frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 18 August 2020
doi: 10.3389/fams.2020.00031

A Witness Function Based Construction of Discriminative Models Using Hermite Polynomials

Hrushikesh N. Mhaskar[1], *Xiuyuan Cheng*[2] *and Alexander Cloninger*[3]*

[1] Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA, United States, [2] Department of Mathematics, Duke University, Durham, NC, United States, [3] Department of Mathematics and Halicioglu Data Science Institute, University of California, San Diego, San Diego, CA, United States

In machine learning, we are given a dataset of the form $\{(x_j, y_j)\}_{j=1}^M$, drawn as i.i.d. samples from an unknown probability distribution μ; the marginal distribution for the x_j's being μ^*, and the marginals of the k^{th} class $\mu_k^*(x)$ possibly overlapping. We address the problem of detecting, with a high degree of certainty, for which x we have $\mu_k^*(x) > \mu_i^*(x)$ for all $i \neq k$. We propose that rather than using a positive kernel such as the Gaussian for estimation of these measures, using a non-positive kernel that preserves a large number of moments of these measures yields an optimal approximation. We use multi-variate Hermite polynomials for this purpose, and prove optimal and local approximation results in a supremum norm in a probabilistic sense. Together with a permutation test developed with the same kernel, we prove that the kernel estimator serves as a "witness function" in classification problems. Thus, if the value of this estimator at a point x exceeds a certain threshold, then the point is reliably in a certain class. This approach can be used to modify pretrained algorithms, such as neural networks or nonlinear dimension reduction techniques, to identify in-class vs out-of-class regions for the purposes of generative models, classification uncertainty, or finding robust centroids. This fact is demonstrated in a number of real world data sets including MNIST, CIFAR10, Science News documents, and LaLonde data sets.

Keywords: generative model, discriminative model, probability estimation, Hermite functions, witness function

AMS2000 Classification: 68Q32, 94A11, 62-07, 62E17, 42C10

Edited by:
Ding-Xuan Zhou,
City University of Hong Kong,
Hong Kong

Reviewed by:
Danilo Costarelli,
University of Perugia, Italy
Sergiy Pereverzyev Jr.,
Innsbruck Medical University, Austria

****Correspondence:***
Alexander Cloninger
acloninger@ucsd.edu

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

Received: *21 May 2020*
Accepted: *09 July 2020*
Published: *18 August 2020*

Citation:
Mhaskar HN, Cheng X and Cloninger A (2020) A Witness Function Based Construction of Discriminative Models Using Hermite Polynomials. Front. Appl. Math. Stat. 6:31. doi: 10.3389/fams.2020.00031

1. INTRODUCTION

A central problem in machine learning is the following. We are given data of the form $\{(\mathbf{x}_j, y_j)\}_{j=1}^M$, where each $\mathbf{x}_j$ is typically in a Euclidean space and y_j is a label associated with $\mathbf{x}_j$. The data is assumed to be sampled independently from an unknown probability distribution μ. The problem is to learn either the **generative model** μ or a functional model for the unknown function $\mathbf{x} \mapsto \mathbb{E}_\mu(y|\mathbf{x})$. A problem germane to both of these interpretations is to learn the generative model for the $\mathbf{x}$-component; i.e., the marginal distribution μ^* of $\mathbf{x}$.

The problem of finding a functional model is essentially a regression problem, and it is customary to assume smoothness on the target function to get good approximation results. However, it is often used also for studying classification problems, where the labels y_j are limited to a finite set, which may be identified with $\{1, \ldots, m\}$, where m is the number of classes. Therefore, the functional model should also have only m possible values.

We have proposed in earlier papers [1, 2] to approximate the function

$$\mathbf{x} \mapsto \arg\max_{1\le j\le m} \chi_j(\mathbf{x}),$$

where χ_j is 1 if the label associated with $\mathbf{x}$ is j, and 0 otherwise. The functions χ_j are manifestly not continuous. Nevertheless, the smoothness assumptions then hold only away from the class boundaries. Using diffusion geometry [see [3] for an early introduction], we can assume that the feature space is selected judiciously to be a data-defined manifold on which the supports of these functions are well separated; at least, the set of discontinuities of χ_j is not too large. The use of localized kernels developed in Maggioni and Mhaskar [2, 4] on the unknown manifold is then expected to yield a good approximation to the χ_j's and hence, a good classification. A theory for this sort of approximation is well-developed in many papers (e.g., Mhaskar and Prestin [5], Maggioni and Mhaskar [2], Filbir and Mhaskar [6], Mhaskar [7]), illustrated with some toy examples in Maggioni and Mhaskar [2], Ehler et al. [8], and in connection with diabetic blood sugar prediction in a clinical data set in Mhaskar et al. [9]. A bottleneck in this approach is the construction of the right eigensystem to approximate with. Another drawback of this strategy is the lack of out-of-sample extension; i.e., the model depends upon the data set at hand, the entire computation needs to be repeated with the appearance of any new data.

In this paper, we explore another alternative. Corresponding to each of these classes, say the j-th class, one can define a probability distribution $\mu_j^*(\mathbf{x})$ giving the probability that the label j is associated with $\mathbf{x}$. These measures can also be thought of as the **discriminative models** $\mu(j|\mathbf{x})\mu^*(\mathbf{x})$. The classification problem is then equivalent to the approximation of the function,

$$W(\mathbf{x}) = \arg\max_{1\le j\le m} \mu(j|\mathbf{x})\mu^*(\mathbf{x}).$$

Unlike the approach in Maggioni and Mhaskar [2], the approximation of W needs to be done without knowing the values of W even at training data points. Instead of these values, we have labeled samples from the probability measure μ^*. Further, we do not make any assumptions on the structure of the support of the measures μ_j^*. In particular, it is possible to have μ_j^* with overlapping support and differing, or even equivalent, densities.

Our approach to this problem is use a generalized version of the *witness function* between distributions Gretton et al. [10]. To motivate, we consider the problem of one-hot classification, where we need to estimate μ_j^* as if there are two classes: labeled 1 if the actual class label is j and -1 otherwise; i.e., we treat two measures at a time, $\nu_1 = \mu_j^*$ and $\nu_{-1} = \sum_{k\neq j}\mu_j^*$. In this motivation, let us assume that the (closed) supports of these measures are disjoint, and that μ^* has a smooth density f on the feature space $\mathbb{R}^q$. We then consider a smooth function F that takes the value 1 on the support of ν_1, -1 on the support of ν_{-1}, and construct an approximation to Ff using the samples from μ^* and the labels associated with the samples. When this approximation is reliably positive at point $\mathbf{x}$, then $\mathbf{x}$ has the label 1, and if it reliably negative, then it has the label -1. Of course, if the approximation is not reliably positive or negative at a point $\mathbf{x}$, then we cannot classify $\mathbf{x}$ confidently.

The main theoretical goal of this paper is to extend this idea to multiclass classification, and to develop rigorous tests to determine reliability. This can be thought of as investigating where the function

$$\mu(W(\mathbf{x})|\mathbf{x})\mu^*(\mathbf{x}) - \arg\max_{j\neq W(\mathbf{x})} \mu(j|\mathbf{x})\mu^*(\mathbf{x})$$

is significantly greater than zero, given only finite samples of μ.

In general, the question of detecting where two distributions deviate, and whether that deviation is significant, is already of great interest both in machine learning and in various scientific disciplines. The approach of building a kernel based witness function has been developed by Gretton et al. [10]. In these works, the witness function that identifies where two samples deviate comes as a biproduct of determining the maximum mean discrepancy (MMD) between the two distributions and create a measure of global deviation between the distributions. The paper Cheng et al. [11] describes how the power of the global test is affected by changing the definition of locality in the kernel. In the present work, rather than using a data dependent orthogonal system, we use the localized kernels developed in Mhaskar [12], Chui and Mhaskar [13], based on Hermite polynomials. This has several advantages.

1. The kernels are well-defined on the entire Euclidean space and their theoretical properties are well-investigated. Therefore, the use of these kernels obviates the problem of out-of-sample extension. Indeed, we use this kernel to generate labels for new, unseen possible points in the feature space.
2. Although the use of these kernels does involve some tunable parameters, the constructions are more robust with respect to the choice of these parameters, compared with the role of the parameters in the diffusion geometry setting.
3. In view of the Mehler identity, these kernels can be computed efficiently even in high dimensions using only the inner products of the data points (cf. section A).
4. It is shown in Mhaskar [14], Chui and Mhaskar [15] that these kernels can be implemented as Gaussian networks with arbitrary accuracy.

This is important for a number of applications, where it is important to highlight why two distributions deviate, and determine whether specific bumps in the witness function are a product of structure or of noise. Each experiment in section 5 relies on identifying these local deviations.

In section 5.1, we demonstrate experimentally that introducing the localized kernel significantly increases the power of detecting local deviations compared to the Gaussian kernel traditionally used in MMD.

An important application of determining the local deviation between distributions is in accurately determining the confidence of predictions or generated points using neural networks. There have been many reported cases of decisions being made by neural networks for points that lie far away from the training data, and yet are assigned high confidence predictive value Guo et al. [16], Hendrycks and Gimpel [17]. Moreover, it has been recently shown that this is an inherent property of ReLU networks Hein et al. [18]. While there have been a number of attempts to alleviate the issue for predictive nets, there hasn't been nearly as much work for determining out-of-distribution regions for generative networks and Variational Autoenconders, where sampling from out-of-distribution regions leads to non-natural images or blurring between multiple classes of images. We discuss avoiding out-of-distribution and out-of-class sampling of Variational Autoencoders in section 5.2. The use of a witness function for model comparison of GANs is studied in Sutherland et al. [19]. However, that paper only considers the raw size of the witness function without establishing a local baseline, and does not provide a theory of how the empirical witness function deviates from the true witness function. Most approaches to mitigating out-of-distribution high confidence prediction require changing the training objective for the network. We instead examine a pretrained VGG-16 network and determine a method for detecting outliers and likely misclassified points in section 5.3.

Certain scientific applications also benefit from recognition of the where two distributions deviate. The topic of clustering documents based off of a co-occurrence of words has been a topic of significant consideration Shahnaz et al. [20], Wang and Mahadevan [21]. However, most approaches based on k-means in an embedding space can be biased by outliers and clusters with small separation. We examine the uses of the witness function for determining in class vs out-of-training distribution points in a term document embedding in section 5.4 using the localized kernel, which exhibits better edges between close or overlapping clusters. Another application is in propensity matching, in which the problem is to identify bias in which groups received or didn't receive treatment in an observational trial. Propensity matching boils down to modeling the probability of being in one class vs the other, traditionally with a logistic regression, and using such probability for subsequent matching of treated and untreated patients Rosenbaum and Rubin [22]. The uses of propensity matching in literature are too numerous to cite here, but we refer readers to the following article outlining both the importance and its drawbacks King and Nielsen [23]. We instead consider a distance based matching using the nonlinear localized kernel in section 5.5, and demonstrate that viewing this problem as finding local deviations of the witness function allows for the benefits of both an unsupervised distance based algorithm like proposed in King and Nielsen [23] and a simple 1D similar to a propensity score that describes the bias in treatment.

To summarize, we illustrate our theory using the following examples:

1. A toy example of detecting the differences, with low false discovery rate, in support of a measure supported on a circle verse one supported on an ellipse (section 5.1),
2. Discovering the boundaries between classes in the latent space of a variational autoencoder, and generating "prototypical" class examples in the MNIST data set (section 5.2),
3. Prospectively quantifying the uncertainty in classification of the VGG-16 network trained on CIFAR10 (section 5.3),
4. Determining robust cluster centroids in a document-term matrix of Science News documents (section 5.4), and
5. Discovering the propensity of treatment for people in the LaLonde job training data set (section 5.5).

We develop the necessary background and notation for Hermite polynomials and associated kernels and other results from the theory of weighted polynomial approximation in section 6.1. Our main theorems are discussed in section 3, and proved in section 6. The algorithms to implement these theorems are given in section 4, and the results of their application in different experiments are given in section 5.

2. NOTATION

A good preliminary source of many identities regarding Hermite polynomials is the book Szegö [24] of Szegö or the Bateman manuscript Bateman et al. [25]. The univariate Hermite polynomials are defined functionally using the recurrence relations

$$xh_{j-1}(x) = \sqrt{\frac{j}{2}}h_j(x) + \sqrt{\frac{j-1}{2}}h_{j-2}(x), \quad j = 2, 3, \cdots,$$
$$h_0(x) = \pi^{-1/4},\ h_1(x) = \sqrt{2}\pi^{-1/4}x. \tag{2.1}$$

We define the Hermite functions by $\psi_j(x) = h_j(x)\exp(-x^2/2)$, $x \in \mathbb{R}, j \in \mathbb{Z}_+$.

