How rotational invariance of common kernels prevents generalization in high dimensions

Konstantin Donhauser [1] **Mingqi Wu** [1] **Fanny Yang** [1]

Abstract

Kernel ridge regression is well-known to achieve minimax optimal rates in low-dimensional settings. However, its behavior in high dimensions is much less understood. Recent work establishes consistency for high-dimensional kernel regression for a number of specific assumptions on the data distribution. In this paper, we show that in high dimensions, the rotational invariance property of commonly studied kernels (such as RBF, inner product kernels and fully-connected NTK of any depth) leads to inconsistent estimation unless the ground truth is a low-degree polynomial. Our lower bound on the generalization error holds for a wide range of distributions and kernels with different eigenvalue decays. This lower bound suggests that consistency results for kernel ridge regression in high dimensions generally require a more refined analysis that depends on the structure of the kernel beyond its eigenvalue decay.

1. Introduction

Traditional analysis establishes good generalization properties of kernel ridge regression when the dimension d is relatively small compared to the number of samples n. These minimax optimal and consistency results, however, do not reflect observations for modern datasets with large d close to n. High dimensional statistical theory (Vaart, 1998; Bühlmann and Van De Geer, 2011; Wainwright, 2019) aims to fill the gap and obtain bounds that are predictive for large d. One popular approach is to consider the high-dimensional asymptotic regime $n, d \to \infty, d/n^\beta \to \gamma$ resulting in bounds that often match the behavior of empirical observations for large finite dimension d quite well.

While recent work (Dobriban and Wager, 2018; Hastie et al., 2019; Bartlett et al., 2020) establishes explicit asymptotic upper bounds for the bias and variance for high-dimensional linear regression, the results for kernel regression are less conclusive in the regime $d/n^\beta \to c$ with $\beta \in (0, 1)$. In particular, even though several papers (Ghorbani et al., 2021; 2020; Liang et al., 2020a) show that the variance decreases with the dimensionality of the data, the bounds on the bias are inconclusive. On the one hand, Liang et al. (2020a) prove asymptotic consistency for ground truth functions with asymptotically bounded Hilbert norms for neural tangent kernels (NTK) and inner product (IP) kernels. In contrast, Ghorbani et al. (2021; 2020) show that for uniform distributions on the product of two spheres, consistency cannot be achieved *unless* the ground truth is a low-degree polynomial. This *polynomial approximation barrier* can also be observed for random feature and neural tangent regression (Mei and Montanari, 2019; Mei et al., 2021a; Ghorbani et al., 2021).

Notably, the two seemingly contradictory consistency results hold for different distributional settings and are based on vastly different proof techniques. While (Liang et al., 2020a) proves consistency for general input distributions including isotropic Gaussians, the lower bounds in the papers (Ghorbani et al., 2021; 2020) are limited to data that is uniformly sampled from the product of two spheres. Hence, it is a natural question to ask whether the polynomial approximation barrier is a more general phenomenon or restricted to the explicit settings studied in (Ghorbani et al., 2021; 2020). Concretely, this paper addresses the following question:

Can we overcome the polynomial approximation barrier when considering different high-dimensional input distributions, eigenvalue decay rates or scalings of the kernel function?

We unify previous distributional assumptions in one proof framework and thereby characterize how the rotational invariance property of common kernels induces a bias towards low-degree polynomials. Specifically, we show that the *polynomial approximation barrier* persists for

- a broad range of common rotationally invariant kernels such as radial basis functions (RBF) with vastly different eigenvalue decay rates, inner product kernels and NTK of any depth (Jacot et al., 2018).

[1] Department of Computer Science, ETH Zürich. Correspondence to: Konstantin Donhauser <donhausk@ethz.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

- general input distributions, including anisotropic Gaussians. The degree of the polynomial depends only on the distribution via $d_{\text{eff}} := \text{tr}(\Sigma_d)/\,||\Sigma_d||_{\text{op}}$ and not on the specific structure of Σ_d. In particular, we cover the distributions studied in previous related works (Ghorbani et al., 2021; 2020; Liang et al., 2020a;b; Liu et al., 2021).

- different scalings τ of the kernel function $k_\tau(x, x') = k(\frac{x}{\sqrt{\tau}}, \frac{x'}{\sqrt{\tau}})$ beyond the classical choice $\tau \asymp d_{\text{eff}}$.

As a result, this paper demonstrates that it is a general high-dimensional phenomenon that for rotational invariant kernels, kernel regression can only consistently learn low-degree polynomials. that the polynomial approximation barrier is a general high-dimensional phenomenon for rotationally invariant kernels; that is, kernel regression with such kernels can only consistently learn functions that are low-degree polynomials.

Rotationally invariant kernels are a natural choice if no prior information on the structure of the ground truth is available, as they treat all dimensions equally. Since our analysis covers a broad range of distributions, eigenvalue decays and different scalings, our results motivate future work to focus on the symmetries, or rather asymmetries, of the kernel incorporating prior knowledge on the structure of the high-dimensional problem (see (Arora et al., 2019; Shankar et al., 2020; Mei et al., 2021b)).

This paper is organized as follows. First of all, we show in Section 2.2 that the bounded norm assumption that previous consistency results (Liang et al., 2020a;b) rely on, is violated as $d \to \infty$ even for simple functions such as $f(x) = e_1^\top x$. We then introduce our generalized setting in Section 2 and present our main results in Section 3 where we show a lower bound on the bias that increases with the dimensionality of the data. Finally, in Section 4 we empirically illustrate how the bias dominates the risk in high dimensions and therefore limits the performance of kernel regression. As a result, we argue that it is crucial to incorporate structural prior knowledge of the ground truth function in high-dimensional kernel learning, even in the noiseless setting. We empirically verify this on real-world sparsely parameterized data.

2. Problem setting

In this section, we briefly introduce kernel regression estimators in reproducing kernel Hilbert spaces and subsequently state our assumptions on the kernel, data distribution and high-dimensional regime.

2.1. Kernel ridge regression

We consider nonparametric regression in a *reproducing kernel Hilbert space* (RKHS, see e.g. (Wahba, 1990; Smola and Schölkopf, 1998)) with functions on the domain $\mathcal{X} \subset \mathbb{R}^d$ induced by a positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. That is, for any set of input vectors $\{x_1, \cdots, x_m\}$ in $\mathcal{X}$, the empirical kernel matrix K with entries $K_{i,j} = k(x_i, x_j,)$ is positive semi-definite. We denote the corresponding inner product of the Hilbert space by $\langle ., . \rangle_k$ and the corresponding norm by $\|.\|_{\mathcal{H}} := \sqrt{\langle ., . \rangle_k}$.

