Generalization in Kernel Regression Under Realistic Assumptions

Daniel Barzilai [1] **Ohad Shamir** [1]

Abstract

It is by now well-established that modern over-parameterized models seem to elude the bias-variance tradeoff and generalize well despite overfitting noise. Many recent works attempt to analyze this phenomenon in the relatively tractable setting of kernel regression. However, as we argue in detail, most past works on this topic either make unrealistic assumptions, or focus on a narrow problem setup. This work aims to provide a unified theory to upper bound the excess risk of kernel regression for nearly all common and realistic settings. When applied to common kernels, our results imply benign overfitting in high input dimensions, nearly tempered overfitting in fixed dimensions, and explicit convergence rates for regularized regression. As a by-product, we obtain time-dependent bounds for neural networks trained in the kernel regime. Our results rely on new relative perturbation bounds for the eigenvalues of kernel matrices, which may be of independent interest. These reveal a self-regularization phenomenon, whereby a heavy tail in the eigendecomposition of the kernel implicitly leads to good generalization.

1. Introduction

It is by now well-established that various families of highly over-parameterized models tend to generalize well, even when perfectly fitting noisy data (Zhang et al., 2021; Belkin et al., 2019). This phenomenon seemingly contradicts the classical intuition of the bias-variance tradeoff, and motivated a large literature attempting to explain it (Bartlett et al., 2020; Hastie et al., 2022).

In particular, a long series of works attempted to understand this phenomenon in the context of kernel methods (Liang & Rakhlin, 2020; Mei et al., 2022; Xiao & Pennington, 2022; Mallinar et al., 2022). This is due both to their classical importance and their relation to over-parameterized neural networks via the Neural Tangent Kernel (NTK) and Gaussian Process Kernel (GPK, also known as NNGP) (Lee et al., 2017; Jacot et al., 2018). However, there is still a large gap between empirical observations and current theoretical analysis. As we argue in detail in Sec. 2, past works tend to either make unrealistic assumptions (often inspired by the analysis of *linear* regression) that do not hold for common kernels of interest, or are limited to a very narrow problem setup. This is not just a technical limitation, but rather, as we will show, may result in an inaccurate analysis for common kernels in practice. In this paper, we provide simple, sharp, and rigorous upper bounds for the generalization error of kernel regression, which hold under realistic assumptions and can be applied to a wide range of kernels and settings.

Specifically, we demonstrate that many kernels have a built-in *self-regularization* property, meaning that the structure of the kernel provides an implicit form of regularization. This is characterized by novel relative deviation bounds on the eigenvalues of kernel matrices, which may be of independent interest and may be useful in many other settings.

We then apply these tools to analyze the generalization performance of regularized and un-regularized kernel regression. Self-regularization causes the kernel to learn a function that generalizes well, even if it can interpolate the data. As such, we provide upper bounds for the excess risk (and its bias and variance components) regardless of the amount of explicit regularization. Importantly, our mild assumptions allow us to apply these bounds to common kernels, including NTKs (and hence provide insights on generalization in neural networks). Specifically, our main results and insights include the following:

Relative concentration bounds for the eigenvalues of kernel matrices (Thm. 3.1). We derive both upper and lower bounds for the eigenvalues of kernel matrices under very mild assumptions which hold for common kernels. In particular, this highlights a self-regularization phenomenon whereby the eigenvalues of the kernel matrix behave as if one added an explicit regularization term to the training objective.

A general-purpose upper bound for the excess risk in

[1] Weizmann Institute of Science. Correspondence to: Daniel Barzilai <daniel.barzilai@weizmann.ac.il>, Ohad Shamir <ohad.shamir@weizmann.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

kernel regression (Thm. 4.1). The assumptions of this bound are very mild, and it can thus be applied to common kernels in a variety of settings. The bound is sharp without further assumptions, and characterizes both the bias and variance up to universal constants. In particular, no assumption is made on the regularization strength, amount of noise, input dimension, or number of samples.

Benign overfitting in high input dimensions (Thm. 5.1), meaning that the excess risk goes to zero despite the presence of noise and lack of explicit regularization. In such a high dimensional setting, the frequencies that can be learned are limited, thus preventing any harmful overfitting. In particular, our results apply to the NTK, showing benign overfitting (and the corresponding convergence rates) for neural networks in the kernel regime when the input dimension is large.

Nearly Tempered overfitting in fixed input dimensions (Thm. 5.2), meaning that the bias goes to zero, and the variance cannot diverge too quickly. As such, when the amount of noise is relatively small, this implies a good excess risk despite a possibly harmful overfitting of noise. As far as we know, this is the first rigorous upper bound for unregularized kernel regression (i.e., min-norm interpolator) in the fixed dimensional setting for generic kernels.

Learning rates for regularized kernel regression (Thm. 5.3), where we bound the bias and variance as a function of the regularization strength. In particular, through a connection with gradient flow, this gives convergence rates for neural networks trained in the kernel regime.

Overall, we hope that our paper will contribute to the development of a rigorous general theory analyzing overfitting in kernel regression and, more generally, in over-parameterized models, under minimal and realistic assumptions.

2. Preliminaries

Let $\mathcal{X}$ be some input space, μ an associated measure and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel, meaning that it admits a spectral decomposition of the form

$$K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{x}'), \tag{1}$$

where $\lambda_i \geq 0$ are the non-negative eigenvalues (not necessarily ordered), and the eigenfunctions ψ_i form an orthonormal basis in $L^2_\mu(\mathcal{X})$. Let $p \in \mathbb{N} \cup \{\infty\}$ denote the number of non-zero eigenvalues, and w.l.o.g let $\phi(\mathbf{x}) := \left(\sqrt{\lambda_i}\psi_i(\mathbf{x})\right)_{i=1}^{p}$ be the non-zero features (with $\lambda_i > 0$) and $\psi(\mathbf{x}) := (\psi_i(\mathbf{x}))_{i=1}^{p}$. Since $\mathbb{E}_x[\psi(\mathbf{x})\psi(\mathbf{x})^\top] = I$, the features admit a diagonal and invertible (uncentered) covariance operator given by $\Sigma := \mathbb{E}_{\mathbf{x}}\left[\phi(\mathbf{x})\phi(\mathbf{x})^\top\right] = \mathrm{diag}(\lambda_1, \lambda_2, \ldots)$. The features are related to the eigenfunctions by $\phi(\mathbf{x}) = \Sigma^{1/2}\psi(\mathbf{x})$, and to the kernel by $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ where the inner product is the standard one.

We will always work in the over-parameterized setting, meaning that throughout the paper, we assume that $p \geq n$. Since oftentimes $p = \infty$, our bounds will not explicitly depend on p (only implicitly through the eigenvalues of Σ).

Let $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subseteq \mathcal{X}$ be a set of n training points drawn i.i.d from μ, $f^* \in L^2_\mu(\mathcal{X})$ some target function, and $y_i = f^*(\mathbf{x}_i) + \epsilon_i$ be the labels, where ϵ_i is any i.i.d noise with mean 0 and variance σ_ϵ^2. Given some regularization parameter $\gamma_n > 0$, *Kernel Ridge Regression* (KRR) corresponds to minimizing the objective

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(f(\mathbf{x}_i) - y_i\right)^2 + \gamma_n \left\|f\right\|_{\mathcal{H}}^2, \tag{2}$$

where $\mathcal{H}$ is the RKHS of K, consisting of functions of the form $f(\mathbf{x}) = \langle \theta, \phi(\mathbf{x}) \rangle$ with $\|\theta\|_2 < \infty$. The minimizer of the KRR problem in Eq. (2) is given by $\hat{f}(\mathbf{x}) := \langle \hat{\theta}(\mathbf{y}), \phi(\mathbf{x}) \rangle$ where

$$\hat{\theta}(\mathbf{y}) := \phi(X)^\top (\mathbf{K} + n\gamma_n I)^{-1} \mathbf{y}, \tag{3}$$

$\mathbf{K}_{ij} := K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel matrix, and using infinite matrix notation $\phi(X) := [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]^\top \in \mathbb{R}^{n \times p}$ are the training features. As $\gamma_n \to 0$, $\hat{\theta}$ tends to the *min-norm interpolator*:

$$\hat{\theta}(y) = \arg\min_{\theta} \|\theta\|_{\mathcal{H}} \text{ s.t. } \mathbf{y} = \phi(X)\theta. \tag{4}$$

We can decompose the target function as $f^*(\mathbf{x}) = \langle \theta^*, \phi(\mathbf{x}) \rangle + P^\perp f^*$ where $\theta^* \in \mathbb{R}^p$ and $P^\perp$ is the orthogonal projection onto the space spanned by the eigenfunctions with 0 eigenvalues (from Eq. (1)). In particular, if K has no zero eigenvalues in Eq. (1), then $P^\perp f^* = 0$. By the orthonormality of ψ_i, it holds that $\|f^*\|_{L^2_\mu(\mathcal{X})} = \|\Sigma^{1/2}\theta^*\|_2 + \|P^\perp f^*\|_{L^2_\mu(\mathcal{X})}$. We do not require f^* to be in the RKHS. We will define the excess risk of KRR as:

$$\begin{aligned} R\left(\hat{\theta}(y)\right) :=& \mathbb{E}_{\mathbf{x},\epsilon}\left[\left(\langle \hat{\theta}(\mathbf{y}), \phi(\mathbf{x}) \rangle - f^*(\mathbf{x})\right)^2\right] \\ =& \mathbb{E}_{\mathbf{x},\epsilon}\left[\left\langle \hat{\theta}(\mathbf{y}) - \theta^*, \phi(\mathbf{x}) \right\rangle^2\right] + \left\|P^\perp f^*\right\|_{L^2_\mu(\mathcal{X})}^2 . \end{aligned} \tag{5}$$

By linearity, the predictor can be decomposed as $\hat{\theta}(\mathbf{y}) = \hat{\theta}(\phi(X)\theta^*) + \hat{\theta}(\boldsymbol{\epsilon})$. Using this, the fact that the noise is independent of $\mathbf{x}$ and letting $\|\mathbf{x}\|_\Sigma := \sqrt{\mathbf{x}^\top \Sigma \mathbf{x}}$, the excess risk from Eq. (5) can be decomposed in terms of bias, variance, and an approximation error as:

$$\begin{aligned} &B := \left\|\hat{\theta}(\phi(X)\theta^*) - \theta^*\right\|_\Sigma^2, \quad V := \mathbb{E}_\epsilon\left[\left\|\hat{\theta}(\boldsymbol{\epsilon})\right\|_\Sigma^2\right] \\ &R\left(\hat{\theta}(y)\right) = B + V + \left\|P^\perp f^*\right\|_{L^2_\mu(\mathcal{X})}^2 . \end{aligned} \tag{6}$$

2.1. Issues With Past Works

There is a vast literature on KRR and linear regression, with many interesting results under various assumptions and settings. However, perhaps surprisingly, there does not appear to be a unified theory that can provide upper bounds for the excess risk of kernel regression for common kernels and for *any* amount of regularization, noise, any input dimension, and any number of samples. We now detail a few aspects of how current bounds are insufficient.

Assumptions That Do Not Hold: Many works rely on assumptions that are common or reasonable for analyzing *linear* regression. However, as we argue below, they are generally inapplicable for kernel regression. These assumptions include that the features $\phi_i(\mathbf{x})$ are Gaussians (Spigler et al., 2020; Jacot et al., 2020; Cui et al., 2021), and that the eigenfunctions $\psi_i(\mathbf{x})$ (sometimes called covariates) are sub-Gaussian, i.i.d finite dimensional and/or have mean 0 (Bartlett et al., 2020; Hastie et al., 2022; Cheng & Montanari, 2022; Tsigler & Bartlett, 2023; Bach, 2023; Cheng et al., 2023). There are also works that make various non-rigorous assumptions common in the statistical physics literature (Bordelon et al., 2020; Gerace et al., 2020; Canatar et al., 2021; Simon et al., 2021; Mallinar et al., 2022), which, despite being observed to be empirically accurate in some settings, do not provide rigorous results.

Unfortunately, none of the assumptions mentioned in the previous paragraph hold for common kernels, making such works incapable of providing rigorous results in common settings. As a simple example, suppose our inputs are one-dimensional standard Gaussians $x \sim \mathcal{N}(0,1)$ and let $K(x,y) = \exp\left(-\gamma(x-y)^2\right)$ be a Gaussian (RBF) kernel. We show in Appendix F that if we pick for simplicity $\gamma = \frac{3}{8}$, then for any $p \geq 3$, the moments of ψ_i diverge as

$$\left(\mathbb{E}\left[|\psi_i(x)|^p\right]\right)^{1/p} \geq \Omega_i\left(\exp\left(\frac{p-2}{4}\cdot i\right)\right) \xrightarrow[i\to\infty]{} \infty, \quad (7)$$

implying that for the classical RBF kernel, not only is $\psi(\mathbf{x})$ not sub-Gaussian, but all moments ≥ 3 diverge. Another simple example is given with inputs distributed uniformly on the unit sphere $\mathbb{S}^{d-1}$, and dot product kernels such as RBF, Laplace and NTK. Under this setting, $\psi_i(\mathbf{x})$ are given by spherical harmonics, for which even in the case of $d = 3$ the third moments diverge as $i \to \infty$ (Han, 2016). Additionally, for dot product kernels, ψ_i are definitely not i.i.d across i, ψ_1 is generally constant and not mean 0, and p may be ∞ (see Appendix G for more details).

A major issue with overly strong assumptions is that they may lead to inaccurate predictions. Specifically, they induce concentration inequalities (e.g bounding the eigenvalues of the empirical covariance matrix) which are tighter than one can typically expect, resulting in risk bounds that may be over-optimistic (see Fig. 2). By contrast, we work under very mild and realistic assumptions, and we do not know of any interesting kernel for which our analysis is not applicable.

Limitation to a Specific Setting: The literature seems to be split into several categories, with different works focusing on incompatible settings. These include:

"High-Dimensional" vs. "Fixed-Dimensional": Many works assume that the input dimension d and the number of samples n both tend towards infinity at a fixed ratio $n = d^\tau$ for some $\tau > 0$ (Dobriban & Wager, 2018; Liang & Rakhlin, 2020; Wu & Xu, 2020; Richards et al., 2021; Ghorbani et al., 2021; Li et al., 2021; Hu & Lu, 2022; Montanari & Zhong, 2022; Mei et al., 2022; Misiakiewicz, 2022; Xiao & Pennington, 2022). By contrast, other lines of work assume a fixed d and $n \to \infty$ (Caponnetto & De Vito, 2007; Steinwart et al., 2009; Li et al., 2023a; Cui et al., 2021; Li et al., 2023b). The techniques and assumptions used by these two lines of work are inherently different, and make the results from the high-dimensional works inapplicable for fixed d and vice versa. We further elaborate on the limitations of differences between these lines of work in Appendix J. Nevertheless, we obtain bounds that are relevant for *any* d, n, regardless of the ratio between them, and in particular, capture interesting phenomena in these two regimes.

Regularized vs. Unregularized: Several works are limited to either the regularized case (Caponnetto & De Vito, 2007; Steinwart et al., 2009; Fischer & Steinwart, 2020; Lin et al., 2020) or the unregularized case (a.k.a min-norm interpolation) (Bartlett et al., 2020; Liang & Rakhlin, 2020; Hastie et al., 2022). This distinction is of course unwanted, and our results provide bounds that can handle both and make the role of the regularization explicit.

Noisy vs. Noiseless: Cui et al. (2021) noted a discrepancy between rates obtained in a noisy setting (when $\sigma_\epsilon > 0$) (Caponnetto & Vito, 2005; Steinwart et al., 2009) vs. a noiseless setting (when $\sigma_\epsilon = 0$) (Spigler et al., 2020). Furthermore, quantifying the effect of the noise is important since even when $\sigma_\epsilon > 0$, one may still obtain a small excess risk if the noise is small. Recent works in the fixed dimensional setting still only manage to provide upper bounds in the noiseless case (Li et al., 2023b). Our analysis handles both cases, separating the bias and variance, and upper bounding both of these separately.

There are also many prior works that bound the eigenvalues of kernel matrices similarly to what we do here (e.g Rosasco et al. (2010); Valdivia (2018)). We provide a detailed discussion in Appendix J, but briefly mention here that these do not yield sufficiently strong bounds for the smallest eigenvalue of the kernel matrix for our needs. As we shall see, this will be crucial for our analysis.

2.2. Additional Notations and Definitions

We use the subscripts $\leq k$ and $> k$ to denote the first $1, \dots, k$ and $k+1, k+2, \dots$ coordinates of a vector (e.g $\phi_{\leq k}(X)$ is an $n \times k$ matrix). Similarly, let $\mathbf{K}_{\leq k} := \phi_{\leq k}(X)\phi_{\leq k}(X)^T$ and $\mathbf{K}_{>k} := \phi_{>k}(X)\phi_{>k}(X)^T$. For an operator T, we use $\mu_i(T)$ to denote its i'th largest eigenvalue (allowing repeated eigenvalues). We use this notation to avoid confusion with the eigenvalues λ_i of Σ. Unless stated otherwise, $\|\cdot\|$ is the standard ℓ^2 norm for vectors and operator norm for operators. We use the standard big-O notation and the $\tilde{\mathcal{O}}(\cdot)$ notation to hide additional logarithmic factors. We may make the problem parameters explicit, e.g $\mathcal{O}_{n,d}$, to mean up to constants that do not depend on n or d.

As in Bartlett et al. (2020), for any $k \in \mathbb{N}$, we define two highly related notions of the effective rank of $\Sigma_{>k}$ as:

$$r_k := r_k(\Sigma) := \frac{\operatorname{tr}(\Sigma_{>k})}{\|\Sigma_{>k}\|}, \quad R_k := R_k(\Sigma) = \frac{\operatorname{tr}(\Sigma_{>k})^2}{\operatorname{tr}\left(\Sigma_{>k}^2\right)}. \tag{8}$$

r_k is the common definition of effective rank, and is related to R_k via $r_k \leq R_k \leq r_k^2$ (Bartlett et al., 2020)[Lemma 5].

2.3. Assumptions

Typically, one must assume something on $\psi(\mathbf{x})$ to obtain various concentration inequalities, meaning that the kernel matrix and empirical covariance matrix will behave as they are "supposed to". Perhaps the most common assumption in previous works is that $\psi(\mathbf{x})$ is sub-Gaussian, requiring the moments of $\psi_i(\mathbf{x})$ to be sufficiently well-behaved for every i. Unfortunately, as discussed earlier, this does not hold for many common kernels, even when the input distribution is "nice." In order to overcome this issue, we present a framework for analyzing kernels under only a mild heavy-tailed condition which can be shown to hold for many common kernels. In particular, we wish that quantities concerning the features will be related to their expected values by a multiplicative constant. By the orthonormality of ψ_i, for any $k \in \mathbb{N}$ one has that $\mathbb{E}[\|\psi_{\leq k}(\mathbf{x})\|^2] = k, \mathbb{E}[\|\phi_{>k}(\mathbf{x})\|^2] = \operatorname{tr}(\Sigma_{>k})$ and $\mathbb{E}\left[\left\|\Sigma_{>k}^{1/2}\phi_{>k}(\mathbf{x})\right\|^2\right] = \operatorname{tr}\left(\Sigma_{>k}^2\right)$. We quantify the distance of the quantities from their expected values by the following definitions:

Definition 2.1. Given $k \in \mathbb{N}$, let $\beta_k \geq \alpha_k \geq 0$ be defined as follows:

$$\alpha_k := \inf_{\mathbf{x}} \left\{ \frac{\|\phi_{>k}(\mathbf{x})\|^2}{\operatorname{tr}(\Sigma_{>k})} \right\}, \tag{9}$$

$$\beta_k := \sup_{\mathbf{x}} \left\{ \frac{\|\psi_{\leq k}(\mathbf{x})\|^2}{k}, \frac{\|\phi_{>k}(\mathbf{x})\|^2}{\operatorname{tr}(\Sigma_{>k})}, \frac{\|\Sigma_{>k}^{1/2}\phi_{>k}(\mathbf{x})\|^2}{\operatorname{tr}\left(\Sigma_{>k}^2\right)} \right\}, \tag{10}$$

where the sup and inf are for a.s any $\mathbf{x}$.

For each term in these definitions, the denominator is the expected value of the numerator, so α_k and β_k quantify how much the features behave as they are "supposed to". Since $\inf \leq \mathbb{E} \leq \sup$, one always has $0 \leq \alpha_k \leq 1 \leq \beta_k$. Upper bounding β_k is often easy, and common examples for kernels with $\beta_k = \mathcal{O}_k(1)$ include dot product kernels such as NTK and polynomial kernels, shift-invariant kernels, and kernels with bounded eigenfunctions $\|\psi(\mathbf{x})\|_\infty < \infty$. α_k can also be lower bounded as $\Omega_k(1)$ for many kernels (e.g dot product kernels); however, a lower bound on α_k may sometimes be more difficult, and as such, many of our bounds will not require any control of α_k. Nevertheless, when $\alpha_k > 0$, in some cases, stronger bounds will be available. We defer a more complete discussion of these definitions, their relation to common kernels, and our claims in this paragraph to Appendix H. Overall, for sufficiently "nice" kernels, one should think of α_k and β_k as generally being $\Theta_k(1)$. For the bounds in this paper, we will not need to control α_k and β_k for every value of k, but rather k can be arbitrarily chosen.

Remark 2.2. Def. (9) and Def. (10) are stated for a.s any $\mathbf{x}$. However, one can weaken the definition for α_k to the training set, so that w.p at least $1 - \delta_k$, $\min_{\mathbf{x}_i \in X} \left\{ \frac{\|\phi_{>k}(\mathbf{x})\|^2}{\operatorname{tr}(\Sigma_{>k})}, \right\} \geq \alpha_k$. In such a case, all bounds that depend on α_k would still hold with probability $1 - \delta_k$.

In some cases, we will need to make the control of β_k explicit via the following regularity assumption.

Assumption 2.3. Either the feature dimension p is finite, or there exists some sequence of natural numbers $(k_i)_{i=1}^\infty \subseteq \mathbb{N}$ with $k_i \xrightarrow[i\to\infty]{} \infty$ s.t. $\beta_{k_i}\operatorname{tr}(\Sigma_{>k_i}) \xrightarrow[i\to\infty]{} 0$.

Because Mercer kernels are trace class, one always has $\operatorname{tr}(\Sigma_{>k_i}) \xrightarrow[i\to\infty]{} 0$. As such, Assumption 2.3 simply states that for infinitely many choices of $k \in \mathbb{N}$, β_k does not increase too quickly. This is of course satisfied by the previous examples of kernels with $\beta_k = \mathcal{O}_k(1)$.

