A Spectral Analysis of Dot-product Kernels

Meyer Scetbon
CREST, ENSAE

Zaid Harchaoui
University of Washington

Abstract

We present eigenvalue decay estimates of integral operators associated with compositional dot-product kernels. The estimates improve on previous ones established for power series kernels on spheres. This allows us to obtain the volumes of balls in the corresponding reproducing kernel Hilbert spaces. We discuss the consequences on statistical estimation with compositional dot product kernels and highlight interesting trade-offs between the approximation error and the statistical error depending on the number of compositions and the smoothness of the kernels.

Introduction

Dot product kernels are important tools to tackle signal or image data in machine learning, statistical estimation, and computational mathematics (Steinwart and Christmann, 2008; Eggermont and LaRiccia, 2001; Wendland, 2005; Dyn et al., 2001). Normalizing signal and image data to lie on a sphere is common in signal processing and computer vision (Mairal et al., 2014a). The shape and the volume of the reproducing kernel Hilbert space is reflected through the decay of the eigenvalues of the associated integral operator.

The spectrum of eigenvalues of an integral operator associated with the Gaussian radial basis function kernel was first presented by Smola et al. (2001). The subject was further explored in several papers (Cho and Saul, 2010; Zwicknagl, 2009; Azevedo and Menegatto, 2014). Recently, dot product kernels have been considered in relation to the theoretical analysis of deep networks (Daniely et al., 2016; Bach, 2017a; Song et al., 2018) and in relation to the design of new kernel-based methods (Mairal et al., 2014b; Mroueh et al., 2015).

We present in this paper general estimates of eigenvalue decay of integral operators associated with dot product kernels of the form

$$K(x,x) := f(\langle x,x'\rangle_{\mathbb{R}^d}) \tag{1}$$

when the function f satisfies regularity conditions on $[-1,1]$. The eigenvalue decay estimates we obtain generalize previous fundamental results on Mercer decompositions and eigenvalue estimates of dot product kernels (Zwicknagl, 2009; Azevedo and Menegatto, 2014).

Spherical harmonic functions are central to our analysis. These special functions arise as the eigenfunctions of the integral operator associated with the simple dot product kernel. We highlight the *relationship between f, the smoothness properties of the dot product kernel K and the rate of decay of the eigenvalues of the associated integral operator*. The conditions we provide are concrete and verifiable, boiling down to conditions related to a Taylor series expansion of the kernel. This allows us to characterize the reproducing kernel Hilbert space, obtain estimates of the effective dimension in statistical estimation of eigendecay and show the learning rates of the regularized least-squares algorithm in all regimes.

The results we present here can potentially be used in a number of contexts. We illustrate them on three examples related to the theoretical analysis of deep networks. These examples allow us to relate the nonlinear activation functions involved in the construction of a deep network to the spectrum of eigenvalues of an integral operator. In particular, we show that, as one iterates the composition of a nonlinear function, the effect on the spectrum is different if the nonlinearity is smooth, as in the case of the exponential or the Swish activation (Ramachandran et al., 2017), or non-smooth, as in the special case of the ReLU activation (Goodfellow et al., 2016).

Furthermore our results also establish sufficient conditions for this family of kernels in (1) to be universal. The universality of a kernel is a key property which guarantees Bayes-consistency (Steinwart and Christmann, 2008). We show that the universality can be related to smoothness properties of the function f.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

All the proofs can be found in the long version (Scetbon and Harchaoui, 2021). We start in Sec. 1 with a refresher on spherical harmonics and eigenspectra of integral operators associated with dot product kernels. In Sec. 2, we present our main results on eigendecay estimates for these integral operators. Table 2 summarizes our results. In Sec. 3, we explore the statistical implications for regularized least-squares. Finally, in Sec. 4, we discuss examples related to deep networks.

Related work We give here a brief overview of the related works. The variety of the related works shows the versatility of dot product kernels in machine learning and related fields and the importance of general results on the eigendecay of integral operators.

Dot product kernels. Smola et al. (2001) provided in a seminal work estimates of eigenvalue decay for simple dot product kernels. Eigendecay estimates for power series kernels were obtained by Zwicknagl (2009); Azevedo and Menegatto (2014) in a particular eigendecay regime. We obtain tight eigendecay estimates in a broad range of regimes. The results we present can also be potentially applied to recent kernel-based alternatives to deep networks (Shankar et al., 2020).

Regularized least-squares. Caponnetto and De Vito (2007) studied regularized least-square in a reproducing kernel Hilbert spaces in the polynomial regime of eigendecay of the spectrum of the integral operator. The polynomial regime is also common in asymptotic statistical results; see also (Gu, 2013) for a review. We extend this line of work by delineating and studying the geometric regime and the super-geometric regime. The analysis requires a careful control of the eigendecay. The tools we develop for this purpose can be of independent interest.

Deep networks in kernel regime. Recent work has shown that a fully connected network, *i.e.*, a multilayer perceptron, trained with gradient descent may behave like a (tractable) kernel method in a certain over-parameterized regime. See (Chen et al., 2020) for instance. The framework we develop here can be applied to such a tangent kernel and obtain the rate of decay of the eigenvalues of the integral operator associated with the kernel. While our theoretical results cover a broad class of activation functions, very recent work has considered the special case of the ReLU activation and developed a tailored analysis for that case. Bietti and Bach (2021) argue that in that case the RKHS remains unchanged regardless of the depth of the neural network; see also (Chen and Xu, 2020). In this paper, we relate the behavior of the coefficients in the Taylor expansion of f to the decay of the eigenvalues of the integral operator. Moreover, we cover all regimes of eigendecay, including the regime corresponding to the ReLU activation.

Kernels on spheres and shallow networks. In (Bach, 2017a), reproducing kernel Hilbert spaces of dot product kernels are used to analyze single-hidden layer neural networks with input data normalized on the sphere. We extend the work of (Bach, 2017a) in that we analyze neural networks with more than one hidden layer in various eigenvalue decay regimes including the geometric and super-geometric ones which were not considered in (Bach, 2017a). Moreover, in contrast to (Bach, 2017a) in which the learning problem is assumed to be realizable, *i.e.*, the target function is assumed to live in the function space, we work with source conditions which allow us to obtain statistical convergence rates under more general assumptions.

Hilbertian envelopes of deep networks. In (Zhang et al., 2016), reproducing kernel Hilbert spaces are used to analyze multi-layer neural networks with smooth activation functions. The family of kernels we consider in Prop. 4.1 generalizes the one studied in (Zhang et al., 2016) as our kernels are adaptive to the nonlinear functions involved in the construction of the network. In (Suzuki, 2018), excess risk bounds for multi-layer perceptrons are presented. The eigendecay estimates we obtain result in estimates of effective dimension or degrees-of-freedom of multi-layer perceptrons.

