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] 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-
Table 1: Trends of the variance $V$ with respect to $n$ in the $n < d$ case. The notation $\Rightarrow$ means $V$ increases with $n$; $\Leftarrow$ for $V$ stays unchanged; and $\nLeftarrow$ for $V$ decreases with $n$, see Figure 1(a) and $r_* := \text{rank}(XX^\top)$.

<table>
<thead>
<tr>
<th>eigenvalue decay</th>
<th>$\lambda = 0$</th>
<th>$\lambda := \bar{c}n^{-\vartheta}$ (KRR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>harmonic decay</td>
<td>$\Rightarrow$</td>
<td>$\Rightarrow$ 1 $\geq \vartheta \geq \frac{1}{2(2-c)}$</td>
</tr>
<tr>
<td></td>
<td>$\Rightarrow$</td>
<td>$r_* &lt; d \leq n_* \Rightarrow r_* \leq n_* \leq d \Rightarrow n_* \leq r_* &lt; d \Rightarrow n_* \leq c &lt; r_* &lt; d$, $\Rightarrow$</td>
</tr>
<tr>
<td>polynomial decay</td>
<td>$\Rightarrow$</td>
<td>$\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$</td>
</tr>
<tr>
<td></td>
<td>$\Rightarrow$</td>
<td>$\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$</td>
</tr>
<tr>
<td>exponential decay</td>
<td>$\Rightarrow$</td>
<td>$\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$ $\Rightarrow$</td>
</tr>
</tbody>
</table>

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

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.

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 $XX^\top/d$, where $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ is the data matrix. So in high dimensions, the eigenvalue decay of $K$ 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 $\lambda$ 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 $r_* := \text{rank}(XX^\top/d)$: the error bound for variance $V$ monotonically increases with $n$ until $n := n_*$, as in the red curve of Figure 1(a). Under a moderate regularization with $n_* \leq r_* < d$, $V$ 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).

dimensions [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 $z := \{(x_i, y_i)\}_{i=1}^n$ with data $x_i \in \mathbb{R}^d$ and responses $y_i \in \mathbb{R}$:

$$f_{z, \lambda} := \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( f(x_i) - y_i \right)^2 + \lambda \langle f, f \rangle_{\mathcal{H}} \right\}, \quad (1)$$

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.
regularization with \( n_s \leq c \) for some constant \( c \), \( \mathcal{V} \) 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].

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 properties 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 \( C \) independent of \( n \) such that \( a(n) \leq Cb(n) \), and analogously for \( \lesssim \) 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 \((x_i, y_i)\) in the training set \( z =\{ (x_i, y_i)\}_{i=1}^n \subseteq \mathbb{Z}^n \) are assumed to be independently drawn from a non-degenerate Borel probability measure \( \rho \) on \( X \times Y \). The target function of \( \rho \) is defined by

\[
\rho(x) = \int_Y y\,d\rho(y | x), \quad x \in X ,
\]

where \( \rho(\cdot | x) \) is the conditional distribution of \( \rho \) at \( x \in X \). Define the response vector \( y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^n \) and the kernel matrix \( K = \{ k(x_i, x_j)\}_{i,j=1}^n \) induced by a positive definite kernel \( k(\cdot, \cdot) \). KRR aims to find a hypothesis \( f : X \to Y \) such that \( f(x) \) is a good approximation of the response \( y \in Y \) corresponding to a new instance \( x \in X \). This is actually an empirical risk minimization in problem (1).

We consider two popular positive definite kernel classes of (i) the inner-product kernel of the form \( k(x_i, x_j) = h(\|x_i - x_j\|)^2/d \). Here \( h(\cdot) : \mathbb{R} \to \mathbb{R} \) is 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 \( \mathcal{E}(f) = \int_X (f(x) - y)^2\,d\rho(x) \) and the empirical risk functional is defined on the training set \( z \), i.e., \( \mathcal{E}_z(f) = \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 \). To measure the estimation quality of \( f_{z,\lambda} \), one natural way is the expected excess risk: \( \mathbb{E}_{y|x}[\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_{\rho})] \). Specifically, in KRR, the expected excess risk admits \( \mathbb{E}_{y|x}[\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_{\rho})] = \mathbb{E}_{y|x}[\|f_{z,\lambda} - f_{\rho}\|^2] \), which is exactly in the weighted \( L^2 \)-space with the norm \( \|f\|^2 = \int_X |f(x)|^2\,d\rho_X(x) \).

