Approximation beats concentration? An approximation view on inference with smooth radial kernels

Mikhail Belkin   MBELKIN@CSE.OHIO-STATE.EDU
Ohio State University, Department of Computer Science and Engineering, Department of Statistics

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Positive definite kernels and their associated Reproducing Kernel Hilbert Spaces provide a mathematically compelling and practically competitive framework for learning from data.

In this paper we take the approximation theory point of view to explore various aspects of smooth kernels related to their inferential properties. We analyze eigenvalue decay of kernels operators and matrices, properties of eigenfunctions/eigenvectors and "Fourier" coefficients of functions in the kernel space restricted to a discrete set of data points. We also investigate the fitting capacity of kernels, giving explicit bounds on the fat shattering dimension of the balls in Reproducing Kernel Hilbert spaces. Interestingly, the same properties that make kernels very effective approximators for functions in their "native" kernel space, also limit their capacity to represent arbitrary functions. We discuss various implications, including those for gradient descent type methods.

It is important to note that most of our bounds are measure independent. Moreover, at least in moderate dimension, the bounds for eigenvalues are much tighter than the bounds which can be obtained from the usual matrix concentration results. For example, we see that eigenvalues of kernel matrices show nearly exponential decay with constants depending only on the kernel and the domain. We call this "approximation beats concentration" phenomenon as even when the data are sampled from a probability distribution, some of their aspects are better understood in terms of approximation theory.

1. Introduction

Modern supervised machine learning is largely based on Empirical Risk Minimization (ERM), a form of functional approximation. Kernel machines perform variants of ERM over Reproducing Kernel Hilbert Spaces (RKHS). RKHS, also known as *native spaces* in the approximation literature, are generalizations of Sobolev spaces and have many attractive mathematical and computational properties. In particular, these spaces correspond to positive definite kernels, such as Gaussian, inverse multiquadrics, or Laplace kernels. Inference in these function spaces is analytically tractable, practically competitive and often leads to convex optimization problems, which can be viewed as linear methods in infinite dimensional Hilbert spaces.

In this paper we take a look at the properties of kernels from the approximation point of view. While there is an extensive and diverse literature on kernel methods and their use in machine learning (including the books Scholkopf and Smola (2001); Shawe-Taylor and Cristianini (2004); Steinwart and Christmann (2008)), we are aware of few works that use powerful results now available in the approximation theory literature, with the notable exception of Rieger and Zwicknagl (2010). There are a number of learning theory results

based on certain assumption about kernel eigenvalue decay. However, there are few analyses showing that specific practically used kernels, e.g., Gaussian kernels, satisfy these assumptions. Moreover, the exact nature of the dependence on the underlying measure has not, to the best of our knowledge, been addressed in the literature. We feel that the approximation point of view provides a rather different perspective on the properties of kernel methods, their strengths and limitations. In particular, we show the following:

1. Eigenvalues of smooth radial kernel operators/matrices decay at a nearly exponential rate[1] with constants depending only on the kernel and the dimension of the space and independent of the underlying data sample/measure. This also implies that kernel matrices/operators corresponding to smooth kernels are uniformly effectively low rank.

2. A function in the reproducing kernel space of a smooth kernel K written in the basis of eigenfunctions of K, must have nearly exponential coefficient decay in L_μ^2 for any measure μ. In particular, μ can be the empirical measure corresponding to a given dataset.

3. Eigenfunctions of a kernel matrix/operator can be nearly exponentially approximated by a linear combination of kernel functions. The span of eigenfunctions corresponding to the top eigenvalues is in a sense invariant to change of measure. Significantly, this is not true for individual eigenvectors, which are strongly influenced by the geometry of the underlying measure.

4. The fat shattering (V_γ) dimension of balls of radius R in the RKHS of smooth kernels is poly-logarithmic in $\frac{R}{\gamma}$. This limits the fitting power of any procedure whose output belongs to a ball of a polynomial size in the RKHS. In particular, this analysis applies to various regularization methods and to gradient descent-like algorithms with bounded step sizes.

5. While reducing the width of a kernel (such as a Gaussian) expands the function space, the RKHS corresponding to a wider kernel is contained in the RKHS of the narrow kernel. Thus combining radial kernels of different bandwidths is unlikely to yield results significantly different from simply using a single kernel with smaller width.

Our results use powerful approximation theory available for radial kernels. At least in moderate dimension they are significantly stronger than learning theory results not relying on these techniques. In particular, sample-independent nearly exponential decay for the eigenvalues of kernel matrices seems counter-intuitive in view of the matrix concentration results with rates of $O(\frac{1}{\sqrt{n}})$ (see the discussion in Section 3).

The paper is organized as follows: in Section 2 we collect some important background definitions and results on kernels, RKHS and approximation theory. In Section 3 we give results on low-rank approximations for kernels and consequences for eigenvalue decay of kernel operators/matrices. We proceed to discuss the "approximation beats concentration" phenomenon. In Section 4 we analyze "Fourier" coefficient decay of RKHS functions in the basis of eigenvectors. We proceed to show approximation properties for top eigenvectors and their spans. Section 5 gives bounds on the fat shattering dimensions for balls in RKHS and discusses implications for regularized kernel algorithms. In Section 6 we address the effect of kernel width. We conclude in Section 7.

