STRONG INDUCTIVE BIASES PROVABLY PREVENT HARMLESS INTERPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical wisdom suggests that estimators should avoid fitting noise to achieve good generalization. In contrast, modern overparameterized models can yield small test error despite interpolating noise — a phenomenon often called "benign overfitting" or "harmless interpolation". This paper argues that the degree to which interpolation is harmless hinges upon the strength of an estimator's inductive bias, i.e., how heavily the estimator favors solutions with a certain structure: while strong inductive biases prevent harmless interpolation, weak inductive biases can even require fitting noise to generalize well. Our main theoretical result establishes tight non-asymptotic bounds for high-dimensional kernel regression that reflect this phenomenon for convolutional kernels, where the filter size regulates the strength of the inductive bias. We further provide empirical evidence of the same behavior for deep neural networks with varying filter sizes and rotational invariance.

1 INTRODUCTION

According to classical wisdom (see, e.g., Hastie et al. (2001)), an estimator that fits noise suffers from "overfitting", and cannot generalize well. A typical solution is to prevent interpolation, that is, stopping the estimator from achieving zero training error and thereby fitting less noise. For example, one can use ridge regularization or early stopping for iterative algorithms to obtain a model that has training error close to the noise level. However, large overparameterized models such as neural networks seem to behave differently: even on noisy data, they may achieve optimal test performance at convergence after interpolating the training data (Nakkiran et al., 2021; Belkin et al., 2019a) — a phenomenon referred to as *harmless interpolation* (Muthukumar et al., 2020) or *benign overfitting* (Bartlett et al., 2020) and often discussed in the context of *double descent* (Belkin et al., 2019a).

To date, we lack a general understanding of when interpolation is harmless for overparameterized models. In this paper, we argue that the *strength of an inductive bias* critically influences whether an estimator exhibits harmless interpolation. An estimator with a strong inductive bias heavily favors "simple" solutions that structurally align with the ground truth (such as sparsity or rotational invariance). Based on well-established high-probability recovery results of sparse linear regression (Tibshirani, 1996; Candes, 2008; Donoho & Elad, 2006), we expect that models with stronger inductive biases generalize better than ones with weaker inductive biases, particularly from noiseless data. In contrast, the effects of inductive biases are much less studied for interpolators of noisy data.

Recently, Donhauser et al. (2022) provided a first rigorous analysis of the effects of inductive bias strength on the generalization performance of linear max-ℓ_p-margin/min-ℓ_p-norm interpolators. In particular, the authors prove that a stronger inductive bias (small $p \to 1$) not solely enhances a model's ability to generalize on noiseless data, but also increases a model's sensitivity to noise — eventually harming generalization when interpolating noisy data. As a consequence, their result suggests that interpolation might not be harmless when the inductive bias is too strong.

In this paper, we confirm the hypothesis and show that strong inductive biases indeed prevent harmless interpolation, while also moving away from sparse linear models. As one example, we consider data where the true labels nonlinearly only depend on input features in a local neighborhood, and vary the strength of the inductive biases via the filter size of convolutional kernels or shallow convolutional neural networks — small filter sizes encourage functions that depend nonlinearly only on local neighborhoods of the input features. As a second example, we also investigate classification for

rotationally invariant data, where we encourage different degrees of rotational invariance for neural networks. In particular,

- we prove a phase transition between harmless and harmful interpolation that occurs by varying the strength of the inductive bias via the filter size of convolutional kernels for kernel regression in the high-dimensional setting (Theorem 1).
- we further show that, for weak inductive biases, not only is interpolation harmless but partially fitting the observation noise is in fact necessary (Theorem 2).
- we show the same phase transition experimentally for neural networks with two common inductive biases: varying convolution filter size, and rotational invariance enforced via data augmentation (Section 4).

2 Related Work

We now discuss three groups of related work and explain how their theoretical results cannot reflect the phase transition between harmless and harmful interpolation for high-dimensional kernel learning.

Low-dimensional kernel learning: Many recent works (Bietti et al., 2021; Favero et al., 2021; Bietti, 2022; Cagnetta et al., 2022) prove statistical rates for kernel regression with convolutional kernels in low-dimensional settings, but crucially rely on ridge regularization. In general, one cannot expect harmless interpolation for such kernels in the low-dimensional regime (Rakhlin & Zhai, 2019; Mallinar et al., 2022; Buchholz, 2022); positive results exist only for very specific adaptive spiked kernels (Belkin et al., 2019b). Furthermore, techniques developed for low-dimensional settings (see, e.g., Schölkopf et al. (2018)) usually suffer from a curse of dimensionality, that is, the bounds become vacuous in high-dimensional settings where the input dimension grows with the number of samples.

High-dimensional kernel learning: One line of research (Liang et al., 2020; McRae et al., 2022; Liang & Rakhlin, 2020; Liu et al., 2021) tackles high-dimensional kernel learning and proves non-asymptotic bounds using advanced high-dimensional random matrix concentration tools from El Karoui (2010). However, those results heavily rely on a bounded Hilbert norm assumption. This assumption is natural in the low-dimensional regime, but misleading in the high-dimensional regime, as pointed out in Donhauser et al. (2021b). Another line of research (Ghorbani et al., 2021; 2020; Mei et al., 2021; Ghosh et al., 2022; Misiakiewicz & Mei, 2021; Mei et al., 2022) asymptotically characterizes the precise risk of kernel regression estimators in specific settings with access to a kernel's eigenfunctions and eigenvalues. However, these asymptotic results are insufficient to investigate how varying the filter size of a convolutional kernel affects the risk of a kernel regression estimator. In contrast to both lines of research, we prove tight non-asymptotic matching upper and lower bounds for high-dimensional kernel learning which precisely capture the phase transition described in Section 3.2.

Overfitting of structured interpolators: Several works question the generality of harmless interpolation for models that incorporate strong structural assumptions. Examples include structures enforced via data augmentation (Nishi et al., 2021), adversarial training (Rice et al., 2020; Kamath et al., 2021; Sanyal et al., 2021; Donhauser et al., 2021a), neural network architectures (Li et al., 2021), pruning-based sparsity (Chang et al., 2021), and sparse linear models (Wang et al., 2022; Muthukumar et al., 2020; Chatterji & Long, 2022). In this paper, we continue that line of research and offer a new theoretical perspective to characterize when interpolation is expected to be harmless.

3 Theoretical results

For convolutional kernels, a small filter size induces a strong bias towards estimators that depend nonlinearly on the input features only via a small patch in the input. This section analyzes the effect of filter size (as an example inductive bias) on the degree of harmless interpolation for kernel ridge regression. For this purpose, we derive and compare tight non-asymptotic bias and variance bounds as a function of filter size for min-norm interpolators and optimally ridge-regularized estimators (Theorem 1). Furthermore, we prove for large filter sizes that not only does harmless interpolation occur (Theorem 1), but fitting some degree of noise is even necessary to achieve optimal test performance (Theorem 2).

3.1 SETTING

We study kernel regression with a (cyclic) convolutional kernel in a high-dimensional setting where the number of training samples n scales with the dimension of the input data d as $n \in \Theta(d^\ell)$. We use the same setting as in previous works on high-dimensional kernel learning such as Misiakiewicz & Mei (2021): we assume that the training samples $\{x_i, y_i\}_{i=1}^n$ are i.i.d. draws from the distributions $x_i \sim \mathcal{U}(\{-1,1\}^d)$, and $y_i = f^\star(x_i) + \epsilon_i$ with ground truth $f^\star$ and noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For simplicity of exposition, we further assume that $f^\star(x) = x_1 x_2 \cdots x_{L^*}$, with L^* specified in Theorem 1.

While the assumptions on the noise and ground truth can be easily extended by following the proof steps in Section 5, generalizing the feature distribution is challenging. Indeed, existing results that establish precise risk characterizations (see Section 2) crucially rely on hypercontractivity of the feature distribution — an assumption so far only proven for few high-dimensional distributions, including the hypersphere (Beckner, 1992), and the discrete hypercube (Beckner, 1975) which we use in this paper. Hypercontractivity is essential to tightly control the empirical kernel matrix within Lemma 3 in Section 5. Generalizations beyond this assumption require the development of new tools in random matrix theory, which we consider important future work.

We consider (cyclic) convolutional kernels with filter size $q \in \{1, \dots, d\}$ of the form

$$\mathcal{K}(x, x') = \frac{1}{d} \sum_{k=1}^{d} \kappa \left(\frac{\langle x_{(k,q)}, x'_{(k,q)} \rangle}{q} \right), \tag{1}$$

where $x_{(k,q)} := [x_{\mathrm{mod}(k,d)} \cdots x_{\mathrm{mod}(k+q-1,d)}]$, and $\kappa \colon [-1,1] \to \mathbb{R}$ is a nonlinear function that satisfies standard regularity assumptions (see Assumption 1 in Appendix B) that hold for instance for the exponential function. Decreasing the filter size q restricts kernel regression solutions to depend nonlinearly only on local neighborhoods instead of the entire input x.

We analyze the kernel ridge regression (KRR) estimator, which is the minimizer of the following convex optimization problem:

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \frac{\lambda}{n} \|f\|_{\mathcal{H}}^2, \tag{2}$$

where $\mathcal{H}$ is the Reproducing Kernel Hilbert space (RKHS) over $\{-1,1\}^d$ generated by the convolutional kernel $\mathcal{K}$ in Equation (1), $\|\cdot\|_{\mathcal{H}}$ the corresponding norm, and $\lambda > 0$ the ridge regularization penalty. [1] In the interpolation limit ($\lambda \to 0$), we obtain the min-RKHS-norm interpolator

$$\hat{f}_0 = \arg\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \quad \text{s.t.} \quad \forall i: \ f(x_i) = y_i. \tag{3}$$

For simplicity, we refer to $\hat{f}_0$ as the kernel ridge regression estimator with $\lambda = 0$. We evaluate all estimators with the expected population risk over the noise, defined as

$$\textbf{Risk}(\hat{f}_\lambda) := \underbrace{\mathbb{E}_x \left[\left(\mathbb{E}_\epsilon[\hat{f}_\lambda(x)] - f^\star(x) \right)^2 \right]}_{:=\textbf{Bias}^2(\hat{f}_\lambda)} + \underbrace{\mathbb{E}_{x,\epsilon} \left[\left(\hat{f}_\lambda(x) - \mathbb{E}_\epsilon[\hat{f}_\lambda(x)] \right)^2 \right]}_{:=\textbf{Variance}(\hat{f}_\lambda)}.$$

3.2 MAIN RESULT

We now present tight upper and lower bounds for the prediction error of kernel regression estimators in the setting from Section 3.1. The resulting rates hold for the high-dimensional regime, that is, when both the ambient dimension d and filter size q scale with n. [2] We defer the proof to Section 5.

[1] Note that previous works show how early-stopped gradient methods on the square loss behave statistically similarly to kernel ridge regression (Raskutti et al., 2014; Wei et al., 2017).

[2] We hide positive constants that depend at most on ℓ and β (defined in Theorem 1) using the standard Bachmann–Landau notation $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, as well as $\lesssim$, $\gtrsim$, and use $c, c_1, \dots$ as generic positive constants.

Theorem 1 (Non-asymptotic prediction error rates)**.** *Let $\ell > 0$, $\beta \in (0,1)$, $\ell_\sigma \in \mathbb{R}$. Assume a dataset and a kernel as described in Section 3.1, with the kernel satisfying Assumption 1. Assume further $n \in \Theta(d^\ell)$, the filter size $q \in \Theta\left(d^\beta\right)$, and $\sigma^2 \in \Theta(d^{-\ell_\sigma})$. Lastly, define $\bar{\delta} := \frac{\ell-1}{\beta} - \lfloor \frac{\ell-1}{\beta} \rfloor$ and $\delta := \frac{\ell-\ell_\lambda-1}{\beta} - \lfloor \frac{\ell-\ell_\lambda-1}{\beta} \rfloor$ for any ℓ_λ. Then, with probability at least $1 - cd^{-\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$ uniformly over all $\ell_\lambda \in [0, \ell-1)$, the KRR estimate $\hat{f}_\lambda$ in Equation* (2) *with* $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$ *satisfies*

$$\textbf{\textit{Variance}}(\hat{f}_\lambda) \in \Theta\left(n^{\frac{-\ell_\sigma-\ell_\lambda}{\ell} - \frac{\beta}{\ell}\min\{\delta, 1-\delta\}}\right).$$

Further, for a ground truth $f^\star(x) = x_1 x_2 \cdots x_{L^*}$ with $L^* \leq \left\lceil \frac{\ell-\ell_\lambda-1}{\beta} \right\rceil$, with probability at least $1 - cd^{-\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$, we have

$$\textbf{\textit{Bias}}^2(\hat{f}_\lambda) \in \Theta\left(n^{-2-\frac{2}{\ell}(-\ell_\lambda-1-\beta(L^*-1))}\right).$$

Finally, by setting $\ell_\lambda = 0$, both rates also hold for the min-RKHS-norm interpolator $\hat{f}_0$ in Eq. (3)*.*

Note how the theorem statement reflects the usual intuition for the effect of noise and regularization strength on bias and variance via the parameter ℓ_λ: With increasing ridge regularization ℓ_λ (and thus increasing λ), the bias increases and the variance decreases. Similarly, as noise increases (and thus decreasing ℓ_σ), the variance increases.

Phase transition as a function of β: In the following, we focus on the impact of the filter size $q \in \Theta\left(d^\beta\right)$ on the risk (sum of bias and variance) via the growth rate β. Recalling that a small filter size (small values of β) corresponds to a strong inductive bias, and vice versa, Figure 1 demonstrates how the strength of inductive bias affects generalization. For illustration, we choose the ground truth $f^\star(x) = x_1 x_2$ so that the assumption on L^* is satisfied for all β. Specifically, Figure 1a shows the rates for the min-RKHS-norm interpolator $\hat{f}_0$ and the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$, where we choose λ_{opt} to minimize the expected population risk $\textbf{Risk}(\hat{f}_{\lambda_{\text{opt}}})$. Furthermore, Figure 1b depicts the (statistical) bias and variance of the interpolator $\hat{f}_0$. At the threshold $\beta^* \in (0,1)$, implicitly defined as the β at which the rates of statistical bias and variance in Theorem 1 match, we can observe the following phase transition:

- For $\beta < \beta^*$, that is, for strong inductive biases, the rates in Figure 1a for the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$ are strictly better than the ones for the corresponding interpolator $\hat{f}_0$. In other words, we are observing *harmful interpolation*.
- For $\beta > \beta^*$, that is, for weak inductive biases, the rates in Figure 1a of the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$ and the min-RKHS-norm interpolator $\hat{f}_0$ match. Hence, we observe *harmless interpolation*.

In the following theorem, we additionally show that interpolation is not only harmless for $\beta > \beta^*$, but the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$ necessarily fits part of the noise and has a training error strictly below the noise level. In contrast, we show that when interpolation is harmful in Figure 1a, that is, when $\beta < \beta^*$, the training error of the optimally ridge-regularized model approaches the noise level.

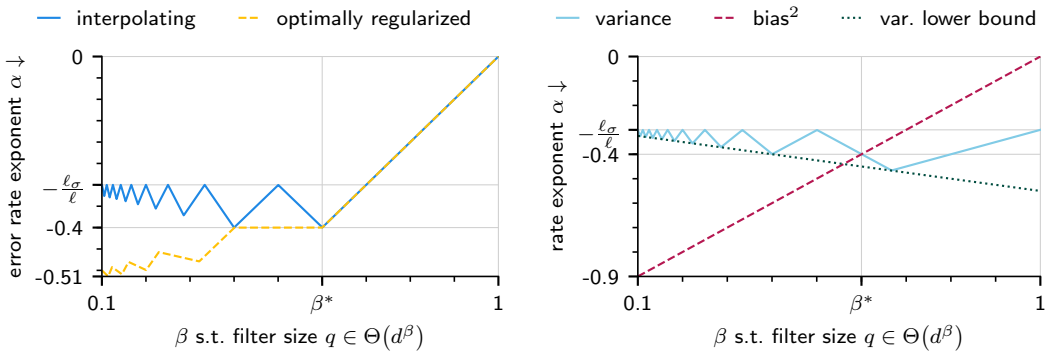
Theorem 2 (Training error (informal))**.** *Let λ_{opt} be such that the expected population risk $\textbf{Risk}(\hat{f}_{\lambda_{opt}})$ is minimal, and let β^* be the threshold where the bias and variance bounds in Theorem 1 are of the same order for the interpolator $\hat{f}_0$ (setting $\ell_\lambda = 0$). Then, the expected training error converges in probability:*

$$\lim_{n,d\to\infty} \frac{1}{\sigma^2}\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_i (\hat{f}_{\lambda_{opt}}(x_i) - y_i)^2\right] \quad \begin{cases} = 1 & \beta < \beta^*, \\ \leq c_\beta & \beta \geq \beta^*, \end{cases}$$

where $c_\beta < 1$ for any $\beta > \beta^*$.

We refer to Appendix D for the proof and a more general statement.

Bias-variance trade-off: We conclude by discussing how the phase transition arises from a (statistical) bias and variance trade-off for the min-RKHS-norm interpolator as a function of β, reflected in



(a) Interpolating vs. regularized model

(b) Variance vs. bias for interpolating model

Figure 1: Illustration of the rates in Theorem 1 for high-dimensional kernel ridge regression as a function of β — the rate of the filter size $q \in \Theta(d^\beta)$. (a) Rate exponent α of the **Risk** $\in \Theta(n^\alpha)$ for the interpolator $\hat{f}_0$ vs. the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$. (b) Rate exponent of the variance and bias for the interpolator $\hat{f}_0$. For both illustrations, we choose $\hat{f}_0$ with $\ell = 2$, $\ell_\sigma = 0.6$, and the ground truth $f^\star(x) = x_1 x_2$. Lastly, β^* denotes the threshold where the bias and variance terms in Theorem 1 match, and where we observe a phase transition between harmless and harmful interpolation.

Theorem 1 when setting $\ell_\lambda = 0$ and illustrated in Figure 1b. While the statistical bias monotonically decreases with decreasing β (i.e., increasing strength of the inductive bias), the variance follows a multiple descent curve with increasing minima as β decreases. Hence, analogous to the observations in Donhauser et al. (2022) for linear max-ℓ_p-margin/min-ℓ_p-norm interpolators, the interpolator achieves the optimal performance at a $\beta \in (0, 1)$, and therefore at a moderate inductive bias. Finally, we note that Liang et al. (2020) previously observed a multiple descent curve for the variance, but as a function of input dimension and without any connection to structural biases.

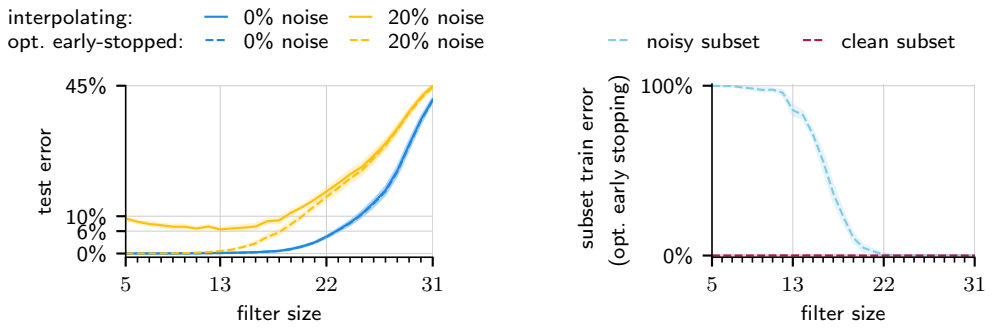
4 Experiments

We now empirically study whether the phase transition phenomenon that we prove for kernel regression persists for deep neural networks with feature learning. More precisely, we present controlled experiments to investigate how the strength of a CNN's inductive bias influences if interpolating noisy data is harmless. In practice, the inductive bias of a neural network varies by way of design choices such as the architecture (e.g., convolutional vs. fully-connected vs. graph networks) or the training procedure (e.g., data augmentation, adversarial training). We focus on two examples: convolutional filter size that we vary via the architecture, and rotational invariance via data augmentation. To isolate the effects of inductive biases and provide conclusive results, we use datasets where we know a priori that the ground truth exhibits a simple structure that matches the networks' inductive bias. See Appendix E for experimental details.

Analogous to ridge regularization for kernels, we use early stopping as a mechanism to prevent noise fitting. Our experiments compare optimally early-stopped CNNs to their interpolating versions. This highlights a trend that mirrors our theoretical results: the stronger the inductive bias of a neural network grows, the more harmful interpolation becomes. These results suggest exciting future work: proving this trend for models with feature learning.