3. Eigenvalues of Kernel Matrices

Since the KRR solution can be written as in Eq. (3), understanding it requires understanding the structure of the kernel matrix $\mathbf{K}$. In particular, we will need tight bounds on its eigenvalues. For a fixed $k \in \mathbb{N}$, it is known that $\mu_k\left(\frac{1}{n}\mathbf{K}\right)$ should tend to λ_k as $n \to \infty$, with known bounds of the form $\left|\mu_k\left(\frac{1}{n}\mathbf{K}\right) - \lambda_k\right| = \mathcal{O}\left(\frac{\operatorname{tr}(\Sigma)}{\sqrt{n}}\right)$ (Rosasco et al., 2010). Unfortunately, these bounds are the same for all $k \leq n$. Since usually $\lambda_k = o\left(\frac{1}{k}\right)$, for most eigenvalues the $\mathcal{O}\left(\frac{\operatorname{tr}(\Sigma)}{\sqrt{n}}\right)$ approximation error is much larger than the eigenvalues themselves, leading to the very weak bound of $0 \leq \mu_k\left(\frac{1}{n}\mathbf{K}\right) \leq \mathcal{O}\left(\frac{\operatorname{tr}(\Sigma)}{\sqrt{n}}\right)$. This is insufficient for multiple reasons. First, the expected decay of eigenvalues in the kernel matrix is not captured. Second, tighter lower bounds are often necessary to ensure the kernel matrix is positive definite and well-conditioned. Control of the smallest eigenvalue is a common working assumption in the NTK literature (Du et al., 2019; Arora et al., 2019; Hu & Lu, 2022) and determines the convergence rate of gradient descent with the corresponding network (Geifman et al., 2023).

We address these issues by providing *relative* perturbation bounds. The general approach is, given some $k \in \mathbb{N}$, to decompose the kernel matrix as $\mathbf{K} = \mathbf{K}_{\leq k} + \mathbf{K}_{>k}$ where the eigenvalues of the "low-dimensional" part $\mathbf{K}_{\leq k}$ should concentrate well, and the "high-dimensional" part $\mathbf{K}_{>k}$ should approximately be $\tilde{\gamma} I$ for some $\tilde{\gamma} > 0$.

Theorem 3.1. *Suppose Assumption 2.3 holds, and that the eigenvalues of* Σ *are given in non-increasing order* $\lambda_1 \geq \lambda_2 \geq \ldots$. *There exist some absolute constants* $c, C, c_1, c_2 > 0$ *s.t for any* $k \leq k' \in [n]$ *and* $\delta > 0$, *w.p at least* $1 - \delta - 4\frac{r_k}{k^4} \exp\left(-\frac{c}{\beta_k}\frac{n}{r_k}\right) - 2\exp\left(-\frac{c}{\beta_k}\max\left(\frac{n}{k}, \log(k)\right)\right)$,

$$\frac{1}{c_2 \beta_k}\mu_k\left(\frac{1}{n}\mathbf{K}\right) \leq \left(1 + \frac{k\log(k)}{n}\right)\lambda_k + \log(k+1)\frac{tr(\Sigma_{>k})}{n},$$

and

$$\mu_k\left(\frac{1}{n}\mathbf{K}\right) \geq c_1 \mathbb{I}_{k,n}\lambda_k + \alpha_k\left(1 - \frac{1}{\delta}\sqrt{\frac{n^2}{R_{k'}}}\right)\frac{tr\left(\Sigma_{>k'}\right)}{n},$$

where $\mathbb{I}_{k,n} = \begin{cases} 1, & \text{if } C\beta_k k\log(k) \leq n \\ 0, & \text{otherwise} \end{cases}$.

Informally, the theorem shows that one can decompose the kernel matrix into a "low-dimensional" part whose eigenvalues are well-behaved and a "high-dimensional" part which is roughly proportional to the identity matrix and thus serves in a similar role as the regularization in Eq. (3). More specifically, when decomposing as $\mathbf{K} = \mathbf{K}_{\leq k} + \mathbf{K}_{>k}$, the "low-dimensional" part $\mathbf{K}_{\leq k}$ satisfies $\mu_k\left(\frac{1}{n}\mathbf{K}_{\leq k}\right) \approx \lambda_k$ for $k \leq \mathcal{O}_n\left(\frac{n}{\log(n)}\right)$. The main tools we use to show this are a variant of Ostrowski's theorem for non-square matrices (Lemma C.1) combined with some concentration of measure arguments (Lemma B.2). By contrast, for the smaller eigenvalues of the kernel matrix where $k = \omega_n\left(\frac{n}{\log(n)}\right)$, one instead has to turn towards the self-regularization induced by the $> k$ features. One should pick k' so that $R_{k'} > \frac{n^2}{\delta^2}$ (see the next paragraph for an example). This implies that the smaller eigenvalues of the kernel matrix can be bounded as $\mu_k\left(\frac{1}{n}\mathbf{K}\right) \gtrsim \frac{\mathrm{tr}(\Sigma_{>k'})}{n}$. If the eigenvalues decay sufficiently slowly, k' can be picked not too large, and this self-regularization becomes significant. Similar behavior has been observed in the asymptotic high-dimensional regime using random matrix theory arguments (Xiao & Pennington, 2022).

As an example, suppose $\lambda_k = \Theta\left(\frac{1}{k\log^{1+a}(k)}\right)$ for some $a > 0$ and $\alpha_k, \beta_k = \Theta(1)$ (a condition satisfied by many common kernels, see Appendix H). Then taking $k' := k'(n) := n^2$, one can easily calculate that $R_{k'} \geq \Omega(n^2\log(n))$ and $\frac{\mathrm{tr}(\Sigma_{>k'})}{n} = \Theta\left(\frac{1}{n\log^a(n)}\right)$. As a result, letting $\tilde{\gamma}_n := \frac{1}{n\log^a(n)}$, Thm. 3.1 implies that for any $k \in [n]$, $\mu_k\left(\frac{1}{n}\mathbf{K}\right) \geq \Omega\left(\mathbb{I}_{k,n}\lambda_k + \tilde{\gamma}_n\right)$. In particular, the smallest eigenvalues can be lower bounded as $\mu_n\left(\frac{1}{n}\mathbf{K}\right) \geq \Omega\left(\frac{1}{n\log^a(n)}\right) \gg \lambda_n$. This result is at first surprising, as the classical intuition arising from works discussed earlier which bound $\left|\mu_k\left(\frac{1}{n}\mathbf{K}\right) - \lambda_k\right|$ would suggest that $\mu_n\left(\frac{1}{n}\mathbf{K}\right) \approx \lambda_n$. One can analogously obtain a matching upper bound up to a $\log(k)$ factor.

The parameter $\tilde{\gamma}$ in the above example plays an identical role in KRR as the actual regularization term γ_n. As such, the kernel actually provides its own regularization, arising from the high dimensionality of the features and the flatness of the eigenvalues. We call this *self-induced regularization*, and it has two significant implications. First, it can be used to derive good bounds on the smallest eigenvalue of a kernel matrix, which as already mentioned, is critical for many applications, and will be used extensively to derive new KRR bounds in the following sections. Second, it can (quite surprisingly) cause the eigenvalues of the kernel matrix to decay at a significantly different rate than λ_k. In particular, when the self-regularization is large enough, the eigenvalues $\mu_k\left(\frac{1}{n}\mathbf{K}\right)$ concentrate around $\lambda_k + \tilde{\gamma}$ up to a multiplicative constant.

4. Excess Risk of Kernel Regression

We now return to bounding the bias and variance of KRR as given by Eq. (6). The strategy will be to pick some $k \leq n$, and treat the $\leq k$ and $> k$ components separately. By the previous section, we expect that $\mathbf{K}_{>k} \approx \tilde{\gamma} I$ and this will serve as a regularization term for KRR. We quantify this by what we call the *concentration coefficient*,

$$\rho_{k,n} := \frac{\|\Sigma_{>k}\| + \mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) + \gamma_n}{\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) + \gamma_n}. \tag{11}$$

Because $\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) = \|\hat{\Sigma}_{>k}\|$ where $\hat{\Sigma}_{>k}$ is the (uncentered) empirical covariance matrix and $\mathbb{E}[\hat{\Sigma}_{>k}] = \Sigma_{>k}$, one should expect that any upper bound on $\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right)$ should be larger than $\|\Sigma_{>k}\|$. As such, the $\|\Sigma_{>k}\|$ term practically affects $\rho_{k,n}$ by at most a factor of 2, and we include it only for technical simplicity within the proofs. Now, if for some k, one shows that $\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) \approx \mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right)$ then $\rho_{k,n}$ can be bounded as $\Theta(1)$. As we shall soon show, in such a case, it will follow that the bias and variance can be well bounded. Although our theory from the previous section provides a bound for $\rho_{k,n}$, we make its role explicit in the bias and variance bounds, since tighter bounds on $\rho_{k,n}$ may be available when there is additional information on the structure of the kernel.

Theorem 4.1. *Let $k \in \mathbb{N}$ and let $\rho_{k,n}$ be as defined in Eq. (11). There exists some absolute constants $c, c', C_1, C_2 > 0$ s.t if $c\beta_k k \log(k) \leq n$, then for every $\delta > 0$, it holds w.p at least $1 - \delta - 16\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ that both the variance and bias can be upper bounded as:*

$$V \leq C_1\rho_{k,n}^2\sigma_\epsilon^2\left(\frac{k}{n} + \min\left(\frac{r_k\left(\Sigma^2\right)}{n}, \frac{n}{\alpha_k^2 R_k(\Sigma)}\right)\right), \tag{12}$$

$$\begin{aligned} B \leq & C_2\rho_{k,n}^3\left(\frac{1}{\delta}\left\|\theta^*_{>k}\right\|^2_{\Sigma_{>k}}\right. \\ & \left. + \left\|\theta^*_{\leq k}\right\|^2_{\Sigma^{-1}_{\leq k}}\left(\gamma_n + \frac{\beta_k tr\left(\Sigma_{>k}\right)}{n}\right)^2\right). \end{aligned} \tag{13}$$

Several comments are in order. First, the optimal choice of k should depend on the concentration coefficient $\rho_{k,n}$ and the eigenvalues λ_i of the kernel. Given these, one can determine an asymptotically optimal k as a function of n. One would typically want to take k to be as small as possible, while still ensuring $\rho_{k,n} \approx 1$. Second, we do not assume here that the eigenvalues λ_i are ordered. This is important because for certain kernels, ordering the eigenvalues is actually quite difficult, for example with NTKs corresponding to popular convolutional architectures (Barzilai et al., 2022). This flexibility will be critical for our analysis in the following section involving dot product kernels. Finally, a control of α_k is not required to obtain bounds for the bias and variance, and is present only in Eq. (12) via the term $\min\left(\frac{r_k\left(\Sigma^2\right)}{n}, \frac{n}{\alpha_k^2 R_k(\Sigma)}\right)$. Under a slight abuse of notation, even when $\alpha = 0$, this term is at most $\frac{r_k\left(\Sigma^2\right)}{n}$. As we shall later show in Thm. 5.3, under sufficient regularization, our bounds on the excess risk will not depend on α_k.

We also note that in the simple case of finite-dimensional linear regression (where $\phi(\mathbf{x}) = \mathbf{x}$) with zero mean and sub-Gaussian $\psi(\mathbf{x}) = \Sigma^{-1/2}\mathbf{x}$, our bounds provide a significant generalization of those of Tsigler & Bartlett (2023)[Theorem 1]. Specifically, they derived similar bounds for a specific k which is hard to determine, under the explicit assumption that the condition number of $\frac{1}{n}\mathbf{K}_{>k} + \gamma_n I$ (similar to $\rho_{k,n}$) is bounded by some constant. Their results only hold for 0-mean, sub-Gaussian, and finite-dimensional ψ_i, and hence are not applicable for many common kernels. The explicit dependence on $\rho_{k,n}$, as well as the ability to choose k freely, will play an important role in the proofs of Thm. 5.1 and Thm. 5.2 in the next sections. Nevertheless, when all of their assumptions are satisfied, including that the condition number of $\frac{1}{n}\mathbf{K}_{>k} + \gamma_n I$ is constant, our bound precisely recovers theirs. Because they showed that their bounds are sharp up to a multiplicative constant, we also obtain that under sufficient conditions, the upper bounds in Thm. 4.1 are also sharp.

5. Applications

5.1. Benign Overfitting in High Dimensions

In order to capture high-dimensional phenomena that likely play a major role in the success of neural networks, it is common to analyze KRR in a high-dimensional setting. Specifically, where n, d both tend towards infinity, with the ratio $\frac{n}{d^\tau} = \Theta(1)$ fixed for some $\tau > 0$. In this chapter, we consider an important class of kernels known as *dot product kernels* of the form $K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x}^\top\mathbf{x}')$ for some function h. One typically has to impose restriction on h for K to be a valid kernel, and as such, we follow the standard assumption that h has a Taylor expansion of the form $h(t) = \sum_{i=0}^\infty a_i t^i$ with $a_i \geq 0$ (Azevedo & Menegatto, 2015; Scetbon & Harchaoui, 2021). We will currently restrict ourselves to $\mathbb{S}^{d-1}$ (and thus $h : [-1, 1] \to \mathbb{R}$) under the uniform distribution. Examples of dot product kernels on $\mathbb{S}^{d-1}$ include NTKs and GPKs of fully-connected networks and fully-connected-ResNets, Laplace kernels, Gaussian (RBF) kernels, and polynomial kernels (Smola et al., 2000; Minh et al., 2006; Bietti & Bach, 2020; Chen & Xu, 2020). For any $d \geq 3$, dot product kernels with inputs uniformly distributed on $\mathbb{S}^{d-1}$ have known Mercer decompositions given by

$$K(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^\infty \frac{\hat{\sigma}_\ell}{N(d, \ell)}\sum_{m=1}^{N(d,\ell)} Y_{\ell,m}(\mathbf{x})Y_{\ell,m}(\mathbf{x}'), \tag{14}$$

where the eigenfunctions $Y_{\ell,m}$ are the m'th spherical harmonic of degree (or frequency) ℓ, $N(d, \ell) = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{d-2}$ is the number of harmonics of each degree, and $\sigma_\ell := \frac{\hat{\sigma}_\ell}{N(d,\ell)}$ are the eigenvalues (Smola et al., 2000). We defer a background on dot product kernels and more involved explanations to Appendix G. We now show that in the high-dimensional regime, any dot product kernel is capable of benign overfitting, i.e achieving an excess risk that approaches zero as $n \to \infty$, without regularization and despite the presence of noise.

Theorem 5.1. *Suppose that as $n, d \to \infty$, $\frac{d^\tau}{n} = \Theta_{n,d}(1)$ for some $\tau \in (0, \infty) \setminus \mathbb{N}$. Let μ be the uniform distribution over $\mathbb{S}^{d-1}$ and K be a dot product kernel given by Eq. (14) s.t $\hat{\sigma}_{\lfloor\tau\rfloor} > 0$ and $\exists \ell > \lfloor 2\tau \rfloor$ with $\hat{\sigma}_\ell \geq 0$ (e.g NTK, Laplace, or RBF). Then for the min norm solution defined in Eq. (4) (given when $\gamma_n \to 0$), for any $\delta > 0$ it holds w.p at least $1 - \delta - o_d\left(\frac{1}{d}\right)$ that*

$$V \leq \sigma_\epsilon^2 \cdot \mathcal{O}_{n,d}\left(\frac{1}{d^{\tau - \lfloor\tau\rfloor}} + \frac{1}{d^{\lfloor\tau\rfloor + 1 - \tau}}\right),$$

$$\begin{aligned} B \leq & \frac{1}{\delta}\mathcal{O}_{n,d}\left(\left\|\theta^*_{>N_d}\right\|^2_{\Sigma_{>N_d}}\right) \\ & + \left\|\theta^*_{\leq N_d}\right\|^2_\infty\left(\max_{\ell \leq \lfloor\tau\rfloor \text{ s.t. } \hat{\sigma}_\ell \neq 0}\frac{1}{\hat{\sigma}_\ell}\right) \cdot \mathcal{O}_{n,d}\left(\frac{1}{d^{2(\tau - \lfloor\tau\rfloor)}}\right). \end{aligned}$$

Where $N_d = \Theta_{n,d}\left(d^{\lfloor\tau\rfloor}\right)$ denotes the number of spherical *harmonics of degree at most $\lfloor\tau\rfloor$ with non-zero eigenvalues, and $\mathcal{O}_{n,d}\big(\|\theta^*_{>N_d}\|^2_{\Sigma_{>N_d}}\big) \leq \mathcal{O}_{n,d}\big(\|\theta^*_{>N_d}\|^2_\infty\big)$.*

Simply put, the variance decays to 0, and the bias approaches $\mathcal{O}_{n,d}\big(\|\theta^*_{>N_d}\|^2_\infty\big)$ for $N_d \approx d^{\lfloor\tau\rfloor}$. More specifically, the rate of decay for the variance depends on τ, with the fastest decay occurring when $\tau = z + \frac{1}{2}$ for some $z \in \mathbb{N}$, and slowest when $\tau \approx z$. This highlights the multiple descent behavior of KRR as discussed in Liang & Rakhlin (2020); Xiao & Pennington (2022). For the bias, $\|\theta^*_{>N_d}\|^2_{\Sigma_{>N_d}}$ is the L^2_μ norm of the projection of f^* onto the spherical harmonics of degree at least $\lceil\tau\rceil$, and $\|\theta^*_{>N^d}\|^2_\infty$ is the maximal projection. The $\max\limits_{\ell\leq\lfloor\tau\rfloor \text{ s.t. } \hat{\sigma}_\ell\neq 0} \frac{1}{\hat{\sigma}_\ell}$ term will typically be $\mathcal{O}_{n,d}(1)$ because often times $\hat{\sigma}_\ell = \Omega_{n,d}(1)$. For example, for the NTK, one has an even stronger statement, $\max\limits_{\ell\leq\lfloor\tau\rfloor \text{ s.t. } \hat{\sigma}_\ell\neq 0} \frac{1}{\hat{\sigma}_\ell} = \mathcal{O}_{n,d}(\frac{1}{d})$ (Cao et al., 2019)[Theorem 4.3]. Thus, whether KRR achieves benign overfitting or not depends on the spectral decomposition of the target function. If θ^* consists of frequencies of at most $\lfloor\tau\rfloor$, then $\|\theta^*_{>N_d}\|^2_\infty = 0$ and thus both the bias and variance tend towards zero, implying benign overfitting. The variance for high-dimensional regression is demonstrated in Fig. 1 for the NTK and polynomial kernel.

The key to this result is that the repeated eigenvalues lead to large effective ranks r_k and R_k, allowing one to take $k = N_d$ (where $\frac{N_d}{n} = \frac{1}{d^{\tau-\lfloor\tau\rfloor}}$) with concentration coefficient $\rho_{k,n} = \Theta(1)$. Notably, there is nothing specific to dot product kernels, and using Thm. 4.1, a similar result can be derived for any kernel with $\rho_{k,n} = \Theta(1)$ for $k \ll n$. The assumptions on $\hat{\sigma}$ are made only for simplicity to avoid degeneracies via convoluted examples involving 0 eigenvalues. We make the role of this assumption clear within the proof, as it can easily be modified. For example, one can obtain similar results when the 0 eigenvalues are the odd frequencies as in an NTK without bias (Basri et al., 2019; Bietti & Bach, 2020)

Our results can naturally be extended to other domains and distributions. Li et al. (2023a) show that the eigenvalues only change by multiplicative constants under suitable changes of measures or diffeomorphisms. One can also exploit the specific structure of certain kernels. For example, NTK kernels and homogeneous polynomial kernels are zonal, meaning that $K(\mathbf{x},\mathbf{x}') = \|\mathbf{x}\|\,\|\mathbf{x}'\|\,K\big(\frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{x}'}{\|\mathbf{x}'\|}\big)$, so results from $\mathbb{S}^{d-1}$ can easily generalize to $\mathbb{R}^d$.

Perhaps the works that provide the results most similar to Thm. 5.1 are the excellent papers of Liang & Rakhlin (2020); Mei et al. (2022); Xiao & Pennington (2022). By comparison, Xiao & Pennington (2022)[Corollary 2] do not provide convergence rates, but rather show that the excess risk approaches $\big\|\theta^*_{>N_d}\big\|^2_{\Sigma_{>N_d}} + o_d(1)$ as $n, d \to \infty$. They assumed that $\hat{\sigma}_\ell$ are $\Theta_d(1)$ independent of d, a condition which is typically not satisfied, e.g. in an NTK. Mei et al. (2022)[Theorem 4] when combined with a "spectral gap condition" (which would also require that $\hat{\sigma}_\ell$ are $\Theta_d(1)$) also implies a bound of the form $\|\theta^*_{>N_d}\|^2_{\Sigma_{>N_d}} + o_d(1)$. It is unclear what their bound implies without this problematic spectral gap assumption. They also impose other strict assumptions, which do not hold for broader domains. For example, they assume that for any $\mathbf{x}_i$, $\frac{\|\phi_{>N_d}(\mathbf{x}_i)\|^2}{\operatorname{tr}\left(\Sigma_{>N_d}\right)} = 1 \pm o_d(1)$. For zonal kernels such as the NTK, this typically will not hold unless all inputs have roughly the same norm. By contrast, our mild assumptions imply that the same results hold in $\mathbb{R}^d$ as discussed above. The results of Liang & Rakhlin (2020)[Theorem 3] are limited to target functions in the RKHS, with a bound that is the same for all θ^*. This is critical since the structure of θ^* is precisely what allows us to characterize when benign overfitting occurs.

Overall, our results are the first to clearly characterize benign overfitting for common kernels, such as NTK.

5.2. Nearly Tempered Overfitting in Fixed Dimensions

We now shift our attention to the fixed dimensional regime. We focus on polynomially decaying eigenvalues, encompassing NTKs and GPKs of common fully-connected architectures (Bietti & Bach, 2020), convolutional and residual architectures (Geifman et al., 2022; Barzilai et al., 2022) as well as the Laplace kernel (Chen & Xu, 2020). For such kernels, various works show lower bounds of the form $\Omega(1)$ for the excess risk for min-norm interpolation (Rakhlin & Zhai, 2019; Haas et al., 2023). Recently Mallinar et al. (2022) distinguished between the regimes where the risk explodes to ∞ (called catastrophic overfitting) vs when the risk remains bounded (called tempered overfitting). The two regimes are significantly different since when the noise is small, kernel regression can still achieve a low risk despite tempered overfitting. Using our tools, we show that when $\lambda_i \approx i^{-1-a}$ for small $a > 0$, such kernels are *nearly tempered*, meaning that the bias goes to 0, and the variance cannot diverge too quickly.