1 Dot Product Kernels and their Spectral Decompositions

Kernels on spheres are ubiquitous in machine learning, statistical estimation, and computational mathematics (Smola et al., 2001; Steinwart and Christmann, 2008; Eggermont and LaRiccia, 2001; Wendland, 2005; Dyn et al., 2001). Simple kernels on spheres date back to the seminal works on reproducing kernels (Schoenberg, 1942). Examples of simple kernels on spheres include homogeneous polynomial kernels, inhomogeneous polynomial kernels, and Vovk's polynomial kernels (Smola et al., 2001; Steinwart and Christmann, 2008). The analysis of a dot product kernel on the sphere hinges upon a Taylor-like expansion which gives, on the one hand, a spectral decomposition, and on the other hand, a Mercer decomposition.

Dot product kernel on the sphere. Let $d \geq 2$ and S^{d-1} be the unit sphere of $\mathbb{R}^d$. A kernel of the form

$$K(x,y) = \sum_{m\geq 0} b_m(\langle x,y\rangle)^m, \quad x,y \in S^{d-1} \tag{2}$$

where $(b_m)_{m\geq 0}$ is an absolutely summable sequence is called a *dot product kernel on the sphere* S^{d-1}.

Kernel	b_m	Space	μ	λ_m
$\exp(-c\|x-y\|_2)$	$b_m \in \mathcal{O}(m^{-3/2})$	S^{d-1}	$d\sigma_{d-1}$	$m^{-d/2}$
$\pi - \arccos(\langle x, x' \rangle)$	$b_m \in \mathcal{O}(m^{-3/2})$	S^{d-1}	$d\sigma_{d-1}$	$m^{-d/2}$
$(2 - \langle x, x' \rangle)^{-1}$	$b_m \in \mathcal{O}(2^{-m})$	S^{d-1}	$d\sigma_{d-1}$	2^{-m}
$\exp(-b\|x - x'\|_2^2)$	$\|b_m/b_{m-1}\| \in \mathcal{O}(m^{-1})$	S^{d-1}	$d\sigma_{d-1}$	$(eb)^m m^{-m+(d-1)/2}$
$\exp(-b(x-x')^2)$	$\|b_m/b_{m-1}\| \in \mathcal{O}(m^{-1})$	$[0,1]$	$\propto \exp(-2ax^2)$	$(b/(a+b))^m$
$1 + \frac{(-1)^{s-1}(2\pi)^{2s}}{(2s)!} B_{2s}(\{x-y\})$	/	$[0,1]$	dx	m^{-2s}

Table 1: Eigendecay rates for different kernels. The kernels above the horizontal line are dot-product kernels on the sphere.

Note that the construction we describe below could be extended to dot product kernels in Hilbert spaces (Schoenberg, 1942). If $b_m \geq 0$ for every $m \geq 0$, then K is a *continuous positive semi-definite kernel* on the sphere S^{d-1} (Pinkus, 2004; Zwicknagl, 2009).

Integral operator. Let $L_2^{d\sigma_{d-1}}(S^{d-1})$ be the space of real square-integrable functions on the sphere S^{d-1} endowed with its induced Lebesgue measure $d\sigma_{d-1}$ and $|S^{d-1}|$ the surface area of S^{d-1}. Given a positive semi-definite dot product kernel K, we define the integral operator on $L_2^{d\sigma_{d-1}}(S^{d-1})$ associated

$$\begin{array}{llcl} T_K & : L_2^{d\sigma_{d-1}}(S^{d-1}) & \rightarrow & L_2^{d\sigma_{d-1}}(S^{d-1}) \\ & \quad f & \rightarrow & \int_{S^{d-1}} K(x,\cdot) f(x) d\sigma_{d-1}(x) \end{array}$$

By continuity of K, $\int_{S^{d-1}} K(x,x) d\sigma_{d-1}(x)$ is finite and T_K is well defined, self-adjoint, positive semi-definite and trace-class (Smale and Zhou, 2007; Steinwart and Christmann, 2008).

Denote H the Reproducing Kernel Hilbert Space (RKHS) associated to K. The spectral theorem for compact operators (Kato, 1995) tells us that for $M \in \mathbb{N} \cup \{+\infty\}$, we have a positive, non-increasing summable sequence $(\eta_m)_{0 \leq m \leq M}$ and a family $(e_m)_{0 \leq m \leq M} \subset H$, such that $(\eta_m^{1/2} e_m)_{0 \leq m \leq M}$ is an orthonormal system in H while $(e_m)_{0 \leq m \leq M}$ is an orthonormal system in $L_2^{d\sigma_{d-1}}(S^{d-1})$ with

$$T_K = \sum_{m=0}^{M} \eta_m \langle ., e_m \rangle e_m \ .$$

where $\langle \cdot, \cdot \rangle$ is in $L_2^{d\sigma_{d-1}}(S^{d-1})$. The system of eigenfunctions of T_K is particularly interesting, yet often unknown analytically, except for special classes of kernels. Our class of kernels is one of them.

Spherical harmonics. Let $P_m(d)$ be the space of homogeneous polynomials of degree m in d variables with real coefficients and $\mathcal{H}_m(d)$ be the space of harmonics polynomials defined by

$$\mathcal{H}_m(d) := \{P \in P_m(d) | \Delta P = 0\}$$

where $\Delta \cdot = \sum_{i=1}^{d} \frac{\partial^2 \cdot}{\partial x_i^2}$ is the Laplace operator on $\mathbb{R}^d$ (Wendland, 2005). Define $H_m(S^{d-1})$ the space of real spherical harmonics of degree m defined as the set of restrictions of harmonic polynomials in $\mathcal{H}_m(d)$ to S^{d-1}. Let also $L_2^{d\sigma_{d-1}}(S^{d-1})$ be the space of (real) square-integrable functions on the sphere S^{d-1} endowed with its induced Lebesgue measure $d\sigma_{d-1}$ and $|S^{d-1}|$ the surface area of S^{d-1}. $L_2^{d\sigma_{d-1}}(S^{d-1})$ endowed with its natural inner product is a separable Hilbert space and the family of spaces $(H_m(S^{d-1}))_{m \geq 0}$, yields a direct sum decomposition (Efthimiou and Frye, 2014) that reads as

$$L_2^{d\sigma_{d-1}}(S^{d-1}) = \bigoplus_{m \geq 0} H_m(S^{d-1}) \tag{3}$$

which means that the summands are closed and pairwise orthogonal. Moreover, each $H_m(S^{d-1})$ has a finite dimension $\alpha_{m,d}$ with $\alpha_{0,d} = 1$, $\alpha_{1,d} = d$ and for $m \geq 2$

$$\alpha_{m,d} = \binom{d-1+m}{m} - \binom{d-1+m-2}{m-2}$$

Therefore for all $m \geq 0$, given any orthonormal basis of $H_m(S^{d-1})$, $(Y_m^1, ..., Y_m^{\alpha_{m,d}})$, we can build an Hilbertian basis of $L_2^{d\sigma_{d-1}}(S^{d-1})$ by concatenating these orthonormal bases. Let us denote in the following $(Y_m^{l_m})_{m,l_m}$ such an Hilbertian basis of $L_2^{d\sigma_{d-1}}(S^{d-1})$.

