^*. The problem is that μ^* being unknown, it is not possible to ensure this requirement.

During this century, manifold learning has sought to ameliorate the situation, with many practical applications. An early introduction to this topic is in the special issue (Chui & Donoho, 2006) of Applied and Computational Harmonic Analysis, edited by Chui and Donoho. In this theory, one assumes that the support of μ^* is an unknown smooth compact connected manifold; for simplicity, even that μ^* is the Riemannian volume measure for the manifold, normalized to be a probability measure. Following, e.g., Belkin and Niyogi (2003, 2004, 2008), Lafon (2004) and Singer (2006), one constructs first a "graph Laplacian" from the

E-mail address: hrushikesh.mhaskar@cgu.edu.

[1] The research is supported in part by National Science Foundation, United States of America grant DMS 2012355.

https://doi.org/10.1016/j.neunet.2020.08.018
0893-6080/

data, and finds its eigen-decomposition. It is proved in the above mentioned 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 eigen-values (respectively, eigen-vectors) 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., Jones, Maggioni, & Schul, 2010; Liao & Maggioni, 2016), based on the so called heat kernel. The theory of function approximation on such manifolds is also well developed (e.g., Ehler, Filbir, & Mhaskar, 2012; Filbir & Mhaskar, 2011; Maggioni & Mhaskar, 2008; Mhaskar, 2010, 2011).

All this work depends upon a two stage procedure — finding the eigen-decomposition of the graph Laplacian and then using approximation in terms of the eigen-vectors/eigen-functions. Once more, this leads to errors not just from the approximation of the target function but also from the approximation of the eigen-decomposition of the Laplace–Beltrami operator itself. In recent years, there are some efforts to explore alternative approaches using deep networks (e.g., Chen, Jiang, Liao, & Zhao, 2019; Chui & Mhaskar, 2018b; Cloninger, Coifman, Downing, & Krumholz, 2015; Schmidt-Hieber, 2019). These papers also take a two-step approach: developing an atlas on the manifold first, and then using some local approximation schemes based on the local coordinate charts.

Our objective in this paper is to develop a single-shot method to solve the problem, knowing only the dimension of the manifold. In particular, we aim not to find any eigen-decomposition nor to learn any atlas on the manifold, but to give a direct construction that starts with the data and constructs an approximation without involving any optimization/training and with guaranteed approximation error estimated in a probabilistic sense. Our approximation can be implemented as a Gaussian network; i.e., a function of the form $\mathbf{x} \mapsto \sum_k a_k \exp(-\lambda|\mathbf{x}-\mathbf{y}_k|^2_{2,Q})$, where $|\cdot|_{2,Q}$ denotes the ℓ^2 norm on $\mathbb{R}^Q$. The size of the data set required depends only on the dimension of the manifold and the smoothness of the target function measured in a technical manner as explained in this paper. We will extend our results to approximation by deep Gaussian networks.

2. Technical introduction and outline

In this section, let us assume that the data $\mathbf{y}_j$ is sampled from some unknown manifold, uniformly with respect to the Riemannian volume element of that manifold. One of the fundamental results in manifold learning is the following theorem of Belkin and Niyogi (2008).

Theorem 2.1. *Let $\mathbb{X}$ be a smooth, compact, q-dimensional submanifold of $\mathbb{R}^Q$, μ^* be its Riemannian volume measure, normalized by $\mu^*(\mathbb{X}) = 1$, and Δ denote the Laplace–Beltrami operator on $\mathbb{X}$. Then for a smooth function $f : \mathbb{X} \to \mathbb{R}$,*

$$\lim_{t\to 0} \frac{1}{t(4\pi t)^{q/2}} \int_{\mathbb{X}} \exp\left(-\frac{|\mathbf{x}-\mathbf{y}|^2_{2,Q}}{t}\right)(f(\mathbf{y})-f(\mathbf{x}))d\mu^*(\mathbf{y}) = \Delta(f)(\mathbf{x}) \tag{2.1}$$

uniformly for $\mathbf{x} \in \mathbb{X}$, where $|\cdot|_{2,Q}$ denotes the ℓ^2 norm on $\mathbb{R}^Q$. Equivalently, uniformly for $\mathbf{x} \in \mathbb{X}$, we have

$$\left|\frac{1}{(4\pi t)^{q/2}} \int_{\mathbb{X}} \exp\left(-\frac{|\mathbf{x}-\mathbf{y}|^2_{2,Q}}{t}\right)(f(\mathbf{y})-f(\mathbf{x}))d\mu^*(\mathbf{y}) - t\Delta(f)(\mathbf{x})\right| = o(t) \tag{2.2}$$

as $t \to 0+$.

From an approximation theory point of view, the theorem is more of a saturation theorem for approximating f on $\mathbb{X}$, analogous to the Voronovskaya estimates for Bernstein polynomials (Lorentz, 2013, Section 1.6.1, See Appendix). Thus, (2.2) states that the rate of approximation of f cannot be better than $\mathcal{O}(t)$, even if f is infinitely differentiable, unless f is in the null space of the Laplace–Beltrami operator. This is to be expected because the Gaussian kernel involved is a positive operator. In particular, this phenomenon holds even if $\mathbb{X}$ is a Euclidean space rather than a manifold. Moreover, the curvature of the manifold contributes to the saturation as well. The Gaussian kernel has many advantages, invariance under translations and rotations is one of the them. This plays a major role in the proof of Theorem 2.1. Nevertheless, it is natural to ask whether another kernel can be found that leads directly to the approximation of the target function f on the manifold from the data without knowing the manifold itself and without having to go through an expensive eigen-decomposition. The curvature of the manifold will still affect the rate of convergence, but when applied to an affine space rather than a manifold, such a construction should lead to approximation without any saturation, without knowing what the affine space is (Remark 3.3).

The main objective of this paper is to demonstrate such a construction using certain localized kernels based on Hermite polynomials (Theorem 3.1). This theorem gives an analogue of Theorem 2.1 to obtain function approximation on an unknown manifold based only on noise-corrupted samples on the manifold, and give estimates on the degree of approximation. In the case when the approximation is done on an affine space rather than a manifold, our construction is free of any saturation, and does not need to know what the affine space is (Theorem 7.1).

To recapture the advantage of the Gaussian kernel, we will study approximation by Gaussian networks. A (shallow) Gaussian network with n neurons has the form $\mathbf{x} \mapsto \sum_{k=1}^n a_k \exp(-\lambda|\mathbf{x}-\mathbf{y}_k|^2_{2,Q})$. A deep Gaussian network is constructed following a DAG structure, where each node (referred to as "channel" in the literature on deep learning) evaluates a Gaussian network. Using the close connection between Hermite polynomials and Gaussian networks (cf. Chui & Mhaskar, 2019; Mhaskar, 1996, 2004), we can translate the result about approximation on the manifold into a result on approximation by shallow Gaussian networks, where the input is assumed to lie on an unknown low dimensional manifold of the nominally high dimensional ambient space (Theorem 5.1). In turn, using a property called "good propagation of errors" (Theorem 5.2), we will "lift" this theorem to estimate the degree of approximation by deep Gaussian networks, where each channel evaluates a Gaussian network on a similarly manifold-based data (Theorem 5.3). The networks themselves are constructed from certain **pre-fabricated networks** in the ambient space to approximate the Hermite functions with a correspondingly high number of neurons. However, we will give an explicit formula for such networks (Proposition 6.6), so that there is **no training required here**. The amount of information used in the final synthesis of the network will depend only on the dimension of the manifold on which the input lives. We consider this to be a step in bringing approximation theory of deep networks closer to the practice, so that the results are proved in the setting of approximation on unknown manifolds analogous to diffusion geometry rather than on known domains.

The statement of the main results in this paper mentioned above require a good deal of background information on the theory of weighted polynomial approximation, which we defer to Section 6. We will state the main results about approximation on a manifold in Section 3, and illustrate them using a simple numerical example in Section 4. We explain our ideas about shallow and deep networks in Section 5. To develop the details required

in the constructions and proofs, we start by summarizing the relevant facts from the theory of weighted polynomial approximation in Section 6. Of particular interest is the approximation of a weighted polynomial using pre-fabricated Gaussian networks whose weights and centers do not depend upon the polynomial, as described in Section 6.3. Our main theorem in the context of approximation on unknown affine spaces is stated and proved in Section 7. The proofs of the results in Sections 3 and 5 are given in Sections 8 and 9 respectively.

3. Approximation on manifolds

In this section, we state our main results on approximation on manifolds. The details and motivations for these constructions will be clearer after reading Sections 6 and 7. The notation on the manifolds is described in Section 3.1, the results themselves are discussed in Section 3.2.

3.1. Definitions

Let $Q \geq q \geq 1$ be integers, $\mathbb{X}$ be a q dimensional, compact, connected, sub-manifold of $\mathbb{R}^Q$ (without boundary), with geodesic distance ρ and volume measure μ^*, normalized so that $\mu^*(\mathbb{X}) = 1$. We will identify the tangent space at $\mathbf{x} \in \mathbb{X}$ with an affine space $\mathbb{T}_{\mathbf{x}}(\mathbb{X})$ in $\mathbb{R}^Q$ passing through $\mathbf{x}$. For any $\mathbf{x} \in \mathbb{X}$, we need to consider in this section three kinds of balls.

$$B_Q(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{R}^Q : |\mathbf{x} - \mathbf{y}|_{2,Q} \leq r\}, \quad B_{\mathbb{T}}(\mathbf{x}, r) := \mathbb{T}_{\mathbf{x}}(\mathbb{X}) \cap B_Q(\mathbf{x}, r), \quad \mathbb{B}(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{X} : \rho(\mathbf{x}, \mathbf{y}) \leq r\}. \tag{3.1}$$

With this convention, the exponential map $\mathcal{E}_{\mathbf{x}}$ at $\mathbf{x} \in \mathbb{X}$ (based on the definition in (do Carmo Valero, 1992, Proposition 2.9)) is a diffeomorphism of an open ball centered at $\mathbf{x}$ in $\mathbb{T}_{\mathbf{x}}(\mathbb{X})$ onto its image in $\mathbb{X}$ such that $\rho(\mathbf{x}, \mathcal{E}_{\mathbf{x}}(\mathbf{u})) = |\mathbf{u} - \mathbf{x}|_{2,Q}$. Since $\mathbb{X}$ is compact, there exists $\iota^* > 0$ such that for ***every*** $\mathbf{x} \in \mathbb{X}$, $\mathcal{E}_{\mathbf{x}}$ is defined on $B_{\mathbb{T}}(\mathbf{x}, \iota^*)$, and $\rho(\mathbf{x}, \mathcal{E}_{\mathbf{x}}(\mathbf{u})) = |\mathbf{u} - \mathbf{x}|_{2,Q}$ for all $\mathbf{u} \in B_{\mathbb{T}}(\mathbf{x}, \iota^*)$.

We now define the smoothness class $W_\gamma(\mathbb{X})$. If $f, g : \mathbb{X} \to \mathbb{R}$, the function $fg : \mathbb{X} \to \mathbb{R}$ is defined as usual by $(fg)(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x})$ for $\mathbf{x} \in \mathbb{X}$. The space $C(\mathbb{X})$ is the space of all continuous real-valued functions on $\mathbb{X}$, equipped with the supremum norm $\|\circ\|_{\mathbb{X}}$. The space $C^\infty(\mathbb{X})$ is the subspace of $C(\mathbb{X})$ comprising all infinitely differentiable functions on $\mathbb{X}$. Let $f \in C(\mathbb{X})$, $\gamma > 0$. We say that $f \in W_\gamma(\mathbb{X})$ if for every $\mathbf{x} \in \mathbb{X}$, and $\phi \in C^\infty(\mathbb{X})$, supported on $\mathbb{B}(\mathbf{x}, \iota^*/2)$, the function $F_{\mathbf{x},\phi} : \mathbb{T}_{\mathbf{x}}(\mathbb{X}) \to \mathbb{R}$ defined by $F_{\mathbf{x},\phi}(\mathbf{u}) := f(\mathcal{E}_{\mathbf{x}}(\mathbf{u}))\phi(\mathcal{E}_{\mathbf{x}}(\mathbf{u}))$ is in $W_\gamma(\mathbb{T}_{\mathbf{x}}(\mathbb{X}))$ in the sense described in Section 7 (See (6.44), (7.3), and (7.4)). We define

$$\|f\|_{W_\gamma(\mathbb{X})} := \sup_{\mathbf{x}\in\mathbb{X}, \|\phi\|_{\mathbb{X}} \leq 1} \|F_{\mathbf{x},\phi}\|_{W_\gamma(\mathbb{T}_{\mathbf{x}}(\mathbb{X}))}. \tag{3.2}$$

If γ is an integer and f is γ times differentiable on $\mathbb{X}$ then $f \in W_\gamma(\mathbb{X})$. The space $W_\gamma(\mathbb{X})$ can contain functions which are not differentiable. For example, we say that $f \in \mathrm{Lip}(\mathbb{X})$ if

$$\|f\|_{\mathrm{Lip}(\mathbb{X})} := \sup_{\mathbf{x},\mathbf{y}\in\mathbb{X}, \mathbf{x}\neq\mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\rho(\mathbf{x}, \mathbf{y})} < \infty.$$

We have $\mathrm{Lip}(\mathbb{X}) \subset W_1(\mathbb{X})$.

Next, we define the approximation operators. The orthonormalized Hermite polynomial h_k of degree k is defined recursively by

$$h_k(x) := \sqrt{\frac{2}{k}} x h_{k-1}(x) - \sqrt{\frac{k-1}{k}} h_{k-2}(x), \qquad k = 2, 3, \ldots, \quad h_0(x) := \pi^{-1/4}, \quad h_1(t) := \sqrt{2}\pi^{-1/4} x. \tag{3.3}$$

We write

$$\psi_k(t) := h_k(t)\exp(-t^2/2), \qquad t \in \mathbb{R}, \ k \in \mathbb{Z}_+. \tag{3.4}$$

The functions $\{\psi_k\}_{k=0}^\infty$ are an orthonormal set with respect to the Lebesgue measure (cf. (6.1)). In the sequel, we fix an infinitely differentiable function $H : [0, \infty) \to [0, 1]$, such that $H(t) = 1$ if $0 \leq t \leq 1/2$, and $H(t) = 0$ if $t \geq 1$. We define for $x \in \mathbb{R}$, $m \in \mathbb{Z}_+$:

$$\mathcal{P}_{m,q}(x) := \begin{cases} \pi^{-1/4}(-1)^m \dfrac{\sqrt{(2m)!}}{2^m m!} \psi_{2m}(x), & \text{if } q = 1, \\ \dfrac{1}{\pi^{(2q-1)/4}\Gamma((q-1)/2)} \displaystyle\sum_{\ell=0}^{m} (-1)^\ell & \\ \times \dfrac{\Gamma((q-1)/2 + m - \ell)}{(m-\ell)!} \dfrac{\sqrt{(2\ell)!}}{2^\ell \ell!} \psi_{2\ell}(x), & \text{if } q \geq 2, \end{cases} \tag{3.5}$$

and the kernel $\widetilde{\Phi}_{n,q}$ for $x \in \mathbb{R}$, $n \in \mathbb{Z}_+$ by

$$\widetilde{\Phi}_{n,q}(x) := \sum_{m=0}^{\lfloor n^2/2 \rfloor} H\left(\frac{\sqrt{2m}}{n}\right) \mathcal{P}_{m,q}(x). \tag{3.6}$$

Constant convention:

In the sequel, $c, c_1, \ldots$ will denote generic positive constants depending upon the dimension and other fixed quantities in the discussion, such as the norm. Their values may be different at different occurrences, even within a single formula. The notation $A \sim B$ means $c_1 A \leq B \leq c_2 B$. ■

3.2. Approximation theorems

The traditional machine learning paradigm is to consider data of the form $\{(\mathbf{y}_j, f(\mathbf{y}_j) + \epsilon_j)\}$, where $\mathbf{y}_j$'s are drawn randomly with respect to μ^* and ϵ_j's are random, mean 0 samples from an unknown distribution. More generally, we assume here a noisy data of the form $(\mathbf{y}, \epsilon)$, with a joint probability distribution τ and assume further that the marginal distribution of $\mathbf{y}$ with respect to τ has the form $d\nu^* = f_0 d\mu^*$ for some $f_0 \in C(\mathbb{X})$. In place of $f(\mathbf{y})$, we consider a noisy variant $\mathcal{F}(\mathbf{y}, \epsilon)$, and denote

$$f(\mathbf{y}) := \mathbb{E}_\tau(\mathcal{F}(\mathbf{y}, \epsilon)|\mathbf{y}). \tag{3.7}$$

Remark 3.1. In practice, the data may not lie on a manifold, but it is reasonable to assume that it lies on a tubular neighborhood of the manifold. Our notation accommodates this — if $\mathbf{z}$ is a point in a neighborhood of $\mathbb{X}$, we may view it as a perturbation of a point $\mathbf{y} \in \mathbb{X}$, so that the noisy value of the target function is $\mathcal{F}(\mathbf{y}, \epsilon)$, where ϵ encapsulates the noise in both the $\mathbf{y}$ variable and the value of the target function. An example is given in Example 4.1. ■

Our approximation process is simple: given by

$$\widehat{F}_{n,\alpha}(Y; \mathbf{x}) := \frac{n^{q(1-\alpha)}}{M} \sum_{j=1}^{M} \mathcal{F}(\mathbf{y}_j, \epsilon_j) \tilde{\Phi}_{n,q}(n^{1-\alpha}|\mathbf{x} - \mathbf{y}_j|_{2,Q}), \quad \mathbf{x} \in \mathbb{R}^Q, \tag{3.8}$$

where $0 < \alpha \leq 1$.