Theorem 5.2. *Let K be a kernel with polynomially decaying eigenvalues $\lambda_i = \Theta_{i,n}(i^{-1-a})$ for some $a > 0$, and assume that $\alpha_k, \beta_k = \Theta_k(1)$. Then for the min norm solution defined in Eq. (4) (given when $\gamma_n \to 0$), for any $\delta > 0$ it holds w.p at least $1 - \delta - \mathcal{O}_n\big(\frac{1}{\log(n)}\big)$ that*

$$V \leq \sigma^2_\epsilon \tilde{\mathcal{O}}_n\left(n^{2a}\right).$$

Moreover, if $\theta^*_i = \mathcal{O}_i\left(\frac{1}{i^r}\right)$ where $r > a$ then under the same probability it also holds that

$$B \leq \frac{1}{\delta}\tilde{\mathcal{O}}_n\left(\frac{1}{n^{\min(2(r-a),2-a)}}\right).$$

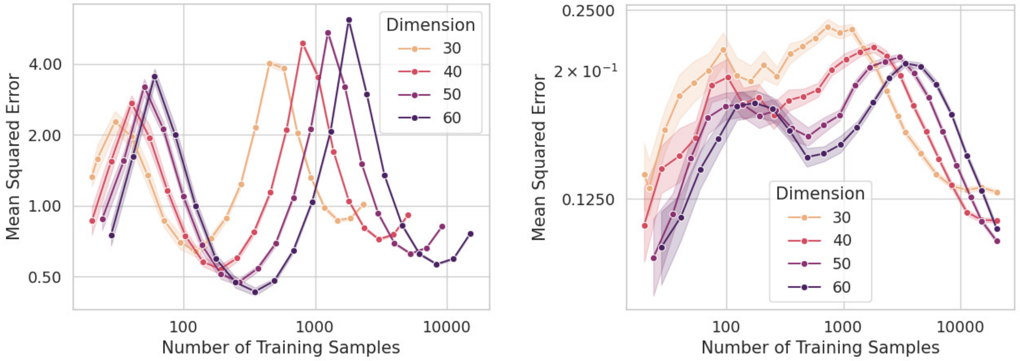

Figure 1. Variance of unregularized Kernel Regression, measured by the MSE for learning a constant 0 function with noise $\epsilon_i \sim \mathcal{N}(0,1)$ and inputs uniformly in $\mathbb{S}^{d-1}$ (log-log scale). Left: Polynomial kernel $K(\mathbf{x},\mathbf{x}') = (1 + \frac{1}{d}\langle \mathbf{x},\mathbf{x}'\rangle)^3$; Right: NTK corresponding to a 3-layer fully-connected network (see Appendix I). As the input dimension grows, the multiple descent phenomenon becomes more pronounced, and the MSE at the "valleys" decreases. The shaded region denotes 95% confidence over 50 runs with 2500 test samples.

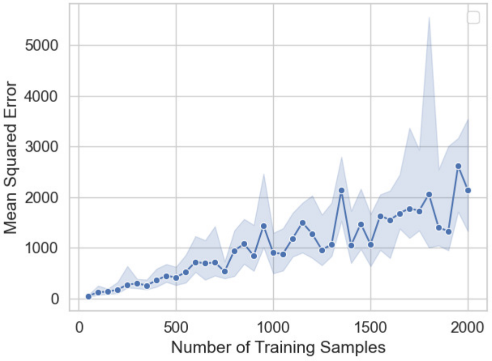

Figure 2. Apparently diverging variance in low dimensions, for a GPK corresponding to a 3-layer fully-connected network (see Appendix I) with inputs distributed uniformly on the unit disk $\{x \in \mathbb{R}^2 : \|x\| \le 1\}$ and noise $\epsilon \sim \mathcal{N}(0,1)$. The solid line denotes the median variance (and not mean, due to extreme values), and the shaded region denotes 95% confidence over 100 trials with 5000 test samples each. This suggests that previous works that inferred $V \le \mathcal{O}(1)$ for kernels with polynomially decaying eigenvalues may be overly optimistic.

When $a \to 0$, the bound for the variance approaches $\tilde{\mathcal{O}}(1)$, and the bound for the bias is nearly $\tilde{\mathcal{O}}\left(\frac{1}{n^{2r}}\right)$. For either the NTK of a fully-connected network or the Laplace kernel, $a = \frac{1}{d-1}$ (Chen & Xu, 2020), so the variance bound becomes $\tilde{\mathcal{O}}\left(n^{\frac{2}{d-1}}\right)$. In fact, when $d \gtrsim \log(n)$ it holds that $n^{\frac{2}{d-1}} \lesssim \text{polylog}(n)$. So when the noise is small, one can expect the excess risk to also be relatively small. The condition on the decay of θ^* is fairly mild, as for any realizable f^* (i.e $f^* \in \mathcal{H}$) it holds that $\|\theta^*\|_2 < \infty$ and thus, under the conditions of the theorem, $r > 1$ and $B < \tilde{\mathcal{O}}(\frac{1}{n^{2-2a}})$.

As far as we know, this is the first rigorous upper bound for the excess risk of the min-norm interpolator in the fixed dimensional setting. Previous bounds were either based on a Gaussian feature assumption or non-rigorous analysis (Cui et al., 2021; Mallinar et al., 2022) and gave $\mathcal{O}\left(n^{-\min(2r+a,2(1+a))}\right)$ and $\sigma_\epsilon^2 \cdot \mathcal{O}(1)$ bounds for the bias and variance respectively. In Fig. 2, we provide a simple example of a common kernel (GPK) that does not appear to adhere to their bounds. The difference between our bounds and theirs is not a limitation of our work but rather due to their strong assumptions and can be quantified by the concentration coefficient $\rho_{k,n}$. Without any special assumptions, we showed that for $k \approx \frac{n}{\log(n)}$, $\rho_{k,n} = \mathcal{O}\left(n^a \text{polylog}(n)\right)$. If one is willing to make stronger assumptions on the features which may not hold in practice (such as Gaussianity) so that $\rho_{k,n} = \Theta(1)$, our bias and variance bounds would improve to $\tilde{\mathcal{O}}\left(n^{-\min(2r+a,2(1+a))}\right)$ and $\sigma_\epsilon^2\tilde{\mathcal{O}}(1)$ respectively, matching their bound up to a polylog factor. When $a \to 0$, the difference is of course very small, implying that one obtains nearly tempered overfitting in the fixed dimensional regime. Unfortunately, common kernels do not have Gaussian features and may suffer from poor concentration in the fixed d regime. Thus, a polylog factor in the bounds is likely inevitable. This explains the observation in Fig. 2, showing that upper bounds that assume Gaussian features may be over-optimistic for common kernels.

5.3. Regularized Regression

A major benefit to our approach is that we can provide bounds for both the regularized and unregularized cases with the same tools. We can thus derive bounds for the classical setup where the regularization γ_n is relatively large.

Theorem 5.3. *Let K be a kernel with polynomially decaying eigenvalues $\lambda_i = \Theta_{i,n}(i^{-1-a})$ for some $a > 0$, and assume that $\beta_k = \mathcal{O}_k(1)$. Further, suppose that the regularization parameter satisfies $\gamma_n = \Theta_n(n^{-1-b})$ for*

$b \in (-1, a)$. *Then for any* $\delta > 0$, *it holds w.p at least* $1 - \delta - o_n(\frac{1}{n})$ *that*

$$V \leq \sigma_\epsilon^2 \cdot \mathcal{O}_n\left(\frac{1}{n^{\frac{a-b}{1+a}}}\right),$$

and if $\theta_i^* = \Theta_{i,n}\left(i^{-r}\right)$ *for some* $r \in \mathbb{R}$ *s.t* $\left\|\Sigma^{1/2}\theta^*\right\|_2 < \infty$ *(necessary for* $f^* \in L_\mu^2(\mathcal{X})$*), then under the same probability it also holds that*

$$B \leq \frac{1}{\delta} \cdot \mathcal{O}_n\left(\frac{1}{n^{(1+b)\min\left(\frac{(2r+a)}{1+a}, 2\right)}}\right),$$

where the $\mathcal{O}$ *is weakened to* $\tilde{\mathcal{O}}$ *if* $r = 1 + \frac{a}{2}$.

The conditions of Thm. 5.3 are very mild, and do not require any control of α_k. Particularly, the kernels mentioned in the previous chapter all satisfy the assumptions here. One can observe a bias-variance tradeoff, where the variance bound improves with increased regularization (smaller b), and the bias bound worsens. Regardless, the excess risk always tends to 0 as $n \to \infty$. The choice of polynomial decay was arbitrary, and bounds for other decays can easily be obtained by modifying the proof.

The result recovers those of Cui et al. (2021) who worked under the heavy Gaussian feature assumption, and Li et al. (2023b) who worked under a Hölder continuity assumption on the kernel as well as an assumption relating to what they called an embedding index. Caponnetto & De Vito (2007) only provide upper bounds for the optimal γ_n, and do not decompose into bias and variance.

5.4. Implications for Neural Networks

At a high level, under suitable initialization and learning rate, training a sufficiently wide neural network for time t with gradient flow is roughly equivalent to kernel regression with the NTK and regularization $\gamma_n = \frac{1}{t}$ (see Appendix K for more details). So by Thm. 5.3, if the eigenvalues of the NTK decay as $\lambda_i = \Theta_{i,n}(\frac{1}{i^{-1-a}})$ and the target function satisfies $\theta_i^* = \Theta_{i,n}\left(i^{-r}\right)$, then as the width of the corresponding network tends towards infinity, the bias and variance after training for time $t := \Theta_n\left(n^s\right)$ for $s \in (0, 1+a)$ approach

$$V \leq \sigma_\epsilon^2 \cdot \mathcal{O}_n\left(\frac{1}{n^{1-\frac{s}{1+a}}}\right), \quad B \leq \mathcal{O}_n\left(\frac{1}{n^{s\min\left(\frac{2r+a}{1+a}, 2\right)}}\right). \tag{15}$$

Neural networks of various architectures exhibit polynomially decaying eigenvalues in the fixed dimensional regime, including fully-connected networks, CNNs, and ResNets (Bietti & Bach, 2020; Geifman et al., 2022; Barzilai et al., 2022). Interestingly, skip connections do not affect the asymptotic rate of decay of the NTK eigenvalues (Barzilai et al., 2022; Belfer et al., 2021) and as a result, ResNets obtain the same rates in Eq. (15) as their non-residual counterparts (i.e if one removes the skip connection).

Similarly, the applications of Thm. 5.2 and Thm. 5.1 to networks that are instead trained to completion (i.e in the $t \to \infty$ limit) are immediate. In particular, one has nearly tempered overfitting in the fixed dimensional regime, and in the high dimensional regime of $\frac{d^\tau}{n} = \Theta(1)$, if f^* consists of frequencies of at most $\lceil \tau \rceil$, then overfitting is benign.

6. Discussion

We studied the eigenvalues of kernel matrices and the generalization properties of KRR under general assumptions and applied these to several common settings. In relation to the rich line of prior works, our main hope is that our work can provide general-purpose tools that could be used in a wide range of future works. In particular, Thm. 3.1 and Thm. 4.1 are stated in a way that can provide bounds in various settings, as demonstrated in Sec. 5.

One direction that we did not fully explore in Sec. 5 is analysis for more general input distributions, beyond a uniform distribution on the sphere. This is because understanding the spectral decomposition of common kernels under general distributions is still an ongoing research direction. Nevertheless, our main theorems (Thm. 3.1, Thm. 4.1) are stated in a general way so that they could be applied to more general distributions in the future.

Another interesting direction for future work is proving corresponding lower bounds, particularly when the eigenvalues decay quickly and the self-induced regularization in the kernel is very small. Such analysis would aid in understanding whether, in some kernels, the variance term increases very slowly, as seen in Fig. 2.

Acknowledgements

This research is supported in part by European Research Council (ERC) grant 754705, by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center and by research grants from the Estate of Harry Schutzman and the Anita James Rosen Foundation.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1370–1378. PMLR, 2019.

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.

Atkinson, K. and Han, W. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

Azevedo, D. and Menegatto, V. A. Eigenvalues of dot-product kernels on the sphere. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 3(1), 2015.

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

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., Geifman, A., Galun, M., and Basri, R. A kernel perspective of skip connections in convolutional networks. *arXiv preprint arXiv:2211.14810*, 2022.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

Belfer, Y., Geifman, A., Galun, M., and Basri, R. Spectral analysis of the neural tangent kernel for deep residual networks. *arXiv preprint arXiv:2104.03093*, 2021.

Belkin, M. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pp. 1348–1361. PMLR, 2018.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Bietti, A. and Bach, F. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020.

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.

Bowman, B. and Montufar, G. F. Spectral bias outside the training set for deep networks in the kernel regime. *Advances in Neural Information Processing Systems*, 35: 30362–30377, 2022.

Braun, M. L. *Spectral properties of the kernel matrix and their relation to kernel methods in machine learning*. PhD thesis, Universitäts-und Landesbibliothek Bonn, 2005.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., and Gu, Q. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Caponnetto, A. and Vito, E. D. Fast rates for regularized least-squares algorithm. 2005.

Chen, L. and Xu, S. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020.

Cheng, C. and Montanari, A. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

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. *arXiv preprint arXiv:2310.00987*, 2023.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Dai, F. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

Dancis, J. A quantitative formulation of sylvester's law of inertia. iii. *Linear Algebra and its Applications*, 80: 141–158, 1986.

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

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Fan, Z. and Wang, Z. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33: 7710–7721, 2020.

Fasshauer, G. E. Positive definite kernels: past, present and future. *Dolomites Research Notes on Approximation*, 4: 21–63, 2011.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Geifman, A., Galun, M., Jacobs, D., and Ronen, B. On the spectral bias of convolutional neural tangent and gaussian process kernels. *Advances in Neural Information Processing Systems*, 35:11253–11265, 2022.

Geifman, A., Barzilai, D., Basri, R., and Galun, M. Controlling the inductive bias of wide neural networks by modifying the kernel's spectrum. *arXiv preprint arXiv:2307.14531*, 2023.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. 2021.

Haas, M., Holzmüller, D., von Luxburg, U., and Steinwart, I. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. *arXiv preprint arXiv:2305.14077*, 2023.

Han, X. Spherical harmonics with maximal l p ($2 < p \leq 6$) norm growth. *The Journal of Geometric Analysis*, 26(1): 378–398, 2016.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

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

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

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. Kernel alignment risk estimator: Risk prediction from training data. *Advances in neural information processing systems*, 33:15568–15578, 2020.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Li, Y., Yu, Z., Chen, G., and Lin, Q. Statistical optimality of deep wide neural networks. *arXiv preprint arXiv:2305.02657*, 2023a.

Li, Y., Zhang, H., and Lin, Q. On the asymptotic learning curves of kernel ridge regression under power-law decay. *arXiv preprint arXiv:2309.13337*, 2023b.

Li, Z., Zhou, Z.-H., and Gretton, A. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. 2020.

Lin, J., Rudi, A., Rosasco, L., and Cevher, V. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Mairal, J. and Vert, J.-P. Machine learning with kernel methods.

Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., and Nakkiran, P. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 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*, 59:3–84, 2022.

Minh, H. Q., Niyogi, P., and Yao, Y. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pp. 154–168. Springer, 2006.

Misiakiewicz, T. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

Montanari, A. and Zhong, Y. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

Nguyen, Q., Mondelli, M., and Montufar, G. F. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pp. 8119–8129. PMLR, 2021.

O'Donnell, R. *Analysis of boolean functions*. Cambridge University Press, 2014.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Rakhlin, A. and Zhai, X. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pp. 2595–2623. PMLR, 2019.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pp. 3889–3897. PMLR, 2021.

Rosasco, L., Belkin, M., and De Vito, E. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010.

Scetbon, M. and Harchaoui, Z. A spectral analysis of dot-product kernels. In *International conference on artificial intelligence and statistics*, pp. 3394–3402. PMLR, 2021.

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. *arXiv preprint arXiv:2110.03922*, 2021.

Smola, A., Ovári, Z., and Williamson, R. C. Regularization with dot-product kernels. *Advances in neural information processing systems*, 13, 2000.

Spigler, S., Geiger, M., and Wyart, M. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

Steinwart, I., Hush, D., and Scovel, C. An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

Steinwart, I., Hush, D. R., Scovel, C., et al. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

Szeg, G. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

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

Valdivia, E. A. Relative concentration bounds for the spectrum of kernel matrices. *arXiv preprint arXiv:1812.02108*, 2018.

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

Wang, Z. and Zhu, Y. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *arXiv preprint arXiv:2109.09304*, 2021.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Xiao, L. and Pennington, J. Precise learning curves and higher-order scaling limits for dot product kernel regression. *arXiv preprint arXiv:2205.14846*, 2022.

Yang, G. and Littwin, E. Tensor programs iib: Architectural universality of neural tangent kernel training dynamics. In *International Conference on Machine Learning*, pp. 11762–11772. PMLR, 2021.

Yang, G. and Salman, H. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A. More Notations

We introduce a few more notations for the appendix, which are not needed in the main text. We let $A_k := \mathbf{K}_{>k} + n\gamma_n I$ and $A := \mathbf{K} + n\gamma_n I$. Additionally, for any $k \leq k' \in \mathbb{N}$ we denote by $k : k'$ the $k, \ldots, k'$ indices, so that, for example, $\phi_{k:k'}(X) = (\phi_k(X), \ldots, \phi_{k'}(X)) \in \mathbb{R}^{n\times(k'-k+1)}$.

B. Concentration Bounds

Lemma B.1. *Let $k \in [n]$, then each of the following holds w.p at least $1 - 2\exp\left(-\frac{1}{2\beta_k^2}n\right)$:*

1. $\frac{1}{2}n\sum_{i>k}\lambda_i^2 \leq tr\left(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top\right) \leq \frac{3}{2}n\sum_{i>k}\lambda_i^2$
2. $\frac{1}{2}kn \leq tr\left(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top\right) \leq \frac{3}{2}kn.$

Proof. For (1), first observe that

$$\begin{aligned}&\text{tr}\left(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top\right)\\ =&\sum_{j=1}^n\left[\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top\right]_{jj} = \sum_{j=1}^n \phi_{>k}(\mathbf{x}_j)^\top\Sigma_{>k}\phi_{>k}(\mathbf{x}_j)\\ =&\sum_{j=1}^n\sum_{i>k}\lambda_i^2\psi_i(\mathbf{x}_j)^2.\end{aligned}$$

We will now show that the conditions for Hoeffding's inequality hold. Let $v_j = \sum_{i>k}\lambda_i^2\psi_i(\mathbf{x}_j)^2$ and $M := \beta_k\sum_{i>k}\lambda_i^2$. By the definition of β_k Eq. (10), we have that for every j, $0 \leq v_j \leq M$. Furthermore, $\mathbb{E}[\sum_{j=1}^n v_j] = n\sum_{i>k}\lambda_i^2$ and so Hoeffding's inequality yields:

$$\mathbb{P}\left(\left|\text{tr}\left(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top\right) - n\sum_{i>k}\lambda_i^2\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{nM^2}\right).$$

Substituting $t = \frac{n}{2}\sum_{i>k}\lambda_i^2$, it holds that w.p at least $1 - 2\exp\left(-\frac{1}{2\beta_k^2}n\right)$,

$$\frac{1}{2}n\sum_{i>k}\lambda_i^2 \leq \text{tr}\left(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top\right) \leq \frac{3}{2}n\sum_{i>k}\lambda_i^2.$$

For (2), the proof is analogous:

$$\begin{aligned}&\text{tr}\left(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top\right)\\ =&\sum_{j=1}^n\left[\psi_{\leq k}(X)\psi_{\leq k}(X)^\top\right]_{jj} = \sum_{j=1}^n\psi_{\leq k}(\mathbf{x}_j)^\top\psi_{\leq k}(\mathbf{x}_j)\\ =&\sum_{j=1}^n\sum_{i=1}^k\psi_i(\mathbf{x}_j)^2 \leq \beta_k kn\end{aligned}$$

Now letting $M' = \beta_k k$ using Hoeffding as before yields

$$\mathbb{P}\left(\left|\text{tr}\left(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top\right) - kn\right| \geq t'\right) \leq 2\exp\left(-\frac{2t'^2}{nM'^2}\right).$$

So picking $t' = \frac{nk}{2}$ we get that w.p at least $1 - 2\exp\left(-n\cdot\frac{1}{2\beta_k^2}\right)$

$$\frac{1}{2}kn \leq \text{tr}\left(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top\right) \leq \frac{3}{2}kn.$$

□

Lemma B.2. *For any $k \in [n]$ there exist some absolute constants $c', c_2 > 0$, s.t the following hold simultaneously w.p at least $1 - 2\exp\left(-\frac{c'}{\beta_k}\max\left(\frac{n}{k}, \log(k)\right)\right)$*

1. $\mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right) \geq \max\left(\sqrt{n} - \sqrt{\frac{1}{2}\max\left(n, \beta_k\left(1+\frac{1}{c'}\right)k\log(k)\right)}\ ,\ 0\right)^2,$
2. $\mu_1\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right) \leq c_2 \max\left(n, \beta_k k\log(k)\right).$

Moreover, there exists some $c > 0$ s.t if $c\beta_k k\log(k) \leq n$ then w.p at least $1 - 2\exp\left(-\frac{c'}{\beta_k}\frac{n}{k}\right)$ and some absolute constant $c_1 > 0$, it holds that

$$c_1 n \leq \mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right) \leq \mu_1\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right) \leq c_2 n.$$

Proof. We will bound the singular values $\sigma_i\left(\psi_{\leq k}(X)\right)$ since

$$\sigma_i(\psi_{\leq k}(X))^2 = \mu_i\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right).$$

$\psi_{\leq k}(X)$ is an $n \times k$ matrix, whose rows $\psi_{\leq k}(\mathbf{x}_j)$ are independent isotropic random vectors in $\mathbb{R}^k$ (where the randomness is over the choice of $\mathbf{x}_j$). Furthermore, by the definition of β_k Eq. (10), for a.s every $\mathbf{x}_i$, $\|\psi_{\leq k}(\mathbf{x}_i)\| \leq \sqrt{\beta_k k}$. As such, from Vershynin (2010)[Theorem 5.41], there is some absolute constant $c' > 0$ s.t for every $t \geq 0$, one has that with probability at least $1 - 2k\exp(-2c't^2)$,

$$\sqrt{n} - t\sqrt{\beta_k k} \leq \sigma_k(\psi_{\leq k}(X)) \leq \sigma_1(\psi_{\leq k}(X)) \leq \sqrt{n} + t\sqrt{\beta_k k}.$$

Now for $t = \sqrt{\frac{1}{2\beta_k}\max\left(\frac{n}{k}, \log(k)\right) + \frac{\log(k)}{2c'}}$ we get that with probability at least $1 - 2\exp\left(-\frac{c'}{\beta_k}\max\left(\frac{n}{k}, \log(k)\right)\right)$ it holds that

$$\begin{aligned}\sigma_1(\psi_{\leq k}(X))^2 \leq& \left(\sqrt{n} + \sqrt{\frac{1}{2}\max\left(n, k\log(k)\right) + k\log(k)\frac{\beta_k}{2c'}}\right)^2\\ \leq& \left(\sqrt{n} + \frac{1}{\sqrt{2}}\sqrt{n + \left(1 + \frac{\beta_k}{c'}\right)k\log(k)}\right)^2\\ \leq& 3n + \left(1 + \frac{\beta_k}{c'}\right)k\log(k),\end{aligned}$$

where the last equality followed from the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$. Because, $\beta_k \geq 1$ Eq. (54), we obtain $\sigma_1(\psi_{\leq k}(X))^2 \leq c_2\max\left(n, \beta_k k\log(k)\right)$ for a suitable $c_2 > 0$, proving point Eq. (2). For the lower bound, we simultaneously have

$$\begin{aligned}\sigma_k(\psi_{\leq k}(X)) \geq& \sqrt{n} - \frac{1}{\sqrt{2}}\sqrt{\frac{1}{2}\max\left(n, k\log(k)\right) + k\log(k)\frac{\beta_k}{2c'}}\\ \geq& \sqrt{n} - \sqrt{\frac{1}{2}\max\left(n, \beta_k\left(1 + \frac{1}{c'}\right)k\log(k)\right)},\end{aligned}$$

Since the singular values are non-negative, the above implies

$$\sigma_k(\psi_{\leq k}(X))^2 \geq \max\left(\sqrt{n} - \sqrt{\frac{1}{2}\max\left(n, \beta_k\left(1 + \frac{1}{c'}\right)k\log(k)\right)}\ ,\ 0\right)^2.$$

proving point Eq. (1).