2.2 Background on RKHS

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

\[
(L_K f)(\cdot) = \int_X k(\cdot, x) f(x) d\rho_X(x), \quad \forall f \in L^2_{\rho_X} \cdot
\]

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 \( \{\psi_i\}_{i=1}^\infty \) are orthogonal basis of \( L^2_{\rho_X}(X) \) and \( \mu_1 \geq \mu_2 \cdots > 0 \) with \( \lim \mu_i = 0 \). Accordingly, by Mercer’s theorem, we have \( k(x, x') = \sum_{i=1}^{\infty} \mu_i \psi_i(x) \psi_i(x) \), and there exists a constant \( \kappa \geq 1 \) such that \( \sup_{x \in X} \sum_{i=1}^{\infty} \mu_i \psi_i^2(x) \leq \kappa^2 \). It holds by \( \kappa := \max\{1, \sup_{x \in X} \sqrt{k(x, x)}\} \). Based on the data matrix \( X \) and the integral operator \( L_K \), the empirical integral operator is given by \( \hat{L}_K(X) = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i) \otimes k(\cdot, x_i) \), 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 $\tilde{K}_{\text{lin}}$ in [26].

<table>
<thead>
<tr>
<th>parameters</th>
<th>inner-product kernels</th>
<th>radial kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$h(0) + h'(0)\frac{|X|^2}{2\sigma^2}$</td>
<td>$h(2\tau) + h''(2\tau)\frac{|X|^2}{2\sigma^2}$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$h'(0) - 2h''(2\tau)$</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$h(\tau) - h(0) - \tau h'(0)$</td>
<td>$h(0) + 2\tau h''(2\tau) - h(2\tau)$</td>
</tr>
<tr>
<td>$\E$</td>
<td>$\tr(\Sigma_d)$</td>
<td>$\tr(\Sigma_d)$</td>
</tr>
</tbody>
</table>

### Entry (1) with a bounded spectral norm $\|\Sigma_d\|_2$.

Assumption 2. (Noise condition [18, 38]) There exists a bounded spectral norm $\|\Sigma_d\|_2$.

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

$$\tilde{K}_{\text{lin}} := \alpha 11^T + \beta \frac{XX^T}{d} + \gamma I + E,$$  

with non-negative parameters $\alpha$, $\beta$, $\gamma$, and the additional matrix $E$ given in Table 2, see some typical examples in Appendix A. Here $\gamma$ is the implicit regularization parameter in kernel estimator that depends on the nonlinear function $h$ in the kernel $k$ and the data structure $\Sigma_d$. According to Eq. (5), denote the shortcut $\tilde{X} := \beta XX^T/d + \alpha 11^T$, we show in high dimensions that, $K$ admits the same eigenvalue decay as $\tilde{X}$ and $XX^T/d$ (see details in Appendix B). Subsequently, we introduce the following quantity function

$$\mathcal{N}_{\tilde{X}} := \tr\left((\tilde{X} + hI_n)^{-2}\tilde{X}\right) = \sum_{i=1}^{n} \frac{\lambda_i(\tilde{X})}{\left[b + \lambda_i(\tilde{X})\right]^2},$$

which is associated with various quantity functions in [16, 58, 13, 11, 42] and, as we shall see, plays an important role in determining the variance behavior. We will discuss at length $\mathcal{N}_{\tilde{X}}$ based on different data eigenvalue decays in Section 3.

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 $f_\rho$) We assume $f_\rho \in \mathcal{H}$. This is a standard assumption in learning theory and assumes that the target function $f_\rho$ defined in Eq. (2) is indeed realizable, see also [26, 37, 30, 32].

Assumption 2. (Noise condition [18, 38]) There exists a standard assumption in learning theory and is in fact weaker than the standard Bernstein condition, e.g., in [39].

Assumption 3. (8+\text{m}-moments [18, 40]) Let $x_i = \Sigma_d^{-1/2}t_i$, where $t_i \in \mathbb{R}^d$ has i.i.d. entries with zero mean, unit variance, and a finite (8+\text{m})-moments, i.e., its entry $t_i(j)$, $1 \leq j \leq d$, satisfies $\E[t_i(j)] = 0$, $\forall [t_i(j)] = 1$, and $\E([t_i(j)]) \leq C d 2^{\text{m}z}$ such that $\E[x_i x_i^T] = \Sigma_d$ with a bounded spectral norm $\|\Sigma_d\|_2$, 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 $K$ in high dimensions can be well approximated by $\tilde{K}_{\text{lin}}$ in spectral norm, i.e., $\|K - \tilde{K}_{\text{lin}}\|_2 \to 0$ as $n, d \to \infty$.

