Kernel regression in high dimensions: Refined analysis beyond double descent

Fanghui Liu
ESAT-STADIUS
KU Leuven
fanghui.liu@kuleuven.be

Zhenyu Liao
ICSI and Department of Statistics
UC Berkeley
zhenyu.liao@berkeley.edu

Johan A.K. Suykens
ESAT-STADIUS
KU Leuven
johan.suykens@esat.kuleuven.be

Abstract

In this paper, we provide a precise characterization of generalization properties of high dimensional kernel ridge regression across the under- and over-parameterized regimes, depending on whether the number of training data n exceeds the feature dimension d. By establishing a bias-variance decomposition of the expected excess risk, we show that, while the bias is (almost) independent of d and monotonically decreases with n, the variance depends on n, d and can be unimodal or monotonically decreasing under different regularization schemes. Our refined analysis goes beyond the double descent theory by showing that, depending on the data eigen-profile and the level of regularization, the kernel regression risk curve can be a double-descent-like, bell-shaped, or monotonic function of n. Experiments on synthetic and real data are conducted to support our theoretical findings.

1 Introduction

Interpolation learning [1, 2, 3] has recently attracted growing attention in the machine learning community. This is mainly because current state-of-the-art neural networks appear to be models of this type: they are able to interpolate the training data while still generalize well on test data, even in the presence of label noise [4]. It has been empirically observed that other models including random features, decision trees, and as simple as linear regression also exhibit similar phenomenon [3, 5, 6]. This is somewhat striking as it goes against the conventional wisdom of *bias-variance trade-off* [7]: predictors that generalize well must trade off the model complexity against training data fitting. The double descent theory [5] resolves this paradox by revisiting the bias-variance trade-off and showing that the model generalization error exhibits a phase transition at the *interpolation point*: moving away from this point on either side tends to reduce the generalization error.

The double descent phenomenon has recently inspired intense theoretical research [1, 8, 9, 10] and has been further extended to multiple descent [11, 12, 13] on various models. One line of work formalized the argument that, even when no explicit regularization is imposed, *implicit regularization* is encoded in the model via the choice of optimization algorithms and techniques, e.g., stochastic gradient descent (SGD) [14], dropout [15], early stopping [16], and ensemble methods [17]. Different from these "external" schemes, the kernel interpolation estimator [18, 19] directly benefits from its intrinsic kernel structure that serves as an *implicit regularization* to help both interpolate and approximate. In fact, (strictly) positive-definite kernels can interpolate an arbitrary number of data points [20], and thus kernel spaces contain (nearly) optimal interpolants [21, 22]. Although the kernel space is rich enough to contain models that generalize well, the generalization property of kernel method, for example how it depends on the choice of kernel, its interplay with the data and the level of regularization, still remains unclear. In particular, the question whether the double descent phenomenon exists in the kernel regression models is still unanswered [18, 23]. As such, refined analyses are needed to have a thorough understanding of kernel estimators, notably in the high dimensional regime of interest. This is indeed the objective of the article.

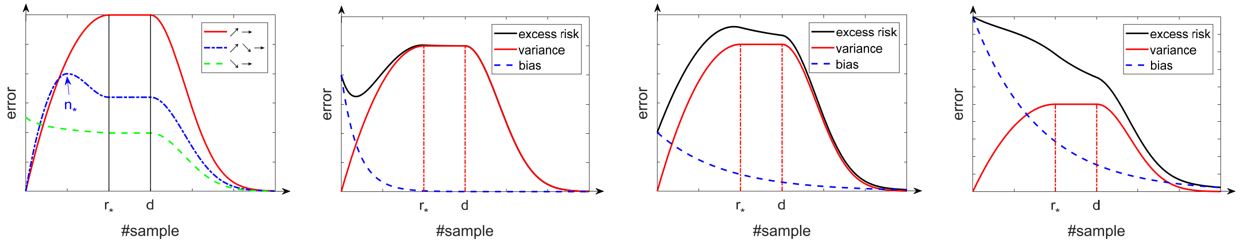
Here, we consider the kernel ridge regression (KRR) estimator [7, 24, 25] in a high dimensional setting with data dimension d and size n both large, and treat the kernel interpolation as a special case of KRR by taking the explicit regularization to be zero. More precisely, by virtue of the linearization of kernel matrices in high di-

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

Table 1: Trends of the variance $\mathtt{V}$ with respect to n in the $n < d$ case. The notation $\nearrow$ means $\mathtt{V}$ increases with n; $\rightarrow$ for $\mathtt{V}$ stays unchanged; and $\searrow$ for $\mathtt{V}$ decreases with n, see Figure 1(a); and $r_* := \operatorname{rank}(\boldsymbol{X}\boldsymbol{X}^\top)$. From left to the right column, the regularization λ increases, and a large λ leads to a small value of peak point $n_* := n_*(\lambda)$, which may even disappear. Note that n_* is different for three eigenvalue decays of $\boldsymbol{X}\boldsymbol{X}^\top/d$. See Section 4.1 for details.

eigenvalue decay	$\lambda = 0$	$\lambda := \bar{c}n^{-\vartheta}$ (KRR)				
harmonic decay	$\nearrow \rightarrow$	$1 \geq \vartheta \geq \frac{1}{2(2-\bar{c})}$	$\vartheta < \frac{1}{2(2-\bar{c})}$			
		$\nearrow \rightarrow$	$r_* < d \leq n_*$	$r_* \leq n_* \leq d$	$n_* \leq r_* < d$	$n_* \leq c < r_* < d$ [1]
			$\nearrow \rightarrow$	$\nearrow \rightarrow$	$\nearrow \searrow \rightarrow$	$\searrow \rightarrow$
polynomial decay	$\nearrow \rightarrow$	$1 \geq \vartheta \geq \frac{1}{1+\frac{1}{2a}}$	$\vartheta < \frac{1}{1+\frac{1}{2a}}$			
		$\nearrow \rightarrow$	$r_* < d \leq n_*$	$r_* \leq n_* \leq d$	$n_* \leq r_* < d$	$n_* \leq c < r_* < d$
			$\nearrow \rightarrow$	$\nearrow \rightarrow$	$\nearrow \searrow \rightarrow$	$\searrow \rightarrow$
exponential decay	$\nearrow \rightarrow$		$r_* < d \leq n_*$	$r_* \leq n_* \leq d$	$n_* \leq r_* < d$	$n_* \leq c < r_* < d$
			$\nearrow \rightarrow$	$\nearrow \rightarrow$	$\nearrow \searrow \rightarrow$	$\searrow \rightarrow$

[1] Here c is some constant such that $n > c$ always holds as n is required to be large in theory and practice.



