CHARACTERIZING OVERFITTING IN KERNEL RIDGELESS REGRESSION THROUGH THE EIGENSPECTRUM

A PREPRINT

Tin Sum Cheng **Aurelien Lucchi**
Department of Mathematics and Computer Science
University of Basel
{tinsum.cheng, aurelien.lucchi}@unibas.ch

Anastasis Kratsios
Department of Mathematics and Statistics,
McMaster University and Vector Institute,
kratsioa@mcmaster.ca

David Belius
Faculty of Mathematics and Computer Science
UniDistance Suisse
david.belius@cantab.ch

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 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 conclusion on overfitting is two-fold: (i) kernel regressors whose eigenspectrum decays polynomially must generalize well, even in the presence of noisy labeled training data; these models exhibit so-called tempered overfitting; (ii) if the eigenspectrum of any kernel ridge regressor decays exponentially, then it generalizes poorly, i.e., it exhibits catastrophic overfitting. This adds to the available characterization of kernel ridge regressors exhibiting benign overfitting as the extremal case where the eigenspectrum of the kernel decays sub-polynomially. Our analysis combines new random matrix theory (RMT) techniques with recent tools in the kernel ridge regression (KRR) literature.

1 Introduction

Kernel regression 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 and Rakhlin [2020], Mei and Montanari [2022], Misiakiewicz [2022], also known as the ultra-high dimensional regime Zou and 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 and Mendelson [2002], require that the weights of the kernel regressor satisfy data-independent a-priori 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.

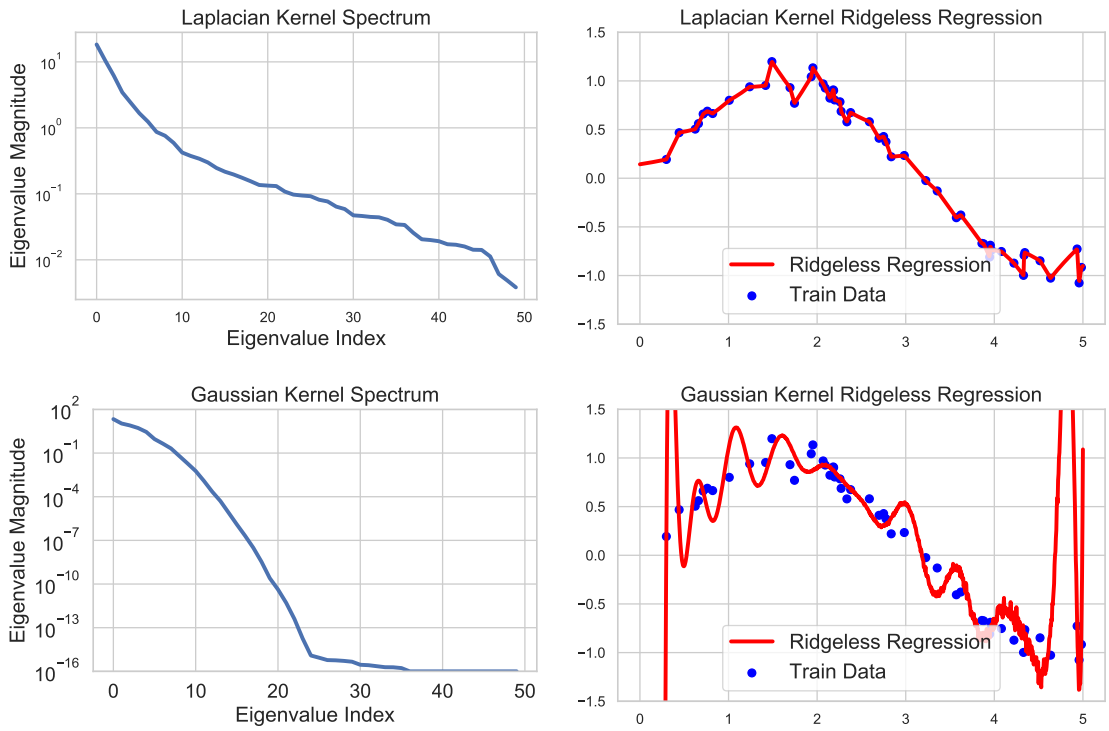

Figure 1: Kernel spectra for Laplacian and Gaussian kernels and their overfitting behaviours.
Tempered Overfitting: The empirical kernel spectra of the Laplacian kernel decay moderately (top left), and so does the quality of its test-set performance as one departs from the training data (top right).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

Contributions This work aims at developing novel test error bounds for KRR in the setting of finite input dimension and sample size.

As a summary, our main contributions are:

1. We obtain a high-probability bound on the condition number of the empirical kernel matrix (Theorem 4.1);
2. We derive *tight* non-asymptotic upper and lower bounds for the test error of the minimum norm interpolant with polynomially decaying spectrum in the fixed input dimension setting. Consequentially, we show that this regime yields *tempered overfitting* (Theorem 4.2);
3. On the other hand, we show that the minimum norm interpolant with exponentially decaying spectrum must exhibit *catastrophic overfitting* (Theorem 4.3).

This mirrors the special case identified in Mallinar et al. [2022], which showed that the neural tangent kernel (NTK) and Laplacian kernel (polynomial spectra) generalize well even without ridge while the Gaussian kernel (exponential spectrum) does not. The correspondence between polynomial and exponential spectral decay rates and tempered and catastrophic overfitting regimes is illustrated in Figure 1.

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. Based on our findings, we formulate a conjecture for future research.

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 the proof of our main results in a simpler setting (under Gaussian Design Assumption A.1) for the sake of simplicity.
7. In Section B, we extend our proof to the general setting (under Sub-Gaussian Design Assumption 3.2).
8. In Section C, 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 and Rakhlin [2020], Bartlett et al. [2020], Tsigler and Bartlett [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 training instances. 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 Gaussian Assumption

Many previous works Jacot et al. [2020], Bordelon et al. [2020], Simon et al. [2021] require the universality assumption on the eigenfunctions evaluated on the training set, namely, the entries are i.i.d. Gaussians, to prove their results on KRR generalization. (See Assumption A.1 in Section 3 for details.) In contrast, we obtain tight bounds on test error in the more widely applicable sub-Gaussian setting. We note that for simplicity of exposition, we first showcase our results under the Gaussian Design Assumption A.1 and, once explaining our proof strategy, we extend our argument to the general sub-Gaussian setting in Section B.

2.2 Test Error on Ridgeless Regression

Many previous works Arora et al. [2019], Liang and 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] have devoted on bounding the KRR test error in different settings. In the context of benign overfitting, a recent related paper Tsigler and 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. The benign overfitting case has already been characterized by Barzilai and Shamir [2023], and we characterize the tempered and catastrophic overfitting cases.

They follow a similar approach by characterizing the overfitting by the kernel spectral decay. However, their analysis is based on a proxy of the test error by Simon et al. [2021] where they use the Gaussian Design Assumption A.1. Our paper recovers their result with Sub-Gaussian Design Assumption 3.2 instead. We refer the reader to Figure 2 for visualization and Definition 3.9 for further details.

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. Our analysis yields tighter bounds for the class of kernels to which our analysis applies than theirs, which is accomplished via tighter bounds on the involved kernel eigenspectrum. Importantly, unlike their results, our analysis provides upper and *matching*

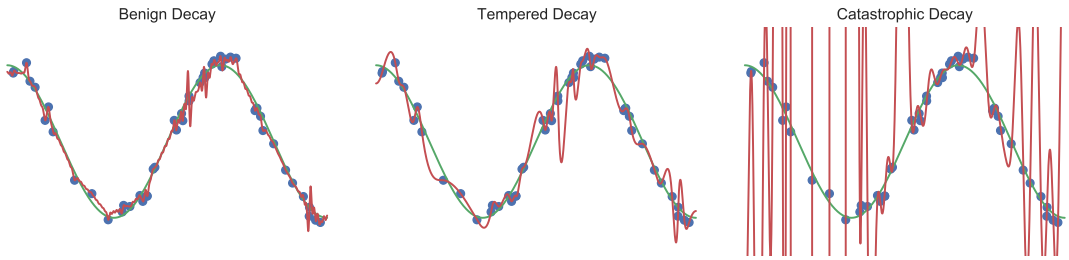

Figure 2: A cartoon description for benign, tempered and catastrophic overfitting regimes. We characterize the tempered and catastrophic overfitting regimes, where the out-of-sample performance is controlled and uncontrollable, respectively.

lower bounds bounds on the test error. Also, our bound on the condition number of the kernel matrix is tight even for exponential decay, while the bound in Barzilai and Shamir [2023] [1] becomes vacuous in that setting.

Though our analysis covers a mildly smaller class of kernels, we expect one could extend our arguments to cover the broader class of kernels studied in their analysis.

Table 1: Comparison with prior works

	Mallinar et al. [2022]	Tsigler and Bartlett [2023]	Barzilai and Shamir [2023]	**This paper**
Assumption on kernel	Gaussian feature	Bound on condition number	Kernel cont. and bdd.	Sub-Gaussian 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:

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

Given a data-distribution μ on the input space $\mathcal{X}$, we 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. In the context of over-parametrized machine learning, we assume the kernel is of finite-rank M and is much larger than the sample size N:

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

[1] 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 and Shamir [2023] is the general setting under which they perform their analysis. However, our results can address the difficult case of exponential decay and the two works are in conclusion complementary of each other.

