Characterizing Overfitting in Kernel Ridgeless Regression Through the Eigenspectrum

Tin Sum Cheng [1] Aurelien Lucchi [1] Anastasios Kratsios [2] David Belius [3]

Abstract

We derive new bounds for the condition number of kernel matrices, which we then use to enhance existing non-asymptotic test error bounds for kernel ridgeless regression (KRR) in the over-parameterized regime for a fixed input dimension. For kernels with polynomial spectral decay, we recover the bound from previous work; for exponential decay, our bound is non-trivial and novel. Our contribution is two-fold: (i) we rigorously prove the phenomena of tempered overfitting and catastrophic overfitting under the sub-Gaussian design assumption, closing an existing gap in the literature; (ii) we identify that the independence of the features plays an important role in guaranteeing tempered overfitting, raising concerns about approximating KRR generalization using the Gaussian design assumption in previous literature.

1. Introduction

Kernel ridge regression (KRR) plays a pivotal role in machine learning since it offers an expressive and rapidly trainable framework for modeling complex relationships in data. In recent years, kernels have regained significance in deep learning theory since many deep neural networks (DNNs) can be understood as converging to certain kernel limits.

Its significance has been underscored by its ability to approximate deep neural network (DNN) training under certain conditions, providing a tractable avenue for analytical exploration of test error and robust theoretical guarantees (Jacot et al., 2018; Arora et al., 2019; Bordelon et al., 2020). The adaptability of kernel regression positions it as a crucial tool in various machine learning applications, making it imperative to comprehensively understand its behavior, particularly concerning overfitting.

Despite the increasing attention directed towards kernel ridge regression, the existing literature predominantly concentrates on overfitting phenomena in either the high input dimensional regime or the asymptotic regime (Liang & Rakhlin, 2020; Mei & Montanari, 2022; Misiakiewicz, 2022), also known as the ultra-high dimensional regime (Zou & Zhang, 2009; Fan et al., 2009). Notably, the focus on asymptotic bounds, requiring the input dimension to approach infinity, may not align with the finite nature of real-world datasets and target functions. Similarly, classical Rademacher-based bounds, e.g. (Bartlett & Mendelson, 2002), require that the weights of the kernel regressor satisfy data-independent bounds, a restriction that is also not implemented in standard kernel ridge regression algorithms. These mismatches between idealized mathematical assumptions and practical implementation standards necessitate a more nuanced exploration of overfitting in kernel regression in a fixed input dimension.

Contributions This work aims to understand the overfitting behaviour for kernel ridge regression (KRR). Our main contributions are summarized as follows:

1. We rigorously derive *tight* non-asymptotic upper and lower bounds for the test error of the minimum norm interpolant under a sub-Gaussian design assumption on the features. While this assumption assumes independence of the features, this point will be relaxed in contribution #3.
2. Consequently, we show that a polynomially decaying spectrum yields *tempered overfitting* (Theorem 4.2), whereas an exponentially decaying spectrum leads to *catastrophic overfitting* (Theorem 4.3), filling a gap in the existing literature.
3. We extend our analysis to the case of sub-Gaussian but possibly dependent features. We discover a qualitative difference in overfitting behavior that was previously unknown in the literature using the Gaussian design

[1] Department of Mathematics and Computer Science, University of Basel, Switzerland [2] Department of Mathematics and Statistics, McMaster University and Vector Institute, Canada [3] Faculty of Mathematics and Computer Science, UniDistance Suisse, Switzerland. Correspondence to: Tin Sum Cheng <tinsum.cheng@unibas.ch>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

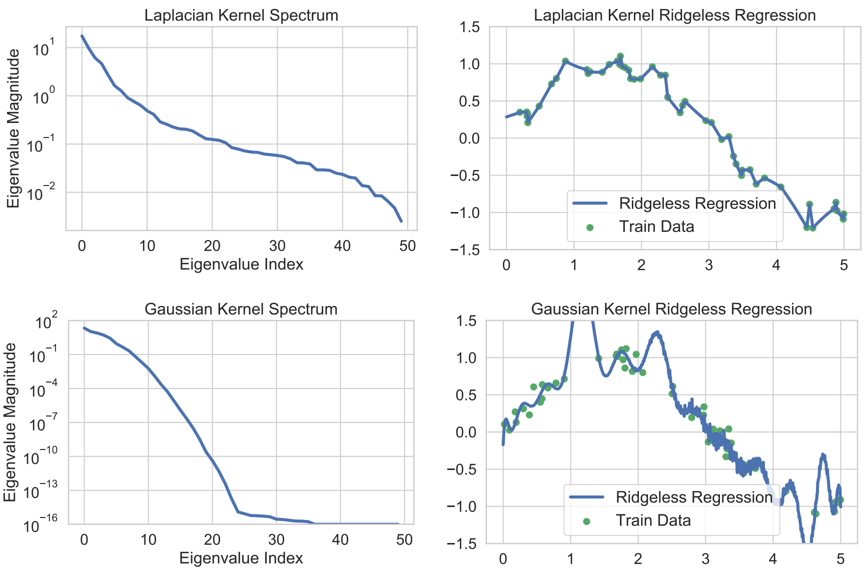

Figure 1. Kernel spectra for Laplacian and Gaussian kernels and their overfitting behaviours.
Tempered Overfitting: The empirical kernel spectrum of the Laplacian kernel decays moderately (top left), and so does the quality of its test-set performance as one departs from the training data (top right).
Catastrophic Overfitting: The Gaussian kernel exhibits rapid spectral decay (bottom left), and so does the reliability of its test-set performance for inputs far from the training data (bottom right).

assumption to approximate KRR test error. This raises concerns that previous literature may have oversimplified the KRR setting by relying on the Gaussian design assumption.

Motivation This paper is motivated by observations made in (Mallinar et al., 2022), where the Laplacian kernel (with a polynomial spectrum) does not suffer from catastrophic overfitting even without ridge regularization, whereas the Gaussian kernel (with an exponential spectrum) does. The correspondence between polynomial and exponential spectral decay rates and the tempered and catastrophic overfitting regimes is illustrated in Figure 1. However, (Mallinar et al., 2022) relied on findings from (Simon et al., 2021), which inevitably depend on the Gaussian design assumption. We aim to explore whether it is possible to characterize overfitting behavior solely based on the kernel eigen-spectrum under a weaker assumption. The first step, which is undertaken in this paper, is to relax the assumption to sub-Gaussian.

Organization of the Paper The structure of this paper is as follows:

1. In Section 2, we discuss how our work differs from previous studies and complements their results. A summary for comparison can be found in Table 1.

2. In Section 3, we state the definitions and assumptions for this paper.

3. In Section 4, we present our main results (Theorems 4.1, 4.2, and 4.3) and interpret their significance, novelty, and improvement compared to previous work.

4. In Section 5, we showcase the empirical results of a simple experiment to validate our findings.

5. In Section 6, we discuss the implications of our contributions in-depth, including their limitations and potential directions for future research.

6. In Section A, we present our proof under the Sub-Gaussian design assumption 3.3.

7. In Section B, we list the technical lemmata used in this paper.

2. Previous Work

Traditional statistical wisdom has influenced classical machine learning models to focus on mitigating overfitting with the belief that doing so maximizes the ability of a model to generalize beyond the training data. However, these traditional ideas have been challenged by the discovery of the “benign overfitting” phenomenon, see e.g (Liang & Rakhlin, 2020; Bartlett et al., 2020; Tsigler & Bartlett, 2023; Haas et al., 2023), in the context of KRR. A key factor is that traditional statistics operate in the under-parameterized setting where the number of training instances exceeds the number of parameters. This assumption is rarely applicable to modern machine learning, where models depend on vastly more parameters than their training instances, and thus, classical statistical thought no longer applies.

2.1. (Sub-)Gaussian Design Assumption

Many previous works (Jacot et al., 2020; Bordelon et al., 2020; Simon et al., 2021; Loureiro et al., 2021; Cui et al., 2021) require the so-called Gaussian design assumption, where isotropic kernel feature vectors are replaced by Gaussian vectors, to prove their results on KRR generalization. (See Assumption 3.4 in Section 3 for details.) In contrast, we obtain tight bounds on test error under the weaker sub-Gaussian design assumption, where we replace the isotropic kernel feature vectors with sub-Gaussian vectors. This seemingly simple extension yields surprisingly many fundamental differences compared to previous work:

1. Generally, a random vector $\mathbf{z} = (z_k)_{k=1}^p \in \mathbb{R}^p$ is isotropic if $\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] = \mathbf{I}_p$, meaning the entries are uncorrelated but possibly dependent on each other. However, for a Gaussian vector, uncorrelatedness implies independence. Therefore, the Gaussian design assumption implicitly requires the independence of features, which, as we will show later, is crucial to the phenomenon of tempered overfitting with a polynomially decaying spectrum.

2. Also, the argument of many previous works relies heavily on the Gaussian design assumption, which cannot be generalized to sub-Gaussian design by any universality argument. For instance, (Simon et al., 2021) utilizes rotational invariant property of Gaussian vectors; (Bordelon et al., 2020; Loureiro et al., 2021; Cui et al., 2021) inevitably require the Gaussian design assumption in the Replica method. By relaxing to the sub-Gaussian assumption, we show that many nice properties of Gaussian vectors, such as rotational invariant property, smoothness/continuity, anti-concentration, are not important to the overfitting behaviour of the ridgeless regression, extending previous results to a more general setting.

3. Last but not least, both Gaussian and sub-Gaussian design assumptions require the feature dimension M to be finite, while it should be infinite in the case of a kernel. For the sake of completeness, we provide a result showing that the overfitting behavior of an infinite rank kernel can be approximated by its finite rank truncation. See Proposition A.13 for details.

2.2. Test Error on Ridgeless Regression

Many previous works (Arora et al., 2019; Liang & Rakhlin, 2020; Bordelon et al., 2020; Bartlett et al., 2020; Simon et al., 2021; Mei et al., 2021; Misiakiewicz, 2022; Bach, 2023; Cheng et al., 2023) are devoted to bounding the KRR test error in different settings. In the context of benign overfitting, a recent related paper (Tsigler & Bartlett, 2023) gives tight non-asymptotic bounds on the ridgeless regression test error ***under the assumption that the condition number of kernel matrix is bounded by some constant***. Our random matrix theoretic arguments successfully allow us to derive tight non-asymptotic bounds for the condition number of the empirical kernel matrix (see Theorem 4.1) and to apply some of their technical tools without their stylized assumptions.

2.3. Overfitting

Recently, (Mallinar et al., 2022) characterized previous results on overfitting, especially in the context of KRR, and classified them into three categories 1) *benign overfitting* meaning that the learned model interpolates the noisy training data while exhibiting a negligible reduction in test performance decline, 2) *tempered overfitting*, which happens when the learned model exhibits a bounded reduction in test set performance due resulting from an interpolation of the training data, and 3) *catastrophic overfitting* which covers the case where the test error is unbounded due to the learned model having interpolated the training data. In this paper, we characterize the tempered and catastrophic overfitting cases, omitting benign overfitting which has been characterized in prior work. According to (Mallinar et al., 2022), even with Gaussian design assumption, benign overfitting occurs only when the spectral eigen-decay is slower than any polynomial decay $\lambda_k = \Theta\left(k^{-1-\epsilon}\right)$ for any constant $\epsilon > 0$, for instance the linear-poly-logarithmic decay $\lambda_k = \Theta\left(k^{-1}\log^{-a}(k)\right)$ for some constant $a > 0$. Such spectral eigen-decay, to the best of our knowledge, does not appear in commonly-known kernels.

2.4. Comparison to other Results

A comparison of our results to the state-of-the-art in the literature is detailed in Table 1. Especially relevant is the comparison to (Barzilai & Shamir, 2023). We note that this paper is in fact a concurrent work, as it was published on arXiv just four weeks prior to the submission deadline for ICML. The strength of (Barzilai & Shamir, 2023) is the general setting under which they perform their analysis. However, our analysis yields tighter bounds than theirs for the class of kernels to which our analysis applies, achieved via tighter bounds on the involved kernel eigenspectrum. Importantly, unlike their results, our analysis provides upper and matching lower bounds on the test error. Additionally, we address the catastrophic behavior with exponential eigendecay, which they have not considered

Table 1. Comparison with prior works

	(Mallinar et al., 2022)	(Tsigler & Bartlett, 2023)	(Barzilai & Shamir, 2023)	**This paper**
Assumption on kernel	Gaus. feature	Bound on condition num.	Concentrated feature	Sub-Gaus. feature
Non-asymptotic bounds	✗	✓	✓	✓
Overfitting for poly. decay	✓	✗	✓	✓
Overfitting for exp. decay	✓	✗	✗	✓

3. Setting

Given a kernel K with reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, we consider the kernel ridge regression (KRR) problem:

$$\min_{f\in\mathcal{H}} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

The solution $\hat{f}$ to the KRR problem, called the kernel ridge regressor, is unique whenever $\lambda > 0$. For $\lambda = 0$ and $\dim(\mathcal{H}) > N$, with minor abuse of notation, we write $\hat{f}$ the norm-minimizing interpolant:

$$\hat{f} \in \operatorname*{arg\,min}_{f(x_i)=y_i,\forall i} \|f\|_{\mathcal{H}}.$$

Given a data-distribution μ on the input space $\mathcal{X}$, using the Mercer theorem we decompose:

$$K(x,x') = \sum_{k=1}^{M} \lambda_k \psi_k(x)\psi_k(x'),$$

where $M \in \mathbb{N}\cup\{\infty\}$ is the kernel rank, λ_k's are the eigenvalues indexed in decreasing order with corresponding eigenfunctions ψ_k's. Hence the (random) kernel matrix $\mathbf{K} = [K(x_i,x_j)]_{i,j}$ can be written concisely in matrix form

$$\mathbf{K} = \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi},$$

where $\mathbf{\Psi} = [\psi_k(x_i)] \in \mathbb{R}^{M\times N}$ is the design block.