For the moreover part, taking $c = \left(1 + \frac{1}{c'}\right)$, we now have by assumption that $\frac{n}{k} \geq c\beta_k\log(k) \geq \log(k)$ (where we used the facts that $c \geq 1$ and $\beta_k \geq 1$), the probability that Eq. (1) and Eq. (2) hold is in fact $1 - 2\exp\left(-\frac{c'}{\beta_k}\frac{n}{k}\right)$.

Furthermore, plugging $c\beta_k k\log(k) \leq n$ into the lower bound Eq. (1) yields

$$\begin{aligned}\mu_k\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right) &\geq \max\left(\sqrt{n}-\sqrt{\frac{1}{2}\max\left(n, c\beta_k k\log(k)\right)}\ ,\ 0\right)^2.\\ &\geq\left(\sqrt{n}-\sqrt{\frac{n}{2}}\right)^2=\left(1-\frac{1}{\sqrt{2}}\right)^2 n.\end{aligned}$$

Similarly, since $\beta_k k\log(k) \leq n$ the upper bound Eq. (2) becomes

$$\mu_1\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right) \leq c_2 n$$

□

Lemma B.3. *For any $k \in [n]$ and $\delta > 0$, it holds w.p at least $1-\delta$ that*

$$\left\|\phi_{>k}(X)\theta^*_{>k}\right\|^2 \leq \frac{1}{\delta}n\left\|\theta^*_{>k}\right\|^2_{\Sigma_{>k}}$$

Proof. Let $v_j = \langle\phi_{>k}(\mathbf{x}_j), \theta^*_{>k}\rangle^2$ so that $\left\|\phi_{>k}(X)\theta^*_{>k}\right\|^2 = \sum_{j=1}^n v_j$. Since $\mathbf{x}_j$ are independent, it holds that v_j are independent random variables with mean:

$$\begin{aligned}\mathbb{E}[v_j] =&\mathbb{E}\left[\left(\sum_{i>k}\sqrt{\lambda_i}\psi_i(\mathbf{x}_j)\theta^*_i\right)^2\right]\\ =&\sum_{i>k}\sum_{>l}\sqrt{\lambda_i}\sqrt{\lambda_l}\theta^*_i\theta^*_l\underbrace{\mathbb{E}_{\mathbf{x}_j}\left[\psi_i(\mathbf{x}_j)\psi_l(\mathbf{x}_j)\right]}_{\delta_{il}}\\ =&\sum_{i>k}\lambda_i(\theta^*_i)^2=\left\|\theta^*\right\|^2_{\Sigma_{>k}}.\end{aligned}$$

So by Markov's inequality:

$$\mathbb{P}\left(\sum_{j=1}^n v_j \geq \frac{1}{\delta}n\left\|\theta^*_{>k}\right\|^2_{\Sigma_{>k}}\right) \leq \delta.$$

□

Lemma B.4. *There exists some absolute constants $c, c', c_1, c_2 > 0$ s.t for any $k \in \mathbb{N}$ with $c\beta_k k\log(k) \leq n$, it holds w.p at least $1-8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ that all of the following hold simultaneously:*

1. $c_1 n\sum_{i>k}\lambda_i^2 \leq$ *tr* $\left(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top\right) \leq c_2 n\sum_{i>k}\lambda_i^2$
2. $c_1 kn \leq$ *tr* $\left(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top\right) \leq c_2 kn$
3. $\mu_k\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right) \geq c_1 n$
4. $\mu_1\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right) \leq c_2 n$

Proof. By Lemma B.1, points (1) and (2) each hold w.p at least $1-2\exp\left(-\frac{1}{2\beta_k^2}n\right)$ so they both hold w.p at least $\left(1-2\exp\left(-\frac{1}{2\beta_k^2}n\right)\right)^2$.

Furthermore, the "moreover" part of Lemma B.2 states that points (3) and (4) hold simultaneously w.p at least $1-2\exp\left(-\frac{c'}{\beta_k}\frac{n}{k}\right)$.

Now the probability for which (1)-(4) all hold simultaneously is at least

$$\left(1-2\exp\left(-\frac{1}{2\beta_k^2}n\right)\right)^2\left(1-2\exp\left(-\frac{c'}{\beta_k}\frac{n}{k}\right)\right)$$
$$\geq 1-8\exp\left(-\min\left(\frac{1}{2\beta_k^2}n,\frac{c'}{\beta_k}\frac{n}{k}\right)\right)\geq 1-8\exp\left(-\min\left(\frac{1}{2\beta_k^2},\frac{c'}{\beta_k}\right)\frac{n}{k}\right)$$

Since $\beta_k \geq 1$ Eq. (54) replacing c' with $\min(\frac{1}{2},c')$ results in the desired bounds holding w.p at least $1-8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$.

□

C. Bounds on the Eigenvalues of Kernel Matrices - Proofs of Results in Sec. 3

C.1. Proof of Thm. 3.1

Theorem 3.1. *Suppose Assumption 2.3 holds, and that the eigenvalues of Σ are given in non-increasing order $\lambda_1 \geq \lambda_2 \geq \ldots$. There exist some absolute constants $c, C, c_1, c_2 > 0$ s.t for any $k \leq k' \in [n]$ and $\delta > 0$, w.p at least $1-\delta-4\frac{r_k}{k^4}\exp\left(-\frac{c}{\beta_k}\frac{n}{r_k}\right)-2\exp\left(-\frac{c}{\beta_k}\max\left(\frac{n}{k},\log(k)\right)\right)$,*

$$\frac{1}{c_2\beta_k}\mu_k\left(\frac{1}{n}\mathbf{K}\right)\leq\left(1+\frac{k\log(k)}{n}\right)\lambda_k+\log(k+1)\frac{tr(\Sigma_{>k})}{n},$$

and

$$\mu_k\left(\frac{1}{n}\mathbf{K}\right)\geq c_1\mathbb{I}_{k,n}\lambda_k+\alpha_k\left(1-\frac{1}{\delta}\sqrt{\frac{n^2}{R_{k'}}}\right)\frac{tr\left(\Sigma_{>k'}\right)}{n},$$

where $\mathbb{I}_{k,n}=\begin{cases}1, & \textit{if } C\beta_k k\log(k)\leq n\\ 0, & \textit{otherwise}\end{cases}$.

Proof. From Lemma C.2, we have that

$$\lambda_k\mu_k\left(D_k\right)+\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right)\leq\mu_k\left(\frac{1}{n}\mathbf{K}\right)\leq\lambda_k\mu_1\left(D_k\right)+\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right),\tag{16}$$

where D_i is as in the formulation of the lemma.

We bound each of the summands in the upper bound separately. From Corollary C.5, it holds w.p at least $1-4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right)$ that for some absolute constants $c', c_2' > 0$,

$$\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right)\leq c_2'\left(\lambda_{k+1}+\beta_k\log(k+1)\frac{\text{tr}\left(\Sigma_{>k}\right)}{n}\right).$$

For the other summand, since $D_i=\frac{1}{n}\psi_{\leq k}(X)^\top\psi_{\leq k}(X)$ Lemma B.2 states that there exists some absolute constants $c'', c_2'' > 0$, s.t w.p at least $1-2\exp\left(-\frac{c''}{\beta_k}\max\left(\frac{n}{k},\log(k)\right)\right)$

$$\lambda_k\mu_1\left(D_i\right)\leq c_2''\frac{1}{n}\max\left(n,\beta_k k\log(k)\right)\lambda_k\leq c_2''\beta_k\left(1+\frac{k\log(k)}{n}\right)\lambda_k,$$

where in the last inequality we used the fact that $\beta_k \geq 1$. So taking $c=\max(c',c'')$, both events hold w.p at least $1-4\frac{r_k}{k^4}\exp\left(-\frac{c}{\beta_k}\frac{n}{r_k}\right)-2\exp(-\frac{c}{\beta_k}\max\left(\frac{n}{k},\log(k)\right))$ and the upper bound from Eq. (16) yields

$$\mu_k\left(\frac{1}{n}\mathbf{K}\right)\leq c_2\beta_k\left(\left(1+\frac{k\log(k)}{n}\right)\lambda_k+\log(k+1)\frac{\text{tr}(\Sigma_{>k})}{n}\right),$$

for some suitable absolute constant $c_2 > 0$. The "moreover" part of this proof analogously follows from the "moreover" part of Lemma B.2, which states that $\mu_k(D_k)\geq c_1$ if $C\beta_k k\log(k)\leq n$, and from the lower bound of Corollary C.5, which holds w.p at least $1-\delta$. □

C.2. Lemmas and Alternative Results for Eigenvalue Bounds

We now provide an extension of Ostrowski's theorem to non-square matrices. Note that the case of $k \leq n$ is relatively easy. However, we also prove the case of $k > n$.

Lemma C.1 (Extension of Ostrowski's Theorem)**.** *Let $i, k \in \mathbb{N}$ satisfy $1 \leq i \leq \min(k, n)$ and $D_k := \frac{1}{n}\psi_{\leq k}(X)\psi_{\leq k}(X)^\top \in \mathbb{R}^{n\times n}$. Suppose that the eigenvalues of Σ are given in non-increasing order $\lambda_1 \geq \lambda_2 \geq \dots$ then*

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) \leq \mu_i\left(\frac{1}{n}\mathbf{K}_{\leq k}\right) \leq \lambda_i\mu_1(D_k).$$

Proof. Let π_1 denote the number of positive eigenvalues of $\frac{1}{n}\mathbf{K}_{\leq k}$ (where in particular $\pi_1 \leq \min(n,k)$). Because the kernel can be decomposed as $\mathbf{K}_{\leq k} = \psi_{\leq k}(X)\Sigma_{\leq k}\psi_{\leq k}(X)^\top$, it follows from Dancis (1986)[Theorem 1.5] that for $1 \leq i \leq \pi_1$,

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) \leq \mu_i\left(\frac{1}{n}\mathbf{K}_{\leq k}\right) \leq \lambda_i\mu_1(D_k).$$

It remains to handle the case where $\pi_1 < i$ (where in particular this means $\pi_1 < \min(n,k)$). By definition of π_1 there are some orthonormal eigenvectors of $\mathbf{K}_{\leq k}$, $v_{\pi_1+1}, \dots, v_n$ with eigenvalues 0. Since $\Sigma \succ 0$, for each such 0 eigenvector v,

$$0 = \left(\psi_{\leq k}^\top(X)v\right)^\top \Sigma\left(\psi_{\leq k}^\top(X)v\right) \implies \psi_{\leq k}^\top(X)v = 0.$$

In particular, D_k has $v_{\pi_1+1}, \dots, v_n$ as 0 eigenvectors and since $D_k \succeq 0$, we obtain that $\mu_{\pi_1+1}(D_k), \dots, \mu_n(D_k) = 0$. So for $i > \pi_1$ we have

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) = 0 = \mu_i\left(\frac{1}{n}\mathbf{K}_{\leq k}\right) \leq \lambda_i\mu_1(D_k).$$

□

Lemma C.2. *Let $i, k \in \mathbb{N}$ satisfy $1 \leq i \leq n$ and $i \leq k$, let $D_k := \frac{1}{n}\psi_{\leq k}(X)\psi_{\leq k}(X)^\top \in \mathbb{R}^{n\times n}$. that the eigenvalues of Σ are given in non-increasing order $\lambda_1 \geq \lambda_2 \geq \dots$ then*

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) + \mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq \mu_i\left(\frac{1}{n}\mathbf{K}\right) \leq \lambda_i\mu_1\left(D_k\right) + \mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right).$$

In particular,

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) \leq \mu_i\left(\frac{1}{n}\mathbf{K}\right) \leq \lambda_i\mu_1\left(D_k\right) + \mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right).$$

Proof. We can decompose $\mathbf{K}$ into the sum of two hermitian matrices by $\mathbf{K} = \mathbf{K}_{\leq k} + \mathbf{K}_{>k}$. By Weyl's theorem (Horn & Johnson, 2012)[Corollary 4.3.15] we can use this decomposition to bound the eigenvalues of $\mathbf{K}$ as:

$$\mu_i\left(\mathbf{K}_{\leq k}\right) + \mu_n\left(\mathbf{K}_{>k}\right) \leq \mu_i(\mathbf{K}) \leq \mu_i\left(\mathbf{K}_{\leq k}\right) + \mu_1\left(\mathbf{K}_{>k}\right). \tag{17}$$

Further, since $\mathbf{K}_{\leq k} = \psi_{\leq k}(X)\Sigma_{\leq k}\psi_{\leq k}(X)^\top$, we use an extension of Ostrowski's theorem, Lemma C.1, to obtain the bound:

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) \leq \mu_i\left(\frac{1}{n}\mathbf{K}_{\leq k}\right) \leq \lambda_i\mu_1(D_k). \tag{18}$$

So combining the two results yields the bounds:

$$\lambda_{i+k-\min(n,k)}\mu_{\min(n,k)}(D_k) + \mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq \mu_i\left(\frac{1}{n}\mathbf{K}\right) \leq \lambda_i\mu_1(D_k) + \mu_1\left(\frac{1}{n}\phi_{>k}(X)\phi_{>k}(X)^\top\right).$$

The "in particular part" now follows from $\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) \geq 0$. □

Lemma C.3. *Suppose Assumption 2.3 holds, and that the eigenvalues of Σ are given in non-increasing order $\lambda_1 \geq \lambda_2 \geq \dots$. Let $k \in \mathbb{N}$ and let r_k be as defined in Def. (8). There exist absolute constant $c, c' > 0$ s.t it holds w.p at least $1 - 4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right)$ that*

$$\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq c\left(\lambda_{k+1} + \beta_k \log(k+1)\frac{tr\left(\Sigma_{>k}\right)}{n}\right).$$

Proof. Let $E_k = \mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right)$, $\hat{\Sigma}_{>k} := \frac{1}{n}\phi_{>k}(X)^\top\phi_{>k}(X)$ and observe that $E_k = \left\|\hat{\Sigma}_{>k}\right\|$. We would ideally like to bound $\left\|\hat{\Sigma}_{>k}\right\|$ using the matrix Chernoff inequality with intrinsic dimension (Tropp et al., 2015)[Theorem 7.2.1]. However, as this inequality was proved for finite matrices, if the dimension of the features is $p = \infty$ we first approximate $\left\|\hat{\Sigma}_{>k}\right\|$, letting $\phi_{k+1:p'}(X) := (\phi_{k+1}(X), \dots, \phi_p'(X))$ for some $p' \in \mathbb{N}$ and $\hat{\Sigma}_{k+1:p'} := \frac{1}{n}\phi_{k+1:p'}(X)^\top\phi_{k+1:p'}(X)$, then E_k can be bounded as:

$$E_k = \left\|\frac{1}{n}\mathbf{K}_{k+1:p'} + \frac{1}{n}\mathbf{K}_{\geq p'}\right\| \leq \left\|\frac{1}{n}\mathbf{K}_{k+1:p'}\right\| + \left\|\frac{1}{n}\mathbf{K}_{\geq p'}\right\| = \left\|\hat{\Sigma}_{k+1:p'}\right\| + E_{p'}. \tag{19}$$

Furthermore, $E_{p'}$ can be bounded as

$$E_{p'} \leq \frac{1}{n}\text{tr}\left(\mathbf{K}_{>p'}\right) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i>p'}\lambda_i\psi_i(\mathbf{x}_j)^2 \leq \beta_{p'}\sum_{i>p'+1}\lambda_i = \beta_{p'}\text{tr}(\Sigma_{>p'}). \tag{20}$$

So, to summarize, either p is finite, in which case we can take $p' = p$ and $E_{p'} = 0$, or p is infinite, in which case $E_{p'} \leq \beta_{p'}\text{tr}(\Sigma_{>p'})$. However, by Assumption 2.3 this implies:

$$\forall u > 0, \exists p' \in \mathbb{N} \text{ s.t. } E_{p'} \leq u. \tag{21}$$

Let $Z_j^{p'} = \frac{1}{n}\phi_{k+1:p'}(\mathbf{x}_j)\phi_{k+1:p'}(\mathbf{x}_j)^\top$ (where $(Z_j^{p'}) \succeq 0$) so that we can decompose the empirical covariance as a sum $\hat{\Sigma}_{k+1:p'} = \sum_{j=1}^{n} Z_j^{p'}$. We will need a bound on both $\mu_1(Z_j^{p'})$ and $\mu_1(\mathbb{E}\hat{\Sigma}_{k+1:p'})$. For the first, we have

$$\mu_1(Z_j^{p'}) = \frac{1}{n}\sum_{i=k+1}^{p'}\lambda_i\psi_i(\mathbf{x}_j)^2 \leq \frac{1}{n}\sum_{i=k+1}^{\infty}\lambda_i\psi_i(\mathbf{x}_j)^2 \leq \underbrace{\frac{\beta_k}{n}\text{tr}(\Sigma_{>k})}_{:=L},$$

where we denote by L the right-hand side. For the bound on $\mu_1(\mathbb{E}\hat{\Sigma}_{k+1:p'})$, it holds that $\mathbb{E}\hat{\Sigma}_{k+1:p'} = \Sigma_{k+1:p'} = \text{diag}(\lambda_{k+1} + 1, \dots, \lambda_p')$ and thus $\mu_1(\mathbb{E}\hat{\Sigma}_{k+1:p'}) = \lambda_{k+1}$.

We have shown that the conditions of Tropp et al. (2015)[Theorem 7.2.1] are satisfied. As such, for $r_{k:p'} := \frac{\text{tr}(\Sigma_{k+1:p'})}{\lambda_{k+1}}$ and any $t \geq 1 + {}^{L}\!/\!{}_{\lambda_{k+1}} = 1 + \frac{\beta_k r_k}{n}$,

$$\mathbb{P}\left(\left\|\hat{\Sigma}_{k+1:p'}\right\| \geq t\lambda_{k+1}\right) \leq 2r_{k:p'}\left(\frac{e^{t-1}}{t^t}\right)^{\lambda_{k+1}/L}.$$

By Eq. (19) it holds that $\left\|\hat{\Sigma}_{k+1:p'}\right\| \geq E_k - E_{p'}$. Using this, the fact that $\frac{\lambda_{k+1}}{L} = \frac{n}{\beta_k r_k}$, and upper bounding $e^{t-1} \leq e^t$, $r_{k:p'} \leq r_k$ yields

$$\mathbb{P}\left(E_k - E_{p'} \geq t\lambda_{k+1}\right) \leq \mathbb{P}\left(\left\|\hat{\Sigma}_{k+1:p'}\right\| \geq t\lambda_{k+1}\right) \leq 2r_k\left(\frac{e}{t}\right)^{tn/\beta_k r_k}.$$

Now we pick $t = e^3 + 2\frac{\beta_k r_k}{n}\log(k+1)$, (which satisfies the requirement of $t \geq 1 + \frac{\beta_k r_k}{n}$). In particular $\frac{e}{t} \leq \frac{1}{e^2}$, and we obtain that:

$$\begin{aligned}\mathbb{P}\left(E_k \geq t\lambda_{k+1} + E_{p'}\right) \leq& 2r_k\left(\frac{1}{e^2}\right)^{\frac{e^3}{\beta_k}\frac{n}{r_k} + 2\log(k+1)}\\ \leq& 2\frac{r_k}{(k+1)^4}\exp\left(-2\frac{e^3}{\beta_k}\frac{n}{r_k}\right).\end{aligned}$$

Furthermore, $E_{p'}$ can be bounded via Eq. (20)As a result, we obtain that for $c' = 2e^3$, $c = e^3$, it holds w.p at least $1 - 4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right)$ that

$$E_k \leq c\left(\lambda_{k+1} + \beta_k \log(k+1)\frac{\text{tr}\left(\Sigma_{>k}\right)}{n} + E_{p'}\right).$$

Notice that the bound on E_k depends on p' only via $E_{p'}$. So by Eq. (21) we are done.

□

Lemma C.4. *Let R_k be as defined in Def. (8). For any $\delta > 0$ it holds w.p at least $1 - \delta$ that for all $1 \leq i \leq n$*

$$\alpha_k \frac{1}{n} tr\left(\Sigma_{>k}\right)\left(1 - \frac{1}{\delta}\sqrt{\frac{n^2}{R_k}}\right) \leq \mu_i\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq \beta_k \frac{1}{n} tr\left(\Sigma_{>k}\right)\left(1 + \frac{1}{\delta}\sqrt{\frac{n^2}{R_k}}\right).$$

Proof. Let $\Lambda_{>k} := \text{diag}(\frac{1}{n}\mathbf{K}_{>k}) \in \mathbb{R}^{n\times n}$ be equal to $\frac{1}{n}\mathbf{K}_{>k}$ on the diagonal and 0 elsewhere, and $\Delta_{>k} := \frac{1}{n}\mathbf{K}_{>k} - \Lambda_{>k}$ be the remainder. $\Lambda_{>k}$ is a diagonal matrix with the i'th value on the diagonal given by $[\Lambda_{>k}]_{ii} = \frac{1}{n}\sum_{\ell>k}\lambda_\ell \psi_\ell(\mathbf{x}_i)^2$. By Def. (9) of α_k and Def. (10) of β_k it holds that

$$\alpha_k \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right) \leq [\Lambda_{>k}]_{ii} \leq \beta_k \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right),$$

which together with the fact that $\Lambda_{>k}$ is diagonal implies

$$\alpha_k \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right) I \preceq \Lambda_{>k} \preceq \beta_k \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right) I. \tag{22}$$

As such, by Weyl's theorem (Horn & Johnson, 2012)[Corollary 4.3.15], we can bound the eigenvalues of $\frac{1}{n}\mathbf{K}_{>k}$ as

$$\alpha_k \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right) + \mu_n\left(\Delta_{>k}\right) \leq \mu_i\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq \beta_k \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right) + \mu_1\left(\Delta_{>k}\right). \tag{23}$$

So in order to bound the eigenvalues of $\frac{1}{n}K_{>k}$, it remains to bound the eigenvalues of $\Delta_{>k}$. We first bound the expectation using

$$\begin{aligned}\mathbb{E}[\|\Delta_{>k}\|] \leq& \mathbb{E}[\|\Delta_{>k}\|_F^2]^{\frac{1}{2}} = \sqrt{\sum_{i,j=1}^{n}\mathbb{E}\left[\left(\frac{1}{n}\langle\phi_{>k}(\mathbf{x}_i), \phi_{>k}(\mathbf{x}_j)\rangle\right)^2\right]} \\ =& \sqrt{\frac{n(n-1)}{n^2}\text{tr}\left(\Sigma_{>k}^2\right)} \leq \sqrt{\text{tr}\left(\Sigma_{>k}^2\right)} = \frac{1}{n}\text{tr}\left(\Sigma_{>k}\right)\sqrt{\frac{n^2}{R_k}}.\end{aligned}$$

By Markov's inequality, it holds that

$$\mathbb{P}\left(\|\Delta_{>k}\| \geq \frac{1}{\delta}\mathbb{E}[\|\Delta_{>k}\|]\right) \leq \delta.$$

Implying that with probability at least $1 - \delta$ it holds that

$$\|\Delta_{>k}\| \leq \frac{1}{\delta}\mathbb{E}[\|\Delta_{>k}\|] \leq \frac{1}{n\delta}\text{tr}\left(\Sigma_{>k}\right)\sqrt{\frac{n^2}{R_k}}.$$

Finally, plugging this back into Eq. (23) completes the proof.