---

1. By nearly exponential we will mean a function of the form $O(\exp(-Cn^{-\alpha}))$, where $C, \alpha > 0$.

2. Spaces and operators associated to positive definite kernels

We first establish some background facts about RKHS, kernels and approximation theory needed for further development. We recommend Wendland (2004) for a comprehensive introduction to the subject. Let Ω be a domain[2] in $\mathbb{R}^d$. Let $K(x, z)$ be a positive definite kernel on $\mathbb{R}^d$. We will denote by $\mathcal{H}$ the Reproducing Kernel Hilbert Space (RKHS) corresponding to the kernel K. Given a probability measure μ on Ω, we can define the integral operator $\mathcal{K}_\mu : L^2_\mu \to L^2_\mu$:

$$\mathcal{K}_\mu f(z) = \int K(x, z) f(x) d\mu \tag{1}$$

It is easy to check that $\mathcal{K}_\mu$ is a self-adjoint operator on L^2_μ and is compact when the kernel $K(\cdot, \cdot)$ is continuous. Notice that while $f \in L^2_\mu$ needs to be defined only on the support of the measure μ, Eq. 1 defines $\mathcal{K}_\mu f$ everywhere on $\mathbb{R}^d$. We will often consider the case when μ is supported on a finite set of points, so the difference between the support of μ and the domain of definition of $\mathcal{K}_\mu f$ is significant. Moreover, it can be shown that the $\mathcal{K}_\mu f \in \mathcal{H}$ for any $f \in L^2_\mu$. Notice also that a function $f \in \mathcal{H}$ gives rise to a function in L^2_μ by restricting it to the support of μ. We will call the restriction operator $\mathcal{R}_\mu : \mathcal{H} \to L^2_\mu$. Thus we will suppress $\mathcal{R}_\mu$, where no ambiguity arises. For example, for $f \in \mathcal{H}$, we will write $\|f\|_{L^2_\mu} := \|\mathcal{R}_\mu f\|_{L^2_\mu}$. Note that $\mathcal{R}_\mu$ does not change the function values at any point, just the domain of the definition and the function space norm. It can be shown by an extension of the argument in Wendland (2004) (Proposition 10.28) that $\mathcal{R}_\mu$ is the adjoint of the kernel operator $\mathcal{K}_\mu : L^2_\mu \to \mathcal{H}$. Specifically, for $f \in \mathcal{H}, g \in L^2_\mu$ we have

$$\langle f, \mathcal{K}_\mu g \rangle_\mathcal{H} = \langle \mathcal{R}_\mu f, g \rangle_{L^2_\mu} \tag{2}$$

Moreover, it turns out that that the square root $\mathcal{K}_\mu^{1/2}$ exists and is an isometric embedding of $L^2_\mu \to \mathcal{H}$. Specifically, for $f, g \in L^2_\mu$ we have:

$$\langle \mathcal{K}_\mu^{1/2} f, \mathcal{K}_\mu^{1/2} g \rangle_\mathcal{H} = \langle f, g \rangle_{L^2_\mu}$$

Kernel matrices. Given a set of points $X = \{x_1, \ldots, x_n\} \subset \Omega$ we can construct the corresponding kernel matrix K_n, $(K_n)_{ij} = \frac{1}{n} K(x_i, x_j)$. Note that K_n can be viewed as a special case of $\mathcal{K}_\mu$, where μ is a uniform discrete measure on X, $\mu = \frac{1}{n} \sum \delta_{x_i}$.

Eigenfunctions/Nystrom extension. The eigenfunctions of the operator $\mathcal{K}_\mu$ are defined by the equation

$$\int K(x, z) e(x) d\mu_x = \lambda e(z), \quad \lambda \in \mathbb{R}$$

Note that $e(z)$ is technically an element of L^2_μ. However, as the image of $\mathcal{K}_\mu$ is actually in $\mathcal{H}$, we can define $e(z)$ in $\mathcal{H}$, and, indeed, on all of $\mathbb{R}^d$:

$$e(z) = \frac{1}{\lambda} \int K(x, z) e(x) d\mu_x$$

This formula is known as the *Nystrom extension*.

---

2. Ω can be taken to be a unit cube or a more general bounded domain.

3

Remark 1 *Note that technically there are two objects corresponding to $e(z)$, $e_{\mathcal{H}}(z) \in \mathcal{H}$ and $e_{L_\mu^2}(z) \in L_\mu^2$. These functions coincide on the support of the measure μ, $e_{\mathcal{H}}(z) = e_{L_\mu^2}(z), z \in \mathrm{supp}(\mu)$, but have different norms in their respective spaces. More precisely (cf., e.g., the discussion in* Rosasco et al. (2010)*), $\mathcal{R}_\mu e_{\mathcal{H}} = e_{L_\mu^2}$, $\quad \frac{1}{\lambda}\mathcal{K}_\mu e_{L_\mu^2} = e_{\mathcal{H}}$. Overloading the notation, we will simply write $e(x)$, while keeping note of the norms.*

The Nystrom extension allows us to directly compare eigenfunctions of operators with different measures μ, potentially with disjoint support. We can compare eigenvectors of different kernel matrices by comparing the Nystrom extensions of the corresponding operators.

The spectral decomposition for kernel operators. We will now establish some properties of eigenfunctions of kernel operators. Let e and e' be two orthogonal eigenfunctions of $\mathcal{K}_\mu$ in L_μ^2. Note that since $\mathcal{K}_\mu$ is self-adjoint any two eigenfunctions corresponding to different eigenvalues are orthogonal. Additionally, as K is a positive definite kernel, there are no eigenfunctions with eigenvalue 0. Using Eq.2 We have

$$\langle \mathcal{K}_\mu e, \mathcal{K}_\mu e' \rangle_{\mathcal{H}} = \langle \mathcal{R}_\mu \mathcal{K}_\mu e, e' \rangle_{L_\mu^2} = \lambda \langle e, e' \rangle_{L_\mu^2} = 0 \tag{3}$$

Hence we see that the Nystrom extensions of eigenfunctions orthogonal in L_μ^2 are also orthogonal in $\mathcal{H}$.

The spectral theorem for compact self-adjoint operators (see, e.g., Reed and Simon (1980), Theorem VI.16) guarantees the existence of an orthogonal Hilbert space basis $e_1, e_2, \ldots$ of eigenfunctions of $\mathcal{K}_\mu$ in L_μ^2. This basis is finite if μ is supported on a finite set (and L_μ^2 is finite-dimensional) and infinite otherwise. The discussion above shows that the basis $e_1, e_2, \ldots$ extends to an orthogonal (possibly partial) basis in $\mathcal{H}$.