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 assume sub-Gaussianity of the eigenfunctions ψ_k's, which is very standard in KRR literature(to name a few, Liang and Rakhlin [2020], Bartlett et al. [2020], Bach [2023], Tsigler and Bartlett [2023], Cheng et al. [2023]):

Assumption 3.2 (Sub-Gaussian Design)**.** *The sub-Gaussian norm of the random variables $\psi_k(x)$'s are uniformly bounded.*

We remark that Assumption 3.2 is much more general and realistic than the Gaussian Design Assumption A.1 which the papers cited in Subsection 2.1 assumed.

However, for simplicity, we will first prove our result in Gaussian Assumption A.1 in Section A; then in Section B, we will extend our proof to sub-Gaussian case accordingly.

In general, the random variables $\psi_k(x)$'s can be dependent to each other. For technical reasons, we require the independence property in **only some parts, but not all** of our statements, which we will state clearly in Sections 4, A and B. This is also quite common for KRR literature Liang and Rakhlin [2020], Bartlett et al. [2020], Bordelon et al. [2020], Simon et al. [2021], Mallinar et al. [2022], Bach [2023], Tsigler and Bartlett [2023].

Assumption 3.3 (Independent Features)**.** *The random variables $\psi_k(x)$'s are independent to each other.*

By the Representer Theorem, we know that

$$\hat{f}(x) = \mathbf{K}_x^\top (\mathbf{K} + \lambda N \mathbf{I}_N)^{-1} \mathbf{y} = \mathbf{K}_x^\top \mathbf{K}^{-1} \mathbf{y},$$

where $\mathbf{K}_x = \{K(x_i, x)\}_{i=1}^N \in \mathbb{R}^N$. The analysis on the test error inevitably converges to the analysis of the kernel matrix $\mathbf{K}$. The first analysis is the condition number of $\mathbf{K}$ given different decays. In this paper, we consider the two types of decay:

Assumption 3.4 (Exponential Decay)**.** *Assume there exists constants $\overline{r} \geq \underline{r} > 0$ and $a > 0$ such that $\underline{r} e^{-ak} \leq \lambda_k \leq \overline{r} e^{-ak}$ for all $k = 1, 2, ..., M$.*

Assumption 3.5 (Polynomial Decay)**.** *Assume there exists constants $\overline{r} \geq \underline{r} > 0$ and $a > 1$ such that $\underline{r} k^{-a} \leq \lambda_k \leq \overline{r} k^{-a}$, $\forall k = 1, ..., M$.*

Next, we bound the test error on regression task:

Assumption 3.6 (Proper Agnostic Learning)**.** *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 the noise ϵ_i's are draws from a centered sub-Gaussian random variable ϵ with variance $\mathbb{E}\left[\epsilon^2\right] = \sigma^2 > 0$. In other words, the target function decomposes as*

$$f^\star = \sum_{k=1}^{M} \gamma_k^* \left(\frac{\psi_k}{\lambda_k^{1/2}} \right),$$

where γ_k^*'s are real numbers satisfying $\sum_{k=1}^M (\gamma_k^*)^2 < \infty$. Under the Interpolation Assumption 3.1, M is finite and this inequality must hold. However, if we consider a sequence of KRR task with growing kernel rank M and dataset N, this inequality has to hold at the limit. In particular, if we write $\boldsymbol{\gamma}^* = (\gamma_k^*)_{k=1}^M \in \mathbb{R}^M$, the RKHS norm of f^* can be written as $\|f^*\|_{\mathcal{H}}^2 = \|\boldsymbol{\gamma}^*\|_2^2$ and its L2 norm as $\|f^*\|_{L_\mu^2} = \|\boldsymbol{\gamma}^*\|_{\mathbf{\Lambda}}^2$, under the notation $\|\mathbf{v}\|_{\mathbf{M}} = \sqrt{\mathbf{v}^\top \mathbf{M} \mathbf{v}}$ with vector $\mathbf{v}$ and matrix $\mathbf{M}$ with conformable dimensions.

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

Definition 3.7 (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*

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

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.

Next, we introduce some important quantities commonly defined in the KRR literature such as Tsigler and Bartlett [2023], Barzilai and Shamir [2023], which also occur several times in our analysis:

Definition 3.8 ((Normalized) effective rank)**.** *Let λ_k's be the eigenvalues of the kernel K indexed in decreasing order. Define the l^{th}- (normalized) effective rank to be*

$$\rho_l \overset{\text{def.}}{=} \frac{\sum_{k>l} \lambda_l}{N\lambda_{l+1}}. \tag{1}$$

For $l = 0$, it is the same as the well-known effective rank of the covariance $\mathbf{\Lambda}$. Also define

$$R_l \overset{\text{def.}}{=} \frac{(\sum_{k>l} \lambda_l)^2}{\sum_{k>l} \lambda_l^2}. \tag{2}$$

We define different overfitting regimes in the KRR context:

Definition 3.9 (Overfitting)**.** *Fix a spectrum λ_k's and denote by $\mathbf{\Lambda}_M \overset{\text{def.}}{=} \operatorname{diag}\{\lambda_k\}_{k=1}^M$, and fix features ψ_k's.*

Let $M = \theta N$ and $\{\mathbf{\Psi}_N\}_{N\in\mathbb{N}}$ be a family of independently drawn datasets. This gives a reproducing kernel Hilbert space (RKHS) corresponding to each kernel

$$K_M \overset{\text{def.}}{=} \mathbf{\Psi}_M^\top \mathbf{\Lambda}_M \mathbf{\Psi}_M$$

indexed by M and let $\mathcal{H}$ be their limit.

For a target function $f^* \in \mathcal{H}$, write f_M^* as its projection onto the subspace $\mathcal{H}_M$. Hence $\sum_{k>M}(\gamma_k^*)^2 \to 0$ as $M \to \infty$.

Write the test error $\mathcal{R}_N = \mathcal{R}_N(f_M^*, \epsilon)$ as a function of the target function $f_M^* \in \mathcal{H}_M$ and the noise random variable ϵ, for each dataset size N and respective model size $M = \theta N$. Assume that the limit

$$L \overset{\text{def.}}{=} \max_{f^*\in\mathcal{H}} \lim_{N\to\infty} \mathcal{R}_N \tag{3}$$

exists. We call the minimum norm interpolant of the random Gaussian Feature Model with spectral decay λ_k's:

(i) a benign overfitting, when $L = 0$;

(ii) a tempered overfitting, when $L \in (0, +\infty)$;

(iii) a catastrophic overfitting, when $L = \infty$,

which holds almost surely in terms of the randomness of the random features, input draw, and noise.

4 Main Result

Our main result consists of three stages. First, we bound the condition number of the kernel matrix $\mathbf{K}$ under Gaussian Assumption A.1 or Sub-Gaussian Assumption 3.1 in Theorem 4.1. Next, we use this result to give an upper bound of the test error in Theorem 4.2. Lastly, we give a matching lower bound of the test error in Theorems 4.2 and 4.3 and conclude the effect of the spectral decay on overfitting.

4.1 Condition Number

Theorem 4.1 (Approximation on the Condition Number)**.** *Suppose Assumptions 3.1 and 3.2 hold.*

1. ***Exponential Decay*** *If Assumptions 3.3 and 3.4 hold, then with high probability, the condition number of the kernel matrix $\mathbf{K}$ is*
$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \asymp \frac{\lambda_1}{\lambda_N} N.$$
2. ***Polynomial Decay****: If Assumption 3.5 holds. Then, with high probability, the condition number of the kernel matrix $\mathbf{K}$ is*
$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \asymp \frac{\lambda_1}{\lambda_N}.$$

Proof Idea: The first part of Theorem 4.1, showing $s_{\max}(\mathbf{K}) \asymp N\lambda_1$, is rather easy for both decays and has been discussed intensively in the past literature Vershynin [2010], Koltchinskii and Lounici [2017], Zhivotovskiy [2021]. For the sake of completeness, we include a proof for our scenario in Lemma A.2 for the Gaussian case and Lemmata B.2 and B.1 for the sub-Gaussian case. The second part of Theorem 4.1, to derive the upper bound of $s_{\min}(\mathbf{K})$, this too is not difficult and it is shown in Lemma A.6 for the Gaussian case and Lemma B.3 for the sub-Gaussian case.

The difficulty lies in the third part of Theorem 4.1; namely, in deriving a tight lower bound of $s_{\max}(\mathbf{K})$. For exponential decay, we have $\lambda_N \gtrsim \lambda_N$ in Lemma A.3 for the Gaussian case and Lemma B.5 for the sub-Gaussian case using a technique from Tao [2012]. In the case of a polynomial decay, we have $\lambda_N \gtrsim N\lambda_N$ in Lemma A.5 for the Gaussian case and Lemma B.4 using RMT results from Rudelson and Vershynin [2008], Vershynin [2010]. The formulation of the above theorem with its proof can be found in Theorem A.7 in the appendix. □

Remarkably, we uncover a novel qualitative difference of $s_{\min}(\mathbf{K})$ between exponential and polynomial decays. Let us compare with [Barzilai and Shamir, 2023, Theorem 1], wherein the authors obtain a lower bound of $s_{\min}(\mathbf{K})$ which holds with high probability:

$$s_{\min}(\mathbf{K}) \geq N\alpha_k \left(1 - \frac{1}{\delta}\sqrt{\frac{N^2}{R_k}}\right) \frac{\sum_{l>k}\lambda_l}{N}, \tag{4}$$

for any integer $k \leq N$, and some constants $\alpha_k > 0$ and $\delta \in (0,1)$, and $R_k = \frac{(\sum_{k>l}\lambda_l)^2}{\sum_{k>l}\lambda_l^2}$ as in Definition 3.8. When the eigenvalues λ_k's are exponential, then R_k becomes a constant independent to k and N. Hence the factor $\left(1 - \frac{1}{\delta}\sqrt{\frac{N^2}{R_k}}\right)$ becomes negative as $N \to \infty$ and the lower bound in line (4) becomes trivial.