We observe tuples of input vectors and response variables (x, y) with $x \in \mathcal{X}$ and $y \in \mathbb{R}$. Given n samples, we consier the ridge regression estimator

$$\hat{f}_\lambda = \underset{f \in \mathcal{H}}{\arg\min} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \qquad (1)$$

with $\lambda > 0$ and the minimum norm interpolator (also called the kernel *ridgeless* estimate)

$$\hat{f}_0 = \underset{f \in \mathcal{H}}{\arg\min}\, \|f\|_{\mathcal{H}} \text{ such that } \forall i : \; f(x_i) = y_i \qquad (2)$$

that can be obtained as the limit of the ridge estimate $\hat{f}_0 = \lim_{\lambda \to 0} \hat{f}_\lambda$. It is well-known that the ridge estimator can attain consistency as $n \to \infty$ for some sequence of λ such that $\frac{\lambda}{n} \to 0$. Motivated by the curiously good generalization properties of neural networks with zero training error, recently some works (Liang et al., 2020a;b; Ghorbani et al., 2021; 2020) have also analyzed the consistency behavior of ridge and ridgeless estimates in the high-dimensional regime.

For evaluation, we assume that the observations are i.i.d. samples from a joint distribution $(x_i, y_i)_{i=1}^n \sim \mathbb{P}_{XY}$ and refer to $f^\star(x) := \mathbb{E}[y \mid X = x]$ as the *ground truth* function that minimizes the population square loss $\mathbb{E}(Y - f(X))^2$. We evaluate the estimator using the population risk conditioned on the input data $\mathbf{X}$

$$\mathbf{R}(\hat{f}_\lambda) := \mathbb{E}_Y \|\hat{f}_\lambda - f^\star\|_{\mathcal{L}_2(\mathbb{P}_X)}^2 = \underbrace{\|\mathbb{E}_Y \hat{f}_\lambda - f^\star\|_{\mathcal{L}_2(\mathbb{P}_X)}^2}_{=: \text{ Bias } \mathbf{B}} + \underbrace{\mathbb{E}_Y \|\mathbb{E}_Y \hat{f}_\lambda - \hat{f}_\lambda\|_{\mathcal{L}_2(\mathbb{P}_X)}^2}_{=: \text{ Variance } \mathbf{V}},$$

where $\mathbb{E}_Y$ is the conditional expectation over the observations $y_i \sim \mathbb{P}(Y|X = x_i)$. In particular, when $y = f^\star(x) + \epsilon$, $\mathbb{E}_Y \hat{f}_\lambda$ is equivalent to the noiseless estimator with $\epsilon = 0$.

Note that consistency in terms of $\mathbf{R}(\hat{f}_\lambda) \to 0$ as $n \to \infty$ can only be reached if the bias vanishes. In this paper, we lower bound the bias $\mathbf{B}$ which, in turn, implies a lower bound on the risk and the inconsistency of the estimator. The theoretical results in Section 3 hold for both ridge regression and minimum norm interpolation.

2.2. Prior work on the consistency of kernel regression

For ridge regression estimates in RKHS, a rich body of work shows consistency and rate optimality when appropriately choosing the ridge parameter λ both in the non-asymptotic setting, e.g. (Caponnetto and De Vito, 2007), and the classical asymptotic setting, e.g. (Christmann et al., 2007), as $n \to \infty$ for fixed d.

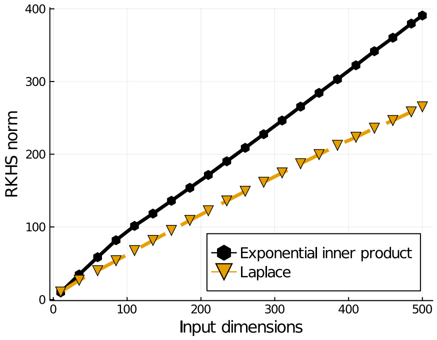


Figure 1: The approximation of the Hilbert norm induced by the Laplace and exponential inner product kernels of the function $f^\star(x) = x_1$ plotted with respect to the dimension d on the space $\mathcal{X} = [0,1]^d$. See Section 4.1 for experimental details.

Similar results have also been shown for high-dimensional asymptotics, where recent papers on minimum norm interpolation in kernel regression (Liang et al., 2020a;b) explicitly show how the bias vanishes as $d, n \to \infty$ when the ground truth function has bounded Hilbert norm. Even though this assumption is perfectly reasonable for a fixed ground truth and Hilbert space, its plausibility is less clear for a sequence of functions as $d \to \infty$.[1] After all, the Hilbert space and thus also the norm changes with d. In fact, we now show that even innocuous functions have diverging Hilbert norm as the dimension increases. The following lemma illustrates this phenomenon for tensor product kernels including exponential inner product kernels (also studied in (Liang et al., 2020a;b)) defined on $x, x' \in \mathcal{X}^{\otimes d} \subset \mathbb{R}^d$.

Lemma 2.1 (Informal)**.** *For any f that is a non-constant sparsely parameterized product function* $f(x) = \prod_{j=1}^m f_j(x_{(j)})$ *for some fixed* $m \in \mathbb{N}_+$,

$$\|f\|_{\mathcal{H}_d} \overset{d\to\infty}{\to} \infty.$$

In words, for simple sequences of sparse product functions, the Hilbert norm diverges as the dimension $d \to \infty$. The precise conditions on the kernel and sequence $\mathcal{H}_d$ of induced Hilbert spaces can be found in Appendix B. Figure 1 illustrates this phenomenon for $f^\star(x) = e_1^\top x$ for the Laplace and exponential inner product kernel.

The discussion so far implies that generalization upper bounds that rely on the bounded Hilbert norm assumption become void even for simple ground truth functions. A natural follow-up question is therefore: Does kernel regression actually fail to consistently learn sparsely parameterized functions or is it an artifact of the analysis in Liang et al. (2020a;b). A recent line of work by Ghorbani et al. (2021; 2020) shows that kernel ridge regression estimates can indeed only consistently learn polynomials of degree at most $\frac{\log n}{\log d_{\text{eff}}}$ as $d, n \to \infty$ (which we refer to as the polynomial approximation barrier). While the results provide some intuition for the behavior of kernel regression, the proofs heavily rely on significant simplifications that only hold for the specific distributional assumptions on the sphere.

In the next sections, we use a different proof technique to show that the polynomial approximation barrier indeed holds for a broad spectrum of data distributions that also capture the distributions studied in the papers (Liang et al., 2020a;b) as well as for an entire range of eigenvalue decay rates of the kernel functions (e.g. polynomial and exponential decay rates) and choices of the scaling τ. As a consequence, our results suggest that the polynomial approximation barrier is strongly tied to the rotational invariance of the kernel function and not specific to uniform distributions on the sphere.