4.1 Filter size of CNNs on synthetic images

In a first experiment, we study the impact of filter size on the generalization of interpolating CNNs. As a reminder, small filter sizes yield functions that depend nonlinearly only on local neighborhoods of the input features. To clearly isolate the effects of filter size, we choose a special architecture on a synthetic classification problem such that the true label function is indeed a CNN with small filter size. In particular, we generate images of size 32×32 containing scattered circles (negative class) and crosses (positive class) with size at most 5×5. Thus, decreasing filter size down to 5×5 corresponds to a stronger inductive bias. Motivated by our theory, we hypothesize that interpolating noisy data is harmful with small filter size, but harmless when using a large filter size.



(a) Test error

(b) Training error of optimally early-stopped models

Figure 2: Convolutional neural network experiments with varying filter size on synthetic image data. (a) For noisy data, small filter sizes (strong inductive bias) induce a gap between the generalization performance of interpolating (yellow solid) vs. optimally early-stopped models (yellow dashed). The gap vanishes as the inductive bias decreases (i.e., filter size increases). For noiseless data, interpolation is always harmless (blue). (b) Training error of the optimally early-stopped model (optimized for test error) on the noisy and clean subsets of a training set with 20% label noise. Under optimal early stopping, models with a strong inductive bias ignore all noisy samples (100% error on the noisy subset), while models with a weak inductive bias fit all noisy training samples (0% error on the noisy subset). All lines show the mean over five random datasets, and shaded areas the standard error; see Section 4.1 for the experiment setup.

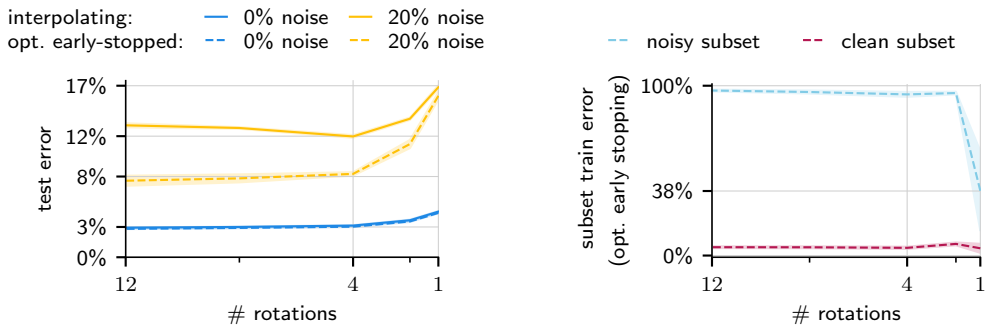
Training setup: In the experiments, we use CNNs with a single convolutional layer, followed by global spatial max pooling and two dense layers. We train those CNNs with different filter sizes on 200 training samples (either noiseless or with 20% label flips) to minimize the logistic loss and achieve zero training error. We repeat all experiments over 5 random datasets with 15 optimizations per dataset and filter size, and report the average 0-1-error for 100k test samples per dataset. For a detailed discussion on the choice of hyperparameters and more experimental details, see Appendix E.1.

Results: First, the noiseless error curves (blue) in Figure 2a confirm the common intuition that the strongest inductive bias (matching the ground truth) at size 5 yields the lowest test error. More interestingly, for 20% training noise (yellow), Figure 2a reveals a similar phase transition as Theorem 1 and confirms our hypothesis: Models with weak inductive biases (large filter sizes) exhibit harmless interpolation, as indicated by the matching test error of interpolating (yellow solid) and optimally early-stopped (yellow dashed) models. In contrast, as filter size decreases, models with a strong inductive bias (small filter sizes) suffer from an increasing gap in test errors when interpolating versus using optimal early stopping. Furthermore, Figure 2b reflects the dual perspective of the phase transition as presented in Theorem 2 under optimal early stopping: models with a small filter size entirely avoid fitting training noise, such that the training error on the noisy training subset equals 100%, while models with a large filter size interpolate the noise.

Difference to double descent: One might suspect that our empirical observations simply reflect another form of double descent (Belkin et al., 2019a). As a CNN's filter size increases (inductive bias becomes weaker), so does the number of parameters and degree of overparameterization. Thus, double descent predicts vanishing benefits of regularization due to model size for weak inductive biases. Nevertheless, we argue that the phenomenon we observe here is distinct, and provide an extended discussion in Appendix E.3. In short, we choose sufficiently large networks and tune their hyperparameters to ensure that all models interpolate and yield small training loss, even for filter size 5 and 20% training noise. To justify this approach, we repeat a subset of the experiments while significantly increasing the convolutional layer width. As the number of parameters increases for a fixed filter size, double descent would predict that the benefits of optimal early stopping vanish. However, we observe that our phenomenon persists. In particular, for filter size 5 (strongest inductive bias), the test error gap between interpolating and optimally early-stopped models remains large.

4.2 Rotational invariance of Wide Residual Networks on satellite images

In a second experiment, we investigate rotational invariance as an inductive bias for CNNs whenever true labels are independent of an image's rotation. Our experiments enforce inductive biases by fitting models on multiple rotated versions of an original training dataset, effectively performing varying



(a) Test error

(b) Training error of optimally early-stopped models

Figure 3: Varying degrees of rotational invariance when fitting Wide Residual Networks on satellite images. (a) For noisy data, a strong bias towards rotational invariance (via the number of augmented rotations) induces a gap between the generalization performance of interpolating (yellow solid) vs. optimally early-stopped models (yellow dashed). The gap decreases as the inductive bias decreases (# rotations decreases). (b) Training error of optimally early-stopped models (w.r.t. the test error) on the noisy and clean subsets of a training set with 20% label noise: For maximum rotational invariance (12 rotations), optimally early-stopped models avoid fitting noisy data (close to 100% training error), yet no rotational invariance (1 rotation) requires fitting noise (less than 50% training error on noisy subset) for optimal generalization. All lines show the mean over five optimization runs, and shaded areas the standard error; see Section 4.2 for the experiment setup.

degrees of data augmentation. [3] As an example dataset with a rotationally invariant ground truth, we classify satellite images from the EuroSAT dataset (Helber et al., 2018) into 10 types of land usage. Because the true labels are independent of image orientation, we expect rotational invariance to be a particularly fitting inductive bias for this task.

Training and test setup: For computational reasons, we subsample the original EuroSAT training set to 7680 raw training and 10k raw test samples. In the noisy case, we replace 20% of the raw training labels with a wrong label chosen uniformly at random. We then vary the strength of the inductive bias towards rotational invariance by augmenting the dataset with an increasing number of k rotated versions of itself. For each sample, we use k equal-spaced angles spanning $360°$, plus a random offset. Note that training noise applies before rotations, so that all rotated versions of the same image share the same label. We then center-crop all rotated images such that they only contain valid pixels. In all experiments, we fit Wide Residual Networks (Zagoruyko & Komodakis, 2016) on the augmented training set for 5 different network initializations. We evaluate the 0-1-error on the randomly rotated test samples to avoid distribution shift effects from image interpolation. All random rotations are the same for all experiments and stay fixed throughout training. See Appendix E.2 for more experimental details. Lastly, we perform additional experiments with larger models to differentiate from double descent; see Appendix E.3 for the results and further discussion.

Results: Similar to the previous subsection, Figure 3a corroborates our hypothesis under rotational invariance: stronger inductive biases result in lower test errors on noiseless data, but an increased gap between the test errors of interpolating and optimally early-stopped models. In contrast to filter size, the phase transition is more abrupt; invariance to $180°$ rotations already prevents harmless interpolation. Figure 3b confirms this from a dual perspective, since all models with some rotational invariance cannot fit noisy samples for optimal generalization.

5 Proof of the main result

The proof of the main result, Theorem 1, proceeds in two steps: First, Section 5.1 presents a fixed-design result that yields matching upper and lower bounds for the prediction error of general kernels under additional conditions. Second, in Section 5.2, we show that the setting of Theorem 1 satisfies those conditions with high probability over dataset draws.

[3] Techniques based on data augmentation can efficiently enforce rotational invariance; see, e.g., Yang et al. (2019).

Notation Assuming that inputs are draws from a data distribution ν (i.e., $x, x' \sim \nu$), we can decompose and divide any continuous, positive semi-definite kernel function as

$$\mathcal{K}(x,x') = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(x') = \underbrace{\sum_{k=1}^{m} \lambda_k \psi_k(x)\psi_k(x')}_{:=\mathcal{K}_{\leq m}(x,x')} + \underbrace{\sum_{k=m+1}^{\infty} \lambda_k \psi_k(x)\psi_k(x')}_{:=\mathcal{K}_{>m}(x,x')}, \tag{4}$$

where $\{\psi_k\}_{k\geq 1}$ is an orthonormal basis of the RKHS induced by $\langle f, g\rangle_\nu := \mathbb{E}_{x\sim\nu}[f(x)g(x)]$ and the eigenvalues λ_k are sorted in descending order. In the following, we write $[\cdot]_{i,j}$ to refer to the entry in row i and column j of a matrix. Then, we define the empirical kernel matrix for $\mathcal{K}$ as $\mathbf{K} \in \mathbb{R}^{n\times n}$ with $[\mathbf{K}]_{i,j} := \mathcal{K}(x_i, x_j)$, and analogously the truncated versions $\mathbf{K}_{\leq m}$ and $\mathbf{K}_{>m}$ for $\mathcal{K}_{\leq m}$ and $\mathcal{K}_{>m}$, respectively. Next, we utilize the matrices $\mathbf{\Psi}_{\leq m} \in \mathbb{R}^{n\times m}$ with $[\mathbf{\Psi}_{\leq m}]_{i,l} := \psi_l(x_i)$, and $\mathbf{D}_{\leq m} := \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. We further use the squared kernel $\mathcal{S}(x,x') := \mathbb{E}_{z\sim\nu}[\mathcal{K}(x,z)\mathcal{K}(z,x')]$, its truncated versions $\mathcal{S}_{\leq m}$ and $\mathcal{S}_{>m}$, as well as the corresponding empirical kernel matrices $\mathbf{S}, \mathbf{S}_{\leq m}, \mathbf{S}_{>m} \in \mathbb{R}^{n\times n}$. Finally, we write $\mu_{\min}(\cdot)$ and $\mu_{\max}(\cdot)$ (or $\|\cdot\|$) to indicate the min and max eigenvalue of a symmetric positive-definite matrix, respectively, and use $\langle\cdot,\cdot\rangle$ to indicate the Euclidean inner product in $\mathbb{R}^d$.

5.1 Generalization bound for fixed-design

First, Theorem 3 provides tight fixed-design bounds for the prediction error.

Theorem 3 (Generalization bound for fixed-design)**.** *Let $\mathcal{K}\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a kernel, and $\{(x_i, y_i)\}_{i=1}^n$ a dataset where $y_i = f^\star(x_i) + \epsilon_i$ for zero-mean i.i.d. noise ϵ_i with variance σ^2 and a ground truth $f^\star$. For any $m \in \mathbb{N}$ such that*

$$\left\| \mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}/n - \mathbf{I}_m \right\| \leq \frac{1}{2} \tag{5}$$

the KRR estimate $\hat{f}_\lambda$ in Equation (2) *for any $\lambda \geq 0$ has a variance upper and lower bounded by*

$$\frac{r_1^2\tau_1^2}{2r_2^2(1.5+r_1)^2}\frac{m}{n} + \frac{\mathbf{Tr}\left(\mathbf{S}_{>m}\right)}{r_2^2\max\{\lambda,1\}^2} \leq \textbf{\textit{Variance}}(\hat{f}_\lambda)/\sigma^2 \leq 6\frac{r_2^2}{r_1^2}\frac{m}{n} + \frac{\mathbf{Tr}\left(\mathbf{S}_{>m}\right)}{r_1^2\max\{\lambda,1\}^2},$$

where $\tau_1 = \min\left\{\frac{n\lambda_m}{\max\{\lambda,1\}}, 1\right\}$, $\tau_2 = \max\left\{\frac{n\lambda_{m+1}}{\max\{\lambda,1\}}, 1\right\}$, $r_1 = \frac{\|\mathbf{K}_{>m}\|+\lambda}{\max\{\lambda,1\}}$, $r_2 = \frac{\mu_{\min}(\mathbf{K}_{>m})+\lambda}{\max\{\lambda,1\}}$. Furthermore, for any ground truth which can be expressed as $f^\star(x) = \sum_{k=1}^m a_k\psi_k(x)$ with $a \in \mathbb{R}_+^m$ and ψ_k as defined via Equation (4)*, the bias is upper and lower bounded by*

$$\frac{r_1^2\tau_1^2}{(1.5+r_1)^2}\max\{\lambda,1\}^2\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2} \leq \textbf{\textit{Bias}}^2(\hat{f}_\lambda) \leq 12\left(r_2^2 + 1.5\frac{r_2^3}{r_1^2}\right)\tau_2\max\{\lambda,1\}^2\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2}.$$

See Appendix A for the proof. Note that this result holds for any fixed-design dataset. We derive the main ideas from the proofs in Bartlett et al. (2020); Tsigler & Bartlett (2020), where the authors establish tight bounds for the min-ℓ_2-norm interpolator on independent sub-Gaussian features.

Remark 1 (Comparison with McRae et al. (2022))**.** *The upper bound for the bias in Theorem 1 from McRae et al. (2022) depends on the suboptimal term $\|\mathbf{D}_{\leq m}^{-1/2}a\|/n$, but also applies to more general ground truths. We improve that bound in Theorem 3 and present a matching lower bound.*

5.2 Proof of Theorem 1

Throughout the remainder of the proof, all statements hold for the setting in Section 3.1 and under the assumptions of Theorem 1; in particular, Assumption 1 on κ, $n \in \Theta(d^\ell)$, and $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$. Furthermore, as in Theorem 1, we use $\delta, \bar{\delta} > 0$ where $\delta := \frac{\ell-\ell_\lambda-1}{\beta} - \left\lfloor\frac{\ell-\ell_\lambda-1}{\beta}\right\rfloor$ and $\bar{\delta} := \frac{\ell-1}{\beta} - \left\lfloor\frac{\ell-1}{\beta}\right\rfloor$. We show that there exists a particular m for which the conditions in Theorem 3 hold and we can control the terms in the bias and variance bounds.

Step 1: Conditions for the bounds in Theorem 3. We first derive sufficient conditions on m such that the conditions on $\mathbf{\Psi}_{\leq m}$ and $f^\star$ in Theorem 3 hold with high probability. The following result from standard random matrix concentration shows that any $m \ll n$ satisfies Equation (5) with high probability.

Lemma 1 (Corollary of Theorem 5.44 in Vershynin (2012))**.** *For d large enough, with probability at least $1 - cd^{-\beta\bar{\delta}}$, all $m \in \mathcal{O}(n \cdot d^{-\beta\bar{\delta}})$ satisfy Equation* (5)*.*

See Appendix C.1 for the proof. Simultaneously, to ensure that $f^\star$ is contained in the span of the first m eigenfunctions, m must be sufficiently large. We formalize this in the following lemma.

Lemma 2 (Condition for the bias bound)**.** *Consider a kernel as in Theorem 1 satisfying Assumption 1, and a ground truth $f^\star(x) = \prod_{j=1}^{L^*} x_j$ with $L^* \geq 1$. Then, for any m such that $\lambda_m \in o\left(\frac{1}{dq^{L^*-1}}\right)$ and d sufficiently large, $f^\star$ is in the span of the first m eigenfunctions and $\left\|\mathbf{D}_{\leq m}^{-1} a\right\| \in \Theta\left(dq^{L^*-1}\right)$.*

See Appendix B.3 for the proof.

Step 2: Concentration of the (squared) kernel matrix. In a second step, we show existence of an m that satisfies the sufficient conditions, and for which the spectrum of the kernel matrix $\mathbf{K}_{>m}$ and squared kernel matrix $\mathbf{S}_{>m}$ concentrates. First, Lemma 3 yields high-probability bounds for a specific set of m.

Lemma 3 (Tight bound conditions)**.** *With probability at least $1 - cd^{-2\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$ uniformly over all $\ell_\lambda \in [0, \ell - 1)$, for any $\lambda \in \mathbb{R}$ and $m \in \mathbb{N}$ with $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$ and $n\lambda_m \in \Theta(\max\{\lambda, 1\})$, we have*

$$r_1, r_2 \in \Theta(1) \qquad \textit{and} \qquad d^{\ell_\lambda} q^{\delta-1} \lesssim \mathbf{Tr}(\mathbf{S}_{>m}) \lesssim d^{\ell_\lambda} q^{-\delta} + d^{\ell_\lambda} q^{\delta-1}.$$

We refer to Appendix C.2 for the proof, which heavily relies on the feature distribution. Note that the results for m in Lemma 3 also imply $\tau_1, \tau_2 \in \Theta(1)$.

Step 3: Completing the proof. Finally, we complete the proof by showing the existence of a particular m that simultaneously satisfies all conditions on m in Lemmas 1 to 3.

Lemma 4 (Eigendecay)**.** *There exists an m such that*

$$n\lambda_m \in \Theta(\max\{\lambda, 1\}) \qquad \textit{and} \qquad m \in \Theta\left(\frac{nq^{-\delta}}{\max\{\lambda, 1\}}\right) \subseteq \mathcal{O}(n \cdot d^{-\beta\bar{\delta}}).$$

Furthermore, assuming $L^* \leq \left\lceil \frac{\ell - \ell_\lambda - 1}{\beta} \right\rceil$, we have $\lambda_m \in o\left(\frac{1}{d \cdot q^{L^*-1}}\right)$.

We refer to Appendix B.4 for the proof. As a result, we can use Lemmas 1 to 4 to instantiate Theorem 3 for the setting in Theorem 1, resulting in the following tight high-probability bounds for variance and bias:

$$d^{-\ell_\lambda} q^{-\delta} + d^{-\ell_\lambda} q^{\delta-1} \lesssim \textbf{Variance}(\hat{f}_\lambda)/\sigma^2 \lesssim d^{-\ell_\lambda} q^{-\delta} + d^{-\ell_\lambda} q^{\delta-1},$$
$$d^{-2(\ell - \ell_\lambda - 1 - \beta(L^*-1))} \lesssim \textbf{Bias}^2(\hat{f}_\lambda) \lesssim d^{-2(\ell - \ell_\lambda - 1 - \beta(L^*-1))}.$$

Reformulating the bounds in terms of n then concludes the proof of Theorem 1.

6 Summary and future work

In this paper, we highlight how the strength of an inductive bias impacts generalization. More precisely, we study when the gap in test error between optimally ridge-regularized or early-stopped models and their interpolating counterparts is zero, that is, when interpolation is harmless. In particular, we prove a phase transition for kernel regression using convolutional kernels with different filter sizes: a weak inductive bias (large filter size) yields harmless interpolation, and even requires fitting noise for optimal test performance, whereas interpolating noise leads to suboptimal generalization for a strong inductive bias (small filter size). Intuitively, this phenomenon arises from a bias-variance trade-off, captured by our main result in Theorem 1: with increasing inductive bias, the risk on noiseless training samples (bias) decreases, while the sensitivity to noise in the training data (variance) increases. Our empirical results on neural networks suggest that this phenomenon might extend to models with feature learning, which opens up an avenue for exciting future work.

REFERENCES

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

William Beckner. Inequalities in fourier analysis. *Annals of Mathematics*, 102:159–182, 1975.

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

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*, 116:15849–15854, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the Conference on Learning Theory (COLT)*, 2019b.

Alberto Bietti. Approximation and learning with deep convolutional models: a kernel perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Simon Buchholz. Kernel interpolation in sobolev spaces is not consistent in low dimensions. In *Proceedings of the Conference on Learning Theory (COLT)*, 2022.

Francesco Cagnetta, Alessandro Favero, and Matthieu Wyart. What can be learnt with wide convolutional networkds? *arXiv preprint arXiv:2208.01003*, 2022.

Emmanuel J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique*, 346:589–592, 2008.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2021.

Niladri S. Chatterji and Philip M. Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research (JMLR)*, 23:1–12, 2022.

Konstantin Donhauser, Alexandru Tifrea, Michael Aerni, Reinhard Heckel, and Fanny Yang. Interpolation can hurt robust generalization even when there is no noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021b.

Konstantin Donhauser, Nicolò Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

David L. Donoho and Michael Elad. On the stability of the basis pursuit in the presence of noise. *Signal Processing*, 86:511–532, 2006.

Noureddine El Karoui. The spectrum of kernel random matrices. *Annals of Statistics*, 38:1–50, 2010.