In multivariate case, we adopt the notation $\mathbf{x} = (x_1, \cdots, x_q)$. The orthonormalized Hermite function is defined by

$$\psi_{\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{q} \psi_{k_j}(x_j). \tag{2.2}$$

In general, when univariate notation is used in multivariate context, it is to be understood in the tensor product sense as above; e.g., $\mathbf{k}! = \prod_{j=1}^q k_j!$, $\mathbf{x}^{\mathbf{k}} = \prod_{j=1}^q x_j^{k_j}$, etc. The notation $|\cdot|_p$ will denote the Euclidean ℓ^p norm, with the subscript omitted when $p = 2$.

We define

$$\mathsf{Proj}_m(\mathbf{x}, \mathbf{y}) = \sum_{|\mathbf{k}|_1 = m} \psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y}) \tag{2.3}$$

Let $H : [0, \infty) \to [0, 1]$ be a C^∞ function, $H(t) = 1$ if $t \in [0, 1/2]$, $H(t) = 0$ if $t \geq 1$. We define the *localized kernel* by

$$\Phi_n(H; \mathbf{x}, \mathbf{y}) = \Phi_n(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k} \in \mathbb{Z}_+^q} H\left(\frac{\sqrt{|\mathbf{k}|_1}}{n}\right)\psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y})$$
$$= \sum_{m=0}^{\infty} H\left(\frac{\sqrt{m}}{n}\right)\mathsf{Proj}_m(\mathbf{x}, \mathbf{y}). \tag{2.4}$$

The localization property is made precise in (6.2).

Function space: For $n > 0$ (not necessarily an integer), let $\Pi_n^q = \mathsf{span}\{\psi_{\mathbf{k}} : |\mathbf{k}|_1 < n^2\}$. An element of Π_n^q has the form $P(\mathbf{x})\exp(-|\mathbf{x}|^2/2)$ for a polynomial P of total degree n^2. If $1 \leq p \leq \infty, f \in L^p$,

$$E_{n,p}(f) = \min_{P \in \Pi_n^q} \|f - P\|_p. \tag{2.5}$$

The symbol X^p denotes the set of all $f \in L^p$ for which $E_{n,p}(f) \to 0$ as $n \to \infty$. Thus, $X^p = L^p$ if $1 \leq p < \infty$, and $X^\infty = C_0$. For $\gamma > 0$, the smoothness class $W_{p,\gamma}$ comprises $f \in X^p$ for which

$$\|f\|_{W_{p,\gamma}} = \|f\|_p + \sup_{n \geq 0} 2^{n\gamma} E_{2^n,p}(f) < \infty. \tag{2.6}$$

In Mhaskar [26], Mhaskar [27], we have given a characterization of the spaces $W_{p,\gamma}$ in terms of the constructive properties of f in terms of divided differences and the bounds near ∞.

Next, we describe local smoothness classes. If $r > 0$, $\mathbf{x}_0 \in \mathbb{R}^q$, we denote the ball of radius r around $\mathbf{x}_0$ by

$$\mathbb{B}(\mathbf{x}_0, r) = \{\mathbf{y} \in \mathbb{R}^q : \|\mathbf{x}_0 - \mathbf{y}\|_\infty \leq r\}. \tag{2.7}$$

If $\mathbf{x}_0 \in \mathbb{R}^q$, $\gamma > 0$ the local smoothness class $W_{p,\gamma}(\mathbf{x}_0)$ comprises functions $f \in L^p$ with the following property: There exists a neighborhood U of $\mathbf{x}_0$ such that for every C^∞ function ϕ supported on U, $\phi f \in W_{p,\gamma}$. We note that the quantity γ is expected to depend upon $\mathbf{x}_0$.

Constant convention: In the sequel, $c, c_1, \cdots$ will denote positive constants depending upon q, H, and other fixed quantities in the discussion, such as the norm. Their values may be different at different occurrences, even within a single formula.

3. RECOVERY OF MEASURES

In the two sample problem, one has samples $\mathcal{C}_j$ from distributions μ_j, $j = 1, 2$, and associates the label 1 with C_1, -1 with $\mathcal{C}_2$. The task of a witness function is to determine if in the neighborhood of a given point μ_1 dominates μ_2 or the other way round, or if they are equal. If both the distributions are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^q$ with smooth densities f_1, f_2, respectively, then Theorem 6.1 suggests that $\sigma_n(f_1 - f_2)$, or its Monte-Carlo estimator using the samples should work as a witness function. If the Lebesgue measure were a probability distribution on $\mathbb{R}^q$, then it would be easy to put probabilistic bounds to make this more precise. Since this is not the case, we take the viewpoint that $\mathcal{C}_1 \cup \mathcal{C}_2$ is a sample from a ground probability distribution μ^* with smooth density f, and F is another smooth function that takes approximately the value 1 on $\mathcal{C}_1$, -1 on $\mathcal{C}_j$. Then a candidate for the witness function is given by

$$\sigma_n(Ff)(\mathbf{x}) = \int_{\mathbb{R}^q} F(\mathbf{y}) f(\mathbf{y}) \Phi_n(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^q} F(\mathbf{y}) \Phi_n(\mathbf{x}, \mathbf{y}) d\mu^*(\mathbf{y}).$$

With this re-formulation, we no longer need to restrict ourselves to two classes, F can be chosen to approximate any number of class values, or can even be just any smooth function. The following theorem makes these sentiments more precise in terms of the Monte-Carlo approximation to the last integral above.

Next, we discuss the robustness of this witness function. For this purpose, we assume noisy data of the form $(\mathbf{y}, \epsilon)$, with a joint probability distribution τ and with μ^* being the marginal distribution of $\mathbf{y}$ with respect to τ. In place of $F(\mathbf{y})$, we consider a noisy variant $\mathcal{F}(\mathbf{y}, \epsilon)$, and denote

$$F(\mathbf{y}) = \mathbb{E}_\tau(\mathcal{F}(\mathbf{y}, \epsilon) | \mathbf{y}). \tag{3.1}$$

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

$$\sigma_n(Ff)(\mathbf{x}) = \mathbb{E}_\tau(\mathcal{F}(\mathbf{y}, \epsilon) \Phi_n(\mathbf{x}, \mathbf{y})). \tag{3.2}$$

A part of our theorem below uses the Lambert function defined by

$$\mathcal{W}(ze^z) = z, \quad W(z) > 1 \text{ if } z \geq 1. \tag{3.3}$$

It is known that

$$\mathcal{W}(x) = \log x - \log\log x + o(1), \quad x \to \infty. \tag{3.4}$$

Theorem 3.1. *Let τ be a probability distribution on $\mathbb{R}^q \times \Omega$ for some sample space Ω, μ^* be the marginal distribution of τ restricted to $\mathbb{R}^q$. We assume that μ^* is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^q$ and denote its density by f. Let $\mathcal{F} : \mathbb{R}^q \times \Omega \to \mathbb{R}$ be a bounded function, and F be defined by (3.1). Let $\mathbf{x}_0 \in \mathbb{R}^q$, $\gamma > 0$, $\delta > 0$, $Ff \in W_{\infty,\gamma}(\mathbf{x}_0)$, and r be chosen such that $\|Ff\|_{\infty,\gamma,\mathbf{x}_0,r} < \infty$ [cf. (6.13)]. Let $M \geq 1$, $Y = \{(\mathbf{y}_1, \epsilon_1), \cdots, (y_M, \epsilon_M)\}$ be a set of random samples chosen i.i.d. from τ, and define*

$$\widehat{F}(\mathbf{x}) = \widehat{F}(Y; \mathbf{x}) = \frac{1}{M} \sum_{j=1}^{M} \mathcal{F}(\mathbf{y}_j, \epsilon_j) \Phi_n(\mathbf{x}, \mathbf{y}_j), \qquad \mathbf{x} \in \mathbb{R}^q. \tag{3.5}$$

Then there exists $c_1 > 0$ such that for every $n \geq 1$ and $r \geq c_1/n^2$,

$$\mathsf{Prob}_\tau \left(\left\| \widehat{F}(Y; \circ) - Ff \right\|_{\infty, \mathbb{B}(\mathbf{x}_0, r)} \geq c_2 \|\mathcal{F}\|_\infty n^q \times \sqrt{\frac{\log(c_3 r^q \exp(q/r) n^{5q}/\delta)}{M}} + n^{-\gamma} \|Ff\|_{\infty,\gamma,\mathbf{x}_0,r} \right) \leq \delta (r/n)^q. \tag{3.6}$$

In particular, writing $B = c_3 r^q \exp(q/r)/\delta$, $\Gamma = (2\gamma + 2q)/(5q)$, and

$$n = C_1 B^{-1/(5q)} \exp\left(\frac{1}{2q + 2\gamma} \mathcal{W}\left(\Gamma B^\Gamma M\right) \right) \sim \left(\frac{M}{\log M} \right)^{1/(2q+2\gamma)}, \tag{3.7}$$

we obtain for $r \geq c_1/n^2$ that

$$\mathsf{Prob}_\tau \left(\left\| \widehat{F}(Y; \circ) - Ff \right\|_{\infty, \mathbb{B}(\mathbf{x}_0, r)} \geq c_4 \frac{\|\mathcal{F}\|_\infty + \|Ff\|_{\infty,\gamma,\mathbf{x}_0,r}}{n^\gamma} \right) \leq \delta (r/n)^q. \tag{3.8}$$

Remark 3.1. In the case when $Ff \in C^\infty[\mathbb{B}(\mathbf{x}_0, r)]$, [in particular, when $Ff \equiv 0$ on $\mathbb{B}(\mathbf{x}_0, r)$], one may choose γ to be arbitrarily large, although the constants involved may depend upon the choice of γ. □

Remark 3.2. We note that the critical cube $[-\sqrt{2}n, \sqrt{2}n]^q$ can be partitioned into $\sim (n/r)^q$ subcubes of radius r. On each of these subcubes, say the subcube with center $\mathbf{x}_0$ the function Ff is in a different smoothness class $\gamma(\mathbf{x}_0)$. Therefore, Theorem 3.1 implies an estimate for the entire critical cube with probability at least $1 - \delta$. □

Remark 3.3. Taking $\mathcal{F} \equiv 1, \widehat{F}$ approximates the generative model μ^*. Thus, our construction can be viewed both as an estimator of the generative model as well as a witness function for the discriminative model. □

4. ALGORITHMS

Our numerical experiments in section 5 are designed to illustrate the use of $\widehat{F}$ defined in (3.5) as a discriminative model for two classes, arising with probabilities with densities $f_1 f$ and $f_2 f$ respectively; i.e., with $F = f_1 - f_2$. The quantity $\mathcal{F}$ representing the difference between the corresponding class labels is a noisy version of F. Intuitively, if $\widehat{F}(\mathbf{x})$ is larger than a positive threshold in a neighborhood of $\mathbf{x}_0$, then f_1 dominates f_2 at $\mathbf{x}_0$, and we may take the label for $\mathbf{x}_0$ to be 1, and similarly if $\widehat{F}(\mathbf{x})$ is smaller than a negative threshold. When $|\widehat{F}(\mathbf{x})|$ is smaller than the threshold, then there is uncertainty in the correct label for $\mathbf{x}_0$, whether because it belongs both to the supports of f_1 and f_2 or because $f(\mathbf{x}_0)$ is small, making the point $\mathbf{x}_0$ itself anomalous. In theory, Theorem 3.1 establishes this threshold as $F(\mathbf{x}_0)f(\mathbf{x}_0) - c\dfrac{\|\mathcal{F}\|_\infty + \|Ff\|_{\infty,\gamma,\mathbf{x}_0,r}}{n^\gamma}$. However, it is not possible to estimate this threshold based only on the given data.

