Precise Learning Curves and Higher-Order Scaling Limits for Dot Product Kernel Regression

Lechao Xiao*
Google Research, Brain Team
xlc@google.com

Hong Hu*
University of Pennsylvania
huhong@wharton.upenn.edu

Theodor Misiakiewicz*
Stanford University
misiakie@stanford.edu

Yue M. Lu
Harvard University
yuelu@seas.harvard.edu

Jeffrey Pennington
Google Research, Brain Team
jpennin@google.com

Abstract

As modern machine learning models continue to advance the computational frontier, it has become increasingly important to develop precise estimates for expected performance improvements under different model and data scaling regimes. Currently, theoretical understanding of the learning curves that characterize how the prediction error depends on the number of samples is restricted to either large-sample asymptotics ($m \to \infty$) or, for certain simple data distributions, to the high-dimensional asymptotics in which the number of samples scales linearly with the dimension ($m \propto d$). There is a wide gulf between these two regimes, including all higher-order scaling relations $m \propto d^r$, which are the subject of the present paper. We focus on the problem of kernel ridge regression for dot-product kernels and present precise formulas for the mean of the test error, bias, and variance, for data drawn uniformly from the sphere with isotropic random labels in the rth-order asymptotic scaling regime $m \to \infty$ with m/d^r held constant. We observe a peak in the learning curve whenever $m \approx d^r/r!$ for any integer r, leading to multiple sample-wise descent and nontrivial behavior at multiple scales.

1 Introduction

Modern machine learning has entered an era in which scaling is arguably the most critical ingredient to improve performance. Recent breakthroughs such as GPT-3 [24] and PaLM [11] have demonstrated that performance of various learning algorithms improves in a *predictable* manner as the amount of data and computational resources used in training increases. The functional relationships between performance and resources are loosely referred to as learning curves. While extrapolation of empirical learning curves is widely used to make predictions about how a model might perform when extra resources become available, a rigorous theoretical understanding is lacking. A fundamental obstacle in developing a detailed theoretical model of such learning curves is that they depend on many moving parts, e.g. the data distribution, the network architecture, the training algorithm, among others. In addition, even in the simplest possible settings, the learning curves can exhibit non-trivial structure that naive scaling laws fail to model, e.g. the well-known double-descent phenomenon [7, 3].

In the past couple years, a large amount of effort from the community has improved our theoretical understanding of such phenomena and in some cases precise characterizations of learning curves have been obtained (see e.g., [21, 1, 31, 34, 28]). These results have helped clarify several puzzling empirical observations, such as the origin of the double-descent peak [2, 27, 13] and linear trends

*LX, HH and TM contributed equally. HH's work was done while at Harvard University.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

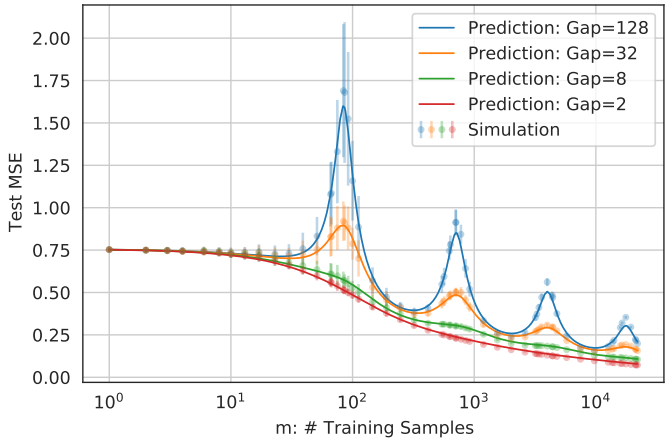


Figure 1: **Precise Sample-wise Learning Curves for One-hidden Layer CNN kernels.** The theoretical predictions (Eq. (18), solid lines) agree with finite-size simulations (markers) across several orders of magnitude and captures cases in which the curves are relatively simple (**monotonically decreasing**, small spectral gap) and complex (**multiple-descent**, large spectral gap). Simulations are obtained from kernel regression with one-layer CNN kernels averaged over 50 runs. The input is of shape $d = d_0 \times p$ with size $d_0 = 14$ and number of patches $p = 6$. We vary the kernels by varying the ratio (aka spectral gap) between consecutive eigenspaces, where the ratio $\text{Gap} \in [\mathbf{2}, \mathbf{8}, \mathbf{32}, \mathbf{128}]$.

between in- and out-of-distribution generalization performance [38, 39, 30], among many others. However, the precise predictions from many of these analyses have been possible only in the linear high-dimensional scaling regime in which the number of training samples m scales linearly with the dimension d, i.e. $m \propto d$. In these asymptotics, the model's effective capacity is limited to linear functions of the features. In contrast, many state-of-the-art models operate in a regime where the amount of data is much larger than the data dimensionality; for example, large text corpora can contain trillions of tokens, whereas the effective input dimensionality of language models is at most millions . Therefore, going beyond the linear scaling regime ($m \propto d$) to higher-order scaling regimes ($m \propto d^r$) is essential in improving our understanding of modern machine learning systems, and is the focus of the current paper.

Several works have investigated the behavior of the learning curves for nonlinear scalings in the dot-product kernel or random features setting, but they have done so only in the noncritical regime $m \not\propto d^r$ [17, 32, 26]. [8, 10] also derive the closed-form predictions of the learning curves for both the critical and the noncritical scalings, but they have done so via nonrigorous statistical physics methods and a "Gaussian equivalence conjecture" [12, 22, 16, 18, 19, 20]. Rigorously extending these results to include the critical regime $m \propto d^r$ is nontrivial, both from the technical perspective, namely, proving a "Gaussian equivalence conjecture", and also from the phenomenological perspective, as we shall see the critical behavior induces nonmonotonicity and multiple sample-wise descents.

In this work, we obtain precise formulas for the sample-wise learning curves in the kernel ridge regression setting for a family of dot-product kernels for spherical input data in the polynomial scaling regimes $m \propto d^r$ for all $r \in \mathbb{N}^*$. This family of kernels includes the neural network Gaussian Process (NNGP) kernels and Neural Tangent Kernels (NTK) associated with multi-layer fully-connected networks or convolutional networks. Both kernels serve as important starting points towards a deeper understanding of neural networks as they often capture the first order learning dynamics of neural networks in certain scaling limits [23, 25, 4].

1.1 Contributions

Our primary contributions are to establish the following, for data drawn uniformly from the sphere:

1. The empirical spectral density of the Gram matrix induced by degree-r spherical harmonics converges to a Marchenko-Pastur distribution when $(d^r/r!)/m$ converges to a positive constant as $d \to \infty$ (Theorem 1);
2. A precise closed-form formula for the sample-wise learning curves for dot-product kernel regression when $m \propto d^r$ for all $r \in \mathbb{N}^*$ as $d \to \infty$ (Theorem 2);
3. Empirically, the theoretical predictions agree with finite-size simulations surprisingly well even in the strong finite-size correction regime (Fig. 1);
4. An extension of the above results to convolutional kernels (Section 5).

Finally, we note that our results also assume the high-degree coefficients of the label function to be random and isotropic; see Eq. (11). It remains an open question to prove similar results[2] when the label function is deterministic.

2 Notation and Setup

Let $\boldsymbol{\mathcal{X}} = \mathbb{S}_{d-1}$ denote the input space, where $\mathbb{S}_{d-1}$ is the unit sphere in $\mathbb{R}^d$ and $\boldsymbol{\mathcal{X}}$ is equipped with the normalized uniform measure σ. We use Δ_d to represent any quantity (a scalar, vector or a matrix) with $|||\Delta_d||| \to 0$ as $d \to \infty$ (in probability if Δ_d is stochastic), where $||| \cdot |||$ can be the absolute value of a scalar, the norm of a vector or the operator norm of a matrix.

Let $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ be the training inputs where the i-th row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^\top$. We assume $\{\boldsymbol{x}_i\}_{i \in [m]}$ is sampled uniformly, iid from $\boldsymbol{\mathcal{X}}$. The label function $f : \mathbb{S}_{d-1} \to \mathbb{R}$ will be defined in Section 4. Let $K = K^{(d)}$ be a dot-product kernel defined on $\mathbb{S}_{d-1} \times \mathbb{S}_{d-1}$, i.e., $K(\boldsymbol{x}, \boldsymbol{x}') = h(\boldsymbol{x}^\top \boldsymbol{x}')$ for some function $h \in [-1, 1] \to \mathbb{R}$. We assume h has the following decomposition

$$h(t) = \sum_{k=1}^{\infty} \hat{h}_k^2 P_k(t), \quad \text{with} \quad \sum_{k=1}^{\infty} \hat{h}_k^2 < \infty \,, \tag{1}$$

where P_k is the k-th order Legendre polynomials in d dimensions. For simplicity, we assume $\hat{\boldsymbol{h}} = (\hat{h}_k)_{k \geq 1}$ is a sequence that is independent of d and $\hat{h}_k \neq 0$ for all $k \leq k_0$ where k_0 is sufficiently large. As such, we can decompose the kernel function using sperical harmonics,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{\infty} \sigma_k^2 \sum_{l \in [N(d,k)]} Y_{k,l}(\boldsymbol{x}) Y_{k,l}(\boldsymbol{x}') = \sum_{k=1}^{\infty} \sigma_k^2 Y_k(\boldsymbol{x})^\top Y_k(\boldsymbol{x}') \,, \tag{2}$$

where $Y_{k,l}$ is the l-th spherical harmonic of degree k, $N(d,k) = d^k/k! + O(d^{k-1})$ is the total number of degree k spherical harmonics in d dimensions, $\sigma_k^2 = \hat{h}_k^2 / N(d,k)$ is the eigenvalue of Y_{kl}, and $Y_k(\boldsymbol{x})$ is the column vector $[Y_{k,l}(\boldsymbol{x})]_{l \in [N(d,k)]}^\top$. We also denote by $Y_k(\boldsymbol{X})$ the $m \times N(d,k)$ matrix whose i-th row is $Y_k(\boldsymbol{x}_i)^\top$.

3 Structure of the Empirical Kernel and Marchenko-Pastur Distribution

The structure of the empirical kernel matrix $K(\boldsymbol{X}, \boldsymbol{X})$ plays a critical role in characterizing the sample-wise test error for the kernel ridge regressor associated to K. We assume the training set size scales polynomially, i.e. $m \sim d^r$ for some positive integer $r \in \mathbb{N}^*$. Decompose this kernel into low-, critical- and high-frequency modes as follows,

$$K(\boldsymbol{X}, \boldsymbol{X}) = \sum_{k<r} \sigma_k^2 Y_k(\boldsymbol{X}) Y_k(\boldsymbol{X})^\top + \sigma_r^2 Y_r(\boldsymbol{X}) Y_r(\boldsymbol{X})^\top + \sum_{k>r} \sigma_k^2 Y_k(\boldsymbol{X}) Y_k(\boldsymbol{X})^\top \,. \tag{3}$$

The low- and high-frequency parts have simple structures since $N(d,k)/m$ either diverges to infinity or converges to zero with rate as least $d^{\pm 1}$, yielding concentration that results in significant simplification. To be precise, for high-frequency modes $k > r$, $Y_k(\boldsymbol{X})$ is a "fat" matrix and

[2] See Sec. 6 for empirical evidences in favor of these results.

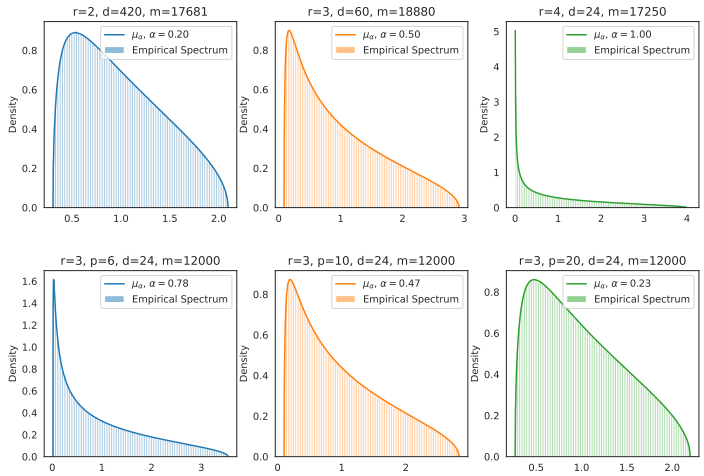


Figure 2: **Marchenko-Pastur Distribution of Spherical Harmonics.** Top: the empirical distribution of product kernels $Y_r(\boldsymbol{X})Y_r(\boldsymbol{X})^\top/N(d,r)$ vs theory prediction from μ_α for various degrees r, input dimensions d and number of samples m as indicated in the titles. Bottom: the empirical distribution of the CNN kernel $Y_r(\boldsymbol{X})Y_r(\boldsymbol{X})^\top/pN(d,r)$ vs theoretical prediction. We fix r, d and m but varying the number of patches $p \in \{6, 10, 20\}$.

$Y_k(\boldsymbol{X})Y_k(\boldsymbol{X})^\top/N(d,k) = \boldsymbol{I}_m + \Delta_d$ where Δ_d vanishes as $d \to \infty$ [32]. Thus, the high-frequency parts behave like a regularizer in the following sense,

$$\sum_{k>r} \sigma_k^2 Y_k(\boldsymbol{X})Y_k(\boldsymbol{X})^\top = \sum_{k>r} \sigma_k^2 N(d,k)\boldsymbol{I}_m + \Delta_d = \left(\sum_{k>r} \hat{h}_k^2\right)\boldsymbol{I}_m + \Delta_d\,. \tag{4}$$

On the other hand, when $k < r$, $Y_k(\boldsymbol{X})$ is a $m \times N(d,k)$ "tall" matrix with $N(d,k)/m = O(d^k/m) = O(d^{-(r-k)}) \to 0$. Similarly, Mei et al. [32] show that $Y_k(\boldsymbol{X})^\top Y_k(\boldsymbol{X})/m = \boldsymbol{I}_{N(d,k)} + \Delta_d$, implying that when restricted to the subspace spanned by low-frequency functions $\{Y_{kl}\}_{k<r}$, the regressor associated to the empirical kernel $K(\boldsymbol{X}, \boldsymbol{X})$ acts like a pure multiplicative scaling.

It remains to understand the critical-frequency mode $Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})$. It turns out that if $N(d,k)/m \to \alpha \in (0,\infty)$, then the empirical spectral measure of the random matrix $Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})/m$ converges to the Marchenko-Pastur distribution μ_α, whose density is given by

$$\mu_\alpha(t) = \left(1 - \frac{1}{\alpha}\right)^+ \delta_0(t) + \frac{\sqrt{(\alpha_+ - t)(t - \alpha_-)}}{2\pi\alpha t}\mathbf{1}_{[\alpha_-,\alpha_+]}(t), \text{ where } \alpha_\pm = (1 \pm \sqrt{\alpha})^2. \tag{5}$$

where $\delta_0(t) = 0$ if $t \neq 0$ else 1. See Fig. 2 for visualizations of μ_α. The $r = 1$ case is obvious as $Y_1(\boldsymbol{X}) = c_d\boldsymbol{X}$ for some normalizing constant c_d and it is clear $\frac{1}{m}Y_1(\boldsymbol{X})^\top Y_1(\boldsymbol{X}) = \frac{c_d^2}{m}\boldsymbol{X}^\top\boldsymbol{X}$ converges to the Marchenko-Pastur distribution μ_α if $d/m \to \alpha \in (0,\infty)$ as $d \to \infty$ [37]. Our first result show that this result continues to hold for all degrees.

Theorem 1. *For fixed $r \in \mathbb{N}$ and $\alpha \in (0,\infty)$, if $N(d,r)/m \to \alpha \in (0,\infty)$ as $d \to \infty$, then the empirical spectral distribution of $\frac{1}{m}Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})$ converges in distribution to the Marchenko-Pastur distribution $\mu_{\mathrm{MP}(\alpha)}$.*

In the top panel of Fig. 2, we generate the empirical spectra[3] of $\frac{1}{m}Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})$ for various values of r, d, and α. The Marchenko-Pastur distribution μ_α perfectly captures the empirical measures of the random matrices $\frac{1}{m}Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})$ for all r considered. We sketch the main steps of the proof of the theorem below; see Sec. B for the whole proof.

[3]In the plot, we generate the spectra of the kernel matrix $Y_r(\boldsymbol{X})Y_r(\boldsymbol{X})^\top$ instead of the covariate matrix $Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})$. Although both of them have the same set of non-zero eigenvalues, the former can be easily implemented via Legendre polynomials $P_r(\boldsymbol{x}^\top\boldsymbol{x}')$.

Sketch of Proof. From Bai and Zhou [5, Theorem 1.1], it suffices to prove concentration of the following quadratic forms: for every sequence of $N(d,k) \times N(d,k)$ matrices $\{\boldsymbol{A}_d\}$ with operator norm $\|\boldsymbol{A}_d\|_{\mathrm{op}} \le 1$, the variance

$$N(d,r)^{-2}\mathbb{V}(Y_r(\boldsymbol{x})^\top \boldsymbol{A}_d Y_r(\boldsymbol{x}) - \mathrm{Tr}(\boldsymbol{A}_d)) \to 0 \quad \text{as} \quad d \to \infty\,. \tag{6}$$

For the purpose of illustration, we assume $\boldsymbol{A} \equiv \boldsymbol{A}_d$ is a diagonal matrix. Then we only need to show

$$N(d,k)^{-2} \sum_{l,l' \in [N(d,k)]} A_{ll} A_{l'l'} (\mathbb{E}_{\boldsymbol{x}} Y_{k,l}^2(\boldsymbol{x}) Y_{k,l'}^2(\boldsymbol{x}) - 1) \to 0. \tag{7}$$

By hypercontractivity of spherical harmonics [6],

$$\mathbb{E}_{\boldsymbol{x}} Y_{k,l}^2(\boldsymbol{x}) Y_{k,l'}^2(\boldsymbol{x}) \le \left(\mathbb{E}_{\boldsymbol{x}} Y_{k,l}(\boldsymbol{x})^4 \mathbb{E}_{\boldsymbol{x}} Y_{k,l'}^4(\boldsymbol{x})\right)^{1/2} \le C_k \mathbb{E}_{\boldsymbol{x}} Y_{k,l}(\boldsymbol{x})^2 \mathbb{E}_{\boldsymbol{x}} Y_{k,l'}(\boldsymbol{x})^2 = C_k, \tag{8}$$

where C_k is some absolute constant. Since $|A_{ll}| \le \|\boldsymbol{A}\|_{\mathrm{op}} \le 1$, we can drop any $o(N(d,l)^2)$ pairs of (l,l') in Eq. (7). We show that for the remaining pairs (l,l'), the eigenfunctions are asymptotically uncorrelated in the sense

$$\mathbb{E}_{\boldsymbol{x}} Y_{k,l}^2(\boldsymbol{x}) Y_{k,l'}^2(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}} Y_{k,l}^2(\boldsymbol{x}) \mathbb{E}_{\boldsymbol{x}} Y_{k,l'}^2(\boldsymbol{x}) + O(d^{-1}) = 1 + O(d^{-1}) \tag{9}$$

which implies Eq. (7). □

4 Generalization Error of Dot-Product Kernel Regression

In this section, we establish the *average* generalization error for the kernel regression in the asymptotic regime $N(d,r)/m \to \alpha$, for some $\alpha \in (0,\infty)$ and $r \ge 1$ fixed. We assume the label function $f \in L^2(\mathbb{S}_{d-1})$ is given by

$$f(\boldsymbol{x}) = \sum_{k \ge 1} \sum_{l \in [N(d,k)]} \hat{f}_{kl} Y_{kl}(\boldsymbol{x}) = \sum_{k \ge 1} \hat{\boldsymbol{f}}_k^\top Y_k(\boldsymbol{x})\,, \tag{10}$$

where $\hat{f}_{kl}$ are the "Fourier" coefficients and $\hat{\boldsymbol{f}}_k = [f_{kl}]_{l \in N(d,k)}^\top$. We need to make a technical assumption that for $k', k \ge r$

$$\mathbb{E}\hat{\boldsymbol{f}}_k = \boldsymbol{0}, \qquad \mathbb{E}\hat{\boldsymbol{f}}_k \hat{\boldsymbol{f}}_k^\top = \frac{\hat{F}_k^2}{N(d,k)} \boldsymbol{I}_{N(d,k)} \quad \text{and} \quad \mathbb{E}\hat{\boldsymbol{f}}_k \hat{\boldsymbol{f}}_{k'}^\top = \boldsymbol{0}_{N(d,k) \times N(d,k')} \tag{11}$$

i.e. $\hat{\boldsymbol{f}}_k$ is centered with isotropic covariance and $\{\hat{\boldsymbol{f}}_k\}_{k \ge r}$ are mutually uncorrelated. Note that we allow $\hat{f}_{kl}$ to be deterministic for $k < r$. We let $\boldsymbol{F} = (\hat{F}_k)_{\ge 1}$ be a fixed sequence with $\sum_{k \ge 1} \hat{F}_k^2 < \infty$, where $\hat{F}_k^2 = \sum_{l \in [N(d,k)]} \hat{f}_{kl}^2$ for $k < r$. For convenience, set $\hat{F}_{>j}^2 = \sum_{k > j} \hat{F}_k^2$ (similarly for $\hat{F}_{\ge j}^2$, $\hat{F}_{\le j}^2$, etc.) and use $\boldsymbol{f}$ to denote the random vector $\{\hat{f}_{kl}\}_{kl}$. Given training inputs $\boldsymbol{X}$ and observed labels $\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_m)$ is the noise, the prediction using kernel function K is given by

$$y(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{X})(K(\boldsymbol{X}, \boldsymbol{X}) + \lambda \boldsymbol{I}_m)^{-1}(f(\boldsymbol{X}) + \boldsymbol{\epsilon})\,. \tag{12}$$

Here $\lambda \ge 0$ is the regularization. As such, the *mean* test error over the random labels is given by

$$\mathrm{Err}(\boldsymbol{X}; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) = \mathbb{E}_{\boldsymbol{f}} \mathrm{Err}(\boldsymbol{X}; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) \quad \text{where} \quad \mathrm{Err}(\boldsymbol{X}; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{\epsilon}} |y(\boldsymbol{x}) - f(\boldsymbol{x})|^2\,. \tag{13}$$

To state our results, we need to introduce two functions χ_B and χ_V which are related to the bias and variance in the generalization error,

$$\chi_B(\alpha, \xi) = \int (1 + \xi t)^{-2} \mu_\alpha(t) dt \quad \text{and} \quad \chi_V(\alpha, \xi) = \alpha \xi^2 \int t(1 + \xi t)^{-2} \mu_\alpha(t) dt\,. \tag{14}$$

Both χ_B and χ_V have closed-form representations; see Sec C.5. Define the effective regularization associated to the r-th order scaling to be

$$\xi_r(\hat{\boldsymbol{h}}, \lambda, \alpha) = \frac{\hat{h}_r^2}{\alpha(\lambda + \hat{h}_{>r}^2)} \tag{15}$$

Finally, we define the bias and variance associated to the r-th order scaling to be

$$\mathrm{B}_r(\alpha) = \mathrm{B}_r(\alpha; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) = \chi_B(\alpha, \xi_r(\hat{\boldsymbol{h}}, \lambda, \alpha)) \hat{F}_r^2 + \hat{F}_{>r}^2 \tag{16}$$

$$\mathrm{V}_r(\alpha) = \mathrm{V}_r(\alpha; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) = \chi_V\left(\alpha, \xi_r(\hat{\boldsymbol{h}}, \lambda, \alpha)\right)\left(\hat{F}_{>r}^2 + \sigma_\epsilon^2\right) \tag{17}$$

The following is our main result, which characterizes the test error in the asymptotic regime $m \propto d^r$.