Next, we introduce two important assumptions in this paper.

Assumption 3.1 (Interpolation)**.** Assume $M \in \mathbb{N}$ and there exists an integer constant $\eta > 1$ (to be determined) such that $M \geq \eta N$. Also, we assume that $\lambda = 0$ and hence $\hat{f}$ denotes the norm-minimizing interpolant.

Remark 3.2. Note that Assumption 3.1 only requires the feature dimension M to be larger than the threshold ηN, and M does not necessarily need to be linear in N. This is different from the so-called proportional regime in (Liang & Rakhlin, 2020; Liu et al., 2021), where the proportion $\frac{M}{N}$ converges to some constant $\gamma > 0$.

Next, we assume sub-Gaussianity of the eigenfunctions, which is very standard in KRR literature (to name a few, (Liang & Rakhlin, 2020; Bartlett et al., 2020; Tsigler & Bartlett, 2023; Bach, 2023)):

Assumption 3.3 (Sub-Gaussian design)**.** Let $M \in \mathbb{N}$. For every $k = 1, ..., M$, the random variable $\psi_k(x)$ is replaced by an independent sub-Gaussian variable with uniformly bounded sub-Gaussian norm.

This is a relaxation of the Gaussian design assumption, which is used in (Bordelon et al., 2020; Cui et al., 2021; Loureiro et al., 2021; Simon et al., 2021):

Assumption 3.4 (Gaussian design)**.** Let $M \in \mathbb{N}$. For every $k = 1, ..., M$, the random variable $\psi_k(x)$ is replaced by an independent standard Gaussian variable.

Under Assumption 3.4, the learning task is simply linear regression with the feature vectors $\psi_k(x)$'s replaced by M-dimensional Gaussian inputs $\boldsymbol{\phi}_k \overset{\text{def.}}{=} \mathbf{\Lambda}^{1/2}\boldsymbol{\psi}_k \sim \mathcal{N}(0, \mathbf{\Lambda}^{1/2})$ for all k.

4. Main Result

The analysis of our main result consists of three steps. First, we bound the condition number of the kernel matrix $\mathbf{K}$ under the interpolation assumption (Assumption 3.1) and sub-Gaussian design assumption (Assumption (Assumption 3.3)) in Theorem 4.1. Next, we use this result to give a tight bound of the test error and conclude the effect of the spectral decay on overfitting in Theorem 4.2. Lastly, we demonstrate the necessity of feature independence by Theorem 4.3. The formal versions of the main theorems can be found in Section A.

4.1. Condition Number

First, we show that the condition number of the kernel matrix is bounded with polynomial and exponential decays differently.

Theorem 4.1 (Bounding the Condition Number)**.** *Suppose $M, N \in \mathbb{N}$ such that $M \geq \eta N$ for some constant $\eta > 1$ that is large enough. Let $\Psi \in \mathbb{R}^{M\times N}$ be a matrix with i.i.d. isotropic random vectors Ψ_i's with independent sub-Gaussian entries as columns. Let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_k)_{k=1}^M \in \mathbb{R}^{M\times M}$ be a diagonal matrix. Then with high probability, the condition number $\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})}$ of the matrix $\mathbf{K} = \mathbf{\Psi}^\top\mathbf{\Lambda}\mathbf{\Psi} \in \mathbb{R}^{M\times N}$ is bounded by:*

1. $\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} = \mathcal{O}_N\left(\frac{\lambda_1}{\lambda_N}\right)$*, if λ_k's decay polynomially;*

2. $\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} = \Theta_N\left(\frac{\lambda_1}{\lambda_N}N\right)$, *if λ_k's decay exponentially and furthermore $\Psi \in \mathbb{R}^{M\times N}$ is a Gaussian random matrix with $M = \eta N$.*

Proof idea: It is well known that, with high probability, $s_{\max}(\mathbf{K}) \asymp N\lambda_1$ for both types of eigen-decay. The major difference between the polynomial and exponential decay is the lower bound of $s_{\min}$. In the former case, we apply random matrix concentration from (Vershynin, 2010) to obtain $s_{\min}(\mathbf{K}) \gtrsim N\lambda_N$; in the latter case, we apply a lemma from (Tao, 2012) and the anti-concentration of Gaussian to obtain $s_{\min}(\mathbf{K}) \asymp N\lambda_N$. The full proof can be found in Lemmata A.6 and Lemmata A.7 in the appendix. □

Intuitively, one might suppose that $\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \approx \frac{\lambda_1}{\lambda_N}$. From Theorem 4.1 and the experiments in Section 5, we can see that polynomial spectral eigen-decays yields the intuitive bound $\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \approx \frac{\lambda_1}{\lambda_N}$ on the condition number. In the latter case, however, the intuitive bound does not hold even if we restrict the feature dimension $M = \eta N$ and the random matrix $\mathbf{\Psi}$ to Gaussian.

While we apply Theorem 4.1 to bound the test error in the rest of the paper, the bound of the condition number can be of independent interest, for instance when studying the convergence properties of gradient-based methods.

4.2. Classifying Overfitting Regimes

The bound on the condition number of the kernel matrix in Theorem 4.1 can be applied to bound the test error of kernel ridgeless regression under the interpolation assumption (Assumption 3.1) and sub-Gaussian design assumption (Assumption 3.3). To this end, we formally define the test error as follows:

Let $y_i = f^\star(x_i) + \epsilon_i$ for all $i = 1, ..., N$, where $f^\star \in \mathcal{H}$ is the target function and ϵ_i's are draws from a centered sub-Gaussian random variable ϵ with variance $\mathbb{E}\left[\epsilon^2\right] = \sigma^2 > 0$. We define the test error (or excess risk) $\mathcal{R}$ to be the mean square error (MSE) between the target function $f^\star$ and the norm-minimizing interpolant $\hat{f}$ of a given fixed dataset of size N averaging out the noise in the dataset:

$$\mathcal{R} \overset{\text{def.}}{=} \mathbb{E}_{x,\epsilon}\left[(f^\star(x) - \hat{f}(x))^2\right]. \tag{1}$$

Now, we present our result on overfitting with polynomial and exponential decays.

Theorem 4.2 (Overfitting with Polynomial and Exponential Eigen-Decay). *Suppose the interpolation assumption (Assumption 3.1) and the sub-Gaussian design assumption (Assumption (Assumption 3.3)) hold.* [1] *Then there exists a constant $C \in (0,1)$ independent of M, N, such that with high probability, the followings hold:*

1. *if λ_k's decays polynomially, then $C \le \mathcal{R} \le C^{-1}$. In other words, a kernel with polynomial decay exhibits tempered overfitting.*

2. *if λ_k's decays exponentially, then $\mathcal{R} \ge CN$. In other words, a kernel with polynomial decay exhibits catastrophic overfitting.*

Proof idea: The proof proceeds similarly to (Tsigler & Bartlett, 2023), where we first use the upper bound on the condition number of the kernel matrix in Theorem 4.1 together with the result from (Tsigler & Bartlett, 2023) to bound the KRR test error from above. Then we apply the result of the matching lower bound from Theorem (Tsigler & Bartlett, 2023) to conclude the statement. The full proof can be found in Corollary A.10 and Theorem A.11 in the appendix. □

Extension of Previous Results Although the two results in Theorem 4.2 are not new in the literature, we have provided a qualitatively better analysis: 1) the upper bound of the test error of Theorem 4.2 is a result we can recover from (Barzilai & Shamir, 2023, Theorem 2), but our probability is of exponential decay which is faster than the Markov type bound of (Barzilai & Shamir, 2023); 2) the tempered overfitting behaviour of kernel with polynomial decay, that is reported in (Cui et al., 2021; Simon et al., 2021; Mallinar et al., 2022), which used the Gaussian design Assumption 3.4. We replace this with the more general sub-Gaussian design assumption (Assumption (Assumption 3.3)).

Benign Overfitting As reported in (Bartlett et al., 2020; Mallinar et al., 2022; Barzilai & Shamir, 2023), if the spectral eigen-decay is much slower than polynomial decay, say $\lambda_k = \Theta_k\left(k^{-1}\log^{-a} k\right)$ for some $a > 0$, then the overfitting is benign under the Gaussian design assumption. However, to the best of our knowledge, there is no natural kernel exhibiting such eigen-decay $\lambda_k = \Theta(k^{-1}\log^{-a} k)$. Thus, we only consider polynomial and exponential eigen-decays, which represent realistic scenarios as in Figure 2. Hence, by Theorem 4.2, the discussion of benign overfitting is out of the scope of this paper.

4.3. Independent versus Dependent Features

We further investigate the reason behind tempered overfitting with polynomial eigen-decay and discover that the independence between the entries of the feature vector $\boldsymbol{\psi}$ plays an important role in bounding the smallest singular

[1] We suppose Assumption (Assumption (Assumption 3.3)) to hold in the sense that the distributions of the regressor $\hat{f}(x) = \mathbf{K}_x^\top \mathbf{K}^{-1}\mathbf{y}$ and the target function $f^*(x)$ evaluated on a random test point x are replaced by those with random variables $\boldsymbol{\psi}^\top \mathbf{\Lambda}^{1/2}(\mathbf{\Psi}^\top\mathbf{\Lambda}\mathbf{\Psi})^{-1}(\mathbf{\Psi}^\top\boldsymbol{\theta}^* + \boldsymbol{\epsilon})$ and $\boldsymbol{\psi}^\top\boldsymbol{\theta}^*$ for some fixed target vector $\boldsymbol{\theta}^* \in \mathbb{R}^M$, noise vector $\boldsymbol{\epsilon} \in \mathbb{R}^N$ and i.i.d. random vectors $\boldsymbol{\psi}$, $\mathbf{\Psi}_i$'s.

value $s_{\min}(\mathbf{K})$ of the kernel matrix K. Specifically, we have

Theorem 4.3 (Smallest singular value with dependent features)**.** *Suppose $M, N \in \mathbb{N}$ such that $M \geq \eta N$ for some constant $\eta > 1$ large enough. Let $\Psi \in \mathbb{R}^{M \times N}$ be a matrix with i.i.d. isotropic random vectors Ψ_i's with (possibly dependent) sub-Gaussian entries as columns. Let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_k)_{k=1}^{M} \in \mathbb{R}^{M \times M}$ be a diagonal matrix with λ_k's decaying polynomially. Then with high probability, we have*

$$s_{\min}(\mathbf{K}) \geq PN\lambda_N,$$

where P is a positive random variable depending on $\mathbf{\Psi}$. If the entries of each Ψ_i are furthermore independent of each other, that is, if the sub-Gaussian design assumption (Assumption 3.3) holds, then there exists a constant $C > 0$, such that with high probability, $P \geq C$, recovering the result in Theorem 4.1.

Proof idea: The argument follows from random matrix concentration from (Vershynin, 2010). The full proof can be found in Lemma A.4 in the appendix. □

In general, when the features are dependent on each other, the random variable P can vanish to zero as $N \to \infty$, rendering the lower bound in Theorem 4.3 vacuous. Hence the argument in Theorem 4.2 would break down in the general feature case (for instance when considering kernel feature vectors $\boldsymbol{\psi} = (\psi_k(x))_{k=1}^{M}$, where the eigenfunctions ψ_k's are generally dependent on each other). Indeed, we discover both tempered and catastrophic overfitting phenomena can occur for kernels with polynomial eigen-decay and dependent feature vectors (see Figure 2). The neural tangent kernel (NTK) for a 1-hidden layer network is defined to be $K(x,z) = x^\top z\kappa_0(x^\top z) + \kappa_1(x^\top z)$ where $\kappa_0(t) \overset{\text{def.}}{=} 1 - \frac{1}{\pi}\arccos(t)$, $\kappa_1(t) \overset{\text{def.}}{=} \frac{1}{\pi}\left(t(\pi - \arccos(t)) + \sqrt{1-t^2}\right)$. According to (Bietti & Mairal, 2019), the NTK exhibits polynomial eigen-decay. This counterexample suggests that the conclusion drawn in (Mallinar et al., 2022) regarding the tempered overfitting with polynomial eigen-decay might be too optimistic.