Interpolant operators and the fill. Given a set $X = \{x_1, \ldots, x_n\} \subset \Omega$ and a RKHS $\mathcal{H}$ with a positive definite kernel K, we can construct the *interpolation operator* $\mathcal{S}_X : \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{S}_X(f) = \arg \min_{g \in \mathcal{H}, \, g(x_i) = f(x_i)} \|g\|_{\mathcal{H}}$$

Setting K_n to be the (positive definite) kernel matrix corresponding to X, there is an explicit formula in terms of the inverse of K_n:

$$\mathcal{S}_X(f)(x) = \sum \alpha_i K(x_i, x), \text{ where } (\alpha_1, \ldots, \alpha_n)^t = K_n^{-1}(f(x_1), \ldots, f(x_n))^t \tag{4}$$

From Eq. 4 it is clear that $\mathcal{S}_X$ is a linear operator, and it can be easily verified that $\mathcal{S}_X(f)(x_i) = f(x_i)$. In fact, $\mathcal{S}_X$ is an orthogonal projection operator, which maps $\mathcal{H}$ to the (finite-dimensional) orthogonal complement to the space of functions vanishing on all points of X.

Another important concept associated to the set X, is the *fill h_X*, which describes how well the set X covers Ω, $h_X = \max_{x \in \Omega} \min_{x_i \in X} \|x - x_i\|$.

Notation for the norm. Given that we will deal with several functional spaces at once and that "same" operators have different norms depending on the range and the domain, we will use the "$\to$" notation. For example, $\|\mathcal{S}_X\|_{\mathcal{H} \to L_\mu^p} := \sup_{f \in \mathcal{H}, \, f \neq 0} \frac{\|\mathcal{S}_X(f)\|_{L_\mu^p}}{\|f\|_{\mathcal{H}}}$ denotes the norm of $\mathcal{S}_X$ as a map from $\mathcal{H}$ to L_μ^2. Note that the operator norm can be defined for any, even *non-linear*, map between two normed spaces in the same way.

4

Approximation Theory. We will now state the key result from the approximation theory which provides a bound on the difference $f - \mathcal{S}_X(f)$ in terms of the fill h_X.

Let $K(x, z)$ be a smooth radial kernel. Specifically, let $K(x, z) = \phi(\|x - y\|)$ and put $f(\cdot) := \phi(\sqrt{\cdot})$. We require that $|f^{(l)}(r)| \leq l!M^l$ for all l large enough and $r > 0$.

Two types of important kernels satisfying these conditions are Gaussian kernels $K(x, z) = \exp\left(-\|x - z\|^2/\sigma^2\right)$ and inverse multiquadric kernels $K(x, z) = (c^2 + \|x - z\|^2)^{-\alpha}, \alpha > 0$ (for the popular Cauchy kernel, $\alpha = 1$).

Theorem A *Under the conditions on the kernel stated above, for any set $X \subset \Omega$ and any $p \in [1, \infty]$ there exist constants $C, C' > 0$, such that*

$$\|\mathcal{R}_\mu - \mathcal{S}_X\|_{\mathcal{H} \to L_\mu^p} < C' \exp(-C/h_X)$$

Theorem A is a special case (and a slight reformulation) of Theorem 11.22 in Wendland (2004) (see also Rieger and Zwicknagl (2010), Theorem 6.1). This is a powerful approximation theory result showing that any function in $\mathcal{H}$ can be accurately reconstructed from its values at a small number of points. The implications are wide-ranging and, perhaps, surprising in their scope.

Remark 2 (Gaussian kernels) *The bound in Theorem A can be made slightly tighter for Gaussian kernels, to be of the form $C' \exp(C \log(h_X)/h_X)$ (see Wendland (2004)). The extra logarithmic factor does not substantially change our discussion and we will not treat this case separately. This leads to slightly looser but more general bounds.*

3. Low rank approximations to kernel operators and their eigenvalues.

We start by showing that any (potentially non-linear) bounded map $\mathcal{T}$ from a Hilbert space[3] to an RKHS $\mathcal{H}$ corresponding to a smooth radial kernel, satisfying the conditions of Theorem A, allows a universal low-rank approximation in L_μ^p. That is, the output of any such map is close to a low-dimensional subspace in $\mathcal{H}$ according to the norm in L_μ^p. This subspace only depends on $\mathcal{H}$ and is independent of μ and $\mathcal{T}$.

Theorem 3 *Suppose $\mathcal{T} : \mathbf{V} \to \mathcal{H}$ is a (not necessarily linear) map from a Hilbert (Banach) space $\mathbf{V}$ to a RKHS of functions on $\mathbb{R}^d$, $\mathcal{H}$. Then there exists a map $\mathcal{T}_n$ from $\mathbf{V}$ to an n-dimensional linear subspace $\mathcal{H}_n \subset \mathcal{H}$, such that*

$$\|\mathcal{T} - \mathcal{T}_n\|_{\mathbf{V} \to L_\mu^p} < C'\|\mathcal{T}\|_{\mathbf{V} \to \mathcal{H}} \exp(-Cn^{-1/d})$$

for some constants $C, C' > 0$ independent of $\mathcal{T}$ and μ. Moreover:
(1) While the map $\mathcal{T}_n$ depends on $\mathcal{T}$, the subspace $\mathcal{H}_n$ is independent of $\mathcal{T}$.
(2) If $\mathcal{T}$ is a linear operator, $\mathcal{T}_n$ is also a linear operator.