(a) trends of variance (b) double descent (c) bell-shaped (d) monotonically decreasing

Figure 1: (a) Trends of variance under different regularization schemes corresponding to Table 1. (b-d) Trends of the risk curve under various bias and variance can be double descent, bell-shaped, and monotonically decreasing.

mensions [18, 26, 27, 28, 29], we disentangle the *implicit* regularization of kernel interpolation estimators in an *explicit* manner. As a result, both implicit and explicit regularization schemes can be systematically studied within the proposed framework. Mathematically, KRR aims to solve the following empirical risk minimization problem on a training set $\boldsymbol{z} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with data $\boldsymbol{x}_i \in \mathbb{R}^d$ and responses $y_i \in \mathbb{R}$:

$$f_{\boldsymbol{z},\lambda} := \operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^n \big(f(\boldsymbol{x}_i) - y_i\big)^2 + \lambda\langle f, f\rangle_{\mathcal{H}} \right\}, \quad (1)$$

where an explicit Tikhonov regularization term induced by a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is added to the least-squares objective. In statistical learning theory [30], the regularization parameter $\lambda > 0$ is generally taken to depend on the sample size n in such a way that $\lim_{n\to\infty}\lambda(n) = 0$. Here we assume that $\lambda := \bar{c}n^{-\vartheta}$ with some $\vartheta \geq 0$ and $0 \leq \bar{c} \leq 1$ to cover the interpolation case.

In this paper, we propose a novel bias-variance decomposition of the KRR expected excess risk, and derive non-asymptotic bounds for both bias and variance. This precise assessment leads to fruitful discussions as a function of different data eigenvalue decays and regularization schemes. Our main findings include:

- We demonstrate that, for data dimension d large, the kernel matrix admits the same eigenvalue decay as $\boldsymbol{X}\boldsymbol{X}^\top/d$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ is the data matrix. So in high dimensions, the eigenvalue decay of $\boldsymbol{K}$ is almost determined by the data, as reflected in our error bound for the bias.
- The explicit regularization $\lambda := \bar{c}n^{-\vartheta}$ largely affects the peak point of the variance: a large λ decreases the model complexity, and thus corresponds to a small value of interpolation point $n_* \equiv n_*(\lambda)$. Table 1 shows that, under a small (or zero) regularization so that $r_* \leq n_*$ with $r_* := \operatorname{rank}(\boldsymbol{X}\boldsymbol{X}^\top/d)$: the error bound for variance $\mathtt{V}$ monotonically increases with n until $n := r_*$, as in the red curve of Figure 1(a). Under a moderate regularization with $n_* \leq r_*$: $\mathtt{V}$ first increases with n until $n := n_*$ and then decreases. In this case, the peak point will move to the left due to $n_* < d$, see the blue curve in Figure 1(a). Under a large

regularization with $n_* \leq c$ for some constant c, $\mathtt{V}$ monotonically decreases with n, as in the green curve of Figure 1(a).

- Our error bounds for the bias and the variance exhibit different characteristics. More specifically, the bias bound is (almost) independent of the data/feature dimension d and monotonically decreases with n at a certain $\mathcal{O}(\lambda)$ (learning) rate as in the classical learning theory [30, 31, 32]. Besides, the variance bound depends on n and d, and exhibits monotonic decreasing or unimodal with n under different regularizations. Hence, the expected excess risk, as the sum of bias and variance, can be double descent (Figure 1(b)), bell-shaped (Figure 1(c)), or monotonic decreasing (Figure 1(d)), depending on the level of *implicit* and *explicit* regularizations. This is in agreement with empirical findings in neural networks [33].
- Our non-asymptotic results show that, for large but fixed d, both the variance and bias tends to zero as $n \to \infty$ under $\lambda := \bar{c}n^{-\vartheta}$, implying that the excess risk approaches zero. Based on this, in the double descent case particularly, the minimum of the expected error in the over-parameterized $n > d$ regime is *lower* than that in the $n < d$ regime. This claim cannot be obtained from [18].

The rest of the paper is organized as follows. We briefly introduce problem settings in Section 2. In Section 3, we present our main results on the generalization property of KRR in high dimensions and briefly sketch the main ideas of the proof. Discussions on the derived error bounds are given in Section 4. In Section 5, we report numerical experiments to support our theoretical results and the conclusion is drawn in Section 6.

2 Problem Settings and Preliminaries

We work in the high dimensional regime for some large d, n with $c \leq d/n \leq C$ for some constants $c, C > 0$. For notational simplicity, we denote by $a(n) \lesssim b(n)$: there exists a constant $\widetilde{C}$ independent of n such that $a(n) \leq \widetilde{C} b(n)$, and analogously for $\asymp$ and $\gtrsim$.

2.1 Kernel Ridge Regression Estimator

Let $X \subseteq \mathbb{R}^d$ be a metric space and $Y \subseteq \mathbb{R}$, the instances $(\boldsymbol{x}_i, y_i)$ in the training set $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in Z^n$ are assumed to be independently drawn from a non-degenerate Borel probability measure ρ on $X \times Y$. The *target function* of ρ is defined by

$$f_\rho(\boldsymbol{x}) = \int_Y y \, \mathrm{d}\rho(y \mid \boldsymbol{x}), \ \boldsymbol{x} \in X \,, \tag{2}$$

where $\rho(\cdot \mid \boldsymbol{x})$ is the conditional distribution of ρ at $\boldsymbol{x} \in X$. Define the response vector $\boldsymbol{y} = [y_1, y_2, \cdots, y_n]^\top \in \mathbb{R}^n$ and the kernel matrix $\boldsymbol{K} = \{k(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^n$ induced by a positive definite kernel $k(\cdot, \cdot)$, KRR aims to find a hypothesis $f : X \to Y$ such that $f(\boldsymbol{x})$ is a good approximation of the response $y \in Y$ corresponding to a new instance $\boldsymbol{x} \in X$. This is actually an empirical risk minimization in problem (1). By denoting $k(\boldsymbol{x}, \boldsymbol{X}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), k(\boldsymbol{x}, \boldsymbol{x}_2), \cdots, k(\boldsymbol{x}, \boldsymbol{x}_n)]^\top \in \mathbb{R}^n$, the closed-form of KRR estimator in Eq. (1) is

$$f_{\boldsymbol{z},\lambda}(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{X})^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} \boldsymbol{y} \,. \tag{3}$$

We consider two popular positive definite kernel classes of (i) the inner-product kernel of the form $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = h\left(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle / d\right)$ and (ii) the *radial* kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = h\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 / d\right)$. Here $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is a nonlinear function that is assumed to be (locally) smooth, as in [26, 18]. Examples include commonly used kernels such as linear kernels, polynomial kernels, Sigmoid kernels, exponential kernels, and Gaussian kernels, to name a few.

The expected (quadratic) risk is defined as $\mathcal{E}(f) = \int_Z (f(\boldsymbol{x}) - y)^2 \mathrm{d}\rho$ and the empirical risk functional is defined on the training set $\boldsymbol{z}$, i.e., $\mathcal{E}_{\boldsymbol{z}}(f) = \frac{1}{n}\sum_{i=1}^n \left(f(\boldsymbol{x}_i) - y_i\right)^2$. To measure the estimation quality of $f_{\boldsymbol{z},\lambda}$, one natural way is the *expected excess risk*: $\mathbb{E}_{y|\boldsymbol{x}}[\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho)]$. Specifically, in KRR, the expected excess risk admits $\mathbb{E}_{y|\boldsymbol{x}}[\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho)] = \mathbb{E}_{y|\boldsymbol{x}} \|f_{\boldsymbol{z},\lambda} - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}}$, which is exactly in the weighted $\mathcal{L}^2$-space with the norm $\|f\|^2_{\mathcal{L}^2_{\rho_X}} = \int_X |f(\boldsymbol{x})|^2 \mathrm{d}\rho_X(\boldsymbol{x})$.