Our main theorem is the following.

Theorem 3.1. *Let $\gamma > 0$, τ be a probability distribution on $\mathbb{X} \times \Omega$ for some sample space Ω such the marginal distribution of τ restricted to $\mathbb{X}$ is absolutely continuous with respect to μ^* with density $f_0 \in W_\gamma(\mathbb{X})$. We assume that*

$$\sup_{\mathbf{x}\in\mathbb{X}, r>0} \frac{\mu^*(\mathbb{B}(\mathbf{x}, r))}{r^q} \leq c. \tag{3.9}$$

Let $\mathcal{F} : \mathbb{X} \times \Omega \to \mathbb{R}$ be a bounded function, f defined by (3.7) be in $W_\gamma(\mathbb{X})$, the probability density $f_0 \in W_\gamma(\mathbb{X})$. Let $M \geq 1$,

$Y = \{(\mathbf{y}_1, \epsilon_1), \dots, (y_M, \epsilon_M)\}$ be a set of random samples chosen i.i.d. from τ. If

$$0 < \alpha < \frac{4}{2+\gamma}, \qquad \alpha \le 1, \tag{3.10}$$

then for every $n \ge 1$, $0 < \delta < 1$ and $M \ge n^{q(2-\alpha)+2\alpha\gamma}\sqrt{\log(n/\delta)}$, we have with τ-probability $\ge 1-\delta$:

$$\left\|\widehat{F}_{n,\alpha}(Y;\circ) - f_0 f\right\|_{\mathbb{X}} \le c_1 \frac{\sqrt{\|f_0\|_{\mathbb{X}}}\|\mathcal{F}\|_{\mathbb{X}\times\Omega} + \|f_0 f\|_{W_\gamma(\mathbb{X})}}{n^{\alpha\gamma}}. \tag{3.11}$$

We record two corollaries of Theorem 3.1 as separate theorems. The first is the approximation of f itself, assuming that $f_0 \equiv 1$.

Theorem 3.2. *With the set-up as in Theorem 3.1, let $f_0 \equiv 1$ (i.e., the marginal distribution of $\mathbf{y}$ with respect to τ is μ^*). Then we have with τ-probability $\ge 1-\delta$:*

$$\left\|\widehat{F}_{n,\alpha}(Y;\circ) - f\right\|_{\mathbb{X}} \le c_1 \frac{\|\mathcal{F}\|_{\mathbb{X}\times\Omega} + \|f\|_{W_\gamma(\mathbb{X})}}{n^{\alpha\gamma}}. \tag{3.12}$$

The second is a consequence analogous to Theorem 2.1.

Theorem 3.3. *With the set-up as in Theorem 3.1, we have with τ-probability $\ge 1-\delta$:*

$$\left\| \frac{n^{q(1-\alpha)}}{M} \sum_{j=1}^{M} \left(\mathcal{F}(\mathbf{y}_j, \epsilon_j) - f(\circ)\right) \tilde{\Phi}_{n,q}(n^{1-\alpha}|\circ - \mathbf{y}_j|_{2,\mathbb{Q}}) \right\|_{\mathbb{X}}$$
$$\le c_1 \frac{\sqrt{\|f_0\|_{\mathbb{X}}}\|\mathcal{F}\|_{\mathbb{X}\times\Omega} + \|f_0 f\|_{W_\gamma(\mathbb{X})}}{n^{\alpha\gamma}}. \tag{3.13}$$

Remark 3.2. To compare the estimate (3.13) with (2.2), which is applicable with $\gamma = 2$, we are tempted to take any $\alpha \in (0,1)$, set $t = n^{-2(1-\alpha)}$, and obtain the upper bound t^A with $A = \alpha/(1-\alpha)$. Clearly, this bound tends to 0 arbitrarily fast with t. However, the estimate (2.2) uses a fixed kernel, while the estimate (3.13) uses a kernel depending upon t. ■

Remark 3.3. Although the curvature of the manifold forces us to put limitations on the rate of convergence in (3.12), this is not a saturation phenomenon. Thus, it is not ruled out that the rate can be much better than that given in (3.12) for non-trivial functions. ■

Remark 3.4. If $\gamma < 2$, we may choose $\alpha = 1$ without knowing the value of γ. The formula (3.8) itself does not require any prior knowledge of the smoothness of f. ■

4. Numerical example

We illustrate the theory using the following simple example. We let $\mathbb{X} \subset \mathbb{R}^3$ to be the helix defined by

$$\mathbf{x}(t) := (\cos(\pi t), \sin(\pi t), \pi t), \qquad 0 \le t \le 2\pi. \tag{4.1}$$

This does not satisfy the conditions of the theorems in Section 3, and we will see an "end point effect" in the errors, but we find it easy to work with this example because of the ease in computing the various quantities like the volume measure (arc-length) : $d\mu^* = (\sqrt{8}\pi^2)^{-1}dt$. The target function f is given by

$$f(\mathbf{x}(t)) := \cos(x_1(t) - x_2(t) + x_3(t)/2) = \cos(\cos(\pi t)$$
$$- \sin(\pi t) - \pi t/2), \qquad 0 \le t \le 2\pi. \tag{4.2}$$

Example 4.1. We consider data of the form

$$\mathcal{F}(\mathbf{y}, \epsilon) := f(\mathbf{y} + \epsilon)\exp(1.125), \tag{4.3}$$

where ϵ is a random normal variable with mean 0 and standard deviation 1.5. The factor $\exp(1.125)$ ensures that the expected value of $\mathcal{F}$ is f. This example illustrates a multiplicative noise as well as additive noise. We may also consider this to be an example where every point $\mathbf{y}$ on the helix is perturbed by a normal noise with mean 0 and standard deviation 1, although we cannot deal directly with the perturbed points in the calculation of $\widehat{F}_{n,\alpha}$. We took $M = 256$, $n = 64$, $\alpha = 1$. The results are reported in Fig. 1 on one trial, as well as the average of $\widehat{F}_{n,\alpha}$ over 100 trials. ■

Example 4.2. We consider data of the form

$$\mathcal{F}(\mathbf{y}, \epsilon) := f(\mathbf{y}) + \epsilon, \tag{4.4}$$

where ϵ is a random normal variable with mean 0 and standard deviation 0.3. We take $M = 1024$, and M samples of $\mathbf{y}$ distributed uniformly according to μ^*. We take $n = 64$, $\alpha = 1$, and compute the quantity $\widehat{F}_{64,1}(Y, \mathbf{x})$ for $\mathbf{x} = \mathbf{x}(t)$, where t ranges over 2048 equidistant points on $[0, 2\pi]$. The results are shown in Fig. 2. ■

5. Gaussian networks

In this section, we describe the consequences of Theorem 3.1 for Gaussian networks. In the case of shallow networks, we can give an explicit construction and error bounds in Section 5.1. In the case of deep networks (Section 5.2), we give only an existence theorem, explaining when the theorem can be described more constructively.

5.1. Shallow networks

Since $\mathcal{P}_{m,q}$ and hence $\widetilde{\Phi}_{n,q}$ are even polynomials of degree $\le n^2$, $\widetilde{\Phi}_{n,q}(n^{1-\alpha}|\circ|_{2,\mathbb{Q}}) \in \Pi_n^{\mathbb{Q}}$. We will see in Remark 6.2 that $\widetilde{\Phi}_{n,q}(n^{1-\alpha}|\circ|_{2,\mathbb{Q}}) = \Phi_{n,q,\mathbb{Q}}(\mathbf{0}, n^{1-\alpha}(\circ))$ for a polynomial kernel $\Phi_{n,q,\mathbb{Q}}$ on $\mathbb{R}^{\mathbb{Q}}$. We may then define a **pre-fabricated** Gaussian network using (6.41)

$$\mathbb{G}^*_{n,q,\mathbb{Q}} := \mathfrak{G}_{\mathbb{Q}}(\Phi_{n,q,\mathbb{Q}}(\mathbf{0}, (n^{1-\alpha}(\circ)_{2,\mathbb{Q}}))). \tag{5.1}$$

Using Corollary 6.2, we then deduce easily the following theorem about Gaussian networks. We note again that there is no training involved here. Even though the number of non-linearities in the network in the following theorem is $\mathcal{O}(Mn^{2\mathbb{Q}})$, this potentially large number of non-linearities is not as much of a problem as it would be if we were to use an optimization procedure to train the network.

Theorem 5.1. *Let (3.9) be satisfied, $\gamma > 0$, τ be a probability distribution on $\mathbb{X} \times \Omega$ for some sample space Ω such the marginal distribution of τ restricted to $\mathbb{X}$ is ν^* with $d\nu^* = f_0 d\mu^*$ for some $f_0 \in W_\gamma(\mathbb{X})$. Let $\mathcal{F} : \mathbb{X} \times \Omega \to \mathbb{R}$ be a bounded function, and f defined by (3.7) be in $W_\gamma(\mathbb{X})$. Let $0 < \delta < 1$, α satisfy (3.10). Let $M \ge 1$, $Y = \{(\mathbf{y}_1, \epsilon_1), \dots, (y_M, \epsilon_M)\}$ be a set of random samples chosen i.i.d. from τ. If*

$$M \ge n^{q(2-\alpha)+2\alpha\gamma}\sqrt{\log(n/\delta)} \tag{5.2}$$

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

$$\left\| \frac{1}{M} \sum_{j=1}^{M} \left(\mathcal{F}(\mathbf{y}_j, \epsilon_j) - f(\circ)\right) \mathbb{G}^*_{n,q,\mathbb{Q}}(\circ - \mathbf{y}_j) \right\|_{\mathbb{X}}$$
$$\le c_1 \frac{\sqrt{\|f_0\|_{\mathbb{X}}}\|\mathcal{F}\|_{\mathbb{X}\times\Omega} + \|f_0 f\|_{W_\gamma(\mathbb{X})}}{n^{\alpha\gamma}}. \tag{5.3}$$

In particular, let

$$\mathbb{G}_{n,q,\mathbb{Q}}(Y;\mathcal{F})(\mathbf{x}) := \frac{1}{M}\sum_{j=1}^{M} \mathcal{F}(\mathbf{y}_j, \epsilon_j)\mathbb{G}^*_{n,q,\mathbb{Q}}(\mathbf{x} - \mathbf{y}_j), \qquad \mathbf{x} \in \mathbb{R}^{\mathbb{Q}}. \tag{5.4}$$

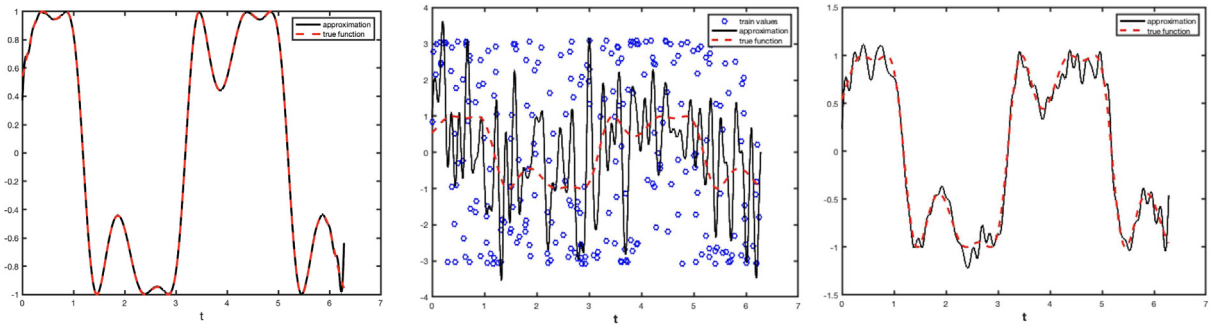

Fig. 1. In all figures, black continuous line is the approximation, dashed line (red in color) is the target function (4.2). Left: Reconstruction without noise using 256 random training points, 2048 equidistant test points, Middle: Estimate in one trial, 256 random training points (dots, blue in color) according to (4.3), 2048 equidistant test points, Right: Average of the estimates in 100 trials, 256 random training points plus noise each, 2048 equidistant test points.

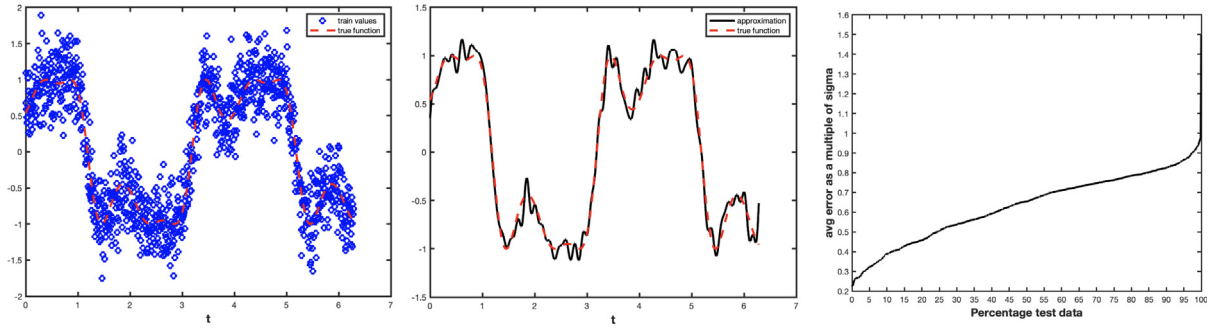

Fig. 2. With f and $\mathcal{F}$ as in (4.2) and (4.4) respectively, $M = 1024$, $n = 64$, $\alpha = 1$. On the x axis are 2048 equidistant samples on $[0, 2\pi]$. Left: The function f in dashed line (red in color), the sampled values $\mathcal{F}$ in dots (blue in color) for one trial, Middle: The function f in dashed line (red in color), the reconstruction $\widehat{F}_{64,1}$ for one trial in black, Right: A cumulative histogram of errors over 50 trials, the point (p, y) signifies that the error is $0.3y$ at $p\%$ of the test data.

If $f_0 \equiv 1$, we have with τ-probability $\geq 1 - \delta$:

$$\left\| \mathbb{G}_{n,q,Q}(Y; \mathcal{F}) - f \right\|_{\mathbb{X}} \leq c_1 \frac{\|\mathcal{F}\|_{\mathbb{X}\times\Omega} + \|f\|_{W_\gamma(\mathbb{X})}}{n^{\alpha\gamma}}. \tag{5.5}$$

5.2. Deep networks

The following discussion about the terminology about the deep networks is based on (almost taken from) the discussion in Mhaskar and Poggio (2016, 2020), and elaborates upon the same. In particular, Fig. 3 is taken from the arxiv version of Mhaskar and Poggio (2016).

A commonly used definition of a deep network is the following. Let $\phi : \mathbb{R} \to \mathbb{R}$ be an activation function; applied to a vector $\mathbf{x} = (x_1, \ldots, x_q)$, $\phi(\mathbf{x}) = (\phi(x_1), \ldots, \phi(x_q))$. Let $L \geq 2$ be an integer, for $\ell = 0, \ldots, L$, let $q_\ell \geq 1$ be an integer ($q_0 = q$), $T_\ell : \mathbb{R}^{q_\ell} \to \mathbb{R}^{q_{\ell+1}}$ be an affine transform, where $q_{L+1} = 1$. A deep network with $L-1$ hidden layers is defined as the compositional function

$$\mathbf{x} \mapsto T_L \phi T_{L-1} \phi (T_{L-2} \cdots \phi(T_0(\mathbf{x})) \cdots). \tag{5.6}$$

This definition has several shortcomings. First, it does not distinguish between a function and the network architecture. As demonstrated in Mhaskar and Poggio (2020), a function may have more than one compositional representation, so that the affine transforms and L are not determined uniquely by the function itself. Second, this notion does not capture the connection between the nature of the target function and its approximation. Third, the affine transforms T_ℓ define a special directed acyclic graph (DAG). It is cumbersome to describe notions of weight sharing, convolutions, sparsity, skipping of layers, etc. in terms of these transforms. Therefore, we have proposed in Mhaskar and Poggio (2016) to separate the architecture from the function itself, and describe a deep network more generally as a directed acyclic graph (DAG) architecture.