5. Experiments

We run several simple experiments to validate our theoretical analysis on overfitting.

5.1. (Sub-)Gaussian Design

First, we validate the main results in Section 4: 1) the bound of the condition number $\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})}$ for polynomial and exponential spectra as predicted in Theorem 1; 2) the tempered and the catastrophic) overfittings for polynomial and exponential spectra.

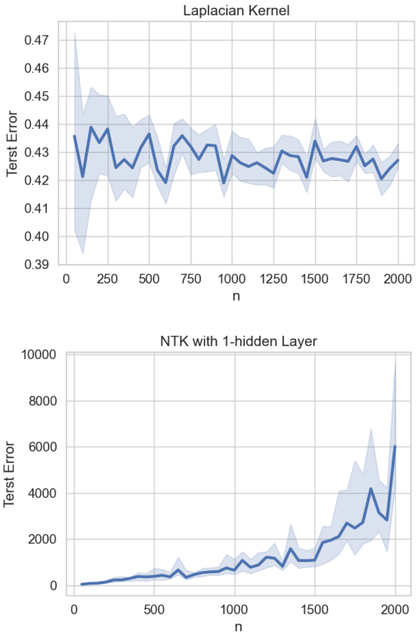

Figure 2. Test error of kernel interpolation on the unit 2-disk against the sample size N. (Top): Laplacian kernel $K(x,z) = e^{-\|x-z\|_2}$ (Bottom): ReLU Neural tangent kernel (NTK) for a 1-hidden layer network

For simplicity, we implement the experiment following Assumption (Assumption 3.4). Let $\boldsymbol{\phi}_k \sim \mathcal{N}(0, \mathbf{\Lambda})$ be i.i.d. Gaussian random vector with covariance $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_k\}$. Write $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ be a matrix with k^{th} column $\boldsymbol{\phi}_k$. For each pair N and $M = 10N$, we run over 20 random samplings for the kernel matrix $\mathbf{\Phi}^\top\mathbf{\Phi}$.

Figure 3 confirms that the condition number of the kernel matrix grows as described in Theorem 4.1: with $\frac{s_{\max}}{s_{\min}} \asymp \frac{\lambda_1}{\lambda_N}$ in the case of a polynomial spectrum and $\frac{s_{\max}}{s_{\min}} \asymp \frac{N\lambda_1}{\lambda_N}$ in the case of an exponential spectrum. To compute the test error, we randomly set the true coefficient $\boldsymbol{\theta}^* \sim \mathcal{N}(0, \mathbf{I}_M)$ and let $y = (\boldsymbol{\theta}^*)^\top\boldsymbol{\phi} + \epsilon$ be the label where $\epsilon \sim \mathcal{N}(0,1)$ is the noise. We evaluate the test error using the mean square error (MSE) between the true label and the ridgeless regression on 1000 random points. For each pair N and $M = 10N$, we run over 20 iterations for the same true coefficient. In Figure 4, we validate Theorem 4.2: the learning curve for polynomial decay is asymptotically bounded by constants; while that for exponential decay increases as $N \to \infty$.

To validate Theorem 4.1 under the sub-Gaussian design assumption (Assumption 3.3), we compare the empirical spectrum of $\mathbf{K} = \mathbf{\Psi}^\top\mathbf{\Lambda}\mathbf{\Psi}$, where the isotropic features $\mathbf{\Psi}_i$ are either Gaussian or uniformly distributed $(\text{unif}[-\sqrt{3}, +\sqrt{3}])$. In Figure 5, we observe that $s_{\min}\left(\frac{1}{n}\mathbf{K}\right)$ remains in the same magnitude as λ_N for both types of features. To observe the effect of feature dependence on the smallest singular value

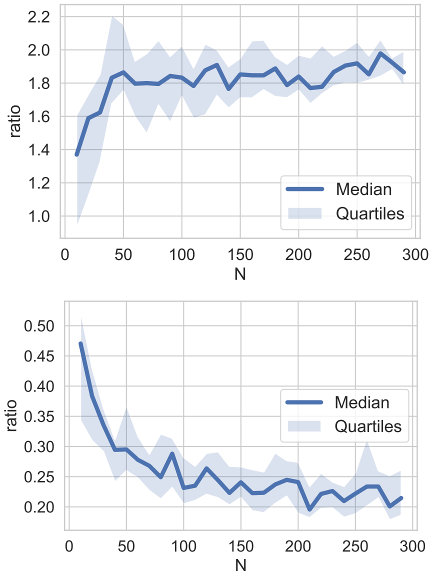

Figure 3. Validation of Theorem 4.1: The ratios $\frac{s_{\max}}{s_{\min}} : \frac{\lambda_1}{\lambda_N}$ for the polynomial spectrum (top) and $\frac{s_{\max}}{s_{\min}} : \frac{N\lambda_1}{\lambda_N}$ for the exponential spectrum (bottom) are asymptotically constant.

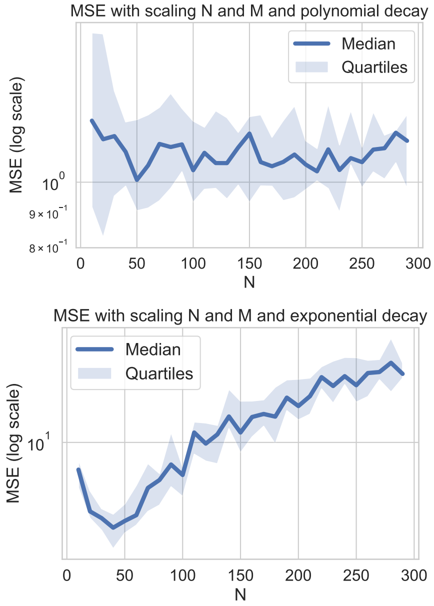

Figure 4. Validation of Theorems 4.2 and 4.3: Learning curves for spectra with polynomial (top) and exponential (bottom) decays.

$s_{\min}(\mathbf{K})$, we conduct experiments with cosine and sine features. In Figure 6, we observe that $s_{\min}(\mathbf{K})$ vanishes, thus validating our findings in Theorem 4.3.

5.2. Dependent Feature

Additionally, when we transition to kernels, we observe that the smallest singular value also diminishes (see Figure 7). In this scenario, the data follows a Gaussian distribution on the real line. Combined with the insights from Figure 2, where the Laplacian kernel displays tempered overfitting under different data distributions, we conclude that *the condition $s_{\min}(\mathbf{K}) \approx \lambda_N$ is sufficient but not necessary for observing tempered overfitting in kernels with polynomial decay*.

6. Discussion

In this section, we discuss the interpretations of our results and their possible extensions.

6.1. Implicit Regularization

Intuitively, given a (possibly infinite rank) PDS kernel K, one decomposes the kernel matrix into: $\mathbf{K} = \mathbf{K}_{\leq l} + \mathbf{K}_{>l}$ where the low-rank part $\mathbf{K}_{\leq l}$ fits the low-complexity target function while the high-rank part $\mathbf{K}_{>l} \approx (\sum_{k>l} \lambda_k)\mathbf{I}_N$ serves as the implicit regularization. Hence the (normalized) effective rank $\rho_l \overset{\text{def.}}{=} \frac{\sum_{k>l}\lambda_l}{N\lambda_{l+1}}$ measures the relative strength of the implicit regularization. With exponential eigen-decay, the effective rank $\rho_l = \Theta(N^{-1}) \ll O(1)$ is negligible, hence one can expect the catastrophic overfitting as the implicit regularization is not strong enough to stop the interpolant using high-frequency eigenfunctions to fit the noise. With polynomial decay, the effective rank $\rho_l = \Theta(1)$ shows that the interpolant would fit the white noise as if it is the target function, hence overfitting is tempered; for even slower decay like linear-poly-logarithmic decay $\lambda_k = \Theta(\frac{1}{k\log^2 k})$ in (Barzilai & Shamir, 2023), the effective rank $\rho_l = \Omega(\log l)$, hence the high-frequency part is heavily regularized and benign overfitting would occur.

6.2. Sub-Gaussian Design

We emphasize that the sub-Gaussian design assumption (Assumption 3.3) represents a significantly weaker assumption compared to the Gaussian design assumption (Assumption 3.4), which enhances the theoretical significance of our paper over previous literature in several ways:

1. We only necessitate the independence of sub-Gaussian variables, not their identical distribution, unlike previous works such as (Bordelon et al., 2020; Cui et al., 2021; Loureiro et al., 2021), which relied on the Replica Method and could not circumvent the Gaussian design assumption.

2. In general, sub-Gaussian vectors lack the rotational invariance property of Gaussian vectors, which was crucial in the analyses of (Simon et al., 2021; Mallinar et al., 2022).

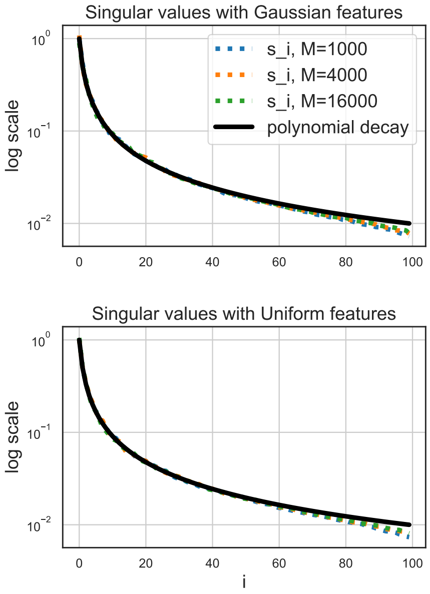

Figure 5. Empirical singular values are close to the eigenspectrum for independent features (top): The features are Gaussian $\psi \sim \mathcal{N}(0, \mathbf{I}_M)$. (bottom): The features are uniformly distributed $\psi \sim (\text{unif}[-\sqrt{3}, +\sqrt{3}])^p$.

3. Sub-Gaussian variables are not required to be continuous, unlike Gaussian vectors, where the continuity of kernels (and consequently the feature vectors) is often assumed in KRR literature (Zhang et al., 2023; Li et al., 2023a;b; Haas et al., 2023).

6.3. Beyond Independent Features

Comparing independent and dependent features, Theorem 4.3 offers insights into the lower bound of the smallest singular value for polynomial eigen-decay and dependent sub-Gaussian features. Consequently, this revelation underscores the qualitative difference in generalization behavior between independent and dependent features: if the sub-Gaussian features are dependent, overfitting can escalate to catastrophic levels (see Figure 2); conversely, independent sub-Gaussian features imply tempered overfitting. As a result, our paper rigorously demonstrates that the theories presented in (Bordelon et al., 2020; Cui et al., 2021; Simon et al., 2021) fail to accommodate the possibility of catastrophic overfitting with polynomial eigen-decay. Therefore, our paper provides insights that are currently absent in the field. It underscores the pivotal role of independence versus dependent features and prompts further inquiry into identifying additional properties of features that influence tempered or catastrophic overfitting.

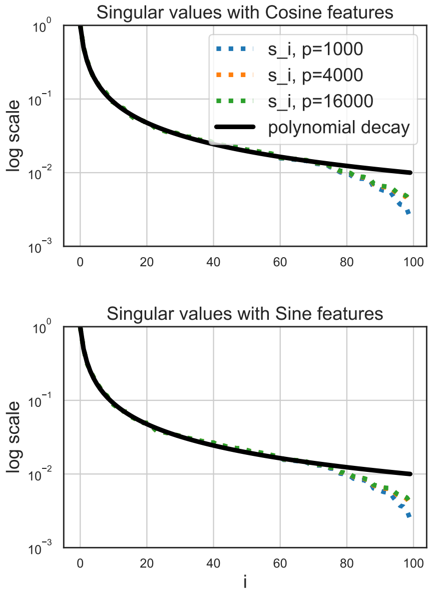

Figure 6. Empirical smallest singular value vanishes for dependent features. (top): The features are cosines $\psi = (\cos(k\cdot))_{k=1}^M$. (bottom): The features are sines $\psi = (\sin(k\cdot))_{k=1}^M$.

6.4. Limitations

Our method currently operates solely under the sub-Gaussian design assumption (Assumption 3.3), presuming the kernel rank is finite, and the features are independent of each other. We acknowledge that this remains distant from the realistic kernel setting.

However, we have the following justifications for the limitations in this paper.