Theorem 2. *Let $\alpha \in (0, \infty)$ and $r \geq 1$ be fixed. Assume $N(d, r)/m \to \alpha$ as $d \to \infty$. Then the average test error is given by*

$$\mathrm{Err}(\boldsymbol{X}; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) = \mathrm{B}_r(\alpha; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) + \mathrm{V}_r(\alpha; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) + \Delta_d, \tag{18}$$

where $\Delta_d \to 0$ in probability.

4.1 Interpretations

We provide some high-level interpretations of the bias term $\mathrm{B_r}$ and variance term $\mathrm{V_r}$.

The Bias. From Eq. (16), the regressor learns all low-frequency modes ($k < r$) but none of the high-frequency modes ($k > r$) as the bias B_r contains no low-frequency modes (i.e. $k < r$) but all high-frequency modes $\hat{F}^2_{>r}$. Importantly, the regressor is progressively learning the critical-frequency mode Y_r as the training size $m = \frac{1}{\alpha} N(d, r)$ increases, i.e. from $\alpha = \infty$ to $\alpha = 0^+$ since $\chi_B(\alpha, \xi_r(\hat{\boldsymbol{h}}, \lambda, \alpha)) \to 1$ if $\alpha \to \infty$ and $\chi_B(\alpha, \xi_r(\hat{\boldsymbol{h}}, \lambda, \alpha)) \to 0$ if $\alpha \to 0^+$. See Fig.3 for the illustration.

The Variance. From Eq. (17), the variance term χ_{V} treats all high-frequency modes $\hat{F}^2_{>r}$ the same as the noise term $\boldsymbol{\epsilon}$. Moreover, $\chi_{\mathrm{V}} \to 0$ as $\alpha \to 0$ or ∞ and is peaked at $\alpha = 1$. The height of the peak depends on the effective regularization ξ_r and it diverges to infinity with rate $\xi_r^{\frac{1}{2}}$ as $\xi_r \to \infty$. Indeed, when $\alpha = 1$ and $\xi^{-1}/2 \leq t \leq \xi^{-1}$, we have $t(1+\xi t)^{-2}\mu_\alpha(t) \propto \xi^{-1/2}$ which implies $\chi_V(1, \xi) \geq \xi^2 \int t(1+\xi t)^{-2}\mu_\alpha(t)\mathbf{1}_{\xi^{-1}/2 \leq t \leq \xi^{-1}} dt \propto \xi^{1/2}$.

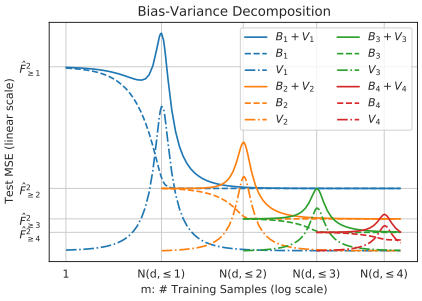


Figure 3: **Multi-scale Bias-Variance Decomposition.** Theoretical predictions of the bias and variance from Eq.16 and Eq.17. For each r, the variance is non-monotonic and has a peak at $N(d, \leq r) = \sum_{k \leq r} N(d, k)$.

Finally, Eq. (18) not only gives precise generalization formula (up to a vanishing term Δ_d) when $m \approx N(d, r) \sim d^r$ but also when $d^{r-1+\delta} \lesssim m \lesssim d^{r-\delta}$ (i.e. when "$\alpha = \infty$") and when $d^{r+\delta} \lesssim m \lesssim m^{r+1-\delta}$ (i.e. when "$\alpha = 0^+$") for any $\delta \in (0, 1/2)$. Indeed, in the non-critical scaling regime $d^{r-1+\delta} \lesssim m \lesssim d^{r-\delta}$, $\alpha = N(d, k)/m \to \infty$ as $d \to \infty$ and

$$\mathrm{B_r}(\alpha = \infty) + \mathrm{V_r}(\alpha = \infty) = \hat{F}^2_{\geq r} + 0 = \hat{F}^2_{\geq r}\,. \tag{19}$$

As such, the regressor learns all low-frequency modes but none of the critical- and high-frequency modes ($k \geq r$), which is consistent with the result in [17, 32]. A similar argument also shows $\mathrm{B_r}(\alpha = 0^+) + \mathrm{V_r}(\alpha = 0^+) = \hat{F}^2_{>r}$, namely, the regressor also learns the r-frequency mode. This observation implies that we can glue together Eq. (18) for $r \geq 1$ and remove all duplicate terms to generate a sample-wise learning curve (LC):

$$\mathrm{LC}(m; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) = \sum_{r \geq 1} \left(\mathrm{B}_r \left(\frac{N(d, r)}{m}; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}} \right) - (r-1)\hat{F}^2_r \right) + \mathrm{V}_r \left(\frac{N(d, \leq r)}{m}; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}} \right) \tag{20}$$

where $N(d, \leq r) = \sum_{k=1}^{r} N(d, k)$. The "$-(r-1)\hat{F}^2_r$" term in the above equation is due to the fact that $\hat{F}^2_r$ is over-counted $(r-1)$ many times (one in each B_k for $k = 1, ..., (r-1)$.) It is worth mentioning that using $\alpha = N(d, \leq r)/m$ rather than $\alpha = N(d, r)/m$ in the variance V_r captures the finite-size correction more accurately. See Eq. (206).

Corollary 1. *If, for $1 \leq r \in \mathbb{N}$, (1) $N(d, r)/m \to \alpha$ for some $\alpha \in (0, \infty)$, or (2) $d^{r-1+\delta} \lesssim m \lesssim d^{r-\delta}$ for some $\delta \in (0, 1/2)$, then*

$$\mathrm{Err}(\boldsymbol{X}; \boldsymbol{f}, \hat{\boldsymbol{h}}, \lambda) = \mathrm{LC}(m; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) + \Delta_d \tag{21}$$

where $\Delta_d \to 0$ in probability as $d \to \infty$.

Recall that for each r, the variance term V_r could diverge to infinity as $\xi_r \to \infty$ at $\alpha = 1$. Thus we might expect a peak in the learning curve for each r, yielding the multiple-descent phenomenon, as shown in Fig.1. However, such phenomena can disappear by making the heights of the peaks small via choosing ξ_r small. We will discuss this point in the experimental section.

4.2 Proof Sketch

The proof of this theorem is quite involved; see Sec.C. For simplicity, we assume the observed labels are noiseless, i.e. $\sigma_\epsilon^2 = 0$. An ingredient is to understand the structure of the operator

$$\boldsymbol{T}_K f(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{X})(K(\boldsymbol{X}, \boldsymbol{X}) + \lambda \boldsymbol{I}_m)^{-1} f(\boldsymbol{X}) \,. \tag{22}$$

The high-level strategy is as follows. We decompose the function into low-, critical- and high-frequency parts $f = f_{<r} + f_r + f_{>r}$. As such, the test error is roughly

$$\mathrm{Err}(\boldsymbol{X}) \approx \mathrm{Err}_{<r}(\boldsymbol{X}) + \mathrm{Err}_r(\boldsymbol{X}) + \mathrm{Err}_{>r}(\boldsymbol{X}) \text{ where } \mathrm{Err}_r(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{f}} \mathbb{E}_{\boldsymbol{x}} |\boldsymbol{T}_K f_r(\boldsymbol{x}) - f_r(\boldsymbol{x})|^2 \,,$$

and similarly for $\mathrm{Err}_{<r}(\boldsymbol{X})$ and $\mathrm{Err}_{>r}(\boldsymbol{X})$. The next step is to estimate each part separately.

Low-frequencies. Using the fact that the low-frequency parts of the kernel function K is almost an isometric operator on the column space of $Y_{<r}(\boldsymbol{X})$, one can show that $\mathbb{E}_{\boldsymbol{x}} |T_K(f_{<r})(\boldsymbol{x}) - f_{<r}(\boldsymbol{x})|^2 = \Delta_d \to 0$ in probability, *pointwisely*.

Critical-frequency. Up to a vanishing term, one can remove all non-critical frequencies in the kernel function K in T_K in the sense of making the following substitutions

$$K(\boldsymbol{x}, \boldsymbol{X}) \to \sigma_r^2 Y_r(\boldsymbol{x})^\top Y_r(\boldsymbol{X})^\top \quad \text{and} \quad K_r(\boldsymbol{X}, \boldsymbol{X}') \to \sigma_r^2 Y_r(\boldsymbol{X}) Y_r(\boldsymbol{X})^\top + \hat{h}_{>r}^2 \boldsymbol{I}_m \,. \tag{23}$$

Thus, with $\gamma = (\lambda + \hat{h}_{>r}^2)$, $T_K f_r(\boldsymbol{x}) - f_r(\boldsymbol{x}) = Y_r(\boldsymbol{x})^\top \boldsymbol{M}_r(\boldsymbol{X}) \boldsymbol{f}_r + \Delta_d$, where

$$\boldsymbol{M}_r(\boldsymbol{X}) = \left(\boldsymbol{I}_{N(d,r)} - \sigma_r^2 Y_r(\boldsymbol{X})^\top (\sigma_r^2 Y_r(\boldsymbol{X}) Y_r(\boldsymbol{X})^\top + \gamma \boldsymbol{I}_m)^{-1} Y_r(\boldsymbol{X}) \right) \,. \tag{24}$$

Taking expectation with respect to $\boldsymbol{x}$ (using orthogonality of $Y_r(\boldsymbol{x})$) and then with respect to $\boldsymbol{f}_r$,

$$\mathbb{E}_{\boldsymbol{f}_r} \mathbb{E}_{\boldsymbol{x}} |Y_r(\boldsymbol{x})^\top \boldsymbol{M}_r(\boldsymbol{X}) \boldsymbol{f}_r|^2 = \mathbb{E}_{\boldsymbol{f}_r} |\boldsymbol{M}_r(\boldsymbol{X}) \boldsymbol{f}_r|^2 = \hat{F}_r^2 \mathrm{Tr}(\boldsymbol{M}_r^2)/N(d,r) \,. \tag{25}$$

Applying the Sherman–Morrison–Woodbury formula and then Theorem 1,

$$\hat{F}_r^2 \mathrm{Tr} \left(Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})/m + m\sigma_r^2/\gamma \boldsymbol{I}_{N(d,r)} \right)^{-2} / N(d,r) \to \hat{F}_r^2 \int \frac{\mu_\alpha(t)}{(t+\xi_r)^2} dt \,.$$

High-frequencies. The cross term $\mathbb{E}_{\boldsymbol{f}} \mathbb{E}_{\boldsymbol{x}} \boldsymbol{T}_K f_{>r}(\boldsymbol{x}) f_{>r}(\boldsymbol{x}) = \Delta_d$ and thus

$$\mathbb{E}_{\boldsymbol{f},\boldsymbol{x}} |\boldsymbol{T}_K f_{>r}(\boldsymbol{x}) - f_{>r}(\boldsymbol{x})|^2 = \mathbb{E}_{\boldsymbol{f},\boldsymbol{x}} |\boldsymbol{T}_K f_{>r}(\boldsymbol{x})|^2 + \mathbb{E}_{\boldsymbol{f},\boldsymbol{x}} |f_{>r}(\boldsymbol{x})|^2 + \Delta_d \,. \tag{26}$$

The second term is equal to $\hat{F}_>^2$. The calculation of the first term is similar to that of the critical frequency above (namely, we remove all high-/low-frequency components in K.)

5 Convolutional Kernels

5.1 One hidden layer

Our analysis can be extended to analyzing NNGP kernel and NT kernel for one-layer convolution [35, 36, 43]. In this case, we assume the input space is $\boldsymbol{\mathcal{X}} = \mathbb{S}_{d_0-1}^p$, where d_0 is the dimension of a patch, p is the number of patches, and $d = pd_0$ is the total dimensions of the inputs. The measure associated to $\boldsymbol{\mathcal{X}}$ is the product of the uniform measure on $\mathbb{S}_{d_0-1}$. We assume that both the filter size and stride of the convolution are equal to d_0. As such, after the first convolutional layer, the input is reduced to a vector of dimension p. We then apply a non-linearity and a dense layer to map this p-dimensional vector to a scalar. The NNGP and NT kernel have the following general form. Let $\boldsymbol{x} = (\boldsymbol{x}_i)_{i\in[p]} \in \boldsymbol{\mathcal{X}}$, where $\boldsymbol{x}_i \in \mathbb{S}_{d_0-1}$ is the i-th patch

$$K(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{p} \sum_{i\in[p]} h(\boldsymbol{x}_i^\top \boldsymbol{x}_i') = \frac{1}{p} \sum_{i\in[p]} \sum_{k\geq 1} \hat{h}_k^2 P_k(\boldsymbol{x}_i^\top \boldsymbol{x}_i') = \sum_{k\geq 1} \frac{\sigma_k^2}{p} \sum_{i\in[p]} \sum_{l\in[N(d_0,k)]} Y_{kl}(\boldsymbol{x}_i) Y_{kl}(\boldsymbol{x}_i') \,.$$

Denote $Y_{kl}^{(i)}(\boldsymbol{x}) = Y_{kl}(\boldsymbol{x}_i)$ and $Y_k(\boldsymbol{x}) = [Y_{kl}^{(i)}(\boldsymbol{x})^\top]_{l\in[N(d_0,k)],i\in[p]}^\top$. Then $Y_k(\boldsymbol{x})$ is the degree k spherical harmonics associated to this kernel, which span a space of dimension $pN(d_0,k)$.

Theorem 3. *Let $r \in \mathbb{N}^*$ and $\alpha \in (0, \infty)$ be fixed. If $pN(d_0, r)/m \to \alpha \in (0, \infty)$ as $d_0 \to \infty$ and the rows of $\boldsymbol{X}$ are sampled uniformly, iid from $\mathbb{S}^p_{d_0-1}$, then the empirical spectral distribution of $\frac{1}{m} Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})$ tends to the Marchenko-Pastur distribution μ_α as $d_0 \to \infty$.*

The assumptions on the label function are similar to that of dot-product kernel, e.g.[4]

$$f(\boldsymbol{x}) = \sum_{k \geq 1} \boldsymbol{f}_k^\top Y_k(\boldsymbol{x}), \quad \text{with} \quad \boldsymbol{f}_k \sim \mathcal{N}\left(0, \frac{\hat{F}_k^2}{pN(d_0, k)} \boldsymbol{I}_{pN(d_0,k)}\right) \text{ if } k \geq r\,, \tag{27}$$

otherwise $\boldsymbol{f}_k$ is deterministic with $\|\boldsymbol{f}_k\|_2^2 = \hat{F}_k^2$.

Theorem 4. *Let $\alpha \in (0, \infty)$ and $r \geq 1$ be fixed. Assume $pN(d_0, r)/m \to \alpha$ as $d_0 \to \infty$. Then the average test error is given by*

$$\mathrm{Err}(\boldsymbol{X}; \lambda, \boldsymbol{F}, \hat{\boldsymbol{h}}) = \mathrm{B}_r(\alpha; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) + \mathrm{V}_r(\alpha; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) + \Delta_{d_0}, \tag{28}$$

where $\Delta_{d_0} \to 0$ in probability as $d_0 \to \infty$.

Corollary 2. *If, for $1 \leq r \in \mathbb{N}$, (1) $pN(d_0, r)/m \to \alpha$ for some $\alpha \in (0, \infty)$, or (2) $pd_0^{r-1+\delta} \lesssim m \lesssim pd_0^{r-\delta}$ for some $\delta \in (0, 1/2)$, then*

$$\mathrm{Err}(\boldsymbol{X}; \boldsymbol{f}, \hat{\boldsymbol{h}}, \lambda) = \mathrm{LC}(m; \lambda, \boldsymbol{f}, \hat{\boldsymbol{h}}) + \Delta_{d_0} \tag{29}$$

where $\Delta_{d_0} \to 0$ in probability as $d_0 \to \infty$.

5.2 Deep Convolutional Kernels

The eigenstructure of general CNN kernels are much more complicated as they depend on both the frequencies (i.e. the order of the polynomials) and the topologies of the networks [42]. To rigorously describe the eigenstructure, a heavy dose of notation must be introduced, which is beyond the scope of the paper. Nevertheless, the approach developed here is readily extended to cover general CNN kernels. We briefly describe the main ideas.

Following [42], we assume the input space is still $\boldsymbol{\mathcal{X}} = \mathbb{S}^p_{d_0-1}$, where p is the number of patches. For simplicity, we assume the network has L convolutional layers and in each layer, the filter size and the stride are all equal to d_0. Thus the spatial dimension of the input is reduced to 1 after L convolutional layers. We then add a non-linearality and a dense layer to generate the logits. The kernel has the following form

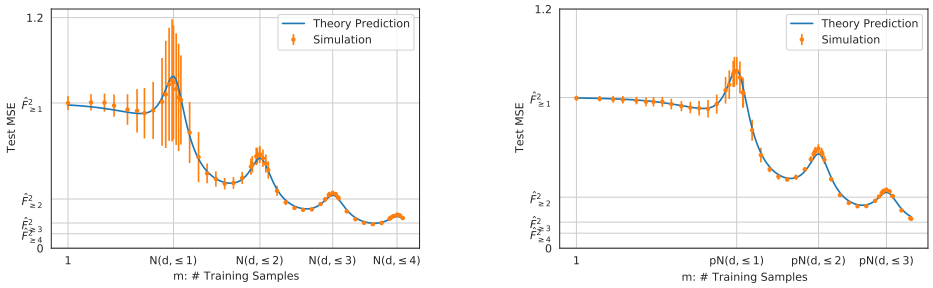
$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\boldsymbol{k} \in \mathbb{N}^p} \sum_{\boldsymbol{l} \in \prod_{i \in [p]} [N(d_0, k_i)]} \sigma^2_{\boldsymbol{k},\boldsymbol{l}} Y_{\boldsymbol{k},\boldsymbol{l}}(\boldsymbol{x}) Y_{\boldsymbol{k},\boldsymbol{l}}(\boldsymbol{x}'), \text{ where } \quad Y_{\boldsymbol{k},\boldsymbol{l}}(\boldsymbol{x}) = \prod_{i \in [p]} Y_{k_i, l_i}(\boldsymbol{x}_i)\,. \tag{30}$$

Unlike dot-product kernels in which $\boldsymbol{k}$ is a scalar and the eigenvalues depend only on $|\boldsymbol{k}|$ (i.e. the frequencies), $\sigma^2_{\boldsymbol{k},\boldsymbol{l}}$ depends on both $|\boldsymbol{k}|$ and the spatial structure of the vector $\boldsymbol{k}$ in a rather complicated manner. Nevertheless, as $d_0 \to \infty$, $\sigma^2_{\boldsymbol{k},\boldsymbol{l}} \sim d_0^{-j_{\boldsymbol{k}}} = d^{-j_{\boldsymbol{k}}/L}$ for some $L \leq j_{\boldsymbol{k}} \in \mathbb{N}$. We can then categorize the eigenvectors according to the decay order of $\sigma^2_{\boldsymbol{k},\boldsymbol{l}}$. Unlike the case of dot-product kernels or the one-hidden layer CNN kernels, in which eigenvectors with same-order eigenvalues are in the same eigenspace (i.e. the eigenvalues are the same), multiple-layer CNN kernels can have *multiple* eigenspaces with the same-order eigenvalues. Although this results in extra challenges (see below), our overall approach carries over. Consider the critical scaling regime $m \sim d^r$, for $r = j/L$ for some $L \leq j \in \mathbb{N}$. Likewise, we can decompose the kernel into low-, critical- and high-frequency parts according to $j_{\boldsymbol{k}} < r$, $j_{\boldsymbol{k}} = r$ and $j_{\boldsymbol{k}} > r$, resp. Following similar assumptions on the labels and eigenvalues, the bias and the variance can be essentially reduced to computing

$$\chi_B = \frac{1}{N_r} \operatorname{Tr} \boldsymbol{R}_r^2 (\boldsymbol{R}_r + Y_r(\boldsymbol{X})^\top Y_r(\boldsymbol{X})/m)^{-2} \tag{31}$$

$$\chi_V = \frac{N_{\leq r}}{m} \frac{1}{N_{\leq r}} \operatorname{Tr}(\boldsymbol{R}_{\leq r} + Y_{\leq r}(\boldsymbol{X})^\top Y_{\leq r}(\boldsymbol{X})/m)^{-2} Y_{\leq r}(\boldsymbol{X})^\top Y_{\leq r}(\boldsymbol{X})/m \tag{32}$$

[4] The Gaussian assumption is unessential. We use it here for convenience.