Let $\mathcal{G}$ be a DAG, with the set of nodes $V \cup \mathbf{S}$, where $\mathbf{S}$ is the set of source nodes, and V that of non-source nodes. For each node $v \in V \cup \mathbf{S}$, we denote its in-degree by $d(v)$. Associated with each $v \in V \cup \mathbf{S}$ is a compact, connected, Riemannian submanifold $\mathbb{X}_v$ of $\mathbb{R}^{d(v)}$ with dimension q_v, metric ρ_v and volume element μ_v^*. We assume further that (3.9) is satisfied with q_v in place of q. Each of the in-edges to each node in $V \cup \mathbf{S}$ represents an input real variable. If $v \in V$, $u \in V \cup \mathbf{S}$, u is called the **child** of v if there is an edge from u to v. The notion of the level of a node is defined as follows. The level of a source node is 0. The level of $v \in V$ is the length of the longest path from the nodes in $\mathbf{S}$ to v.

Each node v is supposed to evaluate a function f_v on its input variables, supplied via the in-edges for v. The value of this function is propagated along the out-edges of v. Each of the source nodes obtains an input from some smooth manifold as described in Section 3. Other nodes can also obtain such an input, but by introducing dummy nodes, it is convenient to assume that only the source nodes obtain an input from the manifold.

Intuitively, we wish to say that the DAG structure implies a compositional structure for the functions involved; for example, if $u_1, \ldots, u_{d(v)}$ are children of v, then the function evaluated at v is $f_v(f_{u_1}, \ldots, f_{u_{d(v)}})$. To make this meaningful, we have to assume some "pooling" operation on the input variables to make sure that the output of the vector valued function $(f_{u_1}, \ldots, f_{u_{d(v)}})$ belongs to $\mathbb{X}_v$. Thus, for example, if the domain of f_v is the cube $[-1, 1]^{d(v)}$, some clipping operation is required; if the domain is the torus in $d(v)$ dimensions then some standard substitutions need to be made (e.g., Mhaskar & Poggio, 2020). We do not know how to specify the pooling operation in the general case of an unknown manifold, but assume that this pooling operation $\pi_v : \mathbb{R}^{d(v)} \to \mathbb{X}_v$ has the following property: For any two sets of

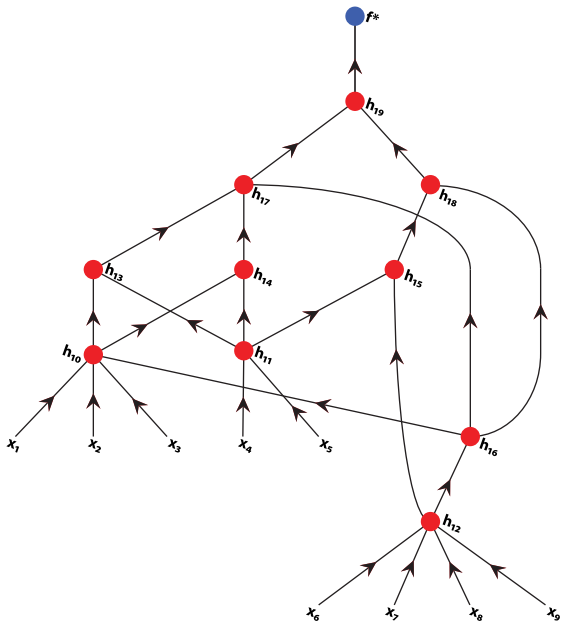

Fig. 3. An example of a $\mathcal{G}$-function (f^* given in (5.8)). The vertices of the DAG $\mathcal{G}$ are the channels of the network. The input to the various channels is indicated by the in-edges of the nodes, and the output of the sink node h_{19} indicates the output value of the $\mathcal{G}$-function, f^* in this example.

functions $\{f_v \in C(\mathbb{X}_v)\}_{v\in V}$, $\{g_v \in C(\mathbb{X}_v)\}_{v\in V}$,

$$\rho_v\left(\pi_v(f_{u_1}(\mathbf{x}_{u_1}),\dots,f_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}})),\pi_v(g_{u_1}(\mathbf{x}_{u_1}),\dots,g_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)$$
$$\le c(v)\sum_{k=1}^{d(v)}\|f_{u_k}-g_{u_k}\|_{\mathbb{X}_{u_k}}, \qquad v\in V. \tag{5.7}$$

A $\mathcal{G}$-function is defined to be a set of functions $\{f_v\}_{v\in V\cup\mathbf{S}}$ such that each $f_v \in C(\mathbb{X}_v)$, and if $v \in V$, $u_1,\dots,u_{d(v)}$ are children of v, then the function evaluated at v is $f_v(\pi_v(f_{u_1},\dots,f_{u_{d(v)}}))$. The individual functions f_v will be called *constituent functions*.

For example, the DAG $\mathcal{G}$ in Fig. 3 (Mhaskar & Poggio, 2016) represents the compositional function

$$\begin{aligned} f^*(x_1,\dots,x_9) = h_{19}(h_{17}(h_{13}(h_{10}(x_1,x_2,x_3,h_{16}(h_{12}(x_6,x_7,x_8,x_9))),\\ h_{11}(x_4,x_5)),\\ h_{14}(h_{10},h_{11}),h_{16}),\\ h_{18}(h_{15}(h_{11},h_{12}),h_{16})). \end{aligned} \tag{5.8}$$

The $\mathcal{G}$-function is $\{h_{10},\dots,h_{19}=f^*\}$.

We assume that there is only one sink node, v^* (or $v^*(\mathcal{G})$) whose output is denoted by f_{v^*} (the *target function*). Technically, there are two functions involved here: one is the final output as a function of all the inputs to all source nodes, the other is the final output as a function of the inputs to the node v^*. We will use the symbol f_{v^*} to denote both with comments on which meaning is intended when we feel that it may not be clear from the context. A similar convention is followed with respect to each of the constituent functions as well. For example, in the DAG of Fig. 3, the function h_{15} can be thought of both as a function of two variables, namely the outputs of h_{11} and h_{12} as well as a function of six variables $x_4,\dots,x_9$. In particular, if each constituent function is a neural network, h_{15} is a shallow network receiving two inputs.

We define the notion of the variables "seen" by a node. If $u \in \mathbf{S}$, then these are the variables input to u. Let $v \in V$, and $u_1,\dots,u_{d(v)}$ be the children of v. If $\mathbf{x}_1,\dots,\mathbf{x}_{d(v)}$ are the inputs seen by $u_1,\dots,u_{d(v)}$, then the inputs seen by v are $(\mathbf{x}_1,\dots,\mathbf{x}_{d(v)})$, where the order is respected. For example, consider the function

$$f^*(x_1,x_2,x_3,x_4)=f(f_1(x_1,x_2),f_2(x_4,x_2),f_3(x_3,x_1)).$$

The inputs seen by the leaves f_1,f_2,f_3 are (x_1,x_2), (x_4,x_2), (x_3,x_1) respectively (not (x_1,x_2), (x_2,x_4), (x_1,x_3)). The inputs seen by f^* are (x_1,x_2,x_3,x_4).

The following theorem enables us to "lift" a theorem about shallow networks to that about deep networks.

Theorem 5.2. *Let $\mathcal{G}$ be a DAG as described above, $\{f_v\}_{v\in V\cup\mathbf{S}}$, $\{g_v\}_{v\in V\cup\mathbf{S}}$ be $\mathcal{G}$-functions, and*

$$\|f_v-g_v\|_{\mathbb{X}_v}\le\varepsilon, \qquad v\in V\cup\mathbf{S}. \tag{5.9}$$

Further assume that for each $v\in V$, $f_v\in Lip(\mathbb{X}_v)$, with $L=\max_{v\in V}\|f_v\|_{Lip(\mathbb{X}_v)}$. Then for the target function, thought of as a compositional function of all the input variables $\mathbf{x}$ to all the nodes in $\mathbf{S}$, we have

$$|f_{v^*}(\mathbf{x})-g_{v^*}(\mathbf{x})|\le c(L,\mathcal{G})\varepsilon. \tag{5.10}$$

Theorem 5.2 allows us to lift Theorem 5.1 to deep networks. In general, we do not know the constituent functions. Also, for any given function and a DAG structure, it may not be possible to devise an algorithm to find the constituent functions uniquely. For example, $(\cos^2 x)^2$ and $(1/4)(1+\cos(2x))^2$ both have the structures $g_1(g_2(x))$ or $f_1(f_2(x))$, both representing the same DAG but with different constituent functions. Thus, even if we may assume that the noise occurs only in the approximation of the target function at the sink node and not in the constituent functions, it seems to be an extremely difficult problem to determine theoretically for any target function what the optimal DAG structure and the input/output for the constituent functions ought to be. Therefore, we have to state our theorem for deep networks only as an existence theorem, in the non-noisy case, not to complicate the notations too much. We assume also that at each node v, the input data is distributed according to the volume measure of $\mathbb{X}_v$.

Theorem 5.3. *Let $\mathcal{G}$ be a DAG as described above, $\{f_v\}$ be a $\mathcal{G}$-function, and we assume that each of the constituent functions $f_v\in W_\gamma(\mathbb{X}_v)\cap Lip(\mathbb{X}_v)$ for some $\gamma>0$, α satisfies (3.10). Let $n\ge 1$. Then there exists a $\mathcal{G}$-function $\{g_v\}$ such that each g_v is a Gaussian network constructed using $\mathcal{O}(n^{q_v(2-\alpha)+2\alpha\gamma}\log n)$ samples of its inputs, such that for any $\mathbf{x}$ seen by v^*,*

$$|f_{v^*}(\mathbf{x})-g_{v^*}(\mathbf{x})|\le c(L,\mathcal{G})n^{-\alpha\gamma}. \tag{5.11}$$

6. Background on weighted polynomials

6.1. Weighted polynomials

A good preliminary source of many identities regarding Hermite polynomials is the book (Szegö, 1975) of Szegö or the Bateman manuscript (Bateman, Erdélyi, Magnus, Oberhettinger, & Tricomi, 1955).

We denote the class of all univariate algebraic polynomials of degree $< n$ by $\mathbb{P}_n$. The orthonormalized Hermite polynomial h_k of degree k is defined recursively by (3.3). With $\psi_k(x)=h_k(x)\exp(-x^2/2)$, one has the orthogonality relation for $k,j\in\mathbb{Z}_+$,

$$\int_{\mathbb{R}}\psi_k(x)\psi_j(x)dx=\begin{cases}1, & \text{if } k=j,\\ 0, & \text{if } k\ne j.\end{cases} \tag{6.1}$$

Using (3.3), it is easy to deduce by induction that

$$\psi_\ell(0)=\begin{cases}\pi^{-1/4}(-1)^{\ell/2}\dfrac{\sqrt{\ell!}}{2^{\ell/2}(\ell/2)!}, & \text{if } \ell \text{ is even},\\ 0, & \text{if } \ell \text{ is odd},\end{cases} \tag{6.2}$$

The Hermite polynomial h_m has m real and simple zeros $x_{k,m}$. Writing

$$\lambda_{k,m} := \left(\sum_{j=0}^{m-1} h_j(x_{k,m})^2\right)^{-1}, \tag{6.3}$$

it is well known (cf. Szegö (1975, Section 3.4)) that

$$\sum_{k=1}^{m} \lambda_{k,m} P(x_{k,m}) = \int_{\mathbb{R}} P(x)\exp(-x^2)dx, \qquad P \in \mathbb{P}_{2m}. \tag{6.4}$$

It is also known (cf. Mhaskar (1996, Theorem 8.2.7), applied with $p = 2$, $b = 0$) that

$$\sum_{k=1}^{m} \lambda_{k,m}\exp(x_{k,m}^2) \le cm^{1/2}. \tag{6.5}$$

The Mehler formula (Andrews, Askey, & Roy, 1999, Formula (6.1.13)) states that

$$\sum_{j=0}^{\infty} \psi_j(y)\psi_j(z)w^j = \frac{1}{\sqrt{\pi(1-w^2)}}\exp\left(\frac{2yzw-(y^2+z^2)w^2}{1-w^2}\right) \times \exp(-(y^2+z^2)/2), \qquad y,z\in\mathbb{R},\ w\in\mathbb{C},\ |w|<1. \tag{6.6}$$

Next, we introduce and review the properties of Hermite polynomials in the multivariate setting. We will need to use spaces with many different dimensions. Therefore, in this section, we will use the symbol d to denote a generic dimension, which will be replaced later by q, Q, q_v, etc.

If $d \ge 2$ is an integer, we define Hermite polynomials on $\mathbb{R}^d$ using tensor products. We adopt the notation $\mathbf{x} = (x_1,\dots,x_d)$. The orthonormalized Hermite function is defined by

$$\psi_{\mathbf{k}}(\mathbf{x}) := \prod_{j=1}^{d} \psi_{k_j}(x_j). \tag{6.7}$$

In general, when univariate notation is used in multivariate context, it is to be understood in the tensor product sense as above; e.g., $\mathbf{k}! = \prod_{j=1}^d (k_j!)$, $\mathbf{x}^{\mathbf{k}} = \prod_{j=1}^d x_j^{k_j}$, etc. The notation $|\cdot|_{p,d}$ will denote the ℓ^p norm on $\mathbb{R}^d$.

For any set $A \subset \mathbb{R}^d$ and $f : A \to \mathbb{R}$, we denote by $C(A)$ the space of all uniformly continuous and bounded functions on A, with the norm $\|f\|_A = \sup_{\mathbf{x}\in A}|f(\mathbf{x})|$. The space $C_0(A)$ is the subspace of all $f \in C(A)$ vanishing at infinity.

We will often use (without mentioning it explicitly) the fact deduced from the univariate bounds proved in Askey and Wainger (1965) that

$$|\psi_{\mathbf{k}}(\mathbf{x})| \le c. \tag{6.8}$$

We will denote by Π_n^d the span of $\{\psi_{\mathbf{k}} : \sqrt{|\mathbf{k}|_{1,d}} < n\}$ and by $\mathbb{P}_n^d$ the space of all algebraic polynomials of total degree $< n$. Thus, if $P \in \Pi_n^d$, then $P(\mathbf{x}) = R(\mathbf{x})\exp(-|\mathbf{x}|_{2,d}^2/2)$ for some $R \in \mathbb{P}_{n^2}^d$. The following proposition lists a few important properties of these spaces (cf. Mhaskar, 1996, 2005, 2017).

Proposition 6.1. *Let $n > 0$, $P \in \Pi_n^d$.*

(a) (Infinite–finite range inequality) *For any $\delta > 0$, there exists $c = c(\delta)$ such that*

$$\|P\|_{\mathbb{R}^d\setminus[-\sqrt{2}n(1+\delta),\sqrt{2}n(1+\delta)]^d} \le c_1 e^{-cn^2}\|P\|_{[-\sqrt{2}n(1+\delta),\sqrt{2}n(1+\delta)]^d} \tag{6.9}$$

(b) (MRS identity) *We have*

$$\|P\|_{\mathbb{R}^d} = \|P\|_{[-\sqrt{2}n,\sqrt{2}n]^d}. \tag{6.10}$$

(c) (Bernstein inequality) *There is a positive constant B depending only on d such that*

$$\left\||\nabla P|_{2,d}\right\|_{\mathbb{R}^d} \le Bn\|P\|_{\mathbb{R}^d}. \tag{6.11}$$

Let $m \ge 1$. For a multi-integer $\mathbf{j}$, $1 \le \mathbf{j} \le m$, we write $\mathbf{x}_{\mathbf{j},m,d} := (x_{j_1,m},\dots,x_{j_d,m})$, and $\lambda_{\mathbf{j},m,d} := \prod_{\ell=1}^d \lambda_{j_\ell,m}$. We observe further that if $P_1, P_2 \in \Pi_m^d$, then $P_1(\mathbf{x})P_2(\mathbf{x}) = R(\mathbf{x})\exp(-|\mathbf{x}|^2)$ for some $R \in \mathbb{P}_{2m^2}^d$. Therefore, (6.4) and (6.5) lead to the following fact, which we formulate as a proposition.