Proof Let $X = (x_1, \cdots, x_n)$ be a finite subset of Ω. Notice that $\mathcal{T} - \mathcal{S}_X \circ \mathcal{T} = (\mathcal{R}_\mu - \mathcal{S}_X) \circ \mathcal{T}$. Using the definition of the norm we see that

$$\|(\mathcal{R}_\mu - \mathcal{S}_X) \circ \mathcal{T}\|_{\mathbf{V} \to L_\mu^p} \leq \|\mathcal{R}_\mu - \mathcal{S}_X\|_{\mathcal{H} \to L_\mu^p} \|\mathcal{T}\|_{\mathbf{V} \to \mathcal{H}}.$$

---

3. A Banach space can be used as well.

Applying Theorem A we obtain $\|\mathcal{T} - \mathcal{S}_X\mathcal{T}\|_{\mathbf{V}\to L_\mu^p} < C'\|\mathcal{T}\|_{\mathbf{V}\to\mathcal{H}}\exp(-C/h_X)$ for some constants $C, C' > 0$. Choosing the set X appropriately (e.g., a d-dimensional grid), we can ensure that $h_X = O(n^{-1/d})$, where d is the dimension of the space. Notice that the image of S_X belongs to a n-dimensional subspace of $\mathcal{H}$ spanned by the functions $K(x_i, \cdot), x_i \in X$. Taking $\mathcal{H}_n = \text{span}\{K(x_1, \cdot), \ldots K(x_n, \cdot)\}$ and $\mathcal{T}_n = \mathcal{S}_X \circ \mathcal{T}$ completes the proof. ∎

Thus the image of a bounded operator to $\mathcal{H}$ can be nearly exponentially approximated by a finite-dimensional subspace independent of $\mathcal{T}$ and μ. To provide a bound on the eigenvalues of kernel operators and matrices we will need the following perturbation result:

Lemma 4 *Suppose $\mathcal{A}$ is a self-adjoint operator on a Hilbert space $\mathbf{V}$, $\mathcal{A} : \mathbf{V} \to \mathbf{V}$ and $\mathcal{A}_n$ is a finite rank operator with rank n, such that $\|\mathcal{A} - \mathcal{A}_n\| < \epsilon$. Then all eigenvalues of $\mathcal{A}$ except for at most n (counting multiplicity) are smaller than ϵ.*

Proof As rank $\mathcal{A}_n = n$, every $n + 1$ dimensional subspace contains a non-zero vector v, such that $\mathcal{A}_n v = 0$. Suppose $\mathcal{A}$ has at least $n + 1$ linearly independent eigenvectors with eigenvalues $\geq \epsilon$. Then there exists a vector $v \neq 0$, in the span of these eigenvectors, such that $\mathcal{A}_n v = 0$. As $\mathcal{A}$ is self-adjoint we can assume that these eigenvectors are orthogonal and hence it is easily seen that $\|\mathcal{A}v\| > \epsilon\|v\|$. We see that $\|\mathcal{A}_n v\| = \|(\mathcal{A} + \mathcal{A}_n - \mathcal{A})v\| > \|\mathcal{A}v\| - \epsilon\|v\| > 0$, which is a contradiction. ∎

We can now apply Theorem 3 to easily obtain a bound on eigenvalues of kernel matrices and operators. Important related work includes Schaback and Wendland (2002); Santin and Schaback (2016), which deal with approximation of spectral properties of integral operators with uniform measure. In contrast, we provide a measure-independent bound, which is key in our learning-theoretic context.

Theorem 5 (Eigenvalue decay) *Let $\kappa = \sup_{x\in\Omega} K(x, x)$. Then for some $C, C' > 0$.*

$$\lambda_i(\mathcal{K}_\mu) \leq \sqrt{\kappa}\, C' \exp(-C i^{1/d}) \tag{5}$$

Proof Consider $\mathcal{K}_\mu$ as an operator from $L_\mu^2 \to \mathcal{H}$. Recall from Section 2 that there exists a basis of eigenfunctions of $\mathcal{K}_\mu$ in L_μ^2 which is also orthogonal in $\mathcal{H}$. Let e_1 be an eigenfunction of $\mathcal{K}_\mu$ with the largest eigenvalue. We have

$$\|\mathcal{K}_\mu\|_{L_\mu^2\to\mathcal{H}} = \frac{\|\mathcal{K}_\mu e_1\|_\mathcal{H}}{\|e_1\|_{L_\mu^2}} = \frac{\lambda_1\|e_1\|_\mathcal{H}}{\|e_1\|_{L_\mu^2}}.$$

Recall now that $\mathcal{K}_\mu$ is adjoint to the restriction operator $\mathcal{R}_\mu : \mathcal{H} \to L_\mu^2$. We have $\|e_1\|_\mathcal{H}^2 = \frac{1}{\lambda_1}\langle e_1, \mathcal{K}_\mu e_1\rangle_\mathcal{H} = \frac{1}{\lambda_1}\|e_1\|_{L_\mu^2}^2$. Hence $\|\mathcal{K}_\mu\|_{L_\mu^2\to\mathcal{H}} = \frac{\lambda_1}{\sqrt{\lambda_1}} = \sqrt{\lambda_1} \leq \sqrt{\text{Tr}(\mathcal{K}_\mu)} \leq \sqrt{\kappa}$.