(a) Dot Product Kernel

(b) One-layer CNN Kernel

Figure 4: **Simulation vs Prediction.** We generate the learning curves obtain from **kernel regression** by densely varying m from 1 to 24000. For each m, we average the MSE over 20 runs. The **closed-form prediction** from Eq. (18) captures the simulations surprising well even for small d. Left: dot product kernel with $d = 24$. Right: one-hidden layer CNN kernel with $d_0 = 20$ and $p = 6$. The spectral gap is $\text{Gap} = 32$ in both plots.

where $\boldsymbol{R}_r$ and $\boldsymbol{R}_{\leq r}$ are diagonal matrices whose entries are determined by the eigenvalues of the critical-frequency modes. In the dot-product kernels or one-hidden layer CNN kernels setting, $\boldsymbol{R}_r/\boldsymbol{R}_{\leq r}$ is a scaled identity matrix and simple, closed-form expressions for the above traces straightforwardly follow from the Marcenko-Pastur distribution. However, for a general diagonal matrix $\boldsymbol{R}_r$ with bounded limiting spectra, $\boldsymbol{R}_r$ does not commute with $Y_r^\top(\boldsymbol{X})Y_r(\boldsymbol{X})$, and a more detailed random matrix analysis is needed. See the supplementary material for more details.

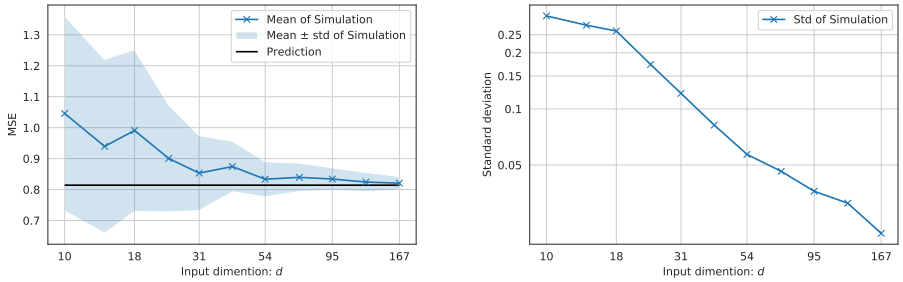
6 Experiments

We provide experiments to show that our learning curves (Eq. (21)) accurately capture empirical sample-wise learning curve even when the ambient dimensions remains small. Even though our theoretical results require averaging the test error over random labels (aka, mean test error), our experimental results suggest this is unnecessary, i.e. the learning curve Eq. (21) can capture the test error accurately for any given draw of label function.

Experimental setup. We generate a polynomial kernel function $h(t) = \sum_{k=1}^{7} \hat{h}_k^2 P_k(t)$, where P_k is the degree-k Legendre polynomial in d dimensions. The kernel function can be efficiently computed via $K(\boldsymbol{x}, \boldsymbol{x}') = h(\boldsymbol{x}^\top \boldsymbol{x}')$. We choose the label function to be $f(\boldsymbol{x}) = \sum_{k=1}^{7} \hat{F}_k y_k(\boldsymbol{x})$, where $y_k(\boldsymbol{x}) = \sum_{j\in[d]} w_{k,j} \prod_{i=j}^{j+k-1} x_i$, and the coefficients $w_{k,j}$ are randomly sampled from a Gaussian and then normalized so that $\mathbb{E}_{\boldsymbol{x}}|y_k(\boldsymbol{x})|^2 = 1$ for each k. Therefore $\mathbb{E}_{\boldsymbol{x}}|f(\boldsymbol{x})|^2 = \sum_{k=1}^{7} \hat{F}_k^2$. For simplicity, we also set $\sigma_\epsilon^2 = 0$ (i.e. noiseless) and $\lambda = 0$ (i.e. ridgeless). Note that when $m \lesssim d^r$, the regressor still contains "effective noise" $\hat{F}_{>r}^2$ from un-learnable high-frequency modes and "effective regularization" $\hat{h}_{>r}^2$. Finally, in our experiments, we choose $\hat{F}_k^2 = k^{-2}$ and $\hat{h}_k^2 = \text{Gap}^{-(k-1)}$, where we will vary the value of the spectral gap: $\text{Gap} = \hat{h}_k^2/\hat{h}_{k+1}^2$. Under this setup, the predicted learning curve $\text{LC}(m) = \text{LC}(m; \text{Gap})$ depends only on the spectral gap of the kernel.

To simulate higher-order scaling ($r \geq 3$), the dimension d has to be very small as we need to invert a sequence of matrices of size ranging from $m = 1$ to $m \propto d^r/r!$. Due to the constraints in compute and memory, the largest m we can have is typically $m_{\max} \approx 25,000$ for one single GPU and d in our experiments is typically around $d = 24$. As such we are in a regime with strong finite-size corrections. Finally, all experiments are run in a single A-100 using Google Cloud Colab Notebooks.

Learning Curves Accurately Capture Simulations. In Fig. 4, we generate the empirical sample-wise learning curve by applying kernel regression Eq. (13) with training set $\boldsymbol{X}$. We vary the training set size m densely in $[1, m_{\max}]$ and for each m we sample 20 independent $\boldsymbol{X}$ to get the **errorbar plot** for the test error. The closed-form **learning curve** is obtained from Eq. (21) and the calculation is done in Sec.C.5. Even in the low-dimensional regime with $d = 24$ for dot-product kernel ($d_0 = 20$ and $p = 6$ for one-hidden layer CNN kernel), the predicted learning curve captures the empirical



(a) Mean of Simulations.

(b) Standard Deviation of Simulations.

Figure 5: **Simulations approach predictions as $d \to \infty$.** The mean and the standard deviation are computed over 32 runs. The predictions and simulations are obtained via the second peak $r = 2$, namely, $m = N(d, \leq 2)$.

learning curve surprisingly well, which has a highly non-trivial multiple-descent behavior. It is worth mentioning that, from the simulation, the deviation of the test error from its mean is relatively large when m is small but vanishes quickly as m becomes larger. This suggests Theorem 2 and Corollary 1 should hold in a pointwise fashion, i.e. without averaging the test error over random labels.

Finite-size correction vanishes as $d \to \infty$. Our theoretical results assume that the input dimension d is sufficiently large and these results are exact when $d = \infty$. To visualize the finite-size correction, we plot the dependence of the correction (between simulations and predictions) on the the input dimension d. Fig. 5 (a) shows that the means of simulations are converging to the theoretical prediction. Fig. 5 (b) shows that the standard deviations are converging zero.

Small Spectral Gap Eliminates Multiple-descent. In Fig. 1, we plot both the predicted learning curves and simulations when Gap ranging in $[2, 8, 32, 128]$. For $1 \leq r \leq 6$, we have $\xi_r = \mathrm{Gap}^{-(r-1)} / \sum_{k=r}^{7} \mathrm{Gap}^{-(k-1)}$ and $\xi_r \approx \mathrm{Gap}$ when Gap is large. Recall that the variance term V_r peaks at $\alpha = 1$ and the peak scales like $\xi_r^{1/2} \approx \mathrm{Gap}^{1/2}$. When Gap is large, e.g. $\mathrm{Gap} = 32, 128$, the variance is also large near $\alpha = 1$, the multiple-descent phenomena are more prominent. On the other hand, when Gap is small, e.g. $\mathrm{Gap} = 8, 2$, such phenomena disappear and learning curves become monotonic.

7 Conclusion

In this work, we establish precise asymptotic formulas for the sample-wise learning curves in the kernel ridge regression setting for a family of dot-product kernels in the polynomial scaling regimes $m \propto d^r$ for all $r \in \mathbb{N}^*$. We demonstrate that these formulas can capture empirical learning curves surprisingly well even in the regime where strong finite-size corrections would be expected. We rigorously prove that the learning curves can be non-monotonic near $m \propto d^r/r!$ for each $r \in \mathbb{N}^*$. There are a couple limitations of our approach which could be improved in future work. The first one is the strong assumption on the distribution of the input data, namely, the uniform distribution on the spherical type of data. In addition, the learning curves are obtained only in the kernel regression setting and extending the results to the random feature setting (see, e.g., [29]) and the feature learning setting [44] would be meaningful future directions.

Acknowledgement

We thank Ben Adlam for providing valuable feedback on a draft. T.M. was supported by NSF through award DMS-2031883 and the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning. T.M. also acknowledge the NSF grant CCF-2006489 and the ONR grant N00014-18-1-2729. The work of Yue M. Lu is supported by a Harvard FAS Dean's competitive fund award for promising scholarship, and by the US National Science Foundation under grant CCF-1910410.

References

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

[2] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

[3] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[4] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.

[6] William Beckner. Sobolev inequalities, the poisson semigroup, and analysis on the sphere sn. *Proceedings of the National Academy of Sciences*, 89(11):4816–4819, 1992.

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

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

[9] Jennifer Bryson, Roman Vershynin, and Hongkai Zhao. Marchenko–pastur law with relaxed independence conditions. *Random Matrices: Theory and Applications*, page 2150040, 2021.

[10] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021.

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[12] Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14):2975, 1999.

[13] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

[14] Reza Rashidi Far, Tamer Oraby, Wlodzimierz Bryc, and Roland Speicher. Spectra of large block matrices. *arXiv preprint cs/0610045*, 2006.

[15] Christopher Frye and Costas J Efthimiou. Spherical harmonics in p dimensions. *arXiv preprint arXiv:1205.3548*, 2012.

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

[17] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

[18] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

[19] Sebastian Goldt, Galen Reeves, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. 2020.

[20] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[21] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[22] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

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

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[25] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[26] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

[27] Licong Lin and Edgar Dobriban. What causes the test error? going beyond bias-variance via anova. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

[28] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[29] Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices. *arXiv preprint arXiv:2205.06308*, 2022.

[30] Horia Mania and Suvrit Sra. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020.

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

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

[33] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[34] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[35] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

[36] Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A Alemi, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. *arXiv preprint arXiv:1912.02803*, 2019.

[37] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[38] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021.

[39] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[41] Denny Wu and Ji Xu. On the optimal weighted $\backslash ell_2$ regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

[42] Lechao Xiao. Eigenspace restructuring: a principle of space and frequency in neural networks. *arXiv preprint arXiv:2112.05611*, 2021.

[43] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

[44] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 7
   (c) Did you discuss any potential negative societal impacts of your work? [No] Pure theoretical work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We did not provide code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. 6
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?[Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Sec. 6
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**

A Appendix Guidelines

The appendix is organized as follows. We prove Theorem 1 and Theorem 3 in Sec. B. In Sec. C, we prove the test errors for dot-product kernels, namely, Theorem 2 and for one-hidden-layer convolutional kernels, namely, Theorem 4. The proof also shows how to reduce the test error of multiple-layer convolutional kernels to evaluating Eq. (31) and Eq. (32). Finally, in Sec.D, we provide additional plots to empirically verify that the finite-size correction becomes smaller as d grows larger.

B Proof of Theorem 1

We begin with some notations. For positive numbers a and b, we use $a \lesssim b$ to mean there is a constant independent of d such that $a \leq Cb$. In addition, $a \sim b$, if $a \lesssim b$ and $b \lesssim a$.

The proof of Theorem 3 is similar. We only present the proof of Theorem 1. Our proof is based on the following result from [5].

Lemma 1 ([5])**.** *Let $\boldsymbol{x}_p \in \mathbb{R}^p$ be random vectors and $\boldsymbol{X} = [\boldsymbol{x}_{p1}, \dots \boldsymbol{x}_{pn}]$ be a $p \times m$ matrix with iid columns. If for every $\{\boldsymbol{A}_p\}_p$, $p \times p$ matrix with uniform operator norm,*

$$\frac{1}{p^2}\mathbb{E}|\boldsymbol{x}_p^T \boldsymbol{A}_p \boldsymbol{x}_p - \mathrm{Tr}(\boldsymbol{A}_p))|^2 \to 0, \tag{33}$$

then the empirical spectral distribution of $\frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T$ converges to μ_α weakly if $p/m \to \alpha \in (0, \infty)$.

We prove a slightly more general version.

Theorem 5. *Let $r \in \mathbb{N}$ and $\alpha \in (0, \infty)$ be fixed. Assume $m = m(d)$ with $N(d, r)/m \to \alpha \in (0, \infty)$ as $d \to \infty$. Let $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{d})$ be a sequence of functions defined on $\mathbb{S}_{d-1}$ such that*

(1.) the cardinality of $\boldsymbol{u}$ satisfies $|\boldsymbol{u}|/d^r \to 0$ as $d \to \infty$;

(2.) the functions in $\boldsymbol{u}$ and Y_r are mutually orthogonal;

(3.) for any unit vector θ, let $\mathbb{E}_{\boldsymbol{x}}|\theta^T Z_r(\boldsymbol{x})|^4 \lesssim 1$ uniformly of d and θ, where $Z_r(\boldsymbol{x})$ is the concatenation of $\boldsymbol{u}(\boldsymbol{x})$ and $Y_r(\boldsymbol{x})$.

Let $Z_r(\boldsymbol{X})$ be the concatenation of $Y_r(\boldsymbol{X})$ and $\boldsymbol{u}(\boldsymbol{X})$. Then the empirical spectral distribution of $\frac{1}{m}Z_r(\boldsymbol{X})^T Z_r(\boldsymbol{X})$ converges in distribution to the Marchenko-Pastur distribution μ_α.

We mainly use the case when $\boldsymbol{u}$ is the empty set, i.e. $Z_r = Y_r$ and the case when $\boldsymbol{u} = [Y_{kl}]_{k<r}^T$, i.e. $Z_r = Y_{\leq r}$.

Proof of Theorem 5. We apply Lemma 1 to $Z_r^T(\boldsymbol{X})$. We only need to show for matrices $\boldsymbol{A} = \boldsymbol{A}^{(d)}$ with $\|\boldsymbol{A}\|_{op} \leq 1$,

$$(|\boldsymbol{u}| + N(d, r))^{-2}\mathbb{E}_x|Z_r(\boldsymbol{x})^T \boldsymbol{A} Z_r(\boldsymbol{x}) - \mathrm{Tr}(\boldsymbol{A})|^2 \to 0 \quad \text{as} \quad d \to \infty\,, \tag{34}$$

or, equivalently

$$N(d, r)^{-2}\mathbb{E}_x|Z_r(\boldsymbol{x})^T \boldsymbol{A} Z_r(\boldsymbol{x}) - \mathrm{Tr}(\boldsymbol{A})|^2 \to 0 \quad \text{as} \quad d \to \infty. \tag{35}$$

since $|\boldsymbol{u}| = o(d^r)$. The assumption $\|\boldsymbol{A}\|_{op} \leq 1$ implies that the absolute values of all entries of $\boldsymbol{A}$ are bounded 1. A key observation in proving the above estimate is that, up to a unitary transformation, almost all functions in $\{Y_{r,l}(\boldsymbol{x})\}_l$ are monomials of the form

$$g_{\boldsymbol{i}}(\boldsymbol{x}) = C_{d,r} \prod_{i \in \boldsymbol{i}} x_i \tag{36}$$

where $\boldsymbol{i} \subseteq [d]$ with $|\boldsymbol{i}| = r$ and $C_{d,r}$ is a normalizing factor such that

$$C_{d,r}^2 \int_{\mathbb{S}_{d-1}} \prod_{i \in \boldsymbol{i}} |x_i|^2 d\boldsymbol{x} = 1\,. \tag{37}$$

We prove later that for any finite integer $r \geq 1$,

$$\int_{\mathbb{S}_{d-1}} \prod_{i \in \boldsymbol{i}} |x_i|^2 d\boldsymbol{x} = \prod_{i \in \boldsymbol{i}} \int_{\mathbb{S}_{d-1}} |x_i|^2 d\boldsymbol{x} + O(d^{-r-1}) = d^{-r} + O(d^{-r-1}) \tag{38}$$

Now we proceed to prove Eq. (35). First note that $\mathbb{G} = \{g_{\boldsymbol{i}} : \boldsymbol{i} \subseteq [d], |\boldsymbol{i}| = r\}$ is an orthonormal set. This can be proved by noticing that if $\boldsymbol{i} \neq \boldsymbol{j}$, then there is $i \in \boldsymbol{i}$ but $i \notin \boldsymbol{j}$. Clearly, the symmetries of the measure on $\mathbb{S}_{d-1}$ implies

$$\int_{\mathbb{S}_{d-1}} g_{\boldsymbol{i}}(\boldsymbol{x}) g_{\boldsymbol{j}}(\boldsymbol{x}) d\boldsymbol{x} = 0\,. \tag{39}$$

We choose $\mathbb{B} = \{b_j\}_{j \in [p]}$ so that $\mathbb{G} \sqcup \mathbb{B}$ forms an orthonormal basis of $Z_r(\boldsymbol{x})$. Note that the cardinality of $\mathbb{B}$ is $o(d^r)$. Indeed,

$$p = |\boldsymbol{u}| + N(d,r) - \binom{d}{r} = |\boldsymbol{u}| + \binom{d+r-2}{r} + \binom{d+r-3}{r-1} - \binom{d}{r} = |\boldsymbol{u}| + O(d^{r-1}) = o(d^r) \tag{40}$$

Thus $\frac{p}{N(d,r)} \to 0$ as $d \to \infty$. After a change of basis, we can assume $Z_r(\boldsymbol{x}) = [\boldsymbol{g}(\boldsymbol{x})^T, \boldsymbol{b}(\boldsymbol{x})^T]^T$, where $\boldsymbol{g} = [g_{\boldsymbol{i}}]_{\boldsymbol{i}}^T$ and $\boldsymbol{b} = [b_j]_{j \in [p]}^T$. Here we use the fact that Eq. (35) holds for all $\boldsymbol{A}$ with uniform operator norms is equivalent to that it holds for $\boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q}$ for all such $\boldsymbol{A}$ and any unitary matrix $\boldsymbol{Q}$. We write,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \tag{41}$$

where $\boldsymbol{A}_{11}$ is the upper left $\binom{d}{r} \times \binom{d}{r}$ block of $\boldsymbol{A}$, $\boldsymbol{A}_{22}$ is the lower right $p \times p$ block of $\boldsymbol{A}$ and the other two blocks are defined similarly. Note that $\|\boldsymbol{A}_{ij}\|_{op} \leq \|\boldsymbol{A}\|_{op} \leq 1$ for $i, j \in \{1, 2\}$. As such, we have

$$N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |Z_r(\boldsymbol{x})^T \boldsymbol{A} Z_r(\boldsymbol{x}) - \mathrm{Tr}(\boldsymbol{A})|^2 \leq 5(I_1 + I_2 + I_3 + I_4 + I_5) \tag{42}$$

where

$$I_1 = N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |\boldsymbol{g}(\boldsymbol{x})^T \boldsymbol{A}_{11} \boldsymbol{g}(\boldsymbol{x}) - \mathrm{Tr}(\boldsymbol{A}_{11})|^2 \tag{43}$$

$$I_2 = N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |\boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{A}_{22} \boldsymbol{b}(\boldsymbol{x})|^2 \tag{44}$$

$$I_3 = N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |\boldsymbol{g}(\boldsymbol{x})^T \boldsymbol{A}_{12} \boldsymbol{b}(\boldsymbol{x})|^2 \tag{45}$$

$$I_4 = N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |\boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{A}_{21} \boldsymbol{g}(\boldsymbol{x})|^2 \tag{46}$$

$$I_5 = N(d,r)^{-2} |\mathrm{Tr}(\boldsymbol{A}_{22})|^2 \tag{47}$$

We prove $I_i \to 0$ for $1 \leq i \leq 5$. The $i = 1$ case is the most difficult and the others are straightforward since $pN(d,r)^{-1} \to 0$. E.g., when $i = 3$

$$I_3 \leq N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{g}(\boldsymbol{x})\|_{l_2}^2 \|\boldsymbol{A}_{12}\|_{op}^2 \|\boldsymbol{b}(\boldsymbol{x})\|_{l_2}^2 \tag{48}$$

$$\leq N(d,r)^{-2} \left(\mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{g}(\boldsymbol{x})\|_{l_2}^4 \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{b}(\boldsymbol{x})\|_{l_2}^4 \right)^{1/2} \tag{49}$$

$$\leq N(d,r)^{-2} \max_{\boldsymbol{i},j} (\mathbb{E}_{\boldsymbol{x}} |g_{\boldsymbol{i}}(\boldsymbol{x})|^4 \mathbb{E}_{\boldsymbol{x}} |b_j(\boldsymbol{x})|^4)^{1/2} pN(d,r) \tag{50}$$

$$= CpN(d,r)^{-1} \to 0 \tag{51}$$

where we have set $C = (\mathbb{E}_{\boldsymbol{x}} |g_{\boldsymbol{i}}(\boldsymbol{x})|^4 \mathbb{E}_{\boldsymbol{x}} |b_j(\boldsymbol{x})|^4)^{1/2}$, which is $O(1)$ due to assumption (3.) in Theorem 5. The bounds for I_2 and I_4 can be obtained similarly. For I_5, we simply use $\mathrm{Tr}(\boldsymbol{A}_{22}) \leq p\|\boldsymbol{A}\|_{op} \leq p$.