□

Corollary C.5. *Suppose Assumption 2.3 holds, and that the eigenvalues of Σ are given in non-increasing order $\lambda_1 \geq \lambda_2 \geq \ldots$. Let $k \in \mathbb{N}$ and let r_k be as defined in Def. (8). There exist absolute constant $c, c' > 0$ s.t it holds w.p at least $1 - 4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right)$ that*

$$\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq c\left(\lambda_{k+1} + \beta_k \log(k+1)\frac{tr\left(\Sigma_{>k}\right)}{n}\right).$$

And for any $k' \in \mathbb{N}$ with $k' > k$, and any $\delta > 0$ it holds w.p at least $1 - \delta$ that

$$\alpha_k \left(1 - \frac{1}{\delta}\sqrt{\frac{n^2}{R_{k'}}}\right) \frac{tr\left(\Sigma_{>k'}\right)}{n} \leq \mu_n \left(\frac{1}{n}\mathbf{K}_{>k'}\right),$$

so that both statements hold w.p at least $1 - \delta - 4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right)$.

Proof. By Weyl's theorem (Horn & Johnson, 2012)[Corollary 4.3.15], for any $k' \geq k, \mu_n(\mathbf{K}_{\geq k}) \geq \mu_n(\mathbf{K}_{\geq k'}) + \mu_n(\mathbf{K}_{k:k'}) \geq \mu_n(\mathbf{K}_{\geq k'})$. So the lower bound comes from Lemma C.4 (with k') and the upper bound comes from Lemma C.3. □

D. Upper bounds for the Risk - Proofs of Results in Sec. 4

D.1. Proof of Thm. 4.1.

The majority of the work was done in lemmas B.4, B.3, D.5 and D.6. Here we essentially combine these results to obtain the desired bounds. Throughout the section, the notations of $A_k := \mathbf{K}_{>k} + n\gamma_n I$ and $A := \mathbf{K} + n\gamma_n I$ as defined in Sec. A will be very common.

Theorem 4.1. *Let $k \in \mathbb{N}$ and let $\rho_{k,n}$ be as defined in Eq. (11). There exists some absolute constants $c, c', C_1, C_2 > 0$ s.t if $c\beta_k k\log(k) \leq n$, then for every $\delta > 0$, it holds w.p at least $1 - \delta - 16\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ that both the variance and bias can be upper bounded as:*

$$V \leq C_1 \rho_{k,n}^2 \sigma_\epsilon^2 \left(\frac{k}{n} + \min\left(\frac{r_k\left(\Sigma^2\right)}{n}, \frac{n}{\alpha_k^2 R_k(\Sigma)}\right)\right), \tag{12}$$

$$\begin{aligned} B \leq & C_2 \rho_{k,n}^3 \left(\frac{1}{\delta}\left\|\theta^*_{>k}\right\|^2_{\Sigma_{>k}}\right. \\ & \left. + \left\|\theta^*_{\leq k}\right\|^2_{\Sigma^{-1}_{\leq k}}\left(\gamma_n + \frac{\beta_k tr\left(\Sigma_{>k}\right)}{n}\right)^2\right). \end{aligned} \tag{13}$$

Proof. The majority of the work is given by lemmas D.1 and D.2. We note a few properties which are immediate, from which the claim will follow:

$$\frac{\mu_1\left(\frac{1}{n}A_k\right)^2}{\mu_n\left(\frac{1}{n}A_k\right)^2} = \left(\frac{\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) + \gamma_n}{\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) + \gamma_n}\right)^2 \leq \rho_{k,n}^2. \tag{24}$$

$$\frac{\left\|\Sigma_{>k}\right\|}{\mu_n\left(\frac{1}{n}A_k\right)} \leq \rho_{k,n}. \tag{25}$$

$$\frac{1}{n\mu_n\left(\frac{1}{n}A_k\right)^2}\sum_{i>k}\lambda_i^2 = \frac{\left\|\Sigma_{>k}\right\|^2}{\mu_n\left(\frac{1}{n}A_k\right)^2} \cdot \frac{r_k(\Sigma^2)}{n} \leq \rho_{k,n}^2 \frac{r_k(\Sigma^2)}{n}. \tag{26}$$

Furthermore, because the trace of a matrix is the sum of its eigenvalues, we obtain

$$\begin{aligned} \mu_1\left(\frac{1}{n}A_k\right)^2 = & \frac{\mu_1\left(\frac{1}{n}A_k\right)^2}{\mu_n\left(\frac{1}{n}A_k\right)^2}\mu_n\left(\frac{1}{n}A_k\right)^2 \leq \rho_{k,n}^2\left(\frac{1}{n}\mathrm{tr}\left(\frac{1}{n}A_k\right)\right)^2 \\ \leq & \rho_{k,n}^2\left(\gamma_n + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i>k}\lambda_i\psi_i(\mathbf{x}_j)^2\right)^2 \leq \rho_{k,n}^2\left(\gamma_n + \frac{\beta_k \mathrm{tr}\left(\Sigma_{>k}\right)}{n}\right)^2. \end{aligned} \tag{27}$$

and similarly

$$\begin{aligned}\mu_n\left(\frac{1}{n}A_k\right)^2 \geq& \frac{\mu_n\left(\frac{1}{n}A_k\right)^2}{\mu_1\left(\frac{1}{n}A_k\right)^2}\mu_1\left(\frac{1}{n}A_k\right)^2 \geq \frac{1}{\rho_{k,n}^2}\left(\frac{1}{n}\text{tr}\left(\frac{1}{n}A_k\right)\right)^2\\ \geq& \frac{1}{\rho_{k,n}^2}\left(\gamma_n + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i>k}\lambda_i\psi_i(\mathbf{x}_j)^2\right)^2 \geq \frac{1}{\rho_{k,n}^2}\left(\gamma_n + \frac{\alpha_k\text{tr}\left(\Sigma_{>k}\right)}{n}\right)^2.\end{aligned} \tag{28}$$

We thus also obtain an alternative bound for Eq. (26) via Eq. (28) as

$$\frac{1}{n\mu_n\left(\frac{1}{n}A_k\right)^2}\sum_{i>k}\lambda_i^2 \leq \rho_{k,n}^2\frac{n\sum_{i>k}\lambda_i^2}{\left(n\gamma_n + \alpha_k\text{tr}\left(\Sigma_{>k}\right)\right)^2} \leq \frac{\rho_{k,n}^2}{\alpha_k^2}\frac{n}{R_k(\Sigma)}. \tag{29}$$

Now for the variance part of the claim, by combining Lemma D.1 with Eq. (24), Eq. (26) and Eq. (29), we obtain that w.p at least $1-\delta-8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ it holds that

$$V \leq C_1\rho_{k,n}^2\sigma_\epsilon^2\left(\frac{k}{n} + \min\left(\frac{r_k\left(\Sigma^2\right)}{n}, \frac{n}{\alpha_k^2 R_k(\Sigma)}\right)\right). \tag{30}$$

For the bias part of the claim, by similarly combining Lemma D.2 with Eq. (24), Eq. (25) and Eq. (27), and using the fact that $\rho_{k,n} > 1$, we obtain that w.p at least $1-\delta-8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$

$$\begin{aligned}B \leq& C_2\Bigg(\|\theta^*_{>k}\|^2_{\Sigma_{>k}}\left(1+\frac{1}{\delta}\left(\rho_{k,n}^2+\rho_{k,n}\right)\right)\\ &+\left\|\theta^*_{\leq k}\right\|^2_{\Sigma_{\leq k}^{-1}}\left(\rho_{k,n}^2\left(\gamma_n+\frac{\beta_k\text{tr}\left(\Sigma_{>k}\right)}{n}\right)^2(1+\rho_{k,n})\right)\Bigg)\\ \leq& C_2\cdot 3\rho_{k,n}^3\left(\frac{1}{\delta}\|\theta^*_{>k}\|^2_{\Sigma_{>k}} + \left\|\theta^*_{\leq k}\right\|^2_{\Sigma_{\leq k}^{-1}}\left(\gamma_n+\frac{\beta_k\text{tr}\left(\Sigma_{>k}\right)}{n}\right)^2\right).\end{aligned} \tag{31}$$

So everything holds w.p at least $1-\delta-16\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ □

Lemma D.1. *There exists some absolute constants $c, c', C_1 > 0$, s.t for any $k \in \mathbb{N}$ with $c\beta_k k\log(k) \leq n$, it holds w.p at least $1-8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ the variance can be upper bounded as:*

$$V \leq C_1\sigma_\epsilon^2\left(\frac{\mu_1\left(\frac{1}{n}A_k\right)^2 k}{\mu_n\left(\frac{1}{n}A_k\right)^2 n} + \frac{1}{n\mu_n\left(\frac{1}{n}A_k\right)^2}\sum_{i>k}\lambda_i^2\right). \tag{32}$$

Proof. A_k is positive definite for any $\gamma_n > 0$ and thus, by lemma $Eq.$ ($D.5$) we have that:

$$V \leq \sigma_\epsilon^2\left(\frac{\mu_1(A_k^{-1})^2\text{tr}(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top)}{\mu_n(A_k^{-1})^2\mu_k\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right)^2} + \mu_1(A_k^{-1})^2\text{tr}(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top)\right).$$

Plugging in the bounds from Lemma B.4, there are some absolute constants $c, c', c_1, c_2 > 0$ s.t for any $k \in \mathbb{N}$ with $c\beta_k k\log(k) \leq n$, it holds w.p at least $1-8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ that

$$\begin{aligned}V \leq& \sigma_\epsilon^2\left(\frac{\mu_1(A_k^{-1})^2 c_2 kn}{\mu_n(A_k^{-1})^2 c_1^2 n^2} + \mu_1(A_k^{-1})^2 c_2 n\sum_{i>k}\lambda_i^2\right)\\ \leq& c_2\left(\frac{1}{c_1^2}+1\right)\sigma_\epsilon^2\left(\frac{\mu_1(A_k^{-1})^2 k}{\mu_n(A_k^{-1})^2 n} + \mu_1(A_k^{-1})^2 n\sum_{i>k}\lambda_i^2\right).\end{aligned}$$

Now taking C_1 accordingly, and the facts that $\mu_1(A_k^{-1}) = \frac{1}{n\mu_n(\frac{1}{n}A_k)}$ and $\mu_n(A_k^{-1}) = \frac{1}{n\mu_1(\frac{1}{n}A_k)}$ complete the proof.

□

Lemma D.2. *There exists some absolute constants $c, c', C_2 > 0$ (where c and c' are the same as in Lemma D.1), s.t for any $k \in \mathbb{N}$ with $c\beta_k k\log(k) \le n$, and $\delta > 0$, it holds w.p at least $1 - \delta - 8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ the bias can be upper bounded as:*

$$\begin{aligned} B \le C_2\Bigg(&\|\theta^*_{>k}\|^2_{\Sigma_{>k}}\left(1+\frac{1}{\delta}\left(\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}+\frac{\|\Sigma_{>k}\|}{\mu_n\left(\frac{1}{n}A_k\right)}\right)\right) \\ &+\|\theta^*_{\le k}\|^2_{\Sigma^{-1}_{\le k}}\left(\mu_1\left(\frac{1}{n}A_k\right)^2\left(1+\frac{\|\Sigma_{>k}\|}{\mu_n\left(\frac{1}{n}A_k\right)}\right)\right)\Bigg). \end{aligned} \tag{33}$$

Proof. Similarly, to the variance term, by lemma Eq. (D.6) we have that

$$\begin{aligned} &\|\theta^* - \hat{\theta}(\phi(X)\theta^*)\|^2_{\Sigma} \\ &\le \|\theta^*_{>k}\|^2_{\Sigma_{>k}} + \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{\mu_1\left(\psi_{\le k}(X)^\top\psi_{\le k}(X)\right)}{\mu_k\left(\psi_{\le k}(X)^\top\psi_{\le k}(X)\right)^2}\|\phi_{>k}(X)\theta^*_{>k}\|^2 \\ &\quad + \frac{\|\theta^*_{\le k}\|^2_{\Sigma^{-1}_{\le k}}}{\mu_n(A_k^{-1})^2\mu_k\left(\psi_{\le k}(X)^\top\psi_{\le k}(X)\right)^2} \\ &\quad + \|\Sigma_{>k}\|\,\mu_1(A_k^{-1})\|\phi_{>k}(X)\theta^*_{>k}\|^2 \\ &\quad + \|\Sigma_{>k}\|\,\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\psi_{\le k}(X)^\top\psi_{\le k}(X))}{\mu_k(\psi_{\le k}(X)^\top\psi_{\le k}(X))^2}\|\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2. \end{aligned}$$

Plugging in the bounds from lemmas Eq. (B.4) and Eq. (B.3), there are some absolute constants $c, c', c_1, c_2 > 0$ s.t for any $k \in \mathbb{N}$ with $c\beta_k k\log(k) \le n$, it holds w.p at least $1 - 8\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$ that

$$\begin{aligned} &\|\theta^* - \hat{\theta}(\phi(X)\theta^*)\|^2_{\Sigma} \\ &\le \|\theta^*_{>k}\|^2_{\Sigma_{>k}} + \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{c_2 n}{c_1^2 n^2}\cdot\frac{1}{\delta}n\,\|\theta^*_{>k}\|^2_{\Sigma_{>k}} \\ &\quad + \frac{\|\theta^*_{\le k}\|^2_{\Sigma^{-1}_{\le k}}}{\mu_n(A_k^{-1})^2 c_1^2 n^2} \\ &\quad + \|\Sigma_{>k}\|\,\mu_1(A_k^{-1})\cdot\frac{1}{\delta}n\,\|\theta^*_{>k}\|^2_{\Sigma_{>k}} \\ &\quad + \|\Sigma_{>k}\|\,\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{c_2 n}{c_1^2 n^2}\|\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2 \\ &\le C_2\left(\|\theta^*_{>k}\|^2_{\Sigma_{>k}}\left(1+\frac{1}{\delta}\left(\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}+n\,\|\Sigma_{>k}\|\,\mu_1(A^{-1})\right)\right)\right. \\ &\qquad \left. + \|\theta^*_{\le k}\|^2_{\Sigma^{-1}_{\le k}}\left(\frac{1}{n^2\mu_n(A_k^{-1})^2}+\|\Sigma_{>k}\|\,\frac{\mu_1(A_k^{-1})}{n\mu_n(A_k^{-1})^2}\right)\right), \end{aligned}$$

where $C_2 > 0$ can be chosen to depend only on c_1 and c_2 (which are absolute constants). Now we can use the facts that $\mu_1(A_k^{-1}) = \frac{1}{n\mu_n(\frac{1}{n}A_k)}$ and $\mu_n(A_k^{-1}) = \frac{1}{n\mu_1(\frac{1}{n}A_k)}$ to complete the proof, since $\mu_1(A^{-1}) \le \mu_1(A_k^{-1}) = \frac{1}{n\mu_n(\frac{1}{n}A_k)}$ and $\frac{1}{n^2\mu_n(A_k^{-1})^2} = \mu_1(A_k^{-1})^2$, and finally $\frac{\mu_1(A_k^{-1})}{n\mu_n(A_k^{-1})^2} = \frac{1}{\mu_n(A_k^{-1})}$. □

D.2. Lemmas for Risk bounds

In Tsigler & Bartlett (2023)[Appendices F,G,H], several inequalities which will be highly useful to us were derived. Unfortunately, they assumed throughout their paper that the features are finite-dimensional, mean zero, and follow some sub-Gaussianity constraint. The proofs from their paper that we need technically do not depend on these constraints. However, for completeness and rigor, we rewrite their proofs here, adjusted where necessary to match our settings. Again, we remind the reader of the notations $A_k := \mathbf{K}_{>k} + n\gamma_n I$ and $A := \mathbf{K} + n\gamma_n I$ as defined in Sec. A.

Lemma D.3. *For any $k \in \mathbb{N}$ it holds that*

$$\hat{\theta}(y)_{\leq k} + \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(y)_{\leq k} = \phi_{\leq k}(X)^\top A_k^{-1} y.$$

Proof. We start with the ridgeless case, where $\hat{\theta}(y)$ is the minimum norm interpolating solution. Note that $\hat{\theta}(y)_{>k}$ is also the minimum norm solution to the equation $\phi_{>k}(X)\theta_{>k} = y - \phi_{\leq k}(X)\hat{\theta}(y)_{\leq k}$, where $\theta_{>k}$ is the variable. Thus, we can write

$$\hat{\theta}(y)_{>k} = \phi_{>k}(X)^\top \left(\phi_{>k}(X)\phi_{>k}(X)^\top\right)^{-1}\left(y - \phi_{\leq k}(X)\hat{\theta}(y)_{\leq k}\right).$$

As such, we obtain that the min norm interpolator is the minimizer of the following:

$$\hat{\theta}(y) = \underset{\theta_{\leq k}}{\arg\min}\, v(\theta_{\leq k}) := \left[\theta_{\leq k}^\top, (y - \phi_{\leq k}(X)\theta_{\leq k})^\top \left(\phi_{>k}(X)\phi_{>k}(X)^\top\right)^{-1}\phi_{>k}(X)\right]$$

As $\theta_{\leq k}$ varies, this vector sweeps an affine subspace of our Hilbert space. The vector $\hat{\theta}(y)_{\leq k}$ gives the minimum norm if and only if for any additional vector $\eta_{\leq k}$ we have $v(\hat{\theta}(y)_{\leq k}) \perp v(\hat{\theta}(y)_{\leq k} + \eta_{\leq k}) - v(\hat{\theta}(y)_{\leq k})$. Let's write out the second vector: $\forall \eta_{\leq k} \in \mathbb{R}^k$

$$v(\hat{\theta}(y)_{\leq k} + \eta_{\leq k}) - v(\hat{\theta}(y)_{\leq k}) = \left[\eta_{\leq k}^\top, -\eta_{\leq k}^\top \phi_{\leq k}(X)^\top \left(\phi_{>k}(X)\phi_{>k}(X)^\top\right)^{-1}\phi_{>k}(X)\right]$$

We see that the above mentioned orthogonality for any $\eta_{\leq k}$ is equivalent to the following:

$$\hat{\theta}(y)_{\leq k}^\top - \left(y - \phi_{\leq k}(X)\hat{\theta}(y)_{\leq k}\right)^\top \left(\phi_{>k}(X)\phi_{>k}(X)^\top\right)^{-1}\phi_{\leq k}(X) = 0,$$
$$\hat{\theta}(y)_{\leq k} + \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(y)_{\leq k} = \phi_{\leq k}(X)^\top A_k^{-1} y,$$

where we replaced $\phi_{>k}(X)\phi_{>k}(X)^\top =: A_k$.

This completes the ridgeless case, and we now move on to the case of $\gamma_n > 0$. We have that

$$\hat{\theta}(y)_{\leq k} = \phi_{\leq k}(X)^\top(\mathbf{K} + n\gamma_n I)^{-1} y = \phi_{\leq k}(X)^\top (A_k + \phi_{\leq k}(X)\phi_{\leq k}(X)^\top)^{-1} y.$$

Which yields

$$\begin{aligned}
&\hat{\theta}(y)_{\leq k} + \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(y)_{\leq k} \\
=&\phi_{\leq k}(X)^\top (A_k + \phi_{\leq k}(X)\phi_{\leq k}(X)^\top)^{-1} y \\
&+ \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\phi_{\leq k}(X)^\top (A_k + \phi_{\leq k}(X)\phi_{\leq k}(X)^\top)^{-1} y \\
=&\phi_{\leq k}(X)^\top A_k^{-1}(A_k + \phi_{\leq k}(X)\phi_{\leq k}(X)^\top)(A_k + \phi_{\leq k}(X)\phi_{\leq k}(X)^\top)^{-1} y \\
=&\phi_{\leq k}(X)^\top A_k^{-1} y.
\end{aligned}$$

□

We now prove a very simple lemma that will help us formalized the intuition that we can split the error into the $\leq k$ and $> k$ components

Lemma D.4. *For any $k \in \mathbb{N}$, and $\mathbf{v} \in \mathbb{R}^{\mathbb{N}}$,*

$$\|\mathbf{v}\|_{\Sigma}^2 = \|\mathbf{v}_{\leq k}\|_{\Sigma_{\leq k}}^2 + \|\mathbf{v}_{>k}\|_{\Sigma_{>k}}^2$$

Proof. We can write $\mathbf{v} = \begin{bmatrix} \mathbf{v}_{\leq k} \\ \mathbf{v}_{>k} \end{bmatrix}$ and since Σ is diagonal $\Sigma = \begin{bmatrix} \Sigma_{\leq k} & 0 \\ 0 & \Sigma_{>k} \end{bmatrix}$ and thus:

$$\|\mathbf{v}\|_{\Sigma}^2 = \begin{bmatrix} \mathbf{v}_{\leq k} & \mathbf{v}_{>k} \end{bmatrix} \begin{bmatrix} \Sigma_{\leq k} & 0 \\ 0 & \Sigma_{>k} \end{bmatrix} \begin{bmatrix} \mathbf{v}_{\leq k} \\ \mathbf{v}_{>k} \end{bmatrix} = \|\mathbf{v}_{\leq k}\|_{\Sigma_{\leq k}}^2 + \|\mathbf{v}_{>k}\|_{\Sigma_{>k}}^2 .$$

□

The next lemma provides a useful upper bound for the variance.

Lemma D.5 (Variance term)**.** *If for some $k \in \mathbb{N}$ the matrix A_k is PD, then*

$$V \leq \sigma_\epsilon^2 \left(\frac{\mu_1(A_k^{-1})^2 tr(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top)}{\mu_n(A_k^{-1})^2 \mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right)^2} + \mu_1(A_k^{-1})^2 tr(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top) \right).$$

Proof. Recall that

$$V = \mathbb{E}_\epsilon\left[\left\|\hat{\theta}(\boldsymbol{\epsilon})\right\|_{\Sigma}^2 \right] = \mathbb{E}_\epsilon\left[\left\|\phi(X)^\top (\mathbf{K} + n\gamma_n I)^{-1}\epsilon\right\|_{\Sigma}^2 \right]$$

By Lemma Eq. (D.4) we can split the variance into $\left\|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\right\|_{\Sigma \leq k}^2$ and $\left\|\hat{\theta}(\boldsymbol{\epsilon}_{>k})\right\|_{\Sigma > k}^2$ and bound these separately.