Azevedo and Menegatto (2014) give a *Mercer decomposition* for a dot product kernel on the sphere of the form (2). Indeed each spherical harmonics of degree

m, $Y_m \in H_m(S^{d-1})$, is an eigenfunction of T_K with associated eigenvalue given by the formula

$$\lambda_m = \frac{|S^{d-2}|\Gamma((d-1)/2)}{2^{m+1}} \sum_{s\geq 0} b_{2s+m} \frac{(2s+m)!}{(2s)!} \frac{\Gamma(s+1/2)}{\Gamma(s+m+d/2)} \,. \tag{4}$$

Mercer's theorem then states that the RKHS H associated to the kernel K is the set of functions $f \in L_2(S^{d-1})$ satisfying

$$f = \sum_{\substack{m\geq 0\\ \lambda_m>0}} \sum_{l_m=1}^{\alpha_{m,d}} a_{m,l_m} Y_m^{l_m} \quad \text{s.t.} \quad \sum_{\substack{m\geq 0\\ \lambda_m>0}} \sum_{l_m=1}^{\alpha_{m,d}} \frac{a_{m,l_m}^2}{\lambda_m} < +\infty \,. \tag{5}$$

From this definition, we see immediately that as the eigenvalues of the integral operator decreases slower, the volume of the RKHS becomes larger. More generally, the eigendecay of the integral operator is central to the understanding of a kernel. Note that, in general, the rate of convergence of a sub-sequence of positive $(\lambda_m)_{m\geq 0}$, ranked in the non-increasing order, is different from the one of $(\eta_m)_{0\leq m\leq M}$. Indeed we need to take into account the *eigenvalue multiplicities* in order to control the eigendecay. A control of eigendecay is usually out of reach, except for specific kernels on specific domains; see (Steinwart and Christmann, 2008) for a survey and an extended discussion. Indeed upper bounding or lower bounding can quickly result in such loose bounds that they are trivial bounds. A careful control of eigenvalues and their multiplicities is essential.

Eigendecay regimes. We distinguish three regimes of decay of eigenvalues: polynomial, geometric, and super-geometric. A polynomial decay corresponds to a rate proportional to m^{-q} with $q > 1$; geometric decay to one proportional to $\exp(-\alpha m^q)$ with $\alpha > 0$ and $q > 0$; super-geometric decay to one faster to geometric decay. We shall see that, depending on the behavior of the coefficients $(b_m)_{m\geq 0}$, dot product kernels relate to one of the above three regimes. In Table 1, we give an overview of dot product kernels on the sphere (Blanchard and Zwald, 2008; Zhang et al., 2016; Bach, 2017a,b) and give the rates of the sequence $(\lambda_m)_{m\geq 0}$ defined in Eq. (4). We also recall the eigendecay for classical kernels from the nonparametric statistics literature (Gu, 2013).

2 Eigenvalue Decay of Dot Product Kernels on the Sphere

We show now how to control the eigenvalue decay of an integral operator associated with a dot product kernel K on the sphere introduced in Eq. 2. We exhibit three regimes: polynomial, geometric and super-geometric. We can be in one or the other regime, depending on the coefficients $(b_m)_{m\geq 0}$ involved. Recall that for such kernels we have an explicit formulation of the eigenvalues $(\lambda_m)_{m\geq 0}$ associated to the integral operator T_K given by (4). In the following we denote $(\eta_m)_{0\leq m\leq M}$ the positive eigenvalues of the integral operator T_K associated to the kernel K ranked in a non-increasing order with their multiplicities, where $M \in \mathbb{N} \cup \{+\infty\}$.

Super-Geometric Decay. A first case of interest is the one studied by Azevedo and Menegatto (2014). There tight estimates for eigenvalues $(\lambda_m)_{m\geq 0}$ are obtained, under the assumption that $|b_m/b_{m-1}| \in O(m^{-\delta})$ when δ is assumed to be strictly bigger than $1/2$. We present here a more general result, holding for any $\delta > 0$.

Proposition 2.1. *If there exists $\delta > 0$ such that*

$$\left|\frac{b_m}{b_{m-1}}\right| \in O(m^{-\delta}) \tag{6}$$

then, denoting $\alpha = 1/(1-2\delta)$, we have

$$\lambda_m \in \begin{cases} \mathcal{O}\left(\frac{b_m}{2^m m^{(d-2)/2}}\right) & \textit{if } \delta \geq 1/2 \\[2ex] \mathcal{O}\left(\frac{m^{\frac{m\delta}{2\alpha}+\frac{1}{\alpha}} b_m}{2^{m+1} m^{(d-2)/2}}\right) & \textit{if } 0 < \delta < 1/2 \end{cases}$$

To control the eigenvalue decay associated with such dot product kernels, one needs to take into account the eigenvalue multiplicities. From the above control, we obtain a tight control of the eigenvalue decay of T_K ranked a non-increasing order with their multiplicities.

Proposition 2.2. *Under the same assumption as Prop. 2.1, $M = +\infty$ and there exists a universal constant $c > 0$ such that*

$$\eta_m \in \mathcal{O}\left(m^{-\frac{\delta}{s} m^{\frac{1}{d-1}}}\right) \textit{ where } s = \frac{4c}{(d-2)!} \,.$$

Geometric Decay. Another case of interest is when the coefficients $(b_m)_{m\geq 0}$ decrease almost geometrically. Indeed we also obtain a tight control of the sequence $(\lambda_m)_{m\geq 0}$ associated and the eigenvalue decay with their multiplicities of the integral operator T_K.