Alessandro Favero, Francesco Cagnetta, and Matthieu Wyart. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Nikhil Ghosh, Song Mei, and Bin Yu. The three stages of learning dynamics in high-dimensional kernel methods. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2001.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 2018.

Sandesh Kamath, Amit Deshpande, Subrahmanyam Kambhampati Venkata, and Vineeth N Balasubramanian. Can we have it all? on the trade-off between spatial and adversarial robustness of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jingling Li, Mozhi Zhang, Keyulu Xu, John Dickerson, and Jimmy Ba. How does a neural network's architecture impact its robustness to noisy labels? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48:1329–1347, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 2683–2711, 2020.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

Andrew D. McRae, Santhosh Karnik, Mark Davenport, and Vidya K. Muthukumar. Harmless interpolation in regression and classification with structured features. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Theodor Misiakiewicz and Song Mei. Learning with convolution and pooling operations in kernel methods. *arXiv e-prints*, art. arXiv:2111.08308, November 2021.

Theodor Misiakiewicz and Song Mei. Learning with convolution and pooling operations in kernel methods. *arXiv preprint arXiv:2111.08308*, 2021.

Vidya K. Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1:67–83, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021:124003, 2021.

Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Proceedings of the Conference on Learning Theory (COLT)*, 2019.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research (JMLR)*, 15: 335–366, 2014.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. How benign is benign overfitting? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Bernhard Schölkopf, Alexander J. Smola, and Francis Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Roman Vershynin. *Compressed Sensing: Theory and Applications*, chapter Introduction to the non-asymptotic analysis of random matrices, pp. 210–268. Cambridge University Press, 2012.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

A Proof of Theorem 3: Generalization bound for fixed-design

We proof Theorem 3 by deriving a closed-form expression for the bias and variance, and bounding each term individually. Due to the closed form, the proof does not rely on any matrix concentration results.

Proof. It is well-known that the KRR problem defined in Equation (2) is solved by the estimator

$$\hat{f}_\lambda(x) = \mathbf{y}^\intercal \mathbf{H}^{-1}\mathbf{k}(x),$$

where $x_i \sim \nu$, $\mathbf{y} \coloneqq \mathbf{f} + \epsilon = [f^\star(x_1) + \epsilon_1, \dots, f^\star(x_n) + \epsilon_n]^\intercal$, $\mathbf{k}(x) \coloneqq [\mathcal{K}(x_1, x), \dots, \mathcal{K}(x_n, x)]^\intercal$, and $\mathbf{H} \coloneqq \mathbf{K} + \lambda \mathbf{I}_n$. For this estimator, both bias and variance exhibit a closed-form expression:

$$
\begin{aligned}
\textbf{Bias}^2(\hat{f}_\lambda) &= \mathbb{E}_{x\sim\nu}[\big(f(x) - \mathbb{E}_\epsilon[(\mathbf{f}+\epsilon)^\intercal \mathbf{H}^{-1}\mathbf{k}(x)]\big)^2] \\
&= \mathbb{E}_{x\sim\nu}[\big(f(x) - \mathbf{f}^\intercal \mathbf{H}^{-1}\mathbf{k}(x)\big)^2] \\
&= \mathbb{E}_{x\sim\nu}\left[f(x)^2\right] - 2\mathbf{f}^\intercal \mathbf{H}^{-1}\mathbb{E}_{x\sim\nu}[f(x)\mathbf{k}(x)] + \mathbf{f}^\intercal \mathbf{H}^{-1}\mathbb{E}_{x\sim\nu}[\mathbf{k}(x)\mathbf{k}(x)^\intercal]\mathbf{H}^{-1}\mathbf{f} \\
&\overset{(i)}{=} a^\intercal a - 2a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{K}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m} a + a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{K}^{-1}\mathbf{S}\mathbf{K}^{-1}\mathbf{\Psi}_{\leq m} a, \qquad (6) \\
\textbf{Variance}(\hat{f}_\lambda)/\sigma^2 &= \frac{1}{\sigma^2}\mathbb{E}_{x\sim\nu}\mathbb{E}_\epsilon[\big((\mathbf{f}+\epsilon)^\intercal \mathbf{H}^{-1}\mathbf{k}(x) - \mathbb{E}_\epsilon[(\mathbf{f}+\epsilon)^\intercal \mathbf{H}^{-1}\mathbf{k}(x)]\big)^2] \\
&= \frac{1}{\sigma^2}\mathbb{E}_{x\sim\nu}\mathbb{E}_\epsilon[\epsilon^\intercal \mathbf{H}^{-1}\mathbf{k}(x)\mathbf{k}(x)^\intercal \mathbf{H}^{-1}\epsilon] \\
&= \frac{1}{\sigma^2}\sum_{i,j}\mathbb{E}[\epsilon_i\epsilon_j]\left[\mathbf{H}^{-1}\mathbb{E}_{x\sim\nu}[\mathbf{k}(x)\mathbf{k}(x)^\intercal]\mathbf{H}^{-1}\right]_{i,j} \\
&\overset{(i)}{=} \mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right). \qquad (7)
\end{aligned}
$$

In (i) we use the definition of the squared kernel $\mathcal{S}$, and that $f^\star(x) = \sum_{k=1}^m a_k \psi_k(x)$, since we are assuming that $f^\star$ is in the span of the first m eigenfunctions, as well as the following simple consequences of the eigenfunctions' orthonormality:

$$
\begin{aligned}
\mathbb{E}_{x\sim\nu}\left[f(x)^2\right] &= \sum_{k,k'=1}^m a_k a_{k'}\mathbb{E}_{x\sim\nu}\left[\psi_k(x)\psi_{k'}(x)\right] = \sum_{k,k'=1}^m a_k a_{k'}\delta_{k,k'} = \|a\|_2^2, \\
\mathbb{E}_{x\sim\nu}\left[[f(x)\mathbf{k}(x)]_i\right] &= \sum_{k=1}^m\sum_{k'=1}^\infty a_k\lambda_{k'}\mathbb{E}_{x\sim\nu}\left[\psi_k(x)\psi_{k'}(x)\right]\psi_{k'}(x_i) \\
&= \sum_k^m a_k\lambda_k\psi_k(x_i) = \left[\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}a\right]_i.
\end{aligned}
$$

We now bound the closed-form expressions of bias and variance individually.

Bias Lemma 5 below yields $\mathbf{S} = \mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal + \mathbf{S}_{>m}$. Hence, the bias decomposes as

$$
\begin{aligned}
\textbf{Bias}^2(\hat{f}_\lambda) &= a^\intercal a - 2a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}a + a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}a \\
&\quad + a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}a \\
&= \underbrace{\left\|\left(\mathbf{I}_m - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\right)a\right\|^2}_{B_1} + \underbrace{a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}a}_{B_2}.
\end{aligned}
$$

First, we rewrite B_1 as

$$
\begin{aligned}
B_1 &= \left\|\left(\mathbf{I}_m - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\right)a\right\|^2 \\
&\overset{(i)}{=} \left\|\left(\mathbf{D}_{\leq m} - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal\left(\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal + \mathbf{H}_{>m}\right)^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\right)\mathbf{D}_{\leq m}^{-1}a\right\|^2 \\
&\overset{(ii)}{=} \left\|\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{D}_{\leq m}^{-1}a\right\|^2, \qquad (8)
\end{aligned}
$$

where (i) uses the decomposition $\mathbf{H} = \mathbf{K}_{\leq m} + \mathbf{H}_{>m}$ with $\mathbf{H}_{>m} \coloneqq \mathbf{K}_{>m} + \lambda\mathbf{I}_n$, and (ii) applies the Woodbury identity.

Second, we can upper-bound B_2 as follows:

$$
\begin{aligned}
B_2 &\leq \|\mathbf{S}_{>m}\| a^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-2} \mathbf{\Psi}_{\leq m} a = \|\mathbf{S}_{>m}\| \left(\mathbf{D}_{\leq m}^{-1} a\right)^\intercal \mathbf{D}_{\leq m} \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-2} \mathbf{\Psi}_{\leq m} \mathbf{D}_{\leq m} \left(\mathbf{D}_{\leq m}^{-1} a\right) \\
&\overset{(i)}{=} \|\mathbf{S}_{>m}\| a^\intercal \mathbf{D}_{\leq m}^{-1} a \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-2} \mathbf{\Psi}_{\leq m} \\
&\qquad \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1} a \\
&\leq \underbrace{\|\mathbf{S}_{>m}\| \left\|\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-2} \mathbf{\Psi}_{\leq m}\right\|}_{C_1} \underbrace{\left\|\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1} a\right\|^2}_{=B_1},
\end{aligned}
$$

where (i) uses Lemma 6. Thus, the bias can be bounded by $B_1(1 + C_1)$ with $C_1 \geq 0$.

We proceed by upper bounding C_1:

$$
\begin{aligned}
1 + C_1 &\leq 1 + n\|\mathbf{S}_{>m}\| \left\|\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}}{n}\right\| \frac{1}{\mu_{\min}\left(\mathbf{H}_{>m}\right)^2} \\
&\overset{(i)}{\leq} 1 + 1.5 \frac{n\lambda_{m+1}\|\mathbf{K}_{>m}\|}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2} \leq 1 + 1.5 \frac{n\lambda_{m+1}\left(\|\mathbf{K}_{>m}\| + \lambda\right)}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2} \\
&\leq 1 + 1.5 \frac{n\lambda_{m+1}}{\max\{\lambda, 1\}} \frac{\max\{\lambda, 1\}^2}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2} \frac{\|\mathbf{K}_{>m}\| + \lambda}{\max\{\lambda, 1\}} \\
&\leq 1 + 1.5 \frac{r_2}{r_1^2} \frac{n\lambda_{m+1}}{\max\{\lambda, 1\}} \\
&\overset{(ii)}{\leq} \left(1 + 1.5 \frac{r_2}{r_1^2}\right) \max\left\{\frac{n\lambda_{m+1}}{\max\{\lambda, 1\}}, 1\right\} \qquad (9) \\
&\overset{(iii)}{=} 3\left(1 + 1.5 \frac{r_2}{r_1^2}\right) \tau_2, \qquad (10)
\end{aligned}
$$

where (i) uses Equation (5) to bound $\|\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}/n\|$, and Lemma 7 to bound $\|\mathbf{S}_{>m}\|$. In (ii) we use the bound: $cx + 1 \leq (c+1)\max\{x, 1\}$. In (iii) we just see that the terms matches the definition of τ_2.

Hence, to conclude the bias bound, we need to bound B_1 from above and below.

Upper bound:

$$
\begin{aligned}
B_1 &\leq \left\|\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right\|^2 \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2} \\
&\leq \frac{\|\mathbf{H}_{>m}\|^2}{\mu_{\min}\left(\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}}{n}\right)^2} \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2} \\
&\overset{(i)}{\leq} 4\left(\|\mathbf{K}_{>m}\| + \lambda\right)^2 \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2} \\
&= 4r_2^2 \max\{\lambda, 1\}^2 \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2} \qquad (11)
\end{aligned}
$$

where (i) follows from Equation (5). This concludes the upper bound on the bias since we can compute:

$$
\textbf{Bias}^2 \leq B_1 + B_2 \leq (1 + C_1) B_1 \overset{(i)}{\leq} 3\left(1 + 1.5 \frac{r_2}{r_1^2}\right) \tau_2 B_1 \overset{(ii)}{\leq} 12\left(1 + 1.5 \frac{r_2}{r_1^2}\right) r_2^2 \tau_2 \max\{\lambda, 1\}^2 \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2},
$$

where in (i) we use the bound in Equation (10), and in (ii) we used Equation (11). This concludes the upper bound on the bias.

Lower bound:

$$
\begin{aligned}
B_1 &\geq \mu_{\min}\left(\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2} \\
&\geq \frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{\left(\frac{1}{n\lambda_m} + \frac{1}{\mu_{\min}(\mathbf{H}_{>m})}\left\|\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{\Psi}_{\leq m}}{n}\right\|\right)^2 n^2} \\
&\overset{(i)}{\geq} \left(\frac{\frac{n\lambda_m}{\max\{\lambda,1\}}}{1.5\frac{\max\{\lambda,1\}}{\mu_{\min}(\mathbf{K}_{>m})+\lambda}\frac{n\lambda_m}{\max\{\lambda,1\}}+1}\right)^2 \max\{\lambda,1\}^2\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2} \\
&\overset{(ii)}{\geq} \left(\frac{r_1}{1.5+r_1}\right)^2 \min\left\{\frac{n\lambda_m}{\max\{\lambda,1\}},1\right\}^2 \max\{\lambda,1\}^2\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2}, \\
&\overset{(iii)}{=} \left(\frac{r_1}{1.5+r_1}\right)^2 \tau_1^2\max\{\lambda,1\}^2\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2},
\end{aligned}
$$

where (i) follows from Equation (5), (ii) from the fact that $\frac{1}{1+cx} \leq \frac{1}{1+c}\min\{x,1\}$, and (iii) the definition of τ_1. Since **Bias**$^2 \geq B_1$, this concludes the lower bound for the bias.

Variance: As for the bias bound, we first apply Lemma 5 to write $\mathbf{S} = \mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^{\intercal} + \mathbf{S}_{>m}$, and decompose the variance in Equation (7) as

$$
\textbf{Variance}(\hat{f}_\lambda)/\sigma^2 = \underbrace{Tr\left(\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-1}\right)}_{V_1} + \underbrace{\mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\right)}_{V_2}. \tag{12}
$$

Next, we rewrite V_1 as follows:

$$
\begin{aligned}
V_1 &= \mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-1}\right) = \mathbf{Tr}\left(\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\right) \\
&\overset{(i)}{=} \mathbf{Tr}\left(\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\right) \\
&= \frac{1}{n}\mathbf{Tr}\left(\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}}{n}\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right).
\end{aligned}
$$

Step (i) follows from Lemma 6.

To bound V_1 and V_2, we use the fact that the trace is the sum of all eigenvalues. Therefore, the trace is bounded from above and below by the size of the matrix times the largest and smallest eigenvalue,

respectively. This yields the following bound for V_1:

$$
\begin{aligned}
V_1 &\leq \frac{m}{n}\left\|\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n}+\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}}{n}\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n}+\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right\| \\
&\overset{(i)}{\leq} \frac{m}{n}4r_2^2\max\{\lambda,1\}^2\frac{\|\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{\Psi}_{\leq m}/n\|}{\mu_{\min}\left(\mathbf{H}_{>m}\right)^2} \overset{(ii)}{\leq} 6\frac{m}{n}r_2^2\frac{\max\{\lambda,1\}}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right)+\lambda\right)^2} \\
&= 6\left(\frac{r_2}{r_1}\right)^2\frac{m}{n}, \\
V_1 &\geq \frac{m}{n}\mu_{\min}\left(\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n}+\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}}{n}\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n}+\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right) \\
&\overset{(iii)}{\geq} \frac{m}{n}\left(\frac{r_1}{1.5+r_1}\right)^2\tau_1^2\max\{\lambda,1\}^2\frac{\mu_{\min}\left(\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{\Psi}_{\leq m}/n\right)}{\left(\|\mathbf{K}_{>m}\|+\lambda\right)^2} \\
&\overset{(iv)}{\geq} \frac{1}{2}\frac{m}{n}\left(\frac{r_1}{1.5+r_1}\right)^2\tau_1^2\frac{\max\{\lambda,1\}^2}{\left(\|\mathbf{K}_{>m}\|+\lambda\right)^2} = \frac{1}{2}\frac{m}{n}\left(\frac{r_1}{r_2(1.5+r_1)}\right)^2\tau_1^2
\end{aligned}
$$

For (i) and (iii), we bound the terms analogously to the upper and lower bound of B_1, and use Equation (5) in (ii) and (iv).

Next, we bound V_2 as

$$
\begin{aligned}
\mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\right) &\leq \mathbf{Tr}\left(\mathbf{S}_{>m}\right)\|\mathbf{H}^{-2}\| = \frac{\mathbf{Tr}\left(\mathbf{S}_{>m}\right)}{\mu_{\min}\left(\mathbf{K}_{\leq m}+\mathbf{H}_{>m}\right)^2}, \\
\mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\right) &\geq \mathbf{Tr}\left(\mathbf{S}_{>m}\right)\mu_{\min}\left(\mathbf{H}^{-2}\right) = \frac{\mathbf{Tr}\left(\mathbf{S}_{>m}\right)}{(\|\mathbf{K}_{>m}\|+\lambda)^2}.
\end{aligned}
$$

This concludes the bound for the variance, and hence the proof. □

A.1 Additional lemmas

Lemma 5 (Squared kernel decomposition)**.** *Let $\mathcal{K}$ be a kernel function that under a distribution ν can be decomposed as $\mathcal{K}(x,x') = \sum_k \lambda_k\psi_k(x)\psi_k(x')$, where $\mathbb{E}_{x\sim\nu}[\psi_k(x)\psi_{k'}(x)] = \delta_{k,k'}$. Then the squared kernel defined as $\mathcal{S}(x,x') = \mathbb{E}_{z\sim\nu}[\mathcal{K}(x,z)\mathcal{K}(x',z)]$ can be written as*

$$
\mathcal{S}(x,x') = \sum_k \lambda_k^2\psi_k(x)\psi_k(x'),
$$

and the kernel matrix $\mathbf{S}$ can be written as $\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^{\intercal} + \mathbf{S}_{>m}$.

Proof. The statement simply follows from

$$
\mathcal{S}(x,x') = \sum_{k,k'}\lambda_k\lambda_{k'}\psi_k(x)\underbrace{\mathbb{E}_{z\sim\nu}\left[\psi_k(z)\psi_{k'}(z)\right]}_{\delta_{k,k'}}\psi_{k'}(x') = \sum_k \lambda_k^2\psi_k(x)\psi_k(x').
$$

□

Lemma 6 (Corollary of Lemma 20 in Bartlett et al. (2020))**.**

$$
\begin{aligned}
\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m} = &\left(\mathbf{D}_{\leq m}^{-1}+\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m} \\
&\left(\mathbf{D}_{\leq m}^{-1}+\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}
\end{aligned}
$$

Proof.

$$\begin{aligned}
\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m} &= \mathbf{D}_{\leq m}^{1/2}\mathbf{D}_{\leq m}^{1/2}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^{1/2}\mathbf{D}_{\leq m}^{1/2} \\
&\overset{(i)}{=} \mathbf{D}_{\leq m}^{1/2}\left(\mathbf{I}_m + \mathbf{D}_{\leq m}^{1/2}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^{1/2}\right)^{-1}\mathbf{D}_{\leq m}^{1/2}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^{1/2} \\
&\qquad \left(\mathbf{I}_m + \mathbf{D}_{\leq m}^{1/2}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^{1/2}\right)^{-1}\mathbf{D}_{\leq m}^{1/2} \\
&= \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1},
\end{aligned}$$

where (i) applies Lemma 20 from Bartlett et al. (2020). □

Lemma 7 (Squared kernel tail)**.** *Let $\mathbf{S}_{>m}$ be the kernel matrix of the squared kernel $\mathcal{S}$. Then,*

$$\|\mathbf{S}_{>m}\| \leq \lambda_{m+1}\|\mathbf{K}_{>m}\|.$$

Proof. We show that for any vector v, $v^{\intercal}\mathbf{S}_{>m}v \leq \lambda_{m+1}v^{\intercal}\mathbf{K}_{>m}v$, which is sufficient to prove the lemma. To do so, we define $\mathbf{\Psi}_k \in \mathbb{R}^{n\times n}$ with $[\mathbf{\Psi}_k]_{i,j} = \psi_k(x_i)\psi_k(x_j)$. Then we can write $\mathbf{K}_{>m} = \sum_{k>m}\lambda_k\mathbf{\Psi}_k$ and $\mathbf{S}_{>m} = \sum_{k>m}\lambda_k^2\mathbf{\Psi}_k$. Since the eigenvalues are in decreasing order, we have $\lambda_k \leq \lambda_{m+1}$ for any $k > m$, and thus

$$v^{\intercal}\mathbf{S}_{>m}v = \sum_{k>m}\lambda_k^2 v^{\intercal}\mathbf{\Psi}_k v \leq \lambda_{m+1}\sum_{k>m}\lambda_k v^{\intercal}\mathbf{\Psi}_k v = \lambda_{m+1}v^{\intercal}\mathbf{K}_{>m}v.$$

□

B CONVOLUTIONAL KERNELS ON THE HYPERCUBE

First, Appendix B.1 provides a way to decompose general functions for features distributed uniformly on the hypercube. Next, Appendix B.2 uses those results to characterize the eigenfunctions and eigenvalues of cyclic convolutional kernels. Finally, Appendices B.3 and B.4 apply this characterization to prove Lemmas 2 and 4, respectively.