2.2 Background on RKHS

Now we characterize the integral operators defined by a kernel. Given a kernel k, its integral operator $L_K : \mathcal{L}^2_{\rho_X} \to \mathcal{L}^2_{\rho_X}$ admits

$$(L_K f)(\cdot) = \int_X k(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) d\rho_X(\boldsymbol{x}), \quad \forall f \in \mathcal{L}^2_{\rho_X} \,. \tag{4}$$

Since L_K is compact, positive definite and self-adjoint, by the spectral theorem (see, Theorem A.5.13 in [34]), there exists countable pairs of eigenvalues and eigenfunctions $\{\mu_i, \psi_i\}_{i=1}^\infty$ of L_K such that $L_K \psi_i = \mu_i \psi_i$, where $\{\psi\}_{i=1}^\infty$ are orthogonal basis of $\mathcal{L}^2_{\rho_X}(X)$ and $\mu_1 \geq \mu_2 \cdots > 0$ with $\lim\limits_{i \to \infty} \mu_i = 0$. Accordingly, by Mercer's theorem, we have $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^\infty \mu_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}')$, and there exists a constant $\kappa \geq 1$ such that $\sup_{\boldsymbol{x} \in X} \sum_{i=1}^\infty \mu_i \psi_i^2(\boldsymbol{x}) \leq \kappa^2$. It holds by $\kappa := \max\{1, \sup_{\boldsymbol{x} \in X} \sqrt{k(\boldsymbol{x}, \boldsymbol{x})}\}$. Based on the data matrix $\boldsymbol{X}$ and the integral operator L_K, the empirical integral operator is given by $L_{K,\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n k(\cdot, \boldsymbol{x}_i) \otimes k(\cdot, \boldsymbol{x}_i)$, which converges to the data-free limit L_K at an $\mathcal{O}(1/\sqrt{n})$ rate [35].

Table 2: Parameters of the linearized kernel $\widetilde{\boldsymbol{K}^{\mathrm{lin}}}$ in [26].

parameters	inner-product kernels	radial kernels
α	$h(0) + h''(0)\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{2d^2}$	$h(2\tau) + 2h''(2\tau)\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{d^2}$
β	$h'(0)$	$-2h'(2\tau)$
γ	$h(\tau) - h(0) - \tau h'(0)$	$h(0) + 2\tau h'(2\tau) - h(2\tau)$
$\boldsymbol{E}$	$\boldsymbol{0}_{n\times n}$	$h'(2\tau)\boldsymbol{A} + \frac{1}{2}h''(2\tau)\boldsymbol{A}\odot\boldsymbol{A}$ [1]

[1] $\boldsymbol{A} := \boldsymbol{1}\boldsymbol{\psi}^\top + \boldsymbol{\psi}\boldsymbol{1}^\top$, where $\boldsymbol{\psi} \in \mathbb{R}^n$ with $\psi_i := \|\boldsymbol{x}_i\|_2^2/d - \tau$ and $\tau := \mathrm{tr}(\boldsymbol{\Sigma}_d)/d$.

3 Main Results

In this section, we state our main result under some basic/technical assumptions, compare it with existing results, and sketch the main ideas of our proof.

3.1 Basic results

To illustrate our analysis, we need the following three standard assumptions.

Assumption 1. *(Existence of f_ρ) We assume $f_\rho \in \mathcal{H}$.*

This is a standard assumption in learning theory and assumes that the target function f_ρ defined in Eq. (2) is indeed realizable, see also [36, 37, 30, 32].

Assumption 2. *(Noise condition [18, 38]) There exists σ such that $\mathbb{E}[(f_\rho(\boldsymbol{x}) - y)^2 \mid \boldsymbol{x}] \leq \sigma^2$, almost surely.*

This is a broad model for the noise in the output y, containing uniformly bounded or sub-Gaussian noise; and is in fact weaker than the standard Bernstein condition, e.g., in [39].

Assumption 3. *((8+m)-moments [18, 40]) Let $\boldsymbol{x}_i = \boldsymbol{\Sigma}_d^{1/2}\boldsymbol{t}_i$, where $\boldsymbol{t}_i \in \mathbb{R}^d$ has i.i.d. entries with zero mean, unit variance, and a finite (8+m)-moments, i.e., its entry $\boldsymbol{t}_i(j)$, $1 \leq j \leq d$, satisfies $\mathbb{E}[\boldsymbol{t}_i(j)] = 0$, $\mathbb{V}[\boldsymbol{t}_i(j)] = 1$, and $\mathbb{E}(|\boldsymbol{t}_i(j)|) \leq Cd^{\frac{2}{8+m}}$ such that $\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\top] = \boldsymbol{\Sigma}_d$ with a bounded spectral norm $\|\boldsymbol{\Sigma}_d\|_2$, for some $m > 0$.*

This is a standard setting in high-dimensional statistics and random matrix theory [26, 38, 18, 2, 27] that assumes that the data are drawn from some not-too-heavy-tailed distribution, with possibly (involved) structure between the entries.

To aid our proof, we need some extra results. In [26], it has been shown that the kernel matrix $\boldsymbol{K}$ in high dimensions can be well approximated by $\widetilde{\boldsymbol{K}^{\mathrm{lin}}}$ in spectral norm, i.e., $\|\boldsymbol{K} - \widetilde{\boldsymbol{K}^{\mathrm{lin}}}\|_2 \to 0$ as $n, d \to \infty$

$$\widetilde{\boldsymbol{K}^{\mathrm{lin}}} := \alpha\boldsymbol{1}\boldsymbol{1}^\top + \beta\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \gamma\boldsymbol{I} + \boldsymbol{E}\,, \tag{5}$$

with non-negative parameters α, β, γ, and the additional matrix $\boldsymbol{E}$ given in Table 2, see some typical examples in Appendix A. Here γ is the *implicit* regularization parameter in kernel estimator that depends on the nonlinear function h in the kernel k and the data structure $\boldsymbol{\Sigma}_d$. According to Eq. (5), denote the shortcut $\widetilde{\boldsymbol{X}} := \beta\boldsymbol{X}\boldsymbol{X}^\top/d + \alpha\boldsymbol{1}\boldsymbol{1}^\top$, we show in high dimensions that, $\boldsymbol{K}$ admits the same eigenvalue decay as $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}\boldsymbol{X}^\top/d$ (see details in Appendix B). Subsequently, we introduce the following quantity function

$$\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b := \mathrm{tr}\left[(\widetilde{\boldsymbol{X}} + b\boldsymbol{I}_n)^{-2}\widetilde{\boldsymbol{X}}\right] = \sum_{i=1}^n \frac{\lambda_i(\widetilde{\boldsymbol{X}})}{\left[b + \lambda_i(\widetilde{\boldsymbol{X}})\right]^2}\,, \tag{6}$$

which is associated with various quantity functions in [16, 38, 18, 41, 42] and, as we shall see, plays an important role in determining the variance behavior. We will discuss at length $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b$ based on different data eigenvalue decays in Section 4.