Proposition 2.3. *If there exist $0 < r < 1$ and $0 < c_2 \leq c_1$ constants such that for all $m \geq 0$*

$$c_2 r^m \leq b_m \leq c_1 r^m \,, \tag{7}$$

b_m	μ_m^ν	$\mathrm{df}_\nu(\lambda)$	Rates $(2 \geq \beta > 1)$
$b_m \in \mathcal{O}(m^{-\alpha})$, $\alpha > 1$	$m^{-\left(\frac{d/2+\alpha-3/2}{d-1}\right)}$	$\lambda^{-\frac{d-1}{d/2+\alpha-3/2}}$	$\ell^{-\frac{\beta}{\beta+q(\alpha,d)}}$, $q(\alpha,d) := \frac{d-1}{d/2+\alpha-3/2}$
$b_m \in \mathcal{O}(r^{-m})$, $1 > r > 0$	$e^{-\frac{(d-1)!}{Q_1}\log(1/r)m^{\frac{1}{d-1}}}$	$\log(\lambda^{-1})^{d-1}$	$\frac{\log(\ell)^{d-1}}{\ell}$
$\lvert b_m/b_{m-1}\rvert \in O(m^{-\delta})$, $\delta > 0$	$m^{-\frac{\delta}{s}m^{\frac{1}{d-1}}}$	$\frac{\log(\lambda^{-1})^{d-1}}{(\log(\log(\lambda^{-1})))^{d-1}}$	$\frac{\log(\ell)^{d-1}}{[\log(\log(\ell))]^{d-1}\ell}$

Table 2: Comparison of the convergence rate of regularized least-squares with a dot product kernel on the sphere.

then there exists constants $C_1, C_2 > 0$ such that

$$C_2\left(\frac{r}{4}\right)^m \leq \lambda_m \leq C_1 r^m \ .$$

Moreover, $M = +\infty$ and there exists universal constants $Q_1 > Q_2 > 0$ such that for all $m \geq 0$

$$C_2 e^{-\frac{(d-1)!}{Q_2}\log(4/r)m^{\frac{1}{d-1}}} \leq \eta_m \leq C_1 e^{-\frac{(d-1)!}{Q_1}\log(1/r)m^{\frac{1}{d-1}}} \ .$$

Polynomial Decay. When $(b_m)_{m\geq 0}$ admits a polynomial decay, we manage to control the rate of the sequence $(\lambda_m)_{m\geq 0}$ associated and the eigenvalue decay with their multiplicities of the integral operator T_K.

Proposition 2.4. *If there exists $\alpha > 1$ such that*

$$b_m \in \mathcal{O}(m^{-\alpha}) \ , \tag{8}$$

then we have

$$\lambda_m \in \mathcal{O}(m^{-d/2-\alpha+3/2}) \ ,$$

and

$$\eta_m \in \mathcal{O}(m^{-d/(2d-2)-\alpha/(d-1)+3/(2d-2)}) \ .$$

Approximation of the RKHS. The eigenvalue decay of the integral operator gives here a concrete notion of the complexity of the function space considered. Roughly speaking, if the $(\eta_m)_{m\geq 0}$ decay rapidly, the kernel K can be well approximated with a small number of terms in the Mercer decomposition. More formally, let (S, d) a metric space, $M \subset S$ and $\epsilon > 0$. The ϵ-covering number of M with respect to the metric d denoted $\mathbf{N}(\epsilon, M, d)$ is the smallest number of elements of an ϵ-cover for M using the metric d. The n-th entropy number of a set M for $n \in \mathbb{N}$ is defined as

$$\varepsilon_n(M) := \inf\{\epsilon\colon\ \mathbf{N}(\epsilon, M, d) \leq n\} \ .$$

Let $\mathcal{L}(E, F)$ be the set of all bounded linear operators T between the normed spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$. The entropy numbers of an operator $T \in \mathcal{L}(E, F)$ are defined as

$$\varepsilon_n(T) := \varepsilon_n(T(B_E))$$

where B_E is the closed unit ball of E. Obtaining a control of $\varepsilon_n(T_K)$ leads to a control of the generalization error of the kernel-based method using the kernel K (Smola et al., 2001). Smola et al. (1999) obtained a control of such quantities when the integral operator associated with the kernel has a polynomial or geometric eigendecay regime. Combining this with our results, we can obtain a control of the entropy numbers associated with dot product kernels.

Corollary 2.1. *Let $1 > r > 0$ and $\alpha > 1$. We have*

$$b_m \in \mathcal{O}(m^{-\alpha}) \implies \varepsilon_n(T_K) \in \mathcal{O}(\log^{-p(\alpha,d)/2}(n))$$

$$\textit{where } p(\alpha, d) = \frac{d/2 + \alpha - 3/2}{d-1} \ ,$$

Furthermore we have

$$b_m \in \mathcal{O}(r^m) \implies |\log(\varepsilon_n(T_K))| \in \mathcal{O}(\log^{1/d}(n)) \tag{9}$$

Recall that for a compact set M in finite dimensional space of dimension d the entropy number is $\varepsilon_n(M) \in \mathcal{O}(n^{-1/d})$. What (9) tells us is that a nonparametric estimator with that function class basically behaves like an estimator defined on a finite-dimensional space. To obtain statistical bounds, all that is left is to substitute the above control into the classical uniform convergence results (Boucheron et al., 2005; Steinwart and Christmann, 2008). In the next section, we focus on regularized least-squares (RLS) with dot product kernels, and, leveraging the eigendecay estimates we obtained in the previous section, we parameterize the statistical bounds in terms of the *effective dimension.*

3 Statistical Bounds for RLS with Dot-product Kernels

We present here general statistical bounds on the performance of regularized least-squares estimator of dot product kernels in all the regimes. These statistical bounds can be used to describe the statistical performance of a regularized least-squares estimator when this estimator can be computed exactly in practice. This applies for instance to the kernel-based deep networks developed by Shankar et al. (2020) and to kernel-based methods with kernels on spheres (Steinwart and

Christmann, 2008). We focus on the approximation error (our results do not assume realizability) and statistical prediction (our results match minimax rates) of regularized least-squares (RLS).

Learning from data. Given a dataset $\mathbf{z} = (x_i, y_i)_{i=1}^{\ell}$ independently sampled from an unknown distribution $\rho(x, y)$ on $Z := \mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$, the goal of the least-squares regression is to estimate the conditional mean function $f_\rho : \mathcal{X} \to \mathbb{R}$ given by $f_\rho(x) := \mathbb{E}(Y|X = x)$. The joint distribution $\rho(x, y)$, the marginal distribution ν, and the conditional distribution $\rho(.|x)$, are related through $\rho(x, y) = \nu(x)\rho(y|x)$. Consider as hypothesis space a Hilbert space H of functions $f : \mathcal{X} \to \mathcal{Y}$. For any regularization parameter $\lambda > 0$ and training set $\mathbf{z} \in Z^\ell$, the regularized least-squares estimator $f_{H,\mathbf{z},\lambda}$ is the solution of

$$\operatorname*{argmin}_{f \in H} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_H^2 \right\} . \tag{10}$$