$$\tilde{K}_{\text{lin}} := \alpha 11^T + \beta \frac{XX^T}{d} + \gamma I + E,$$  

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 $V_1 + V_2$ is the bound of the variance that depends on both $n$ and $d$. Note that $V_2$ monotonically decreases with $n$, and approaches to zero for a large $n$. Therefore, the error bound for $V_1 \approx \frac{1}{2} N^{n\lambda + \gamma}$ is the key part of estimates for the variance and will be discussed in in Section 3 where $n\lambda$ corresponds to the explicit regularization and $\gamma$ the implicit regularization. We will demonstrate that $V_1$ 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 $f_\rho$ 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 $g_{\rho} \in L_{\rho x}^2$ satisfying $\|g_{\rho}\|_{L_{\rho x}^2} \leq R$ 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 $f_{\rho}$ [30, 43, 37], which indicates $f_{\rho}$ belongs to the range space of $L_K^r$. Assumption 1 is the worst case of Assumption 4 by choosing $r = 1/2$ since $\|f\|_{L_{\rho x}^2} = \|L_K^2 f\|_{\mathcal{H}}, \forall f \in L_{\rho x}^2$.

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

$$N(\lambda) := \text{tr} ((L_K + \lambda I)^{-1} L_K) \leq Q^2 \lambda^{-\eta}.$$  

The notation $N(\lambda)$ 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, 63, 37]. Assumption 5 always holds for $\eta = 1$ and $Q = \kappa$ where $\kappa := \max\{1, \sup_{x \in X} \sqrt{\kappa(x, x)}\}$ as $L_K$ is a trace class operator. Its kernel matrix form is $d^0_K := \text{tr} ((K + \lambda I_n)^{-1} K) = \sum_{i=1}^{n} \frac{\lambda_i(K)}{\lambda_i(K) + \lambda}$.

While Assumption 5 can be further refined to obtain a bound that depends on $d$ [40], here we focus on the eigenvalue decay of $K$, 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 $0 < \delta < 1/2$, $\theta = \frac{1}{2} + \frac{2}{\alpha + 1}$, and $d$ large enough, taking $\lambda := \epsilon n^{-\theta}$ with $0 \leq \theta \leq \frac{1}{1 + \eta}$, then for any given $\epsilon > 0$, it holds with probability at least $1 - 2\delta - d^{-2}$

$$E_{y|x} \|f_{x, \lambda} - f_{\rho}\|_{L_{\rho x}^2}^2 \leq n^{-2\theta'} \log \left(\frac{2}{\delta}\right) + V_1 + V_2,$$  

where $V_1$ and $V_2$ are the same as in Theorem 4.

Remark: Compared to classical learning theory results [47] achieving $\mathcal{O}(n^{-\frac{2}{\alpha + 1 + \eta}})$ learning rates, the parameter $\eta$ in our results only effects the selection range of $\lambda$, 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 [38, 91], 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 $K$ and thus can be (almost) independent of $d$, achieving an optimal learning rate $\mathcal{O}(\lambda)$ in a minmax case. This is different from [18] that corresponds to the sum of tailed eigenvalues of $K$. Specifically, if we directly set $\lambda$ to zero, our result for the bias still holds, which can be bounded by $\|L_K, x - L_K\|_{L_{\rho x}^2} \lesssim \mathcal{O}(1/\sqrt{n})$.

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 [21, 12, 23] and the statistical mechanic approach in [51]. In fact, by considering the limiting eigenvalue distribution of $\mathbf{X}\mathbf{X}^\top$ as its Stieltjes transform $\frac{1}{n} \mathbf{X}\mathbf{X}^\top \approx m(-b) - b n^{-1/2}$, 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 $\Sigma_d = I_d$. Finite sample analyses are often based on a finer control of the Stieltjes transform [41] or the effective rank [3, 53, 54]. However, the aforementioned results are generally limited to Gaussian [41, 42] and sub-Gaussian data [3, 53, 54], or Gaussian covariates [52]. Here we consider a much broader family of distributions. Besides, under some specific situations, the regularization parameter $\lambda$ in (generalized) linear regression can be negative [49] or optimal tuned [22, 9] so as to generalize well. Recent research [21, 22, 23] on kernel regression in $n := \mathcal{O}(d^\epsilon)$ 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 $E_{y|x} \|f_{x, \lambda} - f_{\rho}\|_{L_{\rho x}^2}$ in the $L_{\rho x}^2$ 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 $f_{\lambda}$ is actually the minimizer of the following problem $f_{\lambda} = \arg\min_{f \in \mathcal{H}} \{ \|f - f_{\rho}\|_{L_{\rho x}^2}^2 + \lambda \|f\|_{\mathcal{H}}^2 \}.$
Besides, by defining
\[ f_{X, \lambda}(x) = k(x, X)^\top (K + n\lambda I)^{-1} f_\rho(x), \]
we have \( f_{X, \lambda} = (L_K X + \lambda I)^{-1} L_K f_\rho \). Accordingly, the variance-bias decomposition is stated in the following lemma, with proof deferred to Appendix \[ \square \]