For this reason, we introduce the use of the permutation test Pesarin [28]. Permutation test is a parameter-free and nonparametric method of testing hypotheses, namely in this case the null hypothesis that $Ff = 0$ near $\mathbf{x}_0$. Theorem 3.1 shows that if the null hypothesis is true then $|\widehat{F}(\mathbf{x})|$ is smaller than $c\|\mathcal{F}\|_\infty/n^\gamma$ with high probability. In turn, it is easy to create a $\mathcal{F}_1$ for which this is true. We randomly permute the labels of all points $\mathbf{y}_j$ across all classes, reassign these randomized labels $\mathcal{F}_1(\mathbf{y}_j, \epsilon_j)$. Since $\mathcal{F}_1$ represents the class label, this ensures that we know $\|\mathcal{F}_1\|_\infty$ is the same as $\|\mathcal{F}\|_\infty$, but for its expected value F_1, $F_1 f = 0$. Informally, this means we are mixing the distributions of the classes so that when we redraw new collections of points, each collection is being drawn from the same distribution. The benefit of this test in our context is that we can sample this randomized distribution a large number of times, and use that distribution to estimate $cn^{-\gamma}$. This threshold we call $T(\mathbf{x}_j)$, as this is the decision threshold associated to the local area around $\mathbf{x}_0$. If the two classes had equal sampling density around $\mathbf{y}_j$ (i.e. if $Ff = 0$), then if we estimated $T(\mathbf{x}_0)$ correctly, Theorem 3.1 tells us that the probability $\|\widehat{F}\|_{\infty,\mathbb{B}(\mathbf{x}_0,r)}$ exceeds $T(\mathbf{x}_0)$ is small. If, on the other hand, if $\|\widehat{F}\|_{\infty,\mathbb{B}(\mathbf{x}_0,r)} > A \cdot T(\mathbf{x}_0)$ for some constant A associated with $\|Ff\|_{\infty,\gamma,\mathbf{x}_0,r}$ and $\|\mathcal{F}\|_\infty$, then the hypothesis that $Ff = 0$ near $\mathbf{x}_0$ can be rejected.

This suggests two parametric choices in our algorithm, estimating $T(\mathbf{x}_0)$ and A. Estimating $T(\mathbf{x}_0)$ comes down to estimating the threshold that $\|Ff\|_{\infty,\gamma,\mathbf{x}_0,r}(\mathbf{x}_0, r)$ exceeds only $\delta(r/n)^q$ fraction of the time. Because each random permutation is independent, the probability of failure for any permutation to exceed $cn^{-\gamma}$ over N permutations is $(1 - \delta(r/n)^q)^N$. One can choose a desired probability of failure α and compute a number of permutations N. The statistic $T(\mathbf{x}_0)$ in this case would be the maximum size of the local infinity norm across all permutations. If we wish to avoid taking the maximum of N random variables, it is also possible to increase the size of N and take the $1 - \delta(r/n)^q$ quantile of the random variables.

For now, we take A to be a parameter of choice, with the fact that $A \geq 1$, rejection of the $Ff = 0$ assumption scales continuously with A, and A much larger than 1 becomes far too restrictive. Details of the entire algorithm can be found in Algorithm 1 for the two class test, and Algorithm 2 for the multiple class test. This algorithm returns the empirical witness function $\widehat{F}(Z)$, as well as a decision $D(Z)$ as to whether $\widehat{F}(Z)$ is significantly nonzero [i.e., whether $f_1(\mathbf{z}_i) = f_2(\mathbf{z}_i)$ or $f_1(\mathbf{z}_i) \neq f_2(\mathbf{z}_i)$].

For all the experimental sections, we will specify several parameters that are necessary to the algorithm. One parameter is the degree deg of the polynomials used. Note that the parameter n in the kernel Φ_n is given by $n = \sqrt{\mathsf{deg}}$. We also specify A (the tunable parameter to set the level of confidence for determining significant regions), and the scaling factor on the data σ. The scaling factor rescales the data so that $\widetilde{X} = X/\sigma$. One way to consider this scaling is in analogy with the bandwidth of a Gaussian kernel, where $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2) = \exp(-\|\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j\|^2)$; i.e., the variable X is renormalized to have a variance 1. In the context of the witness function, σ no longer represents a variance parameter, but serves the same role as a free parameter.

Algorithm 1: Algorithm for determining significance of empirical witness function using label permutation. $\widehat{F}(\mathbf{z}_j)$ is notation for the empirical witness function at $\mathbf{z}_j$, and $D(\mathbf{z}_j)$ is an indicator for whether $\widehat{F}(\mathbf{z}_j)$ is significantly nonzero.

```
a) Input: Data sets $X$ and $Y$, points $Z = \{\mathbf{z}_1, ..., \mathbf{z}_K\}$ at which to
   inspect significance, level of confidence $A$
b) Output: Estimate of $\widehat{F}(\mathbf{z}_j)$ for all $\mathbf{z}_j \in Z$ and estimate of
   whether $\widehat{F}(\mathbf{z}_j) \neq 0$.
$\alpha \leftarrow 0.05$
$\mathbf{y} \leftarrow X \cup Y$
$M \leftarrow |X| + |Y|$
$c_j \leftarrow \begin{cases} 1, & \text{if } \mathbf{y}_j \in X \\ -1, & \text{if } \mathbf{y}_j \in Y \end{cases}$
$\widehat{F}(\mathbf{z}_j) \leftarrow \frac{1}{M}\sum_{\ell=1}^{M} c_\ell \Phi_n(\mathbf{z}_j, \mathbf{y}_\ell)$
$\rho \leftarrow$ minimal separation between $\mathbf{z}_j$
$p \leftarrow 1 - \alpha(\rho/n)^q$
$N \leftarrow \log(\alpha)/\log(p)$
for $k = 1$ to $N$ do
    $\pi \leftarrow Permutation(M)$
    $F_k(\mathbf{z}_j) \leftarrow \frac{1}{M}\sum_{\ell=1}^{M} c_{\pi(\ell)} \Phi_n(\mathbf{z}_j, \mathbf{y}_\ell)$
end for
$T(\mathbf{z}_j) \leftarrow Percentile(\{F_k(\mathbf{z}_j)\}, p)$
$D(\mathbf{z}_j) \leftarrow \mathbb{1}\left(|\widehat{F}(\mathbf{z}_j)| > A \cdot T(\mathbf{z}_j)\right)$
Return: $\widehat{F}(\mathbf{z}_j), D(\mathbf{z}_j)$
```

Algorithm 2: Algorithm for determining significance of class identifier function using label permutation. Along with returning $\widehat{F}(\mathbf{z}_j)$ and $D(\mathbf{z}_j)$, it also returns L_j, which is the predicted class for $\mathbf{z}_j$.

```
a) Input: Data sets $\{X_i\}_{i=1}^{C}$, points $Z = \{\mathbf{z}_1, ..., \mathbf{z}_K\}$ at which to inspect significance, level of confidence $A$
b) Output: Estimate of $\widehat{F}(\mathbf{z}_j)$ for all $\mathbf{z}_j \in Z$ and estimate of whether the region is dominated by one class
$\alpha \leftarrow 0.05$
$\mathbf{y} \leftarrow \cup_{i=1}^{C} X_i$
$M \leftarrow \sum_{i=1}^{C} |X_i|$
$c_j^{(i)} \leftarrow \begin{cases} 1, & \text{if } \mathbf{y}_j \in X_i \\ 0, & \text{otherwise} \end{cases}$
$\widehat{F}^{(i)}(\mathbf{z}_j) \leftarrow \frac{1}{M} \sum_{\ell=1}^{M} c_\ell^{(i)} \Phi_n(\mathbf{z}_j, \mathbf{y}_\ell)$
$L_j \leftarrow \arg\max_i \widehat{F}^{(i)}(\mathbf{z}_j)$
$\widehat{F}(\mathbf{z}_j) = \widehat{F}^{(L_j)}(\mathbf{z}_j) - \max_{i \neq L} \widehat{F}^{(i)}(\mathbf{z}_j)$
$\rho \leftarrow$ minimal separation between $\mathbf{z}_j$
$p \leftarrow 1 - \alpha(\rho/n)^q$
$N \leftarrow \log(\alpha)/\log(p)$
for $k = 1$ to $N$ do
    $\pi \leftarrow Permutation(M)$
    $\widehat{F}_k^{(i)}(\mathbf{z}_j) \leftarrow \frac{1}{M} \sum_{\ell=1}^{M} c_{\pi(\ell)}^{(i)} \Phi_n(\mathbf{z}_j, \mathbf{y}_\ell)$
    $L \leftarrow \arg\max_i \widehat{F}_k^{(i)}(\mathbf{z}_j)$
    $F_k(\mathbf{z}_j) = \widehat{F}_k^{(L)}(\mathbf{z}_j) - \max_{i \neq L} \widehat{F}_k^{(i)}(\mathbf{z}_j)$
end for
$T(\mathbf{z}_j) \leftarrow Percentile(\{F_k(\mathbf{z}_j)\}, p)$
$D(\mathbf{z}_j) \leftarrow \mathbb{1}\left(|\widehat{F}(\mathbf{z}_j)| > A \cdot T(\mathbf{z}_j)\right)$
Return: $\widehat{F}(\mathbf{z}_j)$, $L_j$, $D(\mathbf{z}_j)$
```

5. EXPERIMENTS

5.1. Toy Examples

We begin the set of experiments with a toy model to demonstrate the benefits of determining the significant regions for the witness function, as well as the benefits of using the localized kernel versus using the Gaussian kernel. We generate two data sets. The first is of the form $(\cos t + \epsilon_1, \sin t + \epsilon_2)$, where t is distributed uniformly on $[0, 2\pi)$ and ϵ_1, ϵ_2 are normal random variables with mean 0 and standard deviation 0.01. The second data set is of the form $[(1 + \varrho)\cos t + \epsilon_3, (1 - \varrho)\sin t + \epsilon_4]$, where t is distributed uniformly on $[0, 2\pi)$ and ϵ_3, ϵ_4 are normal random variables with mean 0 and standard deviation 0.01. See **Figure 1** for visualizations of the data sets with various ϱ. We make random draws of 1,000 points of the first form and 1,000 points of the second form for various sizes of ϱ, and use Algorithm 1 to determine the regions of significant deviation. For this data, we set $\texttt{deg} = 32$, $A = 2$, and $\sigma = 0.5$.

Figure 1 shows the significant areas for varying radii, where the witness function is measured at random points in a ring surrounding the two distributions. Not only does the localized kernel begin to detect significant regions earlier than the Gaussian kernel, the localized kernel is also the only kernel to detect the shorter radii of the ellipse. Also, observe the gap of significance around the points of interaction, in which both distributions are locally similar.

5.2. Data Generation Through VAE

The next set of experiments revolve around estimating regions of space corresponding to given classes, and sampling new points from that given region. This problem has been of great interest in recent years with the growth of generative networks, namely various variants of generative adversarial networks (GANs) Goodfellow et al. [29] and variational autoencoders (VAEs) Kingma and Welling [30]. Each has a low-dimensional latent space in which new points are sampled, and mapped to $\mathbb{R}^q$ through a neural network. While GANs have been more popular in literature in recent years, we focus on VAEs in this paper because it is possible to know the locations of training points in the latent space. A good tutorial on VAEs can be found in Doersch [31].

Our first example is the well-known MNIST data set LeCun et al. [32]. This is a set of handwritten digits $0 \cdots 9$, each scanned as a 28×28 pixel image. There are 50,000 images in the training data set, and 10,000 in the test data.

In order to select the "right" features for this data set, we construct a three layer VAE with encoder $E(\mathbf{x})$ with architecture $784 - 500 - 500 - 2$ and a decoder/generator $G(\mathbf{z})$ with architecture $2 - 500 - 500 - 784$, and for clarity consider the latent space to be the 2D middle layer. In the 2D latent space, we construct a uniform grid on $[-5, 5]^2$. Each of these points can be mapped to the image space via $G(\mathbf{z})$, but there is no guarantee that the reconstructed image appears "real" and no a priori knowledge of which digit will be generated. We display the resulting images $G(\mathbf{z})$ in **Figure 2**, with each digit location corresponding to the location $\mathbf{z}$ in the 2D latent space. However, we also have additional information about the latent space, namely the positions of each of the training points and their associated classes. In other words, we know $\mathbf{z} = E(\mathbf{x})$ for all training data $\mathbf{x}$. The embedding of the training data $E(\mathbf{x})$ is displayed in **Figure 2** as well as the resulting images $G(\mathbf{z})$ for each $\mathbf{z}$ in the 5×5 grid. As one can see, certain regions are dominated by a single digit, other regions contain mixtures of multiple digits, and still other regions are completely devoid of training points (meaning one should avoid sampling from those regions entirely).