B.1 GENERAL FUNCTIONS ON THE HYPERCUBE

This subsection focuses on the main setting in our paper: the hypercube domain $\{-1,1\}^d$ together with the uniform probability density, previously studied in Misiakiewicz & Mei (2021). For any $S \subseteq \{1,\dots,d\}$, we define the polynomial

$$\mathcal{Y}_S(x) \coloneqq \prod_{j\in S}[x]_j \tag{13}$$

of degree $|S|$, where $[x]_j$ is the j-th entry of $x \in \{-1,1\}^d$. It is easy to see that $\{\mathcal{Y}_S\}_{S\subseteq\{1,\dots,d\}}$ is set of orthonormal functions with respect to the inner product $\langle f,g\rangle_{\{-1,1\}^d} \coloneqq \mathbb{E}_{x\sim\mathcal{U}(\{-1,1\}^d)}[f(x)g(x)]$. Those functions play a key role in the remainder of our proof; as it turns out, they are the eigenfunctions of the kernel in Section 3.1. Towards a formal statement, define the polynomials

$$\mathcal{G}_l^{(d)}\left(\frac{\langle x,x'\rangle}{\sqrt{d}}\right) \coloneqq \frac{1}{\mathcal{B}(l,d)}\sum_{|S|=l}\mathcal{Y}_S(x)\mathcal{Y}_S(x'), \tag{14}$$

$$\text{where } \mathcal{B}(l,d) \coloneqq |\{S\subseteq\{1,\dots,d\} \mid |S|=l\}| = \binom{d}{l}. \tag{15}$$

Notice that $\mathcal{G}_l^{(d)}$ only depends on the (Euclidean) inner product of x and x'. Furthermore, $\{\mathcal{G}_l^{(d)}\}_{l=0}^d$ is a set of orthonormal polynomials with respect to the distribution of $\langle x,x'\rangle/\sqrt{d}$. The following lemma shows how such polynomials form an eigenbasis for functions that only depend on the inner product between points in the unit hypercube.

Lemma 8 (Local kernel decomposition)**.** *Let $\kappa\colon \mathbb{R} \to \mathbb{R}$ be any function and $d \in \mathbb{N}_{>0}$. Then, for any $x, x' \in \{-1, 1\}^d$, we can decompose $\kappa(\langle x, x'\rangle/d)$ as*

$$\kappa\left(\frac{\langle x, x'\rangle}{d}\right) = \sum_{l=0}^{d} \xi_l^{(d)} \frac{1}{\mathcal{B}(l,d)} \sum_{|S|=l} \mathcal{Y}_S(x)\mathcal{Y}_S(x'). \tag{16}$$

Proof. Note that the decomposition only needs to hold at the evaluation of κ in the values that $\langle x, x'\rangle/d$ can take, that is, κ computed in $\{-1, -1+2/d, \dots, -2/d+1, 1\}$. Since that set has a cardinality $d+1$, we can write κ as a linear combination of $d+1$ uncorrelated functions. This is the case for $\{\mathcal{G}_l^{(d)}\}_{l=0}^d$ with respect to the distribution of $\langle x, x'\rangle/\sqrt{d}$ by construction. Hence,

$$\kappa\left(\frac{\langle x, x'\rangle}{d}\right) = \sum_{l=0}^{d} c_l \mathcal{G}_l^{(d)}\left(\frac{\langle x, x'\rangle}{\sqrt{d}}\right)$$

for some (unknown) coefficients c_l. Finally, the proof follows by expanding the definition of $\mathcal{G}_l^{(d)}$ in Equation (14) and choosing $\xi_l^{(d)} = c_l$. □

While this lemma is very general, the proof of Theorem 1 requires a mild regularity assumption on κ.

Assumption 1 (Regularity of κ)**.** *A function $\kappa\colon [-1, 1] \to \mathbb{R}$ in the setting of Section 3.1 satisfies the regularity assumption if there exist constants $c_1, c_2 > 0$ and a series of constants $\{c_l > 0\}_{l\geq 0}$ such that, for any $q \geq c_1$, the decomposition $\kappa\left(\langle x, x'\rangle/q\right) = \sum_{l=0}^{q} \xi_l^{(q)} \frac{1}{\mathcal{B}(l,q)} \sum_{|S|=l} \mathcal{Y}_S(x)\mathcal{Y}_S(x')$ from Lemma 8 over inputs $x, x' \in \{-1, 1\}^q$ satisfies*

$$\xi_l^{(q)} \geq c_l \quad \forall l, \qquad \sum_{l\geq 0} \xi_l^{(q)} \leq c_2.$$

For sufficiently high-dimensional inputs x, x', Assumption 1 ensures that the convolutional kernel $\mathcal{K}(x, x')$ in Equation (1) is a valid kernel, that it possesses the so-called "implicit regularization" property , and enables learning of polynomials of any degree. This assumption resembles Assumption 1 from Misiakiewicz & Mei (2021). However, our version is more general: we enable learning of polynomials of any degree, while Assumption 1 from Misiakiewicz & Mei (2021) only covers the degrees of interest. Still, our assumption holds for standard kernels such as the Gaussian kernel. Finally, our assumption does not depend on ℓ. Thus, a kernel satisfying our assumption will also satisfy Assumption 1 from Misiakiewicz & Mei (2021) for any choice of ℓ.

B.2 Convolutional kernels on the hypercube

While the previous subsection considers general functions on the hypercube, we now focus on convolutional kernels and their eigenvalues. This yields the tools to prove Lemmas 2 and 4. In order to characterize eigenvalues, we first introduce the following quantities, also used e.g. in Misiakiewicz & Mei (2021).

Definition 1 (Diameter)**.** *Let $S \subseteq \{1, \dots, d\}$. The diameter of S is*

$$\gamma(S) \coloneqq \max_{i,j\in S} \min\left\{\mathrm{mod}\left(j-i, d\right)+1, \mathrm{mod}\left(i-j, d\right)+1\right\}$$

for $S \neq \emptyset$, and $\gamma(\emptyset) = 0$. Furthermore, we define

$$\mathcal{C}(l, q, d) \coloneqq |\{S \subseteq \{1, \dots, d\} \mid |S| = l, \gamma(S) \leq q\}|. \tag{17}$$

Intuitively, the diameter of S is the smallest number of contiguous feature indices that fully contain S. The following lemma yields an explicit formula for $\mathcal{C}(l, q, d)$, that is, the number of sets of size l with diameter at most q.

Lemma 9 (Number of overlapping sets)**.** *Let $l, q, d \in \mathbb{N}$ with $q \geq 0$ and $l \leq q < d/2$. Then,*

$$\mathcal{C}(l, q, d) = \begin{cases} d\binom{q-1}{l-1} & l > 0, \\ 1 & l = 0 \text{ or } q = 0. \end{cases}$$

Proof. Since the result holds trivially for $q = 0$, $l = 0$, and $l = 1$, we henceforth focus on $l \geq 2$. Let $\tilde{\mathcal{C}}(l, \gamma, d)$ be the number of subsets $S \subseteq \{1, \ldots, d\}$ of cardinality $|S| = l$ with diameter exactly $\gamma(S) = \gamma$. First, consider $\tilde{\mathcal{C}}(2, \gamma, d)$. For each set, we can choose the first element i from d different values, and the second as $\mathrm{mod}\,((i-1) \pm (\gamma - 1), d) + 1$. In this way, since $q < d/2$, we count each set exactly twice. Thus,

$$\tilde{\mathcal{C}}(2, \gamma, d) = \frac{d \cdot 2}{2} = d.$$

Next, consider $l > 2$. We can build all the possible sets by starting with one of the $\tilde{\mathcal{C}}(2, \gamma, d) = d$ sets, and adding the remaining $l - 2$ elements from $\gamma - 2$ possible indices. Hence, every fixed set of size 2 and diameter γ yields $\binom{\gamma-2}{l-2}$ different sets of size l. Furthermore, by construction, every set of size l and diameter γ results from exactly on set of size 2 and diameter γ. Therefore,

$$\tilde{\mathcal{C}}(l, \gamma, d) = d\binom{\gamma - 2}{l - 2}.$$

The result for $l \geq 2$ then follows from summing $\tilde{\mathcal{C}}(l, \gamma, d)$ over all diameters $\gamma \leq q$:

$$\mathcal{C}(l, q, d) = d\sum_{\gamma=l}^{q}\binom{\gamma - 2}{l - 2} \overset{(i)}{=} d\binom{q-1}{l-1},$$

where (i) follows from the hockey-stick identity. □

Now, we focus on cyclic convolutional kernels $\mathcal{K}$ as in Equation (1). First, we restate Proposition 1 from Misiakiewicz & Mei (2021). This proposition establishes that $\mathcal{Y}_S$ for $S \subseteq \{1, \ldots, d\}$ are indeed the eigenfunctions of $\mathcal{K}$, and yields closed-form eigenvalues λ_S up to factors $\xi_{|S|}^{(q)}$ that depend on the inner nonlinearity κ. Then, Lemma 10 uses the regularity assumptions on κ to eliminate the dependency on $\xi_{|S|}^{(q)}$. This characterization of the eigenvalues enables the proof Lemmas 2 and 4.

Proposition 1 (Proposition 1 from Misiakiewicz & Mei (2021))**.** *Let $\mathcal{K}$ be a cyclic convolutional kernel over the unit hypercube as defined in Equation* (1)*. Then we can write*

$$\mathcal{K}(x, x') \coloneqq \frac{1}{d}\sum_{k=1}^{d}\kappa\left(\frac{\langle x_{(k,q)}, {x'}_{(k,q)}\rangle}{q}\right) = \sum_{l=1}^{q}\sum_{\substack{\gamma(S)\leq q \\ |S|=l}}\lambda_S\mathcal{Y}_S(x)\mathcal{Y}_S(x'),$$

with

$$\lambda_S = \xi_{|S|}^{(q)}\frac{q + 1 - \gamma(S)}{d\mathcal{B}(|S|, q)}$$

where $\xi_{|S|}^{(q)}$ are the coefficients of the κ-decomposition (Equation (16)*) over $\{-1, 1\}^{|S|}$. Alternatively,*

$$\mathcal{K}(x, x') = \sum_{k}\lambda_{S_k}\mathcal{Y}_{S_k}(x)\mathcal{Y}_{S_k}(x')$$

where we order all subsets $S_k \subseteq \{1, \ldots, d\}$ with $\gamma(S_k) \leq q$ such that $\lambda_{S_k} \geq \lambda_{S_{k+1}}$. In particular, $\lambda_{S_k} = \lambda_k$.

We refer to Proposition 1 from Misiakiewicz & Mei (2021) for a formal proof. Intuitively, the result follows from applying Lemma 8 for each subset $S \subseteq \{1, \ldots, d\}$. This is possible because crucially, any subset of q features is again distributed uniformly on the q-dimensional unit hypercube. Lastly, the factor $q + 1 - \gamma(S)$ stems from the fact that each eigenfunction $\mathcal{Y}_S$ appears as many times as there are contiguous index sets of size $\gamma(S)$ supported in a fixed contiguous index set of size q. In other words, the term is the number of shifted instances of S supported in a contiguous subset of q features.

While Proposition 1 yields eigenfunctions and eigenvalues of cyclic convolutional kernels, it does not describe the factors $\xi_{|S|}^{(q)}$ which depend on the inner nonlinearity κ. The final lemma of this section is a corollary that characterizes the eigenvalues solely in terms of $|S|$, and further shows that, for d large enough, the eigenvalues decay as $|S|$ grows.

Lemma 10 (Corollary of Proposition 1)**.** *Consider a cyclic convolutional kernel from Proposition 1 with $q \in \Theta(d^\beta)$ for some $\beta \in (0,1)$. Assume that κ satisfies Assumption 1 with $c_1 \leq q$. Then, for any $S \subseteq \{1,\dots,d\}$ such that $\gamma(S) \leq q$, the eigenvalue λ_S corresponding to the eigenfunction $\mathcal{Y}_S(x)$ satisfies*

$$\frac{1}{d \cdot q^{|S|}} \lesssim \lambda_S \lesssim \frac{1}{d \cdot q^{|S|-1}},$$

where $\lesssim$ hides positive constants that depend on $|S|$. Furthermore, there exists an $S \subseteq \{1,\dots,d\}$ such that $\gamma(S) \leq q$ for which $\lambda_S \in \Theta\left(1/d^{1+\beta|S|}\right)$.

Proof. Let $l \coloneqq |S|$ and $r \coloneqq q+1-\gamma(S)$. Since $l \leq \gamma(S) \leq q$, we have

$$1 \leq r \leq q+1-l.$$

Furthermore, because $\mathcal{B}(l,q) = \binom{q}{l}$, the classical bound on the binomial coefficient yields

$$\left(\frac{1}{l}\right)^l q^l \leq \mathcal{B}(l,q) \leq \left(\frac{e}{l}\right)^l q^l.$$

Hence,

$$\begin{aligned}\lambda_S = \frac{\xi_l^{(q)} r}{d\mathcal{B}(l,q)} &\overset{(i)}{\leq} \frac{l^l}{q^l}\frac{c_2 r}{d} \leq c_{l,1}\frac{1}{dq^{l-1}},\\ &\overset{(ii)}{\geq} \frac{l^l}{q^l e^l}\frac{c_l r}{d} \geq c_{l,2}\frac{1}{dq^l},\end{aligned}$$

where Assumption 1 yields in (i) that $\xi_l^{(q)} \leq c_2$, and in (ii) that $\xi_l^{(q)} \geq c_l$. Since $c_{l,1}$ and $c_{l,2}$ only depend on l, this concludes the first part of the proof.

For the second part, any S with $\gamma(S) = q$ yields $r = 1$. Hence, the previous upper bound becomes

$$\lambda_S = \frac{\xi_l^{(q)} r}{d\mathcal{B}(l,q)} \leq \frac{l^l}{q^l}\frac{c_2 r}{d} \leq c_{l,3}\frac{1}{dq^l},$$

where $c_{l,3}$ only depends on l. This yields the second statement, since $dq^l \in \Theta\left(d^{1+\beta|S|}\right)$. □

Finally, using Proposition 1 and Lemma 10, we are able to prove Lemmas 2 and 4.

B.3 Proof of Lemma 2

First, notice that $f^\star = \mathcal{Y}_{S^*}$ for $S^* = \{1,\dots,L^*\}$. Since $|S^*| = \gamma(S^*) = L^*$, applying Proposition 1 yields

$$\lambda_{S^*} = \xi_{L^*}^{(q)} \frac{q+1-L^*}{d\mathcal{B}(L^*,q)} \overset{(i)}{\in} \Theta\left(\frac{q}{dq^{L^*}}\right) \overset{(ii)}{=} \Theta\left(\frac{1}{dq^{L^*-1}}\right) \subseteq \omega(\lambda_m),$$

where in (i) we used assumption Assumption 1 to get that $\xi_{L^*}^{(q)} \in \Theta(1)$, and in (ii) we used the fact that $\lambda_m \in o\left(\frac{1}{dq^{L^*-1}}\right)$. This implies that for d sufficiently large, $\lambda_{S^*} > \lambda_m$. Thus, since the eigenvalues are sorted, $f^\star$ is in the span of the first m eigenfunctions. Finally,

$$\|\mathbf{D}_{\leq m}^{-1} a\| \overset{(i)}{=} \lambda_{S^*}^{-1} \in \Theta\left(dq^{L^*-1}\right),$$

since in (i), the entry of a corresponding to $\mathcal{Y}_{S^*}$ is 1, and all others are 0.

B.4 Proof of Lemma 4

Before proving Lemma 4, we introduce the following quantity:

$$L \coloneqq \left\lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \right\rfloor. \tag{18}$$

Intuitively, L corresponds to the degree of the largest polynomial that a cyclic convolutional kernel as defined in Equation (1) can learn. This quantity plays a key role throughout the proof of Lemma 4, and Lemma 3 later. All eigenfunctions of degree at most L will have an index smaller than m, while all eigenfunctions of degree at least $L+2$ will have an index larger than m. The proof of Lemma 4 shows this statement formally.

Proof of Lemma 4. First, we use Proposition 1 to write the cyclic convolutional kernel as

$$\mathcal{K}(x,x') = \sum_{l=1}^{q} \sum_{\substack{\gamma(S)\leq q \\ |S|=l}} \lambda_S \mathcal{Y}_S(x)\mathcal{Y}_S(x') = \sum_k \lambda_{S_k}\mathcal{Y}_{S_k}(x)\mathcal{Y}_{S_k}(x')$$

where S_i are ordered such that $\lambda_{S_{k+1}} \leq \lambda_{S_k}$.

Here we need to show that we can pick $m \in \mathbb{N}$ such that $n\lambda_{S_m} \in \Theta\left(\max\{\lambda,1\}\right)$, and $m \in \Theta\left(\frac{nq^{-\delta}}{\max\{\lambda,1\}}\right)$. Since $n \in \Theta(d^\ell)$ and $\max\{\lambda,1\} \in \Theta(d^{\ell_\lambda})$, it suffices to show that there exists a m such that

$$\lambda_m \in \Theta\left(\frac{\max\{\lambda,1\}}{n}\right) = \Theta\left(\frac{d^{\ell_\lambda}}{d^\ell}\right) = \Theta\left(\frac{d^{\ell_\lambda}}{d^{1+\ell_\lambda+\beta(L+\delta)}}\right) = \Theta\left(\frac{1}{d\cdot q^{L+\delta}}\right), \tag{19}$$

$$m \in \Theta\left(\frac{nq^{-\delta}}{\max\{\lambda,1\}}\right) = \Theta\left(\frac{d^\ell q^{-\delta}}{d^{\ell_\lambda}}\right) = \Theta\left(\frac{d^{1+\ell_\lambda+\beta(L+\delta)}q^{-\delta}}{d^{\ell_\lambda}}\right) = \Theta(dq^L). \tag{20}$$

By definition of $\lambda_m = \lambda_{S_m}$, this corresponds to showing that there exists $S_m \subseteq \{1,\dots,d\}$ with $\gamma(S_m) \leq q$ such that the eigenvalue corresponding to $\mathcal{Y}_{S_m}$ is $\lambda_{S_m} \in \Theta\left(\frac{1}{d\cdot q^{L+\delta}}\right)$, and $m \in \Theta\left(dq^L\right)$.

Choice of m. The way we find such S_m is different in two cases:

$$S_m = \begin{cases} \{1,\dots,L,\lfloor q+1-q^{1-\delta}\rfloor\} & \delta \in (0,1) \\ \{1,\dots,L,\lfloor q/2\rfloor\} & \delta = 0. \end{cases} \tag{21}$$

First, in both cases, since $\delta < 1$, S_m is well-defined for d— and hence $q \in \Theta(d^\beta)$— sufficiently large. Assume that this is the case. Then, $|S_m| = L+1$ in both cases and

$$\gamma(S_m) = \begin{cases} \lfloor q+1-q^{1-\delta}\rfloor & \delta \in (0,1) \\ \lfloor q/2\rfloor & \delta = 0, \end{cases}$$

by definition.

Rate of λ_{S_m} Using Proposition 1, since we know that $|S_m| = L+1$, we can write the eigenvalue corresponding to $\mathcal{Y}_{S_m}$ as

$$\lambda_{S_m} = \xi^{(q)}_{|S_m|}\frac{q+1-\gamma(S_m)}{d\mathcal{B}(|S_m|,q)} = \xi^{(q)}_{L+1}\frac{q+1-\lfloor q+1-q^{1-\delta}\rfloor}{d\mathcal{B}(L+1,q)}.$$

Numerator: Now we show that in both cases, the numerator is in $\Theta\left(q^{1-\delta}\right)$. In the case where $\delta \in (0,1)$:

$$q+1-\lfloor q+1-q^{-(1-\delta)}\rfloor = -\lfloor -q^{1-\delta}\rfloor = \lceil q^{1-\delta}\rceil \overset{(i)}{\in} \Theta\left(q^{1-\delta}\right),$$

where (i) follows from $\delta < 1$ and q sufficiently large. In the case where $\delta = 0$:

$$q+1-\lfloor q/2\rfloor \leq q+1 \in \mathcal{O}(q)$$
$$q+1-\lfloor q/2\rfloor \geq q/2 \in \Omega(q).$$

Thus, the numerator is in $\Theta(q) = \Theta(q^{1-\delta})$, since in this setting $\delta = 0$.