Applying Theorem 3 to $\mathcal{K}_\mu$, we have $\|\mathcal{K}_\mu - \mathcal{T}_n\|_{L_\mu^2\to L_\mu^2} < C'\sqrt{\kappa}\exp(-Cn^{-1/d})$, where $\mathcal{T}_n = \mathcal{S}_X \circ \mathcal{K}_\mu$ is a linear operator of rank n. Noticing that $\mathcal{K}_\mu : L_\mu^2 \to L_\mu^2$ is self-adjoint, we can apply Lemma 4. That completes the proof. ∎

Remark 6 *Notice that all quantities in the inequality in Theorem 5 are independent of the measure μ. In particular when μ is a finite measure, $\mathcal{K}_\mu$ can be viewed as a matrix. Hence this result provides a uniform bound on the eigenvalue decay, independent of the size of the matrix.*

Approximation beats concentration. Suppose X is a set of n points sampled iid from a probability distribution μ on Ω. Let μ_n denote the empirical measure associated to X. Concentration results for matrices (e.g., Tropp et al. (2015)) combined with spectral perturbation results (e.g., Rosasco et al. (2010)) in the bounds for eigenvalues of the form

$$|\lambda_{\mu,i} - \lambda_{\mu_n,i}| \leq \frac{C}{\sqrt{n}}$$

where C is a constant independent of i. In comparison, from Eq. 5 we see that

$$|\lambda_{\mu,i} - \lambda_{\mu_n,i}| \leq \max(\lambda_{\mu,i}, \lambda_{\mu_n,i}) \leq C' \exp(-C'' i^{1/d})$$

We see that approximation "beats" concentration by providing a tighter bound as long as $O(\exp(-C'' i^{1/d})) < O(\frac{1}{\sqrt{n}})$. In other words we need n to be nearly exponential in the eigenvalue index i for the concentration bounds to be tighter. Moreover, unlike concentration results, approximation also shows that the corresponding eigenvalues must actually be nearly exponentially close to 0. In addition, these approximation-based bounds are measure-independent and do not require any iid-type assumption[4].

This suggests that significant care should be taken when applying concentration-type analyses of kernel methods in the iid setting, as essential inferential properties may become invisible in these statistical analyses. We conjecture that this is one of the reasons why concentration bounds often turn out to be too pessimistic in practice.

4. Spectral characterization of RKHS functions and eigenfunctions of kernel operators.

We will now use Theorem 5 to provide a spectral characterization of RKHS functions in terms of their restrictions to finite (or infinite) sets. This characterization should be viewed as parallel to the classical description of Sobolev spaces in terms of their Fourier coefficients. In particular, we will see that the "Fourier" coefficients of a function from $\mathcal{H}$ written in the basis of eigenvectors of *any* kernel matrix, regardless of the dataset, must show nearly exponential decay with coefficient independent of the measure. This is significant as in many regression/classification problems we can compute these coefficients explicitly from the data. The decay of the coefficients can thus be analyzed empirically.

Theorem 7 (Coefficient decay for functions in RKHS)   *Let $f \in \mathcal{H}$ and consider the restriction of f, $\mathcal{R}_\mu f \in L_\mu^2$. Write the spectral decomposition of $\mathcal{R}_\mu f$ in terms of the eigenfunctions e_i of $\mathcal{K}_\mu$ as*

$$\mathcal{R}_\mu f = \sum a_i e_i, \ a_i = \langle \mathcal{R}_\mu f, e_i \rangle_{L_\mu^2}$$

Then

$$|a_i| \leq \sqrt{\lambda_i} \|f\|_{\mathcal{H}} < C' \exp\left(-C i^{1/d}\right) \|f\|_{\mathcal{H}}$$

for some $C, C' > 0$ independent of μ.

---

4. On the other hand, strong approximation bounds are specific to the smooth kernel setting while concentration results can be applied to a broad class of random matrix problems. Moreover, unlike approximation, concentration results are often dimension-independent.

Proof Recalling Eq.2, showing that operator $\mathcal{K}_\mu$ is adjoint to $\mathcal{R}_\mu$, we have $\langle f, e_i \rangle_{\mathcal{H}} = \langle f, \frac{1}{\lambda_i} \mathcal{K}_\mu e_i \rangle_{\mathcal{H}} = \frac{1}{\lambda_i} \langle \mathcal{R}_\mu f, e_i \rangle_{L_\mu^2}$ and

$$\langle \mathcal{R}_\mu f, e_i \rangle_{L_\mu^2} \leq \lambda_i \langle f, e_i \rangle_{\mathcal{H}} \leq \lambda_i \|e_i\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$$

Notice that
$$\|e_i\|_{\mathcal{H}}^2 = \frac{1}{\lambda_i} \langle e_i, e_i \rangle_{L_\mu^2} = \frac{1}{\lambda_i}$$

Hence by Theorem 5

$$|a_i| = |\langle \mathcal{R}_\mu f, e_i \rangle_{L_\mu^2}| \leq \lambda_i \frac{1}{\sqrt{\lambda_i}} \|f\|_{\mathcal{H}} = \sqrt{\lambda_i} \|f\|_{\mathcal{H}} < \kappa^{1/4} C' \exp(-Ci^{1/d}) \|f\|_{\mathcal{H}}$$

for some constants $C, C' > 0$ independent of the measure μ. ■

Properties of eigenfunctions. We will now proceed with some basic properties of eigenfunctions which follow from the analysis above. The first observation is that any eigenfunction of $\mathcal{K}_\mu$ with a sufficiently large eigenvalue is well-approximated by the span $K(x_i, \cdot), x_i \in X, i = 1, \ldots, n$, where X is chosen as in Theorem 3. Specifically,

Corollary 8 *Let $\lambda e = \mathcal{K}_\mu e$ be an eigenfunction of $\mathcal{K}_\mu$. Then there is a function $e_X \in$ span$\{K(x_1, \cdot), \ldots K(x_n, \cdot)\}$, such that*

$$\|e - e_X\|_{L_\mu^p} \leq \frac{C}{\lambda} \exp(-C' n^{1/d})$$

for some universal constants $C, C' > 0$, and any $p \in [1, \infty]$.

Proof From Theorem 3 we obtain

$$\frac{1}{\lambda} \|\mathcal{K}_\mu e - \mathcal{S}_X \mathcal{K}_\mu e\|_{L_\mu^p} \leq \frac{C}{\lambda} \exp(-C' n^{1/d}).$$

Putting $e_X = \frac{1}{\lambda} \mathcal{S}_X \mathcal{K}_\mu e$ yields the result. ■

Remark 9 *Notice that by taking $p = \infty$ we can make the bound to be pointwise. For example, if μ is a finite set of "data" points, the approximation holds for every point uniformly over the choices of μ. Hence a particular μ is unimportant in this sense.*

Similarly, for two measures μ, ν, the top eigenfunctions/eigenvectors of $\mathcal{K}_\mu$ are nearly contained in the span of the top eigenfunctions of $\mathcal{K}_\nu$, when restricted to the support of ν.