Finite Rank Features To justify whether the finite rank approximation is sufficient to investigate overfitting behaviour, we present the following convergence result of the variance term $\mathcal{V}$. Fix a (infinite-rank) kernel with Mercer decomposition $K = \sum_{k=1}^\infty \lambda_k \psi_k(\cdot)\psi_k(\cdot)$ and a sample of size N. For each integer $M \in \mathbb{N}$, define the truncated kernel $K^{(M)} = \sum_{k=1}^M \lambda_k \psi(\cdot)\psi(\cdot)$. Let $\mathcal{V}$ and $\mathcal{V}(M)$ be the variance terms corresponding to the kernels K and $K^{(M)}$ respectively. Then there exists an integer M_0 such that:

$$|\mathcal{V} - \mathcal{V}(M)| \leq 3\mathcal{V}(M) + \frac{\sigma^2}{N}$$

whenever $M > M_0$. In particular, the variance $\mathcal{V}(M)$ of a finite rank kernel $K^{(M)}$ has the same overfitting behaviour as the original one $\mathcal{V}$. See Proposition A.13 in the appendix for more details.

Concurrent Work We are aware of the concurrent work (Barzilai & Shamir, 2023), which addresses the same prob-

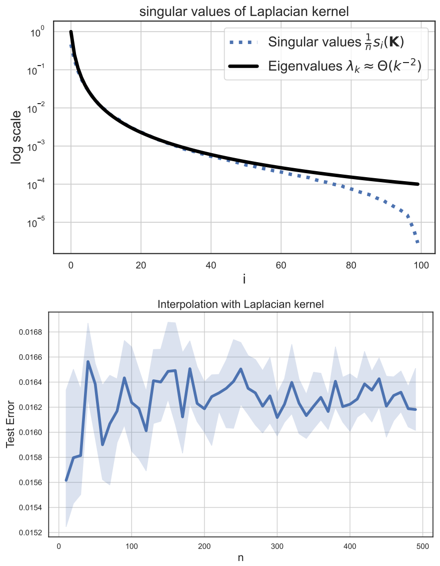

Figure 7. Interpolation with Laplacian kernel $K(x, z) = e^{-|x-z|}$ with inputs $x, z \sim \mathcal{N}(0, 1)$. (Top): The empirical spectrum of Laplacian. The smallest singular value vanishes as in Figure 6. (Bottom): The test error is bounded by some constants as $n \to \infty$, exhibiting tempered overfitting.

lem and offers statements valid for a broader class of features, particularly including kernel features. However, our work serves as a complement to theirs. For instance, our analysis can elucidate overfitting in cases of exponential decay (where their bounds may be vacuous).

6.5. Future Research

There are several obvious possibilities to extend the results of this paper:

1. What causes NTK to exhibit catastrophic overfitting while Laplacian exhibits tempered overfitting? There is more than just the eigen-spectrum that affects the overfitting behaviour, which is worth further investigation.

2. The work (Simon et al., 2021) suggested that the distribution of the kernel features with realistic data is similar to Gaussian. Does this suggest that the data distribution in a realistic dataset leads to independent eigen-functions? As we have seen in our paper, feature independence plays an important role in overfitting behaviour. This might help us to understand more about benign overfitting reported in (Zhang et al., 2017).

3. Controlling the condition number of the kernel matrix in Theorem 4.1 can be of independent interest, for instance when studying the convergence properties of gradient-based methods.

Impact Statement

This paper introduces research aimed at pushing the boundaries of the Machine Learning field. Our work is predominantly theoretical, with minimal direct societal implications. While there are some potential consequences, they are not unique to our study and, therefore, do not warrant specific emphasis in this context.

Acknowledgements

We thank Daniel Barzilai for his insights and his help with the code on NTK.

References

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Bach, F. High-dimensional analysis of double descent for linear regression with random projections. *arXiv preprint arXiv:2303.01372*, 2023.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002. ISSN 1532-4435,1533-7928. doi: 10.1162/153244303321897690.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Barzilai, D. and Shamir, O. Generalization in kernel regression under realistic assumptions. *arXiv preprint arXiv:2312.15995*, 2023.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Cheng, T. S., Lucchi, A., Dokmanić, I., Kratsios, A., and Belius, D. A theoretical analysis of the test error of finite-rank kernel ridge regression. *Annual Conference on Neural Information Processing Systems*, 2023.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10131–10143. Curran Associates, Inc., 2021.

Fan, J., Samworth, R., and Wu, Y. Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, 2009. ISSN 1532-4435,1533-7928.

Haas, M., Holzmüller, D., von Luxburg, U., and Steinwart, I. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension, 2023.

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

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640. PMLR, 2020.

Koltchinskii, V. and Lounici, K. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Li, Y., Zhang, H., and Lin, Q. Kernel interpolation generalizes poorly. *Biometrika*, 2023a.

Li, Y., Zhang, H., and Lin, Q. On the asymptotic learning curves of kernel ridge regression under power-law decay. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3), Jun 2020. ISSN 0090-5364. doi: 10.1214/19-aos1849.

Liu, F., Liao, Z., and Suykens, J. Kernel regression in high dimensions: Refined analysis beyond double descent. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 649–657. PMLR, 13–15 Apr 2021.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., and Nakkiran, P. Benign, tempered, or catastrophic: A taxonomy of overfitting. *Annual Conference on Neural Information Processing Systems*, 2022.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 2021.

Misiakiewicz, T. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression, 2022.

Rudelson, M. and Vershynin, R. The littlewood–offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

Rudelson, M. and Vershynin, R. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

Simon, J. B., Dickens, M., Karkada, D., and DeWeese, M. R. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks, 2021.

Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization, 2017.

Zhang, H., Li, Y., and Lin, Q. On the optimality of misspecified spectral algorithms, 2023.

Zhivotovskiy, N. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29(none):1 – 28, 2024.

Zou, H. and Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37 (4):1733, 2009.

Appendix

With abuse of notations, the constants $c, c_1, c_2, ...$ with small letter c may change from line to line.

Denote $\|\mathbf{v}\|_{\mathbf{M}} \overset{\text{def.}}{=} \sqrt{\mathbf{v}^\top \mathbf{M} \mathbf{v}}$ for any vector $\mathbf{v}$ and matrix $\mathbf{M}$ with appropriate dimension.

For an $M \times N$ matrix $\mathbf{M}$, denote $\mathbf{M}_{\leq l} \in \mathbb{R}^{l \times N}$ its submatrix containing the first l columns; for $M \times M$ square matrix $\mathbf{S}$, denote $\mathbf{M}_{\leq l} \in \mathbb{R}^{l \times l}$ its submatrix containing the first l columns and rows. Matrices with subscripts $\cdot_{l_1:l_2}$ or $\cdot_{>l}$ are defined similarly.

A. Proof

In this section, we will prove the Theorem 4.1 on the condition number of $\mathbf{K}$ under Assumption 3.3.

Let us first restate the sub-Gaussian design assumption (Assumption 3.3):

Assumption A.1 (Sub-Gaussian design)**.** Let $M \in \mathbb{N}$. For every $k \in \mathbb{N}$, the random variable $\psi_k(x)$ is replaced by an independent sub-Gaussian variable in $\mathbb{R}^M$ with uniformly bounded sub-Gaussian norm.

Consider the regressor

$$\hat{f}(x) = \mathbf{K}_x(\mathbf{K} + N\lambda\mathbf{I})^{-1}\mathbf{y} = \boldsymbol{\psi}(x)^\top \mathbf{\Psi}(\mathbf{\Psi}\mathbf{\Lambda}\mathbf{\Psi}^\top + N\lambda\mathbf{I})^{-1}\mathbf{y}$$

Effectively, the Sub-Gaussian Design Assumption replace the vector $\boldsymbol{\psi}(x)$ and the columns $\mathbf{\Psi}_i$ of the matrix $\mathbf{\Psi} \in \mathbb{R}^{M \times N}$ by sub-Gaussian vectors with independent entries. Note that, by setup, those vectors $\boldsymbol{\psi}(x)$ and $\mathbf{\Psi}_i$ are i.i.d. to each other.

A.1. Condition number

The control on the largest singular value $s_{\max}(\mathbf{K})$ of the kernel matrix directly follows from the literature:

Lemma A.2 (bound on largest singular value, Theorem 9 in (Koltchinskii & Lounici, 2017), Theorem 1 in (Zhivotovskiy, 2024))**.** *Suppose Assumption 3.3 holds, that is, there exists some constant $\kappa > 1$ such that*

$$\left\| \langle \mathbf{v}, \mathbf{\Lambda}^{1/2}\boldsymbol{\psi} \rangle \right\|_{\psi_2} \leq \kappa\sqrt{\mathbf{v}^\top \mathbf{\Lambda}\mathbf{v}}$$

for all $\mathbf{v} \in \mathbb{R}^n$, where $\boldsymbol{\psi}$ is the random sub-Gaussian vector with the columns $\mathbf{\Psi}_i$ in $\mathbf{\Psi}$ as its realization, and $\|\cdot\|_{\psi_2}$ denote the sub-Gaussian norm. Then with probability at least $1 - e^{-t}$, it holds that

$$\left\| \frac{1}{N}\mathbf{\Lambda}^{1/2}\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{\Lambda}^{1/2} - \mathbf{\Lambda} \right\|_{op} \leq 20\kappa^2 \|\mathbf{\Lambda}\|_{op} \sqrt{4\rho_0 + \frac{t}{N}}$$

whenever $N \geq 4N\rho_0 + t$, and $\rho_0 \overset{\text{def.}}{=} \frac{\mathrm{Tr}[\mathbf{\Lambda}]}{N\|\mathbf{\Lambda}\|_{op}}$ is the normalized effective rank of $\mathbf{\Lambda}$.

In particular, if $\rho_0 \leq \frac{1}{80(40\kappa^2)^2}$, then with probability at least $1 - e^{-\frac{N}{2(40\kappa^2)^2}}$, it holds that

$$\frac{1}{2}N\|\mathbf{\Lambda}\|_{op} \leq s_{\max}(\mathbf{\Psi}^\top\mathbf{\Lambda}\mathbf{\Psi}) = s_{\max}(\mathbf{\Lambda}^{1/2}\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{\Lambda}^{1/2}) \leq \frac{3}{2}N\|\mathbf{\Lambda}\|_{op}.$$

Remark A.3 (Sub-Gaussian Condition). In (Koltchinskii & Lounici, 2017; Zhivotovskiy, 2024), the random vector ψ_k are required to be centered. However, as mentioned Remark 5.18 in (Vershynin, 2010), centering of a sub-Gaussian random variable X does not change the sub-Gaussian constant by more than 2:

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq 2\|X\|_{\psi_2}.$$

Hence, by possibly changing the constant κ, we drop the requirement of centered random variable in the statement.

Lemma A.4 (Lower bound of smallest singular value for polynomial spectrum)**.** *Suppose Assumption 3.1 holds, $\lambda_k = \Theta_k\left(k^{-1-a}\right)$ for some constant $a > 0$, and each column $\mathbf{\Psi}_i = (\boldsymbol{\psi}_{ki})_{k=1}^M \in \mathbb{R}^M$ is i.i.d. sub-Gaussian isotropic random* *vector, whose entries ψ_{ki} are not necessarily independent. Then there exists constants $c_1, c_2 > 0$ such that, with a probability of at least $1 - 2e^{-c_1 N}$:*

$$s_{\min}(\mathbf{K}) \geq c_2 \min_i \{P_i^2\} \cdot N\lambda_N, \tag{2}$$

where $P_i \overset{\text{def.}}{=} \sqrt{\frac{\sum_{i=N+1}^{M} \psi_{ki}^2}{M-N}}$ is a random variable depending on the inputs x_i's. Furthermore, if Assumption 3.3 holds, then with probability at least $1 - 2Ne^{-c_1 N}$, it holds that

$$s_{\min}(\mathbf{K}) \geq \frac{c_2}{2} N\lambda_N. \tag{3}$$

Proof. By Assumption 3.1, the feature dimension $M \geq \theta N$. First, by Weyl's theorem (Corollary 4.3.15 in (Horn & Johnson, 2012)): $s_{\min}(\mathbf{M}_1) + s_{\min}(\mathbf{M}_2) \leq s_{\min}(\mathbf{M}_1 + \mathbf{M}_2)$ for any any symmetric matrix $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{N \times N}$, we have

$$s_{\min}(\mathbf{K}) = s_{\min}(\mathbf{\Lambda}^{1/2}\mathbf{\Psi}^\top)^2 \geq s_{\min}(\mathbf{\Lambda}_{N:\theta N}^{1/2}\mathbf{\Psi}_{N:\theta N}^\top)^2 \geq \lambda_N \cdot \frac{\lambda_{\theta N}}{\lambda_N} s_{\min}(\mathbf{\Psi}_{N:\theta N})^2$$