Denominator: Notice that the denominator does not depend on δ, thus we use the same technique for both $\delta = 0, \delta \in (0,1)$. Indeed, we use the classical bound on $\mathcal{B}(L+1,q) = \binom{q}{L+1}$:

$$q^{L+1} \lesssim \left(\frac{q}{L+1}\right)^{L+1} \leq \binom{q}{L+1} \leq e^{L+1}\left(\frac{q}{L+1}\right)^{L+1} \lesssim q^{L+1}.$$

Rate of λ_{S_m}: Using both parts, together with $\xi_{L+1}^{(q)} \in \Theta(1)$ by Assumption 1, we can conclude that S_m satisfies

$$\lambda_{S_m} \in \Theta\left(\frac{q^{1-\delta}}{dq^{L+1}}\right) = \Theta\left(\frac{1}{d \cdot q^{L+\delta}}\right).$$

Hence, thanks to Equation (19), we have shown that there exists an m for which $n\lambda_m \in \Theta(\max\{\lambda, 1\})$. We now need to show that for this choice of m, we also have that $m \in \Theta\left(\frac{nq^{-\delta}}{\max\{\lambda,1\}}\right) = \Theta(dq^L)$.

Rate of m. *Upper bound:* From the construction of S_m in Equation (21) it directly follows that $|S_m| = L + 1$.

Furthermore, Lemma 10 shows that, for d sufficiently large, all S_k with $|S_k| > L + 1$ correspond to eigenvalues $\lambda_{S_k} < \lambda_{S_m}$. Thus, m is at most the number of polynomials with degree no larger than $L + 1$:

$$m \leq \sum_{l=0}^{L+1} \mathcal{C}(l, q, d) \overset{(i)}{\in} \mathcal{O}(dq^L),$$

where (i) uses Lemma 9 for d large enough with $\binom{q}{l} \in \Theta(q^l)$.

Lower bound: From the definition of S_m in Equation (21), we have that $\gamma(S_m) \geq \lfloor q/2 \rfloor$. This, combined with Proposition 1, yield that all the indexes of polynomials of degree $L + 1$, but diameter $\leq \lfloor q/2 \rfloor - 1$ appear before m. Hence, thanks to Lemma 9,

$$m \geq \mathcal{C}(L + 1, \lfloor q/2 \rfloor - 1, d) = d\binom{\lfloor q/2 \rfloor - 2}{L} \geq d\left(\frac{\lfloor q/2 \rfloor - 2}{L}\right)^L \in \Omega(dq^L),$$

concluding the proof of the lower bound. Thus, $m \in \Theta(d^L)$, and, thanks to Equation (20), $m \in \Theta\left(\frac{nq^{-\delta}}{\max\{\lambda,1\}}\right)$. This concludes the proof of the existence. To conclude the proof we only need to show additional formulas for the rates.

Rate of m with respect to n. Finally, we only need to show that $\Theta(dq^L) \subseteq \mathcal{O}(nq^{-\delta})$. For this we notice that, since $\bar{L} := \left\lfloor \frac{\ell-1}{\beta} \right\rfloor \geq L$, we have that $m \in \mathcal{O}(dq^{\bar{L}})$. Then, we can write n as

$$n \in \Theta(d^\ell) = \Theta\left(d \cdot d^{\beta\frac{\ell-1}{\beta}}\right) = \Theta\left(dq^{\bar{L}+\bar{\delta}}\right), \tag{22}$$

and hence $m \in \mathcal{O}(dq^{\bar{L}}) = \mathcal{O}(nq^{-\bar{\delta}})$.

Rate of λ_m with respect to L^*. By assumption, $n\lambda_m \in \Theta(\max\{\lambda, 1\})$. Therefore,

$$\lambda_m \in \Theta\left(\frac{\max\{\lambda, 1\}}{n}\right) \overset{(i)}{=} \Theta\left(\frac{d^{\ell_\lambda}}{d \cdot d^{\ell_\lambda} q^{L+\delta}}\right) = \Theta\left(\frac{1}{dq^{L+\delta}}\right),$$

where (i) uses the identity $\ell = 1 + \ell_\lambda + \beta(L + \delta)$. For the remainder, we need to consider two cases depending on δ. If $\delta > 0$, then $L^* \leq L + 1$, and we have

$$\lambda_m \in \mathcal{O}\left(\frac{q^{-\delta}}{dq^{L^*-1}}\right) \subseteq o\left(\frac{1}{dq^{L^*-1}}\right).$$

If $\delta = 0$, then $L^* \leq L$, and we have

$$\lambda_m \in \mathcal{O}\left(\frac{q^{-\delta}}{dq^L}\right) \subseteq \mathcal{O}\left(\frac{1}{dq^{L^*}}\right) \subseteq o\left(\frac{1}{dq^{L^*-1}}\right).$$

In both cases, $\lambda_m \in o\left(\frac{1}{dq^{L^*-1}}\right)$, concluding the proof. □

C MATRIX CONCENTRATION

This section focuses on random matrix theory. First, Appendix C.1 focuses on the large eigenvalues of our kernel, and proves Lemma 1. Next, Appendix C.2 focuses on the tail of the eigenvalues, culminating in the proof of Lemma 3 Lastly, Appendix C.4 establishes some technical tools that we use throughout the proofs.

C.1 PROOF OF LEMMA 1

In this proof, we show that the matrix $\mathbf{\Psi}_{\leq m}^{\mathsf{T}}\mathbf{\Psi}_{\leq m}/n$ concentrates around the identity matrix under appropriate assumptions, thereby establishing Equation (5).

Let us define $\hat{m}$ to be the largest m such that $m \in \mathcal{O}(nd^{-\beta\bar{\delta}})$. The proof consists of applying Theorem 5.44 from Vershynin (2012) to the matrix $\mathbf{\Psi}_{\leq \hat{m}}$. Then a simple argument extend it to any other choice of m.

Bound for $\hat{m}$. To apply Theorem 5.44 from Vershynin (2012) we need to verify the theorem's conditions on the rows of $\mathbf{\Psi}_{\leq \hat{m}}$. In particular, we show that the rows are independent, have a common second moment matrix, and that their norm is bounded. Let $[\mathbf{\Psi}_{\leq \hat{m}}]_{i,:}$ indicate the i-th row of $\mathbf{\Psi}_{\leq \hat{m}} \in \mathbb{R}^{n\times \hat{m}}$. We may write each row entry-wise as

$$[\mathbf{\Psi}_{\leq \hat{m}}]_{i,:} = [\mathcal{Y}_1(x_i)\ \mathcal{Y}_2(x_i)\ \cdots\ \mathcal{Y}_{\hat{m}}(x_i)]^{\mathsf{T}}.$$

First, the rows of $\mathbf{\Psi}_{\leq \hat{m}}$ are independent since each row depends on a different x_i, and we assume data to be i.i.d..

Second, since the eigenfunctions are orthonormal w.r.t. the data distribution, the second moment of the rows is $\mathbb{E}\left[[\mathbf{\Psi}_{\leq \hat{m}}]_{i,:}[\mathbf{\Psi}_{\leq \hat{m}}]_{i,:}^{\mathsf{T}}\right] = \mathbf{I}_{\hat{m}}$ for all rows.

Third, to show that each row has bounded norm, we use the fact that the eigenfunctions $\mathcal{Y}_k$ in Equation (13) over $\{-1,1\}^d$ satisfy $\mathcal{Y}_k(x_i)^2 = 1$ for all k. Thus, the norm of each row is

$$\|[\mathbf{\Psi}_{\leq \hat{m}}]_{i,:}\|_2 = \sqrt{\sum_{k=1}^{\hat{m}} \mathcal{Y}_k(x_i)^2} = \sqrt{\sum_{k=1}^{\hat{m}} 1} = \sqrt{\hat{m}}.$$

We can now apply Theorem 5.44 from Vershynin (2012). For any $t \geq 0$, this yields the following inequality with probability $1 - \hat{m}\exp\{-ct^2\}$:

$$\left\| \frac{\mathbf{\Psi}_{\leq \hat{m}}^{\mathsf{T}}\mathbf{\Psi}_{\leq \hat{m}}}{n} - \mathbf{I}_{\hat{m}} \right\| \leq \max\left\{\|\mathbf{I}_{\hat{m}}\|\Delta, \Delta^2\right\}, \quad \text{where } \Delta = t\sqrt{\frac{\hat{m}}{n}}.$$

Choosing $t = \frac{1}{2}\sqrt{\frac{n}{\hat{m}}}$ yields $\Delta = 1/2 < 1$, and hence $\max\left\{\|\mathbf{I}_{\hat{m}}\|\Delta, \Delta^2\right\} = 1/2$. Lastly, for d large enough, the error probability is

$$\hat{m}\exp\left\{-c\frac{n}{4\hat{m}}\right\} \overset{(i)}{\lesssim} nd^{-\beta\bar{\delta}}\exp\left\{-c'\frac{n}{nd^{-\beta\bar{\delta}}}\right\} \lesssim d^{-\beta\bar{\delta}} \cdot d^{\ell}\exp\{-c'd^{\beta\bar{\delta}}\} \lesssim d^{-\beta\bar{\delta}},$$

where (i) follows from the fact that $m \in \mathcal{O}(n^{\beta\bar{\delta}})$.

Bound for any $m < \hat{m}$. For this it suffices to notice that the matrix $\mathbf{\Psi}_{\leq m}^{\mathsf{T}}\mathbf{\Psi}_{\leq m}$ is a sub-matrix of $\mathbf{\Psi}_{\leq \hat{m}}^{\mathsf{T}}\mathbf{\Psi}_{\leq \hat{m}}$. Hence, also

$$\frac{\mathbf{\Psi}_{\leq m}^{\mathsf{T}}\mathbf{\Psi}_{\leq m}}{n} - \mathbf{I}_m \quad \text{is a sub-matrix of} \quad \frac{\mathbf{\Psi}_{\leq \hat{m}}^{\mathsf{T}}\mathbf{\Psi}_{\leq \hat{m}}}{n} - \mathbf{I}_{\hat{m}}.$$

Thus, uniformly over all $m \leq \hat{m}$,

$$\left\| \frac{\mathbf{\Psi}_{\leq m}^{\mathsf{T}}\mathbf{\Psi}_{\leq m}}{n} - \mathbf{I}_m \right\| \leq \left\| \frac{\mathbf{\Psi}_{\leq \hat{m}}^{\mathsf{T}}\mathbf{\Psi}_{\leq \hat{m}}}{n} - \mathbf{I}_{\hat{m}} \right\| \leq 1/2 \quad \text{w.p.} \geq 1 - cd^{-\beta\bar{\delta}},$$

concluding the proof.

C.2 CONCENTRATION OF THE TERMS AFTER m AND PROOF OF LEMMA 3

In the following, we focus on the concentration of the smallest and largest eigenvalue of the kernel matrix $\mathbf{K}_{>m}$ to prove Lemma 3. However, this proof is involved, and requires some additional tools. In particular, we further decompose $\mathcal{K}_{>m}$ into two kernels $\mathcal{K}_1$ and $\mathcal{K}_2$. We first introduce the two additional kernels, and then show in Lemma 11 that $\mathcal{K}_{>m} = \mathcal{K}_1 + \mathcal{K}_2$.

Definition 2 (1-2 decomposition)**.** *Consider the setting of Theorem 1, and decompose the convolutional kernel $\mathcal{K}$ via Proposition 1 as*

$$\mathcal{K}(x, x') = \sum_k \lambda_{S_k} \mathcal{Y}_{S_k}(x) \mathcal{Y}_{S_k}(x'), \tag{23}$$

where $\{S_k\}_{k>0}$ is a sequence of all subsets $S_k \subseteq \{1, \dots, d\}$ with $\gamma(S_k) \leq q$, ordered such that $\lambda_{S_k} \geq \lambda_{S_{k+1}}$. Lastly, let $m \in \mathbb{N}$ be such that $n\lambda_{S_m} \in \Theta(\max\{\lambda, 1\})$.

We define the index sets

$$\mathcal{I}_1 \coloneqq \{k \in \mathbb{N} \mid k > m \text{ and } |S_i| \leq \bar{L} + 1\},$$
$$\mathcal{I}_2 \coloneqq \{k \in \mathbb{N} \mid |S_k| \geq \bar{L} + 2\},$$

and the corresponding kernels

$$\mathcal{K}_1(x, x') \coloneqq \sum_{k \in \mathcal{I}_1} \lambda_{S_k} \mathcal{Y}_{S_k}(x) \mathcal{Y}_{S_k}(x'), \qquad \mathcal{S}_1(x, x') \coloneqq \sum_{k \in \mathcal{I}_1} \lambda_{S_k}^2 \mathcal{Y}_{S_k}(x) \mathcal{Y}_{S_k}(x')$$
$$\mathcal{K}_2(x, x') \coloneqq \sum_{k \in \mathcal{I}_2} \lambda_{S_k} \mathcal{Y}_{S_k}(x) \mathcal{Y}_{S_k}(x'), \qquad \mathcal{S}_2(x, x') \coloneqq \sum_{k \in \mathcal{I}_2} \lambda_{S_k}^2 \mathcal{Y}_{S_k}(x) \mathcal{Y}_{S_k}(x'),$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are the squared kernels corresponding to $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Furthermore, as in the original kernel decomposition, we define the matrices

$$\mathbf{\Psi}_1 \in \mathbb{R}^{n \times |\mathcal{I}_1|}, \qquad [\mathbf{\Psi}_1]_{i,j} = \mathcal{Y}_{S_{k_j}}(x_i),$$
$$\mathbf{D}_1 \in \mathbb{R}^{|\mathcal{I}_1| \times |\mathcal{I}_1|}, \qquad \mathbf{D}_1 = \operatorname{diag}(\lambda_{S_{k_1}}, \dots, \lambda_{S_{k_{|\mathcal{I}_1|}}}),$$

and the empirical kernel matrices $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n \times n}$ with

$$[\mathbf{K}_1]_{i,j} = \mathcal{K}_1(x_i, x_j) \quad \text{and} \quad [\mathbf{K}_2]_{i,j} = \mathcal{K}_2(x_i, x_j).$$

We further define $\bar{m}$ as the largest eigenvalue corresponding to an eigenfunction $\mathcal{Y}_S$ of degree $|S| \geq \bar{L} + 2$, that is,

$$\bar{m} \coloneqq \min \mathcal{I}_2.$$

Finally, we define

$$\mathcal{Q}_l^{(d,q)}(x, x') \coloneqq \sum_{\substack{\gamma(S) \leq q \\ |S| = l}} \frac{q + 1 - \gamma(S)}{d\mathcal{B}(l, q)} \mathcal{Y}_S(x) \mathcal{Y}_S(x')$$

with the corresponding kernel matrix $\mathbf{Q}_l^{(d,q)}$.

Intuitively, the quantities $\mathbf{\Psi}_1, \mathbf{D}_1$ are the equivalent to $\mathbf{\Psi}_{\leq m}, \mathbf{D}_{\leq m}$ in the original decomposition $\mathcal{K} = \mathcal{K}_{\leq m} + \mathcal{K}_{>m}$. The following lemma establishes that $\mathcal{K}_1$ and $\mathcal{K}_2$ indeed constitute a decomposition of $\mathcal{K}_{>m}$.

Lemma 11 (1-2 decomposition)**.** *In the same setting as Definition 2, for d sufficiently large, we have*

$$\mathcal{K}_{>m}(x, x') = \mathcal{K}_1(x, x') + \mathcal{K}_2(x, x') \quad \text{and} \quad \mathcal{S}_{>m}(x, x') = \mathcal{S}_1(x, x') + \mathcal{S}_2(x, x').$$

Proof. For the decomposition of $\mathcal{K}_{>m}$, we have to show that exactly the eigenfunctions with index larger than m appear in either $\mathcal{K}_1$ or $\mathcal{K}_2$, that is, $\mathcal{I}_1 \cup \mathcal{I}_2 = \{k > m\}$, and that no eigenfunction appears in both $\mathcal{K}_1$ or $\mathcal{K}_2$, that is, $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. Furthermore, since we can write $\mathcal{S}_{>m}(x, x') = \sum_{k>m} \lambda_{S_k}^2 \mathcal{Y}_{S_k}(x) \mathcal{Y}_{S_k}(x')$ by Lemma 5, the same argument also yields the 1-2 decomposition of $\mathcal{S}_{>m}$.

First, from the definition of $\mathcal{I}_1$ and $\mathcal{I}_2$, it follows directly that $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$, that $\mathcal{I}_1 \cup \mathcal{I}_2 \supseteq \{k > m\}$, and that $\mathcal{I}_1 \subseteq \{k > m\}$. Hence, to conclude the proof, we only need to show that $\mathcal{I}_2 \subseteq \{k > m\}$. Since the indices are sorted in decreasing order, we equivalently show that, for d sufficiently large, all eigenvalues λ_{S_k} with $k \in \mathcal{I}_2$ are smaller than $\lambda_{S_m} \in \Theta(\max\{\lambda, 1\}/n)$.

More precisely, we show that $n\lambda_{S_k} \in o(n\lambda_{S_m})$ for any $k \in \mathcal{I}_2$. Applying Lemma 10 yields

$$\begin{aligned} n\lambda_{S_k} &\in \mathcal{O}\left(\frac{n}{d^{1+\beta(|S_k|-1)}}\right) \overset{(i)}{\subseteq} \mathcal{O}\left(\frac{n}{d^{1+\beta(\bar{L}+2-1)}}\right) \\ &= \mathcal{O}\left(\frac{d^{1+\ell_\lambda+\beta(L+\delta)}}{d^{1+\beta(\bar{L}+\delta)}} d^{-\beta(1-\delta)}\right) \\ &= \mathcal{O}(\max\{\lambda, 1\} d^{\beta(L-\bar{L})} d^{-\beta(1-\delta)}) \overset{(ii)}{\subseteq} o(n\lambda_m), \end{aligned}$$

where (i) follows from $|S_k| \geq \bar{L} + 2$, and (ii) uses $\bar{L} \geq L$ and $\delta < 1$. Hence, for d sufficiently large, all $k \in \mathcal{I}_2$ have $\lambda_{S_k} < \lambda_m$ and therefore $k > m$ by definition. □

Using the 1-2 decomposition, we now prove Lemma 3. We defer the auxiliary Lemmas 12 to 15 after the proof, and more technical tools to Appendix C.4.

Proof of Lemma 3. We address the bound r_1 and r_2, and finally bound $\mathbf{Tr}(\mathbf{S}_{>m})$.

Throughout this proof, we write "with high probability" to indicate with probability at least $1 - cq^{-\min\{\bar{\delta}, 1-\bar{\delta}\}}$ for some context-dependent constant c. Furthermore, this proof does not establish the relevant matrix concentration results directly, but combines them from auxiliary lemmas. Furthermore, we use the m corresponding to S_m, and instantiate Definition 2 with that m. Then, Lemma 11 yields the 1-2 decomposition $\mathbf{K}_{>m} = \mathbf{K}_1 + \mathbf{K}_2$, which we will henceforth use.

Bound on r_1 and r_2 Remember the definition of r_1 and r_2:

$$r_1 = \frac{\mu_{\min}(\mathbf{K}_{>m}) + \lambda}{\max\{\lambda, \mathbf{I}\}}, \qquad r_2 \coloneqq \frac{\|\mathbf{K}_{>m}\| + \lambda}{\max\{\lambda, \mathbf{I}\}}.$$

To bound those quantities, we have to bound $\mu_{\min}(\mathbf{K}_{>m})$ and $\|\mathbf{K}_{>m}\|$. For the upper bound on $\|\mathbf{K}_{>m}\|$, we use the triangle inequality on $\|\mathbf{K}_{>m}\| = \|\mathbf{K}_1 + \mathbf{K}_2\|$, and then bound $\|\mathbf{K}_1\|$ and $\|\mathbf{K}_2\|$ individually.

Note that we can write $\mathbf{K}_1 = \mathbf{\Psi}_1 \mathbf{D}_1 \mathbf{\Psi}_1^\intercal$ by definition. Then,

$$\|\mathbf{K}_1\| = \|\mathbf{\Psi}_1 \mathbf{D}_1 \mathbf{\Psi}_1^\intercal\| = n\|\mathbf{D}_1\| \left\| \frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n} \right\| \overset{(i)}{\leq} 1.5n\|\mathbf{D}_1\| \overset{(ii)}{\leq} 1.5n\lambda_m \overset{(iii)}{\in} \Theta\left(\max\{\lambda, 1\}\right),$$

where (i) follows from Lemma 12 with probability at least $1 - c_3 q^{-\bar{\delta}}$, (ii) uses that all eigenvalues of $\mathcal{K}_1$ are at most λ_m by definition, and (iii) follows from $n\lambda_m \in \Theta(\max\{\lambda, 1\})$.