In contrast, our lower bound on $s_{\min}(\mathbf{K}) \gtrsim \lambda_N$, in the exponential decay case, matches its upper bound and is much more accurate than the primitive bound:

$$\begin{aligned} s_{\min}(\mathbf{K}) &\geq s_{\min}(\mathbf{\Psi}_N^\top \mathbf{\Lambda}_N \mathbf{\Psi}_N) \\ &\geq \lambda_N s_{\min}(\mathbf{\Psi}_N^\top \mathbf{\Psi}_N) \\ &= \lambda_N s_{\min}(\mathbf{\Psi}_N)^2 \\ &\gtrsim \lambda_N (N^{-1/2})^2 \qquad (5) \\ &\geq \lambda_N / N, \end{aligned}$$

where in line (5) holds with high probability by Theorem C.10 under the independent features Assumption 3.3.

4.2 Classifying Overfitting Regimes

The result on the condition number of the kernel matrix can be applied to deduce the tempered overfitting phenomenon Mallinar et al. [2022] in kernel interpolation:

Theorem 4.2 (Tempered Overfitting for Kernels with Polynomial Decaying Spectrum)**.** *Suppose Assumptions 3.1, 3.2, 3.3, 3.5, 3.6 hold. Then there exists constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$ such that with probability at least $1 - c_1 e^{-N/c_1} - e^{-c_2 N}$, the following holds* [2]

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

In particular, as $N \to \infty$, we have

$$\mathcal{B} \to 0 \text{ and } \mathcal{V} \in O(1) \qquad \text{w.h.p.}$$

There is a constant $C > 0$ such that: for any $N \in \mathbb{N}$

$$\mathcal{V} = \Omega(1).$$

Hence, $\lim_{N\to\infty} \mathcal{R}_N = \Theta(1)$ and the minimum norm interpolant $\hat{f}$ of a kernel with polynomial decay exhibits tempered overfitting.

Proof Idea: The key insight is to use the upper bound on the condition number of the kernel matrix in Theorem 4.1 together with Theorem C.4 to bound the KRR test error. The asymptotic behaviours follow clearly from Theorems 4.1 and C.5 on lower bounding the test error using Assumption 3.3 and the effective rank in Definition 3.8.

[2] Recall the notation $\|\mathbf{v}\|_{\mathbf{M}} = \sqrt{\mathbf{v}^\top \mathbf{M} \mathbf{v}}$ introduced in Assumption 3.6.

We refer to Theorems A.8 and A.9 for the detailed formulation and proof. □

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 and Shamir, 2023, Theorem 2], but our probability is of exponential decay (in the form of $1 - c_1 e^{N/c_2}$) compared to their Markov type bound (in the form $1 - \delta - c_1 e^{N/c_2}$, where setting $\delta \to 0$ would explode the upper bound); 2) the tempered overfitting behaviour of kernel with polynomial decay, that is first reported in Mallinar et al. [2022], is based on Simon et al. [2021] which used the Gaussian Design Assumption A.1. We replace this by the more general independent sub-Gaussian Design Assumptions 3.2 and 3.3.

Another interesting observation is that adding more smaller eigenvalues to the spectrum does not harm this upper bound more than a constant factor. Let $\tilde{K}(x, x') = \sum_{k=1}^{M+M'} \lambda_k \psi_k(x)\psi_k(x') \overset{\text{def.}}{=} K(x, x') + K_{>M}(x, x')$ be the sum of our original kernel K plus some smaller kernel $K_{>M}$, with λ_k's are still in decreasing order. Then we have

$$\begin{aligned}\frac{s_{\max}(\tilde{\mathbf{K}})}{s_{\min}(\tilde{\mathbf{K}})} &\leq \frac{s_{\max}(\mathbf{K}) + s_{\max}(\mathbf{K}_{>M})}{s_{\min}(\mathbf{K}) + s_{\min}(\mathbf{K}_{>M})} \\ &\leq \frac{s_{\max}(\mathbf{K}) + \sum_{k>M} \lambda_k}{s_{\min}(\mathbf{K})} \\ &\leq c_1 \frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})},\end{aligned}$$

since the sum $\sum_{k>M} \lambda_k \lesssim M\lambda_M < N\lambda_1 \approx s_{\max}(\mathbf{K})$ in either decays. Hence the model size M in the Interpolation Assumption 3.1 can be replaced by any number $\geq \theta N$ to have Theorem 4.2 valid.

However, if the kernel has exponential decay (Assumption 3.4), the upper bound in Theorem C.4 will become vacuous, so one can show that the minimum norm interpolant suffers from catastrophic overfitting:

Theorem 4.3 (Catastrophic Overfitting for Exponential Decay)**.** *Suppose the Assumptions 3.1, 3.2, 3.3, 3.4, 3.6 hold. Then with high probability, we have*

$$\mathcal{V} = \Omega(N).$$

In particular, the minimum norm interpolant $\hat{f}$ of a kernel with exponential decay exhibits catastrophic overfitting.

Proof Idea: It is a simple application of theorem C.5 where we plug in the exponential eigenvalues in the effective rank ρ_k in Definition 3.8. We refer to Theorem A.10 for the detailed formulation and proof. □

Similarly, the catastrophic overfitting behaviour of kernel with exponential decay reported in Mallinar et al. [2022], is based on Simon et al. [2021] which used the Gaussian Design Assumption A.1. We replace this with the more general independent sub-Gaussian Design Assumptions 3.2 and 3.3.

4.3 Conjecture

Based on our observations, we know that the kernel spectrum determines the overfitting behaviour. But its converse remains as an interesting open question: *if a certain kernel exhibits benign tempered, or catastrophic overfitting, can we conclude anything on the kernel spectrum?*

We begin by a formal definition:

Definition 4.4. *We say a decreasing sequence of positive numbers λ_k's follows a:*

- *(i) polynomial decay if there exists a number $a > 1$ such that $\lambda_k = \Theta(k^{-a})$;*
- *(ii) sub-polynomial decay if $\lambda_k = \Omega(k^{-a})$ for any $a > 1$;*
- *(iii) super-polynomial decay if $\lambda_k = O(k^{-a})$ for any $a > 1$*

Conjecture: Let K be a bounded continuous positive definite symmetric (PDS) kernel. Suppose the limit L of test error in line (3) exists. Then the minimum norm interpolant $\hat{f}$ of K follows a:

(i) catastrophic overfitting ($L = \infty$) **if and only if** K has super-polynomial spectral decay (for example, exponential decay or some slower decay such as: $\lambda_k \asymp e^{p(k)}$ where $p(k) = \sum_{i=1}^{N} \beta_i \log(k)^{\alpha_i}$ for some $N \in \mathbb{N}_+$ and $\beta_1, \ldots, \beta_N, \alpha_1, \ldots, \alpha_N \geq 0$.);

(ii) tempered overfitting ($L = \Theta(1)$) **if and only if** K has polynomial spectral decay;

(iii) benign overfitting ($L = 0$) **if and only if** K has sub-polynomial decay. (for example the logarithmic-linear decay $\lambda_k = \Theta(\frac{1}{k \log^{1+a} k})$ in Barzilai and Shamir [2023] or spiked spectrum/covariance in Johnstone [2001]).

Any proofs or counterexamples will have their own significance. Also, most naturally defined kernels have either exponential or quadratic decay. Whether there exist other practical kernels with other decay rates is an interesting question for future research.

5 Experiments

We run a simple experiment to validate our theoretical analysis on overfitting. For simplicity, we implement the experiment by Assumption A.1. Let $\boldsymbol{\phi}_k \sim \mathcal{N}(0, \boldsymbol{\Lambda})$ be i.i.d. Gaussian random vector with covariance $\boldsymbol{\Lambda} = \text{diag}\{\lambda_k\}$ defined in Assumptions 3.4 or 3.5. Write $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ 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 $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$.