Formally, our main results of KRR in a high-dimensional regime are stated as follows.

Theorem 1. *(Basic result) Under Assumptions 1-3, let $0 < \delta < 1/2$, $\theta = \frac{1}{2} - \frac{2}{8+m}$, d large enough, taking the regularization parameter $\lambda := \bar{c}n^{-\vartheta}$ with $0 \leq \vartheta \leq 1/2$, for any given $\varepsilon > 0$, it holds with probability at least $1 - 2\delta - d^{-2}$ with respect to the draw of $\boldsymbol{X}$ that*

$$\mathbb{E}_{y|\boldsymbol{x}}\left\|f_{\boldsymbol{z},\lambda} - f_\rho\right\|_{\mathcal{L}^2_{\rho_X}}^2 \lesssim n^{-\vartheta}\log^4\left(\frac{2}{\delta}\right) + \mathtt{V}_1 + \mathtt{V}_2\,, \tag{7}$$

with $\mathtt{V}_1 := \frac{\sigma^2\beta}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma}$ and the residual term $\mathtt{V}_2$

$$\mathtt{V}_2 := \begin{cases} \dfrac{\sigma^2\log^{2+4\varepsilon} d}{(n\lambda+\gamma)^2 d^{4\theta-1}}, & \textit{for inner-product kernels} \\ \dfrac{\sigma^2}{(n\lambda+\gamma)^2}d^{-2\theta}\log^{1+\varepsilon} d, & \textit{for radial kernels}. \end{cases}$$

Remark: The first term in Eq. (7) is the bound of the bias, which is independent of d and monotonically decreases with n. The sum $\mathtt{V}_1 + \mathtt{V}_2$ is the bound of the variance that depends on both n and d. Note that $\mathtt{V}_2$ monotonically decreases with n, and approaches to zero for a large n. Therefore, the error bound for $\mathtt{V}_1 \asymp \frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma}$ is the key part of estimates for the variance and will be discussed in in Section 4, where $n\lambda$ corresponds to the *explicit* regularization and γ the *implicit* regularization. We will demonstrate that $\mathtt{V}_1$ can be monotonically decreasing or unimodal under different regularization schemes. Such monotonic bias and unimodal variance can lead to various behaviors of the excess risk, including monotonically decreasing, double descent, and bell-shaped risk curve, as illustrated in Figure 1 of introduction.

3.2 Refined result

Based on the basic result, if we consider two additional assumptions, i.e., extending Assumption 1 by considering the regularity of f_ρ and studying spectral decay of k via complexity of $\mathcal{H}$, we can obtain a refined result.

Assumption 4. *(Source condition [30]) For some $0 < r \leq 1$, there exists $g_\rho \in \mathcal{L}^2_{\rho_X}$ satisfying $\|g_\rho\|_{\mathcal{L}^2_{\rho_X}} \leq R$ such that $f_\rho = L_K^r g_\rho$.*

It has been widely used in the literature of learning theory to assess the regularity of f_ρ [30, 43, 37], which indicates f_ρ belongs to the range space of L_K^r. Assumption 1 is the worst case of Assumption 4 by choosing $r = 1/2$ since $\|f\|_{\mathcal{L}^2_{\rho_X}} = \|L_K^{1/2} f\|_{\mathcal{H}}$, $\forall f \in \mathcal{L}^2_{\rho_X}$.

Assumption 5. *(Capacity condition [30]) For any $\lambda > 0$, there exist $Q > 0$ and $\eta \in [0, 1]$ such that*

$$\mathcal{N}(\lambda) := \operatorname{tr}\left((L_K + \lambda I)^{-1} L_K\right) \leq Q^2 \lambda^{-\eta} \,.$$

The notation $\mathcal{N}(\lambda)$ denotes the "effective dimension" and can be regarded as a "measure of size" of the RKHS. This is a natural and widely used assumption in the literature [30, 43, 37]. Assumption 5 always holds for $\eta = 1$ and $Q = \kappa$ where $\kappa := \max\{1, \sup_{\boldsymbol{x} \in X} \sqrt{k(\boldsymbol{x}, \boldsymbol{x})}\}$ as L_K is a trace class operator. Its kernel matrix form is $d^\lambda_{\boldsymbol{K}} := \operatorname{tr}\left((\boldsymbol{K} + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{K}\right) = \sum_{i=1}^n \frac{\lambda_i(\boldsymbol{K})}{\lambda_i(\boldsymbol{K}) + \lambda}$ [44, 45]. While Assumption 5 can be further refined to obtain a bound that depends on d [46], here we focus on the eigenvalue decay of $\boldsymbol{K}$, see Section 4 for details.

Based on the above discussion, we obtain a refined result of Theorem 1 as below.

Theorem 2. *(Refined result) Under Assumptions 2-5, let $0 < \delta < 1/2$, $\theta = \frac{1}{2} - \frac{2}{8+m}$, and d large enough, taking $\lambda := \bar{c} n^{-\vartheta}$ with $0 \leq \vartheta \leq \frac{1}{1+\eta}$, then for any given $\varepsilon > 0$, it holds with probability at least $1 - 2\delta - d^{-2}$*

$$\mathbb{E}_{y|\boldsymbol{x}} \big\| f_{\boldsymbol{z},\lambda} - f_\rho \big\|^2_{\mathcal{L}^2_{\rho_X}} \leq n^{-2\vartheta r} \log^4\left(\frac{2}{\delta}\right) + \mathtt{V}_1 + \mathtt{V}_2 \,, \quad (8)$$

where $\mathtt{V}_1$ and $\mathtt{V}_2$ are the same as in Theorem 1.

Remark: Compared to classical learning theory results [47] achieving $\mathcal{O}(n^{-\frac{2r+1}{2r+1+\eta}})$ learning rates, the parameter η in our results only effects the selection range of λ, which is nearly independent of the learning rates to some extent. That means, the spectral decay of a kernel function k in high dimensions is almost irrelevant to its kernel type. In fact, the eigenvalue decay of the kernel matrix in our model largely depends on the data, which is in essence different from classical learning theory results. Therefore, our result reflects a certain "universality" on the kernel function in high dimensional problems, which shows consistency to [26].

3.3 Related work

We provide non-asymptotic results that systematically analyze both implicit and explicit regularization schemes within a unified framework.