Lemma Eq. (D.3) states that

$$\phi_{\leq k}(X)^\top A_k^{-1}\epsilon = \hat{\theta}(\boldsymbol{\epsilon}_{\leq k}) + \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(\boldsymbol{\epsilon}_{\leq k}).$$

Multiplying the identity by $\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top$ from the left, and using that $\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \hat{\theta}(\boldsymbol{\epsilon}_{\leq k}) \geq 0$ we get

$$\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \phi_{\leq k}(X)^\top A_k^{-1}\epsilon \geq \hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(\boldsymbol{\epsilon}_{\leq k}). \tag{34}$$

The leftmost expression is linear in $\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})$, and the rightmost is quadratic. We use these expressions to bound $\|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\|_{\Sigma_{\leq k}}$. First, we extract that norm from the quadratic part

$$\begin{aligned} \hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(\boldsymbol{\epsilon}_{\leq k}) \geq& \mu_n(A_k^{-1})\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \phi_{\leq k}(X)^\top \phi_{\leq k}(X)\hat{\theta}(\boldsymbol{\epsilon}_{\leq k}) \\ \geq& \mu_n(A_k^{-1})\|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\|_{\Sigma_{\leq k}}^2 \mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right). \end{aligned}$$

Then we can substitute Eq. (34) and apply Cauchy-Schwarz to obtain

$$\begin{aligned} \|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\|_{\Sigma_{\leq k}}^2 \mu_n(A_k^{-1})\mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right) &\leq \hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(\boldsymbol{\epsilon}_{\leq k}) \\ &\leq \hat{\theta}(\boldsymbol{\epsilon}_{\leq k})^\top \phi_{\leq k}(X)^\top A_k^{-1}\epsilon \\ &\leq \|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\|_{\Sigma_{\leq k}} \left\|\psi_{\leq k}(X)^\top A_k^{-1}\epsilon\right\|, \end{aligned}$$

and so

$$\|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\|_{\Sigma_{\leq k}}^2 \leq \frac{\epsilon^\top A_k^{-1}\psi_{\leq k}(X)\psi_{\leq k}(X)^\top A_k^{-1}\epsilon}{\mu_n(A_k^{-1})^2 \mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right)^2}.$$

Since ϵ is independent of X, taking expectation in ϵ only leaves the trace in the numerator:

$$\begin{aligned} \mathbb{E}_\epsilon \|\hat{\theta}(\boldsymbol{\epsilon}_{\leq k})\|_{\Sigma_{\leq k}}^2 \leq& \sigma_\epsilon^2 \frac{\mathrm{tr}(A_k^{-1}\psi_{\leq k}(X)\psi_{\leq k}(X)^\top A_k^{-1})}{\mu_n(A_k^{-1})^2 \mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right)^2} \\ \leq& \sigma_\epsilon^2 \frac{\mu_1(A_k^{-1})^2 \mathrm{tr}(\psi_{\leq k}(X)\psi_{\leq k}(X)^\top)}{\mu_n(A_k^{-1})^2 \mu_k\left(\psi_{\leq k}(X)^\top \psi_{\leq k}(X)\right)^2}, \end{aligned}$$

where we transitioned to the second line by using the fact that tr$(MM'M) \leq \mu_1(M)^2$tr(M') for PD matrices M, M'.

This completes the bound for the first $\leq k$ components, and we now move on to the $> k$ ones. The rest of the variance term is

$$\left\|\Sigma_{>k}^{1/2}\phi_{>k}(X)^\top A^{-1}\epsilon\right\|^2 = \epsilon^\top A^{-1}\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top A^{-1}\epsilon.$$

Since ϵ is independent of X, taking expectation in ϵ only leaves the trace of the matrix:

$$\begin{aligned}\frac{1}{\sigma_\epsilon^2}\mathbb{E}_\epsilon\left\|\Sigma_{>k}^{1/2}\phi_{>k}(X)^\top A^{-1}\epsilon\right\|^2 =&\text{tr}(A^{-1}\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top A^{-1})\\ \leq&\mu_1(A^{-1})^2\text{tr}(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top)\\ \leq&\mu_1(A_k^{-1})^2\text{tr}(\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top).\end{aligned}$$

Here we again used the fact that tr$(MM'M) \leq \mu_1(M)^2$tr(M') for PD matrices M, M' to transition to the second line. We then used $A \succeq A_k$ to infer $\mu_1(A^{-1}) \leq \mu_1(A_k^{-1})$. □

We now move on to bounding the bias term.

Lemma D.6 (Bias term)**.** *Suppose that for some $k < n$ the matrix A_k is PD. Then,*

$$\begin{aligned}&\|\theta^* - \hat{\theta}(\phi(X)\theta^*)\|_\Sigma^2\\ &\leq \|\theta^*_{>k}\|^2_{\Sigma_{>k}} + \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{\mu_1\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right)}{\mu_k\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right)^2}\|\phi_{>k}(X)\theta^*_{>k}\|^2\\ &\quad + \frac{\|\theta^*_{\leq k}\|^2_{\Sigma^{-1}_{\leq k}}}{\mu_n(A_k^{-1})^2\mu_k\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right)^2}\\ &\quad + \|\Sigma_{>k}\|\,\mu_1(A_k^{-1})\|\phi_{>k}(X)\theta^*_{>k}\|^2\\ &\quad + \|\Sigma_{>k}\|\,\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\psi_{\leq k}(X)^\top\psi_{\leq k}(X))}{\mu_k(\psi_{\leq k}(X)^\top\psi_{\leq k}(X))^2}\|\Sigma_{\leq k}^{-1/2}\theta^*_{\leq k}\|^2.\end{aligned}$$

Proof. As before, by Lemma Eq. (D.4) we can bound the $\leq k$ components and the $> k$ components separately. We start by bounding $\|\theta^*_{\leq k} - \hat{\theta}(y)_{\leq k}(\phi(X)\theta^*)\|^2_{\Sigma_{\leq k}}$. By Lemma Eq. (D.3), we have

$$\hat{\theta}(\phi(X)\theta^*)_{\leq k} + \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\hat{\theta}(\phi(X)\theta^*)_{\leq k} = \phi_{\leq k}(X)^\top A_k^{-1}\phi(X)\theta^*.$$

Denote the error vector as $\zeta := \hat{\theta}(\phi(X)\theta^*) - \theta^*$. We can rewrite the equation above as

$$\zeta_{\leq k} + \phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\zeta_{\leq k} = \phi_{\leq k}(X)^\top A_k^{-1}\phi_{>k}(X)\theta^*_{>k} - \theta^*_{\leq k}.$$

Multiplying both sides by $\zeta_{\leq k}^\top$ from the left and using that $\zeta_{\leq k}^\top\zeta_{\leq k} = \|\zeta_{\leq k}\|^2 \geq 0$ we obtain

$$\zeta_{\leq k}^\top\phi_{\leq k}(X)^\top A_k^{-1}\phi_{\leq k}(X)\zeta_{\leq k} \leq \zeta_{\leq k}^\top\phi_{\leq k}(X)^\top A_k^{-1}\phi_{>k}(X)\theta^*_{>k} - \zeta_{\leq k}^\top\theta^*_{\leq k}.$$

Next, divide and multiply by $\Sigma_{\leq k}^{1/2}$ in several places:

$$\begin{aligned}\zeta_{\leq k}^\top\Sigma_{\leq k}^{1/2}\psi_{\leq k}(X)^\top A_k^{-1}\psi_{\leq k}(X)\Sigma_{\leq k}^{1/2}\zeta_{\leq k} \leq&\zeta_{\leq k}^\top\Sigma_{\leq k}^{1/2}\psi_{\leq k}(X)^\top A_k^{-1}\phi_{>k}(X)\theta^*_{>k}\\ &- \zeta_{\leq k}^\top\Sigma_{\leq k}^{1/2}\Sigma_{\leq k}^{-1/2}\theta^*_{\leq k}.\end{aligned}$$

Now we pull out the lowest singular values of the matrices in the LHS and largest singular values of the matrices in the RHS to obtain lower and upper bounds respectively, yielding

$$\begin{aligned}&\|\zeta_{\leq k}\|^2_{\Sigma_{\leq k}}\mu_n(A_k^{-1})\mu_k\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right)\\ &\leq \|\zeta_{\leq k}\|_{\Sigma_{\leq k}}\mu_1(A_k^{-1})\sqrt{\mu_1\left(\psi_{\leq k}(X)^\top\psi_{\leq k}(X)\right)}\|\phi_{>k}(X)\theta^*_{>k}\|\\ &+ \|\zeta_{\leq k}\|_{\Sigma_{\leq k}}\|\theta^*_{\leq k}\|_{\Sigma^{-1}_{\leq k}},\end{aligned}$$

and so

$$\begin{aligned}\|\zeta_{\le k}\|_{\Sigma_{\le k}} \le & \frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})} \frac{\mu_1\left(\psi_{\le k}(X)^\top \psi_{\le k}(X)\right)^{1/2}}{\mu_k\left(\psi_{\le k}(X)^\top \psi_{\le k}(X)\right)} \|\phi_{>k}(X)\theta^*_{>k}\| \\ & + \frac{\|\theta^*_{\le k}\|_{\Sigma^{-1}_{\le k}}}{\mu_n(A_k^{-1})\mu_k\left(\psi_{\le k}(X)^\top \psi_{\le k}(X)\right)}.\end{aligned}$$

This completes the bounds for the $\le k$ components and we now move on to the $> k$ ones. The contribution of the components of ζ, starting from the $k+1$st can be bounded as follows:

$$\begin{aligned}&\|\theta^*_{>k} - \phi_{>k}(X)^\top A^{-1}\phi(X)\theta^*\|^2_{\Sigma_{>k}} \\ &\qquad \le 3\left(\|\theta^*_{>k}\|^2_{\Sigma_{>k}} + \|\phi_{>k}(X)^\top A^{-1}\phi_{>k}(X)\theta^*_{>k}\|^2_{\Sigma_{>k}} + \|\phi_{>k}(X)^\top A^{-1}\phi_{\le k}(X)\theta^*_{\le k}\|^2_{\Sigma_{>k}}\right).\end{aligned}$$

First of all, let's deal with the second term:

$$\begin{aligned}\|\phi_{>k}(X)^\top A^{-1}\phi_{>k}(X)\theta^*_{>k}\|^2_{\Sigma_{>k}} =& \|\Sigma^{1/2}_{>k}\phi_{>k}(X)^\top A^{-1}\phi_{>k}(X)\theta^*_{>k}\|^2 \\ \le& \|\Sigma_{>k}\|\|\phi_{>k}(X)^\top A^{-1}\phi_{>k}(X)\theta^*_{>k}\|^2 \\ =& \|\Sigma_{>k}\|\,(\theta^*_{>k})^\top\phi_{>k}(X)^\top A^{-1} \underbrace{\phi_{>k}(X)\phi_{>k}(X)^\top}_{=A-n\gamma_n I-\phi_{\le k}(X)\phi_{\le k}(X)^\top \preceq A} A^{-1}\phi_{>k}(X)\theta^*_{>k} \\ \le& \|\Sigma_{>k}\|\,(\theta^*_{>k})^\top\phi_{>k}(X)^\top A^{-1}\phi_{>k}(X)\theta^*_{>k} \\ \le& \|\Sigma_{>k}\|\,\mu_1(A_k^{-1})\|\phi_{>k}(X)\theta^*_{>k}\|^2,\end{aligned}$$

where we used that $\mu_1(A_k^{-1}) \ge \mu_1(A^{-1})$ in the last transition.

Now, let's deal with the last term. Note that $A = A_k + \phi_{\le k}(X)\phi_{\le k}(X)^\top$. By the Sherman–Morrison–Woodbury formula,

$$\begin{aligned}A^{-1}\phi_{\le k}(X) =& (A_k^{-1} + \phi_{\le k}(X)\phi_{\le k}(X)^\top)^{-1}\phi_{\le k}(X) \\ =& \left(A_k^{-1} - A_k^{-1}\phi_{\le k}(X)\left(I_k + \phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X)\right)^{-1}\phi_{\le k}(X)^\top A_k^{-1}\right)\phi_{\le k}(X) \\ =& A_k^{-1}\phi_{\le k}(X)\left(I_n - \left(I_k + \phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X)\right)^{-1}\phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X)\right) \\ =& A_k^{-1}\phi_{\le k}(X)\left(I_n - \left(I_k + \phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X)\right)^{-1}\left(I_k + \phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X) - I_k\right)\right) \\ =& A_k^{-1}\phi_{\le k}(X)\left(I_k + \phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X)\right)^{-1}.\end{aligned}$$

Thus,

$$\begin{aligned}&\|\phi_{>k}(X)^\top A^{-1}\phi_{\le k}(X)\theta^*_{\le k}\|^2_{\Sigma_{>k}} \\ =& \|\phi_{>k}(X)^\top A_k^{-1}\phi_{\le k}(X)\left(I_k + \phi_{\le k}(X)^\top A_k^{-1}\phi_{\le k}(X)\right)^{-1}\theta^*_{\le k}\|^2_{\Sigma_{>k}} \\ =& \|\Sigma^{1/2}_{>k}\phi_{>k}(X)^\top A_k^{-1}\psi_{\le k}(X)\left(\Sigma^{-1}_{\le k} + \psi_{\le k}(X)^\top A_k^{-1}\psi_{\le k}(X)\right)^{-1}\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2 \\ \le& \|A_k^{-1/2}\phi_{>k}(X)\Sigma_{>k}\phi_{>k}(X)^\top A_k^{-1/2}\|\mu_1(A_k^{-1/2})^2\frac{\mu_1(\psi_{\le k}(X)^\top\psi_{\le k}(X))}{\mu_k(\psi_{\le k}(X)^\top A_k^{-1}\psi_{\le k}(X))^2}\|\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2 \\ \le& \|\Sigma_{>k}\|\|A_k^{-1/2}\phi_{>k}(X)\phi_{>k}(X)^\top A_k^{-1/2}\|\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\psi_{\le k}(X)^\top\psi_{\le k}(X))}{\mu_k(\psi_{\le k}(X)^\top\psi_{\le k}(X))^2}\|\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2 \\ =& \|\Sigma_{>k}\|\,\|I_n - n\gamma_n A_k^{-1}\|\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\psi_{\le k}(X)^\top\psi_{\le k}(X))}{\mu_k(\psi_{\le k}(X)^\top\psi_{\le k}(X))^2}\|\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2 \\ \le& \|\Sigma_{>k}\|\,\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\psi_{\le k}(X)^\top\psi_{\le k}(X))}{\mu_k(\psi_{\le k}(X)^\top\psi_{\le k}(X))^2}\|\Sigma^{-1/2}_{\le k}\theta^*_{\le k}\|^2,\end{aligned}$$

where in the last transition we used the fact that $I_n - n\gamma_n A_k^{-1}$ is a PSD matrix with norm bounded by 1 for $\gamma_n > 0$.

Putting those bounds together yields the result.

□

E. Applications - Proofs of Results in Sec. 5

E.1. Regularized Case (Thm. 5.3)

Theorem 5.3. *Let K be a kernel with polynomially decaying eigenvalues $\lambda_i = \Theta_{i,n}(i^{-1-a})$ for some $a > 0$, and assume that $\beta_k = \mathcal{O}_k(1)$. Further, suppose that the regularization parameter satisfies $\gamma_n = \Theta_n(n^{-1-b})$ for $b \in (-1, a)$. Then for any $\delta > 0$, it holds w.p at least $1 - \delta - o_n(\frac{1}{n})$ that*

$$V \leq \sigma_\epsilon^2 \cdot \mathcal{O}_n\left(\frac{1}{n^{\frac{a-b}{1+a}}}\right),$$

and if $\theta_i^* = \Theta_{i,n}\left(i^{-r}\right)$ for some $r \in \mathbb{R}$ s.t $\left\|\Sigma^{1/2}\theta^*\right\|_2 < \infty$ (necessary for $f^* \in L_\mu^2(\mathcal{X})$), then under the same probability it also holds that

$$B \leq \frac{1}{\delta} \cdot \mathcal{O}_n\left(\frac{1}{n^{(1+b)\min\left(\frac{(2r+a)}{1+a}, 2\right)}}\right),$$

where the $\mathcal{O}$ is weakened to $\tilde{\mathcal{O}}$ if $r = 1 + \frac{a}{2}$.

Proof. We use Thm. 4.1, which states that there exist some absolute constants $c, c' > 0$ s.t for any $k \in \mathbb{N}$ with $c\beta_k k \log(k) \leq n$ and any $\delta > 0$, Eq. (12) and Eq. (13) hold w.p at least $1 - \delta - 16\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$.

In order to use the theorem, for any n we first have to pick some $k \in \mathbb{N}$ s.t $c\beta_k k \log(k) \leq n$. As such, let $k := k(n) := \left\lceil n^{\frac{1+b}{1+a}} \right\rceil$. The condition $b \in (-1, a)$ implies that $\frac{1+b}{1+a} < 1$, and thus $k(n) = o_n\left(\frac{n}{\log(n)}\right)$, meaning that for sufficiently large n, Thm. 4.1 can be used with this chosen k. Since k is a function of n, the $\mathcal{O}_n$ notation in particular, implies constants w.r.t k.

We now proceed to bounding $\rho_{k,n}$ (as defined in Thm. 4.1). By Lemma E.2 it holds w.p at least $1 - \mathcal{O}_n\left(\frac{1}{k^3}\right)\exp\left(-\Omega_n(\frac{n}{k})\right)$ that

$$\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) = \mathcal{O}_n\left(\lambda_{k+1}\right) = \mathcal{O}_n\left(\left(n^{\frac{1+b}{1+a}}\right)^{-(1+a)}\right) = \mathcal{O}_n\left(n^{-1-b}\right) = \mathcal{O}_n\left(\gamma_n\right). \tag{35}$$

We can bound the event that both Thm. 4.1 hold and Eq. (35) hold as

$$1 - \delta - 16\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right) - \mathcal{O}_n\left(\frac{1}{k^3}\right)\exp\left(-\Omega_n\left(\frac{n}{k}\right)\right) = 1 - \delta - \mathcal{O}_n\left(\frac{1}{n}\right),$$

Where we used the facts that $\frac{c'}{\beta_k^2}\frac{n}{k} = \omega_n\left(\log(n)\right)$. From now on, we assume that both Thm. 4.1 and Eq. (35) indeed hold. Plugging Eq. (35) into the definition of the concentration coefficient Eq. (11) and using $\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) \geq 0$, we obtain the bound

$$\rho_{k,n} = \mathcal{O}_n\left(\frac{\lambda_{k+1} + \gamma_n}{\gamma_n}\right) = \mathcal{O}_n\left(\frac{\gamma_n}{\gamma_n}\right) = \mathcal{O}_n\left(1\right). \tag{36}$$

By Lemma E.1, it holds that $r_k(\Sigma), r_k(\Sigma^2) = \Theta_n\left(k\right)$. So plugging this and Eq. (36) into Thm. 4.1 yields,

$$V/\sigma_\epsilon^2 = \mathcal{O}_n\left(\frac{k}{n} + \frac{r_k(\Sigma^2)}{n}\right) = \mathcal{O}\left(\frac{k}{n}\right) = \mathcal{O}_n\left(\frac{n^{\frac{1+b}{1+a}}}{n}\right) = \mathcal{O}_n\left(n^{\frac{b-a}{1+a}}\right),$$

and

$$B = \frac{1}{\delta}\mathcal{O}_n\left(\underbrace{\|\theta^*_{>k}\|^2_{\Sigma_{>k}}}_{:=T_1}\right) + \mathcal{O}_n\left(\underbrace{\|\theta^*_{\leq k}\|^2_{\Sigma^{-1}_{\leq k}}}_{:=T_2}\underbrace{\left(\gamma_n + \frac{\text{tr}\left(\Sigma_{>k}\right)}{n}\right)^2}_{:=T_3}\right).$$

Following Lemma E.1 it holds that $\text{tr}(\Sigma_{>k}) = \mathcal{O}_n(k \cdot \lambda_k) = \mathcal{O}_n(k \cdot \gamma_n)$ and so

$$T_3 = \mathcal{O}_n\left(\left(\gamma_n + \frac{k}{n}\gamma_n\right)^2\right) = \mathcal{O}_n\left(\gamma_n^2\right) = \mathcal{O}_n\left(\frac{1}{n^{2+2b}}\right).$$

Combining this bound for T_3 with the bounds for T_1, T_2 from Lemma E.4 yields

$$\begin{aligned} B &\leq \begin{cases} \mathcal{O}_n\left(\frac{1}{k^{2r+a}} + \frac{1}{k^{2r-2-a}n^{2(1+b)}}\right) & 2r < 2+a \\ \mathcal{O}_n\left(\frac{1}{k^{2(1+a)}} + \frac{\log(k)}{n^{2(1+b)}}\right) & 2r = 2+a \\ \mathcal{O}_n\left(\frac{1}{k^{2r+a}} + \frac{1}{n^{2(1+b)}}\right) & 2r > 2+a \end{cases} \\ &\leq \begin{cases} \mathcal{O}_n\left(\frac{1}{n^{\frac{(2r+a)(1+a)}{1+b}}}\right) & 2r < 2+a \\ \mathcal{O}_n\left(\frac{\log(n)}{n^{2(1+b)}}\right) & 2r = 2+a \\ \mathcal{O}_n\left(\frac{1}{n^{2(1+b)}}\right) & 2r > 2+a \end{cases}. \end{aligned}$$

□

E.2. Fixed Dimensional Interpolation Case (Thm. 5.2)

Theorem 5.2. *Let K be a kernel with polynomially decaying eigenvalues $\lambda_i = \Theta_{i,n}(i^{-1-a})$ for some $a > 0$, and assume that $\alpha_k, \beta_k = \Theta_k(1)$. Then for the min norm solution defined in Eq. (4) (given when $\gamma_n \to 0$), for any $\delta > 0$ it holds w.p at least $1 - \delta - \mathcal{O}_n\left(\frac{1}{\log(n)}\right)$ that*

$$V \leq \sigma_\epsilon^2 \tilde{\mathcal{O}}_n\left(n^{2a}\right).$$

Moreover, if $\theta^*_i = \mathcal{O}_i\left(\frac{1}{i^r}\right)$ where $r > a$ then under the same probability it also holds that

$$B \leq \frac{1}{\delta}\tilde{\mathcal{O}}_n\left(\frac{1}{n^{\min(2(r-a), 2-a)}}\right).$$

Proof. We use Thm. 4.1, which states that there exist some absolute constants $c, c' > 0$ s.t for any $k \in \mathbb{N}$ with $c\beta_k k \log(k) \leq n$ and any $\delta > 0$, Eq. (12) and Eq. (13) hold w.p at least $1 - \delta - 16\exp\left(-\frac{c'}{\beta_k^2}\frac{n}{k}\right)$.

In order to use the theorem, for any n we first have to pick some $k \in \mathbb{N}$ s.t $c\beta_k k \log(k) \leq n$. Using the fact that $\beta_k \leq C_0$ for some $C_0 > 0$, let $k := k(n) := \frac{n}{\max(cC_0, 1)\log(n)}$ and we also let $k' := k'(n) = n^2 \log^4(n)$. The probability that Thm. 4.1 holds with $k(n)$ now becomes $1 - \delta - \mathcal{O}_n(\frac{1}{n})$. Since k is a function of n, the $\mathcal{O}_n$ notation in particular, implies constants w.r.t k.