Proposition 6.2. *For $m \ge 1$, we have*

$$\sum_{1\le\mathbf{j}\le m^2} \lambda_{\mathbf{j},m^2,d}\exp(|\mathbf{x}_{\mathbf{j},m^2,d}|_{2,d}^2)P_1(\mathbf{x}_{\mathbf{j},m^2,d})P_2(\mathbf{x}_{\mathbf{j},m^2,d}) = \int_{\mathbb{R}^d} P_1(\mathbf{x})P_2(\mathbf{x})d\mathbf{x}, \qquad P_1,P_2\in\Pi_m^d, \tag{6.12}$$

and

$$\sum_{1\le\mathbf{j}\le m^2} \lambda_{\mathbf{j},m^2,d}\exp(|\mathbf{x}_{\mathbf{j},m^2,d}|_{2,d}^2) \le cm^{d/2}. \tag{6.13}$$

6.2. Applications of Mehler identity

The Mehler identity for multivariate Hermite polynomials is expressed conveniently by writing

$$\mathrm{Proj}_{m,d}(\mathbf{x},\mathbf{y}) := \sum_{|\mathbf{k}|_{1,d}=m} \psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y}). \tag{6.14}$$

Using the univariate Mehler identity (6.6), it is then easy to deduce that for $w \in \mathbb{C}$, $|w| < 1$,

$$\begin{aligned}\sum_{\mathbf{k}\in\mathbb{Z}^d} \psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y})w^{|\mathbf{k}|_{1,d}} &= \sum_{m=0}^{\infty} w^m \mathrm{Proj}_{m,d}(\mathbf{x},\mathbf{y})\\ &= \frac{1}{(\pi(1-w^2))^{d/2}} \times \exp\left(\frac{4w\mathbf{x}\cdot\mathbf{y}-(1+w^2)(|\mathbf{x}|_{2,d}^2+|\mathbf{y}|_{2,d}^2)}{2(1-w^2)}\right)\\ &= \frac{1}{(\pi(1-w^2))^{d/2}} \times \exp\left(-\frac{1+w}{1-w}\frac{|\mathbf{x}-\mathbf{y}|_{2,d}^2}{4}-\frac{1-w}{1+w}\frac{|\mathbf{x}+\mathbf{y}|_{2,d}^2}{4}\right)\\ &= \frac{1}{(\pi(1-w^2))^{d/2}} \times \exp\left(-\frac{1+w^2}{2(1-w^2)}\left|\mathbf{x}-\frac{2w}{1+w^2}\mathbf{y}\right|_{2,d}^2\right) \times \exp\left(-\frac{1-w^2}{2(1+w^2)}|\mathbf{y}|_{2,d}^2\right).\end{aligned} \tag{6.15}$$

We note an identity (6.18) which follows immediately from (6.15) by setting $\mathbf{x} = \mathbf{y} = \mathbf{0}$. For integer d (not necessarily positive), we define the sequence $D_{d;r}$ by

$$D_{d;r} := \begin{cases} \pi^{-d/2}(-1)^{r/2}\dfrac{\Gamma(1-d/2)}{\Gamma(1-d/2-r/2)(r/2)!}, & \text{if } r \text{ is even, } d \le 0,\\ \pi^{-d/2}\dfrac{\Gamma(d/2+r/2)}{\Gamma(d/2)(r/2)!}, & \text{if } r \text{ is even, } d \ge 1,\\ 0, & \text{if } r \text{ is odd.}\end{cases} \tag{6.16}$$

This sequence is chosen so as to satisfy

$$\pi^{-d/2}(1-w^2)^{-d/2} = \sum_{r=0}^{\infty} D_{d;r} w^r, \qquad |w| < 1. \tag{6.17}$$

Using the Mehler identity (6.15), we deduce that for any integer $d \ge 1$

$$\begin{aligned}\sum_{r=0}^{\infty} w^{2r} \sum_{|\mathbf{k}|_{1,d}=2r} |\psi_{\mathbf{k}}(\mathbf{0})|^2 &= (\pi(1-w^2))^{-d/2} \\ &= \pi^{-d/2} \sum_{r=0}^{\infty} \frac{\Gamma(d/2+r)}{\Gamma(d/2)r!} w^{2r} \\ &= \sum_{\ell=0}^{\infty} D_{d;\ell} w^{\ell}. \end{aligned} \tag{6.18}$$

In this section, we point out the invariance and localization properties of certain kernels using the Mehler identity.

6.2.1. Rotation invariance

An interesting consequence of the Mehler identity is that the projection $\mathrm{Proj}_{m,d}$ is invariant under rotations. For $d \ge 2$ and any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we may therefore use an appropriate rotation to write

$$\begin{aligned}\mathrm{Proj}_{m,d}(\mathbf{x},\mathbf{y}) = \sum_{j=0}^{m} \mathrm{Proj}_{j,2}((|\mathbf{x}|_{2,d}, 0), (|\mathbf{y}|_{2,d}\cos\theta, |\mathbf{y}|_{2,d}\sin\theta)) \\ \times \sum_{|\mathbf{k}|_{1,d-2}\le m-j} |\psi_{\mathbf{k}}(\mathbf{0})|^2, \end{aligned} \tag{6.19}$$

where $\cos\theta = \mathbf{x}\cdot\mathbf{y}/(|\mathbf{x}|\,|\mathbf{y}|)$, with obvious modifications if $\mathbf{y} = \mathbf{0}$. Hence, we obtain from (6.19) and (6.18) (used with $d-2$ in place of d),

$$\begin{aligned}\mathrm{Proj}_{m,d}(\mathbf{x},\mathbf{y}) = \sum_{j=0}^{m} \mathrm{Proj}_{j,2}((|\mathbf{x}|_{2,d}, 0), (|\mathbf{y}|_{2,d}\cos\theta, |\mathbf{y}|_{2,d}\sin\theta)) \\ \times D_{d-2;m-j}. \end{aligned} \tag{6.20}$$

In the case when $d = 1$, (6.19) takes the form

$$\mathrm{Proj}_{m,1}(x,y) = \psi_m(|x|)\psi_m(|y|\cos\theta), \tag{6.21}$$

where $\cos\theta = xy/(|x|\,|y|) = \mathrm{sgn}(xy)$, $(\mathrm{sgn}(0) = 0)$.

Let $Q \ge q \ge 1$ be integers. We can extend the definition of $\mathrm{Proj}_{m,q}$ to $\mathbf{x}, \mathbf{y} \in \mathbb{R}^Q$ by

$$\begin{aligned}&\mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y}) \\ &:= \begin{cases} \displaystyle\sum_{j=0}^{m} \mathrm{Proj}_{j,2}((|\mathbf{x}|_{2,Q}, 0), (|\mathbf{y}|_{2,Q}\cos\theta, |\mathbf{y}|_{2,Q}\sin\theta)) D_{q-2;m-j}, & \text{if } q = 2, \\ \psi_m(|\mathbf{x}|_{2,Q})\psi_m(|\mathbf{y}|_{2,Q}\cos\theta), & \text{if } q = 1. \end{cases} \end{aligned} \tag{6.22}$$

The relationship between $\mathrm{Proj}_{m,q,Q}$ and $\mathrm{Proj}_{m,Q}$, both defined on $\mathbb{R}^Q$ is given by the following proposition.

Proposition 6.3. *Let $Q > q \ge 2$ be integers. Let $m \ge 0$, and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^Q$.*

(a) *We have*

$$\begin{aligned}\mathrm{Proj}_{m,Q}(\mathbf{x},\mathbf{y}) = \pi^{(q-Q)/2} \sum_{\ell=0}^{\lfloor m/2\rfloor} \binom{(Q-q)/2+\ell-1}{\ell} \\ \times \mathrm{Proj}_{m-2\ell,q,Q}(\mathbf{x},\mathbf{y}). \end{aligned} \tag{6.23}$$

(b) *We have*

$$\mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y}) = \pi^{(Q-q)/2} \sum_{\ell=0}^{\lfloor m/2\rfloor} (-1)^{\ell} \binom{(Q-q)/2}{\ell} \mathrm{Proj}_{m-2\ell,Q}(\mathbf{x},\mathbf{y}). \tag{6.24}$$

Hence, $\mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y})$ is a weighted polynomial in Π_m^Q as a function of $\mathbf{x}$ and $\mathbf{y}$.

(c) *If $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$, then (6.23) and (6.24) both hold also when $q = 1$.*

Proof. In this proof, let $\mathbf{x}' = (|\mathbf{x}|_{2,Q}, 0, \underbrace{0,\dots,0}_{q-2 \text{ times}})$, $\mathbf{y}' = (|\mathbf{y}|_{2,Q}\cos\theta, |\mathbf{y}|_{2,Q}\sin\theta, \underbrace{0,\dots,0}_{q-2 \text{ times}})$. In view of (6.19), we observe that

$$\mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y}) = \mathrm{Proj}_{m,q}(\mathbf{x}',\mathbf{y}'). \tag{6.25}$$

Further, $|\mathbf{x}-\mathbf{y}|_{2,Q} = |\mathbf{x}'-\mathbf{y}'|_{2,q}$, $|\mathbf{x}+\mathbf{y}|_{2,Q} = |\mathbf{x}'+\mathbf{y}'|_{2,q}$. Therefore, the Mehler identity (6.15) shows that

$$\begin{aligned}\sum_{m=0}^{\infty} w^m \mathrm{Proj}_{m,Q}(\mathbf{x},\mathbf{y}) &= \frac{1}{(\pi(1-w^2))^{Q/2}} \exp\left(-\frac{1+w}{1-w}\frac{|\mathbf{x}-\mathbf{y}|_{2,Q}^2}{4}\right. \\ &\qquad \left. -\frac{1-w}{1+w}\frac{|\mathbf{x}+\mathbf{y}|_{2,Q}^2}{4}\right) \\ &= \frac{1}{(\pi(1-w^2))^{(Q-q)/2}} \sum_{m=0}^{\infty} w^m \mathrm{Proj}_{m,q}(\mathbf{x}',\mathbf{y}') \\ &= \frac{1}{(\pi(1-w^2))^{(Q-q)/2}} \sum_{m=0}^{\infty} w^m \mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y}). \end{aligned} \tag{6.26}$$

We now recall the McClaurin expansion for $(1-w^2)^{-(Q-q)/2}$ (cf. (6.17)), multiply the two power series using the Cauchy–Leibnitz formula, and compare the coefficients to arrive at (6.23). Part (b) is proved similarly by observing that

$$\begin{aligned}\sum_{m=0}^{\infty} w^m \mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y}) = \pi^{(Q-q)/2}(1-w^2)^{(Q-q)/2} \sum_{m=0}^{\infty} w^m \\ \times \mathrm{Proj}_{m,Q}(\mathbf{x},\mathbf{y}). \end{aligned} \tag{6.27}$$

If $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$, then $\sin\theta = 0$, so that $\mathbf{x}' = (|\mathbf{x}|_{2,Q}, \underbrace{0,\dots,0}_{q-1 \text{ times}})$, $\mathbf{y}' = (|\mathbf{y}|_{2,Q}\cos\theta, \underbrace{0,\dots,0}_{q-1 \text{ times}})$. Part (c) is then proved using the same calculations as above. ■

Remark 6.1. Clearly, for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^Q$, $\mathrm{Proj}_{m,Q,Q}(\mathbf{x},\mathbf{y}) = \mathrm{Proj}_{m,Q}(\mathbf{x},\mathbf{y})$, $\mathrm{Proj}_{m,q,Q}(\mathbf{x},\mathbf{y}) = \mathrm{Proj}_{m,q,Q}(-\mathbf{x},-\mathbf{y})$ and the kernel $(\mathbf{x},\mathbf{y}) \mapsto \mathrm{Proj}_{m,q,Q}(\mathbf{0},\mathbf{x}-\mathbf{y})$ is both rotation invariant and translation invariant. Using (6.19), (6.2), and (6.18) (used with $d = q-1$) show that for all $q \ge 1$ and $m \in \mathbb{Z}_+$, $\mathrm{Proj}_{2m-1,q,Q}(\mathbf{0},\mathbf{x}) = 0$, and

$$\begin{aligned}\mathrm{Proj}_{2m,q,Q}(\mathbf{0},\mathbf{x}) &= \mathrm{Proj}_{2m,q}(\mathbf{0},\mathbf{x}') \\ &= \sum_{j=0}^{2m} \psi_j(|\mathbf{x}|_{2,Q})\psi_j(0) \sum_{|\mathbf{k}|_{1,q-1}\le 2m-j} |\psi_{\mathbf{k}}(\mathbf{0})|^2 \\ &= \mathcal{P}_m(|\mathbf{x}|_{2,Q}). \quad \blacksquare \end{aligned} \tag{6.28}$$

6.2.2. Localized kernels

In this section, we recall the localization properties of certain kernels. In the sequel, $H : [0,\infty) \to [0,1]$ is a fixed, infinitely

differentiable function, with $H(t) = 1$ if $0 \le t \le 1/2$, $H(t) = 0$ if $t \ge 1$. All constants may depend upon H as well. We define

$$\Phi_{n,d}(H; \mathbf{x}, \mathbf{y}) := \Phi_{n,d}(\mathbf{x}, \mathbf{y}) := \sum_{\mathbf{k}\in\mathbb{Z}_+^d} H\left(\frac{\sqrt{|\mathbf{k}|_{1,d}}}{n}\right)\psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y})$$
$$= \sum_{m=0}^{n^2} H\left(\frac{\sqrt{m}}{n}\right)\mathrm{Proj}_{m,d}(\mathbf{x}, \mathbf{y}), \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \quad (6.29)$$

Using Mehler identity and the Tauberian theorem in Mhaskar (2018, Theorem 4.3), we proved in Chui and Mhaskar (2018a, Lemma 4.1) the following proposition.

Proposition 6.4. *For $n \ge 1$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have*

$$|\Phi_{n,d}(\mathbf{x}, \mathbf{y})| \le \frac{cn^d}{\max(1, (n|\mathbf{x}-\mathbf{y}|_{2,d})^S)}. \quad (6.30)$$

In particular,

$$|\Phi_{n,d}(\mathbf{x}, \mathbf{y})| \le cn^d, \quad (6.31)$$

and for $1 \le p < \infty$,

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\int_{\mathbb{R}^d} |\Phi_{n,d}(\mathbf{x}, \mathbf{y})|^p d\mathbf{y} \le cn^{d(p-1)}. \quad (6.32)$$

We extend the definition of $\Phi_{n,d}$ as follows. Let $Q \ge q \ge 1$ be integers. We define

$$\Phi_{n,q,Q}(\mathbf{x}, \mathbf{y}) := \sum_{m=0}^{\infty} H\left(\frac{\sqrt{m}}{n}\right)\mathrm{Proj}_{m,q,Q}(\mathbf{x}, \mathbf{y}), \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^Q. \quad (6.33)$$

Remark 6.2. In view of Remark 6.1, the kernel $\widetilde{\Phi}_{n,q}$ defined in (3.6) satisfies $\widetilde{\Phi}_{n,q}(|\mathbf{x}|_{2,Q}) = \Phi_{n,q,Q}(\mathbf{0}, \mathbf{x})$. In particular, $\widetilde{\Phi}_{n,q}(|\circ|_{2,Q}) \in \Pi_n^Q$. ■

Proposition 6.5. *Let $S > Q \ge q \ge 2$ be integers. The kernel $\Phi_{n,q,Q}(\mathbf{x}, \mathbf{y}) \in \Pi_n^Q$ as a function of $\mathbf{x}$ and $\mathbf{y}$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^Q$, $n = 1, 2, \ldots$,*

$$|\Phi_{n,q,Q}(\mathbf{x}, \mathbf{y})| \le \frac{cn^q}{\max(1, (n|\mathbf{x}-\mathbf{y}|_{2,Q})^S)}. \quad (6.34)$$

In particular,

$$|\Phi_{n,q,Q}(\mathbf{x}, \mathbf{y})| \le cn^q. \quad (6.35)$$

If $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$, then

$$|\Phi_{n,1,Q}(\mathbf{x}, \mathbf{y})| \le \frac{cn}{\max(1, (n|\mathbf{x}-\mathbf{y}|_{2,Q})^S)}, \qquad |\Phi_{n,1,Q}(\mathbf{x}, \mathbf{y})| \le cn. \quad (6.36)$$

Proof. Let $\mathbf{x}', \mathbf{y}'$ be as in the proof of Proposition 6.3. Since $\Phi_{n,q,Q}(\mathbf{x}, \mathbf{y}) = \Phi_{n,q}(\mathbf{x}', \mathbf{y}')$, this proposition follows directly from Proposition 6.4. ■

Corollary 6.1. *The kernel $\widetilde{\Phi}_{n,q}$ defined in (3.6) satisfies each of the following properties.*

$$|\widetilde{\Phi}_{n,q}(|\mathbf{x}|_{2,Q})| \le \frac{cn^q}{\max(1, (n|\mathbf{x}|_{2,Q})^S)}, \qquad \mathbf{x} \in \mathbb{R}^q, \quad (6.37)$$

$$|\widetilde{\Phi}_{n,q}(|\mathbf{x}|_{2,Q})| \le cn^q,$$
$$|\widetilde{\Phi}_{n,q}(|\mathbf{x}|_{2,Q}) - \widetilde{\Phi}_{n,q}(|\mathbf{y}|_{2,Q})| \le cn^{q+1}\left||\mathbf{x}|_{2,Q} - |\mathbf{y}|_{2,Q}\right|, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^Q. \quad (6.38)$$

Proof. The estimate (6.37) and the first estimate in (6.38) follow from Proposition 6.5 and the fact that $\widetilde{\Phi}_{n,q}(|\mathbf{x}|_{2,Q}) = \Phi_{n,q,Q}(\mathbf{0}, \mathbf{x})$. The second estimate in (6.38) follows from the Bernstein inequality (6.11) applied with $d = 1$ to the univariate polynomial $\widetilde{\Phi}_{n,q}$. ■

6.3. From Hermite polynomials to Gaussian networks

We discuss in this section the close connection between Hermite polynomials and Gaussian networks.