We use Algorithm 2 to determine the "significant region" in the embedding space of each class. In other words, we run Algorithm 2 on $\{E(\mathbf{x}_i)\}_{\mathbf{x}_i \in X}$. For this data, we set $\texttt{deg} = 128$, $A = 2$, and $\sigma = 1$. In **Figure 2**, we display the resulting decoded images $D(\mathbf{x})$ for points $\mathbf{z}_i \in \mathbb{R}^2$ deemed significant by the localized kernel. Note that most of the clearly fake images from **Figure 2** have been removed as non-significant by the witness function. **Figure 2** also computes the same notion of significance with the Gaussian kernel of the same scaling, which clearly has less ability to differentiate boundaries. We can see in **Figure 2** that the Gaussian kernel struggles to differentiate boundaries of classes, and keeps a number of points at the tails of each class distribution. These points are exactly the points that are poorly reconstructed by the model, as the

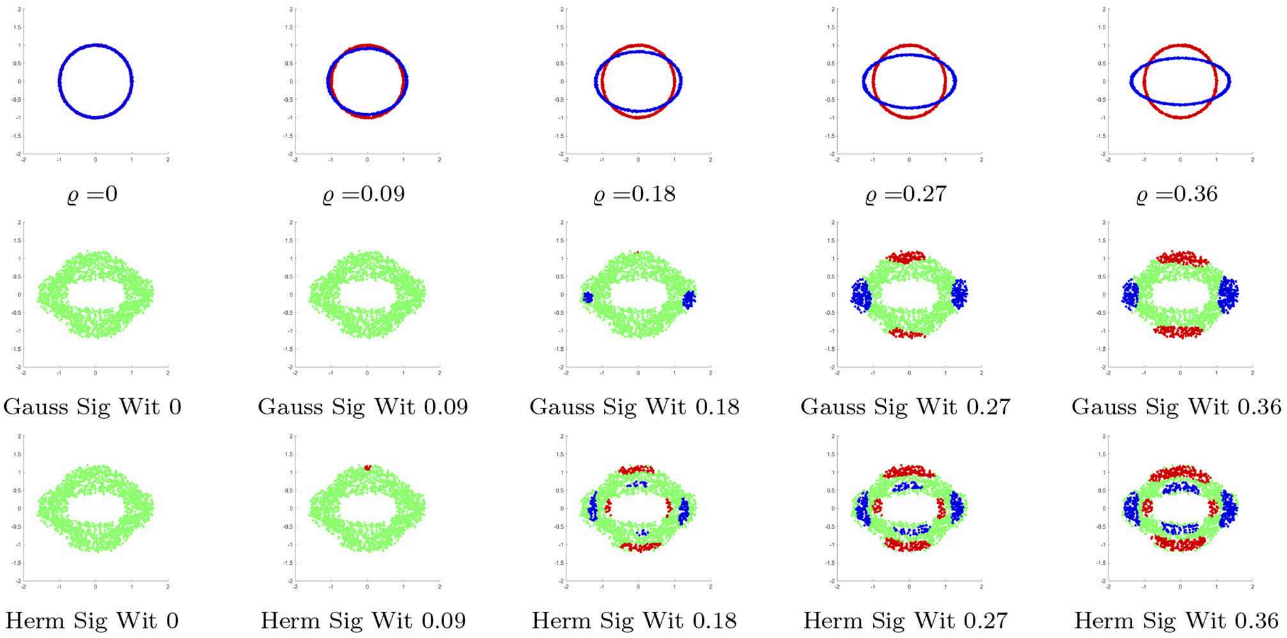


FIGURE 1 | (Top) Examples of the two data sets with varying ϱ, the difference between the principle axis lengths. **(Middle)** Points deemed significant by the Gaussian witness function for given ϱ. **(Bottom)** Points deemed significant by the Hermite witness function for given ϱ. In all figures, green corresponds to 0, meaning neither distribution dominates.

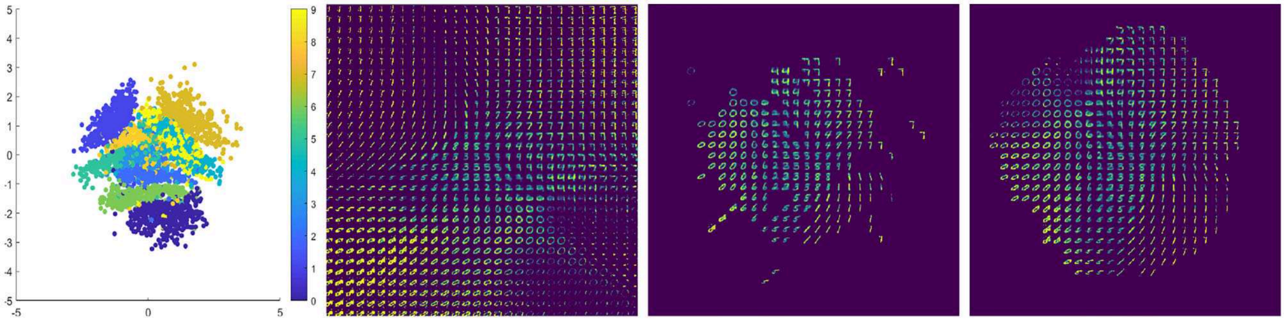
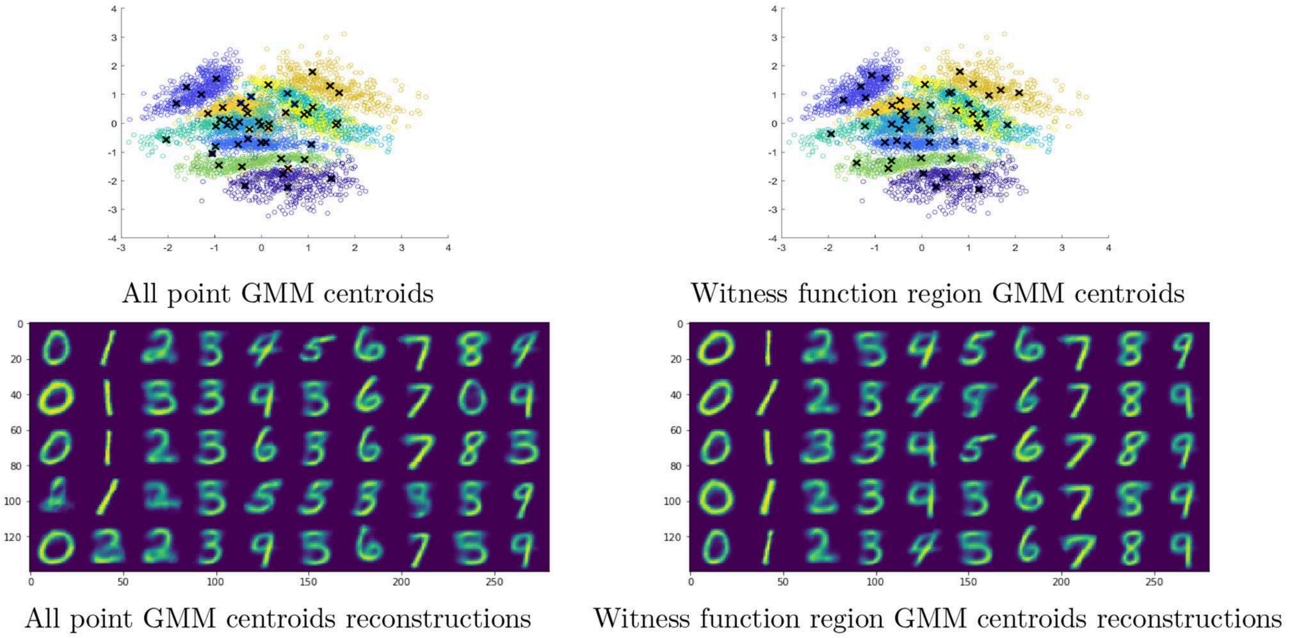


FIGURE 2 | Left to right: (1) Embedding of training data into 2D VAE latent space. (2) Reconstructed images from grid sampling in 2D VAE latent space. (3) Reconstructed images only of grid points that are deemed "significant" by the witness function with the localized kernel (4) Reconstructed images only of grid points that are deemed "significant" by the witness function with the Gaussian kernel.

decoder net hasn't seen a sufficient number of points from the tail regions.

We can also use the significance regions to define "prototypical points" from a given class. We do this by fitting the data from each class with a Gaussian mixture model with five clusters. The means and covariance matrices of each Gaussian are computed through the standard expectation maximization algorithm Reynolds [33]. **Figure 3** shows the centroid values in the embedding of the training data, computed in two different ways. The first approach is to build the mixture model on all points in a given class. In other words, we build a Gaussian mixture model with five clusters on the data $\{E(\mathbf{x}_i) : z_i = j\}$ for each of $j \in \{0, 1, ..., 9\}$. The second approach is to build the mixture model for each class using only those points that are deemed significant by the witness function test. In other words, a mixture model on the restricted dataset $\{E(\mathbf{x}_i) : z_i = j \text{ and } D(\mathbf{x}_i) = 1\}$ for each of $j \in \{0, 1, ..., 9\}$. Due to the structure of the two-dimensional embedding, some of the mixtures for entire classes are pulled toward the origin by a few outliers from the class that are spread across the entire space. This causes overlap between the centroids of the classes considered more difficult to separate in a 2D embedding (4's, 6;'s, 9's), and causes problems in the reconstruction. The centroids of the mixtures for significant regions, on the other hand, have a tendency to remain squarely within neighborhoods of the same class, and their reconstructions $G(\mathbf{z}_i)$ are much clearer.


All point GMM centroids

Witness function region GMM centroids

All point GMM centroids reconstructions

Witness function region GMM centroids reconstructions

FIGURE 3 | Prototypical points from each class of MNIST digits, computed from 2D VAE embedding.

5.3. Determining Significant Class Regions for CNNs

In this section, we consider learning regions of uncertainty in hidden layers of a convolutional neural network. Assessing the uncertainty of classification is an important aspect of machine learning, as certain testing points that are far away from training data, or on the boundary of several classes, should not receive a classification. Even at the last hidden layer of the neural network, there exist points that fall into uncertain regions or boundaries between regions. Our goal is to examine this last hidden layer and prospectively remove uncertain points. In doing so, the goal would be to reduce the set of testing points *without a priori knowledge* of ground truth on the testing data in a way that reduces to final classification error on the test set.

We use the last layer of the VGG-16 pretrained CNN that has been rescaled to the CIFAR10 data set, where the last layer contains $q = 512$ dimensions. The CIFAR10 data set is a collection of 60,000 32 × 32 color images in 10 classes (airplane, bird, cat, deer, etc.), with 6,000 images per class. There are 50,000 training images and 10,000 test images Krizhevsky et al. [34]. VGG-16 is a well known neural network trained on a large set of images called Imagenet, which is rescaled to apply to CIFAR10. VGG-16 has 12 hidden layers, and the architecture and trained weights can be easily downloaded and used Simonyan and Zisserman [35]. VGG-16 attains a prediction error of $\sim 6\%$ on the testing data. Our goal is to detect the fraction of images that are going to be misclassified prior to getting their classification, and thus reduce the prediction error on the remaining images.

We create the witness function on the testing data embedded into this final hidden layer, and determine the threshold by permuting the known labels of the training data. For this data, we set $\mathsf{deg} = 128$ and $\sigma = 7$ (σ was chosen as the median distance to the 100^{th} nearest neighbor in the training data). The parameter A is not set in this section as we are varying it across the data. **Figure 4** shows the decay of the classification error as a function of A, which has a direct correspondence to the overall probability of error. While there is a reduction of the overall number of testing points, the set of points that remain have a smaller classification error than the overall test set. We also show on this reduced set that the label attained by the final layer of the CNN and the estimated label attained by taking the maximum estimated measure across all 10 classes are virtually identical.

Figure 4 also compares the decay of the classification error to the uncertainty in classification as defined by the last layer of the CNN. A CNN outputs a vector, which we call $g(\mathbf{x})$, of the probability a point lies in each of the classes. A notion of uncertainty in the network can be points that have the smallest gap between the prediction of the most likely class $L = \arg\max_i g_i(\mathbf{x})$ and the second highest classification score, which yields an certainty score $g_L(\mathbf{x}) - \max_{j \neq L} g_j(\mathbf{x})$. We sort the testing points by this gap, and remove the first k points with the smallest gap. The larger this gap is, the more certain the CNN should be about the correct classification. As shown in **Figure 4**, our witness function method yields a quicker decay in the classification error as a function of the number of points removed.

A benefit of our approach to quantifying uncertainty is that it explicitly demonstrates that the points classified poorly are those that sit at the boundary between class clusters in the last hidden layer of the network. This means that even at the last layer of the network, misclassified points are still considered "outliers" by the class distributions to which they lie closest.