where we write $\mathbf{\Lambda}^{1/2}\mathbf{\Psi}^\top = \mathbf{\Lambda}_{N:\theta N}^{1/2}\mathbf{\Psi}_{N:\theta N}^\top + (\mathbf{\Lambda}^{1/2}\mathbf{\Psi}^\top - \mathbf{\Lambda}_{N:\theta N}^{1/2}\mathbf{\Psi}_{N:\theta N}^\top)$ and $\cdot_{N:\theta N}$ denote the submatrix with columns ranging from N to θN. Hence with abuse of notation, we replace M by θN in the following argument. Let $\mathbf{R}_i^\top \in \mathbb{R}^{M-N} = \mathbb{R}^{(\theta-1)N}$ be the i-th row of $\mathbf{\Psi}_{>N}$, and $\hat{\mathbf{R}}_i \overset{\text{def.}}{=} \frac{\sqrt{(M-N)}}{\|\mathbf{R}_i\|_2}\mathbf{R}_i$. Note that $\mathbb{E}\left[\|\mathbf{R}_i\|_2^2\right] = (\theta-1)N,\ \forall i = 1, ..., N$. Let $\hat{\mathbf{\Psi}}_{>N} \overset{\text{def.}}{=} (\hat{\mathbf{R}}_i)_{i=1}^N \in \mathbb{R}^{N \times ((\theta-1)N)}$. Now the matrix $\hat{\mathbf{\Psi}}_{>N}^\top$ is an $((\theta-1)N) \times N$ matrix whose columns $\hat{\mathbf{R}}_i$ are independent sub-Gaussian isotropic random matrix with norm $\left\|\hat{\mathbf{R}}_i\right\|_2 = \sqrt{(\theta-1)N}$. Hence, by Theorem B.13, there exists constants $C_8, C_9 > 0$ (depending only on the sub-Gaussian norm of ψ_k) such that, for any $t > 0$, with probability at least $1 - 2e^{-C_8 t^2}$, the inequality holds:

$$s_{\min}(\hat{\mathbf{\Psi}}_{>N}) \geq \sqrt{(\theta-1)N} - C_9\sqrt{N} - t.$$

Set $t = \sqrt{N}$ and $\theta > (C_9 + 2)^2 + 1$, the inequality holds:

$$s_{\min}(\hat{\mathbf{\Psi}}_{>N}) \geq \sqrt{\theta N - N} - C_9\sqrt{N} - \sqrt{N} \geq \sqrt{N},$$

with probability at least $1 - 2e^{-c_8 N}$. Notice that $\mathbf{\Psi}_{>N} = \hat{\mathbf{\Psi}}_{>N}\mathbf{P}$ where $P_i \overset{\text{def.}}{=} \frac{\|\mathbf{R}_i\|_2}{\sqrt{(\theta-1)N}},\ \forall i = 1, ..., N$ and $\mathbf{P} \overset{\text{def.}}{=} \operatorname{diag}\{P_i\}_{i=1}^n \in \mathbb{R}^{N \times N}$ is a random matrix with $\mathbb{E}\left[\mathbf{P}^2\right] = \mathbf{I}_N$. Hence, with high probability,

$$s_{\min}(\mathbf{\Psi}_{>N})^2 \geq s_{\min}(\hat{\mathbf{\Psi}}_{>N})^2 s_{\min}(\mathbf{P})^2 = \min_i\{P_i^2\} s_{\min}(\hat{\mathbf{\Psi}}_{>N})^2 \gtrsim \min_i\{P_i^2\} N.$$

Since $\frac{\lambda_M}{\lambda_N} \asymp \frac{M^{-1-a}}{N^{-1-a}} = \frac{(\theta N)^{-1-a}}{N^{-1-a}} = \theta^{-1-a} \asymp 1$, thus $s_{\min}(\mathbf{K}) \gtrsim \min_i\{P_i^2\} \cdot N\lambda_N$. If, furthermore, Assumption 3.3 holds, then each row vector $\mathbf{R}_i$ has independent sub-Gaussian entries, that is, write $\mathbf{R}_i^\top = \left(z_i^{(k)}\right)_{k=N+1}^{M}$ where each $z_i^{(k)}$ is an independent random variable with sub-Gaussian norm $\leq G$. Then by remark A.3, $P_i^2 = \frac{1}{(\theta-1)N}\sum_{k=N+1}^{M}(z_i^{(k)})^2$ is the average of some independent sub-exponential variables with sub-exponential norms $\leq G^2$. Hence by Lemma B.11, it holds that

$$\mathbb{P}\left\{\left|P_i^2 - 1\right| \geq \delta\right\} \leq 2\exp\left(-C_5 \min\left\{\frac{\delta^2}{G^4}, \frac{\delta}{G^2}\right\}((\theta-1)N)\right)$$

for some absolute constant $C_5 > 0$. Write $c_1 = -C_5 \min\left\{\frac{1}{4G^4}, \frac{1}{2G^2}\right\}(\theta-1)$. Then with a probability at least $1 - 2Ne^{-c_1 N}$, it holds that

$$P_i^2 \geq \frac{1}{2},\ \forall i = 1, ..., N.$$

□

Remark A.5 (Dependence of features). We can see that the effect of the dependence of features is encrypted in the term $\min_i\{P_i^2\}$ in Lemma A.4. Indeed, the smallest singular value of the kernel matrix with dependent features vanishes (see Figure 6) while that with independent features remains in the same magnitude of the theoretical lower bound (see Figure 5).

Hence we can bound the condition number for polynomial eigen-decay:

Lemma A.6 (Condition number for polynomial eigen-decay)**.** *Suppose $M, N \in \mathbb{N}$ such that $M \geq \eta N$ for some constant $\eta > 1$ large enough. Let $\Psi \in \mathbb{R}^{M\times N}$ be a matrix with i.i.d. isotropic random vectors Ψ_i's with (possibly dependent) sub-Gaussian entries as columns. Let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_k)_{k=1}^M \in \mathbb{R}^{M\times M}$ be a diagonal matrix with λ_k's decaying polynomially. Then there exists some constants $c_1, c_2 > 0$, such that for N large enough, with probability $1 - 2Ne^{-c_1 N}$, we have*

$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq c_2 \frac{\lambda_1}{\lambda_N},$$

Proof. Simply combine Lemmata A.2 and A.4. By possibly choosing a larger constant c_1, the claim holds with probability $1 - 2Ne^{-c_1 N}$, for N large enough. □

For exponential eigen-decay, we use another argument:

Lemma A.7 (Condition number for exponential eigen-decay)**.** *Suppose $M = \eta N$ for some integer $\eta > 1$ large enough. Let $\mathbf{K} = \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi} \in \mathbb{R}^{N\times N}$ be a random matrix where $\mathbf{\Psi} \in \mathbb{R}^{M\times N}$ is a Gaussian random matrix, and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_k)_{k=1}^\infty$ is a diagonal matrix with $\lambda_k = \Theta_k\left(e^{-ak}\right)$ for some constant $a > 0$. Then there exist constants $c_1, c_2 > 0$ such that for N large enough, with probability at least $1 - \delta - 2/N$,*

$$c_1 \frac{\lambda_1}{\lambda_N} N \leq \frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq \frac{c_2}{\delta^2} \frac{\lambda_1}{\lambda_N} N.$$

Proof. By Lemma A.2, for N large enough, with probability at least $1 - \frac{1}{N}$, we have

$$\frac{1}{2} N\lambda_1 \leq s_{\max}(\mathbf{K}) \leq \frac{3}{2} N\lambda_1. \tag{4}$$

It remains to show that $s_{\min}$ is bounded above and below at the magnitude of λ_N.

For the upper bound, fix the first $N-1$ vectors $\psi_1, .., \psi_{N-1}$ and pick $v_0 \in \mathbb{S}^{N-1}$ orthogonal to them. Then

$$s_{\min}(\mathbf{K}) = \inf_{v\in\mathbb{S}^{N-1}} \sum_{k=1}^M \lambda_k (\boldsymbol{\psi}_k^\top v)^2 \leq \sum_{k=1}^M \lambda_k (\boldsymbol{\psi}_k^\top v_0)^2 \leq \sum_{k=N}^M \lambda_k (\boldsymbol{\psi}_k^\top v_0)^2.$$

Since the Gaussian is rotational invariant, we have $(\boldsymbol{\psi}_k^\top v_0)^2 \sim \chi^2(1)$. By Lemma B.11, hence we have

$$\mathbb{P}\left\{\left|(\boldsymbol{\psi}_k^\top v_0)^2 - 1\right| \geq t\right\} \leq 2e^{-t^2/8}.$$

Set $t = \sqrt{8\log\frac{2(\theta-1)N}{\delta}}$ and By the union bound, we have

$$\mathbb{P}\left\{\left|(\boldsymbol{\psi}_k^\top v_0)^2 - 1\right| \leq t : N \leq k \leq M\right\} \geq 1 - \sum_{k=N}^M \frac{\delta}{(\theta-1)N} \geq 1 - \delta.$$

Thus with probability of at least $1-\delta$, we have

$$s_{\min}(\mathbf{K}) \leq \sum_{k=N}^M \lambda_k (1+t) = \left(1 + \sqrt{8\log\frac{2(\theta-1)N}{\delta}}\right) \sum_{k=N}^M \lambda_k \tag{5}$$

Since $\lambda_k = \Theta_k\left(e^{-ak}\right)$, there exists some constant $c > 0$ such that

$$\sum_{k=N}^M \lambda_k \leq c\lambda_N;$$

By setting $\delta = \frac{1}{N}$, the factor $\left(1 + \sqrt{8\log\frac{2(\theta-1)N}{\delta}}\right)$ becomes constant in line (5) and hence with probability at least $1 - \frac{1}{N}$,

$$s_{\min}(\mathbf{K}) \leq c\lambda_N \tag{6}$$

for some constant $c > 0$.

For the lower bound, let $\mathbf{K}_N = \sum_{k=1}^N \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \prec \mathbf{K}$. Let $\mathbf{\Lambda}_N = \operatorname{diag}(\lambda_k)_{k=1}^N \in \mathbb{R}^{N\times N}$ and $\mathbf{\Psi}_N = (\boldsymbol{\psi}_k)_{k=1}^N \in \mathbb{R}^{N\times N}$ and set $\mathbf{M} = \mathbf{\Lambda}_N^{1/2}\mathbf{\Psi}_N$ which is invertible almost surely. Note that $\mathbf{K}_N = \mathbf{M}^\top\mathbf{M}$. Let $\mathbf{R}_1, ..., \mathbf{R}_n$ be the rows of $\mathbf{M}$ and let $\mathbf{C}_1, ..., \mathbf{C}_n$ be the columns of $\mathbf{M}^{-1}$. For each $1 \le i \le n$, let $\mathbf{N}_i$ be a unit normal vector orthogonal to the subspace spanned by all rows $\mathbf{R}_1, ..., \mathbf{R}_n$ except $\mathbf{R}_i$.

By Lemma B.10, for any $k \le N$ and $t \in (0, \infty)$,

$$\begin{aligned}
\mathbb{P}\left\{\frac{\lambda_N}{\lambda_k}(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2} \ge t^{-1}e^{-\frac{a}{2}(N-k)}\right\} &= \mathbb{P}\left\{|\boldsymbol{\psi}_k^\top \mathbf{N}_k| \le \sqrt{\frac{\lambda_N}{\lambda_k} t e^{\frac{a}{2}(N-k)}}\right\} \\
&\le \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\lambda_N}{\lambda_k} t e^{\frac{a}{2}(N-k)}} \\
&\le \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\overline{r}e^{-aN}}{\underline{r}e^{-ak}} e^{\frac{a}{4}(N-k)}\sqrt{t}} \\
&= \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\overline{r}}{\underline{r}} e^{-\frac{a}{4}(N-k)}\sqrt{t}}.
\end{aligned}$$

for all $k = 1, ..., N$. By the union bound, we have

$$\begin{aligned}
\mathbb{P}\Bigg\{\underbrace{\frac{\lambda_N}{\lambda_k}(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2} \le t^{-1}e^{-\frac{a}{2}(N-k)} : \forall k = 1, ..., N}_{E}\Bigg\} &\ge 1 - \sum_{k=1}^N \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\overline{r}}{\underline{r}} e^{-\frac{a}{4}(N-k)}\sqrt{t}} \\
&\ge 1 - \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\overline{r}}{\underline{r}} (1-e^{-a/4})^{-1}\sqrt{t}}.
\end{aligned}$$

When the event E happens, we have

$$\sum_{k=1}^N \frac{\lambda_N}{\lambda_k}(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2} \le \sum_{k=1}^N t^{-1}e^{-\frac{a}{2}(N-k)} \le (1-e^{-a/2})^{-1}t^{-1},$$

by Lemma B.9, with probability at least $1 - \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\overline{r}}{\underline{r}}(1-e^{-a/4})^{-1}\sqrt{t}}$, we have

$$s_{\min}(\mathbf{K}) \ge \lambda_N (1-e^{-a/2})t$$

for any $t > 0$. Set $\delta = \frac{2}{\sqrt{2\pi}} \cdot \sqrt{\frac{\overline{r}}{\underline{r}}(1-e^{-a/4})^{-1}\sqrt{t}}$ and we have: with probability at least $1-\delta$,

$$s_{\min}(\mathbf{K}) \ge c\delta^2 \lambda_N \tag{7}$$

for some constant $c > 0$.