Implicit regularization in kernel/linear interpolation: Implicit regularization can be induced by minimum norm solutions in linear interpolation [48, 49], or the curvature of the kernel function in kernel interpolation [18]. Compared to the risk curve in [18] that converges to a non-zero constant, the risk curve in our results tends to zero when $n \gg d$. Hence our result demonstrates that, in the double descent case, the minimum of the expected risk in the second descent is lower than the first descent; while the same claim cannot be obtained from [18]. Besides, under the basic $f_\rho \in \mathcal{H}$ case, our bias bound is based on the eigen-decay (trends) of the kernel matrix $\boldsymbol{K}$ and thus can be (almost) independent of d, achieving an optimal learning rate $\mathcal{O}(\lambda)$ in a minimax case. This is different from [18] that corresponds to the sum of tailed eigenvalues of $\boldsymbol{K}$. Specifically, if we directly set λ to zero, our result for the bias still holds, which can be bounded by $\|L_{K,\boldsymbol{X}} - L_K\|_{\mathcal{L}^2_{\rho_X}} \lesssim \mathcal{O}(1/\sqrt{n})$.

Explicit regularization in kernel/linear regression: We provide non-asymptotic results that refine a series of asymptotic analyses, e.g., the Stieltjes transform approach in [2, 27, 33, 50] and the statistical mechanic approach in [51]. In fact, by considering the limiting eigenvalue distribution of $\boldsymbol{X}\boldsymbol{X}^\top/d$ via its Stieltjes transform $\frac{1}{n}\mathcal{N}^b_{\boldsymbol{X}\boldsymbol{X}^\top/d} \approx m(-b) - bm'(-b)$, for $m(b)$ the solution to the popular Marčenko–Pastur equation [52], our error bound recovers [2, Theorem 5] with $b := \lambda$ and isotropic features $\boldsymbol{\Sigma}_d = \boldsymbol{I}_d$. Finite sample analyses are often based on a finer control of the Stieltjes transform [41] or the effective rank [3, 53]. However, the aforementioned results are generally limited to Gaussian [41, 42] and sub-Gaussian data [3, 53, 54], or Gaussian covariates [55]. Here we consider a much broader family of distributions. Besides, under some specific situations, the regularization parameter λ in (generalized) linear regression can be negative [49] or optimal tuned [42, 9] so as to generalize well. Recent research [21, 12, 23] on kernel regression in $n := \mathcal{O}(d^c)$ shows different trends.

3.4 Proof framework

The proof of our results is fairly technical and lengthy, and we briefly sketch some main ideas of Theorem 2 here. Note that, Theorem 1 is a special case of Theorem 2 by taking $r = 1/2$ and $\eta = 1$. The modified error decomposition, the error bounds of variance for radial kernels, and estimates for bias are the main elements of novelty in the proof.

In order to estimate the error $\mathbb{E}_{y|\boldsymbol{x}} \|f_{\boldsymbol{z},\lambda} - f_\rho\|$ in the $\mathcal{L}^2_{\rho_X}$ space, we need the following intermediate functions. Define $f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho$, where I is the identity operator, then f_λ is actually the minimizer of the following problem $f_\lambda = \operatorname{argmin}_{f \in \mathcal{H}} \left\{ \|f - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} + \lambda \|f\|^2_{\mathcal{H}} \right\}$.

Besides, by defining

$$f_{\boldsymbol{X},\lambda}(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{X})^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} f_\rho(\boldsymbol{x}) \,,$$

we have $f_{\boldsymbol{X},\lambda} = (L_{K,\boldsymbol{X}} + \lambda I)^{-1} L_{K,\boldsymbol{X}} f_\rho$. Accordingly, the variance-bias decomposition is stated in the following lemma, with proof deferred to Appendix C.

Lemma 1. *Let $f_{\boldsymbol{z},\lambda}$ be the minimizer of problem (1), $\mathbb{E}_{y|\boldsymbol{x}} \| f_{\boldsymbol{z},\lambda} - f_\rho \|^2_{\mathcal{L}^2_{\rho_X}}$ can be bounded by*

$$\begin{aligned} \mathbb{E}_{y|\boldsymbol{x}} \big\| f_{\boldsymbol{z},\lambda} - f_\rho \big\|^2_{\mathcal{L}^2_{\rho_X}} &= \mathtt{B} + \mathtt{V} \\ &\leq 2 \left(\| f_{\boldsymbol{X},\lambda} - f_\lambda \|^2_{\mathcal{L}^2_{\rho_X}} + \| f_\lambda - f_\rho \|^2_{\mathcal{L}^2_{\rho_X}} \right) + \mathtt{V} \end{aligned}$$

where the bias $\mathtt{B}$ is defined as

$$\mathtt{B} := \mathbb{E}_{\boldsymbol{x}} \big\| k(\boldsymbol{x}, \cdot)^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} f_\rho(\boldsymbol{X}) - f_\rho \big\|^2_{\mathcal{L}^2_{\rho_X}} \,, \quad (9)$$

where $f_\rho(\boldsymbol{X}) = [f_\rho(\boldsymbol{x}_1), f_\rho(\boldsymbol{x}_1), \cdots, f_\rho(\boldsymbol{x}_n)]^\top \in \mathbb{R}^n$ and the variance $\mathtt{V}$ is defined as

$$\mathtt{V} := \mathbb{E}_{\boldsymbol{x},y} \left\| k(\boldsymbol{x}, \cdot)^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} \boldsymbol{\epsilon} \right\|^2_{\mathcal{L}^2_{\rho_X}} \,, \qquad (10)$$

where $\boldsymbol{\epsilon} := \boldsymbol{y} - f_\rho(\boldsymbol{X})$ satisfying $\mathbb{E}_{y|\boldsymbol{x}}[\boldsymbol{\epsilon}] = 0$.

It is clear that, the variance term does not depend on the target function f_ρ, and the bias is independent of the residual error $\boldsymbol{\epsilon}$. Proof for the bias $\mathtt{B} \lesssim n^{-2\vartheta r} \log^4 \left(\frac{2}{\delta} \right)$ can be found in Appendix D. Proof for the variance $\mathtt{V} \lesssim \mathtt{V}_1 + \mathtt{V}_2$ refers to Appendix E.

4 Discussion on Error Bounds

In this section, we discuss our Theorem 2 for different eigenvalue profiles of $\widetilde{\boldsymbol{X}}$ in the two regimes of $n < d$ and $n > d$. Since $\boldsymbol{K}$ shares the same eigenvalue decay as $\boldsymbol{X}\boldsymbol{X}^\top/d$ and $\widetilde{\boldsymbol{X}}$ (see Proposition 2 in Appendix B), we do not distinguish the eigen-decay of these two data matrices in the subsequent discussions. We first focus on the variance $\mathtt{V}$ that can be unimodal or monotonically decreasing with n under different regularization schemes. Subsequently, we investigate the total risk curve as the sum of bias and variance. Note that $\widetilde{\boldsymbol{X}}$ has different numbers of non-zero eigenvalues under the two regimes, we denote $r_* := \operatorname{rank}(\widetilde{\boldsymbol{X}}) \leq \min\{n, d\}$, which, as we shall see, plays a significant role in characterizing the different cases of our bounds.