2.3. Our problem setting

The framework we study in this paper covers random vectors that are generated from a covariance matrix model, i.e. $X = \Sigma_d^{1/2} W$ with vector W consisting of i.i.d entries and their projections onto the $d-1$-dimensional unit sphere. We now specify all relevant assumptions on the kernel and the data distribution.

Kernels Throughout this paper, we focus on continuous and rotationally invariant kernels. They include a vast majority of the commonly used kernels such as fully connected NTK, RBF and inner product kernels as they only depend on the squared Euclidean norm of x and x' and the inner product $x^\top x'$. We first present one set of assumptions on the kernel functions that are sufficient for our main results.

(A.1) *Rotational invariance and local power series expansion*: The kernel function k is rotationally invariant and there is a function g such that $k(x.x') = g(\|x\|_2^2, \|x'\|_2^2, x^\top x')$. Furthermore, g can be expanded as a power series of the form

$$g(\|x\|_2^2, \|x'\|_2^2, x^\top x') = \sum_{j=0}^{\infty} g_j(\|x\|_2^2, \|x'\|_2^2)(x^\top x')^j \tag{3}$$

that converges for x, x' in a neighborhood of the sphere $\{x \in \mathbb{R}^d \mid \|x\|^2 \in [1-\delta, 1+\delta]\}$ for some $\delta > 0$. Furthermore, all g_i are positive semi-definite kernels.

(A.2) *Restricted Lipschitz continuity:* The restriction of k on $\{(x,x) | x \in \mathbb{R}^d, \|x\|_2^2 \in [1-\delta_L, 1+\delta_L]\}$ is a Lipschitz continuous function for some constant $\delta_L > 0$.

Beyond Assumptions A.1,A.2 our main results hold in fact for a broader range of commonly studied kernels in practice (see Corollary 3.2). In particular, Theorem 3.1 also holds

[1] Although Liu et al. (2021) replace the bounded Hilbert norm assumption with a weaker bounded source condition, we expect this condition not to hold with increasing dimension either. We defer the detailed discussion to future work.

for α-exponential kernels defined as $k(x,x') = \exp(-\|x-x'\|_2^\alpha)$ for $\alpha \in (0,2)$ even though we could not yet show that they satisfy Assumptions A.1-A.2. In Appendix C, we show how the proof of Theorem 3.1 crucially relies on the rotational invariance Assumption C.1 and the fact that the eigenvalues of the kernel matrix K are asymptotically lower bounded by a positive constant. Both conditions are also satisfied by α-exponential kernels and the separate treatment is purely due to the different proof technique used to lower bound the kernel matrix eigenvalues.

Data distribution and scaling We impose the following assumptions on the data distribution.

(B.1) *Covariance model*: We assume that the input data distribution is from one of the following sets

$$\mathcal{Q} = \{\mathbb{P}_X \mid X = \Sigma_d^{\frac{1}{2}} W \text{ with } \forall i : W_{(i)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}, \mathbb{P} \in \mathcal{W}\}$$
$$\mathcal{Q}_{\mathcal{S}^{d-1}} = \{\mathbb{P}_X \mid X = \sqrt{d_{\text{eff}}}\frac{Z}{\|Z\|} \text{ with } Z \sim \mathbb{P} \in \mathcal{Q}\}$$

where $\Sigma_d \in \mathbb{R}^{d\times d}$ is a positive semi-definite covariance matrix and the effective dimension d_{eff} is defined as $d_{\text{eff}} := \text{tr}(\Sigma_d)/\ ||\Sigma_d||_{\text{op}}$. The entries of the random vector W are sampled i.i.d. from a distribution in the set $\mathcal{W}$, containing the standard normal distribution and any zero mean and unit variance distributions with bounded support.

(B.2) *High dimensional regime*: We assume that the effective dimension grows with the sample size n s.t. $d_{\text{eff}}/n^\beta \to c$ for some $\beta, c > 0$.

In words, when $\mathbb{P}_X \in \mathcal{Q}$, the data has covariance Σ_d, and when $\mathbb{P}_X \in \mathcal{Q}_{\mathcal{S}^{d-1}}$, the data can be generated by projecting $Z \sim \mathbb{P}_Z \in \mathcal{Q}$ onto the sphere of radius $\sqrt{d_{\text{eff}}}$. Unlike Ghorbani et al. (2021; 2020), we do not require the random vectors x_i to be *uniformly* distributed on the sphere (see Table 1). In the sequel we assume, without loss of generality, that for simplicity $||\Sigma_d||_{\text{op}} = 1$ and hence $d_{\text{eff}} = \text{tr}(\Sigma_d)$.
In our analysis, the kernel function g does not change for any d. However as $d_{\text{eff}}, n \to \infty$ we need to adjust the scaling of the input as the norm concentrates around $\mathbb{E}\|x\|_2^2 = d_{\text{eff}}$. Hence, we consider the sequence of scale dependent kernels

$$k_\tau(x,x') = g\left(\frac{\|x\|_2^2}{\tau}, \frac{\|x'\|_2^2}{\tau}, \frac{x^\top x'}{\tau}\right) \tag{4}$$

and parameterize the scaling by a sequence of parameters τ dependent on n. In Section 3.1 we study the standard scaling $\frac{\tau}{d_{\text{eff}}} \to c > 0$, before discussing $\frac{\tau}{d_{\text{eff}}} \to 0$ and $\frac{\tau}{d_{\text{eff}}} \to \infty$ respectively in Section 3.2, where we show that the polynomial approximation barrier is not a consequence of the standard scaling.

3. Main Results

We now present our main results that hold for a wide range of distributions and kernels and show that kernel methods can at most consistently learn low-degree polynomials. Section 3.1 considers the case $\tau \asymp d_{\text{eff}}$ while Section 3.2 provides lower bounds for the regimes $\frac{\tau}{d_{\text{eff}}} \to 0$ and $\frac{\tau}{d_{\text{eff}}} \to \infty$.

3.1. Inconsistency of kernel regression for $\tau \asymp d_{\text{eff}}$

For simplicity, we present a result for the case $\tau = d_{\text{eff}}$. The more general case $\tau \asymp d_{\text{eff}}$ follows from the exact same arguments. In the sequel we denote by $\mathcal{P}_{\leq m}$ the space of polynomials of degree at most $m \in \mathbb{N}$.