Finally, Lemma 13 directly yields with probability at least $1 - c_4 q^{-(1-\bar{\delta})}$ that $\|\mathbf{K}_2\| \in \Theta(1)$ and $\mu_{\min}(\mathbf{K}_2) \in \Theta(1)$. Hence, for d large enough, with probability at least $1 - (c_3 q^{-\bar{\delta}} + c_4 q^{-(1-\bar{\delta})}) \geq 1 - cq^{-\min\{\bar{\delta}, 1-\bar{\delta}\}}$, we have

$$\begin{aligned} \|\mathbf{K}_{>m}\| &\leq \|\mathbf{K}_1\| + \|\mathbf{K}_2\| \leq c_1 \max\{\lambda, 1\} \\ \mu_{\min}(\mathbf{K}_{>m}) &\geq \mu_{\min}(\mathbf{K}_2) \geq c_2. \end{aligned}$$

This implies that with high probability:

$$\begin{aligned} \|\mathbf{K}_{>m}\| + \lambda &\in \mathcal{O}(\max\{\lambda, 1\}) \\ \mu_{\min}(\mathbf{K}_{>m}) + \lambda &\in \Omega(\max\{\lambda, 1\}). \end{aligned}$$

This last equation directly implies that, with high probability:

$$\begin{aligned} r_2 &= \frac{\|\mathbf{K}_{>m}\| + \lambda}{\max\{\lambda, 1\}} \in \mathcal{O}(1) \\ r_1 &= \frac{\mu_{\min}(\mathbf{K}_{>m}) + \lambda}{\max\{\lambda, 1\}} \in \Omega(1). \end{aligned}$$

But from the definition of r_1 and r_2, it follows that $r_1 \leq r_2$. Thus, $r_1, r_2 \in \Theta(1)$.

Bound on $\mathbf{Tr}(\mathbf{S}_{>m})$. Finally, we need to show that

$$d^{\ell_\lambda} q^{\delta-1} \lesssim \mathbf{Tr}(\mathbf{S}_{>m}) \lesssim d^{\ell_\lambda} q^{-\delta} + d^{\ell_\lambda} q^{\delta-1} \tag{24}$$

with high probability. To do so, we differentiate whether $\bar{L} = L$ and $\bar{L} > L$. Intuitively, the case $\bar{L} = L$ corresponds to interpolation or mild regularization, because the maximum degree of learnable polynomials with regularization L equals the one without regularization $\bar{L}$. Conversely, $\bar{L} > L$ corresponds to strong regularization.

Case $\bar{L} = L$ (interpolation or mild regularization): In this setting,

$$\delta = \frac{\ell - \ell_\lambda - 1}{\beta} - L = \frac{\ell - 1}{\beta} - \bar{L} - \frac{\ell_\lambda}{\beta} = \bar{\delta} - \frac{\ell_\lambda}{\beta}. \tag{25}$$

First, Lemma 14 yields $\mathbf{Tr}(\mathbf{S}_2) \in \Theta(q^{-(1-\bar{\delta})})$ with probability at least $1 - c_1 q^{-(1-\bar{\delta})}$. Therefore,

$$\mathbf{Tr}(\mathbf{S}_2) \in \Theta(q^{-(1-\bar{\delta})}) = \Theta(q^{-(1-\delta)+\frac{\ell_\lambda}{\beta}}) \overset{(i)}{=} \Theta\left(d^{\ell_\lambda} q^{-(1-\delta)}\right),$$

where (i) follows from Equation (25). This directly yields the lower bound

$$\mathbf{Tr}(\mathbf{S}_{\geq m}) \geq \mathbf{Tr}(\mathbf{S}_2) \in \Omega(d^{\ell_\lambda} q^{-(1-\delta)}).$$

For the upper bound, we also need to consider $\mathbf{Tr}(\mathbf{S}_1)$:

$$\begin{aligned}\mathbf{Tr}(\mathbf{S}_1) &= n\mathbf{Tr}\left(\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\mathbf{D}_1^2\right) \leq n\left\|\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\right\|\mathbf{Tr}(\mathbf{D}_1^2) \\ &\overset{(i)}{\leq} 1.5n \sum_{k\in\mathcal{I}_1} \lambda_{S_k}^2,\end{aligned}$$

where (i) follows from Lemma 12 with probability at least $1 - c_2 q^{-\bar{\delta}}$. Then,

$$\begin{aligned}\mathbf{Tr}(\mathbf{S}_1) &\lesssim n \sum_{k\in\mathcal{I}_1} \lambda_{S_k}^2 \overset{(i)}{\leq} n\lambda_m^2 |\mathcal{I}_1| \\ &\overset{(ii)}{\in} \mathcal{O}\left(\max\{\lambda, 1\}^2 \frac{|\mathcal{I}_1|}{n}\right) \\ &\overset{(iii)}{=} \mathcal{O}\left(d^{2\ell_\lambda} \frac{dq^{\bar{L}}}{dq^{\bar{L}+\bar{\delta}}}\right) = \mathcal{O}\left(d^{2\ell_\lambda} q^{-\bar{\delta}}\right) \overset{(iv)}{=} \mathcal{O}\left(d^{\ell_\lambda} q^{-\delta}\right),\end{aligned}$$

where (i) uses $\lambda_k \leq \lambda_m$ for all $k \in \mathcal{I}_1$ by definition, (ii) uses the assumption $n\lambda_m \in \Theta(\max\{\lambda, 1\})$, and (iv) follows from Equation (25). Furthermore, (iii) uses the following bound of $|\mathcal{I}_1|$:

$$|\mathcal{I}_1| \leq \sum_{l=1}^{\bar{L}+1} \mathcal{C}(l, q, d) \leq \mathcal{C}(\bar{L}+1, q, d) \overset{(i)}{=} d\binom{q-1}{\bar{L}} \overset{(ii)}{\leq} \left(\frac{eq}{\bar{L}}\right)^{\bar{L}} \in \mathcal{O}(d \cdot q^{\bar{L}}),$$

where $\mathcal{C}(l, q, d)$ is defined in Equation (17), (i) follows from Lemma 9 from d large enough, and (ii) is a classical bound on the binomial coefficient. Finally, collecting the upper bounds on $\mathbf{Tr}(\mathbf{S}_1)$ and $\mathbf{Tr}(\mathbf{S}_2)$ yields the desired upper bound

$$\mathbf{Tr}(\mathbf{S}_{\geq m}) \leq \mathbf{Tr}(\mathbf{S}_1) + \mathbf{Tr}(\mathbf{S}_2) \in \mathcal{O}\left(d^{\ell_\lambda} q^{-\delta} + d^{\ell_\lambda} q^{-(1-\delta)}\right)$$

with high probability, and hence the overall bound on $\mathbf{Tr}(\mathbf{S}_{\geq m})$ with probability at least $1 - (c_1 q^{-(1-\bar{\delta})} + c_2 q^{-\bar{\delta}}) \geq 1 - cq^{-\min\{\bar{\delta}, 1-\bar{\delta}\}}$ for the case where $\bar{L} = L$.

Case $\bar{L} > L$ (strong regularization): In this setting, the dominating rate will arise from $\mathbf{Tr}(\mathbf{S}_1)$. We start by linking $\bar{\delta}$ and δ analogous to Equation (25):

$$\delta = \frac{\ell - \ell_\lambda - 1}{\beta} - L = -\frac{\ell_\lambda}{\beta} + \frac{\ell - 1}{\beta} - \bar{L} + \bar{L} - L = \bar{\delta} - \frac{\ell_\lambda}{\beta} + \bar{L} - L. \tag{26}$$

Next, as in the previous case, Lemma 14 yields $\mathbf{Tr}(\mathbf{S}_2) \in \Theta(q^{-(1-\bar{\delta})})$ with probability at least $1 - c_1 q^{-(1-\bar{\delta})}$, and therefore

$$\mathbf{Tr}(\mathbf{S}_2) \in \Theta(q^{-(1-\bar{\delta})}) = \Theta(q^{\bar{\delta}-1}) \overset{(i)}{=} \Theta\left(d^{\ell_\lambda} q^{\delta-(\bar{L}-L)-1}\right) \overset{(ii)}{\subseteq} o\left(d^{\ell_\lambda} q^{-(1-\delta)}\right),$$

where (i) follows from Equation (26), and (ii) from $\bar{L} \geq L+1$. Due to the slower growth in this setting, we consider $\mathbf{Tr}(\mathbf{S}_2)$ only for the upper bound of $\mathbf{Tr}(\mathbf{S}_{\geq m})$, but not for the lower bound. To bound $\mathbf{Tr}(\mathbf{S}_1)$, we first use

$$\begin{aligned}
0.5n \sum_{k \in \mathcal{I}_1} \lambda_{S_k}^2 &\overset{(i)}{\leq} n\mu_{\min}\left(\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\right)\mathbf{Tr}(\mathbf{D}_1^2) \\
&\leq n\mathbf{Tr}\left(\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\mathbf{D}_1^2\right) = \mathbf{Tr}(\mathbf{S}_1) \\
&\leq n\left\|\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\right\|\mathbf{Tr}(\mathbf{D}_1^2) \overset{(ii)}{\leq} 1.5n \sum_{k \in \mathcal{I}_1} \lambda_{S_k}^2,
\end{aligned}$$

where (i) and (ii) follow from Lemma 12 simultaneously with probability at least $1 - c_2 q^{-\bar{\delta}}$. We then decompose the sum over all squared eigenvalues with index in $\mathcal{I}_1$ as

$$n \sum_{k \in \mathcal{I}_1} \lambda_{S_k}^2 = \underbrace{n \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| \leq L+1}} \lambda_{S_k}^2}_{T_1} + \underbrace{n \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| = L+2}} \lambda_{S_k}^2}_{T_2} + \underbrace{n \sum_{l=L+3}^{\bar{L}+1} \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| = l}} \lambda_{S_k}^2}_{T_3}.$$

In the following, we analyze T_1, T_2, and T_3 separately.

Upper bound For the upper bound of $\mathbf{Tr}(\mathbf{S}_{\geq m})$, we bound all three terms from above. First, we upper-bound T_1 as follows:

$$\begin{aligned}
T_1 = n \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| \leq L+1}} \lambda_{S_k}^2 &\overset{(i)}{\leq} \frac{n^2 \lambda_m^2}{n} \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| \leq L+1}} 1 \\
&\leq \frac{n^2 \lambda_m^2}{n} \sum_{l=0}^{L+1} \mathcal{C}(l,q,d) \overset{(ii)}{\in} \mathcal{O}\left(\frac{\max\{\lambda,1\}^2}{n} d q^L\right) \\
&= \mathcal{O}\left(\frac{d^{2\ell_\lambda}}{d \cdot d^{\ell_\lambda} \cdot q^{L+\delta}} d q^L\right) = \mathcal{O}\left(d^{\ell_\lambda} q^{-\delta}\right),
\end{aligned}$$

where (i) follows from $\lambda_k \leq \lambda_m$ for all $k > m$ due to the decreasing order of eigenvalues. Furthermore, (ii) uses $n\lambda_m \in \Theta(\max\{\lambda, 1\})$ and the fact that $\mathcal{C}(l,q,d) \in \mathcal{O}(dq^{l-1})$, which follows from Lemma 9 together with a classical bound on the binomial coefficient: for d large enough,

$$\mathcal{C}(l,q,d) = d\binom{q-1}{l-1} \leq d\left(\frac{e(q-1)}{l-1}\right)^{l-1} \in \mathcal{O}\left(dq^{l-1}\right).$$

Second, the upper bound of T_2 arises as follows:

$$
\begin{aligned}
T_2 = n \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| = L+2}} \lambda_{S_k}^2 &\overset{(i)}{=} n(\xi_{L+2}^{(q)})^2 \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| = L+2}} \left(\frac{q+1-\gamma(S_k)}{d\mathcal{B}(L+2,q)} \right)^2 \\
&\overset{(ii)}{\leq} (\xi_{L+2}^{(q)})^2 \frac{n \cdot q}{d\mathcal{B}(L+2,q)} \sum_{\substack{\gamma(S) \leq q \\ |S| = L+2}} \frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \\
&\overset{(iii)}{\lesssim} \frac{n \cdot q}{d \cdot q^{L+2}} \sum_{\substack{\gamma(S) \leq q \\ |S| = L+2}} \frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \\
&\overset{(iv)}{\lesssim} \frac{d \cdot d^{\ell_\lambda} q^{L+\delta} \cdot q}{d \cdot q^{L+2}} \mathcal{Q}_{L+2}^{(d,q)}(x,x) \\
&\overset{(v)}{\in} \mathcal{O}\left(d^{\ell_\lambda} q^{-(1-\delta)} \right),
\end{aligned}
$$

where (i) follows from the eigenvalue formula in Proposition 1, and (ii) uses $q+1-\gamma(S) \leq q$. Next, (iii) uses Assumption 1 and applies the bound $\mathcal{B}(L+2,q) = \binom{q}{L+2} \leq (eq/(L+2))^{L+2}$. Step (iv) uses $\mathcal{Y}_S(x)\mathcal{Y}_S(x) = 1$ for all S and $x \in \{-1,1\}^d$, together with the definition of $\mathcal{Q}_{L+2}^{(d,q)}$. Lastly, (v) applies Lemma 15.

Third, we upper-bound T_3. For this, let $\tilde{k}$ be the smallest index that corresponds to a set $S_{\tilde{k}}$ with $|S_{\tilde{k}}| = L+3$. Since the eigenvalues are indexed in decreasing order, $\lambda_{S_{\tilde{k}}}$ is the largest eigenvalue corresponding to a set of size $L+3$. Furthermore, for d large enough, Lemma 10 yields that $\lambda_{S_{\tilde{k}}}$ is the maximum eigenvalue for all sets of size at least $L+3$. Then, for d large enough,

$$
\begin{aligned}
T_3 = n \sum_{l=L+3}^{\bar{L}+1} \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| = l}} \lambda_{S_k}^2 &\leq n \lambda_{S_{\tilde{k}}} \sum_{l=L+3}^{\bar{L}+1} \sum_{\substack{k \in \mathcal{I}_1 \\ |S_k| = l}} \lambda_{S_k} \\
&\overset{(i)}{\lesssim} n \lambda_{S_{\tilde{k}}} \\
&\overset{(ii)}{\in} \mathcal{O}\left(n \frac{1}{dq^{L+2}} \right) = \mathcal{O}\left(\frac{d \cdot d^{\ell_\lambda} q^{L+\delta}}{dq^{L+2}} \right) \subseteq \mathcal{O}\left(d^{\ell_\lambda} q^{-(1-\delta)} \right),
\end{aligned}
$$

where (i) follows from the fact that the sum of all the eigenvalues of the kernel must be bounded by a constant, thus the sum of a subset of the eigenvalues must be bounded as well. Furthermore, (ii) follows from Lemma 10 with $|S_{\tilde{k}}| = L+3$.

Lower bound Since all three terms are positive, it suffices to lower-bound T_2:

$$
\begin{aligned}
T_2 &\overset{(i)}{=} n(\xi_{L+2}^{(q)})^2 \sum_{\substack{k\in\mathcal{I}_1\\ |S_k|=L+2}} \left(\frac{q+1-\gamma(S_k)}{d\mathcal{B}(L+2,q)}\right)^2 \\
&\overset{(ii)}{\geq} n(\xi_{L+2}^{(q)})^2 \sum_{\substack{|S|=L+2\\ \gamma(S)\leq\lfloor q/2\rfloor}} \left(\frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)}\right)^2 \\
&\overset{(iii)}{\geq} (\xi_{L+2}^{(q)})^2 \frac{n\cdot q/2}{d\mathcal{B}(L+2,q)} \sum_{\substack{|S|=L+2\\ \gamma(S)\leq\lfloor q/2\rfloor}} \frac{\lfloor q/2\rfloor+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \\
&\overset{(iv)}{\geq} c(\xi_{L+2}^{(q)})^2 \frac{n\cdot q/2}{d\mathcal{B}(L+2,q)} \sum_{\substack{|S|=L+2\\ \gamma(S)\leq\lfloor q/2\rfloor}} \frac{\lfloor q/2\rfloor+1-\gamma(S)}{d\mathcal{B}(L+2,\lfloor q/2\rfloor)} \\
&\overset{(v)}{\geq} c(\xi_{L+2}^{(q)})^2 \frac{n\cdot q/2}{d\mathcal{B}(L+2,q)} \mathcal{Q}_{L+2}^{(d,\lfloor q/2\rfloor)}(x,x) \\
&\overset{(vi)}{\in} \Omega\left(\frac{d\cdot d^{\ell_\lambda} q^{L+\delta}\cdot q}{d\cdot q^{L+2}}\right) = \Omega\left(d^{\ell_\lambda} q^{-(1-\delta)}\right),
\end{aligned}
$$

where (i) follows from Proposition 1. Step (ii) uses that all $k \in \mathcal{I}_1$ have $|S_k| \leq q$ by definition; hence the $k \in \mathcal{I}_1$ with $|S_k| \leq \lfloor q/2\rfloor$ are a subset of $\mathcal{I}_1$. Next, (iii) follows since $q \geq \lfloor q/2\rfloor$ and $q+1-\gamma(S) \geq q/2$. Step (v) uses $\mathcal{Y}_S(x)\mathcal{Y}_S(x) = 1$ for all S and $x \in \{-1,1\}^d$, and (vi) from Lemma 15 with $\lfloor q/2\rfloor$ as filter size. Finally, (iv) follows from the fact that $\mathcal{B}(L+2,q)$ and $\mathcal{B}(L+2,\lfloor q/2\rfloor)$ are of the same order; this follows from a classical bound on the binomial coefficient:

$$
\mathcal{B}(L+2,q) \leq \left(\frac{eq}{L+2}\right)^{L+2} \leq \frac{1}{(2e)^{L+2}}\left(\frac{q/2}{L+2}\right)^{L+2} \leq c\mathcal{B}(L+2,\lfloor q/2\rfloor).
$$

Ultimately, we may summarize all individual bounds for the case $\bar{L} > L$ as follows:

$$
\begin{aligned}
\mathbf{Tr}(\mathbf{S}_{>m}) &= \mathbf{Tr}(\mathbf{S}_1) + \mathbf{Tr}(\mathbf{S}_2) \\
\mathbf{Tr}(\mathbf{S}_2) &\in o(d^{\ell_\lambda} q^{-(1-\delta)}) && \text{with prob. at least } 1 - c_1 q^{-(1-\bar{\delta})} \\
\mathbf{Tr}(\mathbf{S}_1) &= T_1 + T_2 + T_3 && \text{with prob. at least } 1 - c_2 q^{-\bar{\delta}} \\
T_1 &\in \mathcal{O}\left(d^{\ell_\lambda} q^{-\delta}\right) \\
T_2 &\in \Theta\left(d^{\ell_\lambda} q^{-(1-\delta)}\right) \\
T_3 &\in \mathcal{O}\left(d^{\ell_\lambda} q^{-(1-\delta)}\right).
\end{aligned}
$$

Combined, those bounds yield Equation (24) with probability at least $1 - (c_1 q^{-(1-\bar{\delta})} + c_2 q^{-\bar{\delta}}) \geq 1 - cq^{-\min\{\bar{\delta},1-\bar{\delta}\}}$ for the case where $\bar{L} > L$. Since the cases are exhaustive, and hold both with same probability, this concludes the proof of the bound on $\mathbf{Tr}(\mathbf{S}_{>m})$.

The statement of Lemma 3 then follows from taking the union bound on the probabilities over the individual proof parts. □

C.3 Technical Lemma

Finally, we use the following technical lemmas in the proof of Lemma 3.

Lemma 12 (Bound on $\mathbf{K}_1$)**.** *In the setting of Definition 2, uniformly over all choices of λ, we have*

$$
\left\| \frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n} - \mathbf{I} \right\| \leq 1/2,
$$

with probability at least $1 - cq^{-\bar{\delta}}$.

Proof. The proof follows a very similar argument to Lemma 1. Still, minor differences prevent us from using that result directly.

We start by defining

$$\hat{\mathcal{I}}_1 \coloneqq \{k \in \mathbb{N} \mid |S_i| \leq \bar{L} + 1\}.$$

Notice that the definition of $\hat{\mathcal{I}}_1$ does not depend on m, and hence it does not depend on λ. Furthermore, notice that for any m, $\mathcal{I}_1 \subseteq \hat{\mathcal{I}}_1$.

We start by computing the $\left|\hat{\mathcal{I}}_1\right|$:

$$\left|\hat{\mathcal{I}}_1\right| = \sum_{l=0}^{\bar{L}+1} \mathcal{C}(l,q,d) \overset{(i)}{=} \sum_{l=0}^{\bar{L}+1} d\binom{q-1}{l-1} \overset{(ii)}{\leq} d\sum_{l=0}^{\bar{L}+1}\left(\frac{eq}{l-1}\right)^{l-1} \overset{(iii)}{\in} \mathcal{O}(d \cdot q^{\bar{L}}),$$

where (i) follows from Lemma 9 for d large enough, (ii) is a classical bound on the binomial coefficient, and (iii) follows from the fact that the largest degree monomial dominates the others.