4.1 Variance trend for $n < d$

We consider here three eigenvalue decays of $\widetilde{\boldsymbol{X}}$: *harmonic*, *polynomial*, and *exponential decay* [56, 45].

Proposition 1. *Under the three eigenvalue decays in Table 3, denote $r_* = \operatorname{rank}(\widetilde{\boldsymbol{X}})$, then the quantity function $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}}$ with $b := n\lambda + \gamma$ can be bounded by*
1) harmonic decay: $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} \leq \frac{n}{b^2} \ln \frac{n + (r_* + 1)b}{n + b} = \mathcal{O}(\frac{n}{b^2})$.
2) polynomial decay: $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} \leq \frac{\widetilde{C}}{2ab} \left(\frac{n}{b} \right)^{\frac{1}{2a}}$, where $\widetilde{C}$ is some constant.
3) exponential decay: $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} \leq \frac{1}{a} \left(\frac{1}{b + ne^{-a(r_*+1)}} - \frac{1}{b + ne^{-a}} \right)$.

Table 3: Three eigenvalue decays of $\widetilde{\boldsymbol{X}}$.

eigenvalue decay	$\lambda_i(\widetilde{\boldsymbol{X}})$	
	$i \leq r_*$	$i > r_*$
harmonic decay	n/i	0
polynomial decay	ni^{-2a} with $a > 1/2$	
exponential decay	ne^{-ai} with $a > 0$	

Proof. The proof can be found in Appendix F. □

According to Proposition 1, we summarize our results in Table 1 and discuss them as follows:

Harmonic decay: $\mathtt{V}_1 \leqslant \mathcal{O}(\frac{n}{b^2 d})$.

For $\lambda = 0$, i.e., the ridgeless case, we have $b = \gamma = \mathcal{O}(1)$, and $\mathtt{V}_1 \leq \mathcal{O}(\frac{n}{d})$, which indicates $\mathtt{V}_1$ increases with n in the $n < d$ regime. For $\lambda \neq 0$, taking $\lambda := \bar{c} n^{-\vartheta}$, we have $\mathtt{V}_1 \leqslant \mathcal{O}(\frac{n}{d(\bar{c} n^{1-\vartheta} + \gamma)^2})$. To investigate the monotonicity of $g(n) := \frac{n}{d(\bar{c} n^{1-\vartheta} + \gamma)^2}$, define $n_* := \left(\frac{\gamma}{2 - 2\vartheta - \bar{c}} \right)^{\frac{1}{1-\vartheta}}$, we find that, a large λ leads to a small n_*. According to the relationship between r_*, n_*, and d, we can conclude that (see Table 1 and the red curve in Figure 1(a)):

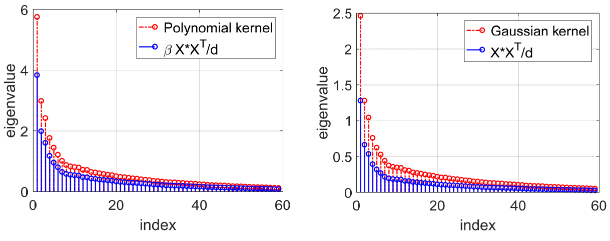
When $\vartheta \geq \frac{1}{2(2-\bar{c})}$, $\mathtt{V}_1$ will increase with n until $n := r_*$ and then remain unchanged when $r_* < n < d$.
When $\vartheta < \frac{1}{2(2-\bar{c})}$, there are various trends as follows:
1) if $d < n_*$, this is the same as the $\vartheta \geq \frac{1}{2(2-\bar{c})}$ case;
2) if $r_* < n_* < d$, $\mathtt{V}_1$ will increase with n until $n := r_*$, and then remain unchanged when $r_* < n < d$;
3) if $n_* < r_* < d$, $\mathtt{V}_1$ will increase with n until $n := n_*$ and then decrease with n until $n := r_*$, and stay unchanged on $r_* < n < d$;
4) If $n_* < c$ such that $n > c$ always holds for some constant c, we have $\mathtt{V}_1$ increases with n until $n := r_*$, and then stays unchanged on $r_* < n < d$. Remark that, for $\gamma < 2 - 2\vartheta - \bar{c}$, we have $n_* < 1$ and thus $n > n_*$, so that $\mathtt{V}_1$ always decreases with n until $n := r_*$.

Polynomial decay: $\mathtt{V}_1 \leqslant \mathcal{O}(\frac{1}{bd} (\frac{n}{b})^{\frac{1}{2a}})$.

Similar to above, define $n_* = \left(\frac{\gamma}{2a\bar{c}[1 - (1 + \frac{1}{2a})\vartheta]} \right)^{\frac{1}{1-\vartheta}}$, we obtain results similar to the case of *harmonic decay*, but with different thresholds: $\vartheta \geq (1 + \frac{1}{2a})^{-1}$ and $\vartheta < (1 + \frac{1}{2a})^{-1}$, see Table 1 for details.

Exponential decay: $\mathtt{V}_1 \leq \frac{\widetilde{C}\beta}{ad} (\frac{1}{b + ne^{-a(r_*+1)}} - \frac{1}{b + ne^{-a}})$.

Here we consider the monotonicity of the function $G(n) := (\frac{1}{b + ne^{-a(r_*+1)}} - \frac{1}{b + ne^{-a}})$ with $b := n\lambda + \gamma$ to



(a) *poly kernel with order 3* (b) *Gaussian kernel*

Figure 2: Top 60 eigenvalues of two kernel matrices and their linearizations on the subset of the *YearPredictionMSD* dataset. Note that the largest eigenvalue λ_1 is not plotted for better display.

study the trend of $\mathtt{V}_1$ regarding to n. Let n_* be the solution of the equation $G'(n) = 0$, then we have the similar conclusion with that of *harmonic decay* and *polynomial decay* by the relationship between n_*, r_*, and d, see Table 1 for details. More specifically, under some certain conditions, $\mathtt{V}_1$ is able to monotonically decrease with n, refer to Appendix F.1 for details.

4.2 Variance trends for $n > d$ and total risk

Different from the above $n < d$ case, the current $n > d$ regime admits that $\widetilde{\boldsymbol{X}}$ has at most d non-zero eigenvalues. In this under-parameterized regime, we are particularly interested in the behavior as $n \to \infty$. In Appendix F.2, we prove that $\mathtt{V}_1$ approaches to zero as $n \to \infty$ under the above three eigenvalue decays.

Based on the above discussions in the $n > d$ and $n < d$ regimes, we conclude that, the variance can be unimodal (small regularization) or decreasing (large regularization) as n grows, which, together with the fact that the bias is monotonically decreasing with n, leads to the following three configurations for the total risk: (i) if the bias dominates at small n and then decays fast (i.e., with a small regularization), we observe a double descent curve as in Figure 1(b); (ii) if the bias dominates but decays slowly (with a large regularization), the risk curve will be monotonic decreasing as in Figure 1(d); (iii) if the variance dominates, a bell-shaped risk curve as in Figure 1(c) will be observed.