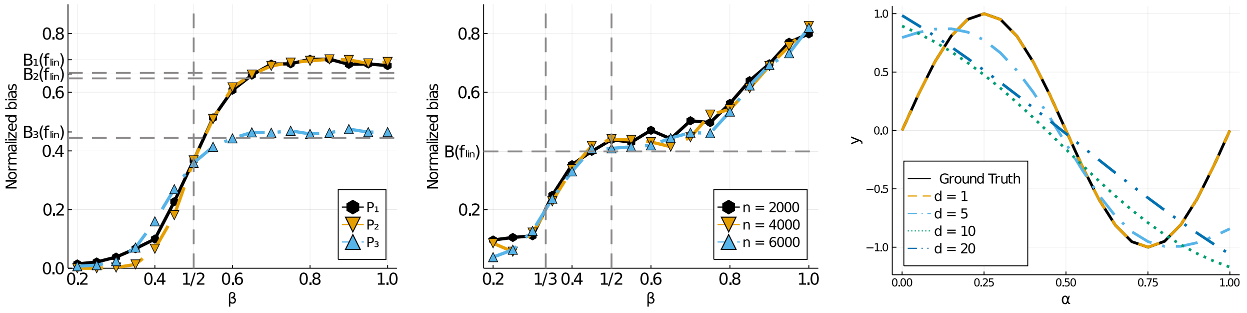
Theorem 3.1 (Polynomial approximation barrier)**.** *Assume that the kernel k, respectively its restriction onto the sphere, satisfies A.1-A.2 or is an α-exponential kernel. Furthermore, assume that the input distribution $\mathbb{P}_X$ satisfies Assumptions B.1-B.2 and that the ground truth function f^* is bounded. Then, for some $m \in \mathbb{N}$ specified below, the following results hold for both the ridge* (1) *and ridgeless estimator* (2) *$\hat{f}_\lambda$ with $\lambda \geq 0$.*

1. *The bias of the kernel estimators $\hat{f}_\lambda$ is asymptotically almost surely lower bounded for any $\epsilon > 0$,*
$$\boldsymbol{B}(\hat{f}_\lambda) \geq \inf_{p \in \mathcal{P}_{\leq m}} \|f^\star - p\|_{\mathcal{L}_2(\mathbb{P}_X)} - \epsilon \quad a.s. \text{ as } n \to \infty. \tag{5}$$

2. *For bounded kernel functions on the support of $\mathbb{P}_X$ the averaged estimator $\mathbb{E}_Y \hat{f}_\lambda$ converges almost surely in $\mathcal{L}_2(\mathbb{P}_X)$ to a polynomial $p \in \mathcal{P}_{\leq m}$,*
$$\left\|\mathbb{E}_Y \hat{f}_\lambda - p\right\|_{\mathcal{L}_2(\mathbb{P}_X)} \to 0 \quad a.s. \text{ as } n \to \infty. \tag{6}$$

More precisely, if g_i is $(\lfloor 2/\beta \rfloor + 1 - i)$-times continuously differentiable in a neighborhood of $(1,1)$ *and there exists $j' > \lfloor 2/\beta \rfloor$ such that $g_{j'}(1,1) > 0$, then the bounds* (5),(6) *hold with $m = 2\lfloor 2/\beta \rfloor$ for $\mathbb{P}_X \in \mathcal{Q}$ and $m = \lfloor 2/\beta \rfloor$ for $\mathbb{P}_X \in \mathcal{Q}_{\mathcal{S}^{d-1}}$.*

The almost sure statements refer to the sequence of matrices $\mathbf{X}$ of random vectors x_i as $n \to \infty$, but also hold true with probability $\geq 1 - n^2 \exp(-C_{\epsilon'} \log(n)^{(1+\epsilon')})$ over the draws of $\mathbf{X}$ (see Lemma C.1 for further details).
The first statement in Theorem 3.1 shows that even with noiseless observations, the estimator $\hat{f}_\lambda$ can consistently learn ground truth functions $f^\star$ that are a polynomial of degree at most m. We refer to m as the β-dependent *polynomial approximation barrier*. Figures 2a and 2b illustrate this barrier on synthetic datasets drawn from different input data distributions (see Section 4.2 for a detailed discussion). Furthermore, the second part of Theorem 3.1 states that the averaged estimator $\mathbb{E}_Y \hat{f}_\lambda$ converges in $\mathcal{L}_2(\mathbb{P}_X)$ to a polynomial of degree at most m when the kernel is bounded on the support of $\mathbb{P}_X$. We provide a closed form expression for the polynomial in the proof of the Theorem 3.1



(a) Ground truth $f^\star = 2x_{(1)}^2$

(b) Ground truth $f^\star = 2x_{(1)}^3$

(c) Degeneration of estimator

Figure 2: (a) and (b): The bias of the minimum norm interpolant $\mathbf{B}(\hat{f}_0)$ normalized by $\mathbf{B}(0)$ as a function of β for (a) different covariance models $\mathbb{P}_1$- $\mathbb{P}_3$ (see Section 4.2) with $n = 4000$ and (b) different choices of n and samples generated from the isotropic Gaussian $\mathbb{P}_1$ with $d = \lfloor n^\beta \rfloor$. The horizontal lines $\mathbf{B}(f_{\text{lin}})$ correspond to the risk of the optimal linear model for the respective input distributions $\mathbb{P}_i$. (c): The minimum norm interpolator $\hat{f}_0$ plotted in the direction $(0, 1/2, \cdots, 1/2) + \alpha e_1$ when fitting noiseless observations with $f^\star(x) = \sin(2\pi x_{(1)})$ and $n = 100$ covariates drawn uniformly from $[0, 1]^d$ for varying d.

in Appendix C.1. Figure 2c illustrates how the estimator degenerates to a linear function as dimension grows. We refer to Theorem C.2 in Appendix C for slightly weaker statements which also apply to unbounded kernels.

The attentive reader might notice that Ghorbani et al. (2020) achieve a lower barrier $m = \lfloor 1/\beta \rfloor$ for their specific setting which implies that our results are not tight. However, we leave this as a future work as the focus of this paper is to demonstrate that the polynomial approximation barrier persists for general covariance model data distributions (Ass. B.1-B.2).

We now present a short proof sketch to provide intuition for the polynomial approximation barrier. The full proof can be found in Appendix C.

Proof sketch The proof of the main theorem is primarily based on the concentration of Lipschitz continuous functions of vectors with i.i.d entries. In particular, we show in Lemma C.1 that

$$\max_i \frac{\left|x_i^\top X\right|}{\tau} \leq n^{-\beta/2}(\log n)^{(1+\epsilon)/2} \quad a.s. \text{ as } n \to \infty, \tag{7}$$

where we use $\operatorname{tr}(\Sigma_d) \asymp n^\beta$. Furthermore, Assumption A.1 and hence the rotational invariance of the kernel, implies that for inner product kernels where g_j are scalars,

$$k_\tau(x_i, X) = \sum_{j=0}^{m} g_j \cdot \left(\frac{x_i^\top X}{\tau}\right)^j + O\left(n^{-\theta}\right) \quad a.s. \text{ as } n \to \infty \tag{8}$$

with θ some constant such that $1 < \theta < (m+1)\frac{\beta}{2}$ that exists because $m \geq \lfloor 2/\beta \rfloor$. Hence, as $n \to \infty$, $k_\tau(x_i, X)$ converge to low-degree polynomials. Using the closed form solution of $\hat{f}_\lambda$ based on the representer theorem we can hence conclude the first statement in Theorem 3.1 if $K + \lambda I \succ cI$ for some constant $c > 0$. The result follows naturally for ridge regression with non vanishing $\lambda > 0$. However for the minimum norm interpolator, we need to show that the eigenvalues of the kernel matrix K themselves are asymptotically lower bounded by a positive non-zero constant. This follows from the additional assumption in Theorem 3.1 and the observation that $(\mathbf{X}^\top \mathbf{X})^{\circ j'} \to I_n$ in operator norm with $\circ$ being the Hadamard product. Finally, the case where g_j are non-constant functions of x, x' requires a more careful analysis and constitute the major bulk of the proof in Appendix C. □

The assumptions in Theorem 3.1 cover a broad range of commonly used kernels. The following corollary summarizes relevant special cases, some of which have also been studied in previous works.