Now we define $\hat{\mathbf{\Psi}}_1 \in \mathbb{R}^{n \times |\hat{\mathcal{I}}_1|}$ in analogous way to how $\mathbf{\Psi}_1$ is defined in Definition 2.

Using the same arguments used already in the proof of Lemma 1, we can show that the rows of $\hat{\mathbf{\Psi}}_1$ are independent, with norm bounded by $|\hat{\mathcal{I}}_1|$, and with expectation of outer product equal to $\mathbf{I}_{|\hat{\mathcal{I}}_1|}$. Again as in Lemma 1, we can apply Theorem 5.44 from Vershynin (2012), and, choosing $t = \frac{1}{n}\sqrt{\frac{n}{|\hat{\mathcal{I}}_1|}}$, we get that:

$$\left\|\frac{\hat{\mathbf{\Psi}}_1^\intercal \hat{\mathbf{\Psi}}_1}{n} - \mathbf{I}_{|\hat{\mathcal{I}}_1|}\right\| \leq 1/2, \quad \text{w.p.} \geq 1 - q^{-\bar{\delta}}.$$

To conclude the proof, it suffices to remember that, uniformly over all choices of λ, and consequential choice of m, $\mathcal{I}_1 \subseteq \hat{\mathcal{I}}_1$. Thus:

$$\left\|\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n} - \mathbf{I}_{|\mathcal{I}_1|}\right\| \leq \left\|\frac{\hat{\mathbf{\Psi}}_1^\intercal \hat{\mathbf{\Psi}}_1}{n} - \mathbf{I}_{|\hat{\mathcal{I}}_1|}\right\| \leq 1/2 \quad \text{w.p.} \geq 1 - cd^{-\beta\bar{\delta}},$$

concluding the proof. □

Lemma 13 (Bound on $\mathbf{K}_2$)**.** *In the setting of Definition 2, assuming that the kernel satisfies Assumption 1, with probability at least* $1 - cq^{-(1-\bar{\delta})}$*, we have*

$$c_1 \leq \mu_{\min}(\mathbf{K}_2) \leq \|\mathbf{K}_2\| \leq c_2$$

for some positive constants c_1, c_2*.*

Proof. For the upper bound on the largest eigenvalue, Proposition 1 yields the following decomposition of $\mathcal{K}_2(x,x')$:

$$\mathcal{K}_2(x,x') = \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \sum_{\substack{\gamma(S)\leq q \\ |S|=l}} \frac{q+1-\gamma(S)}{d\mathcal{B}(l,q)} \mathcal{Y}_S(x)\mathcal{Y}_S(x') = \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \mathcal{Q}_l^{(d,q)}(x,x'),$$

where $\mathcal{Q}_l^{(d,q)}(x,x')$ is defined in Definition 2. Then, using the triangle inequality and positivity of the $\xi_l^{(q)}$ from Assumption 1, we have

$$\begin{aligned}\|\mathbf{K}_2\| &= \left\|\sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \mathbf{Q}_l^{(d,q)}\right\| \leq \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \left\|\mathbf{Q}_l^{(d,q)}\right\| \\ &\overset{(i)}{\leq} 1.5 \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \overset{(ii)}{\leq} 1.5c_3 \qquad \text{w.p.} \geq 1 - cq^{-(1-\bar{\delta})},\end{aligned}$$

where $\mathbf{Q}_l^{(d,q)}$ is the kernel matrix corresponding to $\mathcal{Q}_l^{(d,q)}(x, x')$, (i) uses Lemma 16, and (ii) follows from Assumption 1.

The lower bound follows from

$$\mu_{\min}(\mathbf{K}_2) \geq \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \mu_{\min}\left(\mathbf{Q}_l^{(d,q)}\right) \overset{(i)}{\geq} \frac{1}{2} \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \geq \frac{c_4}{2} \qquad \text{w.p.} \geq 1 - cq^{-(1-\bar{\delta})},$$

where (i) follows from Lemma 16.

Since Lemma 16 yields both the upper and lower bound at once with probability $1 - cq^{-(1-\bar{\delta})}$, this concludes the proof. □

Lemma 14 (Bound on $\mathbf{Tr}(\mathbf{S}_2)$)**.** *In the setting defined in Section 3.1, we have that:*

$$\mathbf{Tr}(\mathbf{S}_2) \in \Theta(q^{-(1-\bar{\delta})}),$$

with probability at least $1 - cq^{-(1-\bar{\delta})}$.

Proof. We start by proofing two technicalities.

Proof that $\|\mathbf{S}_2\| \leq \lambda_{\bar{m}} \|\mathbf{K}_2\|$. Let $\mathbf{v}$ be an arbitrary vector in $\mathbb{R}^n$, then:

$$\begin{aligned}
\|\mathbf{S}_2 \mathbf{v}\| &= \left\| \left(\sum_{k \in \mathcal{I}_2} \lambda_k^2 [\mathbf{\Psi}_2]_{:,\mathbf{k}} [\mathbf{\Psi}_2]_{:,\mathbf{k}}^\intercal \right) \mathbf{v} \right\| \\
&\leq \left\| \max_{k \in \mathcal{I}_2} \{\lambda_k\} \sum_{k \in \mathcal{I}_2} \lambda_k [\mathbf{\Psi}_2]_{:,\mathbf{k}} [\mathbf{\Psi}_2]_{:,\mathbf{k}}^\intercal \mathbf{v} \right\| \\
&\overset{(i)}{=} \lambda_{\bar{m}} \left\| \left(\sum_{k \in \mathcal{I}_2} \lambda_k [\mathbf{\Psi}_2]_{:,\mathbf{k}} [\mathbf{\Psi}_2]_{:,\mathbf{k}}^\intercal \right) \mathbf{v} \right\| \\
&= \lambda_{\bar{m}} \|\mathbf{K}_2 \mathbf{v}\| \leq \lambda_{\bar{m}} \|\mathbf{K}_2\| \|\mathbf{v}\|,
\end{aligned}$$

where in (i) we used the definition of m together with the fact that the eigenvalues are sorted in descending order.

Proof that $\lambda_{\bar{m}} \in \Theta\left(\frac{1}{dq^{\bar{L}+1}}\right)$. To show this, we show that for any k in $\mathcal{I}_2$, $\lambda_k \in \mathcal{O}\left(\frac{1}{dq^{\bar{L}+1}}\right)$ and that there exists a $\tilde{m}$ such that $\lambda_{\tilde{m}} \in \Omega\left(\frac{1}{dq^{\bar{L}+1}}\right)$. Since $\lambda_{\bar{m}}$ is the largest of the eigenvalues $\lambda_{S_k}, k \in \mathcal{I}_2$, those two fact imply that $\lambda_{\bar{m}} \in \Theta\left(\frac{1}{dq^{\bar{L}+1}}\right)$.

Let k be an arbitrary index in $\mathcal{I}_2$, then $|S_k| \geq \bar{L} + 2$. Lemma 10 yields that:

$$\lambda_k \in \mathcal{O}\left(\frac{1}{dq^{|S_k|-1}}\right) \subseteq \mathcal{O}\left(\frac{1}{dq^{L+1}}\right).$$

This concludes the proof that for any k in $\mathcal{I}_2$, $\lambda_k \in \mathcal{O}\left(\frac{1}{dq^{\bar{L}+1}}\right)$.

Now we show that there exist $\tilde{m}$ such that $\lambda_{\tilde{m}} \in \Omega\left(\frac{1}{dq^{\bar{L}+1}}\right)$. To do so, we take $\tilde{m}$ be such that $S_{\tilde{m}} = \{1, 2, \ldots, L+2\}$. Note that $\tilde{m} \in \mathcal{I}_2$ for d sufficiently large, since $|S_{\tilde{m}}| = L + 2$. Then, Proposition 1 yields the following formula for $\lambda_{S_{\tilde{m}}}$:

$$\lambda_{S_{\tilde{m}}} = \xi_{L+2}^{(q)} \frac{(q + 1 - \gamma(S_{\tilde{m}}))}{d\mathcal{B}(L+2, q)} \in \Theta\left(\frac{q}{dq^{L+2}}\right) \subseteq \Theta\left(\frac{1}{dq^{L+1}}\right),$$

Upper bound. The upper bound follows by simply applying the last two results:

$$\mathbf{Tr}(\mathbf{S}_2) \leq n\|\mathbf{S}_2\| \leq n\lambda_{\bar{m}} \|\mathbf{K}_2\| \overset{(i)}{\lesssim} n\lambda_{\bar{m}} \in \Theta(\frac{dq^{\bar{L}+\bar{\delta}}}{dq^{\bar{L}+1}}) = \Theta(q^{-(1-\bar{L})}),$$

where (i) follows with probability at least $1 - cq^{-(1-\bar{\delta})}$ from the Lemma 13.

Lower bound. For the lower bound we need a more refined argument.

$$
\begin{aligned}
\mathcal{S}_2(x,x') &= \sum_{k\in\mathcal{I}_2} \lambda_S^2 \mathcal{Y}_S(x)\mathcal{Y}_S(x') \geq \sum_{\substack{\gamma(S)\leq q\\ |S|=\bar{L}+2}} \lambda_S^2 \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&\geq \sum_{\substack{\gamma(S)\leq \lfloor q/2\rfloor\\ |S|=\bar{L}+2}} \lambda_S^2 \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&\overset{(i)}{=} (\xi_{\bar{L}+2}^{(q)})^2 \sum_{\substack{\gamma(S)\leq \lfloor q/2\rfloor\\ |S|=\bar{L}+2}} \frac{(q+1-\gamma(S))^2}{d^2\mathcal{B}(\bar{L}+2,q)^2} \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&= \frac{(\xi_{\bar{L}+2}^{(q)})^2}{d\mathcal{B}(\bar{L}+2,q)} \sum_{\substack{\gamma(S)\leq \lfloor q/2\rfloor\\ |S|=\bar{L}+2}} (q+1-\gamma(S)) \frac{q+1-\gamma(S)}{d\mathcal{B}(\bar{L}+2,q)} \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&\overset{(ii)}{\geq} \frac{(\xi_{\bar{L}+2}^{(q)})^2 q/2}{d\mathcal{B}(\bar{L}+2,q)} \sum_{\substack{\gamma(S)\leq \lfloor q/2\rfloor\\ |S|=\bar{L}+2}} \frac{\lfloor q/2\rfloor+1-\gamma(S)}{d\mathcal{B}(\bar{L}+2,q)} \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&\overset{(iii)}{\geq} c'_{\bar{L}} \frac{(\xi_{\bar{L}+2}^{(q)})^2 q/2}{d\mathcal{B}(\bar{L}+2,q)} \sum_{\substack{\gamma(S)\leq \lfloor q/2\rfloor\\ |S|=\bar{L}+2}} \frac{\lfloor q/2\rfloor+1-\gamma(S)}{d\mathcal{B}(\bar{L}+2,\lfloor q/2\rfloor)} \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&\overset{(iv)}{=} c'_{\bar{L}} \frac{(\xi_{\bar{L}+2}^{(q)})^2 q/2}{d\mathcal{B}(\bar{L}+2,q)} \mathcal{Q}_{\bar{L}+2}^{(d,\lfloor q/2\rfloor)}(x,x'),
\end{aligned}
$$

where (i) follows Proposition 1. In (ii) we used that as long as $\gamma(S) \leq \lfloor q/2\rfloor$, $q+1-\gamma(S) \geq \frac{q}{2}$ and $q \geq \lfloor q/2\rfloor$. In (iii) we use the bound on the binomial coefficient $\left(\frac{q}{\bar{L}+2}\right)^{\bar{L}+2} \leq \binom{q}{\bar{L}+2} = \mathcal{B}(\bar{L}+2,q) \leq \left(\frac{eq}{\bar{L}+2}\right)^{\bar{L}+2}$ to compute:

$$
\mathcal{B}(\bar{L}+2,q) \leq \left(\frac{eq}{\bar{L}+2}\right)^{\bar{L}+2} \leq \frac{1}{(2e)^{\bar{L}+2}} \left(\frac{q/2}{\bar{L}+2}\right)^{\bar{L}+2} \leq \frac{1}{(2e)^{\bar{L}+2} c_{\bar{L}}} \left(\frac{q/2}{\bar{L}+2}\right)^{\bar{L}+2} \leq c'_{\bar{L}} \mathcal{B}(\bar{L}+2,\lfloor q/2\rfloor),
$$

where $c'_{\bar{L}}$ is a constant that depends on $\bar{L}$. Finally, in (iv), we used the definition of $\mathcal{Q}_{\bar{L}+2}^{(d,\lfloor q/2\rfloor)}$ in Definition 2. This implies that:

$$
\mu_{\min}(\mathbf{S}_2) \geq c'_{\bar{L}} \frac{(\xi_{\bar{L}+2}^{(q)})^2 q/2}{d\mathcal{B}(\bar{L}+2,q)} \mu_{\min}\left(\mathbf{Q}_{\bar{L}+2}^{(d,\lfloor q/2\rfloor)}\right).
$$

Hence, we finally bound $\mathbf{Tr}(\mathbf{S}_2)$:

$$
\begin{aligned}
\mathbf{Tr}(\mathbf{S}_2) &\geq n\mu_{\min}(\mathbf{S}_2) \geq c'_{\bar{L}} n \frac{(\xi_{\bar{L}+2}^{(q)})^2 q/2}{d\mathcal{B}(\bar{L}+2,q)} \mu_{\min}\left(\mathbf{Q}_{\bar{L}+2}^{(d,\lfloor q/2\rfloor)}\right) \\
&\overset{(i)}{\gtrsim} d^{\ell} \frac{q}{dq^{\bar{L}+2}} = \frac{dq^{\bar{L}+\bar{\delta}}}{dq^{\bar{L}+1}} = q^{-(1-\delta)},
\end{aligned}
$$

where (i) follows from the lower bound on $\mu_{\min}\left(\mathbf{Q}_{\bar{L}+2}^{(d,\lfloor q/2\rfloor)}\right)$ found in Lemma 16, and on the lower bound on $\xi_{\bar{L}+2}^{(q)}$ is from Assumption 1. This concludes the proof. □

Lemma 15 (Diagonal elements of $\mathcal{Q}^{(d,q)}$)**.** *For any $x \in \{-1,1\}^d$, for any $d \in \mathbb{N}$, $q \in \mathbb{N}$, $q < d/2$, for any $l \in \mathbb{N}$,*

$$
\mathcal{Q}_l^{(d,q)}(x,x) = 1.
$$

Proof. We compute:

$$\begin{aligned}
\mathcal{Q}_l^{(d,q)}(x,x) &\overset{(i)}{=} \sum_{\substack{\gamma(S)\leq q\\|S|=l}} \frac{q+1-\gamma(S)}{d\mathcal{B}(l,q)}\mathcal{Y}_S(x)^2 \\
&\overset{(ii)}{=} \frac{1}{d\mathcal{B}(l,q)} \sum_{\substack{\gamma(S)\leq q\\|S|=l}} q+1-\gamma(S) = \frac{1}{d\mathcal{B}(l,q)} \sum_{\gamma=l}^{q}(q+1-\gamma) \sum_{\substack{\gamma(S)=\gamma\\|S|=l}} 1 \\
&\overset{(iii)}{=} \frac{1}{d\mathcal{B}(l,q)} \sum_{\gamma=l}^{q} \mathcal{C}(l,\gamma,d) \overset{(iv)}{=} \frac{1}{d\mathcal{B}(l,q)} \sum_{\gamma=l}^{q} d\binom{\gamma-1}{l-1} \overset{(v)}{=} \frac{1}{d\mathcal{B}(l,q)} d\binom{q}{l} \overset{(vi)}{=} 1,
\end{aligned} \tag{27}$$

where (i) is the definition of $\mathcal{Q}_l^{(d,q)}(x,x)$ itself. (ii) follows noticing that the hypercube eigenfunctions squared are always equal to 1, as clear from their definition in Equation (13). (iv) follows from Lemma 9, (v) from the hockey-stick identity and (vi) from the formula to compute $\mathcal{B}$ in Equation (15). The last thing to justify is (iii). For this we notice that $\sum_{\substack{\gamma(S)=\gamma\\|S|=l}} 1$ matches the definition of $\tilde{\mathcal{C}}(l,\gamma,d)$ found in the proof of Lemma 9. We then show the following recurrence:

$$\begin{aligned}
\sum_{\gamma=l}^{q}(q+1-\gamma)\tilde{\mathcal{C}}(l,\gamma,d) &= \sum_{\gamma=l}^{q}\Big(\tilde{\mathcal{C}}(l,\gamma,d) + ((q-1)+1-\gamma)\tilde{\mathcal{C}}(l,\gamma,d)\Big) \\
&\overset{(i)}{=} \mathcal{C}(l,\gamma,d) + \sum_{\gamma=l}^{q-1}((q-1)+1-\gamma)\tilde{\mathcal{C}}(l,\gamma,d),
\end{aligned}$$

where in (i) we use the fact that $\mathcal{C}(l,\gamma,d) = \sum_{\gamma=l}^{q}\tilde{\mathcal{C}}(l,\gamma,d)$, which follows from the definition. Furthermore, the second sum only reaches $q-1$ since the last element would be 0. Applying recursively this formula $q-l$ times, we get:

$$\sum_{\gamma=l}^{q}(q+1-\gamma)\tilde{\mathcal{C}}(l,\gamma,d) = \sum_{\gamma=l}^{q}\mathcal{C}(l,\gamma,d),$$

which explains (iii) in Equation (27). This concludes the proof. □

C.4 Random matrix theory lemmas

In this subsection are contained the key random matrix concentration needed for the proof to go through.

Lemma 16 (Bound on the kernel's tail)**.** *In the same setting as Definition 2:*

$$\forall l \in \mathbb{N}, l \geq \bar{L}+1, \quad \|\mathbf{Q}_l^{(d,q)} - \mathbf{I}_n\| \leq \frac{1}{2},$$

with probability at least $1 - cq^{-(1-\bar{\delta})}$.

Notice that, from this lemma, it naturally follows that $\|\mathbf{Q}_l^{(d,q)}\| \leq 1.5$, and $\mu_{\min}\left(\mathbf{Q}_l^{(d,q)}\right) \geq 1/2$. Now we present the proof:

Proof of Lemma 16. Thanks to Lemma 15, we know that the diagonal elements of $\mathbf{Q}_l^{(d,q)}$ are just 1. Hence, we define $\mathbf{W}_l^{(d,q)} \coloneqq \mathbf{Q}_l^{(d,q)} - \mathbf{I}_n$.

Our objective is now to show that $\|\mathbf{W}_l^{(d,q)}\| \leq 1/2$ at the same time for every $l > \bar{L}+1$ with high probability. We will however use two different strategies for small l and large l. In particular, let us define

$$T \coloneqq \left\lceil \frac{4}{\beta} + 4(\bar{L}+\bar{\delta}) \right\rceil.$$

Notice that T does not depend on d nor ℓ_λ.