In order to bound Eq. (12) and Eq. (13), we begin by bounding $\rho_{k,n}$, which requireds bounding $\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right)$ and $\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right)$. First note that by Bartlett et al. (2020)[Lemma 5] $R_k \geq r_k$ and thus by Lemma E.1 it holds that $R_{k'} = \Omega_n\left(n^2 \log^4(n)\right)$

and $\text{tr}\left(\Sigma_{>k'}\right) = \Omega_n\left((n^2\log^4(n))^{-a}\right)$. By Corollary C.5 it holds w.p at least $1-\frac{1}{\log(n)}$ that,

$$\begin{aligned}\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) \geq & \alpha_k\left(1-\frac{1}{\log(n)}\sqrt{\frac{n^2}{R_{k'}}}\right)\frac{\text{tr}\left(\Sigma_{>k'}\right)}{n} \\ =&\Omega_n\left(\left(1-\log(n)\sqrt{\frac{1}{\log^4(n)}}\right)\frac{\text{tr}\left(\Sigma_{>k'}\right)}{n}\right) \\ =&\Omega_n\left(\frac{(n^2\log^4(n))^{-a}}{n}\right) = \Omega_n\left(n^{-1-2a}\log^{-4a}(n)\right). \end{aligned} \tag{37}$$

For $\mu_1\left(\frac{1}{n}\mathbf{K}\right)$, by Lemma E.2 it holds w.p at least $1-\mathcal{O}_n\left(\frac{1}{k^3}\right)\exp\left(-\Omega_n\left(\frac{n}{k}\right)\right)$ that

$$\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) = \mathcal{O}_n\left(\lambda_{k+1}\right) = \mathcal{O}_n\left(n^{-1-a}\log^{1+a}(n)\right). \tag{38}$$

So Thm. 4.1, Eq. (37) and Eq. (38) all hold simultaneously with probability $1-\delta-\mathcal{O}_n\left(\frac{1}{\log(n)}\right)$, and from now on we assume that this is indeed the case.

By combining Eq. (38) and Eq. (37) we obtain the bound

$$\rho_{k,n} = \mathcal{O}_n\left(\frac{n^{-1-a}\log^{1+a}(n)}{n^{-1-2a}\log^{-4a}(n)}\right) = \tilde{\mathcal{O}}_n\left(n^a\right) \tag{39}$$

And thus by combining Eq. (12), Eq. (39) and the fact that from Lemma E.1 $r_k(\Sigma^2)\lesssim k$, we obtain the bound

$$V/\sigma_\epsilon^2 = \tilde{\mathcal{O}}_n\left(n^{2a}\frac{k}{n}\right) = \tilde{\mathcal{O}}_n\left(n^{2a}\right)$$

and

$$B = \frac{1}{\delta}\tilde{\mathcal{O}}_n\left(n^{3a}\left(\underbrace{\|\theta^*_{>k}\|^2_{\Sigma_{>k}}}_{:=T_1} + \underbrace{\|\theta^*_{\leq k}\|^2_{\Sigma^{-1}_{\leq k}}}_{:=T_2}\underbrace{\left(\frac{\text{tr}\left(\Sigma_{>k}\right)}{n}\right)^2}_{:=T_3}\right)\right).$$

Following Lemma E.1 it holds that $\text{tr}(\Sigma_{>k}) = \tilde{\mathcal{O}}_n(k\cdot\lambda_k) = \tilde{\mathcal{O}}_n(\frac{1}{n^a})$ and so $T_3 = \tilde{\mathcal{O}}_n\left(\frac{1}{n^{2+2a}}\right)$. Combining this bound for T_3 with the bounds for T_1, T_2 from Lemma E.4 yields

$$\begin{aligned} T_1 + T_2T_3 \leq & \begin{cases} \tilde{\mathcal{O}}_n\left(\frac{1}{n^{2r+a}} + \frac{1}{n^{2r-2-a}n^{2(1+a)}}\right) & 2r \leq 2+a \\ \tilde{\mathcal{O}}_n\left(\frac{1}{n^{2r+a}} + \frac{1}{n^{2(1+a)}}\right) & 2r > 2+a \end{cases} \\ =&\tilde{\mathcal{O}}_n\left(\frac{1}{n^{\min(2r+a,2(1+a))}}\right). \end{aligned}$$

Implying that

$$\begin{aligned} B \leq & \frac{1}{\delta}\tilde{\mathcal{O}}_n\left(n^{3a}\frac{1}{n^{\min(2r+a,2(1+a))}}\right) \\ =&\frac{1}{\delta}\tilde{\mathcal{O}}_n\left(\frac{1}{\min\left(n^{2(r-a),2-a}\right)}\right). \end{aligned}$$

□

E.3. High Dimensional Interpolation Case (Thm. 5.1)

Theorem 5.1. *Suppose that as $n, d \to \infty$, $\frac{d^\tau}{n} = \Theta_{n,d}(1)$ for some $\tau \in (0,\infty) \setminus \mathbb{N}$. Let μ be the uniform distribution over $\mathbb{S}^{d-1}$ and K be a dot product kernel given by Eq. (14) s.t $\hat{\sigma}_{\lfloor \tau \rfloor} > 0$ and $\exists \ell > \lfloor 2\tau \rfloor$ with $\hat{\sigma}_\ell \geq 0$ (e.g NTK, Laplace, or RBF). Then for the min norm solution defined in Eq. (4) (given when $\gamma_n \to 0$), for any $\delta > 0$ it holds w.p at least $1 - \delta - o_d\left(\frac{1}{d}\right)$ that*

$$\begin{aligned}
V \leq & \sigma_\epsilon^2 \cdot \mathcal{O}_{n,d}\left(\frac{1}{d^{\tau - \lfloor \tau \rfloor}} + \frac{1}{d^{\lfloor \tau \rfloor + 1 - \tau}}\right), \\
B \leq & \frac{1}{\delta}\mathcal{O}_{n,d}\left(\left\|\theta^*_{>N_d}\right\|^2_{\Sigma_{>N_d}}\right) \\
& + \left\|\theta^*_{\leq N_d}\right\|^2_\infty \left(\max_{\ell \leq \lfloor \tau \rfloor \text{ s.t. } \hat{\sigma}_\ell \neq 0} \frac{1}{\hat{\sigma}_\ell}\right) \cdot \mathcal{O}_{n,d}\left(\frac{1}{d^{2(\tau - \lfloor \tau \rfloor)}}\right).
\end{aligned}$$

Where $N_d = \Theta_{n,d}\left(d^{\lfloor \tau \rfloor}\right)$ denotes the number of spherical harmonics of degree at most $\lfloor \tau \rfloor$ with non-zero eigenvalues, and $\mathcal{O}_{n,d}\left(\|\theta^*_{>N_d}\|^2_{\Sigma_{>N_d}}\right) \leq \mathcal{O}_{n,d}\left(\|\theta^*_{>N_d}\|^2_\infty\right)$.

Proof. Let $\sigma_\ell := \frac{\hat{\sigma}_\ell}{N(d,\ell)}$ be the eigenvalues from Eq. (50). We order ϕ in the natural way, by first taking $\tilde{\phi}(\mathbf{x}) = (\sqrt{\sigma_0}Y_{0,1}, \sqrt{\sigma_1}Y_{1,1}, \ldots, \sqrt{\sigma_1}Y_{1,N(d,1)}, \sqrt{\sigma_2}Y_{2,1}, \ldots)$, and letting ϕ be the same as $\tilde{\phi}$ with zero-valued indices removed (where $\sigma_\ell = 0$). We let ψ be given accordingly.

For any $s \in \mathbb{N}$, and $d \in \mathbb{N}$ let $k_s(d) = \sum_{\ell=0}^s N(d,\ell) \cdot \mathbb{I}_{\hat{\sigma}_\ell}$ where $\mathbb{I}_{\hat{\sigma}_\ell} = \begin{cases} 1 & \hat{\sigma}_\ell > 0 \\ 0 & \text{else} \end{cases}$. Let $\Delta_{>k_s(d)} \in \mathbb{R}^{n \times n}$ be the matrix given by $[\Delta_{>k_s(d)}]_{ij} = \begin{cases} \frac{1}{n}[\mathbf{K}_{>k_s(d)}]_{ij} & i \neq j \\ 0 & i = j \end{cases}$. By Eq. (23) we have that

$$\alpha_{k_s(d)}\frac{1}{n}\text{tr}\left(\Sigma_{>k_s(d)}\right) + \mu_n\left(\Delta_{>k_s(d)}\right) \leq \mu_i\left(\frac{1}{n}\mathbf{K}_{>k_s(d)}\right) \leq \beta_{k_s(d)}\frac{1}{n}\text{tr}\left(\Sigma_{>k_s(d)}\right) + \mu_1\left(\Delta_{>k_s(d)}\right). \quad (40)$$

In order to bound the eigenvalues of $\Delta_{>k_s(d)}$ we will need to control the effective ranks. Let $j_s := \arg\max_{j \geq s} \sigma_j$, then

$$r_{k_s(d)}(\Sigma) = \frac{\sum_{i=s+1}^\infty N(d,i)\sigma_i}{\sigma_{j_{s+1}}} \geq N(d, j_{s+1}) \geq N(d, s+1),$$

where our assumption that $\hat{\sigma}_\ell > 0$ for some $\ell \geq \lfloor 2\tau \rfloor$ ensures that $\sigma_{j_{s+1}} > 0$ for $s \leq \lfloor 2\tau \rfloor$. By Bartlett et al. (2020)[Lemma 5] we also have $R_{k_s(d)}(\Sigma) \geq r_{k_s(d)}(\Sigma)$. Let $k(d) := k_{\lfloor \tau \rfloor}(d)$ and $v(d) := k_{\lfloor 2\tau \rfloor}(d)$. Let $t = \min(\lfloor \tau \rfloor - \tau + 1, \lfloor 2\tau \rfloor - 2\tau + 1) > 0$, then by what we just showed, and using the fact that for any $i \in \mathbb{N}$, $N(d,i) = \Theta_d\left(d^i\right)$, we have the following identities:

$$R_{v(d)}(\Sigma) \geq \Omega_{n,d}\left(d^{\lfloor 2\tau \rfloor + 1}\right) \geq \Omega_{n,d}\left(n^{2+\frac{t}{\tau}}\right), \quad (41)$$

$$r_{k(d)}(\Sigma) \geq \Omega_{n,d}\left(d^{\lfloor \tau \rfloor + 1}\right) \geq \Omega_{n,d}\left(n^{1+\frac{t}{\tau}}\right). \quad (42)$$

We have shown that conditions (A2) and (A3) of Mei et al. (2022)[Proposition 4] hold. Furthermore, condition (A1) holds applying Mei et al. (2022)[Lemma 19] to $\psi_{\leq v(d)}$. As a result, Mei et al. (2022)[Proposition 4] states that for some $t' > 0$,

$$\left\|\Delta_{\geq k(d)}\right\| \leq \mathcal{O}_{n,d}\left(d^{-t'}\right) \cdot \frac{1}{n}\text{tr}\left(\Sigma_{>k(d)}\right).$$

Plugging this into Eq. (40) and using that by the addition theorem Eq. (51), $\alpha_{k(d)} = \beta_{k(d)} = 1$, it holds that

$$\mu_i\left(\frac{1}{n}\mathbf{K}_{>k(d)}\right) = \Theta_{n,d}\left(\frac{1}{n}\text{tr}\left(\Sigma_{>k(d)}\right)\right). \quad (43)$$

As a result, we obtain that for $\rho_{k,n}$ as defined in Thm. 4.1,

$$\rho_{k(d),n} = \frac{\|\Sigma_{>k}\| + \mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) + \gamma_n}{\mu_n\left(\frac{1}{n}\mathbf{K}_{>k}\right) + \gamma_n} = \mathcal{O}_{n,d}\left(\frac{\left(\frac{n}{r_{k(d)}} + 1\right)\frac{1}{n}\text{tr}\left(\Sigma_{>k(d)}\right)}{\frac{1}{n}\text{tr}\left(\Sigma_{>k(d)}\right)}\right) \leq \mathcal{O}_{n,d}\left(1\right).$$

Combining this with Thm. 4.1, it holds that for every $\delta > 0$, w.p at least $1 - \delta - 16\exp\left(-\frac{c'}{\beta_{k(d)}^2}\frac{n}{k(d)}\right)$, both the variance and bias can be upper bounded as

$$V \leq \sigma_\epsilon^2\mathcal{O}_{n,d}\left(\left(\frac{k(d)}{n} + \frac{n}{R_k(\Sigma)}\right)\right) \leq \sigma_\epsilon^2\mathcal{O}_{n,d}\left(\frac{1}{d^{\tau-\lfloor\tau\rfloor}} + \frac{1}{d^{\lfloor\tau\rfloor+1-\tau}}\right). \tag{44}$$

$$B \leq \frac{1}{\delta}\mathcal{O}_{n,d}\left(\left\|\theta^*_{>k(d)}\right\|^2_{\Sigma_{>k(d)}}\right) + \mathcal{O}_{n,d}\left(\left\|\theta^*_{\leq k(d)}\right\|^2_{\Sigma^{-1}_{\leq k(d)}}\left(\frac{\text{tr}\left(\Sigma_{>k(d)}\right)}{n}\right)^2\right). \tag{45}$$

Using the fact that $\frac{c'}{\beta_{k(d)}^2}\frac{n}{k(d)} = \omega_d(\log(d))$ the probability becomes $1 - \delta - o_d\left(\frac{1}{d}\right)$

Now in order to further bound the bias, we first note that by the addition theorem Eq. (51) it holds that

$$\text{tr}\left(\Sigma\right) = \sum_{\ell=0}^{\infty}\sigma_\ell N(d,\ell) = h(1) = \Theta_{n,d}(1). \tag{46}$$

As in the statement of the lemma, let $N_d := k(d)$. Because $i \in \mathbb{N}$, $N(d,i) = \Theta_d\left(d^i\right)$ and by assumption, $\hat{\sigma}_{\lfloor\tau\rfloor} \neq 0$, it holds that $k(d) = \mathcal{O}_{n,d}\left(d^{\lfloor\tau\rfloor}\right)$. Combining this with Eq. (46) and the fact that for all $i \leq k(d)$, $\lambda_i \geq \min_{\ell\leq\lfloor\tau\rfloor \text{ s.t. } \hat{\sigma}_\ell\neq0}\hat{\sigma}_\ell \cdot \Omega_{n,d}\left(\frac{1}{d^{\lfloor\tau\rfloor}}\right)$ the right hand side of Eq. (45) can be bounded as

$$\begin{aligned}
\left\|\theta^*_{\leq k(d)}\right\|^2_{\Sigma^{-1}_{\leq k(d)}}\left(\frac{\text{tr}\left(\Sigma_{>k(d)}\right)}{n}\right)^2 &= \sum_{i\leq k(d)}\frac{(\theta^*_i)^2}{\lambda_i}\left(\frac{\text{tr}\left(\Sigma_{>k(d)}\right)}{n}\right)^2\\
&\leq \frac{k(d)}{\min_{i\leq k(d)}\lambda_i}\left\|\theta^*_{\leq N_d}\right\|^2_\infty\left(\frac{\text{tr}\left(\Sigma\right)}{n}\right)^2\\
&\leq \left\|\theta^*_{\leq N_d}\right\|^2_\infty\frac{1}{\min_{\ell\leq\lfloor\tau\rfloor \text{ s.t. } \hat{\sigma}_\ell\neq0}\hat{\sigma}_\ell}\cdot\mathcal{O}_{n,d}\left(\frac{1}{d^{2(\tau-\lfloor\tau\rfloor)}}\right).
\end{aligned}$$

The left hand side of Eq. (45) can be bounded as

$$\frac{1}{\delta}\left\|\theta^*_{>k(d)}\right\|^2_{\Sigma_{>k(d)}} \leq \frac{1}{\delta}\left\|\theta^*_{>k(d)}\right\|^2_\infty\text{tr}\left(\Sigma_{>k(d)}\right) = \frac{1}{\delta}\mathcal{O}_{n,d}\left(\left\|\theta^*_{>N_d}\right\|^2_\infty\right).$$

So Eq. (45) becomes

$$B \leq \frac{1}{\delta}\mathcal{O}_{n,d}\left(\left\|\theta^*_{>N_d}\right\|^2_\infty\right) + \left\|\theta^*_{\leq N_d}\right\|^2_\infty\max_{\ell\leq\lfloor\tau\rfloor \text{ s.t. } \hat{\sigma}_\ell\neq0}\frac{1}{\hat{\sigma}_\ell}\cdot\mathcal{O}_{n,d}\left(\frac{1}{d^{2(\tau-\lfloor\tau\rfloor)}}\right).$$

□

E.4. Lemmas for Applications

Lemma E.1. *For any $a > 0$,*

1. *If $c_1\frac{1}{i\log^{1+a}(i)} \leq \lambda_i \leq c_2\frac{1}{i\log^{1+a}(i)}$ then $\frac{c_1}{c_2}\frac{1}{a}(k+1)\log(k+1) \leq r_k \leq 1 + \frac{c_2}{c_1}\frac{1}{a}(k+1)\log(k+1)$.*

2. *If* $c_1\frac{1}{i^{1+a}} \leq \lambda_i \leq c_2\frac{1}{i^{1+a}}$ *then* $\frac{c_1}{c_2}\frac{1}{a}(k+1) \leq r_k \leq 1+\frac{c_2}{c_1}\frac{1}{a}(k+1)$.
3. *If* $c_1\frac{1}{e^{ai}} \leq \lambda_i \leq c_2\frac{1}{e^{ai}}$ *then* $\frac{c_1}{c_2}\frac{1}{a} \leq r_k \leq 1+\frac{c_2}{c_1}\frac{1}{a}$.

Proof. The famous integral test for convergence states that for a monotonic decreasing function $f(n)$, it holds for any $k \in \mathbb{N}$ that

$$\int_{k+1}^{\infty} f(x)dx \leq \sum_{i>k} f(i) \leq f(k+1) + \int_{k+1}^{\infty} f(x)dx,$$

We now split into separate cases of eigenvalue decay.

1. If $c_1\frac{1}{i\log^a(i)} \leq \lambda_i \leq c_2\frac{1}{i\log^a(i)}$ then using the fact that $\int_{k+1}^{\infty}\frac{1}{x\log^{1+a}(x)}dx = \frac{1}{a\log^a(k+1)}$ we obtain

   $$r_k \leq 1+\frac{1}{c_1\lambda_{k+1}}\int_{k+1}^{\infty} c_2\frac{1}{x\log^{1+a}(x)}dx \leq 1+\frac{c_2}{c_1}\frac{1}{a}(k+1)\log(k+1),$$

   and

   $$r_k \geq \frac{1}{c_2\lambda_{k+1}}\int_{k+1}^{\infty} c_a\frac{1}{x\log^{1+a}(x)}dx \geq \frac{c_1}{c_2}\frac{1}{a}(k+1)\log(k+1).$$

2. If $c_1\frac{1}{i^{1+a}} \leq \lambda_i \leq c_2\frac{1}{i^{1+a}}$ then using the fact that $\int_{k+1}^{\infty}\frac{1}{x^{1+a}(x)}dx = \frac{1}{a(k+1)^a}$ we obtain that

   $$r_k \leq 1+\frac{1}{c_1\lambda_{k+1}}\int_{k+1}^{\infty} c_2\frac{1}{x^{1+a}(x)}dx \leq 1+\frac{c_2}{c_1}\frac{1}{a}(k+1),$$

   and

   $$r_k \geq \frac{1}{c_2\lambda_{k+1}}\int_{k+1}^{\infty} c_1\frac{1}{x^{1+a}(x)}dx \geq \frac{c_1}{c_2}\frac{1}{a}(k+1).$$

3. If $c_1\frac{1}{e^{ai}} \leq \lambda_i \leq c_2\frac{1}{e^{ai}}$ then using the fact that $\int_{k+1}^{\infty}\exp(-ax)dx = \frac{1}{ae^{a(k+1)}}$ we obtain that

   $$r_k \leq 1+\frac{1}{c_1\lambda_{k+1}}\int_{k+1}^{\infty} c_2\exp(-ax)dx \leq 1+\frac{c_2}{c_1}\frac{1}{a},$$

   and

   $$r_k \geq \frac{1}{c_2\lambda_{k+1}}\int_{k+1}^{\infty} c_1\exp(-ax)dx \geq \frac{c_1}{c_2}\frac{1}{a}.$$

□

Lemma E.2. *Let K be a kernel with polynomially decaying eigenvalues $\lambda_i = \Theta_{i,n}(i^{-1-a})$ for some $a > 0$. Furthermore, suppose that $\frac{\beta_k k\log(k)}{n} = \mathcal{O}_{k,n}(1)$ and that $\beta_k = \mathcal{O}_k(1)$. Then it holds w.p at least $1-\mathcal{O}_{k,n}\left(\frac{1}{k^3}\right)\exp\left(-\Omega_{k,n}(\frac{n}{k})\right)$ that*

$$\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) = \mathcal{O}_{k,n}\left(\lambda_{k+1}\right)$$

Proof. By Lemma E.1, it holds that $r_k(\Sigma), r_k(\Sigma^2) = \Theta_{k,n}(k)$. Now using Corollary C.5 (note that Assumption 2.3 holds since $\beta_k = \mathcal{O}_k(1)$), there exist absolute constants $c, c' > 0$ s.t it holds w.p at least $1-4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right)$ that

$$\begin{aligned}\mu_1\left(\frac{1}{n}\mathbf{K}_{>k}\right) \leq& c\left(\lambda_{k+1}+\beta_k\log(k+1)\frac{\text{tr}\left(\Sigma_{>k}\right)}{n}\right)\\ =&\mathcal{O}_{k,n}\left(\lambda_{k+1}\left(1+\beta_k\log(k+1)\frac{r_k}{n}\right)\right)\\ =&\mathcal{O}_{k,n}\left(\lambda_{k+1}\left(1+\frac{\beta_k k\log(k)}{n}\right)\right) = \mathcal{O}_{k,n}\left(\lambda_{k+1}\right).\end{aligned}$$

Now to bound the probability which this holds, we use the fact that $r_k = \Theta_{k,n}(k)$ together with the fact that $\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right) < 1$ to get the claim holds w.p at least $1 - 4\frac{r_k}{k^4}\exp\left(-\frac{c'}{\beta_k}\frac{n}{r_k}\right) = 1 - \mathcal{O}_{k,n}\left(\frac{1}{k^3}\right)\exp\left(-\Omega_{k,n}(\frac{n}{k})\right)$. □

Lemma E.3. *Let $a \in \mathbb{R}$, $1 < k \in \mathbb{N}$, then*

$$\sum_{i \le k} i^{-a} \le \begin{cases} 1 + k^{1-a} & a < 1 \\ 1 + \log(k) & a = 1 \\ \frac{1}{a-1} & a > 1 \end{cases}$$

Proof. If $a < 0$, then bounding the mean with the maximum yields $\sum_{i \le k} i^{-a} \le k \cdot k^{-a} = k^{1-a}$. Next, if $a \neq 1$, bounding the sum with the integral yields

$$\sum_{i \le k} i^{-a} \le 1 + \int_1^k \frac{1}{x^a} dx = 1 + \frac{1}{a-1} - \frac{k^{1-a}}{a-1}.$$

So if $a < 1$, we obtain a $1 + k^{1-a}$ bound, and if $a > 1$, a $1 + \frac{1}{a-1}$ bound. Lastly, if $a = 1$ then we can similarly bound as

$$\sum_{i \le k} i^{-a} \le 1 + \int_1^k \frac{1}{x} dx = 1 + 1 + \log(k).$$

□

Lemma E.4. *Let $1 < k \in \mathbb{N}$ and suppose that $\lambda_i = \Theta_{i,n}\left(\frac{1}{i^{1+a}}\right)$ for some $a > 0$, and $\theta_i^* = \Theta_{i,n}\left(i^{-r}\right)$ for some $r \in \mathbb{R}$ s.t $f^* \in L_\mu^2(\mathcal{X})$. It holds that*

$$\|\theta_{>k}^*\|_{\Sigma_{>k}}^2 \le \mathcal{O}_{k,n}\left(\frac{1}{k^{2r+a}}\right), \qquad \|\theta_{\le k}^*\|_{\Sigma_{\le k}^{-1}}^2 \le \begin{cases} \mathcal{O}_{k,n}\left(k^{-2r+2+a}\right) & 2r < 2+a \\ \mathcal{O}_{k,n}\left(\log(k)\right) & 2r = 2+a \\ \mathcal{O}_{k,n}\left(1\right) & 2r > 2+a \end{cases}.$$

Proof. The condition that $f^* \in L_\mu^2(\mathcal{X})$ implies $\sum_{i=1}^\infty \theta^* \lambda_i^2 = \|\langle \theta^* \phi(\mathbf{x})\rangle\| < \infty$. The $> k$ part can be bounded using Lemma E.1 as

$$\|\theta_{>k}^*\|_{\Sigma_{>k}}^2 = \sum_{i>k} (\theta_i^*)^2 \lambda_i = \mathcal{O}_{k,n}\left(\sum_{i>k} i^{-2r-1-a}\right) \le \mathcal{O}_{k,n}\left(\frac{1}{k^{2r+a}}\right).$$

The $\le k$ part can be bounded using lemma Lemma E.3 (with $2r - 1 - a$) as

$$\|\theta_{\le k}^*\|_{\Sigma_{\le k}^{-1}}^2 = \sum_{i \le k} \frac{(\theta_i^*)^2}{\lambda_i} = \mathcal{O}_{k,n}\left(\sum_{i \le k} i^{-2r+1+a}\right) \le \begin{cases} \mathcal{O}_{k,n}\left(k^{-2r+2+a}\right) & 2r < 2+a \\ \mathcal{O}_{k,n}\left(\log(k)\right) & 2r = 2+a \\ \mathcal{O}_{k,n}\left(1\right) & 2r > 2+a \end{cases}.$$

□

F. Lack of Sub Gaussianity

Suppose our inputs are one-dimensional standard Gaussians $x \sim \mathcal{N}(0, \sigma^2)$ and let $K(x,y) = \exp\left(-\gamma(x-y)^2\right)$ be the Gaussian (RBF) kernel. Such kernels have known Mercer decompositions (Fasshauer, 2011), and if we pick for simplicity $\sigma = 1$ and $\gamma = \frac{3}{8}$ (meaning that in their notation, $\alpha = \frac{1}{\sqrt{2}}$ and $\epsilon = \sqrt{\frac{3}{8}}$) we obtain that $\psi(x) = (\psi_i(x))_{i=0}^\infty$ is given by:

$$\psi_i(x) = \frac{\sqrt[4]{2}}{\sqrt{2^i i!}} e^{-\frac{x^2}{4}} H_i(x), \tag{47}$$

where $H_i(x) = (-1)^i e^{x^2} \frac{d^i}{dx^i} e^{-x^2}$ is the i'th order (physicist's) Hermite polynomial. Note that in this chapter, for ease of notation, we start counting at $i = 0$.