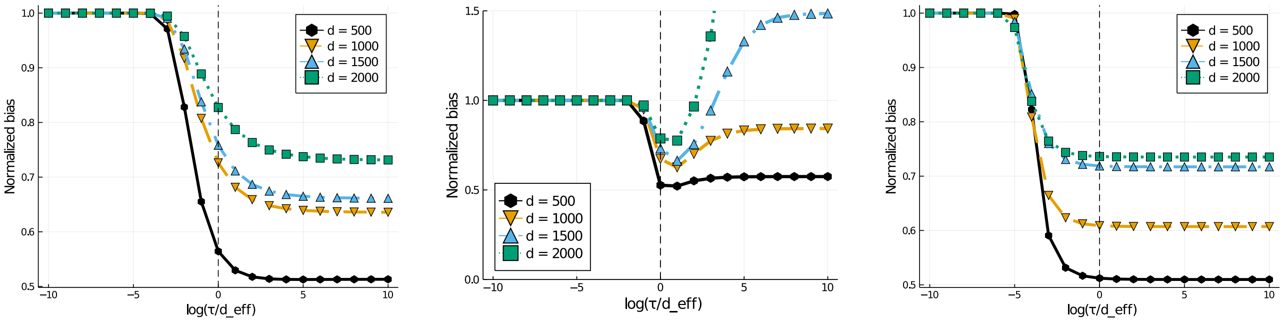
Corollary 3.2. *Theorem 3.1 applies to:*

1. *The exponential inner product kernel*
2. *The α-exponential kernel for $\alpha \in (0, 2]$, including Laplace ($\alpha = 1$) and the Gaussian ($\alpha = 2$) kernels*
3. *The fully-connected NTK of any depth with regular activation functions including the ReLU activation $\sigma(x) = \max(0, x)$*

The precise regularity conditions of the activation functions for the NTK and the proof of the corollary can be found in Appendix C.3.

3.2. Inconsistency of kernel interpolation for $\tau \not\asymp d_{\text{eff}}$

As Section 3.1 focuses on the classical scaling $\frac{\tau}{d_{\text{eff}}} \asymp 1$, an important question remains unaddressed: can we avoid the polynomial approximation barrier with a different scaling? When $\frac{\tau}{d_{\text{eff}}} \to 0$, intuitively the estimate $\hat{f}_\lambda$ converges to the zero function almost everywhere and hence the bias is lower bounded by the $\mathcal{L}_2(\mathbb{P}_x)$-norm of $f^\star$ (see Appendix D.2 for



(a) Laplacian kernel

(b) Gaussian kernel

(c) Exponential inner product kernel

Figure 3: The bias of the minimum interpolant $\mathbf{B}(\hat{f}_0)$ normalized by $\mathbf{B}(0)$ as a function of the normalization constant τ for different choices of $d = d_{\text{eff}}$. The ground truth function is $f(x) = 2x_{(1)}^3$ and $n = 2000$ noiseless observations are fit where the input vectors are sampled from an isotropic Gaussian with $d = \lfloor n^{1/2} \rfloor$.

a rigorous statement). When τ increases faster than d_{eff}, however, the behavior is unclear a priori. Simulations in Figure 3a, 3c suggest that the bias could in fact decrease for $\tau \gg d_{\text{eff}}$ and attain its minimum at the so called *flat limit*, when $\tau \to \infty$. To the best of our knowledge, the next theorem is the first to show that the polynomial approximation barrier persists in the flat limit for RBF kernels with polynomial eigenvalue decay.

Theorem 3.3. *Let k be an RBF kernel with Fourier transform $\hat{k}$ such that for any d, $\lim_{\|\theta\|\to\infty} \|\theta\|^{d+\alpha}\hat{k}(\theta) = c_d > 0$ for some $\alpha \in (0, 2)$. Under the assumptions B.1-B.2 on the data distribution, the bias lower bound* (5) *and polynomial approximation* (6) *hold for the flat limit interpolator $\lim_{\tau\to\infty} \hat{f}_0$ with the same β-dependence for m as in Theorem 3.1, given that $f^\star$ is bounded on the support of $\mathbb{P}_X$.*

In particular, the assumptions hold for instance for the α-exponential kernels with $\alpha \in (0, 2)$ (see (Blumenthal and Getoor, 1960)) and the popular Matern RBF kernels with $\nu < 2$. The proof of the theorem can be found in Appendix D.1 and is based on the flat limit literature on RBFs (Lee et al., 2014; Driscoll and Fornberg, 2002; Schaback, 2005; Larsson and Fornberg, 2005). Finally, we remark that Theorem 3.3 applies only for the interpolating estimator $\hat{f}_0$. However, it is well known that the bias increases with the ridge penalty $\lambda > 0$ and hence attains its minimum at $\lambda = 0$.

3.3. Discussion of theoretical results

Our unifying treatment shows that the polynomial approximation barrier (5) is neither restricted to a few specific distributions nor to a particular choice of the scaling or the eigenvalue decay of the kernel function (see Table 1).

Distribution Assumptions B.1-B.2 allow very general distributions and include the ones in the current literature. In particular, we also cover the settings studied in the papers (Liang et al., 2020a;b; Liu et al., 2021) and consequently put their consistency results into perspective. Besides, our results hold true for *arbitrary* covariance matrices and only depend on the growth rate of the effective dimension, but are independent of the explicit structure of the covariance matrix. This stands in contrast to linear regression where consistency results depend on the "spikiness" of the covariance matrices (Bartlett et al., 2020; Muthukumar et al., 2020).

Scaling Our results do not only apply for the standard choice of the scaling $\tau/d_{\text{eff}} \to c > 0$, but also apply to general RBF kernels in the flat limit scaling, i.e. where $\tau \to \infty$. This case is particularly important since this is where we empirically find the bias to attain its minimum in Figure 3c. We therefore conjecture that the polynomial barrier cannot be overcome with different choices of the scaling.