**Lemma 1.** Let \( f_{X, \lambda} \) be the minimizer of problem \[ \square \], \( \mathbb{E}_{y|x} \| f_{X, \lambda} - f_\rho \|_2^2 \) can be bounded by
\[
\mathbb{E}_{y|x} \| f_{X, \lambda} - f_\rho \|_2^2 = B + V \leq 2 \left( \| f_X - f_\lambda \|_2^2 + \| f_\lambda - f_\rho \|_2^2 \right) + V
\]
where the bias \( B \) is defined as
\[
B := \mathbb{E}_{x} \| k(x, \cdot) \|_2^2 (K + n\lambda I)^{-1} f_\rho(X) - f_\rho \|_2^2 ,
\]
where \( f_\rho(X) = [f_\rho(x_1), f_\rho(x_1), \ldots, f_\rho(x_n)]^\top \in \mathbb{R}^n \) and the variance \( V \) is defined as
\[
V := \mathbb{E}_{x,y} \| k(x, \cdot) \|_2^2 (K + n\lambda I)^{-1} \epsilon \|_2^2 ,
\]
where \( \epsilon := y - f_\rho(X) \) satisfying \( \mathbb{E}_{y|x} \| \epsilon \| = 0 \).

It is clear that, the variance term does not depend on the target function \( f_\rho \), and the bias is independent of the residual error \( \epsilon \). Proof for the bias \( B \leq n^{-2\theta} \log^4 \left( \frac{n}{\theta} \right) \) can be found in Appendix \[ \square \]. Proof for the variance \( V \leq V_1 + V_2 \) refers to Appendix \[ \square \].

## 4 Discussion on Error Bounds

In this section, we discuss our Theorem \[ \square \] for different eigenvalue profiles of \( \tilde{X} \) in the two regimes of \( n < d \) and \( n > d \). Since \( K \) shares the same eigenvalue decay as \( XX^\top /d \) and \( \tilde{X} \) (see Proposition \[ \square \] in Appendix \[ \square \]), we do not distinguish the eigen-decay of these two data matrices in the subsequent discussions. We first focus on the variance \( V \) 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 \( \tilde{X} \) has different numbers of non-zero eigenvalues under the two regimes, we denote \( r_* := \text{rank}(\tilde{X}) \leq \min(n, d) \), 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 \( \tilde{X} \): harmonic, polynomial, and exponential decay \[ \square \].

**Proposition 1.** Under the three eigenvalue decays in Table \[ \square \], denote \( r_* = \text{rank}(\tilde{X}) \), then the quantity function \( N_{\tilde{X}}^{b} \) with \( b := n\lambda + \gamma \) can be bounded by

1) **harmonic decay:** \( N_{\tilde{X}}^{b} \leq \frac{\gamma}{2} \ln \frac{n+(r_*)^2/b}{n+b} = \mathcal{O}(\frac{\gamma}{2b}) \).

2) **polynomial decay:** \( N_{\tilde{X}}^{b} \leq \frac{C}{2b^2} \left( \frac{\gamma}{2} \right)^{\frac{1}{2}} \), where \( C \) is some constant.

3) **exponential decay:** \( N_{\tilde{X}}^{b} \leq \frac{1}{a} \left( \frac{1}{b+ne^{-a(r_*)^2}} - \frac{1}{b+ne^{-a}} \right) \).

**Proof.** The proof can be found in Appendix \[ \square \].

According to Proposition \[ \square \] we summarize our results in Table \[ \square \] and discuss them as follows:

### Harminic decay: \( V_1 \leq \mathcal{O}(\frac{n}{2\theta}) \).

For \( \lambda = 0 \), i.e., the ridgeless case, we have \( b = \gamma = \mathcal{O}(1) \), and \( V_1 \leq \mathcal{O}(\frac{n}{2}) \), which indicates \( V_1 \) increases with \( n \) in the \( n < d \) regime. For \( \lambda \neq 0 \), taking \( \lambda := \bar{c}n^{-\bar{d}} \), we have \( V_1 \leq \mathcal{O}(\frac{n}{d(n-\bar{c}n^{-\bar{d}})}) \). To investigate the monotonicity of \( g(n) := \frac{n}{d(n-\bar{c}n^{-\bar{d}})} \), define \( n_* := \left( \frac{n}{d(n-\bar{c}n^{-\bar{d}})} \right)^{\frac{1}{2\bar{d}}} \), we find that, a large \( \lambda \) leads to a small \( n_* \). According to the relationship between \( r_* \), \( n_* \), and \( d \), we can conclude that (see Table \[ \square \] and the red curve in Figure \[ \square \]):

When \( \theta \geq \frac{1}{2(2-\bar{d})} \), \( V_1 \) will increase with \( n \) until \( n = r_* \) and then remain unchanged when \( r_* < n < d \).