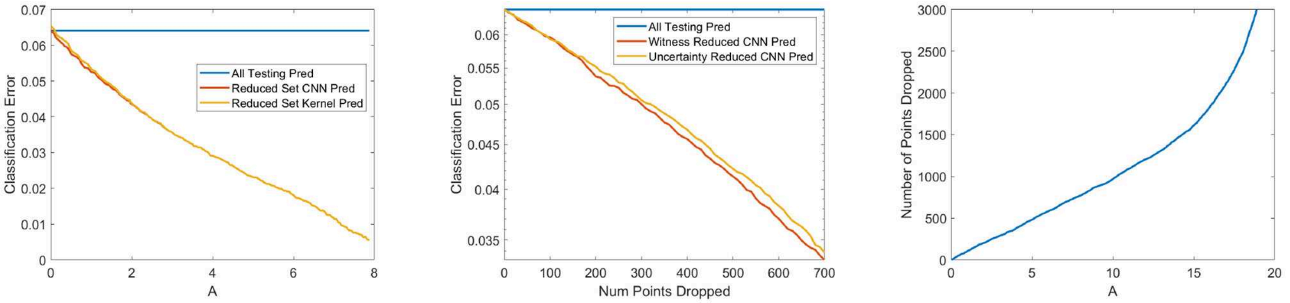


FIGURE 4 | (Left) Classification error on points deemed "significant" as a function of *A*. **(Middle)** Classification error as a function of the number of points removed for being "uncertain," for both the witness function and the size of gap of CNN outputs. These two measures are placed on the same scale by considering the number of points removed. **(Right)** The relationship between the number of points dropped and parameter *A* in our algorithm.

5.4. Term Document Significance

In this section, we consider term-document organization and characterizing types of words that are indicitive of a class of documents. We use the Science News dataset, which consists of 1,046 articles across 8 categories, and their use of 1,153 popular words that appear in the magazine. The categories are: Anthropology, Space, Behavioral Psych, Environmental Science, Life Science, Math/CS, Medicine, and Physics/Tech. There are obvious overlaps of these categories, so not every article in a category will necessarily contain "indicative words" of that category alone.

We begin by taking the top three principle components of the words, and generating a hierarchical topic modeling by splitting the words into four levels of 4, 8, 16, and 32 clusters, respectively. Each document at a given clustering level is encoded by the fraction of its words that fall into a given cluster, and the new features of a document become these histograms across all four levels of word splits. From here, we take the top three principle components of the documents to create a low-dimensional embedding of all documents. This embedding is displayed in **Figure 5**. We then run Algorithm 2 to determine the significant regions for each class. For this data, we set $\mathsf{deg} = 32$, $\sigma = 1$, and $A = 2$.

We sample the embedding at 10,000 random grid points in the embedding space, and display the significant region in **Figure 5**. It is important to note the meaning of these regions of significance, namely that rejection of the null hypothesis at a given point indicates that the concentration of points from one class in that region is well beyond any concentration that would occur due to chance.

In an effort to quantify the significant regions and their benefits, we designed the following simple experiment. For every point, we compare its class to the class of its nearest neighbor, and we record the average classification score across all classes. Namely, for data X and corresponding labels Y, we compute

$$\mathbb{E}_{\mathbf{x}_i \in X}[\delta(y_i, y_{\{\text{NN of } \mathbf{x}_i \text{ in } X\}})],$$
$$\text{and} \quad \mathbb{E}_{\mathbf{x}_i \in X}[\delta(y_i, y_{\{\text{NN of } \mathbf{x}_i \text{ in centroids from } X\}})],$$

where $\delta(\cdot, \cdot)$ is a dirac delta function. We run this experiment for X being all documents, and for X being "significant" documents as deemed by the witness function. The results are in **Table 1**. Clearly, restricting ourselves to the documents deemed significantly within one class greatly increases the reliability of the neighborhood and the computed centroids of the classes.

5.5. Propensity Matching and Non-experiment Sampling

As a final example, we consider the problem of propensity matching and scoring. In this setting, we consider two sets of observable (e.g., non-randomized) data in which one set was given a treatment and the other was not (which serves as a control set). There exists questions around how to determine exactly where these two datasets disagree with one another, and which data points to remove because they are biasing either the treated or control groups (i.e., with their features, they were virtually guaranteed to be either in treatment or in control, and we can't extrapolate treatment effectiveness for them). This is traditionally done with different versions of logistic regression and matching observations with approximately equivalent probabilities of treatment Rubin and Thomas [36].

We address this problem in the context of the canonical LaLonde data set LaLonde [37]. This is a data set of men in the National Supported Work Demonstration who were either given (or not given) on job training for > 9 months. The ultimate goals of this data are to determine the monetary benefits of job training, but we will focus on detecting differences between the groups that were and were not treated. The pre-treatment features of the people are age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), RE74 (earnings in 1974), and RE75 (earnings in 1975). We choose a subset of the data that has information on RE74 following the work of Dehejia-Wahba Dehejia and Wahba [38]. This leaves us with 260 control observations and 185 treated observations.

After z-scoring the 8 dimensional data (i.e., subtract mean and divide by standard deviation), we apply Algorithm 1 comparing the data to itself (rather than constructing a grid in 8D space). We use $\mathsf{deg} = 16$, $\sigma = 1$, and $A = 2$. Here we take $\widehat{F} > 0$ to be treated and $\widehat{F} < 0$ to be control. **Table 2** shows the means of $SigTreat = \{\mathbf{x}_j : \widehat{F}(\mathbf{x}_j) > 0 \text{ and } D(\mathbf{x}_j) = 1\}$ and

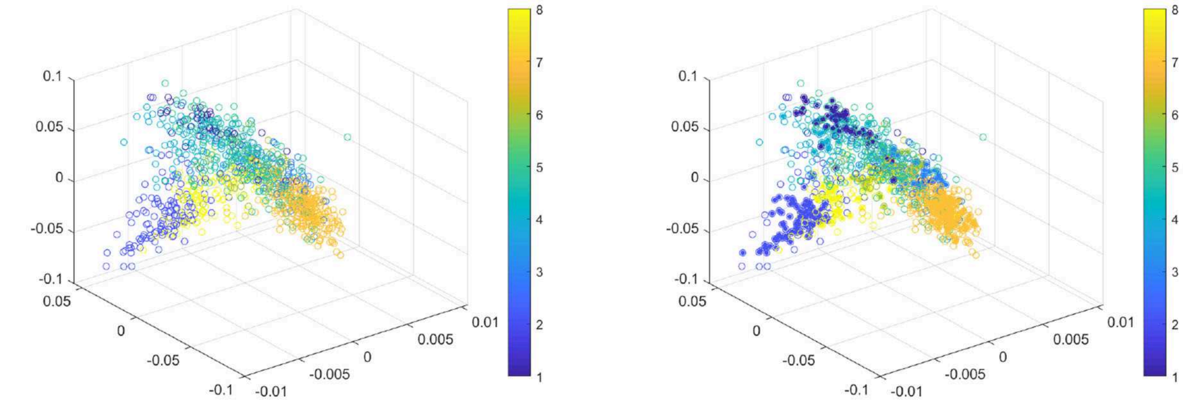


FIGURE 5 | (Left) Hierarchical topic embedding of documents. **(Right)** Embedding highlighted by grid points deemed significantly within one class.

TABLE 1 | Nearest neighbor classification of Science News documents across all documents and across only significant documents.

	Neighbor in given set of docs	Centroid computed from given set of docs
All documents	0.5143	0.5344
Sig. documents	**0.7156**	**0.7635**

$SigControl = \{\mathbf{x}_j : \widehat{F}(\mathbf{x}_j) < 0 \text{ and } D(\mathbf{x}_j) = 1\}$, as well as the means of the treated and control groups independent of significance of the witness function.

Table 2 tells a clear story, namely that the people unlikely to be given job training (i.e., people in *SigControl*) are young, average educated, Hispanic men that are not married, likely don't have a degree, and dropped considerably in income between 1974 and 1975. On the other hand, people that were almost guaranteed to be given job training (i.e., people in *SigTreat*) are older, above average educated, black men that are unmarried, have a degree, and previously made well above average income. These results are consistent with the biases being offered job training that are identified in previous research on propensity matching Dehejia and Wahba [38].

We also display in **Figure 6** the balancing that occurs after removing those people that fall significantly into one group or the other, and display this as a function of A in Algorithm 1. We compare the mean of the remaining control group and the mean of the remaining treated group, and plot the ℓ^2 norm difference between these mean vectors. This demonstrates that, as we remove observations deemed to be significantly in one class or the other, the remaining groups move closer together in mean and become more balanced.

6. PROOFS

The basic idea behind the proof of Theorem 3.1 is the following. Theorem 6.1 gives a deterministic estimate on $\|\sigma_n(Ff) - Ff\|_{\mathbb{B}(\mathbf{x}_0,r)}$. In turn, (3.2) expresses $\sigma_n(Ff)$ as an expected value expression, and $\widehat{F}$ is clearly an estimator of this expression. Therefore, at each point $\mathbf{x}$, one can estimate $\widehat{F}(\mathbf{x}) - \sigma_n(Ff)(\mathbf{x})$ using Hoeffding's inequality. We note that this difference is a weighted polynomial of degree $< n^2$. To convert this estimate into a norm estimate, we need to obtain a finite subset of the ball around $\mathbf{x}_0$ so that the norm of any weighted polynomial of degree $< n^2$ is of the same order of magnitude as the norm on the points of this finite subset, which is therefore called a norming subset. While Corollary 6.1 gives an insight in this direction, a main technical difficulty here is that the norm of the gradient of a weighted polynomial on a cube needs to be estimated by the norm of the polynomial on the cube itself. If one allows the bounds to depend upon the point $\mathbf{x}_0$, then this is a trivial consequence of the Markov inequality. A much deeper argument is needed to obtain the bound independent of $\mathbf{x}_0$. An inequality of this sort is known as the Videnskii inequality proved in the context of trigonometric polynomials on arcs of a circle [39, Chapter 5.1, E.19]. We are not aware of an analog for weighted polynomials.

We prefer to discuss the choice of the norming subset in a greater generality. This is done in section 6.2. The probabilistic estimates on the norms of polynomials are obtained in a more general context as well in section 6.3. The proof of the Videnskii inequality (Theorem 6.3) and Theorem 3.1 are given in section 6.

6.1. Background on Hermite Functions and Weighted Polynomial Approximation

One has the orthogonality relation for $k, j \in \mathbb{Z}_+$,

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

The following lemma [13, Lemma 4.1] lists some important properties of Φ_n (the notation in Chui and Mhaskar [13] is somewhat different; the kernel Φ_n above is $n^q\Phi_n$ in the notation of Chui and Mhaskar [13]).

Lemma 6.1. *Let $S > q$ be an integer.*
(a) *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^q$, $n = 1, 2, \cdots$,*

$$|\Phi_n(\mathbf{x}, \mathbf{y})| \le \frac{cn^q}{\max(1, (n|\mathbf{x} - \mathbf{y}|)^S)}. \tag{6.2}$$

TABLE 2 | Mean value of each feature for the control group and treated group as a whole, and for the subsets of the groups that are deemed to be definitively within one class.

	Age	Years Ed.	Black	Hispanic	Married	No degree	RE74	RE75
Control means	25.1	10.1	0.83	0.11	0.15	0.83	2107.0	1266.9
SigControl means	20.7	10.4	0.33	0.67	0.00	0.89	6325.9	1658.2
Treated means	25.8	10.3	0.84	0.06	0.19	0.71	2095.6	1532.1
sigTreat means	30.6	12.4	1.00	0.00	0.00	0.00	9351.8	2525.9

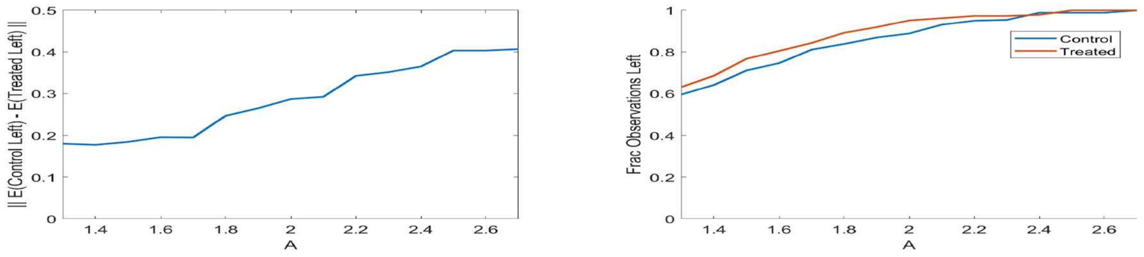


FIGURE 6 | (Left) Difference of means of control and treated groups for LaLonde data after removing points with *D*(*Z*) = 1 for varying *A*. **(Right)** Fraction of observations left in control and treated groups for LaLonde data after removing points with *D*(*Z*) = 1 for varying *A*. For both plots, *A* closer to 1 imples we are liberal with the points being removed, meaning the sets of points remaining will be small but much more similar between treated and control.