Eigenvalue decay Furthermore, by explicitly showing that the polynomial barrier persists for all α-exponential kernels with $\alpha \in (0, 2]$ (that have vastly different eigenvalue decay rates), we provide a counterpoint to previous work that suggests consistency for $\alpha \to 0$. In particular, Belkin et al. (2019) prove minimax optimal rates for the Nadaraya-Watson estimators with singular kernels for fixed dimensions and empirical work by Belkin et al. (2018) and Wyner et al. (2017) suggests that spikier kernels have more favorable performances. Our results, however, suggest that in high dimensions, the effect of eigenvalue decay (and hence "spikiness") may be dominated by asymptotic effects of rotationally invariant kernels. We discuss possible follow-up questions in Section 5.

As a result, we can conclude that the polynomial approximation barrier is a rather general phenomenon that occurs for commonly used rotationally invariant kernels. For ground truths that are inherently higher-degree polynomials that depend on all dimensions, our theory predicts

Functions $f^\star$	Kernels	Domain	Choice of τ	Σ_d	Regime	Paper
$\mathcal{P}_{\leq m}$	IP, α-exp, NTK	$\mathbb{R}^d$, $\mathcal{S}^{d-1}(\sqrt{d_{\text{eff}}})$	$\tau = d_{\text{eff}}$	arbitrary	$d_{\text{eff}} \asymp n^\beta$	Ours
$\mathcal{P}_{\leq m}$	RBF	$\mathbb{R}^d$, $\mathcal{S}^{d-1}(\sqrt{d_{\text{eff}}})$	$\tau \to \infty$	arbitrary	$d_{\text{eff}} \asymp n^\beta$	Ours
$\|f^\star\|_{\mathcal{H}} = O(1)$	IP, NTK	$\mathbb{R}^d$	$\tau = d_{\text{eff}} = d$	I_d	$d \asymp n^\beta$	(Liang et al., 2020a)
$\|f^\star\|_{\mathcal{H}} = O(1)$[2]	IP, RBF	$\mathbb{R}^d$	$\tau = d_{\text{eff}} = d$	$\text{tr}(\Sigma_d)/d \to c$ or $\to 0$	$d \asymp n$	(Liang et al., 2020b), (Liu et al., 2021)
$\mathcal{P}_{\leq m'}$	IP, NTK	$\mathcal{S}^{d-1}(\sqrt{d_{\text{eff}}})$	$\tau = d$	I_d	$d \asymp n^\beta$	(Ghorbani et al., 2021)
$\mathcal{P}_{\leq m'}$	IP, NTK	$\mathcal{S}^{d-1}(\sqrt{d_{\text{eff}}})$	$\tau \approx d_{\text{eff}}$	$UU^\top + d^{-\kappa} I$	$d_{\text{eff}} \asymp n^\beta$	(Ghorbani et al., 2020)

Table 1: Compilation of the different settings studied in the literature and our paper. The left-most column denotes the necessary conditions on the function space of the ground truth $f^\star$ for the corresponding consistency results. Here, $m = 2\lfloor 2/\beta \rfloor$ and $m' = \lfloor 1/\beta \rfloor$.

that consistency of kernel learning with fully-connected NTK, standard RBF or inner product kernels is out of reach if the data is high-dimensional. In practice however, it is possible that not all dimensions carry equally relevant information. In Section 4.3 we show how, in this case, feature selection can be used in such settings to circumvent the bias lower bound. On the other hand, for image datasets like CIFAR-10 where the ground truth is a complex function of all input dimensions, kernels that perform well heavily rely on convolutional structures to break the rotational symmetries such CNTKs or compositional kernels (Arora et al., 2019; Novak et al., 2019; Daniely et al., 2016; Shankar et al., 2020; Mei et al., 2021b)).

4. Experiments

In this section we describe our synthetic and real-world experiments to further illustrate our theoretical results and underline the importance of feature selection in high dimensional kernel learning.

4.1. Hilbert space norm increases with dimension d

In Figure 1, we demonstrate how the Hilbert norm of the simple sparse linear function $f^\star(x) = x_{(1)}$ grows with dimension d as discussed in Section 2.2. We choose the scaling $\tau = d$ and consider the Hilbert space induced by the scaled Gaussian $k_\tau(x, x') = \exp(-\frac{\|x-x'\|^2}{\tau})$, Laplace $k_\tau(x, x') = \exp(-\frac{\|x-x'\|_2}{\sqrt{\tau}})$ and exponential inner product $k_\tau(x, x') = \exp(-\frac{x^T x'}{\tau})$ kernels. To estimate the norm, we draw 7500 i.i.d. random samples with noiseless observations from the uniform distribution on $\mathcal{X} = [0, 1]^d$.

4.2. Illustration of the polynomial barrier

We now provide details for the numerical experiments in Figure 2,3 that illustrate the lower bounds on the bias in Theorem 3.1 and Theorem 3.3. For this purpose, we consider the following three distributions $\mathbb{P}_X$ that satisfy the assumptions of the theorems and are covered in previous works:

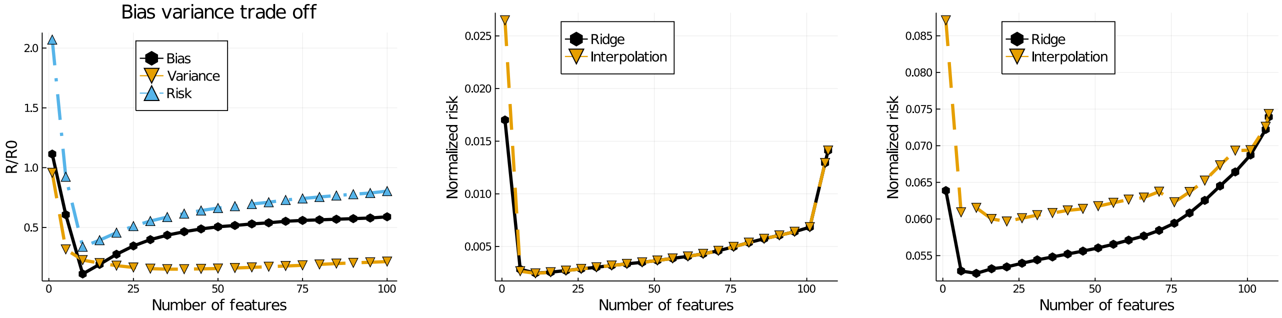
- $\mathbb{P}_1$**:** $X = W$, $W_{(i)} \sim \mathbb{N}(0, 1)$ and $d = \lfloor n^\beta \rfloor$
- $\mathbb{P}_2$**:** $X = \sqrt{d}\frac{W}{\|W\|_2}$, $W_{(i)} \sim \mathbb{N}(0, 1)$ and $d = \lfloor n^\beta \rfloor$
- $\mathbb{P}_3$**:** $X = \Sigma_d^{1/2} W$, $W_{(i)} \sim \text{Uniform}([-u, u]^d)$, with u such that $W_{(i)}$ has unit variance, Σ_d diagonal matrix with entries $(1 - ((i-1)/d)^\kappa)^{1/\kappa}$ and $\kappa \geq 0$ such that $\text{tr}(\Sigma_d) = n^\beta$ and $d = n$