Theorem 10 *Let $\lambda_\mu e_\mu = \mathcal{K}_\mu e_\mu$ be an eigenfunction of $\mathcal{K}_\mu$. Then there is a function $e \in$ span$\{e_1, \ldots, e_k\}$, of eigenfunctions of $\mathcal{K}_\nu$, such that for some constants $C, C' > 0$*

$$\|e_\nu - e\|_{L_\mu^2} \leq \frac{C}{\sqrt{\lambda_\nu}} k^{\frac{d-1}{d}} \exp(-C' k^{-1/d})$$

8

Proof First note that $\|e_\mu\|_\mathcal{H} = \frac{1}{\sqrt{\lambda_\mu}}$. Now write $e_\nu = \sum_{i=1}^\infty a_i e_i$, where e_i are eigefunctions of $\mathcal{K}_\nu$. Put $e = \sum_{i=1}^k a_i e_i$. By Theorem 7, we have for some C, C'

$$\|e_\nu - e\|_{L_\mu^2}^2 \leq \sum_{i=k+1}^\infty a_i^2 \leq C' \frac{1}{\lambda_\mu} \sum_{i=k+1}^\infty \exp(-Ci^{-1/d}).$$

The last sum can be estimated by noticing that

$$\sum_{i=k+1}^\infty \exp(-Ci^{1/d}) < \int_k^\infty \exp(-Cx^{1/d})\,dx = d \int_{k^{1/d}}^\infty e^{-Cz} z^{d-1}\,dz$$

Integrating by parts shows that for k sufficiently large, the last integral is of the order $O(k^{(d-1)/d}\exp(-Ck^{1/d}))$, which completes the proof. ∎

Remark 11 *It is interesting to note that top eigenvectors of $\mathcal{K}_\mu$ contain important information about the structure of the measure μ, e.g., its clustering structure (e.g., Shi et al. (2009)). However, the span of the top eigenvectors is relatively invariant to the measure. Theorem 10 shows that eigenfunctions of ν will not significantly "spill" onto eigenfunctions of μ with much smaller eigenvalues.*

5. The (low) fat shattering dimension of balls in RKHS and its algorithmic implications

Approximation-theoretic results are easily turned into bounds on the fat shattering dimension V_γ for balls in RKHS, which are significantly tighter than those found in the literature. Combining these bounds with some standard learning theory results, we immediately obtain generalization guarantees for a number of regularized kernel methods and algorithms including gradient descent with early stopping. We start by recalling the definition of the fat shattering dimension V_γ for a function class $\mathcal{F}$ (see, e.g., Alon et al. (1997)). We say that a set $x_1, \ldots, x_n$ is γ-shattered by functions from $\mathcal{F}$ if there exist $s_i \in \mathbb{R}, i = 1 \ldots, n$, such that for any assignments of signs $\sigma_i \in \{-1, 1\}$ there is a function f in $\mathcal{F}$, satisfying

$$f(x_i) > s_i + \gamma, \text{if } \sigma_i = 1$$
$$f(x_i) < s_i - \gamma, \text{if } \sigma_i = -1$$

$V_\gamma(\mathcal{F})$ is taken to be *the maximum cardinality* of a set of points that is γ-shattered by functions from $\mathcal{F}$. To clarify the role of s_i's, note that it is easy to verify that V_0-dimension[5] is simply the VC-dimension of the sets $\{(x, t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) < t\}$ in $\mathbb{R}^{d+1}$. Thus V_γ for $\gamma > 0$ is a more demanding version of the VC-dimension appropriate for analyzing real-valued functions.

Theorem 12 (V_γ-dimension of RKHS balls) *Let $\mathcal{B}_R := \{f \in \mathcal{H}, \|f\|_\mathcal{H} < R\}$ be a ball of radius $R > 0$ in $\mathcal{H}$. Under our standard assumptions on the kernel, we have*

$$V_\gamma(\mathcal{B}_R) < O\left(\log^d\left(\frac{R}{\gamma}\right)\right) \tag{6}$$

---

5. Note that 0-shatters and "shatters" are not exactly the same notion.

Proof Let $\mathcal{S}$ be a linear space of functions $\mathbb{R}^d \to \mathbb{R}$. We will say that $\mathcal{S}$ γ-approximates $\mathcal{B}_R$ in the L^∞ norm, if for any $f \in \mathcal{B}_R$, there exists $f_1 \in \mathcal{S}$, s.t. for any x, $|f(x) - f_1(x)| < \gamma$. Suppose we have $\mathcal{S}$ that γ-approximates $\mathcal{B}_R$. By replacing each $f \in \mathcal{B}_R$ with its approximation $f' \in \mathcal{S}$, we see that $\mathcal{S}$ will 0-shatter any set of points $x_1, \ldots, x_N$ whenever $\mathcal{B}_R$ γ-shatters these points.