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,

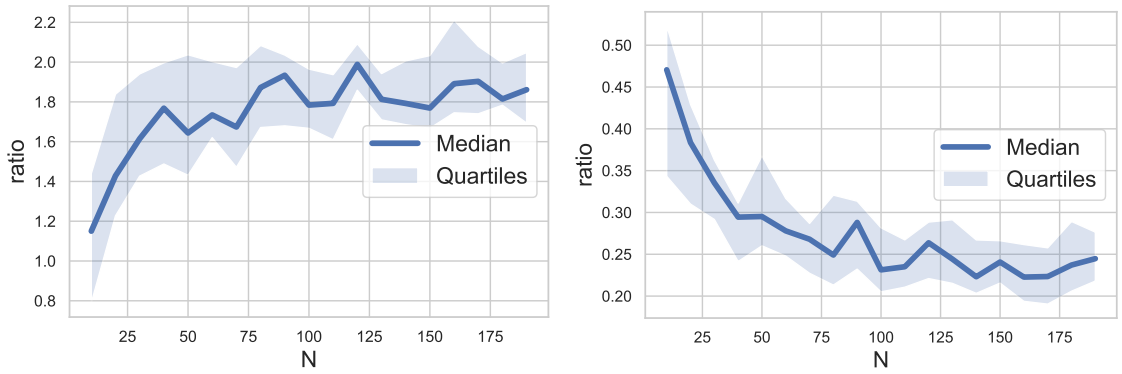

Figure 3: Validation of Theorem 4.1: The ratios $\frac{s_{\max}}{s_{\min}} : \frac{\lambda_1}{\lambda_N}$ for the polynomial spectrum (top) and $\frac{s_{\max}}{s_{\min}} : \frac{N\lambda_1}{\lambda_N}$ for the exponential spectrum (bottom) are asymptotically constant.

we randomly set the true coefficient $\boldsymbol{\gamma}^* \sim \mathcal{N}(0, \mathbf{I}_M)$ and let $y = (\boldsymbol{\gamma}^*)^\top \boldsymbol{\phi} + \epsilon$ 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 Theorems 4.2 and 4.3: the learning curve for polynomial decay is asymptotically bounded by constants; while that for exponential decay increases as $N \to \infty$.

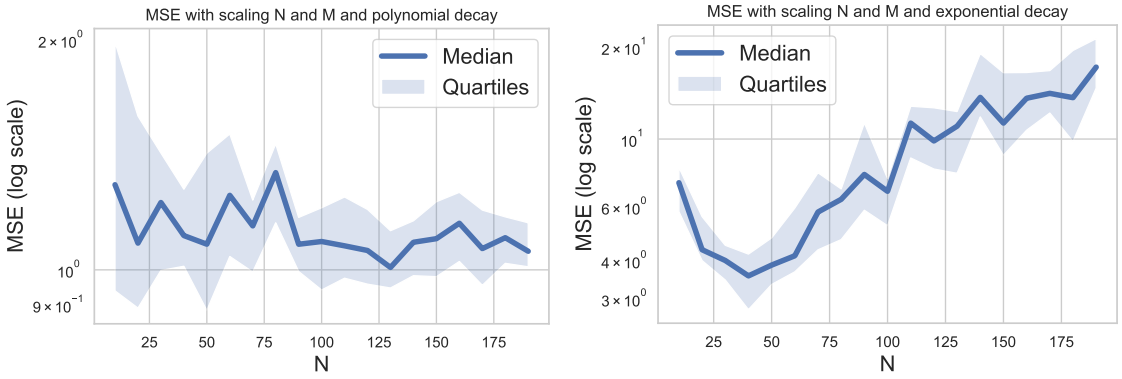

Figure 4: Validation of Theorems 4.2 and 4.3: Learning curves for spectra with polynomial (top) and exponential (bottom) decays.

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. Under the Exponential Decay Assumption 3.4, 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. Under the Polynomial Decay Assumption 3.4, 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 logarithmic-linear decay $\lambda_k = \Theta(\frac{1}{k \log^2 k})$ in Barzilai and Shamir [2023], the effective rank $\rho_l = \Omega(\log l)$, hence the high-frequency part is heavily regularized and benign overfitting would occur. This intuition supports our conjecture.

6.2 Beyond Independent Features

Assumption 3.3 is employed in proving the lower bound of $s_{\min}$ with exponential spectral decay in Lemma B.5 and proving the lower bound of the test error in Theorem C.5. It is natural to assume that ψ_k's are independent as it represents the worst-case scenario for estimation: since $\psi_2, \psi_3, ...$ contain no information about ψ_1, one needs to argue for some probabilistic bound for each k independently and take the union bound at the end. It is also theoretically convenient to assume independence for decoupling cross terms. This will also be of theoretical interest to remove Assumption 3.3 for obtaining lower bounds in future research.

6.3 Limitations

Our method works currently only on kernels with sub-Gaussian features. As mentioned in the introduction, we notice that the concurrent work Barzilai and Shamir [2023] works on the same problem and their statement is valid for a wider class of kernels. But our work is somewhat complementary to theirs, as our analysis can, for instance, explain overfitting for the exponential decay case (while their bounds are vacuous in that case).

6.4 Future Research

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

1. We hypothesize that different types of overfitting only depend on the spectrum. See Subsection 4.3.
2. One can attempt to remove Assumption 3.3 concerning the lower bounds of the test error and $s_{\min}(\mathbf{K})$ with exponential spectral decay.
3. Controlling the condition number of the kernel matrix can be of individual interest in terms of optimization.

References

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems, 31, 2018.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In International Conference on Machine Learning, pages 322–332. PMLR, 2019.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In International Conference on Machine Learning, pages 1024–1034. PMLR, 2020.

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

Song Mei and Andrea Montanari. 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.

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

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

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

Peter L. Bartlett and Shahar Mendelson. 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.

Neil Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. Annual Conference on Neural Information Processing Systems, 2022.

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

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

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

James B Simon, Madeline Dickens, Dhruva Karkada, and Michael R DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. arXiv preprint arXiv:2110.03922, 2021.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. Applied and Computational Harmonic Analysis, 2021.

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

Tin Sum Cheng, Aurelien Lucchi, Ivan Dokmanić, Anastasis Kratsios, and David Belius. A theoretical analysis of the test error of finite-rank kernel ridge regression. Annual Conference on Neural Information Processing Systems, 2023.

Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions. arXiv preprint arXiv:2312.15995, 2023.

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

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. Bernoulli, pages 110–133, 2017.

Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. arXiv preprint arXiv:2108.08198, 2021.

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

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

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. The Annals of statistics, 29(2):295–327, 2001.

Mark Rudelson and Roman Vershynin. 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.

Appendix

A Proof for Gaussian case

For simplicity, for subsection A, we first prove our result with the following assumption instead of Assumption 3.2:

Assumption A.1 (Gaussian Design)**.** *We replace the entries in $\mathbf{\Psi}$ in the kernel matrix $\mathbf{K} = \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi}$ by i.i.d. Gaussian $\mathcal{N}(0,1)$.*

Essentially, the task is just 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.

In section B, we will extend our proof to sub-Gaussian case in a similar flavor.

A.1 Largest Singular Value

The approximation of the largest singular value $s_{\max}(\mathbf{K})$ is well-studied, but for the sake of completeness, we will prove the statement here fundamentally.

Lemma A.2 (Bound on largest singular value)**.** *Suppose Assumption A.1 holds. Suppose either decay assumptions 3.4 or 3.5 holds, there exists some constants $c_1, c_2 > 0$ such that, with a probability at least $1 - 3\delta$, one has*

$$N\lambda_1 \left(1 - \sqrt{\frac{8}{N}\log\frac{2}{\delta}}\right) \le s_{\max}(\mathbf{K}) \le N\lambda_1 \left(1 + \sqrt{\frac{c_1}{N}\log\frac{N}{\delta^2}}\right).$$

for $N > c_2$ large enough.

Proof. We first bound $s_{\max}$ from below. By definition of $s_{\max}$, take $\mathbf{x} = \boldsymbol{\psi}_1 / \|\boldsymbol{\psi}_1\|_2$:

$$s_{\max}(\mathbf{K}) = \sup_{\mathbf{x}\in\mathbb{S}^{N-1}} \sum_{k=1}^{M} \lambda_k (\boldsymbol{\psi}_k^\top \mathbf{x})^2 \ge \sup_{\mathbf{x}\in\mathbb{S}^{N-1}} (\lambda_1 \boldsymbol{\psi}_1^\top \mathbf{x})^2 \ge \lambda_1 (\boldsymbol{\psi}_1^\top \boldsymbol{\psi}_1 / \|\boldsymbol{\psi}_1\|_2)^2 = \lambda_1 \|\boldsymbol{\psi}_1\|_2^2 .$$

Note that the random variable $\|\boldsymbol{\psi}_1\|_2^2 \sim \chi^2(N)$. Indeed, we can use Lemma C.9 to obtain a sharp bound: with probability at least $1 - \delta$,

$$s_{\max}(\mathbf{K}) \ge N\lambda_1 \frac{\|\boldsymbol{\psi}_1\|_2^2}{N} \ge N\lambda_1 \left(1 - \sqrt{\frac{8}{N}\log\frac{2}{\delta}}\right).$$

Now we bound $s_{\max}(\mathbf{K})$ from above. Consider the upper bound by triangle inequality:

$$s_{\max}(\mathbf{K}) = \left\| \sum_{k=1}^{M} \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \right\|_{\text{op}} \le \left\| \sum_{k=1}^{\lfloor \log N \rfloor} \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \right\|_{\text{op}} + \left\| \sum_{k=\lfloor \log N \rfloor + 1}^{M} \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \right\|_{\text{op}} . \tag{6}$$