We primarily use the Laplace kernel with $\tau = \text{tr}(\Sigma_d)$ unless otherwise specified and study two sparse monomials as ground truth functions, $f_1^\star(x) = 2x_{(1)}^2$ and $f_2^\star(x) = 2x_{(1)}^3$. We choose the Laplace kernel because of its numerical stability and good performance on the high dimensional datasets studied in (Belkin et al., 2018; Geifman et al., 2020). Other kernels can be found in Appendix A. In order to estimate the bias $\|\mathbb{E}_Y \hat{f}_0 - f^\star\|_{\mathcal{L}_2(\mathbb{P}_X)}^2$ of the minimum norm interpolant we fit noiseless observations and approximate the expected squared error using 10000 i.i.d. test samples.

In Figures 2a and 2b, we plot the dependence of the bias on the parameter β which captures the degree of high-dimensionality, i.e. how large dimension d is compared to the number of samples n. We vary β by fixing n and increasing d (see also Appendix A for plots for fixed d and varying n). In Figure 2a we demonstrate the important consequence of our unifying framework that the polynomial barrier only depends on the growth of the effective dimension $\text{tr}(\Sigma_d)$ parameterized by β. The horizontal lines that indicate the bias of the optimal linear fit f_{lin}, show how for large β, kernel learning with the Laplace kernel performs just as well as a linear function. Figure 2b depicts the bias curve for different choices of n as a function of β with inputs drawn from $\mathbb{P}_1$. Since the bias curves are identical we conclude that we already enter the asymptotic regime for d, n as low as $d \sim 50$ and $n \sim 2000$.

In Figures 2a,2b we provide conclusive evidence in a synthetic setting that the increasing bias is caused by the polynomial approximation barrier. For a nonlinear ground truth function the polynomial approximation barrier suggests that as $\beta = \log d_{\text{eff}} / \log n$ increases, the bias will align stepwise

[2] (Liu et al., 2021) actually requires the weaker assumption that the source condition parameter $r > 0$.



(a) Bias variance trade-off

(b) Residential housing - original

(c) Residential housing - add. noise

Figure 4: (a): The bias-variance trade-off of the minimum norm interpolant normalized by $\mathbf{B}(0)$ for a synthetic experiment as a function of selected features (see details in Section 4.3). The trends are reversed compared to the usual bias-variance curve as a function of model complexity and reflect our theoretical results: the bias term dominates the risk as dimension increases, while the variance monotonically decreases with dimension. This behaviour can also be observed for the *residential housing* dataset (b) without and (c) with additive synthetic noise where the risk of Ridge regression and interpolation follow similar trends that we hence attribute to the bias.

with the best polynomial of decreasing order. For example, with decreasing β, for the cubic polynomial in Figure 2b, we first start to learn linear functions (first descent in the curve). Since the best degree 2 polynomial approximation of $f_2^\star$ around 0 is a linear function, the curve then enters plateau. When further decreasing β, the risk starts to descend again as it gains the ability to learn degree 3 polynomial functions. Indeed, with decreasing β, for the cubic polynomial in Figure 2b we first learn linear functions (first descent in the curve). Since the best degree 2 polynomial approximation of f_2^* around 0 is a linear function, the curve then enters a plateau before descending to zero indicating that we successfully learn the ground truth.

Figure 3 illustrates how the bias depends on the scaling τ for different β with $n = 2000$. We generate samples using $\mathbb{P}_1$ and the ground truth $f_2^\star$ and plot the bias of the minimum norm interpolator for the Laplace, exponential inner product and Gaussian kernel. For the latter, the minimum is obtained around $\tau = d$. For the Laplace and exponential inner product kernel, the bias achieves its minimum at the flat limit $\tau \to \infty$. Given that our lower bounds hold for both $\tau = d$ and $\tau \to \infty$ (Theorems 3.1 and 3.3), we conjecture that there might not exist an intermediate scaling regime for τ that can break the polynomial approximation barrier.

4.3. Feature selection for high-dim. kernel learning

In this section, we demonstrate how the polynomial approximation barrier limits the performance in real world datasets and how one may overcome this issue using feature selection.

Based on our theoretical results we expect that for sparse ground truths (i.e. that depend only on a few features), the bias follows a U-shape as dimension increases: until all relevant features are included, the bias first decreases before it then starts to increase due to the polynomial approximation barrier that holds for large d when asymptotics start to kick in. Since recent work shows that the variance vanishes in high-dimensional regimes (see e.g. (Liang et al., 2020a; Ghorbani et al., 2020)), we expect the risk to follow a U-shaped curve as well. Hence, performing feature selection could effectively yield much better generalization for sparse ground truth functions. We would like to emphasize that this behavior *is not* due to the classical bias-variance trade-off, since the U-shaped curve can be observed even in the noiseless case where we have zero variance. We now present experiments that demonstrate the U-shape of the risk curve for both synthetic experiments on sparse ground truths and real-world data. We vary the dimensionality d by performing feature selection using the algorithm proposed in the paper (Chen et al., 2017)[3]. In order to study the impact of high-dimensionality on the variance, we add different levels of noise to the observations.

Sparse functions For our synthetic experiments, we draw 500 input samples $x_1, \ldots, x_n$ from $\mathbb{P}_1$ and compute the minimum norm interpolator for $n = 100$ different draws over noisy observations from a sparsely parameterized function $y = 0.5 \sum_{i=1}^4 x_{(2i+1)}^2 - \sum_{i=1}^4 x_{(2i)} + \epsilon$ with uniform noise $\epsilon \sim \mathcal{U}([-10, 10])$. We increase d by adding more dimensions (that are irrelevant for the true function value) and compute the bias and variance of the minimum norm interpolator. We approximate the $\mathcal{L}_2(\mathbb{P}_x)$ norm using Monte Carlo sampling from $\mathbb{P}_1$. In Figure 4a we observe that as we increase the number of selected features, the bias first decreases until all relevant information is included and then increases as irrelevant dimensions are added. This is in line with our asymptotic theory that predicts an increasing bias due to the progressively more restrictive polynomial approximation barrier. Furthermore, the variance monotonically

[3] We expect other approaches that incorporate sparsity such as automatic relevance determination (Neal, 1996; MacKay, 1996) to yield a similar effect.

decreases as expected from the literature. Therefore, the risk follows the U-shaped curve described above.