5 Numerical Results

In this section, experiments are conducted to validate our theoretical results[1]. Polynomial kernel of degree 3 and Gaussian kernel are evaluated on 1) a synthetic dataset that satisfies our technical assumptions and 2) a subset of the *YearPredictionMSD* dataset [57] with 1,000 data samples and $d = 90$, to study our derived error bounds for the bias and variance. More experimental results can be found in Appendix G.

Eigenvalue decay equivalence: Here we study the eigenvalue decay of the original polynomial/Gaussian kernel matrices and their linearization $\boldsymbol{X}\boldsymbol{X}^\top/d$ on the subset of *YearPredictionMSD* dataset. Note that, polynomial kernels $k(\boldsymbol{x}, \boldsymbol{x}') := (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle/d)^p$ admit $\beta := p$ independent of $\boldsymbol{\Sigma}_d$ (see in Table 4), so we use the linearization $\beta\boldsymbol{X}\boldsymbol{X}^\top/d$ for this kernel. Results in Figure 2 demonstrate that, the original nonlinear kernels admit the same eigenvalue decay as $\boldsymbol{X}\boldsymbol{X}^\top/d$. More experimental results on various dataset can be found in Appendix G.1.

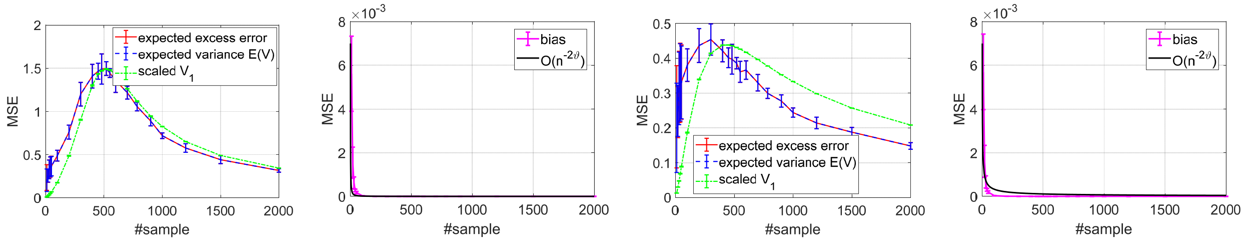
Risk curves on synthetic dataset: To quantitatively assess our derived error bounds for the bias and variance, we generate a synthetic dataset under a known f_ρ, with *harmonic decay* for the data as an illustrating example. More experimental results on different eigenvalue decays refer to Appendix G.2. To be specific, we assume $y_i = f_\rho(\boldsymbol{x}_i) + \varepsilon$ with target function $f_\rho(\boldsymbol{x}) = \sin(\|\boldsymbol{x}\|_2^2)$ and Gaussian noise ε having zero-mean and unit-variance. The feature dimension d is set to 500. The samples are generated from $\boldsymbol{x}_i = \boldsymbol{\Sigma}_d^{1/2}\boldsymbol{t}_i$ (and thus $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{T}^\top\boldsymbol{\Sigma}_d\boldsymbol{T}$ with $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n]^\top$) by the following steps: (i) take $\boldsymbol{\Sigma}_d$ as a diagonal matrix with its diagonal entries following with *harmonic decay*, i.e., $(\boldsymbol{\Sigma}_d)_{ii} \propto n/i$. (ii) take $\boldsymbol{T}$ as a random orthogonal matrix[2] such that $\boldsymbol{T}^\top\boldsymbol{\Sigma}_d\boldsymbol{T}$ also has a harmonic eigen-decay with $\boldsymbol{T}$ having almost i.i.d entries.

Accordingly, the above generation process satisfies Assumption 3, and also $\boldsymbol{X}\boldsymbol{X}^\top/d$ admits the same eigenvalue decay as $\boldsymbol{\Sigma}_d$, which can be used to validate our discussion in Section 4. In this setting, the expected excess risk, the bias, and the variance can be directly computed to validate our derived error bounds. The experimental results are validated across 10 trials. Specifically, to disentangle the *implicit regularization* effect of KRR on the final result, we apply the linearization of the polynomial/Gaussian kernel by setting $\gamma = 0$ in Eq. (5). In this case, the explicit $\lambda := \bar{c}n^{-\vartheta}$ is the only regularization in KRR. In our model, $\bar{c}$ is empirically set to 0.01 to avoid a large λ when n is small.

Figures 3 and 4 show results under the *harmonic decay* setting for the linearization of the polynomial/Gaussian kernel, respectively. We observe that: 1) our error bound $\mathtt{V}_1 \asymp \frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda}$ exhibits the same trend as the true variance; 2) in this case, the variance dominates and we thus obtain a bell-shaped risk curve that first increases and then decreases; 3) as ϑ decreases, λ increases and the peak point of the variance occurs at smaller and smaller n; 4) the bias monotonically decreases with n,

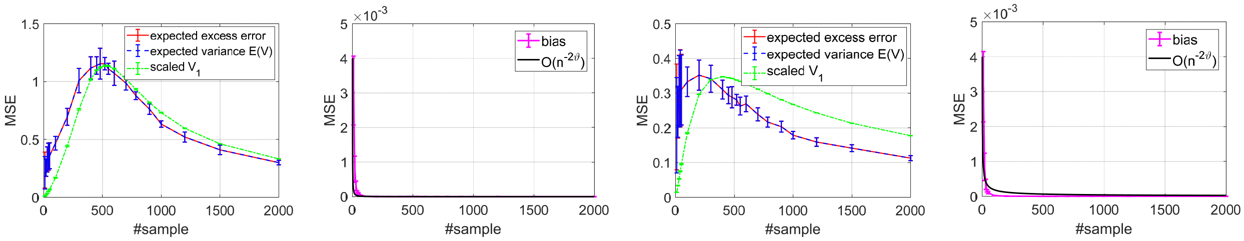
[1] The source code of our implementation can be found in `http://www.lfhsgre.org`.

[2] We generate a random Gaussian matrix and use the QR decomposition to obtain an orthogonal matrix [58].



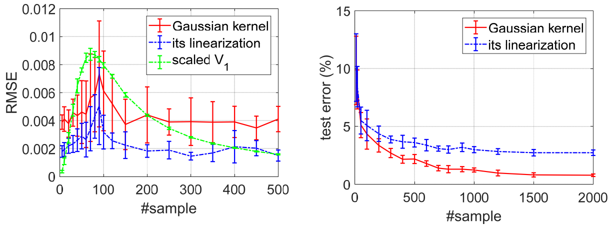
(a) $\vartheta = 2/3$ (b) $\vartheta = 2/3$ (c) $\vartheta = 1/3$ (d) $\vartheta = 1/3$

Figure 3: Harmonic decay of $\widetilde{\boldsymbol{X}}$ with polynomial kernel: MSE of the expected excess risk, the variance in Eq. (10), our derived $\mathtt{V}_1$, the bias in Eq. (9), and our derived convergence rate $\mathcal{O}(n^{-2\vartheta r})$ with $r = 1$ for different ϑ.