Hence V_γ-dimension of $\mathcal{B}_R$ is bounded by the VC-dimension of the subgraph sets $\{(x, t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) < t\}, f \in \mathcal{S}$. If $\mathcal{S}$ is finite-dimensional, rewriting that inequality in a basis of $\mathcal{S}$, we see that these sets can be viewed in as half-planes in $\mathbb{R}^{\dim \mathcal{S}+1}$, passing through the origin in that space. It is well-known that VC-dimension of these is simply the dimension of the space which is $\dim \mathcal{S} + 1$. Thus we get $V_\gamma(B_R) \leq \dim \mathcal{S} + 1$. It remains to find a space $\mathcal{S}$, which γ approximates $\mathcal{B}_R$. From the Theorem 3 (with $p = \infty$ and $\mathbf{V} = \mathcal{H}$), we obtain $C'\|f\|_{\mathcal{H}} \exp(-Cn^{-1/d}) < \gamma$. Solving for n after substituting $\|f\|_H = R$ and taking $\mathcal{S} = \mathcal{H}_n$ completes the proof. ∎

Remark 13 *It should be noted that in contrast to V_γ dimension for $\gamma > 0$, V_0-dimension and, indeed, VC-dimension for the indicator functions $\text{sign}(f), f \in \mathcal{B}_R$ are infinite. This follows easily from the interpolating property of $\mathcal{H}$. Specifically, for any set $(x_i, y_i), y_i \in \{+1, -1\}$, there exists $f \in \mathcal{H}$, such that $f(x_i) = y_i$. Scaling this function by a scalar does not change the corresponding indicator functions but allows to make $\|f\|_{\mathcal{H}}$ arbitrarily small.*

At this point we should compare the bound in Theorem 12 to the literature. The paper Evgeniou and Pontil (1999), gives a bound of the form $V_\gamma(\mathcal{B}_R) = O\left(\frac{R^2}{\gamma^2}\right)$. While our bound is generally much tighter, the result from Evgeniou and Pontil (1999) is dimension independent and also applies to a broad class of RKHS. Bounds on the closely related notion of *covering numbers* for balls in RKHS corresponding to Gaussian kernels are given in Steinwart and Scovel (2007) (Theorem 3.1). However, the bounds there are still polynomial in $\frac{1}{\epsilon}$ (roughly corresponding to our γ). We are not aware of any poly-logarithmic results. Note that an alternative approach to obtain poly-logarithmic bounds for covering numbers similar to those for V_γ dimension in Eq. 6 would be to combine the results based on eigenvalues from Guo et al. (1999) with the eigenvalue bound in our Theorem 5. However, the direct proof is significantly simpler and, arguably, more informative.

Generalization bounds. Assume we are in a standard learning setting where $(x_i, y_i) \in \Omega \times \{-1, 1\}$ is a labeled dataset, L is a Lipshitz loss function and the data are chosen from a probability measure p on $\Omega \times \{-1, 1\}$. Suppose our (otherwise arbitrary) learning algorithm $\mathcal{A}$ outputs functions in $\mathcal{B}_R$. In that case our Theorem 12 together with Alon et al. (1997) immediately imply the following "universal" generalization bound for smooth radial kernels

Theorem 14 (Generalization for kernels) *Let f be the output of our learning algorithm $\mathcal{A}$. Then with high probability*

$$\frac{1}{n}\sum L(f(x_i), y_i) < E_p L(f(x), y) + O\left(\frac{\log^{d/2}(nR)}{\sqrt{n}}\right)$$

This bounds applies to most kernel-based learning algorithms as nearly all of them output a function in a certain ball of radius R. We discuss some of the implications below.

Algorithmic implications and the limitations of kernel methods. We have seen that the poly-logarithmic bound on the fat shattering dimension in Eq. 6 implies broad and strong generalization guarantees given in Theorem 14. The flip side of that is that even mild regularization (i.e., constraining R) imposes severe limitations on the fitting capacity of kernel methods, at least when the dimension d is not very high. [6].

For example, consider the popular "Tikhonov" regularizer of the form $\lambda \|f\|_{\mathcal{H}}^2$. It is easy to see that adding this term in conjunction with a bounded loss function implies that the output of the algorithm belongs to an RKHS ball with radius $O\left(\frac{1}{\sqrt{\lambda}}\right)$. Thus a constant increase to the fitting capacity of $\mathcal{B}_R$ requires a nearly exponential increase of R. That suggests choosing small values of λ, which is consistent with practice, where very small values of λ often produce best results[7].

Another important example is that of *gradient descent* for kernel methods. It is easy to see that in the kernel setting each step of gradient descent increases the norm by at most a constant depending only on the kernel and the loss function. Thus, t steps of gradient descent output a function with norm bounded by $O(t)$. As each step of gradient descent typically requires $O(n^2)$ computations (a matrix-vector multiplication), we see that $R = O(c/n^2)$, where c is the number of operations. Fixing γ (and omitting constants), we see that the dimension of the function space *reachable* by c computations is of the order of $\log^{d/2}(\frac{c}{n^2})$. It follows immediately that at least order of $n^2 e^{n^{2/d}}$ computations are needed to fit arbitrary functions on the data points with accuracy γ.

Note that for square loss, empirical minimization problem is simply matrix inversion, which can be done using only n^3 operations using, e.g., Gaussian elimination. We see that gradient descent compares unfavorably with matrix inversion in terms of computational complexity, when no assumption about function values are made. On the other hand, when the function is in RKHS, or more generally, has rapidly decaying coefficients in the "Fourier basis" of eigenfunctions, subcubic computational complexity can be demonstrated Yao et al. (2007); Raskutti et al. (2014). This functional *algorithmic reach* for gradient descent with smooth kernels is discussed in Ma and Belkin (2017) from the spectral decay point of view. Analyzing V_γ dimension, as we do here, clarifies that point and connects it to other standard capacity measures.

It appears that for many real datasets, gradient descent does require cubic or even super-cubic complexity. Indeed, it could hardly be expected that nature should co-operate by matching the decay of Fourier coefficients for class membership functions to that of kernels chosen primarily for computational reasons!

Remark 15 *Our results suggest that smooth kernels would struggle to fit labels assigned randomly to a set of points, as such a fit would generally require $O(n^2 e^{n^{2/d}})$ operations (aside from the issues of numerical accuracy). Indeed, empirically random assignment are difficult to fit using smooth kernels, while less smooth Laplace kernels fit random labels far more easily Belkin et al. (2018). Interestingly, ReLU neural networks appear to be similar to Laplace kernels, capable of fitting random labels with ease Zhang et al. (2016).*

---

6. While d is the ambient dimension, the effective dimensionality of the data can be much lower.

7. We note that even minimum norm interpolation (i.e., $\lambda = 0$) shows excellent generalization results Belkin et al. (2018). While this finding is compatible with our analysis, it is directly explained by it.