Real-world data We now explore the applicability of our results on real-world data where the assumptions of the theorems are not necessarily satisfied. For this purpose we select datasets where the number of features is large compared to the number of samples. In this section we show results on the regression dataset *residential housing* (RH) with $n = 372$ and $d = 107$ to predict sales prices from the UCI website (Dua and Graff, 2017). Further datasets can be found in Appendix A.3. In order to study the effect of noise, we generate an additional dataset (RH-2) where we add synthetic i.i.d. noise drawn from the uniform distribution on $[-1/2, 1/2]$ to the observations. The plots in Figure 4 are then generated as follows: we increase the number of features using a greedy forward selection procedure (see Appendix A.3 for further details). We then plot the risk achieved by the kernel ridge and ridgeless estimate using the Laplace kernel on the new subset of features.

Figure 4b shows that the risks of the minimum norm interpolator and the ridge estimator are identical, indicating that the risk is essentially equivalent to the bias. Hence our first conclusion is that, similar to the synthetic experiment, the bias follows a U-curve. For the dataset RB-2 in Figure 4c, we further observe that even with additional observational noise, the ridge and ridgeless estimator match, i.e. the bias dominates the risk for large d. We also observe both trends in other high-dimensional datasets discussed in Appendix A.3. Hence we conclude that even for real-world datasets that do not necessarily satisfy the conditions of our bias lower bound, feature selection is crucial for kernel learning for noisy and noiseless observations alike.

We would like to note that this conclusion does not necessarily contradict empirical work that demonstrates good test performance of RBFs on other high-dimensional data such as MNIST. In fact, experiments only indicate that linear or polynomial fitting would do just as well for these datasets which has also been suggested in (Ghorbani et al., 2020).

5. Conclusion and future work

Kernel regression encourages certain structural properties through the RKHS norm induced by the kernel. For example, the eigenvalue decay of α-exponential kernels results in estimators that tend to be smooth (i.e. Gaussian kernel) or more spiky (i.e. small $\alpha < 1$). In fact, it is far less discussed that by choosing a kernel we already implicitly assume certain structure of the data. For instance, rotational invariant kernels are invariant under permutations and hence treat all dimensions equally. Even though rotational invariance is a natural choice when no prior information on the structure of the ground truth is available, this paper shows that the corresponding inductive bias in high dimensions is in fact, restricting the average estimator to a polynomial. In particular, we show in Theorems 3.1 and 3.3 that the lower bound on the bias is simply the projection error of the ground truth function onto the space of polynomials of degree at most $2\lfloor 2/\beta \rfloor$ respectively $\lfloor 2/\beta \rfloor$. Apart from novel technical insights that result from our unified analysis (discussed in Sec. 3.3), our result also opens up new avenues for future research.

Future work Modern datasets which require sophisticated methods like deep neural networks to obtain good predictions are usually inherently non-polynomial and high-dimensional. Hence, our theory predicts that commonly used rotationally invariant kernels cannot perform well for these problems due to a high bias. In particular, our bounds are independent of properties like the smoothness of the kernel function and cannot be overcome by carefully choosing the eigenvalue decay. Therefore, in order to understand why certain highly overparameterized methods generalize well, our results suggest that it is important to understand how prior information can be used to break the rotational symmetry of the kernel function. Examples for recent contributions in this direction are kernels for image datasets relying on convolution structures such as CNTKs (Arora et al., 2019; Novak et al., 2019) or compositional kernels (Daniely et al., 2016; Shankar et al., 2020; Mei et al., 2021b).

Another relevant future research direction is to present a tighter non-asymptotic analysis that allows a more accurate characterization of the estimator in practice. The presented results in this paper are asymptotic statements, meaning that they do not provide explicit bounds for fixed n, d. Therefore, for given finite n, d it is unclear which high-dimensional regime provides the most accurate characterization of the estimator's statistical properties. For instance, our current results do not provide any evidence whether the estimator follows the bias lower bounds for $n = d^\beta$ with $\beta = \log(n)/\log(d)$ or $n = \gamma d$. We remark that the methodology used to prove the statements in this paper could also be used to derive non-asymptotic bounds, allowing us to further investigate this problem. However, we omitted such results in this paper for the sake of clarity and compactness of our theorem statements and proofs and leave this for future work.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedback, Xiao Yang for insightful discussions on the flat limit kernel and Armeen Taeb for his comments on the manuscript.

References

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8141–8150, 2019.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. 117 (48):30063–30070, 2020.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pages 541–549, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1611–1619, 2019.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, 1984.

R. M. Blumenthal and R. K. Getoor. Some theorems on stable processes. *Transactions of the American Mathematical Society*, 95(2):263–273, 1960.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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

Jianbo Chen, Mitchell Stern, Martin J Wainwright, and Michael I Jordan. Kernel feature selection via conditional covariance minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6946–6955, 2017.

Andreas Christmann, Ingo Steinwart, et al. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

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 (NeurIPS)*, volume 29, 2016.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

Tobin A Driscoll and Bengt Fornberg. Interpolation in the limit of increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 43(3-5):413–422, 2002.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Noureddine El Karoui et al. The spectrum of kernel random matrices. *Annals of Statistics*, 38(1):1–50, 2010.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Ronen Basri. On the similarity between the laplace and neural tangent kernels. *arXiv preprint arXiv:2007.01580*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14820–14830, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Annals of Statistics*, 49(2):1029 – 1054, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Khakim Ikramov and N. Savel'eva. Conditionally definite matrices. *Journal of Mathematical Sciences*, 98:1–50, 2000.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Elisabeth Larsson and Bengt Fornberg. Theoretical and computational aspects of multivariate interpolation with increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 49(1):103–130, 2005.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001. ISBN 9780821837924.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Yeon Ju Lee, Charles A Micchelli, and Jungho Yoon. On convergence of flat multivariate interpolation by translation kernels with finite smoothness. *Constructive Approximation*, 40(1):37–60, 2014.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2018.

Tengyuan Liang, A. Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of the Conference on Learning Theory (COLT)*, 2020a.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020b.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 649–657, 13–15 Apr 2021.

David J. C. MacKay. *Bayesian Non-Linear Modeling for the Prediction Competition*, pages 221–234. Springer Netherlands, Dordrecht, 1996. ISBN 978-94-015-8729-7.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021a.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models, 2021b.

D Vitali Milman and Gideon Schechtman. Asymptotic theory of finite dimensional normed spaces. 1986.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996. ISBN 0387947248.

Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Robert Schaback. Multivariate interpolation by polynomials and radial basis functions. *Constructive Approximation*, 21(3):293–317, 2005.

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 International Conference on Machine Learning (ICML)*, volume 119, pages 8614–, 2020.

Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. ISBN 9781108498029.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research (JMLR)*, 18(1):1558–1590, 2017.