It remains to control I_1. To ease the notation, denote $\boldsymbol{B} = \boldsymbol{A}_{11}$. We split $I_1 \leq I_{11} + I_{12}$, where I_{11} and I_{12} are the diagonal and the off-diagonal parts, resp.,

$$I_{11} = 2N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |\sum_{\boldsymbol{i}} B_{\boldsymbol{i}\boldsymbol{i}} (g_{\boldsymbol{i}}^2(\boldsymbol{x}) - 1)|^2 \tag{52}$$

$$I_{12} = 2N(d,r)^{-2} \mathbb{E}_{\boldsymbol{x}} |\sum_{\boldsymbol{i} \neq \boldsymbol{j}} B_{\boldsymbol{i}\boldsymbol{j}} g_{\boldsymbol{i}} g_{\boldsymbol{j}}(\boldsymbol{x})|^2 \tag{53}$$

Bounding the diagonal part I_{11}. Using $\mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}(\boldsymbol{x})^2 = 1$, we have

$$I_{11} = 2N(d,r)^{-2}\mathbb{E}_{\boldsymbol{x}} \sum_{\boldsymbol{i},\boldsymbol{j}} B_{\boldsymbol{i}\boldsymbol{i}} B_{\boldsymbol{j}\boldsymbol{j}} (g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1) \tag{54}$$

We spit the proof into two cases: $\boldsymbol{i} \cap \boldsymbol{j} = \emptyset$ and $\boldsymbol{i} \cap \boldsymbol{j} \neq \emptyset$. The following is the key estimate to handle the first case.

Lemma 2. *If $\boldsymbol{i} \cap \boldsymbol{j} = \emptyset$,*

$$\max_{\boldsymbol{i},\boldsymbol{j}} |\mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1| \lesssim d^{-1} \,. \tag{55}$$

We prove this lemma later. We show how to use this lemma to handle the $\boldsymbol{i} \cap \boldsymbol{j} = \emptyset$ case. Recall that $|B_{\boldsymbol{i},\boldsymbol{j}}| \leq 1$ and the number of tuples $(\boldsymbol{i}, \boldsymbol{j})$ is fewer than $N(d,r)^2$. We have

$$N(d,r)^{-2} |\mathbb{E}_{\boldsymbol{x}} \sum_{\boldsymbol{i},\boldsymbol{j}, \boldsymbol{i}\cap\boldsymbol{j}=\emptyset} B_{\boldsymbol{i}\boldsymbol{i}} B_{\boldsymbol{j}\boldsymbol{j}} (g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1)| \tag{56}$$

$$\leq N(d,r)^{-2} \sum_{\boldsymbol{i},\boldsymbol{j}, \boldsymbol{i}\cap\boldsymbol{j}=\emptyset} \max_{\boldsymbol{i},\boldsymbol{j}} |\mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1| \tag{57}$$

$$\leq \max_{\boldsymbol{i},\boldsymbol{j}} |\mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1| \lesssim d^{-1} \,. \tag{58}$$

We turn to $|\boldsymbol{i} \cap \boldsymbol{j}| = t$, $1 \leq t \leq r$. For each fixed $\boldsymbol{i}$, the number of choices of $\boldsymbol{j}$ is

$$\sum_{1\leq t\leq r} \binom{r}{t}\binom{d-r}{r-t} \lesssim \sum_{1\leq t\leq r} d^{r-t} \sim d^{r-1} \tag{59}$$

As such,

$$2N(d,r)^{-2}\mathbb{E}_{\boldsymbol{x}} \sum_{\boldsymbol{i},\boldsymbol{j}, \boldsymbol{i}\cap\boldsymbol{j}\neq\emptyset} |B_{\boldsymbol{i}\boldsymbol{i}} B_{\boldsymbol{j}\boldsymbol{j}} (g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1)| \tag{60}$$

$$\lesssim N(d,r)^{-2} N(d,r) d^{r-1} \max_{\boldsymbol{i},\boldsymbol{j}} \mathbb{E}_{\boldsymbol{x}} |g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1| \tag{61}$$

$$\lesssim d^{-1} \tag{62}$$

Off-diagonal terms I_{12}. Bounding the off-diagonal terms can be reduced to a combinatorics problem, which is similar to the random tensor model considered in [9]. We need to estimate

$$I_{12} = 2N(d,r)^{-2} \sum_{\boldsymbol{i}\neq\boldsymbol{j}, \boldsymbol{l}\neq\boldsymbol{k}} B_{\boldsymbol{i}\boldsymbol{j}} B_{\boldsymbol{l}\boldsymbol{k}} \mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}(\boldsymbol{x}) g_{\boldsymbol{j}}(\boldsymbol{x}) g_{\boldsymbol{l}}(\boldsymbol{x}) g_{\boldsymbol{k}}(\boldsymbol{x}) \tag{63}$$

By symmetries of the uniform measure on the sphere, we can assume the monomial $g_{\boldsymbol{i}}(\boldsymbol{x}) g_{\boldsymbol{j}}(\boldsymbol{x}) g_{\boldsymbol{l}}(\boldsymbol{x}) g_{\boldsymbol{k}}(\boldsymbol{x})$ has no linear factor, that is the degree of any x_i in this monomial must be at least 2 if not 0. In addition, for such monomials, the Holder inequality and hypercontractivities yield,

$$|\mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}(\boldsymbol{x}) g_{\boldsymbol{j}}(\boldsymbol{x}) g_{\boldsymbol{l}}(\boldsymbol{x}) g_{\boldsymbol{k}}(\boldsymbol{x})| \leq \max_{\boldsymbol{i}} \mathbb{E}_{\boldsymbol{x}} |g_{\boldsymbol{i}}(\boldsymbol{x})|^4 \lesssim \max_{\boldsymbol{i}} (\mathbb{E}_{\boldsymbol{x}} |g_{\boldsymbol{i}}(\boldsymbol{x})|^2)^2 = 1 \,. \tag{64}$$

As such, we only need to show that the growth of the number of such quadruples $(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}, \boldsymbol{l})$, as a function of d, is slower than $N(d,r)^2 \sim d^{2r}$. We proceed to prove this claim. For each fixed $\boldsymbol{i}$, let $t = |\boldsymbol{i} \cap \boldsymbol{j}|$ where $0 \leq t \leq r-1$ ($t \neq r$ since $\boldsymbol{i} \neq \boldsymbol{j}$). Let $J(\boldsymbol{i}; t)$ denote the set of such $\boldsymbol{j}$, whose cardinality is

$$|J(\boldsymbol{i}; t)| = \binom{r}{t}\binom{d-r}{r-t} < r^t d^{r-t} \lesssim d^{r-t} \,. \tag{65}$$

Next we estimate the number of tuples $(\boldsymbol{l}, \boldsymbol{k})$. Let $w = |(\boldsymbol{l} \cup \boldsymbol{k}) \backslash (\boldsymbol{i} \cup \boldsymbol{j})|$. Since $|\boldsymbol{l} \cup \boldsymbol{k} \cup \boldsymbol{i} \cup \boldsymbol{j}| \leq 2r$ and $|\boldsymbol{i} \cup \boldsymbol{j}| = 2r - t$, we have $w \leq t$. The cardinality of choosing such $\boldsymbol{k} \cup \boldsymbol{l}$ cannot exceed

$$\sum_{0\leq w\leq t} \binom{d-(2r-t)}{w} \sum_{v=0}^{2r-w} \binom{2r}{v} \lesssim d^t \,. \tag{66}$$

With $\boldsymbol{k} \cup \boldsymbol{l}$ given, the pair of $(\boldsymbol{k}, \boldsymbol{l})$ cannot exceed $\binom{2r}{r}^2 \lesssim 1$. Thus, with $\boldsymbol{i}, \boldsymbol{j}$ and t fixed, the number of pairs of $(\boldsymbol{k}, \boldsymbol{l})$ is $\lesssim d^t$. Using $|B_{\boldsymbol{lk}}| \leq 1$ and $\max_{\boldsymbol{i}} (\sum_{\boldsymbol{j}} B_{\boldsymbol{ij}}^2)^{1/2} \leq \|\boldsymbol{B}\|_{op} \leq 1$, we have

$$N(d,r)^2 I_{12} \lesssim \sum_{\boldsymbol{i}} \sum_{0 \leq t \leq r-1} \sum_{j \in J(\boldsymbol{i};t)} B_{\boldsymbol{ij}} d^t \tag{67}$$

$$\leq \sum_{\boldsymbol{i}} \sum_{0 \leq t \leq r-1} |J(\boldsymbol{i};t)|^{1/2} (\sum_{\boldsymbol{j} \in J(\boldsymbol{i};t)} B_{\boldsymbol{ij}}^2)^{1/2} d^t \tag{68}$$

$$\lesssim \|\boldsymbol{A}\|_{op} \sum_{\boldsymbol{i}} \sum_{0 \leq t \leq r-1} d^{(r-t)/2} d^t \tag{69}$$

$$\lesssim N(d,r) d^{2r-1/2} \tag{70}$$

which gives $I_{12} \lesssim d^{-1/2}$. □

Proof of Lemma 2. It suffices to prove that for any finite integer $j > 1$,

$$\int_{\mathbb{S}_{d-1}} \prod_{1 \leq t \leq j} x_t^2 d\boldsymbol{x} = d^{-j} + O(d^{-j-1}). \tag{71}$$

Indeed, assuming this estimate, we have $C_{d,j}^{-2} = d^{-j} + O(d^{-j-1})$ and $C_{d,j}^2 = d^j + O(d^{j-1})$. For any $\boldsymbol{i}$ and $\boldsymbol{j}$ with $\boldsymbol{i} \cap \boldsymbol{j} = \emptyset$,

$$\mathbb{E}_{\boldsymbol{x}} g_{\boldsymbol{i}}^2(\boldsymbol{x}) g_{\boldsymbol{j}}^2(\boldsymbol{x}) - 1 = C_{d,r}^4 \int_{\mathbb{S}_{d-1}} \prod_{t \in \boldsymbol{i} \cup \boldsymbol{j}} x_t^2 d\boldsymbol{x} = C_{d,r}^4 C_{d,2r}^{-2} - 1 = O(d^{-1}) \tag{72}$$

It remains to prove Eq. (71). By symmetries,

$$\int_{\mathbb{S}_{d-1}} x_t^2 d\boldsymbol{x} = \frac{1}{d} \int_{\mathbb{S}_{d-1}} \sum_{1 \leq t \leq d} x_t^2 d\boldsymbol{x} = \frac{1}{d} \int_{\mathbb{S}_{d-1}} 1 d\boldsymbol{x} = \frac{1}{d}. \tag{73}$$

By symmetries again,

$$\int_{\mathbb{S}_{d-1}} \prod_{1 \leq t \leq j-1} x_t^2 d\boldsymbol{x} \tag{74}$$

$$= \int_{\mathbb{S}_{d-1}} \prod_{1 \leq t \leq j-1} x_t^2 \left(\sum_{1 \leq i \leq j-1} x_i^2 + \sum_{j \leq i \leq d} x_i^2 \right) d\boldsymbol{x} \tag{75}$$

$$= (d-j+1) \int_{\mathbb{S}_{d-1}} \prod_{1 \leq t \leq j} x_t^2 d\boldsymbol{x} + (j-1) \int_{\mathbb{S}_{d-1}} x_1^2 \prod_{1 \leq t \leq j-1} x_t^2 d\boldsymbol{x} \tag{76}$$

We use hypercontractivities to bound the error term, namely, the second term. Recall that for any $q \geq 2$ and any polynomial defined on the sphere,

$$\left(\int_{\mathbb{S}_{d-1}} |f(\boldsymbol{x})|^q d\boldsymbol{x} \right)^{1/q} \leq (q-1)^{\deg(f)/2} \left(\int_{\mathbb{S}_{d-1}} |f(\boldsymbol{x})|^2 d\boldsymbol{x} \right)^{1/2}. \tag{77}$$

Setting $f(\boldsymbol{x}) = x_t$ (with $\deg(f) = 1$) gives

$$\int_{\mathbb{S}_{d-1}} |x_t|^q d\boldsymbol{x} \leq (q-1)^{q/2} \left(\int_{\mathbb{S}_{d-1}} |x_t|^2 d\boldsymbol{x} \right)^{q/2} = (q-1)^{q/2} d^{-\frac{q}{2}}. \tag{78}$$

By Holder's inequality and symmtries

$$\int_{\mathbb{S}_{d-1}} x_1^2 \prod_{1 \leq t \leq j-1} x_t^2 d\boldsymbol{x} \leq \left(\int_{\mathbb{S}_{d-1}} |x_1|^{2j} d\boldsymbol{x} \prod_{1 \leq t \leq j-1} \int_{\mathbb{S}_{d-1}} |x_t|^{2j} d\boldsymbol{x} \right)^{\frac{1}{j}} \tag{79}$$

$$= \int_{\mathbb{S}_{d-1}} |x_1|^{2j} d\boldsymbol{x} \leq (2j-1)^j d^{-j} \tag{80}$$

Thus

$$\int_{\mathbb{S}_{d-1}} \prod_{1\le t\le j-1} x_t^2 d\boldsymbol{x} = (d-j+1)\int_{\mathbb{S}_{d-1}} \prod_{1\le t\le j} x_t^2 d\boldsymbol{x} + O(d^{-j}) \tag{81}$$

and

$$\int_{\mathbb{S}_{d-1}} \prod_{1\le t\le j} x_t^2 d\boldsymbol{x} = \frac{1}{d-(j-1)}\int_{\mathbb{S}_{d-1}} \prod_{1\le t\le j-1} x_t^2 d\boldsymbol{x} + O(d^{-j-1})\,. \tag{82}$$

Finally, Eq. (71) is a consequence of this estimate and induction.

□

C Generalization

We aim to obtain the asymptotic formulas for the test error in this section. In the high-level, we decompose the empirical kernel $K(\boldsymbol{X},\boldsymbol{X})$ into low-, critical- and high-frequency modes, where we have concentration in the low- and high-frequency parts of the kernel. The test error associated to these two parts are easier to handle. The critical-frequency part is more difficult in which random matrix behaviors emerge, namely, the Marchenko-Pastur distribution. As such, our first step is to remove the contribution in the test error coming from the non-critical frequency parts. After that, the remaining is essentially equivalent to computing the trace of certain functional forms related to the Marchenko-Pastur distribution.

We consider a general setting that includes the dot-product kernels, the one-hidden-layer and the multiple-layer convolutional kernels (NNGP and NT kernels.) In what follows, we use $\Delta_d, \Delta_d', \Delta_d''$, etc. to represent quantities that converge to 0 in probability (the absolute value of a scalar, the norm of a vector, the operator norm of a matrix, etc.), whose exact form may change from line to line.

C.1 Setup

For $d\in\mathbb{N}^*$, let $\boldsymbol{\mathcal{X}}^{(d)}\subseteq\mathbb{R}^d$ be the input space associated with a probability measure $\sigma^{(d)}$ and a kernel function $K^{(d)}$. Assume the kernel function has the following eigen-structure

$$K^{(d)}(\boldsymbol{x},\boldsymbol{x}') = \sum_{k\ge1}\sum_{n\in[E_k]}(\sigma_{kn}^{(d)})^2\sum_{l\in N_{kn}^{(d)}}\phi_{knl}^{(d)}(\boldsymbol{x})\phi_{knl}^{(d)}(\boldsymbol{x}') \tag{83}$$

in the sense $K^{(d)}$, as the integral operator from $L^2(\boldsymbol{\mathcal{X}}^{(d)},\sigma^{(d)})$ to itself,

$$K^{(d)}\phi_{knl}^{(d)}(\boldsymbol{x}) = \int K^{(d)}(\boldsymbol{x},\boldsymbol{x}')\phi_{knl}^{(d)}(\boldsymbol{x}')\sigma^{(d)}(d\boldsymbol{x}') = (\sigma_{kn}^{(d)})^2\phi_{knl}^{(d)}(\boldsymbol{x}). \tag{84}$$

Here $\{\phi_{knl}^{(d)}\}_{knl}$ is an orthonormal basis of $L^2(\boldsymbol{\mathcal{X}}^{(d)},\sigma^{(d)})$. We also assume $K^{(d)}$ is a trace-class operator, i.e.,

$$\sum_{knl}\langle K^{(d)}\phi_{knl}^{(d)},\phi_{knl}^{(d)}\rangle = \sum_{k\ge1}\sum_{n\in[E_k]} N_{kn}^{(d)}(\sigma_{kn}^{(d)})^2 < \infty\,. \tag{85}$$

In the above notations, we use the triplet (k,n,l) to index the eigenfunctions $\phi_{knl}^{(d)}$. The tuple (k,n) determines the eigenspace, whose eigenvalue is of the form "$(\sigma_{kn}^{(d)})^2 = C_n d^{-s_k} + \text{Lower Order}$" and l lists all eigenfunctions in the kn-eigenspace. We make the following assumptions.

Kernel Assumptions.

(1.) **Spectral Gap.** There are $\delta_0>0$ and a sequence of strictly increasing positive real numbers $\{s_k\}$ with $|s_k-s_{k-1}|\ge\delta_0$ for all $k\ge2$ such that

$$(\sigma_{kn}^{(d)})^2 \sim d^{-s_k} \sim (N_{kn}^{(d)})^{-1} \tag{86}$$

Moreover, $\{E_k\} \subseteq \mathbb{N}^*$ is independent from d which grows at most exponentially. We also assume that there is a sequence of real numbers $\{\hat{h}_{kn}^2\}_{kn}$ with $\hat{h}_{kn}^2 \neq 0$ unless k is sufficiently large and

$$\sum_k \sum_{n\in[E_k]} \hat{h}_{kn}^2 < \infty \quad \text{and} \quad (\sigma_{kn}^{(d)})^2 N_{kn}^{(d)} = \hat{h}_{kn}^2 \quad \text{as} \quad d \to \infty \tag{87}$$

(2.) **Hypercontractivity Inequalities.** For any $p \geq 2$ there are constant $C_{p,k}$ such that for any function f in the closure of $\mathrm{Span}\{\phi_{jnl}^{(d)}\}_{j\leq k}$

$$\|f\|_p \leq C_{p,k}\|f\|_2 \tag{88}$$

(3.) **Concentration of Quadratic Forms.** Let $\phi_k^{(d)}(\boldsymbol{x})$ denote the column vector consists of elements $\{\phi_{knl}^{(d)}(\boldsymbol{x})\}_{l\in[N_{kn}(d)n\in[E_k]}$. For every sequence of matrices $\{\boldsymbol{A}^{(d)}\}$ with uniformly bounded operator norm,

$$\left(\sum_{n\in[E_k]} N_{kn}^{(d)}\right)^{-2} \mathbb{E}_{\boldsymbol{x}}|\phi_k^{(d)}(\boldsymbol{x})^\top \boldsymbol{A}^{(d)}\phi_k^{(d)}(\boldsymbol{x}) - \mathrm{Tr}\,\boldsymbol{A}^{(d)}|^2 \to 0 \quad \text{as} \quad d \to \infty. \tag{89}$$

(4.) **Addition Theorem.** For $k \in \mathbb{N}^*$ and $n \in [E_k]$ and $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}^{(d)}$

$$\sum_{l\in[N_{kn}^{(d)}]} \phi_{knl}^{(d)}(\boldsymbol{x})^2 = N_{kn}^{(d)} \tag{90}$$

Let us briefly explain the assumptions. The **Spectral Gap** assumption basically says, we can classify the eigenvectors into countably many categories indexed by $k \in \mathbb{N}^*$. In the k-th category, it has exactly E_k many eigenspaces, each of them has dimensions $\sim d^{s_k}$ and eigenvalues d^{-s_k}. It also implies the number of eigenfunctions with eigenvalues $\lesssim d^{-s_k}$ is $\sim d^{s_k}$. Assumptions (1.), (2.) and (4.) together are stronger than those in Theorem 6 in Mei et al. [32] (and slightly less technical), which allow us to apply kernel concentration from Mei et al. [32]. In particular, they imply concentration of the low- and high-frequency parts of the empirical kernel $K^{(d)}(\boldsymbol{X}, \boldsymbol{X})$. Finally, Assumption (3) is designed to meet the requirements in Lemma 1, which allows us to claim Marchenko-Pastur type behavior of the gram matrix induced by the feature map ϕ_k. We provide a couple examples.