Proposition 6.6. *Let $m \ge 1$, $\mathbf{k} \in \mathbb{Z}_+^d$, and for $|\mathbf{k}|_{1,d} < m^2$, $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathfrak{G}_{\mathbf{k},m,d}(\mathbf{x}) := \left(\frac{3}{2\pi}\right)^{d/2} 3^{|\mathbf{k}|_{1,d}/2} \sum_{1\le\mathbf{j}\le 2m^2} \lambda_{\mathbf{j},2m^2,d} \times \exp(3|\mathbf{x}_{\mathbf{j},2m^2,d}|_{2,d}^2/4)\psi_{\mathbf{k}}(\mathbf{x}_{\mathbf{j},2m^2,d}) \times \exp\left(-\left|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{x}_{\mathbf{j},2m^2,d}\right|_{2,d}^2\right). \quad (6.39)$$

Then

$$\max_{|\mathbf{k}|_{1,d}<m} \left\|\psi_{\mathbf{k}} - \mathfrak{G}_{\mathbf{k},m,d}\right\|_{\mathbb{R}^d} \le cm^{d-2}3^{-m^2/2}. \quad (6.40)$$

Clearly, the number of neurons in the network $\mathfrak{G}_{\mathbf{k},m,d}$ is $\mathcal{O}(m^{2d})$.

Proof. This proof is the same as that in Chui and Mhaskar (2019, Lemma 4.2) and Mhaskar (2004, Lemma 4.1). Using the last expression in (6.15) with $w = 1/\sqrt{3}$, we obtain

$$\sum_{\mathbf{k}\in\mathbb{Z}_+^d} \psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{u})3^{-|\mathbf{k}|_{1,d}/2} = \left(\frac{3}{2\pi}\right)^{d/2} \exp\left(-\left|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{u}\right|_{2,d}^2\right) \times \exp(-|\mathbf{u}|^2/4).$$

In this proof, we denote by $\nu_{m,d}^*$ the measure that associates the mass $\lambda_{\mathbf{j},2m^2,d}\exp(|\mathbf{x}_{\mathbf{j},2m^2,d}|_{2,d}^2)$ with the point $\mathbf{x}_{\mathbf{j},2m^2,d}$ for $1 \le j_1, \ldots, j_d \le 2m^2$. Therefore, using Proposition 6.2 with $m\sqrt{2}$ in place of m, we obtain

$$\psi_{\mathbf{k}}(\mathbf{x}) = 3^{|\mathbf{k}|_{1,d}/2}\left(\frac{3}{2\pi}\right)^{d/2} \int_{\mathbb{R}^d} \exp\left(-\left|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{u}\right|_{2,d}^2\right)\psi_{\mathbf{k}}(\mathbf{u}) \times \exp(-|\mathbf{u}|^2/4)d\nu_{m,d}^*(\mathbf{u}) - 3^{|\mathbf{k}|_{1,d}/2}\int_{\mathbb{R}^d} \sum_{|\mathbf{j}|_{1,d}\ge 2m^2} \psi_{\mathbf{k}}(\mathbf{u})\psi_{\mathbf{j}}(\mathbf{x}) \times \psi_{\mathbf{j}}(\mathbf{u})3^{-|\mathbf{j}|_{1,d}/2}d\nu_{m,d}^*(\mathbf{u}).$$

The first term on the right hand side above is $\mathfrak{G}_{\mathbf{k},m,d}$. The second term is estimated using (6.8) and (6.13) (applied with $m\sqrt{2}$ in place of m) exactly as in the proof of Chui and Mhaskar (2019, Lemma 4.2). We omit the details. ■

The following corollary is easy to deduce (cf. Mhaskar, 2004, Proposition 4.1). If $P = \sum_{|\mathbf{k}|_{1,d}<m^2} b_{\mathbf{k}}\psi_{\mathbf{k}} \in \Pi_m^d$, we define

$$\mathfrak{G}_d(P) := \sum_{|\mathbf{k}|_{1,d}<m^2} b_{\mathbf{k}}\mathfrak{G}_{\mathbf{k},m,d}. \quad (6.41)$$

Corollary 6.2. *Let $m \ge 1$, $P \in \Pi_m^d$. Then*

$$\|P - \mathfrak{G}_d(P)\|_{\mathbb{R}^d} \le c_1 m^c 3^{-m^2/2}\|P\|_{\mathbb{R}^d}. \quad (6.42)$$

We note that the centers and the number of neurons in the network $\sum_{|\mathbf{k}|_{1,d}<m^2} b_{\mathbf{k}}\mathfrak{G}_{\mathbf{k},m,d}$ *are independent of* P*. In particular, the number of neurons is* $\mathcal{O}(m^{2d})$*.*

6.4. Function approximation

In this section, we describe some results on approximation of functions on $\mathbb{R}^d$. If $f \in C_0(\mathbb{R}^d)$, we define its degree of approximation by

$$E_n(\mathbb{R}^d; f) := \min_{P\in\Pi_n^d} \|f - P\|_{\mathbb{R}^d}. \tag{6.43}$$

For $\gamma > 0$, the smoothness class $W_\gamma(\mathbb{R}^d)$ comprises $f \in C_0(\mathbb{R}^d)$ for which

$$\|f\|_{W_\gamma(\mathbb{R}^d)} := \|f\|_{\mathbb{R}^d} + \sup_{n\ge 0} 2^{n\gamma} E_{2^n}(\mathbb{R}^d; f) < \infty. \tag{6.44}$$

We need some results from Mhaskar (1996, 2003), reformulated in the form stated in Theorem 6.1. To state this theorem, we need some notation first. First, for $\delta \in (0, 1]$, $\mathbf{x} \in \mathbb{R}^d$, $1 \le k \le d$, we write

$$Q'_{k,\delta}(\mathbf{x}) := \min(\delta^{-1}, |x_k|). \tag{6.45}$$

For $t > 0$ and integer $j \ge 0$, the forward difference of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\Delta^j_{k,t}(f)(\mathbf{x}) := \sum_{\ell=0}^{j} (-1)^{j-\ell} \binom{j}{\ell} f(x_1, \ldots, x_{k-1}, x_k + \ell t, x_{k+1}, \ldots, x_d).$$

and for integers $r \ge 1$

$$\omega_r(f, \delta) := \sum_{k=1}^{d} \sum_{j=0}^{r} \delta^{r-j} \sup_{|t|\le\delta} \|(Q'_{k,\delta})^{r-j} \Delta^j_{k,t}(f)\|_{\mathbb{R}^d}. \tag{6.46}$$

Remark 6.3. If $\lambda > 0$, $f_\lambda(\mathbf{x}) = f(\mathbf{x}/\lambda)$, then $\Delta^j_{k,t}(f_\lambda)(\mathbf{x}) = \Delta^j_{k,t/\lambda}(f)(\mathbf{x}/\lambda)$, and $Q'_{k,\delta}(\mathbf{x}) = \lambda Q'_{k,\lambda\delta}(\mathbf{x}/\lambda)$. Using the fact that $\delta \mapsto \delta Q_{k,\delta}(\mathbf{x})$ is non-decreasing for every $\mathbf{x}$, it is not difficult to deduce that

$$\begin{aligned}\omega_r(f_\lambda, \delta) &= \sum_{k=1}^{d} \sum_{j=0}^{r} \delta^{r-j} \sup_{|t|\le\delta} \|(Q'_{k,\delta})^{r-j} \Delta^j_{k,t}(f_\lambda)\|_{\mathbb{R}^d} \\ &= \sum_{k=1}^{d} \sum_{j=0}^{r} (\lambda\delta)^{r-j} \sup_{|u|\le\delta/\lambda} \|(Q'_{k,\lambda\delta})^{r-j} \Delta^j_{k,u}(f)\|_{\mathbb{R}^d} \\ &\le \omega_r(f, \max(\lambda, 1/\lambda)\delta).\end{aligned} \tag{6.47}$$

Theorem 6.1. *Let* $f \in C_0(\mathbb{R}^d)$*,* $r \ge 1$*,* $0 < \gamma < r$*. Then*

(a) *For* $n \ge 1$*,*

$$E_n(\mathbb{R}^d; f) \le c\omega_r(f, 1/n). \tag{6.48}$$

(b) *The function* $f \in W_\gamma(\mathbb{R}^d)$ *if and only if* $\omega_r(f, \delta) = \mathcal{O}(\delta^\gamma)$ *for* $0 < \delta \le 1$*. In fact,*

$$\|f\|_{W_\gamma(\mathbb{R}^d)} \sim \|f\|_{\mathbb{R}^d} + \sup_{0<\delta\le 1} \delta^{-\gamma}\omega_r(f, \delta). \tag{6.49}$$

Proof. The theorem is already contained in the results in Mhaskar (2003), but we need to reconcile notation and explain why. In Mhaskar (2003, Formulas (42),(43)) we have defined a univariate K-functional and a pre-modulus of smoothness for $g(\mathbf{x}) = \exp(|\mathbf{x}|^2_{2,d}/2)f(\mathbf{x})$ applied to the kth component of $\mathbf{x}$, $k = 1, \ldots, d$. The K-functional obtained in this way is denoted in Mhaskar (2003, Formula (21)) by $K_{r,k}$. Likewise, the quantity denoted by ω_r in Mhaskar (2003) is the kth summand of the right hand side of (6.46). Our definition of $Q'_{k,\delta}$ is slightly different from that in Mhaskar (2003) (where it is defined to be $\min(\delta^{-1}, (1 + x_k^2)^{1/2})$). However, our $Q'_{k,\delta}$ as defined in (6.45) satisfies $Q'_{k,\delta} \sim \min(\delta^{-1}, (1 + x_k^2)^{1/2})$, Therefore, Mhaskar (2003, Theorem 5.1, Proposition 4.5) lead to the statement of this theorem. ■

Remark 6.4. If $\gamma = r + \beta$, where $r \ge 0$ is an integer and $0 < \beta \le 1$, $f \in C_0^r(\mathbb{R}^d)$ and satisfies

$$\sup_{|\mathbf{u}|_{2,d}\le\delta} \|f^{(r)}(\circ + \mathbf{u}) - f^{(r)}\|_{\mathbb{R}^d} + \delta \left\|\min(\delta^{-1}, |\circ|_{2,d}) f^{(r)}\right\|_{\mathbb{R}^d} \le c(f)\delta^\beta, \tag{6.50}$$

for every derivative $f^{(r)}$ of order r, then $\omega_r(f, \delta) = \mathcal{O}(\delta^\gamma)$ for $0 < \delta \le 1$, and $f \in W_\gamma(\mathbb{R}^d)$. If $f \in C_0^r(\mathbb{R}^d)$ is compactly supported, and every derivative $f^{(r)}$ of order r satisfies

$$\sup_{|\mathbf{u}|_{2,d}\le\delta} \|f^{(r)}(\circ + \mathbf{u}) - f^{(r)}\|_{\mathbb{R}^d} \le c(f)\delta^\beta,$$

then $f \in W_\gamma(\mathbb{R}^d)$. In particular, if f is compactly supported and satisfies a Lipschitz condition, then $f \in W_1(\mathbb{R}^d)$, and therefore, also $f \in W_\gamma(\mathbb{R}^d)$ for every $\gamma \in (0, 1)$. ■

We define

$$\sigma_n(\mathbb{R}^d; f)(\mathbf{x}) := \int_{\mathbb{R}^q} \Phi_{n,d}(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}, \quad f \in C_0(\mathbb{R}^d), \ n > 0, \ \mathbf{x} \in \mathbb{R}^d. \tag{6.51}$$

The following proposition is routine to prove using Proposition 6.4:

Proposition 6.7. (a) *If* $n > 0$ *and* $P \in \Pi^d_{n/\sqrt{2}}$*, then* $\sigma_n(\mathbb{R}^d; P) = P$*.*

(b) *If* $f \in C_0(\mathbb{R}^d)$*,* $n > 0$*, then*

$$\begin{aligned}&\|\sigma_n(\mathbb{R}^d; f)\|_{\mathbb{R}^d} \le c\|f\|_{\mathbb{R}^d}, \\ &E_n(\mathbb{R}^d; f) \le \|f - \sigma_n(\mathbb{R}^d; f)\|_{\mathbb{R}^d} \le cE_{n/\sqrt{2}}(\mathbb{R}^d; f).\end{aligned} \tag{6.52}$$

7. Approximation on affine spaces

In the sequel, we fix integers $Q \ge q \ge 1$.

Let $\mathbb{Y}$ be a q-dimensional affine subspace of $\mathbb{R}^Q$, passing through a point $\mathbf{x}_0 \in \mathbb{R}^Q$. Then there exists a rotation operator $\mathcal{R}$ on $\mathbb{R}^Q$ depending only on $\mathbb{Y}$ such that any point $\mathbf{x} \in \mathbb{Y}$ can be expressed in the form (with $\mathbf{0}_{Q-q} = (0, \ldots, 0) \in \mathbb{R}^{Q-q}$)

$$\mathbf{x} =: \mathbf{x}_0 + \mathcal{R}(\mathbf{u}, \mathbf{0}_{Q-q}) \qquad \mathbf{u} := \mathbf{u}(\mathbf{x}) := (u_1(\mathbf{x}), \ldots, u_q(\mathbf{x})). \tag{7.1}$$

With an abuse of notation, we will write this as $\mathbf{x} = \mathbf{x}_0 + \mathcal{R}(\mathbf{u})$. In this section only, the function $F : \mathbb{R}^q \to \mathbb{R}$ is defined by

$$F(\mathbf{u}) := f(\mathbf{x}_0 + \mathcal{R}(\mathbf{u})), \tag{7.2}$$

we define

$$E_n(\mathbb{Y}; f) := E_n(\mathbb{R}^q; F). \tag{7.3}$$

Similarly, if $\gamma > 0$, then $f \in W_\gamma(\mathbb{Y})$ if $F \in W_\gamma(\mathbb{R}^q)$; i.e., $f \in W_\gamma(\mathbb{Y})$ if $f \in C_0(\mathbb{Y})$ and

$$\|f\|_{W_\gamma(\mathbb{Y})} := \|F\|_{W_\gamma(\mathbb{R}^q)} < \infty. \tag{7.4}$$

In terms of the points $\mathbf{x} = \mathbf{x}_0 + \mathcal{R}(\mathbf{u}, \mathbf{0}_{Q-q}) \in \mathbb{Y}$, the class of approximants of functions on $\mathbb{Y}$ have the form $\mathbf{x} \mapsto P(\mathbf{x}) \exp(-|\mathbf{x} - \mathbf{x}_0|^2/2)$, where $P \in \mathbb{P}^Q_{n^2}$. If we are interested only in approximation on $\mathbb{Y}$, we may decide to use some standard point, such as the best approximation to $\mathbf{0} \in \mathbb{R}^Q$ from $\mathbb{Y}$. This section

is meant to be preparatory to Section 8 where the results in this section will be used with $\mathbb{Y}$ replaced by the tangent space $\mathbb{T}_{\mathbf{x}_0}(\mathbb{X})$ to a manifold $\mathbb{X}$. With this goal in mind, our definition is more natural. We note that if f is supported on a compact neighborhood of $\mathbf{x}_0$, then F is supported on a compact neighborhood of $\mathbf{0} \in \mathbb{R}^q$. Therefore, for such functions, we may use Theorem 6.1 (and Remark 6.4) with F and get the estimates where the constants do not depend upon $\mathbf{x}_0$, although the space of approximants does.

Our goal in this section is to study the analogue of Proposition 6.7 in the context of approximation on $\mathbb{Y}$.

We denote the volume measure of $\mathbb{Y}$ by $\mu_{\mathbb{Y}}$, and for $f \in C_0(\mathbb{Y})$, $\lambda > 0$, $\mathbf{x} = \mathbf{x}_0 + \mathcal{R}(\mathbf{u})$,

$$\sigma_{n,\lambda}(\mathbb{Y}; f)(\mathbf{x}) := \sigma_{n,\lambda}(\mathbf{x}_0, \mathbb{Y}; f)(\mathbf{x}) := \lambda^q \int_{\mathbb{Y}} \Phi_{n,q,Q}(\lambda(\mathbf{x}-\mathbf{x}_0), \lambda(\mathbf{y}-\mathbf{x}_0)) f(\mathbf{y}) d\mu_{\mathbb{Y}}(\mathbf{y}). \tag{7.5}$$

Theorem 7.1. *Let $Q \ge q \ge 1$ be integers, $\mathbb{Y}$ be a q-dimensional affine subspace of $\mathbb{R}^Q$, passing through $\mathbf{x}_0 \in \mathbb{R}^Q$, $f \in C_0(\mathbb{Y})$, $\lambda > 0$. Then*

$$\left\| \sigma_{n,\lambda}(\mathbb{Y}; f) - f \right\|_{\mathbb{Y}} \le c E_{n/\sqrt{2}}(\mathbb{Y}; f(\mathbf{x}_0 + \mathcal{R}((\circ - \mathbf{x}_0)/\lambda))). \tag{7.6}$$

In particular, if $\gamma > 0$, $f \in W_\gamma(\mathbb{Y})$, $\lambda \ge 1$, then

$$\left\| \sigma_{n,\lambda}(\mathbb{Y}; f) - f \right\|_{\mathbb{Y}} \le c \|f\|_{W_\gamma(\mathbb{Y})} (\lambda/n)^\gamma. \tag{7.7}$$

Here, all the constants are independent of λ.