Combining Eq. (4), (6) and (7), we obtain the claim. □

A.2. Test Error

In this paper, we use the bias-variance decomposition to analyse the test error $\mathcal{R}$, which is common in literature: (Bartlett et al., 2020; Tsigler & Bartlett, 2023; Bach, 2023; Li et al., 2023a;b).

With abuse of notation, we write $f^*(\mathbf{X}) \in \mathbb{R}^N$ to be the evaluation of f^* on the training set $\mathbf{X} = (x_i)_{i=1}^N$.

Definition A.8 (Bias-Variance Decomposition of test error)**.** Given the test error $\mathcal{R} \overset{\text{def.}}{=} \mathbb{E}_{x,\epsilon}\left[(f^\star(x) - \hat{f}(x))^2\right]$ be the test error. Define the *bias*

$$\mathcal{B} \overset{\text{def.}}{=} \mathbb{E}_x\left[(f^\star(x) - \mathbf{K}_x^\top \mathbf{K}[f^\star(\mathbf{X})])^2\right],$$

which measures how accurately the KRR approximates the true target function $f^\star$. The *variance*, defined as the difference

$$\mathcal{V} = \mathcal{R} - \mathcal{B}$$

quantifies the impact which overfitting to noise has on the test error.

Together with Theorem B.6 from (Tsigler & Bartlett, 2023), we obtain the first statement of Theorem 4.1:

Theorem A.9. *Suppose the interpolation assumption (Assumption 3.1) and sub-Gaussian design assumption (Assumption 3.3) hold. Then there exists some constants $c_1, c_2, c_3, c_4, c_5, c_6$ such that with probability at least $1 - c_1 e^{-N/c_1} - 2Ne^{-c_2 N}$, we have*

$$\begin{aligned}
\mathcal{B} &\leq c_3 \|\theta^*_{>\lfloor N/c_1 \rfloor}\|^2_{\mathbf{\Lambda}_{>\lfloor N/c_1 \rfloor}} + c_4 \|\boldsymbol{\theta}^*_{\leq \lfloor N/c_1 \rfloor}\|^2 \lambda_{\lfloor N/c_1 \rfloor}; \\
\mathcal{V} &\leq c_5 + \frac{c_6}{N}.
\end{aligned}$$

where the target function f^* is given by the inner product $\langle \boldsymbol{\theta}^*, \mathbf{\Lambda}^{1/2} \cdot \rangle$, where $\boldsymbol{\theta}^* \in \mathbb{R}^M$ is a deterministic vector with $\|\boldsymbol{\theta}^*\|_{\mathbf{\Lambda}} < \infty$. In particular, there exists a constant $C > 0$ independent to N such that

$$\mathcal{R} = \mathcal{B} + \mathcal{V} \leq C.$$

Proof. By Theorem B.6, there exists a constant $c > 0$, such that for any $l \leq N/c$, with probability of at least $1 - ce^{-N/c}$, the bias and the variance is bounded by:

$$\begin{aligned}
\mathcal{B}/c &\leq \|\boldsymbol{\theta}^*_{>l}\|^2_{\mathbf{\Lambda}_{>l}} \left(1 + \frac{s_1(\mathbf{K}_l^{-1})^2}{s_N(\mathbf{K}_l^{-1})^2} + N\lambda_{l+1} s_1(\mathbf{K}_l^{-1}) \right) \\
&\quad + \|\boldsymbol{\theta}^*_{\leq l}\|^2_{\mathbf{\Lambda}^{-1}_{\leq l}} \left(\frac{1}{N^2 s_N(\mathbf{K}_l^{-1})^2} + \frac{\lambda_{l+1}}{N} \frac{s_1(\mathbf{K}_l^{-1})}{s_N(\mathbf{K}_l^{-1})^2} \right) \\
\mathcal{V}/c &\leq \frac{s_1(\mathbf{K}_l^{-1})^2}{s_N(\mathbf{K}_l^{-1})^2} \frac{l}{N} + N s_1(\mathbf{K}_l^{-1})^2 \sum_{k>l} \lambda_k^2,
\end{aligned}$$

where $\mathbf{K}_l \overset{\text{def.}}{=} \mathbf{\Psi}^\top_{>l} \mathbf{\Lambda}_{>l} \mathbf{\Psi}_{>l}$, $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*_{\leq l} \oplus \boldsymbol{\theta}^*_{>l}$ is the splitting of the target function coefficient. Take $l = \lfloor N/c \rfloor$. Since $\lambda = 0$, so $s_N(\mathbf{K}_l^{-1}) = s_1(\mathbf{K}_l)^{-1}$ and $s_1(\mathbf{K}_l^{-1}) = s_N(\mathbf{K}_l)^{-1}$. Hence $\frac{s_1(\mathbf{K}_l^{-1})^2}{s_N(\mathbf{K}_l^{-1})^2} = \frac{s_{\max}(\mathbf{K}_l)^2}{s_{\min}(\mathbf{K}_l)^2}$. Since $\mathbf{K}_l$ is just another kernel matrix with rank $(M - l)$, by modifying Lemma A.6 w.r.t. the right-shifted polynomial decay, with probability $1 - 2Ne^{-c_2 N}$, we have

$$\frac{s_{\max}(\mathbf{K}_l)}{s_{\min}(\mathbf{K}_l)} \lesssim \frac{\lambda_{l+1}}{\lambda_{l+N}} \lesssim \frac{l^{-a}}{(l+N)^{-a}} = (1 + N/l)^a \leq (1 + N/(N/c))^a = (1+c)^a,$$

and

$$N\lambda_{l+1} s_1(\mathbf{K}_l^{-1}) = N\lambda_{l+1} s_N(\mathbf{K}_l)^{-1} \lesssim N \frac{\lambda_{l+1}}{N\lambda_{l+N}} = \frac{\lambda_{l+1}}{\lambda_{l+N}},$$

then we can bound the bias term using Theorem B.6:

$$\begin{aligned}
\mathcal{B}/c &\leq c_1 \|\theta^*_{>l}\|^2_{\mathbf{\Lambda}_{>l}} + \|\boldsymbol{\theta}^*_{\leq l}\|^2_{\mathbf{\Lambda}^{-1}_{\leq l}} \left(\frac{c_2 N^2 \lambda^2_{l+1}}{N^2} + \frac{\lambda_{l+1}}{N} \frac{c_3 N^2 \lambda^2_{l+1}}{N\lambda_{l+N}} \right) \\
&\leq c_1 \|\theta^*_{>l}\|^2_{\mathbf{\Lambda}_{>l}} + c_2 \|\boldsymbol{\theta}^*_{\leq l}\|^2_{\mathbf{\Lambda}^{-1}_{\leq l}} \lambda_l^2 \\
&\leq c_1 \|\theta^*_{>l}\|^2_{\mathbf{\Lambda}_{>l}} + c_2 \|\boldsymbol{\theta}^*_{\leq l}\|^2 \lambda_l.
\end{aligned}$$

Similarly, we can write the variance term into:

$$\begin{aligned}
\mathcal{V}/c &\leq c_3\frac{l}{N} + \frac{N}{N^2\lambda_{l+N}^2}\sum_{k>l}\lambda_k^2 \\
&\leq c_3\frac{l}{N} + \frac{c_4}{N^2\lambda_{l+N}^2}\int_l^\infty t^{-2a}dt \\
&= c_3\frac{l}{N} + \frac{c_4}{N^2\lambda_{l+N}^2}l^{-2a+1} \\
&= c_3\frac{l}{N} + \frac{c_4}{N^2(l+N)^{-2a}}l^{-2a+1} \\
&= c_3 + \frac{c_4}{N},
\end{aligned}$$

since $l = \lfloor N/c \rfloor$. □

Corollary A.10 (tempered overfitting)**.** *There exists a constants $C \in (0,1)$ independent to N such that, with the same probability as in Theorem A.9, we have*

$$C \leq \mathcal{R} \leq C^{-1}.$$

Proof. It is a direct consequence of Theorems B.6 and B.7 about the non-asymptotic upper bound of the variance $\mathcal{V}$ and its matching lower bound. In more details, we compute the (normalized) effective rank:

$$\rho_l \overset{\text{def.}}{=} \frac{1}{N\lambda_{l+1}}\sum_{k=l+1}^{M}\lambda_k \asymp \frac{N\lambda_{l+1}}{N\lambda_{l+1}} \asymp 1$$

for all $l = 1, ..., M-1$. Hence the condition (i) in Theorem B.7 would hold for some $l = N/c_1$ where $c_1 > 1$. Then we apply Theorem B.7 for polynomial decay: there exists constants C, C', with a probability at least $1 - Ce^{-N/C}$, we have

$$\mathcal{V} \geq C'\left(\frac{l}{N} + \frac{N\sum_{k>l}\lambda_k^2}{\left(\sum_{k>l}\lambda_k\right)^2}\right) = \Omega\left(\frac{l}{N} + \frac{N\int_l^\infty t^{-2a}dt}{\left(\int_l^\infty t^{-2a}\right)^2}\right) = \Omega(1).$$

Hence, combining the result with Theorem A.9, there exists constants $c_1, c_2 > 0$ independent to N such that $c_1 \leq \mathcal{R} \leq c_2$ with the probability stated in Theorem A.9. Take $C = \max\{(c_1^{-1}+1)^{-1}, (c_2+1)^{-1}\}$ to conclude the claim. □

Theorem A.11 (catastrophic overfitting)**.** *Suppose interpolation assumption (Assumption 3.1) and sub-Gaussian design assumption (Assumption 3.4) hold. Then there exists some constants $c_1, c_2 > 0$ independent to N such that with probability at least $1 - c_1e^{-N/c_1}$, we have*

$$\mathcal{R} \geq c_2N.$$

Proof. We argue with Theorem B.7 again. We compute the (normalized) effective rank:

$$\rho_l \overset{\text{def.}}{=} \frac{1}{N\lambda_{l+1}}\sum_{k=l+1}^{M}\lambda_k \asymp \frac{\lambda_{l+1}}{N\lambda_{l+1}} \asymp \frac{1}{N}$$

for all $l = 1, ..., M-1$. Hence the condition (i) or (ii) in Theorem B.7 would hold for some $l < N$. Then we apply Theorem B.7 for exponential decay: there exists a constant C, C', with a probability at least $1 - Ce^{-N/C}$, we have

$$\mathcal{R} \geq \mathcal{V} \geq C'\left(\frac{l}{N} + \frac{N\sum_{k>l}\lambda_k^2}{\left(\sum_{k>l}\lambda_k\right)^2}\right) = \Omega\left(\frac{Ne^{-2al}}{e^{-2al}}\right) = \Omega(N).$$

□

It may be of independent interest for the trivial bound of $s_{\min}$.

Lemma A.12 (Trivial bound of the smallest singular value)**.** *Suppose the entries of the feature vector $\boldsymbol{\psi} \in \mathbb{R}^p$ are i.i.d. draws of a sub-Gaussian variable. (In particular, Assumption 3.3 holds.) Then there exists constants $c_1, c_2 > 0$ such that, for any $\epsilon > 0$, with probability at least $1 - c_1\epsilon - e^{-c_2 N}$, it holds that*

$$s_{\min}(\mathbf{K}) \geq \frac{\epsilon^2 \lambda_N}{N}.$$

Proof. Observe that

$$s_{\min}(\mathbf{K}) \geq s_{\min}(\mathbf{\Psi}_{\leq N}^\top \mathbf{\Lambda}_{\leq N} \mathbf{\Psi}_{\leq N}) \geq \lambda_N s_{\min}(\mathbf{\Psi}_{\leq N}^\top \mathbf{\Psi}_{\leq N}) = \lambda_N s_{\min}(\mathbf{\Psi}_{\leq N})^2 \geq \lambda_N \frac{\epsilon^2}{N} \geq \frac{\epsilon^2 \lambda_N}{N}.$$

where second last inequality holds with probability at least $1 - c_1\epsilon - e^{-c_2 N}$ by Theorem B.12. □

A.3. Finite rank approximation of kernels

In this subsection, we discuss approximating the overfitting behavior of kernel ridge regression using truncated kernels. This serves as a justification for the finite rank assumption in both Gaussian and sub-Gaussian design assumptions.

Since we focus on overfitting behaviour, where the variance term $\mathcal{V}$ dominates over the bias term $\mathcal{B}$ (see (Cui et al., 2021; Li et al., 2023a) for details), we show that the variance term from the infinite rank kernel is close to that from its high-rank truncation.