Example 1 (Dot-product Kernels)**.** *When $\boldsymbol{\mathcal{X}}^{(d)} = \mathbb{S}_{d-1}$ and $K^{(d)}$ is the dot-product kernel, we have $E_k = 1$, $s_k = k$, $N_{kn} = N(d,k) \sim d^k/k!$, and $\phi_{knl} = Y_{kl}$ (note that $n = 0$ since $E_k = 1$.) Note that by the Addition Theorem of spherical harmonics (Theorem 4.11 in Frye and Efthimiou [15]),*

$$\sum_{l\in[N(d,k)]} Y_{kl}(\boldsymbol{x})^2 = N(d,k) \tag{91}$$

Example 2 (One-hidden-layer Convolutional Kernels)**.** *Sightly more general setting is the one-layer convolutional kernel (NNGP or NT kernels). In this case, $\boldsymbol{\mathcal{X}}^{(d)} = \mathbb{S}_{d_0-1}^p$ where p is the number of patches and the input dimension is $d = pd_0$. We can set either $p = O(1)$ (i.e. independent of $d_0 \to \infty$) or $p \sim d^{\alpha_p}$ for some $\alpha_p > 0$. This kernel is essentially the sum of p dot-product kernels. As such, $E_k = 1$, $N_{kn}(d) = pN(d_0,k) \sim pd_0^k/k!$ and $(\sigma_{kn}^u pd)^2 \sim (pd_0^k)^{-1}$. If $p \sim d^{\alpha_p}$ and $d_0 \sim d^{\alpha_{d_0}}$ with $\alpha_{d_0} + \alpha_p = 1$, we have $s_k = \alpha_p + k\alpha_{d_0}$ and d^{-s_k} is the decay rate of the k-th order spherical harmonics.*

Example 3 (Multiple-layer Convolutional Kernels)**.** *General convolutional kernels are much more complicated [42]. In this case, $\boldsymbol{\mathcal{X}}^{(d)} = \mathbb{S}_{d_0-1}^p$ where p is the number of patches and the input dimension is $d = pd_0$. We additionally assume, $p = k_0^{L-1}$ for some $k_0 \in \mathbb{N}^*$ and the network has L convolutional layers with filter size and strides being the same in each layer (equal to d_0 in the first layer and to k_0 for the remaining $(L-1)$ layers.) The eigenstructures of such kernels are studied in Xiao [42]. The eigenfunctions are tensor products of spherical harmonics defined on copies of $\mathbb{S}_{d_0-1}$,*

$$Y_{\boldsymbol{k},\boldsymbol{l}}(\boldsymbol{x}) = \prod_{i\in[p]} Y_{k_i l_i}(\boldsymbol{x}_i) \tag{92}$$

The eigenvalues are more complicated to compute as they depend on both the frequencies of $Y_{\boldsymbol{k},\boldsymbol{l}}$ and the topologies of the networks. When $d_0 \propto d^{\alpha_{d_0}}$ and $k_0 \propto d^{\alpha_{k_0}}$ with $\alpha_{d_0} + (L-1)\alpha_{k_0} = 1$ and $\alpha_{k_0}, \alpha_{d_0} > 0$, the eigenvalue of $Y_{\boldsymbol{kl}}$ is $\propto d^{-(\mathscr{F}(\boldsymbol{k})+\mathscr{S}(\boldsymbol{k}))}$, as $d \to \infty$. Here $\mathscr{F}(\boldsymbol{k}) \equiv |\boldsymbol{k}|\alpha_{d_0}$ is the frequency index of $Y_{\boldsymbol{kl}}$ and $\mathscr{S}(\boldsymbol{k}) = J_{\boldsymbol{k}}\alpha_{k_0}$ is the spatial index, where $J_{\boldsymbol{k}}$ is the number of edges in the sub-tree connecting all interacting patches (i.e. $k_i \neq 0$) to the output; see Xiao [42] for more details.

As there can possibly exist $\boldsymbol{k}$ and $\boldsymbol{k}'$ with $\mathscr{F}(\boldsymbol{k})+\mathscr{S}(\boldsymbol{k}) = \mathscr{F}(\boldsymbol{k}')+\mathscr{S}(\boldsymbol{k}')$ (i.e., same order of decay) but $(\mathscr{F}(\boldsymbol{k}), \mathscr{S}(\boldsymbol{k})) \neq (\mathscr{F}(\boldsymbol{k}'), \mathscr{S}(\boldsymbol{k}'))$ (i.e. different space-frequency combination), there can exist more than one eigenspaces whose eigenvalues decay to zero with the same rate $d^{-(\mathscr{F}(\boldsymbol{k})+\mathscr{S}(\boldsymbol{k}))}$, but with different leading coefficients. This is the main reason why we need to allow $|E_k| > 1$ in Eq. (83).

Next we discuss the assumptions on the label function. Let $\boldsymbol{X}$ be the training set with $m \sim d^{s_r}$ many training samples for some $r \in \mathbb{N}^*$ fixed. Then let the ground true label function to be

$$f(\boldsymbol{x}) = \sum_{k\in\mathbb{N}^*} \sum_{n\in[E_k]} \sum_{l\in[N_{kn}^{(d)}]} \hat{f}_{knl}\phi_{knl}^{(d)}(\boldsymbol{x}). \tag{93}$$

Let $N_k^{(d)} = \sum_{n\in[E_k]} N_{kn}^{(d)}$. We assume, for $k \geq r$, $\hat{\boldsymbol{f}}_{kn} = \{\hat{f}_{knl}\}_{l\in[N_{kn}^{(d)}]}$ is a random vector with

$$\mathbb{E}\hat{\boldsymbol{f}}_{kn} = \boldsymbol{0} \quad \text{and} \quad \mathbb{E}\hat{\boldsymbol{f}}_{kn}\hat{\boldsymbol{f}}_{kn}^\top = \frac{\hat{F}_{kn}^2}{N_{kn}^{(d)}}\boldsymbol{I}_{N_{kn}^{(d)}}. \tag{94}$$

and $\{\hat{\boldsymbol{f}}_{kn}\}_{n\in[E_k],k\geq r}$ are mutually independent. One concrete example is

$$\hat{\boldsymbol{f}}_{kn} \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\hat{F}_{kn}^2}{N_{kn}^{(d)}}\boldsymbol{I}_{N_{kn}^{(d)}}\right). \tag{95}$$

For $k < r$, we assume the coefficients are deterministic with $\sum_l \hat{f}_{knl}^2 = \hat{F}_{kn}^2$, $\sum_n \hat{F}_{kn}^2 = \hat{F}_k^2$ and

$$\sum_{k\in\mathbb{N}^*} \hat{F}_k^2 < \infty \tag{96}$$

Our goal is to compute the average test error over the random labels defined above in the scaling limit $m \sim d^{s_r}$.

C.2 Structure of the Empirical Kernels

For convenience, denote

$$\phi_{\leq k}^{(d)}(\boldsymbol{x}) = [\phi_{jnl}^{(d)}(\boldsymbol{x})]_{l\in[N_{jn}^{(d)}],1\leq j\leq k,n\in[E_j],}^\top \tag{97}$$

$$N_k^{(d)} = \sum_{n\in[E_k]} N_{kn}^{(d)} \tag{98}$$

$$N_{\leq k}^{(d)} = \sum_{1\leq j\leq k} N_{\leq j}^{(d)} \tag{99}$$

Let $\boldsymbol{x}_i^\top$ be the i-th row of the training matrix $\boldsymbol{X}$. Similarly,

$$Z_k(\boldsymbol{X}) = [\phi_k^{(d)}(\boldsymbol{x}_0), \ldots, \phi_k^{(d)}(\boldsymbol{x}_{m-1})]^\top \qquad Z_{\leq k}(\boldsymbol{X}) = [\phi_{\leq k}^{(d)}(\boldsymbol{x}_0), \ldots, \phi_{\leq k}^{(d)}(\boldsymbol{x}_{m-1})]^\top \tag{100}$$

$$\boldsymbol{\Lambda}_k = \operatorname{diag}\left([(\sigma_{kn}^{(d)})^2\boldsymbol{I}_{N_{kn}^{(d)}}]_{n\in[E_k]}\right) \qquad \boldsymbol{\Lambda}_{\leq k} = \operatorname{diag}\left([(\sigma_{jn}^{(d)})^2\boldsymbol{I}_{N_{jn}^{(d)}}]_{n\in[E_j],1\leq j\leq k}\right) \tag{101}$$

Note that $Z_k(\boldsymbol{X})$ ($Z_{\leq k}(\boldsymbol{X})$) is an $m \times N_k^{(d)}$ ($m \times N_{\leq k}^{(d)}$) matrix and $\boldsymbol{\Lambda}_k$ ($\boldsymbol{\Lambda}_{\leq k}$) is an $N_k^{(d)} \times N_k^{(d)}$ ($N_{\leq k}^{(d)} \times N_{\leq k}^{(d)}$) diagonal matrix. The followings are defined similarly,

$$Z_{<k}(\boldsymbol{X}), \quad Z_{kn}(\boldsymbol{X}), \quad \boldsymbol{\Lambda}_{<k}, \quad \boldsymbol{\Lambda}_{kn}, \quad N_{<k}^{(d)}, \quad N_{kn}^{(d)}. \tag{102}$$

Next, we decompose the train-train kernel into two parts: the $\leq r$ frequency part and the $> r$ frequency parts,

$$K^{(d)}(\boldsymbol{X},\boldsymbol{X}) = \sum_{k\in\mathbb{N}^*} Z_k(\boldsymbol{X})\boldsymbol{\Lambda}_k Z_k(\boldsymbol{X})^\top = Z_{\leq r}(\boldsymbol{X})\boldsymbol{\Lambda}_{\leq} Z_{\leq r}(\boldsymbol{X})^\top + \sum_{k\geq r+1} Z_k(\boldsymbol{X})\boldsymbol{\Lambda}_k Z_k(\boldsymbol{X})^\top \tag{103}$$

$$= Z_{\leq r}(\boldsymbol{X})\boldsymbol{\Lambda}_{\leq} Z_{\leq r}(\boldsymbol{X})^\top + \sum_{k\geq r+1}\sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^2 Z_{kn}(\boldsymbol{X}) Z_{kn}(\boldsymbol{X})^\top \tag{104}$$

$$\equiv K_{\leq r}^{(d)}(\boldsymbol{X},\boldsymbol{X}) + K_{>r}^{(d)}(\boldsymbol{X},\boldsymbol{X}) \tag{105}$$

Assumptions (1.) (2.) allow us to apply kernel concentration [17, 32], which implies that the low-frequency and high-frequency parts of the empirical kernels are concentrated. By saying concentration in the high-frequency part, we mean

Claim 1. *Let*

$$\Delta_{kn}^{(d)} \equiv \frac{1}{N_{kn}^{(d)}} Z_{kn}(\boldsymbol{X}) Z_{kn}(\boldsymbol{X})^\top - \boldsymbol{I}_m\,. \tag{106}$$

Then

$$\mathbb{E}\sum_{k>r}\sum_{n\in[E_k]} \|\Delta_{kn}^{(d)}\|_{\mathrm{op}} \to 0 \tag{107}$$

The proof of this claim essentially follows from the arguments and results in Theorem 6 of [32]; see Sec. C.7. Thus

$$K_{>r}^{(d)}(\boldsymbol{X},\boldsymbol{X}) = \sum_{k\geq r+1}\sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^2 N_{kn}^{(d)} (\boldsymbol{I}_m + \Delta_{kn}^{(d)}) \tag{108}$$

$$= \sum_{k\geq r+1} \hat{h}_k^2 \boldsymbol{I}_m + \sum_{k\geq r+1}\sum_{n\in[E_k]} \hat{h}_{kn}^2 \Delta_{kn}^{(d)} \tag{109}$$

Denote

$$\hat{h}_{>r}^2 = \sum_{k\geq r+1}\sum_{n\in[E_k]} N_{kn}^{(d)} (\sigma_{kn}^{(d)})^2 \sim 1 \tag{110}$$

$$\Delta_{>r}^{(d)} = \sum_{k\geq r+1}\sum_{n\in[E_k]} \hat{h}_{kn}^2 \Delta_{kn}^{(d)}\,, \tag{111}$$

we can write

$$K_{>r}^{(d)}(\boldsymbol{X},\boldsymbol{X}) = \hat{h}_{>r}^2 \boldsymbol{I}_m + \Delta_{>r}^{(d)}, \quad \text{where} \quad \mathbb{E}\|\Delta_{>r}^{(d)}\|_{\mathrm{op}} \to 0 \tag{112}$$

By saying the low-frequency part of the kernel concentrates, we mean (Theorem 6 [32])

$$\frac{1}{m} Z_<(\boldsymbol{X})^\top Z_<(\boldsymbol{X}) = \boldsymbol{I}_{N_<^{(d)}} + \Delta_{<r}^{(d)}, \quad \text{where} \quad \mathbb{E}\|\Delta_{<r}^{(d)}\|_{\mathrm{op}} \to 0 \tag{113}$$

Finally, Lemma 1 and **Assumption (3.)** imply that if $N_r^{(d)}/m \to \alpha \in (0,\infty)$, then the empirical measure of the critical part of the kernel matrix $\frac{1}{m} Z_r(\boldsymbol{X})^\top Z_r(\boldsymbol{X})$, and the low-and-critical frequency parts $\frac{1}{m} Z_{\leq r}(\boldsymbol{X})^\top Z_{\leq r}(\boldsymbol{X})$ converge to the Marchenko-Pastur distribution μ_α weakly by Lemma 1. In particular, $\|\frac{1}{m} Z_r(\boldsymbol{X})^\top Z_r(\boldsymbol{X})\|_{\mathrm{op}} + \|\frac{1}{m} Z_{\leq r}(\boldsymbol{X})^\top Z_{\leq r}(\boldsymbol{X})\|_{\mathrm{op}} = O(1)$ in probability as $d \to \infty$.

For convenience, we summarize the structure of the empirical kernel in the following.

Corollary 3. *Assume* **Assumptions (1.-4.)**. *Let $r \in \mathbb{N}^*$ and $\alpha > 0$ be fixed and $m = m^{(d)}$ be such that $N_r^{(d)}/m \to \alpha \in (0,\infty)$ as $d \to \infty$. Let $\boldsymbol{X}$, of shape $m \times d$, be the training set matrix whose rows are drawn, uniformly, iid from $\boldsymbol{\mathcal{X}}^{(d)}$. Then we have the following structure for the empirical*

kernel matrix

High-frequency Features: $$K^{(d)}_{>r}(\boldsymbol{X},\boldsymbol{X}) = \hat{h}^2_{>r}\boldsymbol{I}_m + \Delta^{(d)}_{>r}, \tag{114}$$

Low-frequency Features: $$\frac{1}{m}Z_<(\boldsymbol{X})^\top Z_<(\boldsymbol{X}) = \boldsymbol{I}_{N^{(d)}_<} + \Delta^{(d)}_{<r} \tag{115}$$

Critical-frequency Features: $$\textit{the empirical measure of } \frac{1}{m}Z_r(\boldsymbol{X})^\top Z_r(\boldsymbol{X}) \to \mu_\alpha \tag{116}$$

Low-and-critical-frequency Features: $$\textit{the empirical measure of } \frac{1}{m}Z_{\le r}(\boldsymbol{X})^\top Z_{\le r}(\boldsymbol{X}) \to \mu_\alpha \tag{117}$$

Let $0 \le \lambda = O(1)$ be the regularization and $\gamma = \lambda + \hat{h}^2_{>r}$ be the effective regularization. To ease the notations, denote

$$\overline{\boldsymbol{Z}}_< = \frac{1}{\sqrt{m}}Z_{<r}(\boldsymbol{X}) \qquad \overline{\boldsymbol{Z}}_\le = \frac{1}{\sqrt{m}}Z_{\le r}(\boldsymbol{X}) \qquad \overline{\boldsymbol{Z}}_> = \frac{1}{\sqrt{m}}Z_{>r}(\boldsymbol{X}) \qquad \overline{\boldsymbol{Z}}_r = \frac{1}{\sqrt{m}}Z_r(\boldsymbol{X})$$
$$\overline{\boldsymbol{\Lambda}}_< = \gamma^{-1}m\boldsymbol{\Lambda}_< \qquad \overline{\boldsymbol{\Lambda}}_\le = \gamma^{-1}m\boldsymbol{\Lambda}_\le \qquad \overline{\boldsymbol{\Lambda}}_> = \gamma^{-1}m\boldsymbol{\Lambda}_> \qquad \overline{\boldsymbol{\Lambda}}_r = \gamma^{-1}m\boldsymbol{\Lambda}_r$$

Clearly, Corollary 3 and the assumption on the spectra imply that in probability as $d \to \infty$,

$$\|\overline{\boldsymbol{Z}}_<\|_{\rm op} + \|\overline{\boldsymbol{Z}}_\le\|_{\rm op} + \|\overline{\boldsymbol{Z}}_r\|_{\rm op} + \|\overline{\boldsymbol{\Lambda}}_<\|_{\rm op} + \|\overline{\boldsymbol{\Lambda}}_\le\|_{\rm op} + \|\overline{\boldsymbol{\Lambda}}_r\|_{\rm op} \lesssim 1 \tag{118}$$

Then we can write $K^{(d)}$ and $K^{(d)}_\lambda$ as

$$K^{(d)}_\lambda(\boldsymbol{X},\boldsymbol{X}) \equiv K^{(d)}(\boldsymbol{X},\boldsymbol{X}) + \lambda\boldsymbol{I}_m = \gamma(\overline{\boldsymbol{Z}}_\le\overline{\boldsymbol{\Lambda}}_\le\overline{\boldsymbol{Z}}_\le^\top + \boldsymbol{I}_m) + \Delta^{(d)}_{>r} \equiv \boldsymbol{K} + \Delta^{(d)}_{>r} \tag{119}$$

where

$$\boldsymbol{K} = \gamma(\overline{\boldsymbol{Z}}_\le\overline{\boldsymbol{\Lambda}}_\le\overline{\boldsymbol{Z}}_\le^\top + \boldsymbol{I}_m) \tag{120}$$

Then by Sherman–Morrison–Woodbury formula

$$\boldsymbol{K}^{-1} = \gamma^{-1}\left(\boldsymbol{I}_m - \overline{\boldsymbol{Z}}_\le\left(\overline{\boldsymbol{\Lambda}}_\le^{-1} + \overline{\boldsymbol{Z}}_\le^\top\overline{\boldsymbol{Z}}_\le\right)^{-1}\overline{\boldsymbol{Z}}_\le^\top\right) = \gamma^{-1}\left(\boldsymbol{I}_m - \overline{\boldsymbol{Z}}_\le\overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{Z}}_\le^\top\right) \tag{121}$$

where

$$\overline{\boldsymbol{D}} = \overline{\boldsymbol{\Lambda}}_\le^{-1} + \overline{\boldsymbol{Z}}_\le^\top\overline{\boldsymbol{Z}}_\le\,. \tag{122}$$

The matrix $\overline{\boldsymbol{D}}$ plays a critical role in the remaining analysis. We have the following control regarding its eigenvalues, which says the eigenvalues of $\overline{\boldsymbol{D}}$ are away from 0 and ∞

Lemma 3. *There are constants $0 < \lambda_1 < \lambda_2$ independent of d such that, in probability as $d \to \infty$,*

$$\lambda_1\boldsymbol{I} \prec \overline{\boldsymbol{D}} \prec \lambda_2\boldsymbol{I} \tag{123}$$

We will prove the lemma later in Sec.C.6.

Note that

$$K^{(d)}_\lambda(\boldsymbol{X},\boldsymbol{X})^{-1} = (K^{(d)}_\lambda(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{K})\boldsymbol{K}^{-1} = (\boldsymbol{I}_m + \boldsymbol{K}^{-1}\Delta^{(d)}_{>r})^{-1}\boldsymbol{K}^{-1} = (\boldsymbol{I}_m + \Delta'_d)\boldsymbol{K}^{-1} \tag{124}$$

where $\|\Delta'_d\|_{\rm op} \to 0$ in probability since $\|\boldsymbol{K}^{-1}\|_{\rm op} \le \gamma^{-1}$ and $\|\Delta^{(d)}_{>r}\|_{\rm op} \to 0$ in probability.

Similarly, we can write

$$M(\boldsymbol{X},\boldsymbol{X}) \equiv \mathbb{E}K(\boldsymbol{X},x)K(\boldsymbol{X},x)^\top \tag{125}$$
$$= \sum_{1\le k\le r}\sum_{n\in[E_k]}(\sigma^{(d)}_{kn})^4 Z_k(\boldsymbol{X})Z_k(\boldsymbol{X})^\top + \sum_{k>r}\sum_{n\in[E_k]}(\sigma^{(d)}_{kn})^4 Z_{kn}(\boldsymbol{X})Z_{kn}(\boldsymbol{X})^\top \tag{126}$$
$$= m^{-1}\gamma^2\overline{\boldsymbol{Z}}_\le\overline{\boldsymbol{\Lambda}}^2_\le\overline{\boldsymbol{Z}}_\le^\top + m^{-1}\Delta''_d \tag{127}$$

where

$$\Delta_d'' \equiv m \sum_{k>r} \sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^4 Z_{kn}(\boldsymbol{X}) Z_{kn}(\boldsymbol{X})^\top = m \sum_{k>r} \sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^4 N_{kn}^{(d)} (\boldsymbol{I}_m + \Delta_{kn}^{(d)}) \tag{128}$$

We have $\|\Delta_d''\|_{\mathrm{op}} \to 0$ in probability as $d \to \infty$, since

$$m \sum_{k>r} \sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^4 N_{kn}^{(d)} \sim m \sum_{k>r} d^{-s_k} \sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^2 N_{kn}^{(d)} \tag{129}$$

$$\lesssim d^{s_r} d^{-s_{r+1}} \sum_{k>r} \sum_{n\in[E_k]} (\sigma_{kn}^{(d)})^2 N_{kn}^{(d)} \tag{130}$$

$$\leq d^{s_r} d^{-s_{r+1}} \sum_{k>r} \sum_{n\in[E_k]} \hat{h}_{kn}^2 \lesssim d^{-\delta_0} \to 0 \tag{131}$$

Finally, the above estimates imply

$$\boldsymbol{H} \equiv K_\gamma^{(d)}(\boldsymbol{X},\boldsymbol{X})^{-1} M(\boldsymbol{X},\boldsymbol{X}) K_\gamma^{(d)}(\boldsymbol{X},\boldsymbol{X})^{-1} \tag{132}$$

$$= (\boldsymbol{I}_m + \Delta_d') \boldsymbol{K}^{-1} M(\boldsymbol{X},\boldsymbol{X}) \boldsymbol{K}^{-1} (\boldsymbol{I}_m + \Delta_d') \tag{133}$$

$$= m^{-1} (\boldsymbol{I}_m + \Delta_d') \boldsymbol{K}^{-1} (\gamma^2 \overline{\boldsymbol{Z}}_\leq \overline{\boldsymbol{\Lambda}}_\leq^2 \overline{\boldsymbol{Z}}_\leq^\top + \Delta_d'') \boldsymbol{K}^{-1} (\boldsymbol{I}_m + \Delta_d') \tag{134}$$

$$= m^{-1} \left(\overline{\boldsymbol{Z}}_\leq \left(\overline{\boldsymbol{\Lambda}}_\leq^{-1} + \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \right)^{-2} \overline{\boldsymbol{Z}}_\leq^T + \Delta_d''' \right) \tag{135}$$

$$= m^{-1} \left(\overline{\boldsymbol{Z}}_\leq \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^T + \Delta_d''' \right) \tag{136}$$

with the error term $\|\Delta_d'''\|_{\mathrm{op}} \to 0$ in probability.