Proof. Since the kernel $\Phi_{n,q,Q}$ is invariant under rotations, it is easy to verify that for $\mathbf{x} = \mathbf{x}_0 + \mathcal{R}\mathbf{u} \in \mathbb{Y}$,

$$\sigma_{n,\lambda}(\mathbb{Y}; f)(\mathbf{x}) = \int_{\mathbb{R}^q} \Phi_{n,q}(\lambda\mathbf{u}, \mathbf{v}) F(\mathbf{v}/\lambda) d\mathbf{v}.$$

Hence, (7.6) follows from Proposition 6.7. The estimate (7.7) follows from Remark 6.3. ■

8. Proofs of the theorems in Section 3

For any $\mathbf{x} \in \mathbb{X}$, we need to consider in this section three kinds of balls, defined in (3.1):

$B_Q(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{R}^Q : |\mathbf{x}-\mathbf{y}|_{2,Q} \le r\}$, $B_{\mathbb{T}}(\mathbf{x}, r) := \mathbb{T}_{\mathbf{x}}(\mathbb{X}) \cap B_Q(\mathbf{x}, r)$, $\mathbb{B}(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{X} : \rho(\mathbf{x}, \mathbf{y}) \le r\}$.

Clearly, if $r \le \iota^*$, then $\mathbb{B}(\mathbf{x}, r) = \mathcal{E}_{\mathbf{x}}(B_{\mathbb{T}}(\mathbf{x}, r))$.

The following proposition is not difficult to prove using definitions and Taylor expansions (cf. Belkin & Niyogi, 2008). In this section, we will simplify the notation to write $d\mathbf{u}$ in place of $d\mu_{\mathbb{T}_{\mathbf{x}}(\mathbb{X})}(\mathbf{u})$.

Proposition 8.1. *There exists a constant $C^* > 0$ depending only on $\mathbb{X}$ such that each of the following statements holds for every $\mathbf{x} \in \mathbb{X}$.*

(a) *We have*

$$\left| |\mathbf{x} - \mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q} - \rho(\mathbf{x}, \mathcal{E}_x(\mathbf{u})) \right| = \left| |\mathbf{x} - \mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q} - |\mathbf{x}-\mathbf{u}|_{2,Q} \right| \le C^* \rho(\mathbf{x}, \mathcal{E}_x(\mathbf{u}))^3, \quad \mathcal{E}_{\mathbf{x}}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}, \iota^*). \tag{8.1}$$

(b) *If $\delta \le \iota^*$ then*

$$|\mathcal{E}_{\mathbf{x}}(\mathbf{u}) - \mathbf{u}|_{2,Q} \le C^* \delta^2, \qquad \mathcal{E}_{\mathbf{x}}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}, \delta), \tag{8.2}$$

(c) *If $\delta \le \iota^*$ then*

$$\int_{\mathbb{B}(\mathbf{x},\delta)} |d\mu^*(\mathcal{E}_{\mathbf{x}}(\mathbf{u})) - d\mathbf{u}| \le C^* \delta^{q+2}. \tag{8.3}$$

Proof. In this proof only, let $\mathbf{r}$ be any geodesic passing through $\mathbf{x}$, parametrized by the arclength s from $\mathbf{x}$, and g be the metric tensor of $\mathbb{X}$. Then, using the fact that $|\mathbf{r}'(s)|_{2,Q} = 1$, and $\mathbf{r}'(s) \cdot \mathbf{r}''(s) = 0$, it is easy to deduce using Taylor expansions that for $|s| \le \iota^*$,

$$||\mathbf{r}(s) - \mathbf{x}|_{2,Q}^2 - s^2| \le cs^4; \quad \text{i.e., } 1 - \frac{|\mathbf{r}(s)-\mathbf{x}|_{2,Q}^2}{s^2} \le cs^2.$$

Since $1 - |\mathbf{r}(s)-\mathbf{x}|_{2,Q}/s \le 1 - |\mathbf{r}(s)-\mathbf{x}|_{2,Q}^2/s^2$, this proves (8.1). The estimate (8.2) follows from the fact that $\mathbf{r}(s) = \mathcal{E}_{\mathbf{x}}(\mathbf{x} + s\mathbf{r}'(0))$ and a simple estimate using Taylor theorem. The estimate (8.3) follows from the well known fact that in exponential coordinates $\sqrt{\det(g)} = 1 + \mathcal{O}(\delta^2)$ in $\mathbb{B}(\mathbf{x}, \delta)$ if $\delta \le \iota^*$. ■

Corollary 8.1. *There exists $C_1^* > 0$ depending only on $\mathbb{X}$ such that for every $\mathbf{x}, \mathbf{y} \in \mathbb{X}$,*

$$|\mathbf{x}-\mathbf{y}|_{2,Q} \le \rho(\mathbf{x}, \mathbf{y}) \le C_1^* |\mathbf{x}-\mathbf{y}|_{2,Q}. \tag{8.4}$$

In particular, for $r > 0$,

$$\mathbb{B}(\mathbf{x}, r) \subseteq \mathbb{B}_Q(\mathbf{x}, r) \subseteq \mathbb{B}(\mathbf{x}, C_1^* r), \tag{8.5}$$

and (3.9) is equivalent to

$$\sup_{\mathbf{x}\in\mathbb{X}, r>0} \frac{\mu^*(\mathbb{B}_Q(\mathbf{x}, r))}{r^q} \le c. \tag{8.6}$$

Proof. In this proof only, let $a = \min((2C^*)^{-1/2}, \iota^*/2)$. Then for $\rho(\mathbf{x}, \mathbf{y}) \le a$, (8.1) shows that

$$0 \le 1 - \frac{|\mathbf{x}-\mathbf{y}|_{2,Q}}{\rho(\mathbf{x}, \mathbf{y})} \le C^* \rho(\mathbf{x}, \mathbf{y})^2 \le 1/2.$$

Therefore,

$$|\mathbf{x}-\mathbf{y}|_{2,Q} \le \rho(\mathbf{x}, \mathbf{y}) \le 2|\mathbf{x}-\mathbf{y}|_{2,Q}, \qquad \text{if } \rho(\mathbf{x}, \mathbf{y}) \le a. \tag{8.7}$$

In this proof only, let $A = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{X} \times \mathbb{X} : \rho(\mathbf{x}, \mathbf{y}) \ge a\}$. Then A is a compact set and the function $(\mathbf{x}, \mathbf{y}) \mapsto |\mathbf{x}-\mathbf{y}|_{2,Q}/\rho(\mathbf{x}, \mathbf{y})$, being continuous on A, attains its (necessarily positive) minimum. Thus, there exists c such that

$$|\mathbf{x}-\mathbf{y}|_{2,Q} \le \rho(\mathbf{x}, \mathbf{y}) \le c|\mathbf{x}-\mathbf{y}|_{2,Q}, \qquad \text{if } \rho(\mathbf{x}, \mathbf{y}) \ge a.$$

Together with (8.7), this leads to (8.4), and hence to (8.5). ■

To motivate the construction of the operator for approximation, our idea is to transfer the target function locally at each point to the tangent space at that point. Therefore, we use the operator defined as in Section 7. In the present situation, at any point $\mathbf{x}$ at which the approximation is desired, the affine space passes through the point $\mathbf{x}$ itself, which plays the dual role of $\mathbf{x}_0$ in Section 7. While there is only one parameter t in Theorem 2.1, our construction allows us to have two parameters to control localization: the parameter n controlling the degree of the polynomials involved and an additional parameter to control scaling. Recalling that $\Phi_{n,q,Q}(\mathbf{x}, \mathbf{y}) = \Phi_{n,q,Q}(-\mathbf{x}, -\mathbf{y})$ we can define our operator as a convolution as follows.

$$\sigma_{n,\lambda}(\mathbb{X}; f)(\mathbf{x}) := \lambda^q \int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0}, \lambda(\mathbf{y}-\mathbf{x})) f(\mathbf{y}) d\mu^*(\mathbf{y}) = \lambda^q \int_{\mathbb{X}} \widetilde{\Phi}_{n,q,Q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) f(\mathbf{y}) d\mu^*(\mathbf{y}). \tag{8.8}$$

Our first theorem is the analogue of Theorem 7.1 when $\mathbb{X}$ is a manifold instead of an affine space.

Theorem 8.1. *Let $\gamma > 0$, $f \in W_\gamma(\mathbb{X})$, $0 < \alpha \le 1$, $\alpha < 4/(\gamma+2)$. Then for $n \ge 1$, $\lambda = n^{1-\alpha}$,*

$$\|f - \sigma_{n,\lambda}(\mathbb{X}; f)\|_{\mathbb{X}} \le c n^{-\alpha\gamma} \|f\|_{W_\gamma(\mathbb{X})}. \tag{8.9}$$

It is convenient to summarize some details of the proof of this theorem in the form of the following lemma.

Lemma 8.1. *Let $\mathbf{x} \in \mathbb{X}$, $g \in C(\mathbb{X})$ be supported on $\mathbb{B}(\mathbf{x}, \iota^*/8)$, $G(\mathbf{u}) = g(\mathcal{E}_{\mathbf{x}}(\mathbf{u}))$, $\gamma > 0$, $0 < \alpha \le 1$, $\alpha < 4/(\gamma+2)$. Then for $n \ge 1$, $\lambda = n^{1-\alpha}$,*

$$\left| \lambda^q \int_{\mathbb{X}} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) g(\mathbf{y}) d\mu^*(\mathbf{y}) - \lambda^q \int_{\mathbb{T}_{\mathbf{x}}(\mathbb{X})} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q}) G(\mathbf{u}) d\mathbf{u} \right| \le c n^{-\alpha\gamma} \|g\|_{\mathbb{X}}, \tag{8.10}$$

where G is extended outside $\mathbb{B}_{\mathbb{T}}(\mathbf{x}, \iota^*/8)$ as a zero function.

Proof. First, we summarize our choices of various parameters. In this proof only, let

$$\delta = n^{-((2-\alpha)(q+1)+\alpha\gamma)/(q+3)},$$

so that for sufficiently large n,

$$\delta < \min(1, \iota^*/6), \quad n^{q+1}\lambda^{q+1}\delta^{q+3} = n^{-\alpha\gamma}, \quad n\lambda\delta = n^{(4-\alpha\gamma-2\alpha)/(q+3)} \uparrow \infty. \tag{8.11}$$

We choose

$$S \ge \frac{(q(2-\alpha)+\alpha\gamma+1)(q+3)}{4-\alpha\gamma-2\alpha}, \quad (\Rightarrow n^q\lambda^q(n\lambda\delta)^{-S} \le n^{-\alpha\gamma-1}). \tag{8.12}$$

We now assume further that n is large enough so that with C^* as in Proposition 8.1, $C^*\delta^2 \le \delta/2$.

Next, we summarize the implications of our choices on the distances on the manifold, tangent space, and the ambient space. If $\mathbf{y} \in \mathbb{B}(\mathbf{x}, \iota^*/8) \cap B_Q(\mathbf{x}, \delta)$, $\mathbf{u} \in \mathbb{T}_{\mathbf{x}}(\mathbb{X})$, $\mathbf{y} = \mathcal{E}_{\mathbf{x}}(\mathbf{u})$, then (8.2) shows that

$$|\mathbf{x}-\mathbf{u}|_{2,Q} \le |\mathbf{x}-\mathbf{y}|_{2,Q} + |\mathcal{E}_{\mathbf{x}}(\mathbf{u})-\mathbf{u}|_{2,Q} \le \delta + C^*\delta^2 \le (3/2)\delta, \qquad \rho(\mathbf{x},\mathbf{y}) \le 3\delta < \iota^*/2. \tag{8.13}$$

Thus,

$$E_\delta := \mathcal{E}_{\mathbf{x}}^{-1}(\mathbb{B}(\mathbf{x}, \iota^*/8) \cap B_Q(\mathbf{x}, \delta)) \subseteq B_{\mathbb{T}}(\mathbf{x}, 3\delta/2). \tag{8.14}$$

If $\mathbf{u} \in B_{\mathbb{T}}(\mathbf{x}, \iota^*/8)$ then $\mathcal{E}_{\mathbf{x}}(\mathbf{u})$ is well defined. If $\mathbf{u} \in B_{\mathbb{T}}(\mathbf{x}, \iota^*/8) \setminus E_\delta$, then (8.2), (8.1) show that

$$|\mathbf{x}-\mathbf{u}|_{2,Q} \ge |\mathbf{x}-\mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q} - |\mathcal{E}_{\mathbf{x}}(\mathbf{u})-\mathbf{u}|_{2,Q} \ge \delta - C^*\delta^2 \ge \delta/2. \tag{8.15}$$

With this preparation, we are now ready to start with the main estimates. Without loss of generality, we assume that $\|g\|_{\mathbb{X}} = 1$. Since g is supported on $\mathbb{B}(\mathbf{x}, \iota^*/8)$, we find that (cf. (8.12), (6.37))

$$\begin{aligned} &\int_{\mathbb{X}\setminus B_Q(\mathbf{x},\delta)} |\widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) g(\mathbf{y})| d\mu^*(\mathbf{y}) \\ &= \int_{\mathbb{B}(\mathbf{x},\iota^*/8)\setminus B_Q(\mathbf{x},\delta)} |\widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) g(\mathbf{y})| d\mu^*(\mathbf{y}) \\ &\le cn^q(n\lambda\delta)^{-S} \le cn^{-\alpha\gamma-1}\lambda^{-q}. \end{aligned} \tag{8.16}$$

Using (6.38) and (8.1), we deduce that for $\mathbf{y} = \mathcal{E}_{\mathbf{x}}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}, \iota^*/8) \cap B_Q(\mathbf{x}, \delta)$,

$$\begin{aligned} &|\widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q}) - \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q})| \\ &\quad \le cn^{q+1}\lambda \left| |\mathbf{x}-\mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q} - |\mathbf{x}-\mathbf{u}|_{2,Q} \right| \le n^{q+1}\lambda\delta^3. \end{aligned} \tag{8.17}$$

The estimates (8.13) and (3.9) lead further to

$$\left| \int_{\mathbb{B}(\mathbf{x},\iota^*/8)\cap B_Q(\mathbf{x},\delta)} d\mu^*(\mathbf{y}) - \int_{E_\delta} d\mathbf{u} \right| \le \left| \int_{E_\delta} |d\mu^*(\mathcal{E}_{\mathbf{x}}(\mathbf{u})) - d\mathbf{u}| \right| \le c\delta^{q+2}. \tag{8.18}$$

In view of (8.11), (8.17) and (8.18), we deduce that

$$\begin{aligned} &\left| \int_{\mathbb{B}(\mathbf{x},\iota^*/8)\cap B_Q(\mathbf{x},\delta)} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) g(\mathbf{y}) d\mu^*(\mathbf{y}) - \int_{E_\delta} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q}) G(\mathbf{u}) d\mathbf{u} \right| \\ &= \left| \int_{E_\delta} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q}) G(\mathbf{u}) d\mu^*(\mathcal{E}_{\mathbf{x}}(\mathbf{u})) - \int_{E_\delta} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q}) G(\mathbf{u}) d\mathbf{u} \right| \\ &\le \left| \int_{E_\delta} \left(\widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathcal{E}_{\mathbf{x}}(\mathbf{u})|_{2,Q}) - \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q}) \right) G(\mathbf{u}) d\mathbf{u} \right| + cn^q\delta^{q+2} \\ &\le cn^{q+1}\lambda\delta^{q+3} = cn^{-\alpha\gamma}\lambda^{-q}. \end{aligned} \tag{8.19}$$

The localization estimate (6.37) shows (cf. (8.12)) that

$$\left| \int_{\mathbb{T}_{\mathbf{x}}(\mathbb{X})\setminus B_{\mathbb{T}}(\mathbf{x},\iota^*/8)} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q}) G(\mathbf{u}) d\mathbf{u} \right| \le cn^q(n\lambda)^{-S} \le cn^{-\alpha\gamma-1}\lambda^{-q}. \tag{8.20}$$

Invoking the localization estimate (6.37) and (8.11), (8.15) again, we deduce that

$$\left| \int_{B_{\mathbb{T}}(\mathbf{x},\iota^*/8)\setminus E_\delta} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q}) G(\mathbf{u}) d\mathbf{u} \right| \le cn^q(n\lambda\delta)^{-S} \le cn^{-\alpha\gamma-1}\lambda^{-q}. \tag{8.21}$$

The estimates (8.16), (8.19), (8.20) and (8.21) lead to (8.10). ∎

We are now in a position to prove Theorem 8.1.