In the following, the input space $\mathcal{X}$ is the sphere S^{d-1} and the hypothesis space considered is the Hilbert space H associated with the dot product kernel K with coefficients $(b_m)_{m \geq 0}$. Define the integral operator on $L_2^{d\nu}(S^{d-1})$ associated as $T_\nu(f)(y) = \int_{S^{d-1}} K(x, y) f(x) d\nu(x)$ and denote $(\mu_m^\nu)_{0 \leq m \leq M}$ its positive eigenvalues ranked in a non-increasing order with their multiplicities, where $M \in \mathbb{N} \cup \{+\infty\}$. The analysis of the convergence rates of RLS relies on the control of the effective dimension defined as

$$\mathrm{df}_\nu(\lambda) := \mathrm{Tr}\left((T_\nu + \lambda)^{-1} T_\nu\right) = \sum_{m=0}^{M} \frac{\mu_m^\nu}{\mu_m^\nu + \lambda} .$$

In the following, we manage to obtain tight estimates of the df_ν when $M = +\infty$ and $(\mu_m^\nu)_{0 \leq m \leq M}$ has a geometric decay or a super-geometric one. Note that Caponnetto and De Vito (2007) previously obtained such a control in the polynomial decay regime. Applying these controls to the results obtained in the previous section allows us to deduce the convergence rates of RLS for dot product kernels in all the regimes. Table 2 summarizes the control of the quantities of interest as well as the convergence rates obtained for RLS associated to dot product kernels in the different regimes.

We work here under general assumptions on the set of probability measures ρ on $S^{d-1} \times \mathcal{Y}$.

Assumptions 3.1. ***[Probability measures on $S^{d-1} \times \mathcal{Y}$].*** *Let $\mathcal{P}$ a set of probability measures on S^{d-1}. Furthermore, let $B, B_\infty, L, \sigma > 0$ be some constants and $0 < \beta \leq 2$ a parameter. Then we denote by $\mathcal{F}_{B,B_\infty,L,\sigma,\beta}(\mathcal{P})$ the set of all probability measures ρ on $S^{d-1} \times \mathcal{Y}$ with the following properties.*
(i) $\nu \in \mathcal{P}$
(ii)$\int_{S^{d-1}\times\mathcal{Y}} y^2 d\rho(x, y) < \infty$, $\|f_\rho\|^2_{L_\infty^{d\nu}} \leq B_\infty$
(iii) There exists $g \in L_2^{d\nu}(S^{d-1})$ such that $f_\rho = T_\nu^{\beta/2} g$ and $\|g\|_\rho^2 \leq B$
(iv) there exist $\sigma > 0$ and $L > 0$ such that $\int_{\mathcal{Y}} |y - f_\rho(x)|^m d\rho(y|x) \leq \frac{1}{2} m! \sigma^2 L^{m-2}$.

For $\omega \geq 1$, we denote by $\mathcal{W}_\omega$ the set of all probability measures ν on S^{d-1} which satisfying $d\nu/d\sigma_{d-1} < \omega$. Furthermore, we introduce for a constant $\omega \geq 1 > h > 0$, $\mathcal{W}_{\omega,h} \subset \mathcal{W}_\omega$ the set of probability measures ν on S^{d-1} which additionally satisfy $d\nu/d\sigma_{d-1} > h$. In the following we denote $\mathcal{G}_{\omega,\beta} := \mathcal{F}_{H_N,B,B_\infty,L,\sigma,\beta}(\mathcal{W}_\omega)$ and $\mathcal{G}_{\omega,h,\beta} := \mathcal{F}_{H_N,B,B_\infty,L,\sigma,\beta}(\mathcal{W}_{\omega,h})$.

Geometric Case We consider the case corresponding to a geometric eigendecay. Here the coefficients $(b_m)_{m \geq 0}$ in the Taylor decomposition decrease almost geometrically. The first goal is to obtain a control the of the effective dimension associated with the integral operator T_ν.

Proposition 3.1. *Let $\omega > 0$ and $\nu \in \mathcal{W}_\omega$. If there exists $0 < r < 1$ such that*

$$b_m \in \mathcal{O}(r^m) , \tag{11}$$

Then there exists a constant $Q > 0$ such that all $0 < \lambda \leq e^{-1}$ we have

$$\mathrm{df}_\nu(\lambda) \leq Q \log(\lambda^{-1})^{d-1}$$

From the above control, we are now able to show the convergence rates for nonparametric regression in the geometric regime.

Theorem 3.1. *Let us assume that there exists $0 < r < 1$ such that the sequence $(b_m)_{m \geq 0}$ satisfies:*

$$b_m \in \mathcal{O}(r^m) \tag{12}$$

Let also $w \geq 1$ and $0 < \beta \leq 2$. Then there exists a constant $C > 0$ independent of β such that for any $\rho \in \mathcal{G}_{\omega,\beta}$ and $\tau \geq 1$ we have:

- *If $\beta > 1$, then there exists $\ell_{\tau,\beta} > 0$ such that for all $\ell \geq \ell_\tau$ and $\lambda_\ell = \frac{1}{\ell^{1/\beta}}$, with a ρ^ℓ-probability $\geq 1 - e^{-4\tau}$ it holds*

$$\|f_{H_N,\mathbf{z},\lambda} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^{d-1}}{\ell}$$

- *If $\beta = 1$, then there exists $\ell_\tau > 0$ such that for all $\ell \geq \ell_\tau$ and $\lambda_\ell = \frac{\log(\ell)^\mu}{\ell}$, $\mu > d - 1 > 0$, with a ρ^ℓ-probability $\geq 1 - e^{-4\tau}$ it holds*

$$\|f_{H_N,\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^\mu}{\ell}$$

- *If $\beta < 1$, then there exists $\ell_{\tau,\beta} > 0$ such that for all $\ell \geq \ell_\tau$ and $\lambda_\ell = \frac{\log(\ell)^{\frac{d-1}{\beta}}}{\ell}$, with a ρ^ℓ-probability $\geq 1 - e^{-4\tau}$ it holds*

$$\|f_{H_N,\mathbf{z},\lambda_\ell} - f_\rho\|_\rho^2 \leq 3C\tau^2 \frac{\log(\ell)^{d-1}}{\ell^\beta}$$

Note that we have for any $0 < \beta \leq 2$ an explicit formulation of $\ell_{\tau,\beta}$ which depend to the constants of the problem, τ and β but we decide to hide them to simplify the exposition of the results. Moreover the rates obtained for RLS are optimal in the minimax sense and therefore no better rate can be obtained within this nonparametric learning framework.

Super-Geometric Case. Let us now consider the case corresponding to a super-geometric eigendecay. As in the geometric case, we start by obtaining a control of $\mathrm{df}_\nu(\lambda)$ associated with T_ν in this regime.