Recall that a vector Y is said to be sub-Gaussian if

$$\sup_{u:\|u\|=1} \sup_{p\geq 1} \frac{1}{\sqrt{p}} \left(\mathbb{E}\left[|\langle u, Y\rangle|^p\right]\right)^{1/p} < \infty.$$

In particular, taking $Y = \psi$ and $u = e_i$ we get that:

$$\begin{aligned}\mathbb{E}\left[|\langle u, Y\rangle|^p\right] =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |\psi_i(x)|^p \, e^{-\frac{x^2}{2}} dx \\ =& \frac{2^{\frac{p}{4}-\frac{1}{2}}}{\sqrt{\pi}\left(2^i i!\right)^{p/2}} \int_{-\infty}^{\infty} |H_i(x)|^p \exp\left(-\left(\frac{p}{4}+\frac{1}{2}\right)x^2\right) dx \end{aligned} \tag{48}$$

Thus, if for a fixed p, The value of Eq. (48) diverges to infinity with i, it would imply that ψ is not sub-Gaussian.

We will thus aim to lower bound this term. To do so, we begin by bounding the Hermite polynomials using Szeg (1939)[Theorem 8.22.9], which states that for any $\delta > 0$, and any $x = \sqrt{2i+1}\cos(\phi)$ where $\delta \leq \phi \leq \pi - \delta$, we have the uniform approximation:

$$\begin{aligned} e^{-\frac{x^2}{2}} H_i(x) =& 2^{\frac{i}{2}+\frac{1}{4}} \sqrt{i!} (\pi i)^{-\frac{1}{4}} \\ & \times \underbrace{\sin(\phi)^{-\frac{1}{2}}}_{:=A} \left(\underbrace{\sin\left(\frac{3\pi}{4} + \left(\frac{2i+1}{4}\right)(\sin(2\phi) - 2\phi)\right)}_{:=B} + \mathcal{O}(i^{-1}) \right). \end{aligned} \tag{49}$$

We now wish to bound B. Since $\sin(\phi) \geq 0.5$ for $\phi \in [\frac{1}{6}\pi, \frac{5}{6}\pi]$ then we can lower bound B by 0.5 when

$$\frac{3\pi}{4} + \left(\frac{2i+1}{4}\right)(\sin(2\phi) - 2\phi) \in \left[\frac{1}{6}\pi, \frac{5}{6}\pi\right].$$

This is equivalent to:

$$-\frac{1}{6(2i+1)}\pi \leq \phi - \frac{\sin 2\phi}{2} \leq \frac{7}{6(2i+1)}\pi.$$

Since $\phi \geq 0$, we have (via the sin Taylor expansion) that $\phi - \frac{\phi^3}{6} \leq \sin(\phi) \leq \phi$ (meaning $-\phi \leq -\frac{\sin 2\phi}{\phi} \leq -\phi + \frac{8\phi^3}{6}$) and so the lower bound holds trivially and the upper bound holds when $\phi \leq \sqrt[3]{\frac{7}{8(2i+1)}\pi}$.

We can also lower bound A trivially by 1. Furtheremore, for i sufficiently large the $\mathcal{O}(i^{-1})$ is at least $-\frac{1}{4}$. So overall we obtain that for $\phi \in \left[\delta, \sqrt[3]{\frac{7}{8(2i+1)}\pi}\right]$ and $x = \sqrt{2i+1}\cos(\phi)$, $A(B + \mathcal{O}(i^{-1})) \geq \frac{1}{4}$, and Eq. (49) can be lower bounded as:

$$H_i(x) \geq \frac{1}{4} 2^{\frac{i}{2}+\frac{1}{4}} \sqrt{i!} (\pi i)^{-\frac{1}{4}} e^{\frac{x^2}{2}} = \frac{1}{4}\left(\frac{2}{\pi}\right)^{\frac{1}{4}} 2^{\frac{i}{2}} \sqrt{i!} i^{-\frac{1}{4}} e^{\frac{x^2}{2}}.$$

So for any $p \in \mathbb{N}$, we can lower bound the p'th power of H_i as

$$H_i(x)^p \geq \frac{1}{4^p}\left(\frac{2}{\pi}\right)^{\frac{p}{4}} \left(2^i i!\right)^{p/2} i^{-\frac{p}{4}} e^{\frac{px^2}{2}}.$$

Denoting $a_i = \sqrt{2i+1}\cos\left(\sqrt[3]{\frac{7}{8(2i+1)}\pi}\right)$ and $b_i = \sqrt{2i+1}\cos(\delta)$ we can bound our expected value in $Eq.$ (48) by:

$$\begin{aligned}
\mathbb{E}\left[|\langle u, Y\rangle|^p\right] =& \frac{2^{\frac{p}{4}-\frac{1}{2}}}{\sqrt{\pi}\left(2^i i!\right)^{p/2}} \int_{-\infty}^{\infty} |H_i(x)|^p \exp\left(-\left(\frac{p}{4}+\frac{1}{2}\right)x^2\right) dx \\
\geq& \left(\frac{2}{\pi}\right)^{\frac{p}{4}-\frac{1}{2}} \frac{2^{\frac{p}{4}}}{4^p i^{p/4}} \int_{a_i}^{b_i} \exp\left(\left(\frac{p}{4}-\frac{1}{2}\right)x^2\right) dx \\
\geq& \Omega_i\left(i^{-\frac{3}{2}} \int_{a_i}^{b_i} \exp\left(\frac{p-2}{4}x^2\right) dx\right) \\
\geq& \Omega_i\left(i^{-\frac{3}{2}}(b_i - a_i)\exp\left(\frac{p-2}{4}a_i^2)\right)\right)
\end{aligned}$$

By continuity in δ we can take $b_i = \sqrt{2i+1}\cos(0) = \sqrt{2i+1}$ and by using the inequality (via the $\cos$ Maclaurin expansion) $\cos(t) \leq 1 - \frac{t^2}{2} + o(t^2)$ we get

$$\begin{aligned}
b_i - a_i =& \sqrt{2i+1}(1 - \cos\left(\sqrt[3]{\frac{7}{8(2i+1)}\pi}\right) \\
\geq& \sqrt{2i+1}\left(\frac{1}{2}\sqrt[3]{\frac{7}{8(2i+1)}\pi}^{2} - o\left(\sqrt[3]{\frac{7}{8(2i+1)}\pi}^{2}\right)\right) \\
=& \Omega_i(\sqrt{i}i^{-\frac{2}{3}}) = \Omega_i(i^{-\frac{1}{6}}).
\end{aligned}$$

Finally, since for sufficiently large i, $a_i^2 > \frac{3}{2}i$ (since the $\cos$ part of a_i tends to 1), for any $p \geq 3$ we obtain

$$\mathbb{E}\left[|\langle u, Y\rangle|^p\right] = \Omega_i\left(i^{-\frac{3}{2}}i^{-\frac{1}{6}}\exp\left(\frac{p-2}{4}\cdot\frac{3}{2}i\right)\right) = \Omega_i\left(\exp\left(\frac{p-2}{4}\cdot i\right)\right) \xrightarrow[i\to\infty]{} \infty.$$

This implies that ψ is not sub-Gaussian.

G. Background on Dot Product and Zonal Kernels

A Kernel K is called a dot product kernel if $K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x}^\top\mathbf{x}')$ for some $h : \mathbb{R} \to \mathbb{R}$ which has a Taylor expansion of the form $h(t) = \sum_{i=0}^{\infty} a_i t^i$ with $a_i \geq 0$. Importantly, K depends only on $\mathbf{x}^\top\mathbf{x}'$. With inputs uniformly distributed on $\mathbb{S}^{d-1}$, this family of kernels includes the NTK, Laplace kernel, Gaussian (RBF) kernel, and polynomial kernel (Minh et al., 2006; Bietti & Bach, 2020; Chen & Xu, 2020). We emphasize that for an L layer fully connected network $f(\mathbf{x};\theta)$, KRR with respect to the corresponding GPK $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_\theta[f(\mathbf{x};\theta) \cdot f(\mathbf{x}';\theta)]$ (also called Conjugate Kernel or NNGP Kernel) is equivalent to training the final layer while keeping the weights of the other layers at their initial values (Lee et al., 2017). Furthermore, KRR with respect to the NTK $\Theta\left(\mathbf{x}, \mathbf{x}'\right) = \mathbb{E}_\theta\left[\left\langle \frac{\partial f(\mathbf{x};\theta)}{\partial\theta}, \frac{\partial f\left(\mathbf{x}';\theta\right)}{\partial\theta}\right\rangle\right]$ is equivalent to training the entire network (Jacot et al., 2018).

Under a uniform distribution on $\mathbb{S}^{d-1}$, the domain of h is $[-1, 1]$, and for any $d \geq 3$ dot product kernels exhibit the Mercer decomposition

$$K(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^{\infty} \frac{\hat{\sigma}_\ell}{N(d,\ell)} \sum_{m=1}^{N(d,\ell)} Y_{\ell,m}(\mathbf{x}) Y_{\ell,m}(\mathbf{x}'), \tag{50}$$

where the eigenfunctions $Y_{\ell,m}$ are the m'th spherical harmonic of degree (or frequency) ℓ, $N(d,\ell) = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{d-2}$ is the number of harmonics of each degree, and $\sigma_\ell := \frac{\hat{\sigma}_\ell}{N(d,\ell)}$ are the eigenvalues (Smola et al., 2000). Each spherical harmonic can be defined via restrictions of homogeneous polynomials to the unit sphere, with the degree (or frequency) of the spherical harmonic corresponding to the degree of said polynomials. When $d \gg \ell$, $N(d,\ell) = \Theta_d(d^\ell)$ and when

$\ell \gg d, N(d,\ell) = \Theta_\ell(\ell^{d-2})$. Importantly, all spherical harmonics $Y_{\ell,m}$ of the same degree ℓ share the same eigenvalue σ_ℓ, and as a result, there are many repeated eigenvalues. For background on spherical harmonics, see Dai (2013); Atkinson & Han (2012); Smola et al. (2000). In order to write the kernel as Eq. (1), we can order ϕ in the natural way, by first taking $\tilde{\phi}(\mathbf{x}) = (\sqrt{\sigma_0}Y_{0,1}, \sqrt{\sigma_1}Y_{1,1}, \ldots, \sqrt{\sigma_1}Y_{1,N(d,1)}, \sqrt{\sigma_2}Y_{2,1}, \ldots)$, and letting ϕ be the same as $\tilde{\phi}$ with zero-valued indices removed (where $\sigma_\ell = 0$). We let ψ be given accordingly. We note that $\psi_1 = Y_{0,1}$ is a constant function.

The famous addition theorem (Dai, 2013)[1.2.8 and 1.2.9] implies that for any $d \geq 3, \mathbf{x}, \mathbf{x}' \in \mathbb{S}^{d-1}$ and $\ell \geq 0$,

$$\sum_{m=1}^{N(d,\ell)} Y_{\ell,m}(\mathbf{x})Y_{\ell,m}(\mathbf{x}) = N(d,\ell). \tag{51}$$

For any $\ell \in \mathbb{N}$, let $N(d, \leq \ell) = \sum_{j=1}^{\ell} N(d,\ell)$. The addition theorem Eq. (51) in particular implies that the eigenfunctions ψ_i are highly correlated, and definitely not i.i.d. Importantly, Eq. (51) implies that

$$\text{For any } \ell \in \mathbb{N}, k := N(d, \leq \ell), \text{ it holds that } \beta_k = \alpha_k = 1. \tag{52}$$

Furthermore, for any $k \in \mathbb{N}$, let $\ell_k = \max\{\ell \in \mathbb{N} \cup \{0\} \text{ s.t. } N(d, \leq \ell) \leq k\}$, so that $N(d, \leq \ell_k) \leq k \leq N(d, \leq \ell_k + 1)$. If momentarily we consider the case when $\hat{\sigma}_\ell \neq 0$ for all ℓ, then from Eq. (51), it holds that for any $\mathbf{x} \in \mathbb{S}^{d-1}$,

$$\Theta_k(1) = \frac{N(d, \leq \ell_k)}{N(d, \leq \ell_{k+1})} \leq \frac{\|\psi_{\leq k}(\mathbf{x})\|^2}{k} \leq \frac{N(d, \leq \ell_{k+1})}{N(d, \leq \ell_k)} = \Theta_k(1).$$

Implying that $\frac{\|\psi_{\leq k}(\mathbf{x})\|^2}{k} = \Theta_k(1)$. A similar argument yields

$$1 - \frac{\hat{\sigma}_{\ell_k}}{\sum_{\ell=\ell_k}^{\infty} \hat{\sigma}_\ell} = \frac{\sum_{\ell=\ell_k+1}^{\infty} \hat{\sigma}_\ell}{\sum_{\ell=\ell_k}^{\infty} \hat{\sigma}_\ell} \leq \frac{\|\phi_{>k}(\mathbf{x})\|^2}{\operatorname{tr}(\Sigma_{>k})} \leq \frac{\sum_{\ell=\ell_k}^{\infty} \hat{\sigma}_\ell}{\sum_{\ell=\ell_k+1}^{\infty} \hat{\sigma}_\ell} \leq 1 + \frac{\hat{\sigma}_{\ell_k}}{\sum_{\ell=\ell_k+1}^{\infty} \hat{\sigma}_\ell},$$

which analogously to Lemma E.1 will typically be $\Theta_k(1)$ if the decay of $\hat{\sigma}$ is at most exponential (but may be slower). This is the case for common kernels such as NTK, Laplace and RBF, and for such kernels we obtain:

$$\alpha_k, \beta_k = \Theta_k(1). \tag{53}$$

H. Examples of Kernels That Fit Our Framework

Here, we provide some simple examples of kernels that fit our framework. Namely, that β_k and possibly α_k (as defined in Def. (2.1)) can be bounded. First, note that for each of the terms in Def. (2.1), the denominator is the expected value of the numerator, so α_k and β_k quantify how much the features behave as they are "supposed to". Since $\inf \leq \mathbb{E} \leq \sup$, one always has

$$0 \leq \alpha_k \leq 1 \leq \beta_k. \tag{54}$$

A control on β_k is usually easier than one on α_k. Nevertheless, bounding α_k may be made easier by Remark 2.2. We also mention that bounds on α_k, β_k in one domain can often be extended to others. See Sec. 5.1 for details.

- Dot Product Kernels on $\mathbb{S}^{d-1}$: A complete treatment of such kernels is given in Appendix G.
- Kernels With Bounded Eigenfunctions: If $\psi_i^2(\mathbf{x}) < M$ for any $i, \mathbf{x}$ the it trivially holds that $\beta_k \leq M$ for any $k \in \mathbb{N}$. Analogously, if $\psi_i^2 \geq M'$ then $\alpha_k \geq M'$. This may be weakened to a high probability lower bound (see Remark 2.2).
- RBF and shift-invariant kernels in $X \subseteq \mathbb{R}^d$: The features ϕ_i for an RBF kernel on $X \subseteq \mathbb{R}^d$ with nonempty interior (i.e $X^\circ \neq \emptyset$) are given by (Steinwart et al., 2006)[Theorem 3.7]. If for simplicity $X \subseteq [-1, 1]$, then ϕ_i are bounded, implying that ψ_i are also bounded. Hence, by the previous item, $\beta_k = \mathcal{O}_{k,n}(1)$. A simple and easy-to-understand construction of the Mercer Decomposition for general shift-invariant kernels on $[0, 1]$ is provided in Mairal & Vert.
- Kernels on the Hypercube $\{-1, 1\}^d$: With a uniform distribution, the hypercube has a Fourier decomposition given by monomials (O'Donnell, 2014). As a result, for kernels of the form $K(\mathbf{x}, \mathbf{x}') = h\left(\frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|\mathbf{x}\|\|\mathbf{x}'\|}, \frac{\|\mathbf{x}\|^2}{d}, \frac{\|\mathbf{x}'\|^2}{d}\right)$ for some $h : \mathbb{R}^3 \to \mathbb{R}$, the eigenfunctions ψ_i are given by monomials (Yang & Salman, 2019). In particular, for any i, $\psi_i^2 \equiv 1$ and thus $\alpha_k = \beta_k = 1$ for any k.

I. Experiments

We plot the variance for a 3-layer fully connected NTK and polynomial kernel in Fig. 1 and 3-layer fully connected GPK in Fig. 2. Background on the NTK and GPK is given in Appendix G; however, we note here that there is a closed form for the expectations (Jacot et al., 2018; Lee et al., 2019; Bietti & Bach, 2020), which we used when computing the figures. First, let

$$\kappa_0(u) := \frac{1}{\pi}(\pi - \arccos(u)), \qquad \kappa_1(u) := \frac{1}{\pi}\left(u\left(\pi - \arccos(u)\right) + \sqrt{1-u^2}\right).$$

The L layer GPK on $\mathbb{S}^{d-1}$ is equal to

$$K_{\text{GPK}}^{(L)}(\mathbf{x}, \mathbf{x}') := \kappa_1\left(K_{\text{GPK}}^{(L-1)}(\mathbf{x}, \mathbf{x}')\right), \qquad K_{\text{GPK}}^{(0)}(\mathbf{x}, \mathbf{x}') := \mathbf{x}^\top \mathbf{x}',$$

and the L layer NTK on $\mathbb{S}^{d-1}$ is

$$\Theta^{(L)}(\mathbf{x}, \mathbf{x}') := \Theta^{(L-1)}(\mathbf{x}, \mathbf{x}')\kappa_0\left(K_{\text{GPK}}^{(L-1)}(\mathbf{x}, \mathbf{x}')\right) + K_{\text{GPK}}^{(L)}(\mathbf{x}, \mathbf{x}'), \qquad \Theta_{\text{GPK}}^{(0)}(\mathbf{x}, \mathbf{x}') := K_{\text{GPK}}^{(0)}(\mathbf{x}, \mathbf{x}').$$

J. Further Details on Related Works

We now continue the discussion from Sec. 2.1.

Regarding the differences between the high-dimensional and low-dimensional settings, we note that the techniques and assumptions used by these two lines of work are inherently different, and make the results from the high-dimensional works inapplicable for fixed d and vice versa. For example, high-dimensional works typically rely on tools from random matrix theory, which require d and n to be tied and are inapplicable for a fixed d. By contrast, low-dimensional works have bounds that depend on the properties of the fixed RKHS, and often assume a fixed polynomial decay for the eigenvalues λ_i. This not only excludes kernels with an exponential decay such as RBF (Minh et al., 2006) but is also problematic, for example, for analyzing the NTK with high-dimensional inputs, since the polynomial decay only begins when the eigenvalue index is $i \gg \text{poly}(d)$ (Cao et al., 2019). By contrast, we obtain bounds that are relevant for *any* d, n, regardless of the ratio between them, and in particular, capture interesting phenomena in these two regimes.

Regarding works that bound the eigenvalues of kernel matrices similarly to what we do here, Braun (2005); Rosasco et al. (2010); Valdivia (2018) provide generic bounds; however, they are not sufficiently strong for many applications and, in particular, often do not yield nontrivial bounds for the smallest eigenvalue of the kernel matrix. As we shall see, this will be crucial for our analysis. Fan & Wang (2020); Montanari & Zhong (2022) provide lower bounds for the smallest eigenvalue when the input dimension is linear in the number of samples and tends towards infinity. For fully-connected NTKs, Oymak & Soltanolkotabi (2020); Wang & Zhu (2021) provide bounds for two-layer networks, and Nguyen et al. (2021) provide bounds for deep networks for large input dimensions. Belkin (2018) gives bounds for radial kernels such as RBF.

K. NTK - Neural Network Correspondence

For many architectures, under suitable initialization and learning rate, gradient decent with sufficiently wide neural networks is equivalent to kernel regression with the NTK (Jacot et al., 2018; Lee et al., 2019; Yang & Littwin, 2021). Specifically, for a neural network $f(\mathbf{x}, \theta)$, one can typically bound its distance from its first order Taylor approximation $f^{\text{lin}}(\mathbf{x}, \theta)$ at time t of gradient flow as $\sup_{t\geq 0}\left|f(\mathbf{x}, \theta_t) - f^{\text{lin}}(\mathbf{x}, \theta_t)\right| \leq O\left(\frac{1}{\sqrt{\text{width}}}\right)$ (Lee et al., 2019; Bowman & Montufar, 2022). Furthermore, training $f^{\text{lin}}(\mathbf{x}, \theta)$ for time t is roughly equivalent to kernel regression with regularization $\gamma_n = \frac{1}{t}$ (Ali et al., 2019). By combining the two, one can easily bound the difference in generalization errors between neural networks trained for time t and kernel regression with the NTK and regularization $\gamma_n = \frac{1}{t}$.