Proof of Theorem 8.1. Let $\mathbf{x} \in \mathbb{X}$. Let $\phi \in C^\infty(\mathbb{X})$ be chosen so that $\phi(\mathbf{y}) = 1$ if $\mathbf{y} \in \mathbb{B}(\mathbf{x}, \iota^*/16)$, $\phi(\mathbf{y}) = 0$ if $\mathbf{y} \in \mathbb{X} \setminus \mathbb{B}(\mathbf{x}, \iota^*/8)$, and $0 \le \phi(\mathbf{y}) \le 1$ for $\mathbf{y} \in \mathbb{X}$. Then the function $f\phi$ is supported on $\mathbb{B}(\mathbf{x}, \iota^*/8)$, and hence, the function $F : \mathbb{T}_{\mathbf{x}}(\mathbb{X}) \to \mathbb{R}$ defined by $F(\mathbf{u}) := f(\mathcal{E}_{\mathbf{x}}(\mathbf{u}))\phi(\mathcal{E}_{\mathbf{x}}(\mathbf{u}))$ is in $W_\gamma(\mathbb{T}_{\mathbf{x}}(\mathbb{X}))$. Clearly,

$$\|F\|_{\mathbb{T}_{\mathbf{x}}(\mathbb{X})} \le \|f\|_{\mathbb{X}}, \qquad \|F\|_{W_\gamma(\mathbb{T}_{\mathbf{x}}(\mathbb{X}))} \le \|f\|_{W_\gamma(\mathbb{X})}.$$

We choose $S > q + (\alpha\gamma+1)/(2-\alpha)$, and write $a = \iota^*/(16C_1^*)$, where C_1^* is the constant defined in Corollary 8.1. Then, the inclusion (8.5) and the localization property (6.37) show that

$$\begin{aligned} &\left| \int_{\mathbb{X}} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) \, (1-\phi(\mathbf{y})) f(\mathbf{y}) d\mu^*(\mathbf{y}) \right| \\ &\quad = \left| \int_{\mathbb{X}\setminus\mathbb{B}(\mathbf{x},\iota^*/16)} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q})(1-\phi(\mathbf{y})) \times f(\mathbf{y}) d\mu^*(\mathbf{y}) \right| \\ &\quad \le \int_{\mathbb{X}\setminus\mathbb{B}_Q(\mathbf{x},a)} \left| \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q})(1-\phi(\mathbf{y})) \times f(\mathbf{y}) \right| d\mu^*(\mathbf{y}) \\ &\quad \le cn^{q-S}n^{-(1-\alpha)S}\|f\|_{\mathbb{X}} \le cn^{-\alpha\gamma-1}\lambda^{-q}\|f\|_{\mathbb{X}}. \end{aligned} \tag{8.22}$$

In view of Lemma 8.1,

$$\left|\int_{\mathbb{X}} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q})\phi(\mathbf{y})f(\mathbf{y})d\mu^*(\mathbf{y}) - \int_{\mathbb{T}_{\mathbf{x}}(\mathbb{X})} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q})F(\mathbf{u})d\mathbf{u}\right| \le cn^{-\alpha\gamma}\lambda^{-q}\|f\|_{\mathbb{X}}, \tag{8.23}$$

so that

$$\left|\int_{\mathbb{X}} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q})f(\mathbf{y})d\mu^*(\mathbf{y}) - \int_{\mathbb{T}_{\mathbf{x}}(\mathbb{X})} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{u}|_{2,Q})F(\mathbf{u})d\mathbf{u}\right| \le cn^{-\alpha\gamma}\lambda^{-q}\|f\|_{\mathbb{X}}. \tag{8.24}$$

Since $F(\mathbf{x}) = f(\mathbf{x})$, (7.7) in Theorem 7.1 now shows that

$$\left|\lambda^q \int_{\mathbb{X}} \widetilde{\Phi}_{n,q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q})f(\mathbf{y})d\mu^*(\mathbf{y}) - f(\mathbf{x})\right| \le c(n/\lambda)^{-\gamma}\|F\|_{W_\gamma(\mathbb{T}_{\mathbf{x}}(\mathbb{X}))} \le cn^{-\alpha\gamma}\|f\|_{W_\gamma(\mathbb{X})}. \tag{8.25}$$

This proves (8.9). ■

Our next objective in this section is to obtain the following discretization of Theorem 8.1 based on noise-corrupted random samples of f as in Theorem 3.1.

The proof of Theorem 3.1 is included in that of the following theorem, together with Theorem 8.1 applied with $f_0 f$ in place of f.

Theorem 8.2. *We assume the setup as in Theorem 3.1. Then for every $n \ge 1$ and $M \ge n^{q(2-\alpha)+2\alpha\gamma}\log(n/\delta)$ we have with $\lambda = n^{1-\alpha}$,*

$$\mathrm{Prob}_\tau\left(\left\|\widehat{F}_{n,\alpha}(Y;\circ) - \sigma_{n,\lambda}(\mathbb{X};f_0 f)\right\|_{\mathbb{R}^Q} \ge c\sqrt{\|f_0\|_{\mathbb{X}}}\|\mathcal{F}\|_{\mathbb{X}\times\Omega}n^{-\alpha\gamma}\right) \le \delta. \tag{8.26}$$

The proof of Theorem 8.2 requires some preparation. We start with the following concentration inequality (Boucheron, Lugosi, & Massart, 2013, Section 2.7).

Proposition 8.2 (*Bernstein Concentration Inequality*)**.** *Let $Z_1, \dots, Z_M$ be independent real valued random variables such that for each $j = 1, \dots, M$, $|Z_j| \le R$, and $\mathbb{E}(Z_j^2) \le V$. Then for any $t > 0$,*

$$\mathrm{Prob}\left(\left|\frac{1}{M}\sum_{j=1}^M (Z_j - \mathbb{E}(Z_j))\right| \ge t\right) \le 2\exp\left(-\frac{Mt^2}{2(V+Rt/3)}\right). \tag{8.27}$$

In order to apply Proposition 8.2, we need to estimate the second moment of $\mathcal{F}(\mathbf{y},\epsilon)\widetilde{\Phi}_{n,q,Q}(\lambda|\mathbf{x}-\mathbf{y}|_{2,Q}) = \mathcal{F}(\mathbf{y},\epsilon)\Phi_{n,q,Q}(\mathbf{0}, \lambda(\mathbf{x}-\mathbf{y}))$ for every $\mathbf{x} \in \mathbb{R}^Q$. This is done in the following lemma.

Lemma 8.2. *We have*

$$\lambda^{2q}\sup_{\mathbf{x}\in\mathbb{R}^Q}\int_{\mathbb{X}\times\Omega} |\mathcal{F}(\mathbf{y},\epsilon)\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))|^2 d\tau(\mathbf{y},\epsilon) \le c(n\lambda)^q\|\mathcal{F}\|_{\mathbb{X}\times\Omega}^2\|f_0\|_{\mathbb{X}}. \tag{8.28}$$

Proof. Let $\mathbf{x} \in \mathbb{R}^Q$. We need only to estimate

$$\int_{\mathbb{X}\times\Omega} \left|\mathcal{F}(\mathbf{y},\epsilon)\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))\right|^2 d\tau(\mathbf{y},\epsilon) \le \|\mathcal{F}\|_{\mathbb{X}\times\Omega}^2\|f_0\|_{\mathbb{X}}\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))^2 d\mu^*(\mathbf{y}). \tag{8.29}$$

Using Proposition 6.3 and (8.6), and keeping in mind that $\lambda \ge 1$, we deduce that

$$\begin{aligned}
&\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))^2 d\mu^*(\mathbf{y}) \\
&= \int_{\mathbb{X}\cap\mathbb{B}_Q(\mathbf{x},1/(n\lambda))} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))^2 d\mu^*(\mathbf{y}) \\
&+ \sum_{k=0}^\infty \int_{\mathbb{X}\cap(\mathbb{B}_Q(\mathbf{x},2^{k+1}/(n\lambda))\setminus\mathbb{B}_Q(\mathbf{x},2^k/(n\lambda)))} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))^2 d\mu^*(\mathbf{y}) \\
&\le cn^{2q}\left\{\mu^*(\mathbb{B}_Q(\mathbf{x},1/(n\lambda))) + \sum_{k=0}^\infty 2^{-2kS} \times \mu^*(\mathbb{B}_Q(\mathbf{x},2^{k+1}/(n\lambda)) \setminus \mathbb{B}_Q(\mathbf{x},2^k/(n\lambda)))\right\} \\
&\le cn^q\lambda^{-q}\left\{1 + \sum_{k=0}^\infty 2^{-k(2S-q)}\right\} \le cn^q\lambda^{-q}. \quad ■
\end{aligned}$$

The proof of Theorem 8.2 requires an estimation of a quantity of the form

$$\sup_{\mathbf{y}_1,\dots,\mathbf{y}_M\in\mathbb{X}} \left\|\frac{\lambda^q}{M}\sum_{j=1}^M \mathcal{F}(\mathbf{y}_j,\epsilon_j)\Phi_{n,q,Q}(\mathbf{0},\lambda(\circ-\mathbf{y}_j)) - \sigma_{n,\lambda}(\mathbb{X};f_0 f)\right\|_{\mathbb{R}^Q}$$

in terms of the maximum of the function involved at finitely many points. The following lemma accomplishes this by considering the difference between two measures on $\mathbb{X}$: one that associates the mass $(1/M)\mathcal{F}(\mathbf{y}_j,\epsilon_j)$ with each $\mathbf{y}_j$, and other given by $f(\mathbf{y})d\nu^*(\mathbf{y}) = f(\mathbf{y})f_0(\mathbf{y})d\mu^*(\mathbf{y})$. We will denote the total variation of a measure ν by $\|\nu\|_{TV}$. The total variation of the difference between the two measures mentioned above is clearly $\le 2\|\mathcal{F}\|_{\mathbb{X}\times\Omega}$.

Lemma 8.3. *Let $S > Q+2$, λ be as in Theorem 8.1. There exist $c^* = c^*(S) > 0$ and a finite set $\mathcal{D}^* \subset \mathbb{R}^Q$ with $|\mathcal{D}^*| \sim n^{c^*}$ such that for any measure ν on $\mathbb{X}$,*

$$\left\|\lambda^q\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\circ-\mathbf{y}))d\nu(\mathbf{y})\right\|_{\mathbb{R}^Q} \le \max_{\mathbf{x}\in\mathcal{D}^*}\left|\lambda^q\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))d\nu(\mathbf{y})\right| + cn^{-S}\|\nu\|_{TV}. \tag{8.30}$$

Proof. We assume n to be large enough so that $\mathbb{X} \subset [-\sqrt{2}n, \sqrt{2}n]^Q$. Then Proposition 6.5 (used with $2S$ in place of S) shows that

$$\sup_{\mathbf{x}\in\mathbb{R}^Q\setminus[-2n,2n]^Q}\left|\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))d\nu(\mathbf{y})\right| \le cn^{Q-2S}\|\nu\|_{TV} \le cn^{-S}\|\nu\|_{TV}. \tag{8.31}$$

Therefore,

$$\left\|\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\circ-\mathbf{y}))d\nu(\mathbf{y})\right\|_{\mathbb{R}^Q} \le \sup_{\mathbf{x}\in[-2n,2n]^Q}\left|\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))d\nu(\mathbf{y})\right| + cn^{-S}\|\nu\|_{TV}. \tag{8.32}$$

Next, we observe that for any $\mathbf{y} \in \mathbb{R}^Q$

$$\nabla_{\mathbf{x}}\left(\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))\right) = \lambda\left(\nabla_{\mathbf{x}}\Phi_{n,q,Q}(\mathbf{0},\circ)\right)(\lambda(\mathbf{x}-\mathbf{y})).$$

Therefore, for any $\mathbf{x} \in \mathbb{R}^Q$,

$$\left|\nabla_{\mathbf{x}}\left(\int_{\mathbb{X}} \Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))d\nu(\mathbf{y})\right)\right| \le \lambda\int_{\mathbb{X}} \left|\left(\nabla_{\mathbf{x}}\Phi_{n,q,Q}(\mathbf{0},\circ)\right)(\lambda(\mathbf{x}-\mathbf{y}))\right| d|\nu|(\mathbf{y}).$$

Using the Bernstein inequality Proposition 6.1(c), we conclude that

$$\sup_{\mathbf{x}\in\mathbb{R}^Q}\left|\nabla_{\mathbf{x}}\left(\int_{\mathbb{X}}\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))d\nu(\mathbf{y})\right)\right| \le cn^{q+1}\lambda\|\nu\|_{TV} = cn^{q+2-\alpha}\|\nu\|_{TV}.$$

and hence, for any $\mathbf{x},\mathbf{z}\in\mathbb{R}^Q$,

$$\left|\lambda^q\int_{\mathbb{X}}\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))d\nu(\mathbf{y})-\lambda^q\int_{\mathbb{X}}\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{z}-\mathbf{y}))d\nu(\mathbf{y})\right| \le cn^{(q+1)(2-\alpha)}\|\nu\|_{TV}|\mathbf{x}-\mathbf{z}|_{\infty,Q}. \tag{8.33}$$

We now let $\mathcal{D}^*$ be a finite subset of $[-2n,2n]^Q$ such that

$$\max_{\mathbf{x}\in[-2n,2n]^Q}\min_{\mathbf{z}\in\mathcal{D}^*}|\mathbf{x}-\mathbf{z}|_{\infty,Q} \le n^{-(q+1)(2-\alpha)-S}, \tag{8.34}$$

and observe that $|\mathcal{D}^*|\sim n^{Q((q+1)(2-\alpha)+S)}$. The estimate (8.30) is easy to deduce using (8.32), (8.33), and (8.34). ■

With this preparation, we now prove Theorem 8.2, and hence, Theorem 3.1.

Proof of Theorem 8.2 (and Theorem 3.1). Let $\mathbf{x}\in\mathbb{R}^Q$. We consider the random variables

$$Z_j(\mathbf{x}) = \lambda^q\mathcal{F}(\mathbf{y}_j,\epsilon_j)\Phi_{n,q,Q}(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}_j)). \tag{8.35}$$

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

$$\mathbb{E}_\tau(\lambda^q\mathcal{F}(\mathbf{y},\epsilon)\Phi_n(\mathbf{0},\lambda(\mathbf{x}-\mathbf{y}))) = \sigma_{n,\lambda}(\mathbb{X};f_0f)(\mathbf{x}). \tag{8.36}$$

The estimate (6.35) implies that $|Z_j|\le c(n\lambda)^q\|\mathcal{F}\|_{\mathbb{X}\times\Omega}$. Further, Lemma 8.2 yields $\mathbb{E}_\tau(Z_j^2)\le c(n\lambda)^q\|\mathcal{F}\|^2_{\mathbb{X}\times\Omega}\|f_0\|_{\mathbb{X}}$. Therefore, we deduce using Proposition 8.2 that for any $t\in(0,1)$,

$$\mathrm{Prob}_\tau\left(\left|\frac{1}{M}\sum_{j=1}^M Z_j(\mathbf{x})-\sigma_{n,\lambda}(\mathbb{X};f_0f)(\mathbf{x})\right|\ge t\|\mathcal{F}\|_{\mathbb{X}\times\Omega}\|f_0\|_{\mathbb{X}}/2\right) \le 2\exp\left(-c\frac{M\|f_0\|_{\mathbb{X}}t^2}{(n\lambda)^q}\right). \tag{8.37}$$

In view of Lemma 8.3, we have for $S\ge Q+2+\alpha\gamma$,

$$\mathrm{Prob}_\tau\left(\left\|\frac{1}{M}\sum_{j=1}^M Z_j-\sigma_{n,\lambda}(\mathbb{X};f_0f)\right\|_{\mathbb{R}^Q}\ge t\|\mathcal{F}\|_{\mathbb{X}\times\Omega}\|f_0\|_{\mathbb{X}} + c_2n^{-S}\|\mathcal{F}\|_{\mathbb{X}\times\Omega}\right) \le c_1n^{c^*}\exp\left(-c\frac{M\|f_0\|_{\mathbb{X}}t^2}{(n\lambda)^q}\right). \tag{8.38}$$

We recall that $n\lambda=n^{2-\alpha}$ and choose

$$t = c_3\sqrt{\frac{n^{q(2-\alpha)}}{M\|f_0\|_{\mathbb{X}}}\log(n/\delta)}$$

for a suitable constant to make the right hand side of (8.38) to be $\le\delta$, to obtain

$$\mathrm{Prob}_\tau\left(\left\|\frac{1}{M}\sum_{j=1}^M Z_j-\sigma_{n,\lambda}(\mathbb{X};f_0f)\right\|_{\mathbb{R}^Q} \ge c_2\|\mathcal{F}\|_{\mathbb{X}\times\Omega}\left(\sqrt{\frac{n^{q(2-\alpha)}\|f_0\|_{\mathbb{X}}}{M}\log(n/\delta)}+n^{-S}\right)\right)\le\delta. \tag{8.39}$$

We now observe that since $1=\int_{\mathbb{X}}f_0d\mu^*$, and $\mu^*(\mathbb{X})=1$, $\|f_0\|_{\mathbb{X}}\ge 1$. Therefore, choosing $M\ge n^{q(2-\alpha)+2\alpha\gamma}\sqrt{\log(n/\delta)}$, we arrive at (8.26). ■

Theorem 3.2 is obtained immediately from Theorem 3.1 by setting $f_0\equiv 1$. To obtain Theorem 3.3, we use Theorem 3.1 once as stated and again with $\mathcal{F}(Y;\circ)\equiv 1$ to get an approximation to f_0.