Proposition A.13 (Finite rank kernel)**.** *Let K be a PDS kernel with Mercer decomposition*

$$K(x, x') = \sum_{k=1}^{\infty} \lambda_k \psi_k(x) \psi_k(x')$$

with strictly positive eigenvalues $\lambda_1 \geq \lambda_2 \geq ...$ and corresponding eigenfunction ψ_k's. For any integer $M > N$, define its truncation:

$$K^{(M)}(x, x') = \sum_{k=1}^{M} \lambda_k \psi_k(x) \psi_k(x').$$

Denote by $\mathcal{V}$ the variance corresponding to the kernel K, and by $\mathcal{V}(M)$ that corresponding to $K^{(M)}$. Fix a random sample $\{x_i\}_{i=1}^N$ of size N. Then there exists an integer $M_0 > N$, such that

$$|\mathcal{V} - \mathcal{V}(M)| \leq 3\mathcal{V}(M) + \frac{\sigma^2}{N}$$

whenever $M > M_0$.

Proof. Consider the variance expression in Lemma B.15, we have:

$$\mathcal{V} = \sigma^2 \operatorname{Tr}\left[(\mathbf{\Psi}^\top \mathbf{\Lambda}^2 \mathbf{\Psi})(\mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi})^{-2}\right], \quad \mathcal{V}(M) = \sigma^2 \operatorname{Tr}\left[(\mathbf{\Psi}_{\leq M}^\top \mathbf{\Lambda}_{\leq M}^2 \mathbf{\Psi}_{\leq M})(\mathbf{\Psi}_{\leq M}^\top \mathbf{\Lambda}_{\leq M} \mathbf{\Psi}_{\leq M})^{-2}\right].$$

To simplify the notation, let

$$\mathbf{A}_1 = \mathbf{\Psi}_{\leq M}^\top \mathbf{\Lambda}_{\leq M} \mathbf{\Psi}_{\leq M}, \quad \mathbf{\Delta}_1 = \mathbf{\Psi}_{>M}^\top \mathbf{\Lambda}_{>M} \mathbf{\Psi}_{>M}$$
$$\mathbf{A}_2 = \mathbf{\Psi}_{\leq M}^\top \mathbf{\Lambda}_{\leq M}^2 \mathbf{\Psi}_{\leq M}, \quad \mathbf{\Delta}_2 = \mathbf{\Psi}_{>M}^\top \mathbf{\Lambda}_{>M}^2 \mathbf{\Psi}_{>M}.$$

Note that the matrices $\mathbf{A}_1, \mathbf{A}_2, \mathbf{\Delta}_1, \mathbf{\Delta}_2$ depends on M and are PDS a.s. Write the singular values as functions of M:

$$p_1(M) = s_{\max}(\mathbf{\Delta}_1), \quad q_1(M) = s_{\min}(\mathbf{A}_1)$$
$$p_2(M) = s_{\max}(\mathbf{\Delta}_2), \quad q_2(M) = s_{\min}(\mathbf{A}_2).$$

By Lemma B.14, both p_1 and p_2 are decreasing functions in M, and both q_1 and q_2 are increasing functions in M. Moreover, by the entry-wise convergence of $K^{(M)}(x,x') \to K(x,x')$, we have a.s.: [2]

$$\begin{aligned}\lim_{M\to\infty} p_1(M) = 0, &\quad \lim_{M\to\infty} q_1(M) = Q_1,\\ \lim_{M\to\infty} p_2(M) = 0, &\quad \lim_{M\to\infty} q_2(M) = Q_2.\end{aligned}$$

where $Q_1 \overset{\text{def.}}{=} s_{\min}(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi}) > 0$ and $Q_2 \overset{\text{def.}}{=} s_{\min}(\mathbf{\Psi}^\top \mathbf{\Lambda}^2\mathbf{\Psi}) > 0$ a.s.

Fix an $\epsilon > 0$, then there exists an integer $M_0 > N$ such that

$$\begin{aligned}p_1(M) \le \epsilon, &\quad q_1(M) \ge \frac{Q_1}{2},\\ p_2(M) \le \epsilon, &\quad q_2(M) \ge \frac{Q_2}{2},\end{aligned}$$

whenever $M > M_0$. In such case, we have:

$$\begin{aligned}(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2} &= \left((\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right)^2\\ &= \left((\mathbf{A}_1^{-1} - \mathbf{A}_1^{-1}\mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right)^2\\ &= \mathbf{A}_1^{-2}\left(\mathbf{I} - \mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right)^2,\end{aligned}$$

where we use the identity of matrix inverse difference: $(\mathbf{M}_1 + \mathbf{M}_2)^{-1} = \mathbf{M}_1^{-1} - \mathbf{M}_1^{-1}\mathbf{M}_2(\mathbf{M}_1 + \mathbf{M}_2)^{-1}$. Then,

$$\begin{aligned}\left\|\left(\mathbf{I} - \mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right)^2\right\|_{\text{op}} &\le \left\|\mathbf{I} - \mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right\|_{\text{op}}^2\\ &\le \left(1 + \left\|\mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right\|_{\text{op}}\right)^2\\ &\le \left(1 + \left\|\mathbf{\Delta}_1\right\|_{\text{op}}\left\|(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right\|_{\text{op}}\right)^2\\ &\le \left(1 + \left\|\mathbf{\Delta}_1\right\|_{\text{op}}\left\|\mathbf{A}_1^{-1}\right\|_{\text{op}}\right)^2\\ &= \left(1 + s_{\max}(\mathbf{\Delta}_1)s_{\min}(\mathbf{A}_1)^{-1}\right)^2\\ &\le \left(1 + \epsilon\cdot\frac{2}{Q_1}\right)^2.\end{aligned}$$

Hence

$$\begin{aligned}\mathcal{V}/\sigma^2 &= \operatorname{Tr}\left[(\mathbf{\Psi}^\top \mathbf{\Lambda}^2\mathbf{\Psi})(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}\right]\\ &= \operatorname{Tr}\left[(\mathbf{A}_2 + \mathbf{\Delta}_2)(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}\right]\\ &= \operatorname{Tr}\left[\mathbf{A}_2(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}\right] + \operatorname{Tr}\left[\mathbf{\Delta}_2(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}\right]\\ &= \operatorname{Tr}\left[\mathbf{A}_2\mathbf{A}_1^{-2}\left(\mathbf{I} - \mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1}\right)^2\right] + \operatorname{Tr}\left[\mathbf{\Delta}_2(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}\right]\\ &\le \left\|(\mathbf{I} - \mathbf{\Delta}_1(\mathbf{A}_1 + \mathbf{\Delta}_1)^{-1})^2\right\|_{\text{op}}\operatorname{Tr}\left[\mathbf{A}_2\mathbf{A}_1^{-2}\right] + \left\|\mathbf{\Delta}_2\right\|_{\text{op}}\left\|(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}\right\|_{\text{op}}\operatorname{Tr}\left[\mathbf{I}_N\right]\\ &\le \left(1 + \epsilon\cdot\frac{2}{Q_1}\right)^2\mathcal{V}(M)/\sigma^2 + p_2(M)s_{\min}(\mathbf{\Psi}^\top \mathbf{\Lambda}\mathbf{\Psi})^{-2}N\\ &\le \left(1 + \epsilon\cdot\frac{2}{Q_1}\right)^2\mathcal{V}(M)/\sigma^2 + \epsilon Q_1^{-2}N,\end{aligned}$$

[2] In more details, we have $0 \le \lim_{M\to\infty} s_{\max}(\mathbf{\Delta}_1) \le \lim_{M\to\infty}\left\|\mathbf{\Delta}_1\right\|_F \to 0$; by Lemma B.14 and Weyl's interlacing Theorem, $s_{\min}(\mathbf{K}) \ge \lim_{M\to\infty} s_{\min}(\mathbf{A}_1) \ge \lim_{M\to\infty}(s_{\min}(\mathbf{K}) - s_{\max}(\mathbf{\Delta}_1)) \ge s_{\min}(\mathbf{K}) - \lim_{M\to\infty}(s_{\max}(\mathbf{\Delta}_1)) \to s_{\min}(\mathbf{K})$. Argue similarly for p_2 and q_2.

where we use that fact that $\mathrm{Tr}[\mathbf{M}_1\mathbf{M}_2] \leq \|\mathbf{M}_1\|_{\mathrm{op}} \mathrm{Tr}[\mathbf{M}_2]$ for any PDS matrix $\mathbf{M}_1$ in the first inequality. Now set $\epsilon = \min\{\frac{Q_1}{2}, \frac{1}{Q_1^2 N^2}\}$, we have

$$|\mathcal{V} - \mathcal{V}(M)| \leq 3\mathcal{V}(M) + \frac{\sigma^2}{N}.$$

□

In Proposition A.13, we can see that for each fixed sample of size N, we can find a truncation level M large enough so that the decay of the variance $\mathcal{V}$ is of the same magnitude of $\mathcal{V}(M)$. The extra term does not play an important role in the case of analysing tempered overfitting where $\mathcal{V} = \Theta\left(\sigma^2\right)$ or catastrophic overfitting where $\mathcal{V} \to \infty$.

B. Technical Lemmata

This section contains known results from previous work that we use for our main theorems.

Proposition B.1 (Proposition 2.5 in (Rudelson & Vershynin, 2008))**.** *Let G be a $n \times k$ matrix whose entries are independent centered random variables with variances at least 1 and fourth moments bounded by B. Let $K \geq 1$. Then there exist $C_1, C_2 > 0$ and $\delta_0 \in (0,1)$ that depend only on B and K such that if $k < \delta_0 n$ then*

$$\mathbb{P}\left\{\inf_{v \in \mathbb{S}^{k-1}} \|Gv\|_2 \leq C_1 n^{1/2}, \|G\|_{op} \leq K n^{1/2}\right\} \leq e^{-C_2 n}.$$

If the random variable is sub-Gaussian, the condition on the operator norm $\|G\|_{op} \leq K n^{1/2}$ can be dropped.

Theorem B.2 (Corollary 5.35 in (Vershynin, 2010))**.** *Let $\mathbf{A}$ be an $N \times n$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - \exp\left(-t^2/2\right)$, we have*

$$\sqrt{N} - \sqrt{n} - t \leq s_{\min}(\mathbf{A}) \leq s_{\max}(\mathbf{A}) \leq \sqrt{N} + \sqrt{n} + t.$$

Theorem B.3 (Theorem 5.39 and Remark 5.40 in (Vershynin, 2010))**.** *Let $\mathbf{A}$ be an $N \times n$ matrix with independent rows $\mathbf{A}_i$ of sub-Gaussian random vector with covariance $\mathbf{\Sigma} \overset{\text{def.}}{=} \mathbb{E}\left[\mathbf{A_i}\mathbf{A_i}^\top\right] \in \mathbb{R}^{n \times n}$. Then there exists constants $C_3, C_4 > 0$ (depending only on the sub-Gaussian norm of entries of $\mathbf{A}$), such that for any $t \geq 0$, with probability at least $1 - 2e^{-C_3 t^2}$, we have*

$$\left\|\frac{1}{N}\mathbf{A}^\top\mathbf{A} - \mathbf{\Sigma}\right\|_{op} \leq \max\{\delta, \delta^2\} \|\Sigma\|_{op}.$$

where $\delta = C_4\sqrt{\frac{n}{N}} + \frac{t}{N}$. In particular, if $\mathbf{\Sigma} = \mathbf{I}_n$, we have

$$\sqrt{N} - \sqrt{C_4 n} - t \leq s_{\min}(\mathbf{A}) \leq s_{\max}(\mathbf{A}) \leq \sqrt{N} + \sqrt{C_4 n} + t.$$

Theorem B.4 (Theorem 9 (modified) in (Koltchinskii & Lounici, 2017))**.** *Let $\mathbf{A}$ be an $N \times n$ matrix whose columns are i.i.d. sub-Gaussian centered random vectors with covariance $\mathbf{\Sigma}$. Then there exists a constant $C > 0$, such that, for any $t \geq 1$, with probability at least $1 - e^{-t}$, it holds that*

$$\left\|\frac{1}{n}\mathbf{A}\mathbf{A}^\top - \mathbf{\Sigma}\right\|_{op} \leq C\|\mathbf{\Sigma}\|_{op} \min\left\{\sqrt{\rho}, \rho, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

where $\rho = \frac{\mathrm{Tr}[\mathbf{\Sigma}]}{n\|\mathbf{\Sigma}\|_{op}}$ is the (re-scaled) effect rank of the covairance $\mathbf{\Sigma}$.

Remark B.5 (Dimension-free bound). Theorem B.4 differs from Theorem B.3 in that the bound in the former contains both dimensions N and n, while the latter only contains n.