In particular,

$$|\Phi_n(\mathbf{x},\mathbf{y})| \le cn^q. \tag{6.3}$$

(b) *For* $\mathbf{x}\in\mathbb{R}^q$*,* $n = 1, 2, \cdots$*,* $1 \le p < \infty$

$$\int_{\mathbb{R}^q} |\Phi_n(\mathbf{x},\mathbf{y})|^p d\mathbf{y} \le cn^{q(1-1/p)}. \tag{6.4}$$

The following proposition lists a few important properties of the spaces Π_n^q (cf. Mhaskar [26], Mhaskar [40], Mhaskar [12]).

Proposition 6.1. *Let* $n > 0$*,* $P \in \Pi_n^q$*,* $1 \le p \le \infty$*.*
(a) (Infinite-finite range inequality) *For any* $\delta > 0$*, there exists* $c = c(\delta, p)$ *such that*

$$\|P\|_{p,\mathbb{R}^q\setminus[-\sqrt{2}n(1+\delta),\sqrt{2}n(1+\delta)]^q} \le c_1 e^{-cn^2}\|P\|_{p,[-\sqrt{2}n(1+\delta),\sqrt{2}n(1+\delta)]^q} \tag{6.5}$$

(b) (MRS identity) *We have*

$$\|P\|_\infty = \|P\|_{\infty,[-\sqrt{2}n,\sqrt{2}n]^q}. \tag{6.6}$$

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

$$\||\nabla P|\|_p \le Bn\|P\|_p. \tag{6.7}$$

In view of Proposition 6.1, we refer to the cube $[-\sqrt{2}n, \sqrt{2}n]^q$ as the critical cube. When $q = 1$, it is often called the MRS (Mhaskar-Rakhmanov-Saff) interval. The following corollary is easy to prove using Proposition 6.1, parts (b) and (c).

Corollary 6.1. *Let* $n > 0$*,* $\mathcal{C} \subset I_{n,q} = [-\sqrt{2}n, \sqrt{2}n]^q$ *be a finite set satisfying*

$$\max_{\mathbf{x}\in I_{n,q}} \min_{\mathbf{y}\in\mathcal{C}} \|\mathbf{x}-\mathbf{y}\| \le 1/(2Bn). \tag{6.8}$$

Then for any $P \in \Pi_n^q$*,*

$$\max_{\mathbf{y}\in\mathcal{C}} |P(\mathbf{y})| \le \|P\|_\infty \le 2\max_{\mathbf{y}\in\mathcal{C}} |P(\mathbf{y})|. \tag{6.9}$$

There exists a set $\mathcal{C}$ *as above with* $|\mathcal{C}| \sim n^{2q}$*.*

We define

$$\sigma_n(f)(\mathbf{x}) = \int_{\mathbb{R}^q} \Phi_n(\mathbf{x},\mathbf{y})f(\mathbf{y})d\mathbf{y}, \qquad f \in L^1+L^\infty,\ n > 0,\ \mathbf{x}\in\mathbb{R}^q. \tag{6.10}$$

The following proposition is routine to prove using Lemma 6.1(b) with $p = 1$:

Proposition 6.2. (a) *If* $n > 0$ *and* $P \in \Pi_{n/\sqrt{2}}$*, then* $\sigma_n(P) = P$*.*
(b) *If* $f \in L^p$*,* $n > 0$*, then*

$$\|\sigma_n(f)\|_p \le c\|f\|_p, \quad E_{n,p}(f) \le \|f-\sigma_n(f)\|_p \le cE_{n/\sqrt{2},p}(f). \tag{6.11}$$

(c) *We have*

$$\begin{aligned}\|f\|_{W_{p,\gamma}} &\sim \|f\|_p + \sup_{n\ge 0} 2^{n\gamma}\|f-\sigma_{2^n}(f)\|_p \\ &\sim \sup_{n\ge 0} 2^{n\gamma}\|\sigma_{2^{n-1}}(f)-\sigma_{2^n}(f)\|_p.\end{aligned} \tag{6.12}$$

The following characterization of local smoothness classes can be obtained by imitating arguments in Mhaskar [12].

Theorem 6.1. *Let* $1 \le p \le \infty$*,* $f \in X^p$*,* $\gamma > 0$*,* $\mathbf{x}_0 \in \mathbb{R}^q$*. The following statements are equivalent:*
(a) $f \in W_{p,\gamma}(\mathbf{x}_0)$*.*
(b) *There exists* $r = r(f, \mathbf{x}_0, p, \gamma) > 0$ *such that*

$$\sup_{n\ge 0} n^\gamma \|f-\sigma_n(f)\|_{p,\mathbb{B}(\mathbf{x}_0,r)} < \infty.$$

(c) *There exists* $r_1 = r_1(f, \mathbf{x}_0, p, \gamma) > 0$ *such that*

$$\sup_{n\ge 0} 2^{n\gamma}\|\sigma_{2^{n-1}}(f) - \sigma_{2^n}(f)\|_{p,\mathbb{B}(\mathbf{x}_0,r_1)} < \infty.$$

If $\mathbf{x}_0 \in \mathbb{R}^q$, $1 \le p \le \infty$, $f \in W_{p,\gamma}(\mathbf{x}_0)$, and part (b) of Theorem 6.1 holds with $\mathbb{B}(\mathbf{x}_0, r)$ for some $r > 0$, we define

$$\|f\|_{p,\gamma,\mathbf{x}_0,r} = \|f\|_p + \sup_{n\ge 0} n^\gamma \|f - \sigma_n(f)\|_{p,\mathbb{B}(\mathbf{x}_0,r)}, \tag{6.13}$$

where we note that this defines a norm since we have used the norm of f on the entire $\mathbb{R}^q$ as one of the summands above.

6.2. Norming Sets

If $\mathbb{X}$ is a topological space, we denote by $C_0(\mathbb{X})$ the space of all continuous real valued functions on $\mathbb{X}$ vanishing at infinity, equipped with the supremum norm.

Definition 6.1. Let $\mathbb{X}$ be a topological space, $W \subset C_0(\mathbb{X})$. A set $\mathcal{C} \subset \mathbb{X}$ is called a **norming set** for W if there exists a finite number $c > 0$ such that

$$\sup_{x\in\mathbb{X}} |f(x)| \le c \sup_{y\in\mathcal{C}} |f(y)|, \qquad f \in W. \tag{6.14}$$

We denote by $\mathfrak{n}(W, \mathcal{C})$ the infimum of all the numbers c that work above.

Definition 6.2. Let $\mathbb{Y}$ be a metric space with metric ρ. If $\epsilon > 0$, a subset $V \subseteq \mathbb{Y}$ is called an ϵ-cover of $\mathbb{Y}$ if

$$\sup_{f\in\mathbb{Y}} \inf_{g\in V} \rho(f, g) \le \epsilon.$$

Proposition 6.3. *Let* $\mathbb{X}$ *be a topological space,* W *be a linear subspace of* $C_0(\mathbb{X})$.
(a) *If there exists a finite norming set* $\mathcal{C}$ *for* W *then* W *is finite dimensional, and* $\dim(W) \le |\mathcal{C}|$.
(b) *Let* W *be finite dimensional,*

$$B_W = \{f \in W : \|f\|_{\infty,\mathbb{X}} = 1\},$$

and $\{f_1, \cdots, f_N\}$ *be a* $1/4$*-cover for* B_W. *Then there exists a norming set* $\mathcal{C}$ *for* W *with* $|\mathcal{C}| \le N$, *and* $\mathfrak{n}(\mathcal{C}, W) \le 2$.
(c) *Let* $\mathbb{X}$ *be a compact metric space with metric* ρ, W *be a finite dimensional linear subspace of* $C_0(\mathbb{X})$, *and there exist* $L = L(W) > 0$ *such that*

$$|f(x) - f(y)| \le L\|f\|_{\infty,\mathbb{X}}\rho(x, y), \qquad f \in W. \tag{6.15}$$

If $\mathcal{C}$ *is a* $1/(2L)$*-cover for* $\mathbb{X}$, *then* $\mathcal{C}$ *is a norming set for* W *with* $\mathfrak{n}(\mathcal{C}, W) \le 2$.

Proof.

If $\mathcal{C}$ is a finite norming subset for W then the mapping $W \to \mathbb{R}^{|\mathcal{C}|}$ given by $f \mapsto (f(x))_{x\in\mathcal{C}}$ is injective. This proves part (a).

To prove part (b), let $\{f_1, \cdots, f_N\}$ be a $1/4$-cover for B_W. Since each of the functions $f_j \in C_0(\mathbb{X})$, there exists $x_j \in \mathbb{X}$ (not necessarily distinct) for which $|f_j(x_j)| = \|f_j\|_{\infty,\mathbb{X}}$. We let $\mathcal{C} = \{x_1, \cdots, x_N\}$. If $f \in B_W$, and $\|f - f_j\|_{\infty,\mathbb{X}} \le 1/4$, then

$$\begin{aligned}1 &= \|f\|_{\infty,\mathbb{X}} \le \|f - f_j\|_{\infty,\mathbb{X}} + \|f_j\|_{\infty,\mathbb{X}} \le 1/4 + |f_j(x_j)| \\ &\le 1/4 + |f_j(x_j) - f(x_j)| + |f(x_j)| \le 1/2 + \max_{1\le j\le N} |f(x_j)|.\end{aligned}$$

This proves part (b).

Let the hypothesis of part (c) be satisfied, and $\mathcal{C}$ be a $1/(2L)$ covering for $\mathbb{X}$. If $f \in W$ and $\|f\|_{\infty,\mathbb{X}} = |f(x^*)|$, then there exists $z \in \mathcal{C}$ with $\rho(x^*, z) \le 1/(2L)$. Then (6.15) implies that

$$\begin{aligned}\|f\|_{\infty,\mathbb{X}} &\le |f(z)| + |f(x^*) - f(z)| \le \max_{y\in\mathcal{C}} |f(y)| + L\rho(x^*, z)\|f\|_{\infty,\mathbb{X}} \\ &\le \max_{y\in\mathcal{C}} |f(y)| + (1/2)\|f\|_{\infty,\mathbb{X}}.\end{aligned}$$

This proves part (c). □

6.3. Probabilistic Estimates

We recall Hoeffding's inequality [41, Appendix A(3)]:

Proposition 6.4. *If* $Y_1, \cdots, Y_M$ *are independent random variables with mean* 0 *such that* $a_j \le Y_j \le b_j$ *for* $j = 1, \cdots, M$, *then for* $\eta > 0$,

$$\mathsf{Prob}\left(\left|\frac{1}{M}\sum_{j=1}^M Y_j\right| \ge \eta\right) \le 2\exp\left(-\frac{M^2\eta^2}{\sum_{j=1}^M (b_j - a_j)^2}\right). \tag{6.16}$$

Next, we wish to develop a general theorem estimating the probability of a lower bound on the supremum norm of a random family of functions analogous to [42, Lemma 6.2].