C.3 Reduction I: Reducing the MSE to Traces

We will repeatedly use the following simple results.

Lemma 4. *Let $\boldsymbol{u}$ be a random vector with $\mathbb{E}\boldsymbol{u} = \boldsymbol{0}$ and $\mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top = \sigma^2 \boldsymbol{I}_k$. Then for any $k \times k$ deterministic matrix $\boldsymbol{A}$,*

$$\mathbb{E}_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u} = \sigma^2 \operatorname{Tr}(\boldsymbol{A}) \tag{137}$$

$$\mathbb{E}_{\boldsymbol{u}} \|\boldsymbol{A}\boldsymbol{u}\|_2^2 = \sigma^2 \operatorname{Tr}(\boldsymbol{A}^\top \boldsymbol{A}) \,. \tag{138}$$

Next, we compute the loss by decomposing it as follows. Recall that the observed labels is $f(\boldsymbol{X}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is the iid noise term, which is centered and has variance σ_ϵ^2. Thus the average test error is

$$\begin{aligned}
&\operatorname{Err}(\boldsymbol{X}; \lambda, \boldsymbol{F}, \boldsymbol{h}) \\
=&\mathbb{E}_{\boldsymbol{f},\boldsymbol{\epsilon},\boldsymbol{x}} \left| f(\boldsymbol{x}) - K^{(d)}(\boldsymbol{x},\boldsymbol{X}) K_\gamma^{(d)}(\boldsymbol{X},\boldsymbol{X})^{-1} (f(\boldsymbol{X}) + \boldsymbol{\epsilon}) \right|^2 \\
=&\mathbb{E}_{\boldsymbol{f}} \mathbb{E}_{\boldsymbol{x}} f^2(\boldsymbol{x}) - 2\mathbb{E}_{\boldsymbol{f}} \mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x}) K^{(d)}(\boldsymbol{x},\boldsymbol{X}) K_\gamma^{(d)}(\boldsymbol{X},\boldsymbol{X})^{-1} f(\boldsymbol{X}) + \mathbb{E}_{\boldsymbol{f}} f^\top(\boldsymbol{X}) \boldsymbol{H} f(\boldsymbol{X}) + \sigma_\epsilon^2 \operatorname{Tr}(\boldsymbol{H}) \\
\equiv&\mathbb{E}_{\boldsymbol{f}} I_1 + \mathbb{E}_{\boldsymbol{f}} I_2 + \mathbb{E}_{\boldsymbol{f}} I_3 + I_4.
\end{aligned}$$

Here

$$I_1 = \mathbb{E}_{\boldsymbol{x}} f^2(\boldsymbol{x}) \tag{139}$$

$$I_2 = -2\mathbb{E}_{\boldsymbol{f}} \mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x}) K^{(d)}(\boldsymbol{x},\boldsymbol{X}) K_\gamma^{(d)}(\boldsymbol{X},\boldsymbol{X})^{-1} f(\boldsymbol{X}) \tag{140}$$

$$I_3 = \mathbb{E}_{\boldsymbol{f}} f^\top(\boldsymbol{X}) \boldsymbol{H} f(\boldsymbol{X}) \tag{141}$$

$$I_4 = \sigma_\epsilon^2 \operatorname{Tr}(\boldsymbol{H}) \tag{142}$$

We estimate each I_i individually.

Estimate I_1. We simply keep it unchanged at the moment.

Estimate I_2. Note that

$$\mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x}) K^{(d)}(\boldsymbol{x}, \boldsymbol{X}) = \sum_k \sum_{n,l} \hat{f}_{knl} (\sigma_{kn}^{(d)})^2 \phi_{knl}^{(d)}(\boldsymbol{X})^\top = \sqrt{m} \hat{\boldsymbol{f}}_\leq^\top \boldsymbol{\Lambda}_\leq \overline{\boldsymbol{Z}}_\leq^\top + \sum_{k>r} \sqrt{m} \hat{\boldsymbol{f}}_k^\top \boldsymbol{\Lambda}_k \overline{\boldsymbol{Z}}_k^\top \tag{143}$$

$$= \gamma \frac{1}{\sqrt{m}} \hat{\boldsymbol{f}}_\leq^\top \overline{\boldsymbol{\Lambda}}_\leq \overline{\boldsymbol{Z}}_\leq^\top + \sum_{k>r} \gamma \frac{1}{\sqrt{m}} \hat{\boldsymbol{f}}_k^\top \overline{\boldsymbol{\Lambda}}_k \overline{\boldsymbol{Z}}_k^\top \tag{144}$$

where $\hat{\boldsymbol{f}}_\leq$ is the column vector with elements $\{\hat{f}_{knl}\}_{k \leq r}$ and $\hat{\boldsymbol{f}}_<$, $\hat{\boldsymbol{f}}_>$, $\hat{\boldsymbol{f}}_k$, etc. are defined similarly. In addition,

$$f(\boldsymbol{X}) = \sqrt{m} \overline{\boldsymbol{Z}}_\leq \hat{\boldsymbol{f}}_\leq + \sqrt{m} \sum_{k>r} \overline{\boldsymbol{Z}}_k \hat{\boldsymbol{f}}_k \tag{145}$$

We then use the fact that $\boldsymbol{f}_>$ is centered to eliminate all cross terms between $\hat{\boldsymbol{f}}_\leq$ and $\hat{\boldsymbol{f}}_>$. Denote $\mathbb{E}_>$, $\mathbb{E}_k$ and $\mathbb{E}_{kn}$ the expectation operator over $\hat{\boldsymbol{f}}_>$, over $\hat{\boldsymbol{f}}_k$ and over $\hat{\boldsymbol{f}}_{kn}$ resp. Under this notation, $\mathbb{E}_{\boldsymbol{f}} = \mathbb{E}_r \mathbb{E}_>$. Using Eq. (124) and Eq. (125),

$$\mathbb{E}_> I_2 = -2 \hat{\boldsymbol{f}}_\leq^\top \overline{\boldsymbol{D}}^{-1} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \hat{\boldsymbol{f}}_\leq - 2\gamma \mathbb{E}_> \sum_{k>r} \hat{\boldsymbol{f}}_k^\top \quad \left(\overline{\boldsymbol{\Lambda}}_k \overline{\boldsymbol{Z}}_k^\top K_\gamma^{(d)}(\boldsymbol{X}, \boldsymbol{X})^{-1} \overline{\boldsymbol{Z}}_k \right) \hat{\boldsymbol{f}}_k + \Delta_d \tag{146}$$

for some $\Delta_d \to 0$ in probability. The second term goes to zero since, for each $k > r$ and $n \in E_k$

$$\begin{aligned}
&\mathbb{E}_{kn} \hat{\boldsymbol{f}}_{kn}^\top \left(\overline{\boldsymbol{\Lambda}}_{kn} \overline{\boldsymbol{Z}}_{kn}^\top K_\gamma^{(d)}(\boldsymbol{X}, \boldsymbol{X})^{-1} \overline{\boldsymbol{Z}}_{kn} \right) \hat{\boldsymbol{f}}_{kn} \\
=& \hat{F}_{kn}^2 / N_{kn}^{(d)} \operatorname{Tr} \overline{\boldsymbol{\Lambda}}_{kn} \left(\overline{\boldsymbol{Z}}_{kn}^\top K_\gamma^{(d)}(\boldsymbol{X}, \boldsymbol{X})^{-1} \overline{\boldsymbol{Z}}_{kn} \right) \\
=& \hat{F}_{kn}^2 / N_{kn}^{(d)} \gamma^{-1} m (\sigma_{kn}^{(d)})^2 \operatorname{Tr} \left(\overline{\boldsymbol{Z}}_{kn}^\top K_\gamma^{(d)}(\boldsymbol{X}, \boldsymbol{X})^{-1} \overline{\boldsymbol{Z}}_{kn} \right) \\
\leq& \gamma^{-1} \hat{F}_{kn}^2 / N_{kn}^{(d)} m (\sigma_{kn}^{(d)})^2 \| K_\gamma^{(d)}(\boldsymbol{X}, \boldsymbol{X})^{-1} \|_{\mathrm{op}} \operatorname{Tr} \left(\overline{\boldsymbol{Z}}_{kn} \overline{\boldsymbol{Z}}_{kn}^\top \right) \\
\lesssim& \hat{F}_{kn}^2 / N_{kn}^{(d)} m (\sigma_{kn}^{(d)})^2 \operatorname{Tr}(\frac{1}{m} N_{kn}^{(d)} \frac{1}{N_{kn}^{(d)}} \boldsymbol{Z}_{kn} \boldsymbol{Z}_{kn}^\top) \\
=& \hat{F}_{kn}^2 m (\sigma_{kn}^{(d)})^2 \frac{1}{m} \operatorname{Tr}(\boldsymbol{I}_m + \Delta_{kn}^{(d)}) \\
\lesssim& m (\sigma_{kn}^{(d)})^2 (1 + \| \Delta_{kn}^{(d)} \|_{\mathrm{op}}) \\
=& m / N_{kn}^{(d)} \hat{h}_{kn}^2 (1 + \| \Delta_{kn}^{(d)} \|_{\mathrm{op}}) \\
\sim& d^{s_r - s_k} \hat{h}_{kn}^2 (1 + \| \Delta_{kn}^{(d)} \|_{\mathrm{op}})
\end{aligned}$$

Clearly, the sum over $k > r$ and $n \in [E_k]$ of the above is bounded by $\lesssim d^{-\delta_0}$ in probability. Thus

$$\mathbb{E}_> I_2 = -2 \left(\hat{\boldsymbol{f}}_\leq^\top \overline{\boldsymbol{D}}^{-1} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \hat{\boldsymbol{f}}_\leq \right) + \Delta_d \tag{147}$$

Estimate I_3. Again, we use the fact that cross terms have mean zero

$$\mathbb{E}_> I_3 = f^\top(\boldsymbol{X}) \boldsymbol{H} f(\boldsymbol{X}) = m \left(\hat{\boldsymbol{f}}_\leq^\top \overline{\boldsymbol{Z}}_\leq^\top \boldsymbol{H} \overline{\boldsymbol{Z}}_\leq \hat{\boldsymbol{f}}_\leq + \sum_{k>r} \mathbb{E}_k \hat{\boldsymbol{f}}_k^\top \overline{\boldsymbol{Z}}_k^\top \boldsymbol{H} \overline{\boldsymbol{Z}}_k \hat{\boldsymbol{f}}_k \right) \tag{148}$$

$$= \hat{\boldsymbol{f}}_\leq^\top \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \hat{\boldsymbol{f}}_\leq + \sum_{k>r} \mathbb{E}_k \hat{\boldsymbol{f}}_k^\top \overline{\boldsymbol{Z}}_k^\top \overline{\boldsymbol{Z}}_\leq \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_k \hat{\boldsymbol{f}}_k + \Delta_d \tag{149}$$

$$\equiv I_{3,1} + I_{3,2} + \Delta_d \tag{150}$$

Note that

$$I_{3,2} = \sum_{k>r}\sum_{n} \mathbb{E}_{kn} \operatorname{Tr} \overline{\boldsymbol{Z}}_{\leq} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \hat{\boldsymbol{f}}_{kn} \boldsymbol{f}_{kn}^{\top} \overline{\boldsymbol{Z}}_{kn} \overline{\boldsymbol{Z}}_{kn}^{\top} \tag{151}$$

$$= \sum_{k>r}\sum_{n} \hat{F}_{kn}^2 / N_k^{(d)} \operatorname{Tr} \overline{\boldsymbol{Z}}_{\leq} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{kn} \overline{\boldsymbol{Z}}_{kn}^{\top} \tag{152}$$

$$= \sum_{k>r}\sum_{n} \hat{F}_{kn}^2 / m \operatorname{Tr} \overline{\boldsymbol{Z}}_{\leq} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} (\boldsymbol{I}_m + \Delta_{kn}^{(d)}) \tag{153}$$

$$= m^{-1} \sum_{k>r}\sum_{n} \hat{F}_{kn}^2 \left(\operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \right) (1 + \Delta_d) \tag{154}$$

$$= m^{-1} \hat{F}_{>r}^2 \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} + \Delta_d' \tag{155}$$

where $\Delta_d' \to 0$ in probability since $\|\boldsymbol{D}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq}\|_{\mathrm{op}} \lesssim 1$ in probability.

Estimate I_4. The I_4 is similar to $I_{3,2}$ above and we have

$$I_4 = m^{-1} \sigma_\epsilon^2 \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} + \Delta_d \tag{156}$$

All Together. Combining all terms we have

$$\mathbb{E}_{>} I_1 + I_2 + I_3 + I_4 \tag{157}$$

$$= \left\| \left(\boldsymbol{I} - \overline{\boldsymbol{D}}^{-1} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \right) \hat{\boldsymbol{f}}_{\leq} \right\|_2^2 + \left(1 + m^{-1} \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \right) \hat{F}_{>r}^2 + m^{-1} \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \sigma_\epsilon^2 + \Delta_d \tag{158}$$

$$= \left\| \overline{\boldsymbol{D}}^{-1} \overline{\boldsymbol{\Lambda}}_{\leq}^{-1} \hat{\boldsymbol{f}}_{\leq} \right\|_2^2 + \left(1 + m^{-1} \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \right) \hat{F}_{>r}^2 + m^{-1} \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \sigma_\epsilon^2 + \Delta_d \tag{159}$$

$$\equiv T_1 + T_2 + T_3 + \Delta_d \tag{160}$$

where

$$T_1 = \left\| \overline{\boldsymbol{D}}^{-1} \overline{\boldsymbol{\Lambda}}_{\leq}^{-1} \hat{\boldsymbol{f}}_{\leq} \right\|_2^2 \tag{161}$$

$$T_2 = \left(1 + m^{-1} \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \right) \hat{F}_{>r}^2 \tag{162}$$

$$T_3 = m^{-1} \operatorname{Tr} \overline{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_{\leq}^{\top} \overline{\boldsymbol{Z}}_{\leq} \sigma_\epsilon^2 \tag{163}$$

As such, it remains to handle

$$\mathbb{E}_r (T_1 + T_2 + T_3) = \mathbb{E}_r T_1 + T_2 + T_3, . \tag{164}$$

C.4 Reduction II: Reducing Traces to Integrals

Recall that $\overline{\boldsymbol{\Lambda}}_{\leq}$ is a diagonal matrix with elements $\gamma^{-1} m (\sigma_{kn}^{(d)})^2$, whose multiplicity is $N_{kn}^{(d)}$ for $k \leq r$ and $n \in [E_k]$. When $k = r$, $\gamma^{-1} m (\sigma_{rn}^{(d)})^2 \sim 1$, otherwise $\gamma^{-1} m (\sigma_{kn}^{(d)})^2 \sim d^{s_r - s_k}$. For convenience, we let

$$N_< \equiv N_<^{(d)}, \quad N_= \equiv N_r^{(d)}, \quad N_\leq \equiv N_\leq^{(d)} \tag{165}$$

Therefore,

$$\overline{\boldsymbol{\Lambda}}_{\leq}^{-1} = \begin{bmatrix} \Delta & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} \tag{166}$$

where Δ is an $N_< \times N_<$ diagonal matrix whose entries are $\gamma m^{-1} (\sigma_{kn}^{(d)})^{-2} \sim d^{-(s_r - s_k)}$, and $\boldsymbol{R}$ is a $N_= \times N_=$ diagonal matrix whose entries are $\gamma m^{-1} (\sigma_{rn}^{(d)})^{-2} \sim 1$. As such, we claim that we

can replace $\hat{\boldsymbol{f}}_{\leq}$ by $[\mathbf{0}, \hat{\boldsymbol{f}}_{=}]$ in estimating T_1 and replace the Δ in $\overline{\boldsymbol{\Lambda}}_{\leq}^{-1}$ by any $\rho \boldsymbol{I}_{N_<}$ for any finite non-negative constant ρ in estimating T_2. The first claim is obvious as, by Lemma 3,

$$\|\overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{\Lambda}}_{\leq}^{-1}[\boldsymbol{f}_<, \mathbf{0}]^\top\|_2 \leq \|\overline{\boldsymbol{D}}^{-1}\|_{\mathrm{op}}\|\overline{\boldsymbol{\Lambda}}_{\leq}^{-1}[\boldsymbol{f}_<, \mathbf{0}]^\top\|_2 \lesssim \lambda_1^{-1}(m\sigma_{r-1}^2\gamma^{-1})\|\boldsymbol{f}_<\|_2 \lesssim d^{-(s_r - s_k)} \to 0 \tag{167}$$

To prove the second claim regarding estimating T_2, denote

$$\tilde{\boldsymbol{D}} = \begin{bmatrix} \rho \boldsymbol{I}_{N_<} & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} + \overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq} \tag{168}$$

We claim that, in probability,

$$m^{-1}\operatorname{Tr}\left((\overline{\boldsymbol{D}}^{-2} - \tilde{\boldsymbol{D}}^{-2})\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\right) \tag{169}$$

$$= m^{-1}\operatorname{Tr}\left((\overline{\boldsymbol{D}}^{-1}(\overline{\boldsymbol{D}}^{-1} - \tilde{\boldsymbol{D}}^{-1}) + (\overline{\boldsymbol{D}}^{-1} - \tilde{\boldsymbol{D}}^{-1})\tilde{\boldsymbol{D}}^{-1})\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\right) \to 0 \tag{170}$$

We only bound the first term as the second term can be handled similarly.

$$m^{-1}\operatorname{Tr}\left(\overline{\boldsymbol{D}}^{-1}(\overline{\boldsymbol{D}}^{-1} - \tilde{\boldsymbol{D}}^{-1})\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\right) \tag{171}$$

$$= m^{-1}\operatorname{Tr}\left((\overline{\boldsymbol{D}}^{-1} - \tilde{\boldsymbol{D}}^{-1})\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\overline{\boldsymbol{D}}^{-1}\right) \tag{172}$$

$$= m^{-1}\operatorname{Tr}\left(\overline{\boldsymbol{D}}^{-1}(\tilde{\boldsymbol{D}} - \overline{\boldsymbol{D}})\tilde{\boldsymbol{D}}^{-1}\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\overline{\boldsymbol{D}}^{-1}\right) \tag{173}$$

$$= m^{-1}\operatorname{Tr}\left((\tilde{\boldsymbol{D}} - \overline{\boldsymbol{D}})\tilde{\boldsymbol{D}}^{-1}\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\overline{\boldsymbol{D}}^{-2}\right) \tag{174}$$

Then we use the facts that (1) the upper right $N_< \times N_<$ block matrix of $(\tilde{\boldsymbol{D}} - \overline{\boldsymbol{D}})$ is a diagonal matrix whose entries are in $[0, 1]$ and the three remaining block matrices are zeros, and (2) all entries in $\tilde{\boldsymbol{D}}^{-1}\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\overline{\boldsymbol{D}}^{-2}$ are bounded above by a constant (each matrix in $\tilde{\boldsymbol{D}}^{-1}\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\overline{\boldsymbol{D}}^{-2}$ has bounded operator norm[5]) to conclude that

$$\left|m^{-1}\operatorname{Tr}\left(\overline{\boldsymbol{D}}^{-1}(\overline{\boldsymbol{D}}^{-1} - \tilde{\boldsymbol{D}}^{-1})\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\right)\right| \lesssim m^{-1}N_< \tag{175}$$

Thus

$$T_2 = \left(1 + m^{-1}\operatorname{Tr}\tilde{\boldsymbol{D}}^{-2}\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}\right)\hat{F}_{>r}^2 + \Delta_d \tag{176}$$

which will be handled later.