When \( \theta < \frac{1}{2(2-\bar{d})} \), there are various trends as follows:

1) if \( d < n_* \), this is the same as the \( \theta \geq \frac{1}{2(2-\bar{d})} \) case;
2) if \( r_* < n_* < d \), \( V_1 \) will increase with \( n \) until \( n = r_* \), and then remain unchanged when \( r_* < n < d \);
3) if \( n_* < r_* < d \), \( V_1 \) 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 \( V_1 \) increases with \( n \) until \( n = r_* \), and then stays unchanged on \( r_* < n < d \).

**Polynomial decay:** \( V_1 \leq \mathcal{O}(\frac{1}{b(1+(n+b)^{-\bar{d}})}) \).

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

**Exponential decay:** \( V_1 \leq \mathcal{O}(\frac{1}{b+ne^{-a(r_*)^2} - \frac{1}{b+ne^{-a}}}) \).

Here we consider the monotonicity of the function \( G(n) := \left( \frac{1}{b+ne^{-a(r_*)^2}} - \frac{1}{b+ne^{-a}} \right) \) with \( b := n\lambda + \gamma \) to
study the trend of $V_1$ 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, $V_1$ is able to monotonically decrease with $n$, refer to Appendix F.2 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 $\bar{X}$ 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 $V_1$ 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. 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 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.3.

Eigenvalue decay equivalence: Here we study the eigenvalue decay of the original polynomial/Gaussian kernel matrices and their linearization $X^T X/d$ on the subset of YearPredictionMSD dataset. Note that, polynomial kernels $k(x, x') := (1 + \langle x, x' \rangle/d)^p$ admit $\beta := p$ independent of $\Sigma_d$ (see in Table 1), so we use the linearization $\bar{X}X^T /d$ for this kernel. Results in Figure 2 demonstrate that, the original nonlinear kernels admit the same eigenvalue decay as $X^T X/d$. More experimental results on various dataset can be found in Appendix G.3.

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 $f_\rho$, 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 $y_i = f_\rho(x_i) + \varepsilon$ with target function $f_\rho(x) = \sin(||x||^2)$ and Gaussian noise $\varepsilon$ having zero-mean and unit-variance. The feature dimension $d$ is set to 500. The samples are generated from $x_i = \Sigma_d^{1/2} t_i$ (and thus $X^T X = T^T \Sigma_d T$ with $T = [t_1, t_2, \ldots, t_n]^T$) by the following steps: (i) take $\Sigma_d$ as a diagonal matrix with its diagonal entries following with harmonic decay, i.e., $(\Sigma_d)_{ii} \propto n/i$. (ii) take $T$ as a random orthogonal matrix such that $T^T \Sigma_d T$ also has a harmonic eigen-decay with $T$ having almost i.i.d entries.

Accordingly, the above generation process satisfies Assumption 3 and also $X^T X/d$ admits the same eigenvalue decay as $\Sigma_d$, 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. 3. In this case, the explicit $\lambda := \tilde{c} n^{-\vartheta}$ is the only regularization in KRR. In our model, $\tilde{c}$ is empirically set to 0.01 to avoid a large $\lambda$ 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 $V_1 \propto 1/d \lambda N n^\alpha X$ 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 $\vartheta$ decreases, $\lambda$ increases and the peak point of the variance occurs at smaller and smaller $n$; 4) the bias monotonically decreases with $n$. We generate a random Gaussian matrix and use the QR decomposition to obtain an orthogonal matrix.  

1 The source code of our implementation can be found in http://www.lfhsgre.org.
Kernel regression in high dimensions: Refined analysis beyond double descent

Figure 3: Harmonic decay of $\tilde{X}$ with polynomial kernel: MSE of the expected excess risk, the variance in Eq. (10), our derived $V_1$, the bias in Eq. (9), and our derived convergence rate $O(n^{-2r})$ with $r = 1$ for different $\vartheta$.

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

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

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 $V_1 \approx \frac{1}{\pi} \frac{1}{\lambda^2} \tilde{X}$ 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 $\gamma$ 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


Kernel regression in high dimensions: Refined analysis beyond double descent