The first term in line (6) is

$$\begin{aligned}
\left\| \sum_{k=1}^{\lfloor \log N \rfloor} \lambda_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \right\|_{\text{op}} &= \sup_{\mathbf{x}\in\mathbb{S}^{N-1}} \sum_{k=1}^{\lfloor \log N \rfloor} \lambda_k (\boldsymbol{\psi}_k^\top \mathbf{x})^2 \\
&\le \sup_{\mathbf{x}\in\mathbb{S}^{N-1}} \sum_{k=1}^{\lfloor \log N \rfloor} \lambda_1 (\boldsymbol{\psi}_k^\top \mathbf{x})^2 \\
&= N\lambda_1 \left\| \frac{1}{N} \sum_{k=1}^{\lfloor \log N \rfloor} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \right\|_{\text{op}} ,
\end{aligned}$$

where the last line can be controlled by Theorem C.2: with probability at least $1 - \delta$,

$$\left\| \frac{1}{N} \sum_{k=1}^{\lfloor \log N \rfloor} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top \right\|_{\text{op}} \le \left(1 + \sqrt{\frac{1}{N}\log\frac{N}{\delta^2}}\right)^2$$

By Lemma C.9 and union bound, with probability

$$\mathbb{P}\left\{\frac{\|\boldsymbol{\psi_k}\|_2^2}{N}-1<\sqrt{\frac{8}{N}\log\frac{2\theta N}{\delta}}\ :\ \forall k=1,...,M\right\}\geq 1-\sum_{k=1}^{M}\frac{\delta}{M}\geq 1-\delta.$$

When this happens, we have the second term in line (6) upper bounded by:

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{k=\lfloor\log N\rfloor+1}^{M}\lambda_k\boldsymbol{\psi}_k\boldsymbol{\psi}_k^\top\right\|_{\text{op}}\\
&\leq\sum_{k=\lfloor\log N\rfloor+1}^{M}\lambda_k\left(1+\sqrt{\frac{8}{N}\log\frac{2\theta N}{\delta}}\right)\\
&\leq\left(1+\sqrt{\frac{8}{N}\log\frac{2\theta N}{\delta}}\right)\sum_{k=\lfloor\log N\rfloor+1}^{M}\lambda_k\\
&\leq\begin{cases}\frac{c_2}{N}\left(1+\sqrt{\frac{8}{N}\log\frac{2\theta N}{\delta}}\right) & \text{, if the Exponential Decay Assumption 3.4 holds,}\\ \frac{c_2}{(\log N)^{a-1}}\left(1+\sqrt{\frac{8}{N}\log\frac{2\theta N}{\delta}}\right) & \text{, if the polynomial Decay Assumption 3.5 holds,}\end{cases}
\end{aligned}$$

which is $\ll N\lambda_1\sqrt{\frac{1}{N}\log\frac{N}{\delta^2}}$ for N large enough. □

A.2 Smallest Singular Value

For the smallest singular value, we divide the cases into different decays. First, we suppose the Exponential Decay Assumption 3.4 holds.

Lemma A.3 (Lower bound of $s_{\min}$ for exponential spectral decay)**.** *Suppose Assumption A.1 and 3.4 hold. Then there exists some constants $c_1>0$ such that, with a probability of at least $1-\delta$, we have*

$$s_{\min}(\mathbf{K})\geq c_1\delta^2\lambda_N.$$

Proof. By Lemma C.7 and C.8, for any $k\leq N$ and $t\in(0,\infty)$,

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

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

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

When the event E happens, we have

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

by Lemma C.7, 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

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

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 obtain the claim. □

For the polynomial decay, we first observe a simple fact:

Lemma A.4. *If we write $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 \in \mathbb{R}^{M\times M}$ into a sum of two matrices orthogonal to each other: $\mathbf{A}_1^\top \mathbf{A_2} = 0$, then we have*

$$\inf_{v\in\mathbb{S}^{N-1}} \|(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{\Psi} v\|_2^2 \geq \max\left\{ \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_1\mathbf{\Psi} v\|_2^2, \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_2\mathbf{\Psi} v\|_2^2 \right\}.$$

Proof.

$$\begin{aligned}
\inf_{v\in\mathbb{S}^{N-1}} \|(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{\Psi} v\|_2^2 &= \inf_{v\in\mathbb{S}^{N-1}} \left\{ \|\mathbf{A}_1\mathbf{\Psi} v\|_2^2 + \|\mathbf{A}_2\mathbf{\Psi} v\|_2^2 + v^\top \mathbf{\Psi}^\top \mathbf{A}_1\mathbf{A}_2\mathbf{\Psi} v \right\} \\
&\geq \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_1\mathbf{\Psi} v\|_2^2 + \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_2\mathbf{\Psi} v\|_2^2 + \inf_{v\in\mathbb{S}^{N-1}} v^\top \mathbf{\Psi}^\top \mathbf{A}_1^\top \mathbf{A}_2\mathbf{\Psi} v \\
&= \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_1\mathbf{\Psi} v\|_2^2 + \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_2\mathbf{\Psi} v\|_2^2 \\
&\geq \max\left\{ \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_1\mathbf{\Psi} v\|_2^2, \inf_{v\in\mathbb{S}^{N-1}} \|\mathbf{A}_2\mathbf{\Psi} v\|_2^2 \right\}.
\end{aligned}$$

□

Using the above lemma, we have:

Lemma A.5 (Lower bound on smallest singular value for the polynomial decay)**.** *Suppose Assumptions A.1 and 3.5 hold. Suppose Assumption 3.1 holds with the constant $\theta = \lceil 1/\delta_0 \rceil$ for the δ_0 in Proposition C.1. Then where there exists some constants $c_1, c_2 > 0$ such that with a probability at least $1 - e^{-c_2 N}$:*

$$s_{\min}(\mathbf{K}) \geq c_1 N \lambda_N. \tag{7}$$

Proof. We write $\mathbf{\Lambda}^{1/2} = \begin{pmatrix} \mathbf{\Lambda}_{\leq N}^{1/2} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{\Lambda}_{>N}^{1/2} \end{pmatrix}$, we can begin bounding the smallest singular value of $\mathbf{K}$ using:

$$s_{\min}(\mathbf{K}) = \inf_{v\in\mathbb{S}^{N-1}} \left\| \mathbf{\Lambda}^{1/2}\mathbf{\Psi} v \right\|_2^2 \geq \inf_{v\in\mathbb{S}^{N-1}} \left\| \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{\Lambda}_{>N}^{1/2} \end{pmatrix} \mathbf{\Psi} v \right\|_2^2 = \inf_{v\in\mathbb{S}^{N-1}} \left\| \mathbf{\Lambda}_{>N}^{1/2} \bar{\mathbf{\Psi}} v \right\|_2^2$$

where $\bar{\Psi} \in \mathbb{R}^{(M-N)\times N}$ is the submatrix of Ψ corresponding to the last $(M - N)$ rows. Under the polynomial decay assumption 3.5, the ratio between the largest and the smallest eigenvalues in $\mathbf{\Lambda}_{>M}$ is bounded by some constant: $\frac{\lambda_{N+1}}{\lambda_M} \leq \frac{\overline{r}N^{-a}}{\underline{r}M^{-a}} = \frac{\overline{r}}{\underline{r}}\frac{(\theta N)^a}{N^a} = \frac{\overline{r}}{\underline{r}}\theta^a$, hence the above square norm is closed to some scaled random standard Gaussian matrix: by Proposition C.1, there exists some absolute constant $C_1, C_2 > 0$ such that, with a probability of at most $e^{-C_2 N}$,

$$\inf_{v\in\mathbb{S}^{N-1}} \left\| \mathbf{\Lambda}_{>N}^{1/2} \bar{\mathbf{\Psi}} v \right\|_2^2 \leq \lambda_M \frac{\lambda_{N+1}}{\lambda_M} \inf_{v\in\mathbb{S}^{N-1}} \left\| \bar{\mathbf{\Psi}} v \right\|_2^2 \leq \lambda_M \cdot \frac{\overline{r}}{\underline{r}}\theta^a \cdot \inf_{v\in\mathbb{S}^{N-1}} \left\| \bar{\mathbf{\Psi}} v \right\|_2^2 \leq \lambda_M \cdot \frac{\overline{r}}{\underline{r}}\theta^a \cdot C_1(M - N),$$

where we set $n = M - N, k = N$ and $\theta = \lceil 1/\delta_0 \rceil$ (see the proposition for the definitions), and the constant C_1 is as stated in Proposition C.1. Then we have

$$\lambda_M \cdot \frac{\overline{r}}{\underline{r}}\theta^a \cdot C_1(M - N) = \lambda_{\theta N} \cdot \frac{\overline{r}}{\underline{r}}\theta^a \cdot C_1(\theta - 1)N \leq C_1 \frac{\overline{r}}{\underline{r}}(\theta - 1)\lambda_N N$$

Set $c_1 = C_1 \frac{\overline{r}}{\underline{r}}(\theta - 1)$, $c_2 = C_2$ as in Proposition C.1 and we are done. □

Note that discarding the first N rows is not pessimistic as it seems: the lower bound is sharp as it matches the upper bound:

Lemma A.6 (Upper bound on smallest singular value)**.** *With notation above, we have, with a probability of at least* $1-\delta$*:*

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

In particular, there exists constants $c_1, c_2 > 0$ *such that, with a probability of at least* $1-1/N$*:*