9. Proof of the theorems in Section 5

Proof of Theorem 5.1. Theorem 5.1 follows easily from Theorem 8.2 and Corollary 6.2.

Proof of Theorem 5.2. Let $v\in V$, and $u_1,\dots,u_{d(v)}$ be the children of v, and $\mathbf{x}_1,\dots,\mathbf{x}_{d(v)}$ be the inputs seen by these in that order. Let $\mathbf{x}$ be the corresponding input seen by v. Then using the Lipschitz condition on f_v and the property (5.7), we obtain

$$\begin{aligned}
|f_v(\mathbf{x})-g_v(\mathbf{x})| &= \left|f_v\left(\pi_v\left((f_{u_1}(\mathbf{x}_{u_1}),\dots,f_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\right)\right.\\
&\quad\left.-g_v\left(\pi_v\left((g_{u_1}(\mathbf{x}_{u_1}),\dots,g_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\right)\right|\\
&\le \left|f_v\left(\pi_v\left((f_{u_1}(\mathbf{x}_{u_1}),\dots,f_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\right)\right.\\
&\quad\left.-f_v\left(\pi_v\left((g_{u_1}(\mathbf{x}_{u_1}),\dots,g_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\right)\right|\\
&\quad+\left|f_v\left(\pi_v\left((g_{u_1}(\mathbf{x}_{u_1}),\dots,g_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\right)\right.\\
&\quad\left.-g_v\left(\pi_v\left((g_{u_1}(\mathbf{x}_{u_1}),\dots,g_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\right)\right|\\
&\le \|f_v\|_{\mathrm{Lip}(\mathbb{X}_v)}\rho_v\left(\pi_v(f_{u_1}(\mathbf{x}_{u_1}),\dots,f_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}})),\right.\\
&\qquad\left.\pi_v(g_{u_1}(\mathbf{x}_{u_1}),\dots,g_{u_{d(v)}}(\mathbf{x}_{u_{d(v)}}))\right)\\
&\quad+\|f_v-g_v\|_{\mathbb{X}_v}\\
&\le c(v)L\sum_{k=1}^{d(v)}\|f_{u_k}-g_{u_k}\|_{\mathbb{X}_{u_k}}+\|f_v-g_v\|_{\mathbb{X}_v}\le c(L,\mathcal{G})\varepsilon.
\end{aligned} \tag{9.1}$$

We now use induction on the level of v. Thus, if $v^*\in\mathbf{S}$, then the "shallow network" estimate implied in Theorem 5.1 is already the one which we want. Suppose the theorem is proved for the DAGs for which the sink node is at level $\ell\ge 0$. If $v\in V$, so that its level $\ell\ge 1$, then its children are at level $\ell-1\ge 0$. For each of the children, say u, we consider the subgraph $\mathcal{G}_u$ of $\mathcal{G}$ comprising only those nodes and edges that culminate in u as the sink node. We then apply the theorem to each of these subgraphs, and then use (9.1) to conclude that the statement is true for the subgraph $\mathcal{G}_v$ of $\mathcal{G}$ comprising only those nodes and edges that culminate in v as the sink node. ■

Remark 9.1. Suppose we consider a shallow Gaussian network acting on a 2^s dimensional manifold of $\mathbb{R}^Q$. The number of samples required to obtain an accuracy of $n^{-\alpha\gamma}$ predicted by Theorem 5.1 is $\mathcal{O}(n^{2^s(2-\alpha)+2\alpha\gamma}\log n)$. On the other hand, suppose the target function has a compositional structure according to a binary tree, but in addition, for any $v\in V$ with children u_1,u_2, the image of (f_{u_1},f_{u_2}) forms a curve in $\mathbb{R}^2$. Then the number of samples required to get the same accuracy with the corresponding network is only $\mathcal{O}(n^{2-\alpha+2\alpha\gamma}\log n)$ at each level. In fact, it seems likely that this is the number of samples in the original submanifold of $\mathbb{R}^Q$ itself, since the input variables external to the machine are given only at the source nodes. ■

10. Conclusions

We have given a direct solution to the problem of function approximation if the data is sampled from a compact, smooth, connected Riemannian manifold, **without knowing the manifold itself**, except for its dimension. Our construction avoids the evaluation of an eigen-decomposition of a matrix or otherwise the need to compute the local charts on the manifold. Also, the construction avoids any optimization/training in the classical paradigm.

Our construction is universal; i.e., can be used for any target function without any assumption on its prior. The approximation error is estimated in the probabilistic sense, and of course, depends upon the smoothness of the target function. In the case when the data is taken from an affine space, our approximation error does not suffer from any saturation, but can be as small as the smoothness of the target function allows. In the general case, the curvature of the manifold imposes some limitations on how well we can estimate the degree of approximation, but there is no saturation in the sense that if the degree of approximation is better for a function, then it must be "trivial" in some sense.

We have extended our results to the case of deep Gaussian networks. However, in this context, they are not completely constructive unless the constituent functions in the DAG defining the deep network are known.

Declaration of competing interest

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

Appendix. Saturation phenomenon

The notation in this section is not the same as that in the rest of the paper, except that $\|\cdot\|_A$ will denote the supremum norm on a set A. A detailed discussion of saturation phenomena in approximation theory can be found in Butzer and Nessel (2011). Intuitively, an *approximation process* on a metric space A is a sequence of operators $U_n : C(A) \to C(A)$ such that $U_n(f) \to f$ uniformly on A. The process is saturated with the rate $\{\delta_n\}$ if $\|U_n(f) - f\|_A = o(\delta_n)$ as $n \to \infty$ implies that f is *trivial* in some sense (classically $U_n(f) = f$) and there exists a non-trivial function f for which $\|U_n(f) - f\|_A = \mathcal{O}(\delta_n)$. We are unable to find in the literature a precise definition that covers the many applications where this phenomenon holds. As remarked earlier, Theorem 2.1 is one example. We give two other examples.

Example A.1. For $f \in C([-1, 1])$, the Bernstein polynomial is defined by

$$B_n(f)(x) := \sum_{k=0}^{n} \binom{n}{k} f(k/n) x^k (1-x)^{n-k}, \qquad x \in [-1, 1],$$

$n = 0, 1, \ldots$.

The Voronovskaya theorem (Lorentz, 2013, Section 1.6.1) states that if $f \in C^2([-1, 1])$ then uniformly in $x \in [-1, 1]$,

$$\lim_{n\to\infty} \left| n\left(B_n(f)(x) - f(x)\right) - f''(x)\frac{x(1-x)}{2} \right| = 0.$$

Thus, $f \in C^2([-1, 1])$, $\|B_n(f) - f\|_{[-1,1]} = \mathcal{O}(1/n)$ and if $\|B_n(f) - f\|_{[-1,1]} = o(1/n)$ then $f''(x) = 0$ for $x \in (-1, 1)$, so that f is a linear function. ■

Example A.2. A function $S : [-1, 1] \to \mathbb{R}$ is called *piecewise constant* with n break-points if there are points $t_0 = -1 < t_1 < \cdots < t_{n+1} = 1$ such that S is a constant on each (t_j, t_{j+1}), $j = 0, \ldots, n$. We denote the class of all piecewise constants with n break-points by $\mathcal{S}_n$, and define for $f \in C([-1, 1])$,

$$\sigma_n(f) := \inf_{S\in\mathcal{S}_n} \|f - S\|_{[-1,1]}.$$

We note that the break-points of the approximating function may depend upon the target function f. It is known (DeVore & Lorentz, 1993, Chapter 12, Theorem 4.3, Corollary 4.4) that if f has a bounded total variation on $[-1, 1]$ then $\sigma_n(f) = \mathcal{O}(1/n)$. Moreover, if $f \in C([-1, 1])$ and $\sigma_n(f) = o(1/n)$ then f is a constant. ■

List of Symbols

$\mathbb{B}_Q(\mathbf{x}, r)$, $\mathbb{B}_{\mathbb{T}}(\mathbf{x}, r)$, $\mathbb{B}(\mathbf{x}, r)$ Defined in (3.1)
$\Delta^j_{k,\ell}$, $Q'_{k,\delta}$, ω_r Section 6.4
ι^* Inradius of $\mathbb{X}$
λ Scaling factor, typically, $n^{1-\alpha}$
$\lambda_{k,m}$, $\lambda_{\mathbf{k},m}$ Quadrature weights, Section 6.1
$\mathbb{G}^*_{n,q,Q}$, $\mathbb{G}_{n,q,Q}$ Special Gaussian network (5.1), (5.4)
$\mathbb{P}^d_n$, Π^d_n Polynomial spaces Section 6.1
$\mathcal{E}_{\mathbf{x}}$ Exponential map at $\mathbf{x} \in \mathbb{X}$, $\mathcal{E}_{\mathbf{x}} : \mathbb{T}_{\mathbf{x}}(\mathbb{X}) \to \mathbb{X}$
$\mathcal{G}$ DAG for deep networks, Section 5.2
$\mathcal{P}_{m,q}$, $\widetilde{\Phi}_{n,q}$ Univariate polynomials defined in (3.5), (3.6)
$\mathfrak{G}_{\mathbf{k},m,d}$, $\mathfrak{G}_Q$ Basic Gaussian networks (6.39), (6.41)
$\mathsf{Proj}_{m,d}$, $\mathsf{Proj}_{m,q,Q}$ Projection kernels (6.14), (6.22)
$\mathrm{Lip}(\mathbb{X})$ Lipschitz functions on $\mathbb{X}$
μ^* Volume measure on $\mathbb{X}$
$\Phi_{n,d}$, $\Phi_{n,q,Q}$ Localized kernels (6.29), (6.33)
π_v Pooling operation Section 5.2
ρ Metric on $\mathbb{X}$
σ_n, $\sigma_{n,\lambda}$ Approximation operators (6.51), (7.5), (8.8)
τ Probability distribution for the data
$\mathbb{T}_{\mathbf{x}}(\mathbb{X})$ Tangent space to $\mathbb{X}$ at $\mathbf{x}$
$\mathbb{X}$ Manifold
$\mathbb{Y}$ Affine space
d Generic dimension, Section 6
$d(v)$ Ambient dimension at vertex v Section 5.2
$E_n(A; f)$ Degree of approximation of f on A
f, $\mathcal{F}$, $\widehat{F}_{n,\alpha}$ Target function, observations, and estimator
f_0 Density of the marginal distribution
H Low pass filter Section 3.1
h_k, ψ_k, $\psi_{\mathbf{k}}$ Orthonormalized Hermite polynomial, Hermite function, tensor product Hermite function
n, α Parameters in approximation
Q Dimension of the ambient space
q Dimension of affine space or manifold
q_v Dimension in Section 5.2
S Large integer controlling localization
V, $\mathbf{S}$ Non-source, source vertices Section 5.2
v_v Constituent function at v Section 5.2
$W_\gamma(A)$ Smoothness class on A
$x_{k,m}$, $\mathbf{x}_{\mathbf{k},m}$ Quadrature nodes Section 6.1

References

Andrews, G. E., Askey, R., & Roy, R. (1999). *Special functions, Vol. 71*. Cambridge university press.

Askey, R., & Wainger, S. (1965). Mean convergence of expansions in Laguerre and Hermite series. *American Journal of Mathematics*, *87*(3), 695–708.

Bateman, H., Erdélyi, A., Magnus, W., Oberhettinger, F., & Tricomi, F. G. (1955). *Higher transcendental functions, Vol. 2*. New York: McGraw-Hill.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396.

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, *56*(1–3), 209–239.

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

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Butzer, P. L., & Nessel, R. J. (2011). *Fourier analysis and approximation, Vol. 40*. Academic Press.

Chen, M., Jiang, H., Liao, W., & Zhao, T. (2019). Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. arXiv preprint arXiv:1908.01842.

Chui, C. K., & Donoho, D. L. (2006). Special issue: Diffusion maps and wavelets. *Applied and Computational Harmonic Analysis*, *21*(1).
Chui, C. K., & Mhaskar, H. N. (2018a). A Fourier-invariant method for locating point-masses and computing their attributes. *Applied and Computational Harmonic Analysis*, *45*, 436–452.
Chui, C. K., & Mhaskar, H. N. (2018b). Deep nets for local manifold learning. *Frontiers in Applied Mathematics and Statistics*, *4*, 12.
Chui, C. K., & Mhaskar, H. N. (2019). A unified method for super-resolution recovery and real exponential-sum separation. *Applied and Computational Harmonic Analysis*, *46*(2), 431–451.
Cloninger, A., Coifman, R. R., Downing, N., & Krumholz, H. M. (2015). Bigeometric organization of deep nets. arXiv preprint arXiv:1507.00220.
Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *American Mathematical Society. Bulletin*, *39*, 1–49.
Cucker, F., & Zhou, D. X. (2007). *Learning theory: An approximation theory viewpoint, Vol. 24*. Cambridge University Press.
DeVore, R. A., & Lorentz, G. G. (1993). *Constructive approximation, Vol. 303*. Springer Science & Business Media.
do Carmo Valero, M. P. (1992). *Riemannian geometry*. Birkhäuser.
Ehler, M., Filbir, F., & Mhaskar, H. N. (2012). Locally learning biomedical data using diffusion frames. *Journal of Computational Biology*, *19*(11), 1251–1264.
Filbir, F., & Mhaskar, H. N. (2011). Marcinkiewicz–Zygmund measures on manifolds. *Journal of Complexity*, *27*(6), 568–596.
Girosi, F., & Poggio, T. (1990). Networks and the best approximation property. *Biological Cybernetics*, *63*(3), 169–176.
Jones, P. W., Maggioni, M., & Schul, R. (2010). Universal local parametrizations via heat kernels and eigenfunctions of the Laplacian. *Annales Academiæ Scientiarium Fenniæ. Mathematica*, *35*, 131–174.
Lafon, S. S. (2004). *Diffusion maps and geometric harmonics* (Ph.D. thesis), Yale: Yale University.
Liao, W., & Maggioni, M. (2016). Adaptive geometric multiscale approximations for intrinsically low-dimensional data. arXiv preprint arXiv:1611.01179.
Lorentz, G. G. (2013). *Bernstein polynomials*. American Mathematical Soc..
Maggioni, M., & Mhaskar, H. N. (2008). Diffusion polynomial frames on metric measure spaces. *Applied and Computational Harmonic Analysis*, *24*(3), 329–353.
Mhaskar, H. N. (1996). *Introduction to the theory of weighted polynomial approximation, Vol. 56*. Singapore: World Scientific.
Mhaskar, H. N. (2003). On the degree of approximation in multivariate weighted approximation. In *Advanced problems in constructive approximation* (pp. 129–141). Springer.
Mhaskar, H. N. (2004). When is approximation by Gaussian networks necessarily a linear process? *Neural Networks*, *17*(7), 989–1001.
Mhaskar, H. N. (2005). A Markov-Bernstein inequality for Gaussian networks. In *Trends and applications in constructive approximation* (pp. 165–180). Springer.
Mhaskar, H. N. (2010). Eignets for function approximation on manifolds. *Applied and Computational Harmonic Analysis*, *29*(1), 63–87.
Mhaskar, H. N. (2011). A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. *Neural Networks*, *24*(4), 345–359.
Mhaskar, H. N. (2017). Local approximation using Hermite functions. In *Progress in approximation theory and applicable complex analysis* (pp. 341–362). Springer.
Mhaskar, H. N. (2018). A unified framework for harmonic analysis of functions on directed graphs and changing data. *Applied and Computational Harmonic Analysis*, *44*(3), 611–644.
Mhaskar, H. N., & Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, *14*(06), 829–848.
Mhaskar, H. N., & Poggio, T. (2020). An analysis of training and generalization errors in shallow and deep networks. *Neural Networks*, *121*, 229–241.
Schmidt-Hieber, J. (2019). Deep ReLU network approximation of functions on a manifold. arXiv preprint arXiv:1908.00695.
Singer, A. (2006). From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, *21*(1), 128–134.
Szegö, G. (1975). *Colloquium publications: vol. 23, Orthogonal polynomials*. Providence: American mathematical society.