Theorem 6.2. *Let* $\mathbb{X}$ *be a topological space,* W *be a linear subspace of* $C_0(\mathbb{X})$. *We assume that there is a finite norming set* $\mathcal{C}$ *for* W. *Let* $(\Omega, \mathcal{B}, \mu)$ *be a probability space, and* $Z : \Omega \to W$. *We assume further that for any* $x \in \mathbb{X}$, $\omega \in \Omega$, $|Z(\omega)(x)| \le R$ *for some* $R > 0$. *Then for any* $\delta > 0$, *integer* $M \ge 1$, *and independent sample* $\omega_1, \cdots, \omega_M$, *we have*

$$\begin{aligned}&\mathsf{Prob}_\mu\left(\sup_{x\in\mathbb{X}}\left|\frac{1}{M}\sum_{j=1}^M Z(\omega_j)(x)\right.\right. \\ &\left.\left. - \ \mathbb{E}_\mu(Z(\circ)(x))\right| \ge 4\mathfrak{n}(W, \mathcal{C})R\sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{M}}\right) \le \delta.\end{aligned} \tag{6.17}$$

Proof. Let $x \in \mathbb{X}$, and $Y_j = Z(\omega_j)(x) - \mathbb{E}_\mu(Z(\circ)(x))$. Then $|Y_j| \le 2R$, and Hoeffding's inequality implies that for any $\eta > 0$,

$$\begin{aligned}&\mathsf{Prob}_\mu\left(\left|\frac{1}{M}\sum_{j=1}^M Z(\omega_j)(x) - \mathbb{E}_\mu(Z(\circ)(x))\right| \ge \frac{\eta}{2\mathfrak{n}(W, \mathcal{C})}\right) \\ &\le 2\exp\left(-\frac{M\eta^2}{16\mathfrak{n}(W, \mathcal{C})^2R^2}\right).\end{aligned} \tag{6.18}$$

We observe that since W is a finite dimensional linear space, the functions

$$x \mapsto \frac{1}{M}\sum_{j=1}^{M} Z(\omega_j)(x) - \mathbb{E}_\mu(Z(\circ)(x))$$

are in W for all choices of the ω_j's. We now apply (6.18) with each $y \in \mathcal{C}$ in place of x to conclude that

$$\begin{aligned}
&\mathsf{Prob}_\mu\left(\sup_{x\in\mathbb{X}}\left|\frac{1}{M}\sum_{j=1}^{M} Z(\omega_j)(x) - \mathbb{E}_\mu(Z(\circ)(x))\right| \ge \eta\right)\\
&\le \bigcup_{y\in\mathcal{C}} \mathsf{Prob}_\mu\left(\left|\frac{1}{M}\sum_{j=1}^{M} Z(\omega_j)(y) - \mathbb{E}_\mu(Z(\circ)(y))\right| \ge \frac{\eta}{2\mathfrak{n}(W,\mathcal{C})}\right)\\
&\le 2|\mathcal{C}|\exp\left(-\frac{M\eta^2}{16\mathfrak{n}(W,\mathcal{C})^2R^2}\right).
\end{aligned}$$

We set the last expression above to δ and solve for η to obtain (6.17). □

The following lemma states the asymptotics in (3.4) for the Lambert function in (3.3) in a form that is easily applicable in the proof of (3.8) in Theorem 3.1.

Lemma 6.2. *Let $y, \alpha, \beta, A, B > 0$. The solution of the equation*

$$Ax^\alpha \log(Bx^\beta) = y \tag{6.19}$$

is given by

$$\begin{aligned}
x &= B^{-1/\beta}\exp\left(\frac{1}{\alpha}W\left(\frac{B^{\alpha/\beta}}{A}\frac{\alpha}{\beta}y\right)\right)\\
&\approx \left(\frac{\alpha}{\beta A}\right)^{1/\alpha}\left\{\frac{y}{\log y + \log\left(\frac{\alpha B^{\alpha/\beta}}{\beta A}\right)}\right\}^{1/\alpha},
\end{aligned} \tag{6.20}$$

where $\approx$ denotes an asymptotic relationship.

Proof. We multiply both sides of (6.19) by α/β, and write $z = \log\left(B^{\alpha/\beta}x^\alpha\right)$ to obtain

$$e^z = B^{\alpha/\beta}x^\alpha, \quad ze^z = \frac{\alpha}{\beta}\frac{B^{\alpha/\beta}}{A}y.$$

The solution of this equation leads to the first expression on the right hand side of (6.20). The asymptotic expression in (3.4) leads to the second expression. □

6.4. Proof of Theorem 3.1

We wish to obtain a Videnskii inequality for weighted polynomials; i.e., an analog of (6.7) for derivatives on a ball. We note that the following bound is much worse than (6.7), but is adequate for our purpose.

Theorem 6.3. *Let $n \ge 1$, $0 < r < n^3/2$, $\mathbf{x}_0 \pm 2r \in [-2n, 2n]^q$, $P \in \Pi_n^q$. Then there exists constant $C > 0$ independent of r, n, and $\mathbf{x}_0$, and depending linearly on q such that*

$$\||\nabla P|\|_{\infty,\mathbb{B}(\mathbf{x}_0,r)} \le Cn^4\|P\|_{\infty,\mathbb{B}(\mathbf{x}_0,r)} \times \begin{cases} re^{1/r}, & \text{if } r \ge c/n^2,\\ (1/r)\exp(8n^2r) \le c_1/r, & \text{otherwise.}\end{cases} \tag{6.21}$$

Proof. It is sufficient to prove (6.21) for $q = 1$; the general case follows by applying the univariate inequality one component at a time. We will simplify our notation by writing $m = n^2$ in this proof. Let $P \in \Pi_n^1$, and Q be a (univariate) polynomial of degree $< m$ such that $P(x) = Q(x)\exp(-x^2/2)$. Without loss of generality, let $\|P\|_{\infty,[x_0-r,x_0+r]} = 1$.

Let $\|Q\|_{\infty,[x_0-r,x_0+r]} = |Q(x^*)|$. In view of Markov's inequality ([43, Chapter VI, section 6]), we obtain for $x \in [x_0 - r, x_0 + r]$

$$\begin{aligned}
|P'(x)| &= |Q'(x) - xQ(x)|\exp(-x^2/2)\\
&\le \frac{m^2}{r}\|Q\|_{\infty,[x_0-r,x_0+r]}\exp(-x^2/2) + 2\sqrt{m}\\
&\le 2\sqrt{m} + \frac{m^2}{r}|Q(x^*)|\exp(-(x^*)^2/2)\exp\left(\frac{1}{2}|x^*-x||x^*+x|\right)\\
&\le 2n + \frac{n^4}{r}\exp(4nr).
\end{aligned} \tag{6.22}$$

Thus,

$$\|P'\|_{\infty,[x_0-r,x_0+r]} \le 2\frac{m^2}{r}\exp(4nr), \qquad 0 < r < n^3/2. \tag{6.23}$$

This proves, in particular, (6.21) in the case when $0 < r \le c/n^2$ for *every* $c > 0$.

We now assume that $r \ge c/n^2$ for c to be chosen later. We consider the function $G : \mathbb{C} \to \mathbb{C}$ defined by

$$G(z) = \left\{\frac{z - x_0 + \sqrt{(z-x_0)^2 - r^2}}{r}\right\}^m \exp\left(-\frac{(z+x_0)\sqrt{(z-x_0)^2-r^2}}{2}\right), \tag{6.24}$$

where the principle branch of the square root is chosen (so that $\sqrt{\zeta} \to \infty$ as $\zeta \to \infty$). The mapping

$$w = \frac{z - x_0 + \sqrt{(z-x_0)^2 - r^2}}{r} \tag{6.25}$$

maps the exterior of $[x_0 - r, x_0 + r]$ to the exterior of the complex unit disc, and obviously, $|w| = 1$ if $z \in [x_0 - r, x_0 + r]$. So, $|G(z)| = 1$ if $z \in [x_0 - r, x_0 + r]$. Hence, the function

$$F(z) = \frac{Q(z)\exp(-z^2/2)}{G(z)}$$

is analytic in $\mathbb{C} \cup \{\infty\} \setminus [x_0 - r, x_0 + r]$ and $|F(z)| \le 1$ on $[x_0 - r, x_0 + r]$. Therefore, the maximum modulus principle shows that

$$|P(z)| = |Q(z)\exp(-z^2/2)| \le |G(z)|, \qquad z \in \mathbb{C}\setminus[x_0 - r, x_0 + r]. \tag{6.26}$$

We wish to use this bound on the interior of the ellipse $\mathcal{E}$ defined by $|w| = 1 + 1/(rm)$. Since

$$z = x_0 + \frac{r}{2}(w + 1/w),$$

we calculate that

$$(z - x_0)^2 - r^2 = \frac{r^2}{4}\left((w + w^{-1})^2 - 4\right) = \frac{r^2}{4}(w - w^{-1})^2,$$

so that

$$(z + x_0)\sqrt{(z - x_0)^2 - r^2} = x_0 r(w - w^{-1}) + \frac{r^2}{4}(w^2 - w^{-2}).$$

Therefore, on the ellipse $\mathcal{E}$ parameterized by $w = (1 + (rm)^{-1})\exp(i\theta)$,

$$\begin{aligned}&\left|\Re e\left((z + x_0)\sqrt{(z - x_0)^2 - r^2}\right)\right| \\ &= \left|\left(\frac{2x_0}{m}\cos\theta + \frac{4r}{m}\cos(2\theta)\right)(1 + O(1/r^2m^2))\right| \le c\frac{|x_0 + 2r|}{m}.\end{aligned}$$

Thus, if $|x_0 + 2r| \le 2\sqrt{m}$, the estimate (6.26) shows that for z on and inside $\mathcal{E}$,

$$|P(z)| \le (1 + (rm)^{-1})^m \exp\left(c_1/\sqrt{m}\right). \tag{6.27}$$

Now, let $x \in [x_0 - r, x_0 + r]$. It is easily calculated that the minimum distance between $\mathcal{E}$ and $[x_0 - r, x_0 + r]$ is at least $2c_2(rm^2)^{-1}$ for a constant c_2 independent of r and m as long as $rm \ge c$. Hence, the complex disc bounded by the contour $\Gamma : \{\zeta : |\zeta - x| = c_2(rm^2)^{-1}\}$ is contained in the interior of $\mathcal{E}$. On this disc, the bound (6.27) holds. Therefore, the Cauchy integral formula leads to

$$|P'(x)| = \frac{1}{2\pi}\left|\oint_\Gamma \frac{P(\zeta)}{(\zeta - x)^2}d\zeta\right| \le c_3 rm^2 e^{1/r}.$$

Recalling that $m = n^2$, this leads to (6.21) when $q = 1$. As explained earlier, this completes the proof. □

PROOF OF THEOREM 3.1.

Let $\mathbf{x}_0 \in \mathbb{R}^q$, $n \ge 1$, $r \ge c/n^2$ where c is as in Theorem 6.3. We apply Theorem 6.2 with $\mathbb{X} = \mathbb{B}(\mathbf{x}_0, r)$, Π_n^q (restricted to $\mathbb{X}$) in place of W, $\mathbb{R}^q \times \Omega$ in place of Ω, τ as in Theorem 3.1. We take

$$Z(z, \epsilon)(\mathbf{x}) = \mathcal{F}(\mathbf{z}, \epsilon)\Phi_n(\mathbf{x}, \mathbf{z}),$$

where $(\mathbf{z}, \epsilon)$ is a random sample from τ. Then the quantity R in Theorem 6.2 can be chosen to be $c_1\|\mathcal{F}\|_\infty n^q$. Moreover, (3.2) shows that,

$$\mathbb{E}_\tau(Z)(\mathbf{x}) = \sigma_n(Ff)(\mathbf{x}).$$

In view of Theorem 6.3, there exists a norming set $\mathcal{C} \subset \mathbb{B}(\mathbf{x}_0, r)$ for W with $|\mathcal{C}| \sim \exp(q/r)(n^4r^2)^q$ and $\mathfrak{n}(\mathcal{C}, W) \le 2$. Theorem 6.2 used with $\delta(r/n)^q$ in place of δ shows that

$$\begin{aligned}&\mathsf{Prob}_{\mu\times\tau}\left(\left\|\frac{1}{M}\sum_{j=1}^{M}\mathcal{F}(\mathbf{y}_j, \epsilon_j)\Phi_n(\circ, \mathbf{y}_j) - \sigma_n(Ff)\right\|_{\infty,\mathbb{B}(\mathbf{x}_0,r)}\right. \\ &\left.\ge c_1\|\mathcal{F}\|_\infty n^q\sqrt{\frac{\log(c_3 r^q \exp(q/r)n^{5q}/\delta)}{M}}\right) \le \delta(r/n)^q.\end{aligned} \tag{6.28}$$

Since $Ff \in W_{\infty,\gamma}(\mathbf{x}_0)$, Theorem 6.1 shows that

$$\|Ff - \sigma_n(Ff)\|_{\infty,\mathbb{B}(\mathbf{x}_0,r)} \le n^{-\gamma}\|Ff\|_{\infty,\gamma,\mathbf{x}_0,r}. \tag{6.29}$$

Therefore, (6.28) leads to (3.6).

To prove (3.8), we solve for n the equation

$$n^{2q+2\gamma}\log(Bn^{5q}) = M,$$

where B is defined in the statement of Theorem 3.1. Lemma 6.2, used with $A = 1$, $\alpha = 2q + 2\gamma$, $\beta = 5q$ shows that the solution is given by (3.7). Therefore, with this choice of n, (3.6) implies (3.8). □

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: MNIST, Lalonde, and SciNews data sets are all available on UCI Machine Learning repository.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by NSF DMS-1818945 and DMS-1819222. AC was also supported by NSF DMS-2012266, and Sage Foundation Grant 2196. XC was also supported by NSF DMS-1820827, NIH Grant R01GM131642, and the Alfred P. Sloan Foundation. HM was supported by NSF DMS-2012355. This manuscript has been released as a pre-print [44].