- *given that Assumption 3.4 (exponential decay) holds, we have* $s_{\min}(\mathbf{K}) \leq c_1\lambda_N$*,*
- *given that Assumption 3.5 (polynomial decay) holds, we have* $s_{\min}(\mathbf{K}) \leq c_2 N\lambda_N$*.*

Proof. 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 C.9, 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

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

If the Exponential Decay Assumption 3.4 holds, we have

$$\sum_{k=N}^{M}\lambda_k \leq c_1\lambda_N;$$

if the polynomial Decay Assumption 3.5 holds, we have

$$\sum_{k=N}^{M}\lambda_k \leq c_2 N\lambda_N.$$

By setting $\delta = \frac{1}{N}$, the factor $\left(1+\sqrt{8\log\frac{2(\theta-1)N}{\delta}}\right)$ becomes constant in line (8). □

A.3 Condition Number and Test Error

Theorem A.7 (Bound on condition number of kernel matrix)**.** *Suppose Assumptions 3.1 and A.1 hold.*

1. *If, furthermore, the exponential decay assumption 3.4 holds, then there exists some constants* c_1, c_2, c_3 *such that, with probability at least* $1-\delta-c_1/N$*, the condition number of the kernel matrix* $\mathbf{K}$ *is*

$$c_2\frac{\lambda_1}{\lambda_N}N \leq \frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq \frac{c_3}{\delta^2}\frac{\lambda_1}{\lambda_N}N.$$

2. *If, furthermore, the polynomial decay assumption 3.5 holds,then there exists some constants* c_1, c_2, c_3, c_4 *such that with probability at least* $1-c_1/N-e^{-c_2 N}$*, the condition number of the kernel matrix* $\mathbf{K}$ *is*

$$c_3\frac{\lambda_1}{\lambda_N} \leq \frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq c_4\frac{\lambda_1}{\lambda_N}.$$

Proof. For exponential decay, from Lemma A.2, set $\delta = 1/N$, there exists constants c_1, c_2, c_3 such that with probability of at least $1 - c_2/N$, $c_3 N\lambda_1 \leq s_{\max} \leq c_1 N\lambda_1$; from Lemma A.3, there exists a constant c_4 with probability of at least $1 - \delta$, we have $s_{\min} \geq c_4\delta^2\lambda_N$. By the union bound, we can bound the condition number of the kernel matrix $\mathbf{K}$ from above: with probability at least $1 - \delta - c_2/N$

$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq \frac{c_1 N\lambda_1}{c_4\delta^2\lambda_N}.$$

From Lemma A.6, there exists a constant c_5 such that with probability of at least $1 - 1/N$,e $s_{\min} \leq c_5\lambda_N$, hence By the union bound, with probability of at least $1 - (c_2 + 1)/N$,

$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \geq \frac{c_3 N\lambda_1}{c_5\lambda_N}.$$

Combining the above results and renaming the constants, there exist constants c_1, c_2, c_3 such that with probability at least $1 - \delta - c_1/N$,

$$c_2\frac{N\lambda_1}{\lambda_N} \leq \frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq \frac{c_3}{\delta^2}\frac{N\lambda_1}{\lambda_N}.$$

For polynomial decay, from Lemma A.2, set $\delta = 1/N$, there exists constants c_1, c_2, c_3 such that with probability of at least $1 - c_2/N$, $c_3 N\lambda_1 \leq s_{\max} \leq c_1 N\lambda_1$; from Lemma A.5, there exists a constant c_4, c_5 with probability of at least $1 - e^{-c_4 N}$, we have $s_{\min} \geq c_5 N\lambda_N$. By the union bound, with probability at least $1 - c_2/N - e^{-c_4 N}$,

$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq \frac{c_1 N\lambda_1}{c_5 N\lambda_N}$$

From Lemma A.6, with probability at least $1 - 1/N$, we have $s_{\min} \leq c_6\lambda_N N$ and hence, with probability at least $1 - (c_2 + 1)/N$,

$$\frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \geq \frac{c_3 N\lambda_1}{c_6 N\lambda_N}$$

Combining the above results and renaming the constants, there exists constants c_1, c_2, c_3, c_4 such that with probability at least $1 - c_1/N - e^{-c_2 N}$,

$$c_3\frac{\lambda_1}{\lambda_N} \leq \frac{s_{\max}(\mathbf{K})}{s_{\min}(\mathbf{K})} \leq c_4\frac{\lambda_1}{\lambda_N}.$$

□

Given that we have bounded the condition number of the kernel matrix, we can now bound the test error.

Theorem A.8. *Suppose Assumptions 3.1, 3.5, 3.6 and A.1 hold. Then there exists some constants $c_1, c_2, c_3, c_4, c_5, c_6$ such that with probability at least $1 - c_1 e^{-N/c_1} - e^{-c_2 N}$, we have*

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

Proof. By Theorem C.4, take $l = \lfloor N/c_1\rfloor$. Since $\lambda = 0$, so $s_N(\mathbf{A}_l^{-1}) = s_1(\mathbf{A}_l)^{-1}$ and $s_1(\mathbf{A}_l^{-1}) = s_N(\mathbf{A}_l)^{-1}$. Hence $\frac{s_1(\mathbf{A}_l^{-1})^2}{s_N(\mathbf{A}_l^{-1})^2} = \frac{s_{\max}(\mathbf{A}_l)^2}{s_{\min}(\mathbf{A}_l)^2}$. Since $\mathbf{A}_l$ is just another kernel matrix with rank $(M - l)$, by modifying Theorem A.7 w.r.t. the right-shifted polynomial decay, with high probability, we have

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

and

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

then we can bound the bias term using Theorem C.4:

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

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 $l=\lfloor N/c_1\rfloor$.

□

Theorem A.9. *Suppose Assumptions 3.1, 3.5, 3.6 hold. Suppose Assumption A.1 holds or both Assumptions 3.2 and 3.3 hold. Then there exists constants $C>0$ such that with probability at least $1-Ce^{-N/C}$,*

$$\mathcal{V}=\Omega(1).$$

Proof. 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 C.5 would hold for some $l=N/c_1$ where $c_1>1$. Then we apply Theorem C.5 for polynomial decay: there exists constants C,C', with a probability at least $1-Ce^{-N/C}$, we have

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

□

If the kernel has exponential decay (Assumption 3.4) instead, the upper bound in Theorem C.4 will become vacuous. Instead, this upper bound is sharp in the sense that one can show the kernel with exponential decay suffers from catastrophic overfitting:

Theorem A.10. *Suppose Assumptions 3.1, 3.4, 3.6 hold. Suppose Assumption A.1 holds or both Assumptions 3.2 and 3.3 hold. Then there exists some constants c such that with probability at least $1-Ce^{-N/C}$, we have*

$$\mathcal{V}=\Omega(N).$$

Proof. 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 C.5 would hold for some $l<N$. Then we apply Theorem C.5 for exponential decay: there exists a constant C,C', with a probability at least $1-Ce^{-N/C}$, we have

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

□

B Proof for Sub-Gaussian case

In this section, we will prove the Theorem 4.1 on the condition number of $\mathbf{K}$ under Assumption 3.2. Note that the remaining arguments in Theorem 4.2 and 4.3 follow.

Writing the kernel matrix as $\mathbf{K} = \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi}$, where the rows $\boldsymbol{\psi}_k$ are isotropic sub-Gaussian random vectors. Then the control on the largest singular value directly follows from Theorem C.3:

Lemma B.1 (Upper bound on largest singular value)**.** *There exists constants c_1, c_2 depending only on the sub-Gaussian norm of $\boldsymbol{\psi}_k$, such that, with probability at least $1 - 2e^{-c_1 N}$, we have*

$$s_{\max}(\mathbf{K}) \leq c_2 N \lambda_1.$$

Proof. Set $\mathbf{A} = \mathbf{\Lambda}^{1/2} \mathbf{\Psi}$ and $t = \sqrt{N}$ in Theorem C.3, then with a probability at least $1 - 2e^{-C_3 N}$, we have

$$\left\| \frac{1}{N} \mathbf{K} - \mathbf{\Lambda} \right\|_{\text{op}} \leq \max\{\delta, \delta^2\} \lambda_1$$

where $\delta = C_4 \sqrt{\frac{N}{M}} + \frac{\sqrt{N}}{\sqrt{N}} = \frac{C_4}{\sqrt{\theta}} + 1$ is a constant, since $\theta = \frac{M}{N}$ is a constant by Assumption 3.1. Set $c_1 = C_3$ and $c_2 = \max\{\delta, \delta^2\}$ to conclude the proof. □

The lower bound on $s_{\max}(\mathbf{K})$ follows a similar argument as in the Gaussian case in Lemma A.2, replacing the concentration of chi-square by that of sub-exponential variables:

Lemma B.2 (Lower Bound on Largest Singular Value)**.** *There exists constants $c_1, c_2 > 0$, such that, with probability at least $1 - 2e^{-c_1 N}$, we have*

$$s_{\max}(\mathbf{K}) \geq c_2 N \lambda_1.$$

Proof. By definition of $s_{\max}$, take $\mathbf{x} = \boldsymbol{\psi}_1 / \|\boldsymbol{\psi}_1\|_2$:

$$s_{\max}(\mathbf{K}) = \sup_{\mathbf{x} \in \mathbb{S}^{N-1}} \sum_{k=1}^{M} \lambda_k (\boldsymbol{\psi}_k^\top \mathbf{x})^2 \geq \sup_{\mathbf{x} \in \mathbb{S}^{N-1}} (\lambda_1 \boldsymbol{\psi}_1^\top \mathbf{x})^2 \geq \lambda_1 (\boldsymbol{\psi}_1^\top \boldsymbol{\psi}_1 / \|\boldsymbol{\psi}_1\|_2)^2 = \lambda_1 \|\boldsymbol{\psi}_1\|_2^2 .$$

Note that the random variable $\|\boldsymbol{\psi}_1\|_2^2$ is sub-exponential with mean N and sub-exponential norm $\leq \frac{N}{\sqrt{2}}$ as $\boldsymbol{\psi}_1$ is a sub-Gaussian random vector. By Lemma C.9, set B to be the sub-exponential norm of $\psi_k(x)^2$ and $\delta = \frac{1}{2}$, with probability at least $1 - 2e^{-C_5 \min\{\frac{1}{4B^2}, \frac{1}{2B}\} N}$,

$$s_{\max}(\mathbf{K}) \geq N \lambda_1 \frac{\|\boldsymbol{\psi}_1\|_2^2}{N} \geq N \lambda_1 \left(1 - \frac{1}{2}\right) = \frac{1}{2} N \lambda_1.$$

Set $c_1 = C_5 \min\{\frac{1}{4B^2}, \frac{1}{2B}\}$ and $c_2 = \frac{1}{2}$ to conclude the proof. □

Lemma B.3 (Upper bound on smallest singular value)**.** *There exists constants $C_5 > 0$ such that, with a probability of at least $1 - \delta$:*

$$s_{\min}(\mathbf{K}) \leq (1 + t) \sum_{k=N}^{M} \lambda_k,$$

where $t = \min\left\{ \sqrt{B^2 C_5^{-1} \log \frac{2(\theta-1)N}{\delta}}, B C_5^{-1} \log \frac{2(\theta-1)N}{\delta} \right\}$, B is the maximum sub-exponential norm of of the centered variables $\psi_k(x) - 1$ and C_5 is an absolute constant in Lemma C.9. In particular, there exists constants $c_1, c_2 > 0$ such that, with a probability of at least $1 - 1/N$:

- *given that Assumption 3.4 (exponential decay) holds, we have $s_{\min}(\mathbf{K}) \leq c_1 \lambda_N$,*
- *given that Assumption 3.5 (polynomial decay) holds, we have $s_{\min}(\mathbf{K}) \leq c_2 N \lambda_N$.*

Proof. 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 $(\boldsymbol{\psi}_k^\top v_0)^2$ is sub-exponential, By Lemma C.9, hence we have

$$\mathbb{P}\left\{\left|(\boldsymbol{\psi}_k^\top v_0)^2 - 1\right| \geq t\right\} \leq 2e^{-C_5 \min\{\frac{t^2}{B^2}, \frac{t}{B}\}}.$$

Set $t = \min\left\{\sqrt{B^2 C_5^{-1} \log\frac{2(\theta-1)N}{\delta}}, BC_5^{-1}\log\frac{2(\theta-1)N}{\delta}\right\}$ and by 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

$$s_{\min}(\mathbf{K}) \leq \sum_{k=N}^{M} \lambda_k(1+t) = (1+t)\sum_{k=N}^{M}\lambda_k. \tag{9}$$

If the Exponential Decay Assumption 3.4 holds, we have

$$\sum_{k=N}^{M}\lambda_k \leq c_1\lambda_N;$$

if the polynomial Decay Assumption 3.5 holds, we have

$$\sum_{k=N}^{M}\lambda_k \leq c_2 N\lambda_N.$$

By setting $\delta = \frac{1}{N}$, the factor $(1+t)$ in line (9) becomes constant. □

Lemma B.4 (Lower bound of smallest singular value for polynomial spectrum)**.** *Suppose Assumption 3.5 holds. There exists constants $c_1, c_2 > 0$ such that, with a probability of at least $1 - 2e^{-c_1 N}$:*

$$s_{\min}(\mathbf{K}) \geq c_2\lambda_N N. \tag{10}$$

Proof. We write $\boldsymbol{\Lambda}^{1/2} = \begin{pmatrix}\boldsymbol{\Lambda}_{\leq N}^{1/2} & 0\\ 0 & 0\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & \boldsymbol{\Lambda}_{>N}^{1/2}\end{pmatrix}$, we can begin bounding the smallest singular value of $\mathbf{K}$ using:

$$s_{\min}(\mathbf{K}) = \inf_{v\in\mathbb{S}^{N-1}}\left\|\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Psi}v\right\|_2^2 \geq \inf_{v\in\mathbb{S}^{N-1}}\left\|\begin{pmatrix}0 & 0\\ 0 & \boldsymbol{\Lambda}_{>N}^{1/2}\end{pmatrix}\boldsymbol{\Psi}v\right\|_2^2 = \inf_{v\in\mathbb{S}^{N-1}}\left\|\boldsymbol{\Lambda}_{>N}^{1/2}\bar{\boldsymbol{\Psi}}v\right\|_2^2$$

where $\bar{\Psi}\in\mathbb{R}^{(M-N)\times N}$ is the submatrix of Ψ corresponding to the last $(M-N)$ rows. Under the polynomial decay assumption 3.5, the ratio between the largest and the smallest eigenvalues in $\boldsymbol{\Lambda}_{>M}$ is bounded by some constant: $\frac{\lambda_{N+1}}{\lambda_M} \leq \frac{\overline{r}N^{-a}}{\underline{r}M^{-a}} = \frac{\overline{r}}{\underline{r}}\frac{(\theta N)^a}{N^a} = \frac{\overline{r}}{\underline{r}}\theta^a$, hence the above square norm is closed to some scaled random matrix with independent sub-Gaussian column: by Theorem C.3, there exists some constants $C_3, C_4 > 0$ such that, with a probability of at most $2e^{-C_3 N}$,

$$\inf_{v\in\mathbb{S}^{N-1}}\left\|\boldsymbol{\Lambda}_{>N}^{1/2}\bar{\boldsymbol{\Psi}}v\right\|_2^2 \leq \lambda_M\frac{\lambda_{N+1}}{\lambda_M}\inf_{v\in\mathbb{S}^{N-1}}\left\|\bar{\boldsymbol{\Psi}}v\right\|_2^2 \leq \lambda_M\cdot\frac{\overline{r}}{\underline{r}}\theta^a\cdot\inf_{v\in\mathbb{S}^{N-1}}\left\|\bar{\boldsymbol{\Psi}}v\right\|_2^2 \leq \lambda_M\cdot\frac{\overline{r}}{\underline{r}}\theta^a\cdot\left(\sqrt{M-N}-\sqrt{C_4N}-\sqrt{N}\right)^2,$$

where we set $t = \sqrt{N}$, and the constant C_3, C_4 is as stated in Theorem C.3. Then the term $\left(\sqrt{M-N}-\sqrt{C_4N}-\sqrt{N}\right)^2$ is smaller than or equal to c_2N for some constant $c_2 > 0$. Hence we have

$$\lambda_M\cdot\frac{\overline{r}}{\underline{r}}\theta^a\cdot\left(\sqrt{M-N}-\sqrt{C_4N}-\sqrt{N}\right)^2 = \lambda_{\theta N}\cdot\frac{\overline{r}}{\underline{r}}\theta^a\cdot\left(\sqrt{M-N}-\sqrt{C_4N}-\sqrt{N}\right)^2 \leq c_2\lambda_N N.$$

Set $c_1 = C_3$ to conclude the proof. □

or exponential decay, we have to assume independent sub-Gaussian features:

Lemma B.5 (Lower bound of smallest singular value for exponential spectrum)**.** *Suppose Assumption 3.3 holds and the eigenfunctions ψ_k's are centered and let $B \overset{\text{def.}}{=} \max_k \mathbb{E}_x\left[\psi_k(x)^4\right]$ the maximum 4th moment of the eigenfunction ψ_k. Then there exists some constants $c_1, c_2 > 0$ such that, with probability at least $c_1 B^{-N}$,*

$$s_{\min}(\mathbf{K}) \geq c_2 \lambda_N.$$

Proof. First, we assume that each feature ψ_k is centered. Then for each $k = 1, ..., N$, write $\mathbf{N}_k = (n_i)_{i=1}^N \in \mathbb{R}^N$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}\left[(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2\right] &= \mathbb{E}_{\mathbf{X}}\left[\sum_{i,j=1}^N n_i n_j \psi_k(x_i)\psi_k(x_j)\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\sum_{i=1}^N n_i^2 \psi_k(x_i)^2\right] + \mathbb{E}_{\mathbf{X}}\left[\sum_{i\neq j} n_i n_j \psi_k(x_i)\psi_k(x_j)\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\sum_{i=1}^N n_i^2 \psi_k(x_i)^2\right] + \sum_{i\neq j} \mathbb{E}\left[n_i n_j \psi_k(x_i)\mathbb{E}_{x_j}\left[\psi_k(x_j)\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\sum_{i=1}^N n_i^2 \psi_k(x_i)^2\right] \\
&= \sum_{i=1}^N \mathbb{E}_{x_j : j\neq i}\left[n_i^2\right] \mathbb{E}_{x_i}\left[\psi_k(x_i)^2\right] \\
&= \mathbb{E}_{x_j : j\neq i}\left[\sum_{i=1}^N n_i^2\right] \\
&= 1.
\end{aligned}$$