It remains to handle T_1. We make two steps of reductions in estimating $\mathbb{E}_r T_1$. The first one is to replace $\overline{\boldsymbol{\Lambda}}_{\leq}^{-1}$ by

$$\tilde{\boldsymbol{\Lambda}}_{\leq}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} \tag{177}$$

and the second one is to replace $\overline{\boldsymbol{\Lambda}}_{\leq}^{-1} + \overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}$ by

$$\boldsymbol{W} \equiv \begin{bmatrix} \boldsymbol{I}_{N_\leq} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{I}_{N_<} & \overline{\boldsymbol{Z}}_{<}^\top \overline{\boldsymbol{Z}}_{=} \\ \overline{\boldsymbol{Z}}_{=}^\top \overline{\boldsymbol{Z}}_{<} & \overline{\boldsymbol{Z}}_{=}^\top \overline{\boldsymbol{Z}}_{=} + \boldsymbol{R} \end{bmatrix} \tag{178}$$

Here we have applied

$$\overline{\boldsymbol{Z}}_{<}^\top \overline{\boldsymbol{Z}}_{<} = \boldsymbol{I}_{N_<} + \Delta_d \tag{179}$$

The reason we could do so is exactly the same as we replaced $\overline{\boldsymbol{D}}$ by $\tilde{\boldsymbol{D}}$ above as we only perturb the entries in the upper $N_< \times N_<$ block by $O(1)$.

[5]Recall that $\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}$ follows the Marchenko-Pastur distribution.

Note that $\boldsymbol{W}$ is symmetric and is also strictly positive definite, i.e. the minimal eigenvalue of $\boldsymbol{W}$ is $\gtrsim 1$; see the proof in Sec.C.6. Thus by the Schur complement,

$$\mathbb{E}_r T_1 = \left\| \begin{bmatrix} \boldsymbol{I}_{N_<} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{R} \end{bmatrix} \begin{bmatrix} 0 \\ \hat{\boldsymbol{f}}_= \end{bmatrix} \right\|_2^2 + \Delta_d \tag{180}$$

$$= \mathbb{E}_r \left\| \begin{bmatrix} -\boldsymbol{B}(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R} \hat{\boldsymbol{f}}_= \\ (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R} \hat{\boldsymbol{f}}_= \end{bmatrix} \right\|_2^2 + \Delta_d \tag{181}$$

By the fact that $\hat{\boldsymbol{f}}_=$ is mean zero and isotropic, we have the above equal to

$$\mathbb{E}_r T_1 = \operatorname{Tr}\left(\boldsymbol{R}(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{B} (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R} + \boldsymbol{R}(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} \boldsymbol{R} \right) \hat{F}_r^2 / N_= + \Delta_d \tag{182}$$

$$= \operatorname{Tr}(\boldsymbol{R}\boldsymbol{C}^{-2}\boldsymbol{R}) \hat{F}_r^2 / N_= + \Delta_d' + \Delta_d \tag{183}$$

where

$$\Delta_d' = \operatorname{Tr}\left(\boldsymbol{R}(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{B} (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R} \right) / N_= + \tag{184}$$

$$\operatorname{Tr}\left(\boldsymbol{R}((\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} - \boldsymbol{C}^{-2}) \boldsymbol{R} \right) / N_= . \tag{185}$$

We claim that $\Delta_d' \to 0$ in probability. For the first term, we have

$$\operatorname{Tr}\left(\boldsymbol{R}(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{B} (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R} \right) / N_= \tag{186}$$

$$= \operatorname{Tr}\left((\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R}^2 (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{B} \right) / N_= \tag{187}$$

$$= \|(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{R}^2 (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1}\|_{\mathrm{op}} \operatorname{Tr}\left(\boldsymbol{B}^\top \boldsymbol{B} \right) / N_= \tag{188}$$

$$\leq \|(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1}\|_{\mathrm{op}} \|\boldsymbol{R}^2\|_{\mathrm{op}} \|(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1}\|_{\mathrm{op}} \operatorname{Tr}\left(\boldsymbol{B}^\top \boldsymbol{B} \right) / N_= \tag{189}$$

$$\lesssim \operatorname{Tr}\left(\boldsymbol{B}^\top \boldsymbol{B} \right) / N_= \sim N_< / N_= \to 0 \tag{190}$$

in probability as $d \to \infty$. We have used

$$\|\boldsymbol{R}\|_{\mathrm{op}} \lesssim 1 \tag{191}$$

$$\|(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1}\|_{\mathrm{op}} \leq \|\boldsymbol{W}^{-1}\|_{\mathrm{op}} \lesssim 1 \tag{192}$$

$$\frac{1}{N_=} \operatorname{Tr}\left(\boldsymbol{B}^\top \boldsymbol{B} \right) \sim N_< / N_= . \tag{193}$$

The last one holds because $\boldsymbol{B}^\top \boldsymbol{B}$ is a rank $N_<$ matrix with operator norm $\lesssim 1$. Note that this also implies that $\boldsymbol{B}^\top \boldsymbol{B}$ has at most $N_<$ many non-zero singular values, which is upper bounded by $\lesssim 1$. Using Von Neumann's trace inequalities, for any matrix $\boldsymbol{A}$, we have

$$|\operatorname{Tr} \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{B}| \leq \sum_j \sigma_j(\boldsymbol{A}) \sigma_j(\boldsymbol{B}^\top \boldsymbol{B}) \lesssim N_< \|\boldsymbol{A}\|_{\mathrm{op}} \tag{194}$$

where $\sigma_j(\boldsymbol{A})$ is the j-th (in descending order) singular value of a matrix $\boldsymbol{A}$. Now we proceed to control the second term. Note that

$$(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} - \boldsymbol{C}^{-2} \tag{195}$$

$$= (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} - (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{C}^{-1} + (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{C}^{-1} - \boldsymbol{C}^{-2} \tag{196}$$

$$= (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{C}^{-1} + (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{C}^{-2} \tag{197}$$

As such, by Eq. (194) we have

$$|\operatorname{Tr} \boldsymbol{R}(\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{C}^{-1} \boldsymbol{R}| / N_= \tag{198}$$

$$= |\operatorname{Tr} \boldsymbol{C}^{-1} \boldsymbol{R}^2 (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2} \boldsymbol{B}^\top \boldsymbol{B}| / N_= \tag{199}$$

$$\lesssim N_< / N_= \|\boldsymbol{C}^{-1} \boldsymbol{R}^2 (\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{B})^{-2}\|_{\mathrm{op}} \to 0 . \tag{200}$$

The other term can be bounded similarly. This finishes the proof of $\Delta_d' \to 0$ in probability. To sum up, we have the test error to be

$$\operatorname{Err}(\boldsymbol{X}; \lambda, \boldsymbol{F}, \boldsymbol{h}) = \operatorname{Tr}\left(\boldsymbol{R}^2 \boldsymbol{C}^{-2} \right) / N_= \hat{F}_r^2 + \left(1 + m^{-1} \operatorname{Tr} \tilde{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \right) \hat{F}_{>r}^2 + \tag{201}$$

$$m^{-1} \sigma_\epsilon^2 \operatorname{Tr} \tilde{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq + \Delta_d \tag{202}$$

Generalization Error via Marchenko-Pastur The next step is to reduce the traces to the integral form when $d \to \infty$. That is evaluating the followings as $d \to \infty$,

$$\operatorname{Tr}\left(\boldsymbol{R}^2 \boldsymbol{C}^{-2}\right)/N_= \quad \text{and} \quad m^{-1}\sigma_\epsilon^2 \operatorname{Tr} \tilde{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \tag{203}$$

We begin with the simpler case $E_r = 1$ and then consider $E_r > 1$.

The $E_r = 1$ case. I.e., there is only one eigenspace with eigenvalues $\sim d^{-s_r}$. This is the case for one-hidden layer convolutional kernels and dot-product kernels. In this case, $n = 0$ and

$$\boldsymbol{R} = (\overline{\xi}_r^{(d)})^{-1} \boldsymbol{I}_{N_\leq}, \quad \text{with} \quad \overline{\xi}_r^{(d)} = \gamma^{-1} m (\sigma_{rn}^{(d)})^2 = \frac{m}{N_=} \hat{h}_r^2 \gamma^{-1} \to \overline{\xi}_r = \alpha^{-1} \hat{h}_r^2 \gamma^{-1} \tag{204}$$

Choosing $\rho = (\overline{\xi}_r^{(d)})^{-1}$ and applying Theorem 1, we have when[6] $N_=/m \to \alpha \in (0, \infty)$

$$\overline{\xi}_r^{-2} \operatorname{Tr}\left(\boldsymbol{C}^{-2}\right)/N_= \longrightarrow \int (1 + \overline{\xi}_r t)^{-2} \mu_\alpha(t) dt \tag{205}$$

$$\frac{N_\leq}{m} \frac{1}{N_\leq} \operatorname{Tr} \tilde{\boldsymbol{D}}^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \longrightarrow \alpha \overline{\xi}_r^2 \int t (1 + \overline{\xi}_r t)^{-2} \mu_\alpha(t) dt \tag{206}$$

Therefore,

$$\operatorname{Err}(\boldsymbol{X}; \lambda, \boldsymbol{F}, \boldsymbol{h}) = \left(\hat{F}_r^2 \cdot \int \frac{\mu_\alpha(t)}{(1 + \overline{\xi}_r t)^2} dt + \hat{F}_{>r}^2 \right) + \left(\hat{F}_{>r}^2 + \sigma_\epsilon^2 \right) \cdot \alpha \overline{\xi}_r^2 \int \frac{t \mu_\alpha(t)}{(1 + \overline{\xi}_r t)^2} dt + \Delta_d \tag{207}$$

Both integrals have closed form formulas and they are computed in Sec.C.5.

The $E_r > 1$ Case. Recall that $\boldsymbol{R}$ is a diagonal matrix with entries $\gamma(m(\sigma_{rn}^{(d)})^2)^{-1}$ whose multiplicity is $N_{rn}^{(d)}$. We assume the limiting density exist

$$\gamma(m(\sigma_{rn}^{(d)})^2)^{-1} \to \gamma \alpha \hat{h}_{rn}^{-2} \quad \text{and} \quad N_{rn}^{(d)} / \sum_{n \in [E_k]} N_{rn}^{(d)} \to \tau_{rn} \tag{208}$$

and let $\nu_{\boldsymbol{h}}(r)$ denote this distribution. For convenience, we still $\boldsymbol{R}$ to represent a (sequence of) diagonal matrix with limiting spectral $\nu_{\boldsymbol{h}}(r)$. By our assumptions on $\boldsymbol{h}$, the support of $\nu_{\boldsymbol{h}}(r)$ is bounded away from 0 and ∞. Thus, ignoring vanishing correction term between $\overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq$ and $\overline{\boldsymbol{Z}}_=^\top \overline{\boldsymbol{Z}}_=$, we need to compute the limit of the following

$$\frac{1}{N_\leq} \operatorname{Tr} \boldsymbol{R}^2 (\boldsymbol{R} + \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq)^{-2} = \boldsymbol{R}^{1/2} (1 + \boldsymbol{R}^{-1/2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \boldsymbol{R}^{-1/2})^{-1} \boldsymbol{R}^{-1} (1 + \boldsymbol{R}^{-1/2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq \boldsymbol{R}^{-1/2})^{-1} \boldsymbol{R}^{1/2} \tag{209}$$

$$\frac{1}{N_\leq} \operatorname{Tr} (\boldsymbol{R} + \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq)^{-2} \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq = \frac{1}{N_\leq} \operatorname{Tr} \left((\boldsymbol{R} + \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq)^{-1} - (\boldsymbol{R} + \overline{\boldsymbol{Z}}_\leq^\top \overline{\boldsymbol{Z}}_\leq)^{-2} \boldsymbol{R} \right) \tag{210}$$

.

To evaluate the limit, we may need extra assumptions on the eigenfunctions $\phi_{knl}^{(d)}$ to ensure $\boldsymbol{R}$ and $\boldsymbol{Z}_\leq$ are asymptotically free. Nevertheless, under the freeness assumption, computing self-consistent equations that characterize the asymptotic values of the trace objects in Eq. (209) and Eq. (210) is then straightforward using tools from operator-valued free probability [33]. We do not elaborate on the details here, but refer the reader so many related works for examples of how to apply these tools [14, 1, 2, 38, 39].

C.5 Computing the Integrals.

It remains to compute the above integrals. Note that

$$\overline{\xi}_r^2 \int \frac{t \mu_\alpha(t)}{(1 + \overline{\xi}_r t)^2} dt = \overline{\xi}_r \left(\int \frac{\mu_\alpha(t)}{(1 + \overline{\xi}_r t)} dt - \int \frac{\mu_\alpha(t)}{(1 + \overline{\xi}_r t)^2} dt \right) \tag{211}$$

[6] Note that $N_\leq = N_=(1 + o(1))$

As such we only need to compute, for $k = 1$ and $k = 2$,

$$\zeta_\alpha(\xi; \alpha, k) = \int \frac{\mu_\alpha(t)}{(1+\xi t)^k} dt \tag{212}$$

Note that one only needs $\zeta_\alpha(\xi; \alpha, 1)$ as $\zeta_\alpha(\xi; \alpha, k)$ can be obtained from $\zeta_\alpha(\xi; \alpha, k-1)$ by taking derivative w.r.t. $\bar{\xi}$. Denote $b_\pm = (1 \pm \sqrt{\alpha})^2$ and $\Delta = \alpha_+ - \alpha_-$. Then

$$\mu_\alpha(t) = \left(1 - \frac{1}{\alpha}\right)^+ \delta_0(t) + \frac{\sqrt{(\alpha_+ - t)(t - \alpha_-)}}{2\pi\alpha t} \mathbf{1}_{[\alpha_-, \alpha_+]}(t) dt \tag{213}$$

With $b = (1 + \bar{\xi}_r \alpha_-)/(\bar{\xi}_r(\alpha_+ - \alpha_-))$ and $c = \bar{\xi}_r \alpha_- / (\bar{\xi}_r(\alpha_+ - \alpha_-)) = \alpha_- / (\alpha_+ - \alpha_-)$,

$$\int (1 + \bar{\xi}_r t)^{-k} \mu_\alpha(t) dt \tag{214}$$

$$= \left(1 - \frac{1}{\alpha}\right)^+ + \int_{[\alpha_-, \alpha_+]} (1 + \bar{\xi}_r t)^{-k} \frac{\sqrt{(\alpha_+ - t)(t - \alpha_-)}}{2\pi\alpha t} dt \tag{215}$$

$$= \left(1 - \frac{1}{\alpha}\right)^+ + \frac{1}{2\pi\alpha\bar{\xi}_r} (\bar{\gamma}(\alpha_+ - \alpha_-))^{1-k} \int_0^1 (b+t)^{-k} (c+t)^{-1} \sqrt{t(1-t)} dt \tag{216}$$

Thanks to wolframalpha.com, we have, after doing some algebra,

$$\int_0^1 \frac{\sqrt{t(1-t)}}{(t+b)(t+c)} dt = \pi \left(-1 + \frac{1 + b + c}{\sqrt{bc(1+b)(1+c)}}\right) \tag{217}$$

$$\int_0^1 \frac{\sqrt{t(1-t)}}{(t+b)^2(t+c)} dt = \frac{\pi}{2\sqrt{b^2+b}((b+c+2bc) + 2\sqrt{(b+1)(c+1)bc})} \tag{218}$$

C.6 Proof of Lemma 3.

Note that this lemma is trivial if $\lim_{d\to\infty} m/N_{\le r}^{(d)} = \alpha^{-1} > 1$ as $\overline{\boldsymbol{Z}}_\le^\top \overline{\boldsymbol{Z}}_\le$ follows the Marchenko-Pastur distribution, and the smallest eigenvalues is bounded from below by $\alpha_- = (1 - \sqrt{\alpha})^2$. When $\alpha^{-1} \le 1$, we need to use the regularization term $\overline{\boldsymbol{\Lambda}}_\le^{-1}$. We provide the details below.

Recall that $\overline{\boldsymbol{Z}}_\le^\top = [\overline{\boldsymbol{Z}}_<^\top, \overline{\boldsymbol{Z}}_r^\top]$, where $\overline{\boldsymbol{Z}}_<^\top$ is a $m \times N_{<r}^{(d)}$ matrix consisting of low frequency modes and $\overline{\boldsymbol{Z}}_r^\top$ is a $m \times N_r^{(d)}$ is a matrix consisting of critical frequency modes. We have $N_{<r}^{(d)} \sim d^{s_{r-1}}$, $N_r^{(d)} \sim N_{\le r}^{(d)} \sim d^{s_r}$ and

$$\overline{\boldsymbol{Z}}_<^\top \overline{\boldsymbol{Z}}_< = \boldsymbol{I}_{N_{<r}^{(d)}} + \Delta_d \tag{219}$$

where $\mathbb{E}\|\Delta_d\|_{op} \to 0$ as $d \to \infty$ in probability. Let $\boldsymbol{u} = [\beta_< \boldsymbol{e}_<^\top, \beta_r \boldsymbol{e}_r^\top]^\top$ be a unit vector in $\mathbb{R}^{N_{\le r}^{(d)}}$, where $\boldsymbol{e}_<$ and $\boldsymbol{e}_r$ are unit vectors in $\mathbb{R}^{N_{<r}^{(d)}}$ and $\mathbb{R}^{N_r^{(d)}}$ resp., and $\beta_<^2 + \beta_r^2 = 1$. We want to show that for some $\lambda_1 > 0$

$$\lambda_1 \le \boldsymbol{u}^\top D \boldsymbol{u} = \boldsymbol{u}^\top \overline{\boldsymbol{\Lambda}}_\le^{-1} \boldsymbol{u} + \boldsymbol{u}^\top \overline{\boldsymbol{Z}}_\le^\top \overline{\boldsymbol{Z}}_\le \boldsymbol{u} \tag{220}$$

Note that the entries in $\overline{\boldsymbol{\Lambda}}_\le^{-1}$ corresponding to the critical-frequencies are $m(\sigma_{nr}^{(d)})^2 \sim 1$. Thus there is a constant $c > 0$ such that

$$\boldsymbol{u}^\top \overline{\boldsymbol{\Lambda}}_\le^{-1} \boldsymbol{u} \ge c\beta_r^2 \tag{221}$$

In addition, if $C \equiv \|\overline{\boldsymbol{Z}}_r \boldsymbol{e}_r\|_2$ then $C \le 2\alpha_+$ in probability. Thus by the triangle inequality,

$$\boldsymbol{u}^\top D \boldsymbol{u} \ge c\beta_r^2 + \|\beta_< \overline{\boldsymbol{Z}}_< \boldsymbol{e}_< + \beta_r \overline{\boldsymbol{Z}}_r \boldsymbol{e}_r\|_2^2 \tag{222}$$

$$\ge c\beta_r^2 + (\|\beta_< \overline{\boldsymbol{Z}}_< \boldsymbol{e}_<\|_2 - \|\beta_r \overline{\boldsymbol{Z}}_r \boldsymbol{e}_r\|_2)^2 \tag{223}$$

$$\ge c\beta_r^2 + ((1 - \Delta_d)|\beta_<| - C|\beta_r|)^2 \tag{224}$$

where $\Delta_d \to 0$ in probability. If $C|\beta_r| < \frac{1}{2}|\beta_<|$, the above is greater than $(1/2 - \Delta_d)\beta_<^2 + c\beta_r^2 \gtrsim 1$; otherwise $C|\beta_r| \geq \frac{1}{2}|\beta_<|$ and $\boldsymbol{u}^\top D\boldsymbol{u} \geq c\beta_r^2 \geq c(\frac{1}{2C}\beta_<)^2$ and as a result

$$\boldsymbol{u}^\top D\boldsymbol{u} \geq 2c\beta_r^2/2 \geq (c\beta_r^2 + c(\frac{1}{2C}\beta_<)^2)/2 \geq c\min(1, \frac{1}{2C})^2/2 \gtrsim 1 \tag{225}$$

The other direction is easier as both $\overline{\boldsymbol{\Lambda}}_{\leq}^{-1}$ and $\overline{\boldsymbol{Z}}_{\leq}^\top \overline{\boldsymbol{Z}}_{\leq}$ have operator norms bounded above.

C.7 Proof of Claim 1

The proof is split into two part: the ultra-high frequency parts $k \geq j_0$ and the median-high-frequency part, $r < k < j_0$. The first part is done by a moment-based calculation and the second part is done by matrix concentration [40].