(a) $\vartheta = 2/3$ (b) $\vartheta = 2/3$ (c) $\vartheta = 1/3$ (d) $\vartheta = 1/3$

Figure 4: Harmonic decay of $\widetilde{\boldsymbol{X}}$ in the Gaussian kernel case. The legend is the same as Figure 3.



(a) *YearPredictionMSD* (b) *MNIST* (digits 3 vs. 7)

Figure 5: The test performance of the kernel interpolation estimator and its linearization one.

which corresponds to our error bound for the bias at a certain $\mathcal{O}(n^{-2\vartheta r})$ rate in Theorem 2 by taking $r = 1$ as the used f_ρ is smooth enough to achieve a good approximation error; 5) in our high-dimensional regimes, different kernels lead to the same convergence rates of the bias, which verifies our results but is different from those in classical learning theory.

Risk curves on the real-world datasets: Figure 5(a) shows the relative mean squared error (RMSE) of kernel ridgeless regression and its linearization in Eq. (5) on a subset (1,000 examples) of the *YearPredictionMSD* dataset averaged on 10 trials. Figure 5(b) shows the classification accuracy of such two methods on the *MNIST* dataset [59]. To evaluate the effectiveness of our error bounds, we plot the re-scaled $\mathtt{V}_1 \asymp \frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{\gamma}$ with $\lambda = 0$. It can be found that, kernel interpolation estimator generalizes well due to the *implicit regularization*, i.e., $\gamma \neq 0$, which also exhibits a bell-shaped risk curve as our theoretical results suggest. However, in Figure 5(b), the risk curve monotonically decreases with n on the *MNIST* dataset [59], and at the same time kernel interpolation estimator and its linearization appear to generalize well. This observation may due to the *implicit regularization* parameter γ in Eq. (5) (of 10^{-3} order on this dataset) that plays a fundamental role of "self-regularization". Accordingly, the proposed analysis provides access to the high-dimensional classification problem that may establish more involved behavior than double descent, despite a clear mismatch between real-world data and the technical Assumption 3, thereby conveying a strong practical motivation for the present analysis.

6 Conclusion

We derived non-asymptotic expressions for the expected excess risk of kernel ridge regression estimators in the under- and over-determined regimes. The used linearization technique of nonlinear smooth kernel allows us to discuss the impact of *implicit* and *explicit* regularization in a systematic manner. Our refined analysis demonstrates that the monotonic bias and unimodal variance are able to exhibit various trends of risk curves. Since it is enough to require that the kernel function is differentiable in a neighborhood, our results further extend to the case of Laplace kernels [60].

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. This work was supported in part by Research Council KU Leuven: Optimization frameworks for deep kernel machines C14/18/068; Flemish Government: FWO projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations), PhD/Postdoc grant. This research received funding from the Flemish Government (AI Research Program). This work was supported in part by Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms), EU H2020 ICT-48 Network TAILOR (Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization), Leuven.AI Institute.

References

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

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

[3] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *the National Academy of Sciences*, 2020.

[4] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *the National Academy of Sciences*, 116(32):15849–15854, 2019.

[6] Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan A.K. Suykens. Random features for kernel approximation: A survey in algorithms, theory, and beyond. *arXiv preprint arXiv:2004.11154*, 2020.

[7] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[8] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462, 2020.

[9] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

[10] Zhenyu Liao, Romain Couillet, and Michael Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *Neural Information Processing Systems*, 2020.

[11] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

[12] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Annual Conference on Learning Theory*, pages 1–32, 2019.

[13] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[14] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[16] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1370–1378, 2019.

[17] Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In *International Conference on*

Artificial Intelligence and Statistics, pages 3525–3535, 2020.

[18] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel “ridgeless” regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

[19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics*, pages 1611–1619, 2019.

[20] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Annals of Statistics*, 2019.

[22] Weilin Li. Generalization error of minimum weighted norm and kernel interpolation. *arXiv preprint arXiv:2008.03365*, 2020.

[23] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1–11, 2020.

[24] Johan A.K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines.* World Scientific, 2002.

[25] Fanghui Liu, Lei Shi, Xiaolin Huang, Jie Yang, and Johan A.K. Suykens. Analysis of regularized least squares in reproducing kernel kreĭn spaces. *Machine Learning*, pages 1–20, 2021.

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

[27] Khalil Elkhalil, Abla Kammoun, Xiangliang Zhang, Mohamed-Slim Alouini, and Tareq Al-Naffouri. Risk convergence of centered kernel ridge regression with large dimensional data. *IEEE Transactions on Signal Processing*, 68:1574–1588, 2020.

[28] Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *the International Conference on Machine Learning*, pages 3063–3071, 2018.

[29] Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019.

[30] Felipe Cucker and Dingxuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

[31] Cheng Wang and Ding-Xuan Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.

[32] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

[33] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *the International Conference on Machine Learning*, 2020.

[34] Ingo Steinwart and Christmann Andreas. *Support Vector Machines.* Springer Science and Business Media, 2008.

[35] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

[36] Dominic Richards and Patrick Rebeschini. Optimal statistical rates for decentralised non-parametric regression with linear speed-up. In *Advances in Neural Information Processing Systems*, pages 1216–1227, 2019.

[37] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

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

[39] Gilles Blanchard and Nicole Krämer. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Information Processing Systems*, pages 226–234, 2010.

[40] Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2020.

[41] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. In *Advances in Neural Information Processing Systems*, pages 1–9, 2020.

[42] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

[43] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013.

[44] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *the 34th International Conference on Machine Learning*, pages 253–262, 2017.

[45] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random Fourier features. In *the 36th International Conference on Machine Learning*, pages 3905–3914, 2019.

[46] Nicolò Pagliana, Alessandro Rudi, Ernesto De Vito, and Lorenzo Rosasco. Interpolation and learning with scale dependent kernels. *arXiv preprint arXiv:2006.09984*, 2020.

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

[48] Michał Dereziński, Feynman Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019.

[49] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

[50] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640, 2020.

[51] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Statistical mechanics of generalization in kernel regression. *arXiv preprint arXiv:2006.13198*, 2020.

[52] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[53] Geoffrey Chinot and Matthieu Lerasle. Benign overfitting in the large deviation regime. *arXiv preprint arXiv:2003.05838*, 2020.

[54] Emmanuel Caron and Stephane Chretien. A finite sample analysis of the double descent phenomenon for ridge function estimation. *arXiv preprint arXiv:2007.12882*, 2020.

[55] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. *arXiv preprint arXiv:2006.06386*, 2020.

[56] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

[57] Chih-Chung Chang. LibSVM data: Classification, regression, and multi-label. *http://www. csie. ntu. edu. tw/˜cjlin/libsvmtools/datasets/*, 2008.

[58] Felix Xinnan Yu, Ananda Theertha Suresh, Krzysztof Choromanski, Daniel Holtmannrice, and Sanjiv Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems*, pages 1975–1983, 2016.

[59] Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *the IEEE*, 86(11):2278–2324, 1998.

[60] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623, 2019.

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

[62] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

[63] Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *Journal of Machine Learning Research*, 18(1):4237–4261, 2017.

[64] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075, 2013.