REFERENCES

1. Mhaskar HN. Approximation properties of a multilayered feedforward artificial neural network. *Adv Comput Math.* (1993) **1**:61–80. doi: 10.1007/BF02070821
2. Maggioni M, Mhaskar HN. Diffusion polynomial frames on metric measure spaces. *Appl Comput Harmon Anal.* (2008) **24**:329–53. doi: 10.1016/j.acha.2007.07.001
3. Chui CK, Donoho DL. Special issue: diffusion maps and wavelets. *Appl Comput Harm Anal.* (2006) **21**:1–2. doi: 10.1016/j.acha.2006.05.005
4. Mhaskar HN. A unified framework for harmonic analysis of functions on directed graphs and changing data. *Appl Comput Harm Anal.* (2018) **44**:611–44. doi: 10.1016/j.acha.2016.06.007
5. Mhaskar HN, Prestin J. Polynomial frames: a fast tour. In: Chui CK, Neamtu M, Schumaker LL, editors. *Approximation Theory XI.*. Brentwood, TN: Nashboro Press (2004). p. 101–32.
6. Filbir F, Mhaskar HN. Marcinkiewicz–Zygmund measures on manifolds. *J Complex.* (2011) **27**:568–96. doi: 10.1016/j.jco.2011.03.002
7. Mhaskar HN. Eignets for function approximation on manifolds. *Appl Comput Harmon Anal.* (2010) **29**:63–87. doi: 10.1016/j.acha.2009.08.006
8. Ehler M, Filbir F, Mhaskar HN. Locally learning biomedical data using diffusion frames. *J Comput Biol.* (2012) **19**:1251–64. doi: 10.1089/cmb.2012.0187
9. Mhaskar HN, Pereverzyev SV, van der Walt MD. A deep learning approach to diabetic blood glucose prediction. *Front Appl Math Stat.* (2017) **3**:14. doi: 10.3389/fams.2017.00014
10. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. *J Mach Learn Res.* (2012) **13**:723–73.
11. Cheng X, Cloninger A, Coifman RR. Two-sample statistics based on anisotropic kernels. *Inf Inference: J IMA* (2017). doi: 10.1093/imaiai/iaz018
12. Mhaskar HN. Local approximation using Hermite functions. In: *Progress in Approximation Theory and Applicable Complex Analysis*. Springer (2017). p. 341–62. doi: 10.1007/978-3-319-49242-1_16
13. Chui CK, Mhaskar HN. A Fourier-invariant method for locating point-masses and computing their attributes. *Appl Comput Harmon Anal.* (2018) **45**:436–52. doi: 10.1016/j.acha.2017.08.010
14. Mhaskar HN. When is approximation by Gaussian networks necessarily a linear process? *Neural Netw.* (2004) **17**:989–1001. doi: 10.1016/j.neunet.2004.04.001
15. Chui CK, Mhaskar HN. A unified method for super-resolution recovery and real exponential-sum separation. *Appl Comput Harmon Anal.* (2018) **46**:431–51. doi: 10.1016/j.acha.2017.12.007
16. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599* (2017).
17. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).
18. Hein M, Andriushchenko M, Bitterwolf J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *arXiv preprint arXiv:1812.05720v1* (2018). doi: 10.1109/CVPR.2019.00013
19. Sutherland DJ, Tung HY, Strathmann H, Ramdas A, Smola A, Gretton A. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488* (2016).
20. Shahnaz F, Berry MW, Pauca VP, Plemmons RJ. Document clustering using nonnegative matrix factorization. *Inform Process Manage.* (2006) **42**:373–86. doi: 10.1016/j.ipm.2004.11.005
21. Wang C, Mahadevan S. Multiscale analysis of document corpora based on diffusion models. In: *IJCAI*. Pasadena, CA (2009). p. 1592–7.
22. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* (1983) **70**:41–55. doi: 10.1093/biomet/70.1.41
23. King G, Nielsen R. Why propensity scores should not be used for matching. *Copy at http://j. mp/1sexgVw Download Citation BibTex Tagged XML Download Paper*, 378, 2016.
24. Szegö G. Orthogonal polynomials. In: *Colloquium Publications/American Mathematical Society*. Vol. 23. Providence, RI (1975).
25. Bateman H, Erdélyi A, Magnus W, Oberhettinger F, Tricomi FG. *Higher Transcendental Functions*. Vol. 2. New York, NY: McGraw-Hill (1955).
26. Mhaskar HN. *Introduction to the Theory of Weighted Polynomial Approximation*. Vol. 56. Singapore: World Scientific (1996).
27. Mhaskar HN. On the degree of approximation in multivariate weighted approximation. In: *Advanced Problems in Constructive Approximation*. Basel: Birkhäuser (2003). p. 129–41.
28. Pesarin F. *Multivariate Permutation Tests: With Applications in Biostatistics*. Vol. 240. Chichester: Wiley (2001).
29. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. Montreal, QC (2014). p. 2672–80.
30. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
31. Doersch C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
32. LeCun Y, Cortes C, Burges C. *Mnist Handwritten Digit Database.* AT&T Labs [Online] (2010). Available online at: http://yann.lecun.com/exdb/mnist
33. Reynolds D. Gaussian mixture models. *Encyclop Biometr.* (2015) **741**:827–32. doi: 10.1007/978-1-4899-7488-4_196
34. Krizhevsky A, Nair V, Hinton G. *The Cifar-10 Dataset.* (2014). Available online at: http://www.cs.toronto.edu/~kriz/cifar.html
35. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
36. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics.* (1996) **52**:249–64. doi: 10.2307/2533160
37. LaLonde RJ. Evaluating the econometric evaluations of training programs with experimental data. *Am Econ Rev.* (1986) 604–20.
38. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc.* (1999) **94**:1053–62. doi: 10.1080/01621459.1999.10473858
39. Borwein P, Erdélyi T. *Polynomials and Polynomial Inequalities*. Vol. 161. New York, NY: Springer Science & Business Media (2012).
40. Mhaskar HN. A Markov-Bernstein inequality for Gaussian networks. In: de Bruin MG, Mache DH and Szabados J, editors. *Trends and Applications in Constructive Approximation*. Basel: Springer (2005). p. 165–80.
41. Pollard D. *Convergence of Stochastic Processes*. New York, NY: Springer Science & Business Media (2012).
42. Gia QL, Mhaskar H. Localized linear polynomial operators and quadrature formulas on the sphere. *SIAM J Numer Anal.* (2009) **47**:440–66. doi: 10.1137/060678555
43. Natanson IP. *Constructive Function Theory*. New York, NY: Ungar (1964).
44. Mhaskar HN, Cloninger A, Cheng X. A witness function based construction of discriminative models using hermite polynomials. *arXiv preprint arXiv:1901.02975* (2019).
45. Andrews GE, Askey R, Roy R. *Special Functions.* Vol. 71. Cambridge: Cambridge University Press (1999).

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

A. IMPLEMENTATION OF THE KERNELS

In this sub-section, we describe the construction of the kernels Φ_n. We remark firstly that Next, we recall the Mehler identity (cf. [45, Formula (6.1.13)] in the univariate case), valid for $w \in \mathbb{C}$, $|w| < 1$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^q$:

$$\sum_{\mathbf{k}\in\mathbb{Z}^q} \psi_{\mathbf{k}}(\mathbf{x})\psi_{\mathbf{k}}(\mathbf{y})w^{|\mathbf{k}|_1} = \sum_{m=0}^{\infty} w^m \mathsf{Proj}_m(\mathbf{x},\mathbf{y})$$
$$= \frac{1}{(\pi(1-w^2))^{q/2}} \exp\left(\frac{4w\mathbf{x}\cdot\mathbf{y} - (1+w^2)(|\mathbf{x}|^2 + |\mathbf{y}|^2)}{2(1-w^2)}\right). \tag{A.1}$$

It is clear that the projections $\mathsf{Proj}_m(\mathbf{x},\mathbf{y})$ depend only on $\mathbf{x}\cdot\mathbf{y}$, $\mathbf{x}\cdot\mathbf{x}$, and $\mathbf{y}\cdot\mathbf{y}$; in particular, that they are rotation independent. Denoting by θ the acute angle between $\mathbf{x}$ and $\mathbf{y}$, we may thus write

$$\mathsf{Proj}_m(\mathbf{x},\mathbf{y})$$
$$= \sum_{j=0}^{m} \psi_j(|\mathbf{x}|)\psi_j(|\mathbf{y}|\cos\theta) \sum_{\ell=0}^{m-j} \psi_\ell(0)\psi_\ell(|\mathbf{y}|\sin\theta) \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^{q-2},\\ |\mathbf{k}|_1=m-j-\ell}} |\psi_{\mathbf{k}}(\mathbf{0})|^2. \tag{A.2}$$

Using the Mehler identity (A.1), we deduce that for $q \geq 3$,

$$\sum_{r=0}^{\infty} w^{2r} \sum_{\substack{|\mathbf{k}|_1=2r\\ \mathbf{k}\in\mathbb{Z}_+^{q-2}}} |\psi_{\mathbf{k}}(\mathbf{0})|^2 = (\pi(1-w^2))^{-(q-2)/2}$$
$$= \pi^{1-q/2} \sum_{r=0}^{\infty} \frac{\Gamma(q/2+r-1)}{\Gamma(q/2-1)r!} w^{2r}.$$

Hence,

$$\mathsf{Proj}_m(\mathbf{x},\mathbf{y})$$
$$= \sum_{j=0}^{m} \psi_j(|\mathbf{x}|)\psi_j(|\mathbf{y}|\cos\theta) \sum_{\ell=0}^{m-j} \psi_\ell(0)\psi_\ell(|\mathbf{y}|\sin\theta) D_{q-2;m-j-\ell}, \tag{A.3}$$

where

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

and

$$D_{q-2;r} = \begin{cases} \pi^{1-q/2}\dfrac{\Gamma(q/2+r/2-1)}{\Gamma(q/2-1)(r/2)!}, & \text{if } r \text{ is even, } q \geq 3, \\ 0, & \text{if } r \text{ is odd, } q \geq 3, \\ 1, & \text{if } q \leq 2. \end{cases} \tag{A.5}$$

Remark A.1. A completion of squares shows that

$$(1+w^2)(|\mathbf{x}|^2 + |\mathbf{u}|^2) - 4w\mathbf{x}\cdot\mathbf{u}$$
$$= (1+w^2)\left|\mathbf{x} - \frac{2w}{1+w^2}\mathbf{u}\right|^2 + \frac{(1-w^2)^2}{1+w^2}|\mathbf{u}|^2. \tag{A.6}$$

With the choice $w = 1/\sqrt{3}$, the Mehler identity (A.1) then shows that

$$\sum_{m=0}^{\infty} 3^{-m/2}\mathsf{Proj}_m(\mathbf{x},\mathbf{u})$$
$$= \left(\frac{3}{2\pi}\right)^{q/2} \exp\left(-|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{u})|^2\right) \exp(-|\mathbf{u}|^2/4).$$

It is now easy to calculate using orthogonality that

$$\Phi_n(\mathbf{x},\mathbf{y}) = \left(\frac{3}{2\pi}\right)^{q/2} \int_{\mathbb{R}^q} \exp\left(-|\mathbf{x} - \frac{\sqrt{3}}{2}\mathbf{y})|^2\right)$$
$$\exp(-|\mathbf{y}|^2/4)\widetilde{\Phi}_n(\mathbf{y},\mathbf{u})d\mathbf{u}, \tag{A.7}$$

where

$$\widetilde{\Phi}_n(\mathbf{y},\mathbf{u}) = \sum_{m=0}^{\infty} 3^{m/2} H\left(\frac{\sqrt{m}}{n}\right) \mathsf{Proj}_m(\mathbf{y},\mathbf{u}). \tag{A.8}$$

A careful discretization of (A.7) using a Gauss quadrature formula for Hermite weights exact for polynomials of degree $3n^2$ leads to an interpretation of the kernels Φ_n in terms of Gaussian networks with fixed variance and fixed centers, independent of the data. We refer to Mhaskar [14], Chui and Mhaskar [15] for the details.

There is no training required for these networks. However, the mathematical results are applicable only to this specific construction. Treating the theorem as a pure existence theorem and then trying to find a Gaussian network by traditional training will not work. □