Remark 16 *Our bounds for V_γ dimension do not imply that the corresponding VC-dimension is small. Indeed, as noted above, VC-dimension of indicator functions from a ball in RKHS is infinite. However, a function with a small RKHS norm, corresponding to a random assignments of labels on a set of data points must take values which are exponentially small on all data points. We conjecture that despite their bounded norm, most of such functions are outside of the computational reach of polynomially many steps of gradient descent.*

6. Kernels of different width

We will now briefly discuss the influence of the width (or shape) parameter for kernels from the approximation point of view. It is intuitive that "narrow" kernels have better fitting capacity. In particular data can be trivially represented as a sum of δ-functions, which can be thought of as radial kernels of width zero. However, it is not apriori clear whether choosing a different kernel width can result in a significantly different function space. We will see that making the width of the kernel larger simply shrinks the corresponding RKHS space without adding any new functions. The proof relies on a quite simple Fourier domain description of RKHS. Despite its usefulness, this characterization does not seem to be widely known in the learning literature. While we will state the theorem for Gaussian kernels, it also applies to any radial kernel with Fourier transform that decays fast enough, with the precise condition clear from the proof (see Appendix A).

Theorem 17 *let $K_1(x, z) = \phi(\|x - z\|)$ be a Gaussian kernel (with ϕ a one-dimensional Gaussian) and let $K_2(x, z) = \phi(\|x - z\|/\sigma)$, where $0 < \sigma < 1$. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the corresponding RKHS. Then*
1. $\mathcal{H}_1 \subset \mathcal{H}_2$.
2. For any $R > 0$, the ball of radius R in $\mathcal{H}_1$, $\mathcal{B}_R(\mathcal{H}_1)$ is contained in $\mathcal{B}_{\sigma^{-d/2}R}(\mathcal{H}_2)$ the ball of radius $\frac{1}{\sigma^{d/2}}R$ in $\mathcal{H}_2$. On the other hand, $\mathcal{B}_R(\mathcal{H}_2) \not\subset \mathcal{H}_1$.

7. Conclusions

The main goal of this note is to bring the powerful tools of approximation theory to kernel learning. Approximation theory provides a different perspective on a number of important inferential problems, yielding results which are difficult to obtain using the more standard concentration-based analyses, and are sometimes much tighter. We have not tried to specify constants and their dependence on the parameters of the kernel, including the kernel width. This can be done explicitly, and is an important aspect of understanding kernel methods. Furthermore, fundamentally, we need to understand how these approximation-based results relate to the intrinsic dimensionality of the data. Finally, it is crucial to understand the interface between approximation and concentration. We believe that combining these modes of analysis, and better understanding the regimes where one of them becomes dominant, can yield significant further insight into kernel inference and, likely, other machine learning problems.

References

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4): 615–631, 1997.

M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. *ArXiv e-prints*, 2018.

Theodoros Evgeniou and Massimiliano Pontil. On the v-γ dimension for regression in reproducing kernel hilbert spaces. In *Algorithmic Learning Theory*, pages 106–117. Springer, 1999.

Ying Guo, Peter L Bartlett, John Shawe-Taylor, and Robert C Williamson. Covering numbers for support vector machines. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 267–277. ACM, 1999.

Siyuan Ma and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. *arXiv preprint arXiv:1703.10622, accepted to NIPS 2017*, 2017.

G. Raskutti, M. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *JMLR*, 15(1):335–366, 2014.

Michael Reed and Barry Simon. Functional analysis, vol. i, 1980.

Christian Rieger and Barbara Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32(1):103, 2010.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(Feb):905–934, 2010.

Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

Robert Schaback and Holger Wendland. Approximation by positive definite kernels. In *Advanced Problems in Constructive Approximation*, pages 203–222. Springer, 2002.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: eigenspace of convolution operators and clustering. *The Annals of Statistics*, 37, 6B:3960–3984, 2009.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, pages 575–607, 2007.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Holger Wendland. Scattered data approximation, 2004.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approx.*, 26(2):289–315, 2007.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016. URL http://arxiv.org/abs/1611.03530.

Appendix A. Proof of theorem 17

The result follows easily from the Fourier characterization of RKHS, see Wendland (2004)(Theorem 10.12). If $K(x,z) = \psi(x-z)$ is a translation-invariant kernel the RKHS norm of f in the RKHS $\mathcal{H}$ corresponding to K can be written as

$$\|f\|_{\mathcal{H}}^2 = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\mathcal{F}(f)(\omega)|^2}{\mathcal{F}(\psi)(\omega)} d\omega \tag{7}$$

Here $\mathcal{F}$ denotes the Fourier transform. Recall now that (from the scaling property of Fourier transform) $\mathcal{F}(\phi(x/\sigma))(\omega) = \sigma^d \mathcal{F}(\phi)(\sigma\omega)$. Since Fourier transform of a Gaussian is also a Gaussian and $\sigma < 1$, we obtain

$$\frac{\mathcal{F}(\phi(x/\sigma))(\omega)}{\mathcal{F}(\phi)(\omega)} = \frac{1}{\sigma^d}\frac{\mathcal{F}(\phi(x))(\sigma\omega)}{\mathcal{F}(\phi)(\omega)} \geq \frac{1}{\sigma^d}$$

Hence using Eq. 7, we see that
$$\|f\|_{\mathcal{H}_2} \leq \frac{1}{\sigma^{d/2}}\|f\|_{\mathcal{H}_1}$$

This completes the proof except for the claim that $\mathcal{B}_R(\mathcal{H}_2) \not\subset \mathcal{H}_1$, which is straightforward to see in the Fourier domain.