Theorem B.6 (Theorem 2.5 in (Tsigler & Bartlett, 2023))**.** *Suppose Assumption 3.3 holds. Let $\mathbf{A}_l = \lambda \mathbf{I}_N + \sum_{k=l+1}^M \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \in \mathbb{R}^{N \times N}$. Then there exists a constant $c > 0$, such that for any $l < N/c$, with probability of at least $1 - ce^{-N/c}$, if $\mathbf{A}_l$ is positive definite, then*

$$\begin{aligned}\mathcal{B}/c &\leq \|\boldsymbol{\theta}^*_{>l}\|^2_{\mathbf{\Lambda}_{>l}}\left(1 + \frac{s_1(\mathbf{A}_l^{-1})^2}{s_N(\mathbf{A}_l^{-1})^2} + N\lambda_{l+1}s_1(\mathbf{A}_l^{-1})\right)\\ &\quad + \|\boldsymbol{\theta}^*_{\leq l}\|^2_{\mathbf{\Lambda}_{\leq l}^{-1}}\left(\frac{1}{N^2 s_N(\mathbf{A}_l^{-1})^2} + \frac{\lambda_{l+1}}{N}\frac{s_1(\mathbf{A}_l^{-1})}{s_N(\mathbf{A}_l^{-1})^2}\right)\\ \mathcal{V}/c &\leq \frac{s_1(\mathbf{A}_l^{-1})^2}{s_N(\mathbf{A}_l^{-1})^2}\frac{l}{N} + N s_1(\mathbf{A}_l^{-1})^2 \sum_{k>l}\lambda_k^2,\end{aligned}$$

where $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*_{\leq l} \oplus \boldsymbol{\theta}^*_{>l}$ is the splitting of the target function coefficient; and $\|\mathbf{v}\|_{\mathbf{M}} \overset{\text{def.}}{=} \sqrt{\mathbf{v}^\top\mathbf{M}\mathbf{v}}$ for any vector $\mathbf{v}$ and matrix $\mathbf{M}$ with appropriate dimension.

Theorem B.7 (Lemma 7 and Theorem 10 in (Tsigler & Bartlett, 2023))**.** *Suppose sub-Gaussian design assumption 3.3 holds. In addition, fix constants $A > 0, B > \frac{1}{N}$ and suppose either (i) the (normalized) effective rank $\rho_l \overset{\text{def.}}{=} \frac{1}{N\lambda_{l+1}} \sum_{k=l+1}^{M} \lambda_k \in (A, B)$; or (ii) $l = \min\{\ell : \rho_\ell > B\}$. Then there exists a constant C, C', such that if $l < N/C$, with a probability at least $1 - Ce^{-N/C}$, we have*

$$\mathcal{V} \geq C' \left(\frac{l}{N} + \frac{N \sum_{k>l} \lambda_k^2}{\left(\sum_{k>l} \lambda_k\right)^2} \right).$$

Lemma B.8 (Negative second moment identity, Exercise 2.7.3 in (Tao, 2012))**.** *Let $\mathbf{M}$ be an invertible $n \times n$ matrix, let $\mathbf{R}_1, ..., \mathbf{R}_n$ be the rows of $\mathbf{M}$ and let $\mathbf{C}_1, ..., \mathbf{C}_n$ be the columns of $\mathbf{M}^{-1}$. For each $1 \leq i \leq n$, let $\mathbf{N}_i$ be a unit normal vector orthogonal to the subspace spanned by the all rows $\mathbf{R}_1, ..., \mathbf{R}_n$ except $\mathbf{R}_i$. Then we have*

$$\|\mathbf{C}_i\|_2^2 = (\mathbf{R}_i^\top \mathbf{N}_i)^{-2} \text{ and } \sum_{i=1}^{n} s_i(\mathbf{M})^{-2} = \sum_{i=1}^{n} (\mathbf{R}_i^\top \mathbf{N}_i)^{-2}.$$

Proof. Note that $\mathbf{R}_i^\top \mathbf{C}_j = \delta_{ij}$ and the rows $\mathbf{R}_i$'s spans the space $\mathbb{R}^N$. Hence we have $\mathbf{C}_i = \pm \|\mathbf{C}_i\|_2 \mathbf{N}_i$ for all i and $\|\mathbf{C}_i\|_2^2 = (\mathbf{R}_i^\top \mathbf{C}_i / \mathbf{R}_i^\top \mathbf{N}_i)^2 = (\mathbf{R}_i^\top \mathbf{N}_i)^{-2}$ which proves the first statement. For the second statement, note that

$$\sum_{i=1}^{n} \lambda_i(\mathbf{M})^{-2} = \sum_{i=1}^{n} \lambda_i(\mathbf{M}^{-1})^2 = \mathrm{Tr}[(\mathbf{M}^{-1})^\top (\mathbf{M}^{-1})] = \sum_{i=1}^{n} \|\mathbf{C}_i\|_2^2 = \sum_{i=1}^{n} (\mathbf{R}_i^\top \mathbf{N}_i)^{-2}.$$

□

Lemma B.9 (lower bound of $s_{\min}$)**.** $\mathbf{K}_N = \sum_{k=1}^{N} \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \prec \mathbf{K}$*. Let $\mathbf{\Lambda}_N = \mathrm{diag}(\lambda_k)_{k=1}^{N} \in \mathbb{R}^{N \times N}$ and $\mathbf{\Psi}_N = (\boldsymbol{\psi}_k)_{k=1}^{N} \in \mathbb{R}^{N \times N}$ and set $\mathbf{M} = \mathbf{\Lambda}_N^{1/2} \mathbf{\Psi}_N$ which is invertible almost surely. Note that $\mathbf{K}_N = \mathbf{M}^\top \mathbf{M}$. Let $\mathbf{R}_1, ..., \mathbf{R}_n$ be the rows of $\mathbf{M}$ and let $\mathbf{C}_1, ..., \mathbf{C}_n$ be the columns of $\mathbf{M}^{-1}$. For each $1 \leq i \leq n$, let $\mathbf{N}_i$ be a unit normal vector orthogonal to the subspace spanned by the all rows $\mathbf{R}_1, ..., \mathbf{R}_n$ except $\mathbf{R}_i$. we have*

$$s_{\min}(\mathbf{K}) \geq \frac{\lambda_N}{\sum_{k=1}^{N} \frac{\lambda_N}{\lambda_k} (\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2}}.$$

Proof. Since $s_{\min} \geq s_N(\mathbf{K}_N)$, WLOG: assume $M = N$. Then by Lemma B.8,

$$s_N(\mathbf{K}_N)^{-1} \leq \sum_{k=1}^{N} s_k(\mathbf{K}_N)^{-1} = \sum_{k=1}^{N} s_k(\mathbf{M})^{-2} = \sum_{k=1}^{N} \left(\sqrt{\lambda_k} \boldsymbol{\psi}_k^\top \mathbf{N}_k \right)^{-2},$$

where $\mathbf{N}_k$ denote a unit normal vector orthogonal to the subspace spanned by the all rows $\mathbf{R}_1, ..., \mathbf{R}_n$ of $\mathbf{M}$ except $\mathbf{R}_i$. Hence

$$s_{\min} \geq s_N(\mathbf{K}_N) \geq \frac{\lambda_N}{\sum_{k=1}^{N} \frac{\lambda_N}{\lambda_k} (\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2}}. \tag{8}$$

□

Lemma B.10 (Anti-Concentration Result For Gaussian Laws)**.** *Let g be a standard Gaussian variable, then*

$$\mathbb{P}\{|g| \leq t\} \leq \frac{2t}{\sqrt{2\pi}}, \ \forall t \geq 0. \tag{9}$$

Lemma B.11 (Sub-Exponential Deviation, see Corollary 5.17 in (Vershynin, 2010))**.** *Let $N \in \mathbb{N}$. Let $X_1, ..., X_N$ be independent centered random variables with sub-exponential norms bounded by B. Then for any $\delta > 0$,*

$$\mathbb{P}\left\{ |\sum_{i=1}^{N} X_i| > \delta N \right\} \leq 2 \exp\left(-C_5 \min\left\{ \frac{\delta^2}{B^2}, \frac{\delta}{B} \right\} N \right),$$

where $C_5 > 0$ is an absolute constant.

In particular, if $X \sim \chi(N)$ is the Chi-square distribution, then $\mathbb{P}\left\{ |\frac{X}{N} - 1| > t \right\} \leq 2e^{-Nt^2/8}$, $\forall t \in (0, 1)$.

Theorem B.12 (Theorem 1.1 in (Rudelson & Vershynin, 2009) /Theorem 5.38 in (Vershynin, 2010))**.** *Let $\mathbf{A}$ be an $N \times n$ random matrix whose entries are i.i.d. sub-Gaussian random variables with zero mean and unit variance. Then there exists constants $C_6 > 0, C_7 \in (0,1)$ such that for any $\delta > 0$,*

$$\mathbb{P}\left\{s_{\min}(\mathbf{A}) \leq \delta(\sqrt{N} - \sqrt{n-1})\right\} \leq (C_6\delta)^{N-n+1} + C_7^N.$$

In particular, if $N = n$,

$$s_{\min}(\mathbf{A}) \gtrsim N^{-1/2}$$

with high probability.

Theorem B.13 (Theorem 5.58 in (Vershynin, 2010))**.** *Let $\mathbf{A}$ be an $N \times n$ matrix ($N \geq n$) with independent columns $\mathbf{A}_i \in \mathbb{R}^N$ of sub-Gaussian isotropic random vector with with $\|\mathbf{A}_i\|_2 = \sqrt{N}$ almost surely. Then there exists constants $C_8, C_9 > 0$ (depending only on the sub-Gaussian norm of entries of $\mathbf{A}$), such that for any $t \geq 0$, with probability at least $1 - 2e^{-C_8 t^2}$, we have*

$$\sqrt{N} - C_9\sqrt{n} - t \leq s_{\min}(\mathbf{A}) \leq s_{\max}(\mathbf{A}) \leq \sqrt{N} + C_9\sqrt{n} + t.$$

Lemma B.14 (Corollary 4.3.12 in (Horn & Johnson, 2012))**.** *Let $\mathbf{M}_1, \mathbf{M}_2$ are symmetric matrix. If $\mathbf{M}_2$ is positive semi-definite, then*

$$s_{\max}(\mathbf{M}_1) \leq s_{\max}(\mathbf{M}_1 + \mathbf{M}_2), \quad s_{\min}(\mathbf{M}_1) \leq s_{\min}(\mathbf{M}_1 + \mathbf{M}_2).$$

Lemma B.15 (Variance expression)**.** *Recall the definition of the variance $\mathcal{V} \overset{\text{def.}}{=} \mathcal{R} - \mathcal{B}$. In the case of kernel ridgeless regression where $\lambda = 0$ and the kernel matrix $\mathbf{K}$ can be written as $\mathbf{K} = \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi}$ by Mercer decomposition, the variance admits the following expression:*

$$\mathcal{V} = \sigma^2 \operatorname{Tr}\left[(\mathbf{\Psi}^\top \mathbf{\Lambda}^2 \mathbf{\Psi})(\mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi})^{-2}\right].$$

Proof. By definition,

$$\begin{aligned}
\mathcal{V} &= \mathcal{R} - \mathcal{B} \\
&= \mathbb{E}_{x,\epsilon}\left[(f^*(x) - \hat{f}(x))^2\right] - \mathbb{E}_{x,\epsilon}\left[(f^*(x) - \mathbf{K}_x^\top \mathbf{K}^{-1} f^*(\mathbf{X}))^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[(\mathbf{K}_x^\top \mathbf{K}^{-1} \boldsymbol{\epsilon})^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\boldsymbol{\epsilon}^\top \mathbf{K}^{-1} \mathbf{K}_x \mathbf{K}_x^\top \mathbf{K}^{-1} \boldsymbol{\epsilon}\right] \\
&= \mathbb{E}_{\epsilon}\left[\boldsymbol{\epsilon}^\top \mathbf{K}^{-1} \mathbf{\Psi}^\top \mathbf{\Lambda}^2 \mathbf{\Psi} \mathbf{K}^{-1} \boldsymbol{\epsilon}\right] \\
&= \mathbb{E}_{\epsilon}\left[\operatorname{Tr}\left[\mathbf{K}^{-1} \mathbf{\Psi}^\top \mathbf{\Lambda}^2 \mathbf{\Psi} \mathbf{K}^{-1} \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right]\right] \\
&= \sigma^2 \operatorname{Tr}\left[\mathbf{K}^{-1} \mathbf{\Psi}^\top \mathbf{\Lambda}^2 \mathbf{\Psi} \mathbf{K}^{-1}\right] \\
&= \sigma^2 \operatorname{Tr}\left[(\mathbf{\Psi}^\top \mathbf{\Lambda}^2 \mathbf{\Psi})(\mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi})^{-2}\right].
\end{aligned}$$

□