Let $B \overset{\text{def.}}{=} \max_k \mathbb{E}_x\left[\psi_k(x)^4\right]$ the 4th moment of the eigenfunction ψ_k. Then we can compute:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}\left[(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^4\right] &= \sum_{i,j} \mathbb{E}_{\mathbf{X}}\left[n_i^2 n_j^2 \psi_k(x_i)^2 \psi_k(x_j)^2\right] \\
&= \sum_{i,j} \mathbb{E}_{\mathbf{X}}\left[n_i^2 n_j^2\right] \mathbb{E}_{\mathbf{X}}\left[\psi_k(x_i)^2 \psi_k(x_j)^2\right] \\
&\leq \max_i \left\{\mathbb{E}\left[\psi_k(x_i)^2\right]^2, \mathbb{E}\left[\psi_k(x_i)^4\right]\right\} \sum_{i,j} \mathbb{E}_{\mathbf{X}}\left[n_i^2 n_j^2\right] \\
&= B\mathbb{E}_{\mathbf{X}}\left[\|\mathbf{N}_k \mathbf{N}_k^\top\|_F^2\right] \\
&= B\mathbb{E}_{\mathbf{X}}\left[\mathrm{Tr}[(\mathbf{N}_k \mathbf{N}_k^\top)^2]\right] \\
&= B\mathbb{E}_{\mathbf{X}}\left[\mathrm{Tr}[(\mathbf{N}_k \mathbf{N}_k^\top)]\right] \\
&= B\mathbb{E}_{\mathbf{X}}\left[\mathbf{N}_k^\top \mathbf{N}_k\right] \\
&= B\mathbb{E}_{\mathbf{X}}\left[1\right] = B,
\end{aligned}$$

since $B = \max_k \mathbb{E}_x\left[\psi_k(x)^4\right] \geq \max_k \mathbb{E}_x\left[\psi_k(x)^2\right]^2 = 1$. By Paley-Zygmund inequality,

$$\begin{aligned}
\mathbb{P}\left\{(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2 \geq t\mathbb{E}\left[(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2\right]\right\} &\geq (1-t)^2 \frac{\mathbb{E}\left[(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2\right]^2}{\mathbb{E}\left[(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^4\right]} \\
\mathbb{P}\left\{(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2 \geq t\right\} &\geq (1-t)^2/B.
\end{aligned}$$

Set $t = e^{-\frac{a}{2}(N-k+1)}$ and we have

$$\begin{aligned}\mathbb{P}\left\{\underbrace{(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2 \geq e^{-\frac{a}{2}(N-k+1)} : \forall k = 1, ..., N}_{E}\right\} &\geq \prod_{k=1}^{N}\left(1 - e^{-\frac{a}{2}(N-k+1)}\right)^2 / B \\ &= B^{-N}\left(\prod_{k=1}^{N}\left(1 - e^{-\frac{a}{2}(N-k+1)}\right)\right)^2 \\ &\geq B^{-N}\left(1 - \sum_{k=1}^{N} e^{-\frac{a}{2}(N-k+1)}\right)^2 \\ &= c_1 B^{-N}\end{aligned}$$

for some constant $c_1 > 0$ depending only on a. When this event E happens, we have

$$\sum_{k=1}^{N}\frac{\lambda_N}{\lambda_k}(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2} \leq \sum_{k=1}^{N}\frac{\lambda_N}{\lambda_k}e^{\frac{a}{2}(N-k+1)} \lesssim \sum_{k=1}^{N} e^{-\frac{a}{2}(N-k)}.$$

Hence there exists some constant $c_2 > 0$ such that $\frac{1}{\sum_{k=1}^{N}\frac{\lambda_N}{\lambda_k}(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2}} \geq c_2$. By Lemma C.7, we have

$$s_{\min}(\mathbf{K}) \geq c_2 \lambda_N.$$

□

Remark B.6. *The first observation is that the probability decays exponentially with respect to N when the maximum fourth moment $B > 1$. If additional assumptions are introduced, such as anticoncentration:*

$$\mathbb{P}\left\{(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2 \leq t\right\} \leq c_1 t$$

where $c_1 > 0$ is a constant, one can readily argue as in Lemma A.3 to obtain a much better probability. Specifically, applying an anticoncentration result for $(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2$ to Lemma C.7 suffices. However, for the sake of generality, we retain Lemma B.5 with minimal assumptions and exponential decaying probability.

Remark B.7. *Assumption 3.3 is purely technical. We conducted an experiment using cosine features $\psi_k = \cos(k\cdot)$ (which are dependent on each other) on random points on a circle and empirically computed the term $(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2$. Refer to Figure 5 for the results.*

We end Section B here as the above Lemmata suffice to prove the sub-Gaussian version of Theorem A.7 with respective probability. The remaining two Theorems 4.2 and 4.3 follow in a similar flavor as in the Gaussian case.

C Technical Lemmata

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

Proposition C.1 (Proposition 2.5 in Rudelson and 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 C.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 C.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 $\boldsymbol{\Sigma} \overset{\text{def.}}{=} \mathbb{E}\left[\mathbf{A_i}\mathbf{A_i}^\top\right] \in \mathbb{R}^{n \times n}$. Then there exists constants*

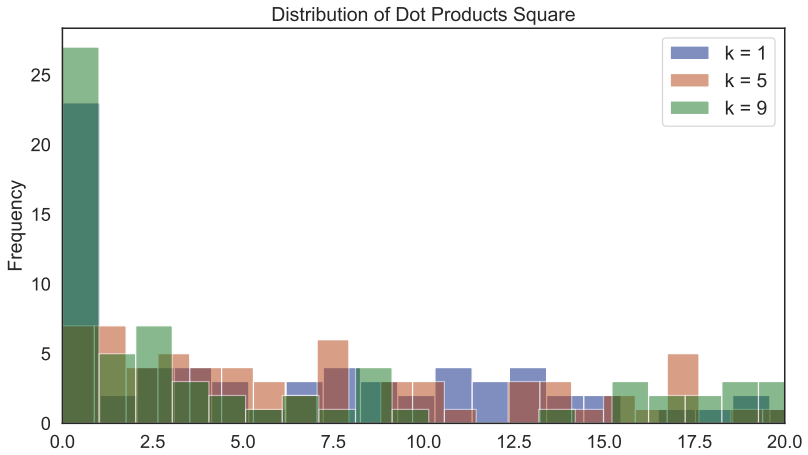

Figure 5: Empirical distribution of the dot-product square $(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2$, where the feature vectors $\boldsymbol{\psi}_k = (\cos(kx_i))_{i=1}^N$ represent cosines evaluated on the training set $\{x_i\}_{i=1}^N$, and $\mathbf{N}_k$ is a unit normal vector of the hyperplane spanned by ψ_l 's for $l \neq k$. We observe a heavy-tailed distribution for each frequency k, indicating the anti-concentration property of $(\boldsymbol{\psi}_k^\top \mathbf{N}_k)^2$ in a dependent-feature setting.

$C_3, C_4 > 0$ (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

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

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 C.4 (Theorem 2.5 in Tsigler and Bartlett [2023])**.** *Suppose Assumption 3.2 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*

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

where $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*_{\leq l} \oplus \boldsymbol{\gamma}^*_{>l}$ 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 C.5 (Lemma 7 and Theorem 10 in Tsigler and Bartlett [2023])**.** *Suppose Assumptions 3.2 and 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 C.6 (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 C.7 (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* $\boldsymbol{\Lambda}_N = \mathrm{diag}(\lambda_k)_{k=1}^N \in \mathbb{R}^{N\times N}$ *and* $\boldsymbol{\Psi}_N = (\boldsymbol{\psi}_k)_{k=1}^N \in \mathbb{R}^{N\times N}$ *and set* $\mathbf{M} = \boldsymbol{\Lambda}_N^{1/2} \boldsymbol{\Psi}_N$ *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 the all rows* $\mathbf{R}_1, ..., \mathbf{R}_n$ *except* $\mathbf{R}_i$*. we have*

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

Proof. Since $s_{\min} \ge s_N(\mathbf{K}_N)$, WLOG: assume $M = N$. Then by Lemma C.6,

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

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

$$s_{\min} \ge s_N(\mathbf{K}_N) \ge \frac{\lambda_N}{\sum_{k=1}^N \frac{\lambda_N}{\lambda_k} (\boldsymbol{\psi}_k^\top \mathbf{N}_k)^{-2}}. \tag{11}$$

□

Lemma C.8 (Anti-Concentration Result For Gaussian Laws)**.** *Let* g *be a standard Gaussian variable, then*

$$\mathbb{P}\{|g| \le t\} \le \frac{2t}{\sqrt{2\pi}}, \ \forall t \ge 0. \tag{12}$$

Lemma C.9 (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* $\delta > 0$*,*

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

where $C_5 > 0$ *is an absolute constant.*

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

Theorem C.10 (Theorem 1.1 in Rudelson and 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* $\delta > 0$*,*

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

In particular, if $N = n$*, we have*

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

with high probability.