Proposition 3.2. *Let $\omega > 0$ and $\nu \in \mathcal{W}_\omega$. If there exist $0 < \delta < 1$ such that*

$$\left|\frac{b_m}{b_{m-1}}\right| \in \mathcal{O}(m^{-\delta})$$

Then there exists a constant $Q > 0$ such that all $0 < \lambda \leq e^{-1}$ we have:

$$df_\nu(\lambda) \leq Q \frac{\log(\lambda^{-1})^{d-1}}{(\log(\log(\lambda^{-1})))^{d-1}}$$

From the above control, we also obtain the convergence rates for nonparametric regression in the super-geometric regime. Table 2 shows the rates obtained in that regime when $1 < \beta \leq 2$.

Polynomial Case. Caponnetto and De Vito (2007) obtained the optimal convergence rates of RLS under the *assumption* that the eigenvalue of the integral operator T_ν admits a polynomial decay for a given kernel K. Combining their results and the one obtained in Prop. 2.4 gives the convergence rates of RLS with dot product kernels in the polynomial regime. See Table 2 for rates obtained. As expected, the convergence rate becomes faster as the complexity of the model shrinks, *i.e.*, the convergence rate of the super-geometric regime is faster than the one obtained in the geometric regime; the latter rate is therefore faster than the one in the polynomial regime.

3.1 Numerical illustrations

In Figure 1, we compare the theoretical rates of RLS estimator with the actual ones in the different regimes. We use here a similar setup to the one of Bietti and Bach (2021, Sec. 4). In each regime, we consider a specific dot-product kernel. More precisely, for the polynomial, geometric and super-geometric regimes, the kernels considered are respectively, $k(x, y) = \exp(-c\|x - y\|)$, $k(x, y) = (2 - \langle x, y\rangle)^{-1}$ and $k(x, y) = \exp(-c\|x - y\|^2)$. To compare the rates, we consider randomly sampled inputs on the unit sphere S^3 in 4 dimensions, and generate outputs according to a target function living in the associated RKHS. The regularization parameter of RLS is chosen according the theoretical rules given in the paper. The actual performance (red curve) is computed on 10,000 test datapoints. The x-axis corresponds to the number of training datapoints. The blue curve corresponds to the theoretical upper rates obtained in our paper. We see that the theoretical rates we obtain match (up to a constant factor) the actual rates of RLS when the number of datapoints is sufficiently large.

4 Examples related to deep nets

We give two other applications of the theoretical results from the previous sections related to multi-layer perceptrons (MLP). Before introducing the applications, Let us first recall the definition of an MLP.

Multi-layer perceptrons. We refer to here as a multi-layer perceptron a fully-connected deep neural network (Shalev-Shwartz and Ben-David, 2014). Let $\mathcal{X}$ the input space be a subset of $\mathbb{R}^d$, N the number of hidden layers, $\boldsymbol{\sigma} := (\sigma_k)_{k=1}^N$ a sequence of nonlinear activation functions and $\mathbf{m} := (m_k)_{k=1}^N$ a sequence of integers corresponding to the width of the hidden layers. Let us also introduce the width m_0 of the input layer which is just the dimension of the input, and m_{N+1} which is the width of the ouput layer supposed to be 1 here. Then any function defined by a MLP is parameterized by weight matrices $\mathbf{W} := (W^k)_{k=1}^{N+1}$ where $W^k \in \mathbb{R}^{m_{k-1}\times m_k}$ and can be recovered as follows. Let $x \in \mathcal{X}$, define $\mathcal{N}^0(x) := x$ and for $k \in \{1, \dots, N\}$, denote $W^k := (w_1^k, ..., w_{m_k}^k)$ where for all $j \in \{1, \dots, m_k\}$ $w_j^k \in \mathbb{R}^{m_{k-1}}$. Then, for all $k \in \{1, ..., N\}$, define the k^{th} layer as

$$\mathcal{N}^k(x) := (\sigma_k(\langle\mathcal{N}^{k-1}(x), w_1^k\rangle), ..., \sigma_k(\langle\mathcal{N}^{k-1}(x), w_{m_k}^k\rangle))$$

Finally the function associated to the MLP with weights $\mathbf{W}$ is defined as $\mathcal{N}(x, \mathbf{W}) := \langle\mathcal{N}^N(x), W^{N+1}\rangle_{\mathbb{R}^{m_N}}$. We shall denote $\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma},\mathbf{m}}$ the function space defined by all functions $\mathcal{N}(\cdot, \mathbf{W})$ defined as above on $\mathcal{X}$ for any choice of $\mathbf{W}$. We shall also consider the union space

$$\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma}} := \bigcup_{\mathbf{m}\in\mathbb{N}_*{}^N} \mathcal{F}_{\mathcal{X},\boldsymbol{\sigma},\mathbf{m}}.$$

We assume in the following that the input data is on the unit sphere ($\mathcal{X} = S^{d-1}$) which is a common assumption in the literature (Elad, 2010).

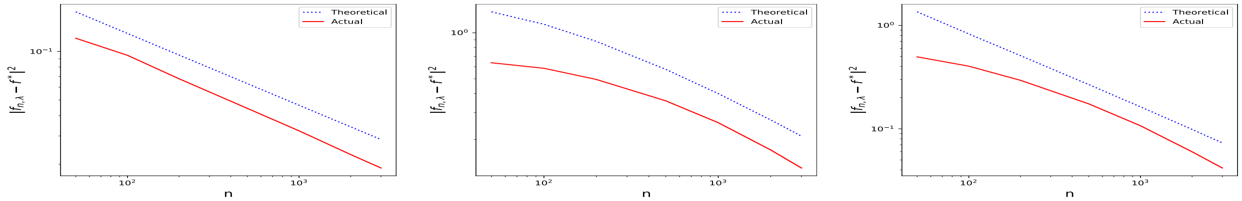

Figure 1: Comparison of the actual rates with the theoretical ones of the regularized least-squares estimator in the three different regimes. *Left:* Polynomial case. *Middle:* Geometric case. *Right:* Super-geometric case.