Controlling the Ultra-high-frequency. Recall that

$$\Delta_{kn}^{(d)} \equiv \frac{1}{N_{kn}^{(d)}} Z_{kn}(\boldsymbol{X}) Z_{kn}(\boldsymbol{X})^\top - \boldsymbol{I}_m \,. \tag{226}$$

By **Assumption (4.)**, the diagonals are zero and we have

$$\Delta_{kn}^{(d)} = [Z_{kn}(\boldsymbol{x}_i)^\top Z_{kn}(\boldsymbol{x}_j)/N_{kn}^{(d)}]_{i,j\in[m],i\neq j} \tag{227}$$

Then

$$\begin{aligned}
\mathbb{E}\|\Delta_{kn}^{(d)}\|_{\text{op}}^2 \leq \mathbb{E}\|\Delta_{kn}^{(d)}\|_{\text{F}}^2 =& \mathbb{E}\sum_{i\neq j} |Z_{kn}(\boldsymbol{x}_i)^\top Z_{kn}(\boldsymbol{x}_j)/N_{kn}^{(d)}|^2 && (228)\\
=& \frac{1}{(N_{kn}^{(d)})^2}\mathbb{E}\sum_{i\neq j}\sum_{l,l'} \phi_{knl}^{(d)}(\boldsymbol{x}_i)\phi_{knl}^{(d)}(\boldsymbol{x}_j)\phi_{knl'}^{(d)}(\boldsymbol{x}_i)\phi_{knl'}^{(d)}(\boldsymbol{x}_j) && (229)\\
=& \frac{1}{(N_{kn}^{(d)})^2}\mathbb{E}\sum_{i\neq j}\sum_{l} \phi_{knl}^{(d)}(\boldsymbol{x}_i)^2\phi_{knl}^{(d)}(\boldsymbol{x}_j)^2 && (230)\\
=& \frac{1}{N_{kn}^{(d)}} m(m-1) \leq \frac{1}{N_{kn}^{(d)}} m^2 && (231)
\end{aligned}$$

Recall that E_k grows at most exponentially, i.e. $E_k \leq C^k$ for some constant C. Thus, choosing d large enough such that $Cd^{-\delta_0/4} < 1$ and summing over $k > j_0 \equiv [4s_r/\delta_0 + 4] + 4$,

$$\begin{aligned}
\mathbb{E}\sum_{k>j_0}\sum_{n\in E_k} \|\Delta_{kn}^{(d)}\|_{\text{op}} \lesssim& \sum_{k>j_0} C^k (m^2/N_{kn}^{(d)})^{1/2} && (232)\\
\lesssim& \sum_{k>j_0} C^k d^{-s_k/2+s_r} && (233)\\
\leq& \sum_{k>j_0} C^k d^{-k\delta_0/2+s_r} && (234)\\
\lesssim& \sum_{k>j_0} d^{-k\delta_0/4+s_r} \lesssim d^{-j_0\delta_0/4+s_r} \lesssim d^{-\delta_0} && (235)
\end{aligned}$$

Controlling the Median-high-frequency. It remains to show, for some $\epsilon > 0$

$$\mathbb{E}\sum_{r<k\leq j_0}\sum_{n\in E_k} \|\Delta_{kn}^{(d)}\|_{\text{op}} \lesssim d^{-\epsilon} \,. \tag{236}$$

As there are only finitely many terms in this sum, we only need to show that for each k and n,

$$\mathbb{E}\|\Delta_{kn}^{(d)}\|_{\text{op}} \lesssim d^{-\epsilon} \,.$$

We use the following theorem from Vershynin [40] regarding matrix concentration to prove this claim.

Theorem 6 (Theorem 5.62 Vershynin [40]). *Let $\boldsymbol{A}$ be a $N \times m$ ($N \geq m$) matrix whose columns A_j are independent isotropic random vectors in $\mathbb{R}^N$ with $\|A_j\|_2 = N$ a.s. Let K be defined as*

$$K = \frac{1}{N}\mathbb{E}\max_{j\leq m}\sum_{i\in[m],i\neq j}|A_j^\top A_i|^2 \tag{237}$$

Then

$$\mathbb{E}\|\boldsymbol{A}^\top\boldsymbol{A}/N - \boldsymbol{I}_m\|_{\mathrm{op}} \lesssim \sqrt{K\log(m)/N} \tag{238}$$

We apply this theorem to $\boldsymbol{A} = Z_{kn}(\boldsymbol{X})^\top$. The columns of $Z_{kn}(\boldsymbol{X})^\top$ are $A_j = \phi_{kn}^{(d)}(\boldsymbol{x}_j)$, $j \in [m]$ which are independent. Let $N = N_{kn}^{(d)}$. By Assumption (4.),

$$A_j^\top A_j = \sum_l \phi_{knl}^{(d)}(\boldsymbol{x}_j)^2 = N. \tag{239}$$

We claim that $K \lesssim_{k,q} m^{1+\frac{1}{q}}$ for any $q \geq 1$. Indeed, let

$$B_j = \sum_{i\in[m],i\neq j}|A_j^\top A_i|^2 \tag{240}$$

We then remove the maximal function by paying an $m^{1/q}$ factor

$$K = \frac{1}{N}\mathbb{E}\max_{j\leq m}B_j \leq \frac{1}{N}m^{1/q}|\mathbb{E}B_j^q|^{1/q} = \frac{1}{N}m^{1/q}\left|\mathbb{E}(\sum_{i\in[m],i\neq j}|A_j^\top A_i|^2)^{2q/2}\right|^{1/q} \tag{241}$$

Next we apply the Minkowski inequality to swap the L^{2q}-norm and the l^2-norm,

$$K \leq \frac{1}{N}m^{1/q}\sum_{i\in[m],i\neq j}(\mathbb{E}|A_j^\top A_i|^{2q})^{1/2q\times 2} \leq \frac{1}{N}m^{1/q+1}(\mathbb{E}|A_j^\top A_i|^{2q})^{1/2q\times 2} \leq C_{k,q}^2 m^{1/q+1} \tag{242}$$

if $(\mathbb{E}|A_j^\top A_i|^{2q})^{1/2q\times 2} \leq C_{k,q}^2 N$, which is done by hypercontractivities below. Indeed, for $\boldsymbol{x}_j$ fixed, $Z_{rn}(\boldsymbol{x}_j)^\top Z_{rn}(\boldsymbol{x}_i)$ is a linear combination of $\phi_{knl}^{(d)}$, **Assumption (2.)** gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_i}|A_j^\top A_i|^{2q} =& \mathbb{E}_{\boldsymbol{x}_i}|Z_{rn}(\boldsymbol{x}_j)^\top Z_{rn}(\boldsymbol{x}_i)|^{2q} && (243)\\
\leq& \left(C_{k,q}(\mathbb{E}_{\boldsymbol{x}_i}|Z_{rn}(\boldsymbol{x}_j)^\top Z_{rn}(\boldsymbol{x}_i)|^2)^{1/2}\right)^{2q} && (244)\\
=& C_{k,q}^{2q}N^q && (245)
\end{aligned}$$

where we applied

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_i}|Z_{rn}(\boldsymbol{x}_j)^\top Z_{rn}(\boldsymbol{x}_i)|^2 =& \mathbb{E}_{\boldsymbol{x}_i}|\sum_l \phi_{knl}^{(d)}(\boldsymbol{x}_i)\phi_{knl}^{(d)}(\boldsymbol{x}_j)|^2 && (246)\\
=& \mathbb{E}_{\boldsymbol{x}_i}\sum_{ll'}\phi_{knl}^{(d)}(\boldsymbol{x}_i)\phi_{knl}^{(d)}(\boldsymbol{x}_j)\phi_{knl'}^{(d)}(\boldsymbol{x}_i)\phi_{knl'}^{(d)}(\boldsymbol{x}_j) && (247)\\
=& \sum_l \phi_{knl}^{(d)}(\boldsymbol{x}_j)^2 && (248)\\
=& N && (249)
\end{aligned}$$

Therefore with $\boldsymbol{A} = Z_{kn}(\boldsymbol{X})^\top$, we have

$$\mathbb{E}\|\Delta_{kn}^{(d)}\|_{\mathrm{op}} = \mathbb{E}\|Z_{kn}(\boldsymbol{X})Z_{kn}(\boldsymbol{X})^\top/N_{kn}^{(d)} - \boldsymbol{I}_m\|_{\mathrm{op}} \lesssim_{k,q} \sqrt{m^{1+1/q}\log m/N_{kn}^{(d)}}\,. \tag{250}$$

For each fixed $k > r$, $s_k - s_r \geq \delta_0$, by choosing q sufficiently large (depending on k and δ_0), we have

$$\mathbb{E}\|\Delta_{kn}^{(d)}\|_{\mathrm{op}} \lesssim_k d^{-\delta_0/2}\,. \tag{251}$$

D Additional Plots

To simulate the learning curves for higher-order scalings, e.g. $r = 4$, we must chose d small. As such, we are in a strong finite-size correction regime. In this section, we vary d to visualize the finite-size effect of the predictions. Note that for larger d ($= 60$ here), we can only simulate up to the quadratic scaling. For smaller d ($d = 10$), we observe noticeable finite-size correction. However, the predictions match the simulations quite well. When d become larger $d = 60$, the predicted learning curves match the simulation almost perfectly.

E Further Analysis

E.1 Reducing finite-size Effect.

There are two non-obvious improvements in our results that lead to near perfect agreements between simulations and predictions even for small d. The first one is to use $m = N(d, \leq r)$ to compute r-th peak vs. $m = N(d, r)$ (or $m = d^r/r!$). As it is shown in Fig. 9 (a), using $m = N(d, r)$ as the peak in the theoretical prediction, the prediction is a bit off to the left. The second one is to use the sum over all contributions from all critical scaling $m = N(d, \leq r)$ (i.e., Eq. (21) rather than the contribution from a single critical scaling (i.e., Eq. (18).) As it is shown in Fig. 9 (b), the predictions are a bit smaller than the simulations when using the latter. These two improvements together lead to accurate agreement Fig. 9 (c).

E.2 Choosing the number of peaks by choosing the right regularization.

Recall that the height of the r-th variance term scales like

$$\xi_r(\hat{\boldsymbol{h}}, \lambda, 1)^{1/2} = \left(\frac{\hat{h}_r^2}{\lambda + \hat{h}_{>r}^2} \right)^{1/2}. \tag{252}$$

If $\hat{h}_r^2 \gg \hat{h}_{>r}^2$ and $\lambda \leq \hat{h}_{>r}^2$, then $\xi_r(\hat{\boldsymbol{h}}, \lambda)^{1/2}$ is large, which could lead to a peak at $m = N(d, \leq r)$. To eliminate this peak, we could choose $\lambda \sim \hat{h}_r^2$ which implies $\xi_r(\hat{\boldsymbol{h}}, \lambda, 1) \lesssim 1$. We verify this observation in Fig. 10. When $\lambda = 0$, the unregularized learning curve have 4 peaks. By increasing λ to $1e-7, 1e-5, 1e-3, 1e-1$, the number of peaks are reduced to 3, 2, 1, 0, respectively. A similar result has also been observed in a linear design setting [41]. The similarity between the linear design in [41] and the nonlinear design here shouldn't be surprising, as we prove a "Gaussian equivalence conjecture," which implies that the polynomial scalings are essential "replicas" of linear designs with different scales.

E.3 Natural Data vs. Spherical Data

We compare the spectrum of the NTKs of CIFAR10 associated with three architectures (FCN: fully-connected networks, CNN-VEC: convolutional networks without pooling, and CNN-GAP: convolutional networks with a global average pooling) against the one-layer convolutional kernels with spherical-type of data. Recall that the larger spectral gap between eigenspaces triggers the multiple-descent phenomenon. This phenomenon disappears, and the learning curve becomes monotonic when the spectral gap is small. Fortunately, for CIFAR10, the spectrum of the NTKs are continuous, and the learning curves are monotonic (power-law decay.) As such, there is still a gap between our results/assumptions and natural data.

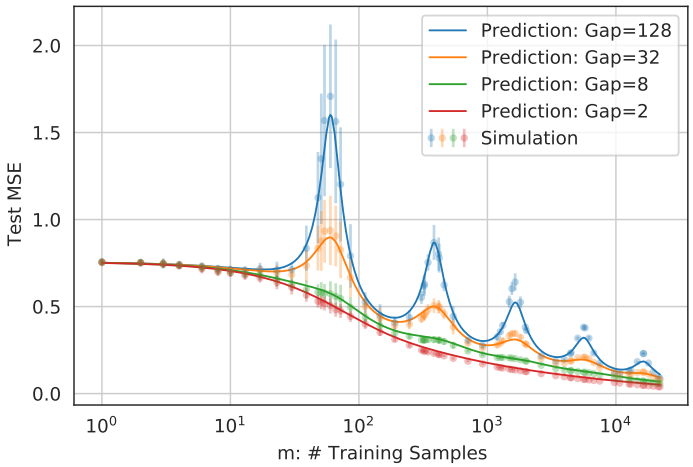


Figure 6: **Tiny** $d = 10$

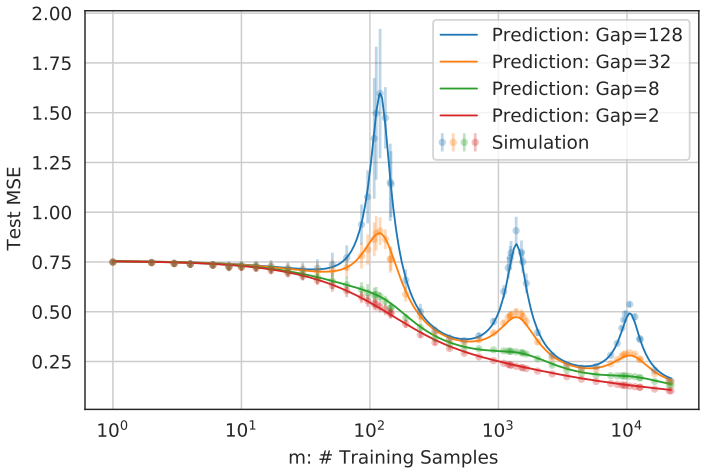


Figure 7: **Small** $d = 20$

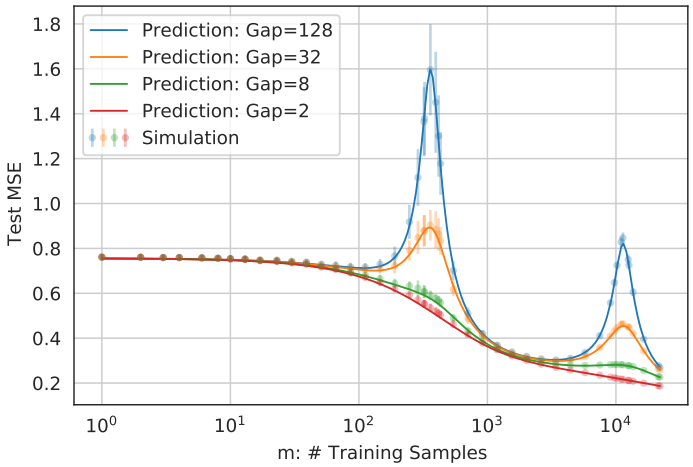
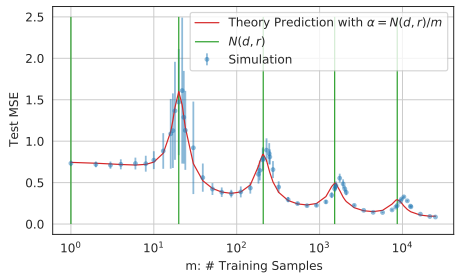
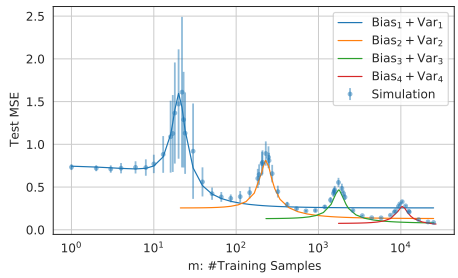


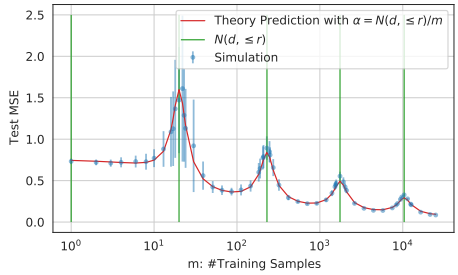
Figure 8: **Large** $d = 60$



(a) Using $\alpha = N(d, r)/m$

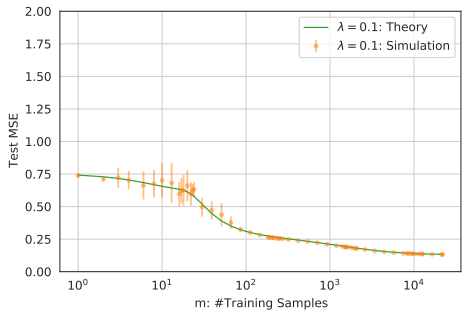


(b) Using Eq. (18)

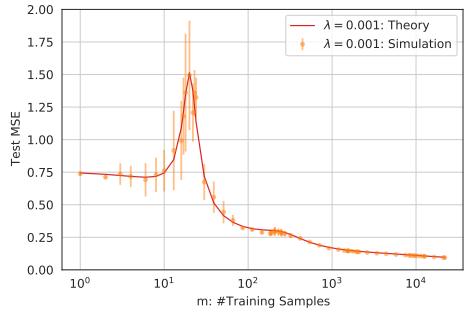


(c) Using Eq. (21)

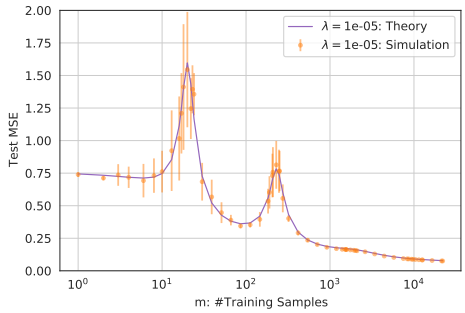
Figure 9: **Two improvements reduce the finite-size effect.** (a) The theoretical prediction is a bit off to the left when estimating α using $N(d, r)/m$. (b) The prediction from Eq. (18) is a bit smaller than the simulation due to the finite-size effect. (c) Almost perfect agreement between the prediction and the simulation after two improvements (1) replacing Eq. (18) by Eq. (21) and (2) estimating α with $N(d, \leq r)/m$ rather than $N(d, r)/m$. Here $d = 20$ and $p = 1$, i.e., inner product kernel.



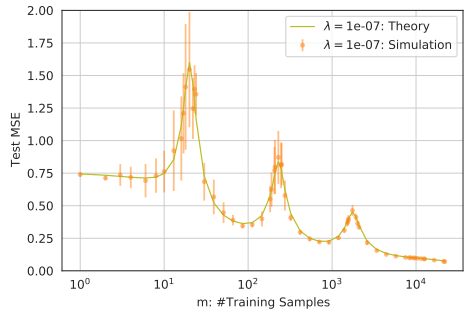
(a) No peak: $\lambda = 0.1$



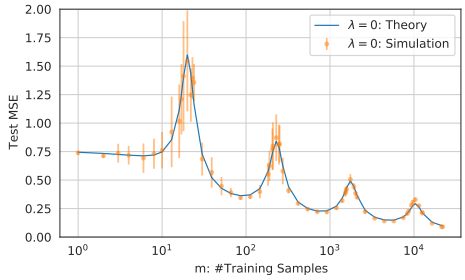
(b) One peak: $\lambda = 10^{-3}$



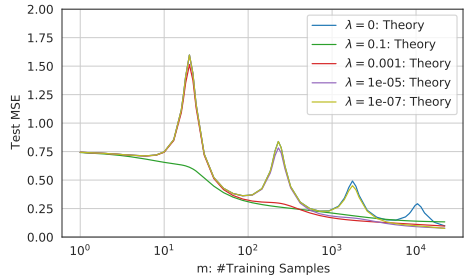
(c) Two peaks: $\lambda = 10^{-5}$



(d) Three peaks: $\lambda = 10^{-7}$

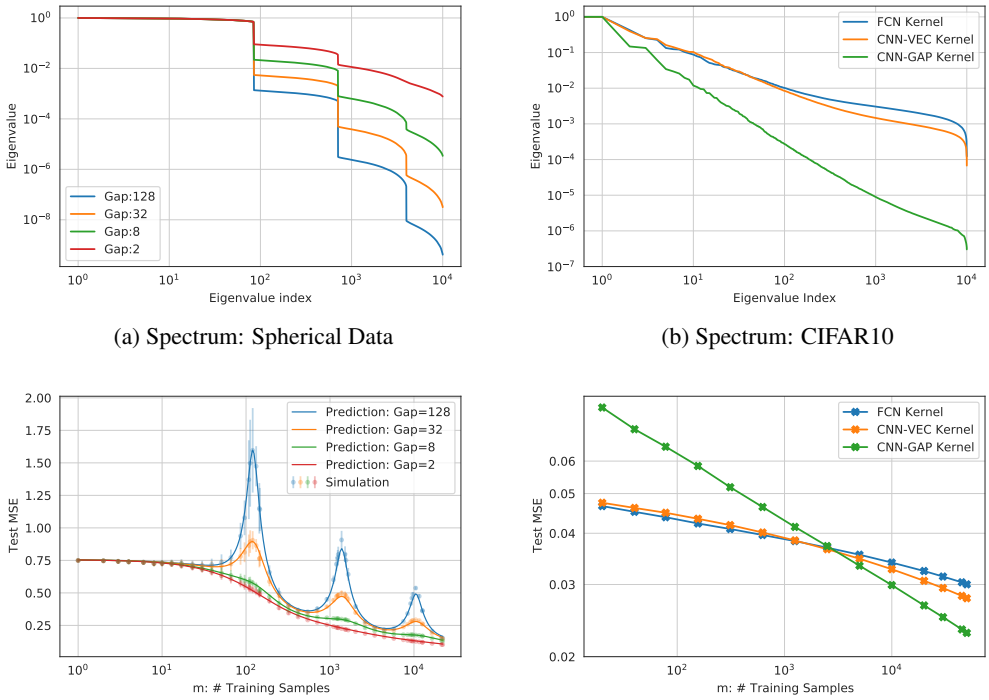


(e) Four peaks: $\lambda = 0$



(f) # Peaks vs Regularization.

Figure 10: **Controlling the number of peaks by varying the strength of regularization.**



(a) Spectrum: Spherical Data

(b) Spectrum: CIFAR10

(c) Test MSE: Spherical Data

(d) Test MSE: CIFAR10

Figure 11: **Spectrum (top) and learning curves (bottom) of Spherical data (left) vs. CIFAR10 (right.)** The spectrum of CIFAR10 has a power-law decay and does not contain any sizable spectral gap, which is the main cause of the multiple-descent phenomena. The learning curves of CIFAR10 have power-law decay for all three kernels.