Neural Tangent Kernels. Learning the weights of a network using gradient methods results in a non-convex problem. However, in a specific over-parameterized regime, it may be shown that gradient descent can reach a global minimum while keeping weights very close to random initialization. More precisely, for a network $\mathcal{N}(x, \mathbf{W})$ initialized with $\mathbf{W}_0$, learning in the infinitely width regime is then equivalent to a kernel method with a specific kernel referred to as a neural tangent kernel (Chen et al., 2020) and defined as

$$K_{\text{NTK}}(x, x') := \lim_{\mathbf{m}\to+\infty} \langle \nabla_{\mathbf{W}}\mathcal{N}(x, \mathbf{W}_0), \nabla_{\mathbf{W}}\mathcal{N}(x', \mathbf{W}_0)\rangle.$$

Bietti and Mairal (2019b) show that, when the input space is the unit sphere, the neural tangent kernel associated to an MLP is a dot product kernel. More precisely, consider the case where for all $i \neq j$, $\sigma_i = \sigma_j$ for simplicity and denote σ the nonlinear activation considered. Moreover let $(a_i^{(1)})_{i\geq 0}$ the coefficients in the decomposition of σ in the basis of Hermite polynomials, $(a_i^{(0)})_{i\geq 0}$ the coefficients in the decomposition of the first-order derivative σ' of σ (assuming that σ is differentiable) in the basis of Hermite polynomials, and define $f_1(x) := \sum_{i\geq 0}(a_i^{(1)})^2 x^i$ and $f_2(x) := \sum_{i\geq 0}(a_i^{(0)})^2 x^i$. Then by defining $K_1^{\text{NTK}}(x) = K_1(x) = x$ and for all $i = 2, \ldots, N$,

$$\begin{aligned} K_i(x) &= f_1(K_{i-1}(x)) \\ K_i^{\text{NTK}}(x) &= K_{i-1}^{\text{NTK}}(x) f_0(K_{i-1}(x)) + K_i(x)\ , \end{aligned}$$

we obtain that $K_{\text{NTK}}(x, x') = K_N^{\text{NTK}}(\langle x, x'\rangle)$. Therefore $K_{\text{NTK}}(x, x')$ is a dot product kernel and our results from Sec. 2-3 can be applied. In particular, we can obtain estimates of the eigendecay of the integral operator associated with that kernel in all possible regimes of eigendecay. Such results can be applied for example to control the convergence the idealized gradient descent algorithm for a two-layer MLP. The convergence analysis of Cao et al. (2019, Th. 4.2) can be used. The convergence results suggests that the magnitude of the projected residuals is driven by the magnitude of the p_k-th eigenvalue of the integral operator associated with the NTK. Therefore, during training by gradient descent, a two-layer MLP with a large enough width learns the target function along the eigenfunctions of the integral operator associated with the NTK corresponding to the larger eigenvalues. Moreover this convergence is faster in the polynomial regime than in the geometric regime; and faster in the geometric regime than in the super-geometric regime.

Hilbertian Envelope of Smooth Multi-layer Perceptrons The mapping defined by a multi-layer perceptron can be embedded into an appropriate reproducing kernel Hilbert space with respect to the nonlinear activations involved in the network architecture. Moreover the *kernel induced by an MLP is a dot product kernel* of the form (2) where $(b_m)_{m\geq 0}$ is completely determined by the non linear activation functions $(\sigma_i)_{i=1}^N$ involved in the network. When the input space is the unit sphere S^{d-1} of $\mathbb{R}^d$ with $d \geq 2$, our results from Sec. 2-3 can be again applied now to the specific RHKS related to this network. Note that this RKHS is a different object than the one associated with a neural tangent kernel.

We show that there exists an RKHS containing the function space $\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma}}$ for any smooth activation functions $\boldsymbol{\sigma} := (\sigma_i)_{i=1}^N$. Moreover, for well chosen activation maps, the kernel is a *universal kernel* on $\mathcal{X}$ in the sense of Sriperumbudur et al. (2011). The universality property endows a kernel with interesting theoretical properties.

Proposition 4.1. *Let $\mathcal{X}$ be any subspace of $\mathbb{R}^d$, $N \geq 1$, $(\sigma_i)_{i=1}^N$ functions which admits a Taylor decomposition on $\mathbb{R}$. Moreover let $(f_i)_{i=1}^N$ be the sequence of functions such that for every $i \in \{1, ..., N\}$:*

$$f_i(x) = \sum_{n\geq 0} \frac{|\sigma_i^{(n)}(0)|}{n!} x^n \tag{13}$$

Then the RKHS H_N of the kernel, K_N defined on $\mathcal{X}\times\mathcal{X}$ by

$$K_N(x, x') := f_N \circ ... \circ f_1(\langle x, x'\rangle_{\mathbb{R}^d}) \tag{14}$$

contains the function space $\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma}}$. If we assume in addition that for every $i \in \{1, ..., N\}$ and $n \in \mathbb{N}$, $\sigma_i^{(n)}(0) \neq 0$, then the kernel K_N is cc-universal.

The RKHS H_N can be seen as an *Hilbertian envelope* of the function space $\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma}}$. Note that the RKHS we define above does *not* require that networks are infinitely wide *i.e.* that all layers of the network are infinitely large, as in some previous works (Daniely et al., 2016; Du et al., 2018). Indeed, for any number of weights $\mathbf{m} := (m_i)_{i=1}^N$, the function space $\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma},\mathbf{m}}$ lies inside the RKHS we have just defined. This is an important difference with previous works where RKHS constructions were used to approach the function spaces related to deep networks.

There are several consequences to the Proposition above. A direct consequence fact is that $\inf_{f\in H_N}\mathbb{E}[(f(X)-Y)^2] \leq \inf_{f\in\mathcal{F}_{\mathcal{X},\boldsymbol{\sigma}}}\mathbb{E}[(f(X)-Y)^2]$. In other words, the minimum expected risk in H_N is a straightforward lower bound on the minimum expected risk in F. A second consequence is that the kernel K_N associated with H_N defined above is universal. Therefore, Bayes-consistency holds for common loss functions and the Hilbert space embedding of probability distributions is injective under general assumptions (Steinwart and Christmann, 2008).

We would like to underscore that, contrary to a common misconception, many kinds of activations functions other than ReLU activation functions have been used with great success by practitioners in a number of applications; see (Eger et al., 2018) for a recent account.

Eigendecay and depth. Thanks to our results, when the input lies on the unit sphere, obtaining the eigendecay of the integral operator associated with a kernel, hence the shape and the volume of the RKHS enveloping the MLP function space, boils down to finding the rate of decay of the coefficients in the Taylor decomposition of the kernel. However, as one overlays layers over layers, iterating compositions of nonlinear functions on top of the dot product, the rate of decay of the coefficients $(b_m)_{m\geq 0}$ changes. For example, as one performs the composition of the exponential function $f_1 := \exp(x)$ with the square function $f_2 := x^2$ (yielding a two-layer network), we get $b_m^{(2)} = 2^m/m!$, while if we had considered only the exponential function (yielding a single-layer network) we would have got simply $b_m^{(1)} = 1/m!$. Generally, as one performs compositions of functions, each coefficient b_m increases hence λ_m increases, resulting in a growth of the RKHS.

Convergence Rates and Network Depth. In the geometric and super-geometric case, we can show that increasing the depth of the network does not affect the statistical rates as soon as the resulting kernel obeys the same regime. Indeed, in the geometric regime (resp. super-geometric), the statistical rates obtained in Sec. 3 do not depend on the parameter $0 < r < 1$ (resp. $\delta > 0$). Therefore while the resulting composed kernel still obeys the same regime, the statistical rates remain the same. For example, in the previous example we obtain that $b_m^{(2)} = 2^m/m!$, therefore $b_{m+1}^{(2)}/b_m^{(2)} = 2m$. Moreover we also have that $b_{m+1}^{(1)}/b_m^{(1)} = m$, therefore both are still in the same regime and the statistical rates for both networks are the same. The two observations above suggest that, from this viewpoint, increasing the depth of a network can increase the size of the target space, *i.e.*, the set of realizable functions, while the statistical rates appear to remain the same at least in the geometric and super-geometric regime.

What about ReLUs? As shown in (Daniely et al., 2016), a ReLU network with N layers can be approximated by the kernel K_N introduced in Prop. 4.1 where for all $i = 1,\dots,N$ $f_i(x) = g(x) := \frac{1}{\pi}(\pi - \arccos(x))$. To clearly make the distinction, note that here the function space generated by the ReLU network *is not included* in the RKHS built associated to K_N, whereas the function space generated by the smooth MLP defined with instead the activation function g at each layer is included. As arccos admits a Taylor decomposition and the coefficients admits a polynomial decay.

Bietti and Bach (2021) argue in a very recent work that, in that specific case of deep neural networks with ReLU activation functions, increasing the depth does not change the eigendecay of the associated integral operator. Our theoretical results encompass the polynomial regime of decay of eigenvalues that is characteristic of deep neural networks with ReLU activation functions. We can then obtain estimates of the effective dimension and statistical rates for regularized least-squares. Contrasting various viewpoints on ReLU networks is an interesting venue for future work (Ongie et al., 2020).

Conclusion. We have analyzed the eigenvalue decay of integral operators associated with dot product kernels on Euclidean spheres. Depending on the behavior of the coefficients in the Taylor series expansion of the kernel, we have distinguished three regimes of decay of eigenvalues: polynomial, geometric, and super-geometric. In each eigendecay regime, we have provided tight effective dimension estimates as well as learning rates for regularized least-squares. We have further illustrated our results through examples inspired from recent theoretical analyses of deep neural networks.

Acknowledgments

The authors would like to thank A. Rudi and L. Rosasco for pointing out relevant literature. We acknowledge support from NSF DMS 2023166, CCF 2019844, CIFAR "Learning in Machines and Brains" progam, and "Chaire d'excellence de l'IDEX Paris Saclay".

References

D. Azevedo and V. A. Menegatto. Sharp estimates for eigenvalues of integral operators generated by dot product kernels on the sphere. *Journal of Approximation Theory*, 177:57–68, 2014.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017b.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. In *9th International Conference on Learning Representations*, 2021.

Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20:25:1–25:49, 2019a.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 32, 2019b.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4): 971–1013, 2018.

Gilles Blanchard and Laurent Zwald. Finite-dimensional projection for classification and statistical learning. *IEEE transactions on information theory*, 54(9):4169–4182, 2008.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *CoRR*, abs/1912.01198, 2019.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368, 2007.

Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same RKHS. *CoRR*, abs/2009.10683, 2020.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13363–13373, 2020.

Youngmin Cho and Lawrence K. Saul. Large-margin classification in infinite neural networks. *Neural Computation*, 22(10):2678–2697, 2010.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

Ronald DeVore, Gerard Kerkyacharian, Dominique Picard, and Vladimir Temlyakov. Approximation methods for supervised learning. *Foundations of Computational Mathematics*, 6(1):3–58, 2006.

Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1339–1348. PMLR, 2018.

N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, editors. *Multivariate approximation and applications.* Cambridge UP, 2001.

Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions.* World Scientific Publishing, 2014.

Steffen Eger, Paul Youssef, and Iryna Gurevych. Is it time to swish? comparing deep learning activation functions across NLP tasks. In *Proc. EMNLP*. ACL, 2018.

P. P. B. Eggermont and V. N. LaRiccia. *Maximum penalized likelihood estimation. Vol. I.* Springer Series in Statistics. Springer-Verlag, New York, 2001.

Michael Elad. *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing.* Springer, 2010.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* The MIT Press, 2016.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer, 2013.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer series in statistics. Springer, 2002.

F. (Francis) Hirsch. *Elements of functional analysis.* Graduate texts in mathematics ; 192. Springer, New York, 1999.

Roger A Horn and Charles R Johnson. *Matrix analysis.* Cambridge university press, 2012.

Tosio Kato. *Perturbation theory for linear operators.* Classics in Mathematics. Springer-Verlag, Berlin, 1995.

Julien Mairal, Francis R. Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Found. Trends Comput. Graph. Vis.*, 8(2-3):85–283, 2014a.

Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in neural information processing systems*, pages 2627–2635, 2014b.

Youssef Mroueh, Stephen Voinea, and Tomaso A Poggio. Learning with group invariant features: A kernel perspective. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *8th International Conference on Learning Representations*, 2020.

Allan Pinkus. Strictly positive definite functions on a real inner product space. *Advances in Computational Mathematics*, 20(4):263–271, 2004.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017.

Saburou Saitoh. *Integral transforms, reproducing kernels and their applications*, volume 369. CRC Press, 1997.

Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. *CoRR*, abs/2002.12640, 2021.

I. J. Schoenberg. Positive definite functions on spheres. *Duke Math. J.*, 9:96–108, 1942. ISSN 0012-7094.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning. From theory to algorithms.* Cambridge UP, 2014.

Vaishaal Shankar, Alex Fang, Wenshuo Guo, Sara Fridovich-Keil, Jonathan Ragan-Kelley, Ludwig Schmidt, and Benjamin Recht. Neural kernels without tangents. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8614–8623. PMLR, 2020.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Alex J. Smola, Robert C. Williamson, Sebastian Mika, and Bernhard Schölkopf. Regularized principal manifolds. In Paul Fischer and Hans Ulrich Simon, editors, *COLT*. Springer, 1999.

Alex J. Smola, Zoltan L. Ovari, and Robert C. Williamson. Regularization with dot-product kernels. In *Advances in neural information processing systems*, pages 308–314, 2001.

Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer, 2008.

Taiji Suzuki. Fast generalization error bound of deep learning from a kernel perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1397–1406, 2018.

Holger Wendland. *Scattered data approximation.* Cambridge UP, 2005.

R. C. Williamson, A. J. Smola, and B. Scholkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, 2001.

Yuchen Zhang, Jason D Lee, and Michael I Jordan. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.

Yuchen Zhang, Percy Liang, and Martin J Wainwright. Convexified convolutional neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4044–4053, 2017.

Barbara Zwicknagl. Power series kernels. *Constructive Approximation*, 29(1):61–84, 2009.