Terms $l < T$: here we use Lemma 17. To successfully apply it, we need to find an appropriate $\mathcal{M}^{(l,q)}$, and show that the assumptions of the Lemma holds. The first assumption has already been checked, since we showed that $\mathrm{diag}(\mathbf{Q}_l^{(d,q)}) = \mathbf{I}_n$. For the second one we can compute, for any $i \neq j$:

$$\begin{aligned}
&\left|\mathbb{E}_{x_j}\left[[\mathbf{W}_l^{(d,q)}]_{i,j}[\mathbf{W}_l^{(d,q)}]_{j,k}\right]\right| \\
&= \left|\frac{1}{(d\mathcal{B}(l,q))^2}\sum_{\substack{\gamma(S)\leq q\\|S|=l}}\sum_{\substack{\gamma(S')\leq q\\|S'|=l}}(q+1-\gamma(S))(q+1-\gamma(S'))\mathcal{Y}_S(x_i)\mathbb{E}\left[\mathcal{Y}_S(x_j)\mathcal{Y}_{S'}(x_j)\right]\mathcal{Y}_{S'}(x_k)\right| \\
&\overset{(i)}{=} \frac{1}{(d\mathcal{B}(l,q))^2}\left|\sum_{\substack{\gamma(S)\leq q\\|S|=l}}(q+1-\gamma(S))^2\mathcal{Y}_S(x_i)\mathcal{Y}_S(x_k)\right| \\
&\leq \frac{q}{d\mathcal{B}(l,q)}\frac{1}{d\mathcal{B}(l,q)}\left|\sum_{\substack{\gamma(S)\leq q\\|S|=l}}(q+1-\gamma(S))\mathcal{Y}_S(x_i)\mathcal{Y}_S(x_k)\right| \\
&= \left(\frac{d\mathcal{B}(l,q)}{q}\right)^{-1}\left|\left[\mathbf{W}_l^{(d,q)}\right]_{i,k}\right| \leq \left(\frac{d\mathcal{B}(l,q)}{lq}\right)^{-1}\left|\left[\mathbf{W}_l^{(d,q)}\right]_{i,k}\right|,
\end{aligned}$$

where in (i) we use that the eigenfunctions are orthonormal. We then chose $\mathcal{M}^{(l,q)} = d\mathcal{B}(l,q)/lq$. The extra l factor will be needed for the third assumption to hold. We now move to the third assumption. As in Lemma 17, let $p \in \mathbb{N}, p \geq 2$, then:

$$\begin{aligned}
\mathbb{E}\left[|[\mathbf{W}_l^{(d,q)}]_{i,j}|^p\right]^{\frac{1}{p}} &\overset{(i)}{\leq} (p-1)^{l/2}\sqrt{\mathbb{E}\left[[\mathbf{W}_l^{(d,q)}]_{i,j}^2\right]} \leq \sqrt{p^l\sum_{\substack{\gamma(S)\leq q\\|S|=l}}\frac{(q+1-\gamma(S)^2}{(d\mathcal{B}(l,q))^2}} \\
&\overset{(ii)}{\leq} \sqrt{p^l\frac{q^2}{(d\mathcal{B}(l,q))^2}\mathcal{C}(l,q,d)} = \sqrt{p^l\frac{lq}{d\mathcal{B}(l,q)}}\sqrt{\frac{q\mathcal{C}(l,q,d)}{ld\mathcal{B}(l,q)}} \overset{(iii)}{=} \sqrt{\frac{p^l}{\mathcal{M}^{(l,q)}}},
\end{aligned}$$

where (i) follows from hypercontractivity as in Lemma 18, (ii) from the definition of $\mathcal{C}(l,q,d)$. (iii) follows from, Lemma 9, the formula of $\mathcal{B}(l,q)$ in Equation (15), the formula $\binom{q-1}{l-1} = \frac{q}{l}\binom{q}{l}$, and the definition of $\mathcal{M}^{(l,q)}$.

Finally, we can apply Lemma 17 and we get that for any $p \in \mathbb{N}, p > 2$:

$$\Pr\left(\|\mathbf{W}_l^{(d,q)}\| > 1/2\right) \leq c^p p^{3p} n\left(\frac{n}{\mathcal{M}^{(l,q)}}\right)^p + c_2^p\left(\frac{n}{\mathcal{M}^{(l,q)}}\right)^2.$$

In particular, if we take $p \geq \frac{2}{\bar{\beta}} + \frac{1}{(1-\bar{\delta})\beta} - (\bar{L}+2-\bar{\delta})$, then we get that:

$$\Pr\left(\|\mathbf{W}_l^{(d,q)}\| > 1/2\right) \leq c_l q^{-2(1-\bar{\delta})}.$$

We can than take the union bound for all $l = \bar{L}+1, \ldots, T$, and we get:

$$\Pr\left(\forall l \in \{\bar{L}+2, \ldots, T-1\},\ \|\mathbf{W}_l^{(d,q)}\| > 1/2\right) \leq q^{-2(1-\bar{\delta})}\sum_{l=\bar{L}+1}^{T-1} c_l \leq C_{\bar{L}} q^{-2(1-\bar{\delta})},$$

where $C_{\bar{L}}$ depends only on $\bar{L}$ and T, which are fixed in our setting. We remark that $\bar{L}$ does not depend on ℓ_λ.

Since we cannot control the dependency of c_l on l, with this strategy, we cannot bound $\|\mathbf{W}_l^{(d,q)}\|$ for all l at the same time, but only for a finite number of different values of l, i.e. all $l \leq T$. We present another technique for the rest of the terms:

Terms $l \geq T$: for this section we can just bound the off-diagonal part using the Frobenius norm:

$$\|\mathbf{W}_l^{(d,q)}\| \leq \sqrt{\sum_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2} \leq \sqrt{n^2 \max_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2} \overset{(i)}{\leq} \sqrt{1/4} = 1/2, \tag{28}$$

where in (i) we use that $\max_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2 \leq 1/4n^2$. Now we compute the probability that this is indeed the case:

$$\begin{aligned}
\Pr\left(\max_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2 \leq 1/4n^2\right) &\leq n^2 \Pr\left([\mathbf{W}_l^{(d,q)}]_{i,j}^2 \leq 1/4n^2\right) \\
&\overset{(i)}{\leq} 4n^4 \mathbb{E}\left[[\mathbf{W}_l^{(d,q)}]_{i,j}^2\right] \\
&\leq \frac{4n^4}{\mathcal{M}^{(l,q)}} = \frac{4ln^4q}{d\mathcal{B}(l,q)} \overset{(ii)}{\leq} 4l^{l+1}Cd^{4+4\beta(\bar{L}+\bar{\delta})-1-\beta l},
\end{aligned}$$

where in (i) we used Markov inequality (also known as Chebychev for second moments), and in (ii) we used rate of n and the bound $\mathcal{B}(l,q) = \binom{q}{l} \geq (q/l)^l$. Notice that the constant C does not depend on l. Now, we can compute the probability that $\max_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2 \leq 1/4n^2$ for all $l \geq T$ at the same time. Thanks to union bound this happens with probability $1 - \Delta$, where:

$$\begin{aligned}
\Delta &\leq \sum_{l=T}^{q} 4l^{l+1}Cd^{4+4\beta(\bar{L}+\bar{\delta})-1-\beta l} \\
&= 4Cd^{3+4\beta(\bar{L}+\bar{\delta})-\beta T}\sum_{l=T}^{\infty} lq^{-(l-T)} \\
&= 4Cd^{3+4\beta(\bar{L}+\bar{\delta})-\beta T}\left(\sum_{l'=0}^{\infty}(l'+T)q^{-l'}\right) \\
&\overset{(i)}{=} 4Cd^{3+4\beta(\bar{L}+\bar{\delta})-\beta T}\left(\frac{1}{(1-1/q)^2} + \frac{T-1}{1-1/q}\right) \leq 8TCd^{3+4\beta(\bar{L}+\bar{\delta})-\beta T},
\end{aligned}$$

where in (i) we used the convergence of a classical series. To conclude the proof, we remember that $T > \frac{4}{\beta} + 4(\bar{L}+\bar{\delta})$, hence, it is easy to verify that $\Delta \in \mathcal{O}(d^{-1})$. Which implies that:

$$\Pr\left(\forall l \geq T,\ \max_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2 \leq 1/4n^2\right) \leq 1 - cd^{-1}.$$

But thanks to Equation (28), the fact that $\max_{i,j}[\mathbf{W}_l^{(d,q)}]_{i,j}^2 \leq 1/4n^2$ implies that $\|\mathbf{W}_l^{(d,q)}\| \leq 1/2$, hence:

$$\Pr\left(\forall l \geq T,\ \|\mathbf{W}_l^{(d,q)}\| \leq 1/2\right) \leq 1 - cd^{-1}.$$

Concluding the proof. □

The next statement is a non-asymptotic version of Proposition 3 from Ghorbani et al. (2021).

Lemma 17 (Graph argument)**.** *Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a random matrix such that:*

- $\mathbf{W}_{i,i} = 0,\ \forall i \in 1,\dots,n.$
- *There exists a $\mathcal{M} > 0$, such that for any $i \neq j \in 1,\dots,n$:*

$$\left|\mathbb{E}_{x_j}[\mathbf{W}_{ij}\mathbf{W}_{jk}]\right| \leq \frac{1}{\mathcal{M}}|\mathbf{W}_{ik}|.$$

- *There exists a l such that, for any $p \in \mathbb{N}, p \geq 2$:*

$$\mathbb{E}\left[|\mathbf{W}_{ij}|^p\right]^{1/p} \leq \sqrt{\frac{p^l}{\mathcal{M}}}.$$

Then, we have that, for any $p \in \mathbb{N}, p > 2$:

$$\Pr\left(\|\mathbf{W}_l\| > 1/2\right) \leq c_1^p p^{3p} n \left(\frac{n}{\mathcal{M}}\right)^p + c_2^p \left(\frac{n}{\mathcal{M}}\right)^2,$$

where c_1 and c_2 depend on l.

Proof. Repeating the steps in the proof of Proposition 3 from Ghorbani et al. (2021), we get

$$\mathbb{E}[\|\mathbf{W}\|^{2p}] \leq \mathbb{E}\left[\mathbf{Tr}(\mathbf{W}^{2p})\right] \leq (cp)^{3p} \frac{n^{p+1}}{\mathcal{M}^p} + C^p \left(\frac{n}{\mathcal{M}}\right)^2.$$

Note that the proof in Ghorbani et al. (2021) assumes $\mathcal{M}$ to be in the order of d^l. We get rid of this assumption and keep $\mathcal{M}$ explicit. Furthermore, Ghorbani et al. (2021) use their Lemma 4 during their proof, but we use our Lemma 18 instead.

Now, we use Markov inequality to get a high probability bound:

$$\begin{aligned}\Pr(\|\mathbf{W}\| \geq 1/2) &= \Pr(\|\mathbf{W}\|^{2p} \geq (1/2)^{2p}) \\ &\leq \frac{\mathbb{E}[\|\mathbf{W}\|^{2p}]}{(1/2)^{2p}} = \left(\frac{c^3}{1/2}\right)^p p^{3p} \frac{n^{p+1}}{\mathcal{M}^p} + \left(\frac{C}{(1/2)^2}\right)^p \left(\frac{n}{\mathcal{M}}\right)^2.\end{aligned}$$

Which, renaming the constants, yields the thesis. □

Lemma 18 (Hypercontractivity)**.** *For any $l \in \mathbb{N}$, for any $p \geq 2$:*

$$\mathbb{E}\left[|\mathcal{Q}_l^{(d,q)}(x, x')|^p\right]^{1/p} \leq (p-1)^{l/2} \sqrt{\mathbb{E}\left[(\mathcal{Q}_l^{(d,q)}(x, x'))^2\right]},$$

where $\mathcal{Q}_l^{(d,q)}(x, x')$ is defined in Definition 2.

Proof. Let z be the entry-wise product between x and y. For any S, $\mathcal{Y}_S(x)\mathcal{Y}_S(x')$ depends only on z:

$$\mathcal{Y}_S(x)\mathcal{Y}_S(x') = \left(\prod_{i \in S} [x]_i\right)\left(\prod_{i \in S} [x']_i\right) = \prod_{i \in S} [x]_i [x']_i = \prod_{i \in S} [z]_i.$$

Hence, given the definition of $\mathcal{Q}_l^{(d,q)}(x, x')$, also $\mathcal{Q}_l^{(d,q)}(x, x')$ depends only on z. Finally, note that $z \sim \mathcal{U}(\{-1, 1\}^d)$. Hence, we can use hypercontractivity (Beckner, 1975), as for instance in Lemma 4 from Misiakiewicz & Mei (2021), to conclude the proof. □

D Optimal Regularization and Training Error

D.1 Optimal Regularization

In the main text we often refer to the optimal regularization λ_{opt}, however it is not clear how this regularization term can be computed. While we do not expect to compute it exactly, we show a way to compute the rate of it.

First we introduce some additional notation:

$$\begin{aligned}\eta_v(\ell_\lambda; \ell, \ell_\sigma, \beta) &\coloneqq \frac{-\ell_\sigma - \ell_\lambda}{\ell} - \frac{\beta}{\ell} \min\{\delta, 1 - \delta\}, \\ \eta_b(\ell_\lambda; \ell, L^*, \beta) &\coloneqq -2 - \frac{2}{\ell}(-\ell_\lambda - 1 - \beta(L^* - 1)), \\ \eta(\ell_\lambda; \ell, \ell_\sigma, L^*, \beta) &\coloneqq \max\left\{\eta_v(\ell_\lambda; \ell, \ell_\sigma, \beta), \eta_b(\ell_\lambda; \ell, L^*, \beta)\right\},\end{aligned}$$

where the dependency with ℓ_λ is also through $\delta = \frac{\ell - \ell_\lambda - 1}{\beta} - \left\lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \right\rfloor$. Note that Intuitively, in the setting of Theorem 1, we have that with high probability,

$$\begin{aligned}\textbf{Variance}(\hat{f}_\lambda) &\in \Theta(n^{\eta_v(\ell_\lambda; \ell, \ell_\sigma, \beta)}), \\ \textbf{Bias}^2(\hat{f}_\lambda) &\in \Theta(n^{\eta_b(\ell_\lambda; \ell, L^*, \beta)}), \\ \textbf{Risk}^2(\hat{f}_\lambda) &\in \Theta(n^{\eta(\ell_\lambda; \ell, \ell_\sigma, L^*, \beta)}).\end{aligned}$$

Hence, those quantities represent the exponent of the rate of variance, bias, and risk. We then define:

$$\ell_{\lambda_{\min}} := \arg\min_{\ell_\lambda \in [0,\bar{\ell}]} \eta(\ell_\lambda; \ell, \ell_\sigma, \beta),$$
$$\eta_{\min} := \min_{\ell_\lambda \in [0,\bar{\ell}]} \eta(\ell_\lambda; \ell, \ell_\sigma, \beta),$$
$$\bar{\ell} := \ell - 1 - \beta(L^* - 1).$$

Note that $\ell_{\lambda_{\min}}$ is a set that might have cardinality larger than one, but it cannot be empty. This is because η is a continuous function, and $[0, \bar{\ell}]$ is a closed set.

The quantity $\bar{\ell}$ defines the scope of our minimization searching domain, guaranteeing that the constraint on L^* in Theorem 1 is respected. For this reason, we are not considering the case where the optimal regularization parameter is larger than $\bar{\ell}$. Yet, simple intuition reveals that such a strong regularization parameter yields models where the risk does not even vanish. Thus, we still claim that looking for the optimum within $[0, \bar{\ell}]$ is sufficient.

The next Lemma goes more in detail about the properties of η, η_b, η_v.

Lemma 19 (Properties of η)**.** *Assuming that $\ell_\sigma > 0, \beta \in (0,1), L^* \in \left[1, \lceil \frac{\ell-1}{\beta} \rceil\right] \cap \mathbb{N}$. Let $\bar{\ell} = \ell - 1 - \beta(L^* - 1)$, then the following holds:*

1. *$\eta_v(l; \ell, \ell_\sigma, \beta)$ is a monotonically non-increasing function with respect to l, and $\eta_b(l; \ell, L^*, \beta)$ is a monotonically increasing one.*
2. *If there exists $\tilde{l} \in [0, \bar{\ell}]$ such that $\eta_v(\tilde{l}; \ell, \ell_\sigma, \beta) = \eta_b(\tilde{l}; \ell, L^*, \beta)$, then $\eta(\tilde{l}; \ell, \ell_\sigma, L^*, \beta) = \eta_{\min}$, otherwise, $\eta(0; \ell, \ell_\sigma, L^*, \beta) = \eta_{\min}$.*
3. *Let $\bar{l}$ be larger than l for any $l \in \ell_{\lambda_{min}}$, then:*
$$\eta(\bar{l}; \ell, \ell_\sigma, L^*, \beta) = -2 - \frac{2}{\ell}(-\bar{l} - 1 - \beta(L^* - 1)).$$
4. *Let $\underline{l}$ be smaller than l for any $l \in \ell_{\lambda_{min}}$, let c_{ℓ_σ} be a constant that depends only on ℓ_σ, ℓ, β. Then, if $\eta(\underline{l}; \ell, \ell_\sigma, \beta) \leq \eta_{\min} + c_{\ell_\sigma}$:*
$$\eta(\underline{l}; \ell, \ell_\sigma, \beta) = \eta_{\min} + \frac{2}{\ell}(\ell_{\lambda_{opt}} - \underline{l}).$$

Rate of the optimal regularization parameter vs. optimal rate of the regularization parameter: for the optimal regularization parameter let $\ell_{\lambda_{\text{opt}}}$ be such that $\lambda_{\text{opt}} \in \Theta(d^{\ell_{\lambda_{\text{opt}}}})$. Then it is not clear a priori that $\ell_{\lambda_{\text{opt}}}$ minimizes η. However, Lemma 20 bridges the two terms and guarantees that this is the case with high probability.

Lemma 20 (Optimal regularization)**.** *In the same setting as in Theorem 1, for d sufficiently large, with probability at least $1 - cd^{-\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$, there exists $l \in \ell_{\lambda_{min}}$ such that*

$$\max\{\lambda_{opt}, 1\} \in \Theta\left(d^l\right).$$

Finding optimal parameters: now that we understand the importance of finding the minimal rate, we notice that Lemma 19 suggests a simple strategy to find a $l \in \ell_{\lambda_{\min}}$. We start by searching the intersection of η_v and η_b in $[0, \bar{\ell}]$: if it is found, then the intersection point is optimal, otherwise just choose $l = 0$. Note that the intersection is unique and, if exists, it is easy to compute since η_v is monotonically non-decreasing, and η_b is monotonically increasing.

Multiple minimizer issue: the last caveat here is that $\ell_{\lambda_{\min}}$ might contain multiple values, and, in that case we are not guaranteed that our algorithm retrieves the one which is the rate of λ_{opt}. It might instead find a different $l \in \ell_{\lambda_{\min}}$. Yet, Theorem 1 guarantees that both the optimally regularized estimator and any estimator regularized with $\lambda \in d^l, l \in \ell_{\lambda_{\min}}$, have a risk vanishing with the same rate $n^{\eta_{\min}}$ with high probability. This made it possible to plot the rate of the optimally regularized estimator in Figure 1a.

Closed formula solutions: because of the phenomenon of multiple descent, it is hard to write both $\ell_{\lambda_{\text{opt}}}$ or β^* in an easily readable closed form. Still, simple optimization algorithms make it easy to retrieve such quantities.

Proof of Lemma 20. Let $\ell_{\lambda_{\text{opt}}}$ be defined such that:

$$\max\{\lambda_{\text{opt}}, 1\} \in \Theta\left(d^{\ell_{\lambda_{\text{opt}}}}\right).$$

Combining the bound on the bias and the variance in Theorem 1, we get that for any ℓ_λ, $\lambda \in \Theta(\ell_\lambda)$,

$$\begin{aligned}
\textbf{Risk}^2(\hat{f}_\lambda) &= \textbf{Bias}^2(\hat{f}_\lambda) + \textbf{Variance}^2(\hat{f}_\lambda) \\
&\in \Theta\left(n^{-2-\frac{2}{\ell}(-\ell_\lambda - 1 - \beta(L^*-1))} + n^{\frac{-\ell_\sigma - \ell_\lambda}{\ell} - \frac{\beta}{\ell}\min\{\delta, 1-\delta\}}\right) \\
&\in \Theta\left(n^{\eta(\ell_\lambda; \ell, \ell_\sigma, L^*, \beta)}\right), \quad \text{with prob. at least } 1 - cd^{-\min\{\bar{\delta}, 1-\bar{\delta}\}}. \qquad (29)
\end{aligned}$$

Throughout the rest of the proof we will write $\eta(\ell_\lambda)$ to indicate $\eta(\ell_\lambda; \ell, \ell_\sigma, L^*, \beta)$, for simplicity. We will also not repeat that each step true with probability at least $1 - cd^{-\min\{\bar{\delta}, 1-\bar{\delta}\}}$, but we are implying it. Now we want to prove that $\eta(\ell_{\lambda_{\text{opt}}})$ is close to $\eta_{\min}$. Hence, we notice that Equation (29) applied to $\ell_{\lambda_{\text{opt}}}$ and any $l \in \ell_{\lambda_{\min}}$ implies that:

$$\begin{aligned}
\textbf{Risk}^2(\hat{f}_l) &\leq c_2 n^{\eta(l)} = c_2 n^{\eta_{\min}}, \\
\textbf{Risk}^2(\hat{f}_{\lambda_{\text{opt}}}) &\geq c_1 n^{\eta(\lambda_{\text{opt}})},
\end{aligned}$$

where $0 < c_1 < c_2$ are the constants hidden by Θ in Equation (29). Now, we use the optimality of λ_{opt}:

$$\begin{aligned}
& c_1 n^{\eta(\lambda_{\text{opt}})} \leq \textbf{Risk}^2(\hat{f}_{\lambda_{\text{opt}}}) \leq \textbf{Risk}^2(\hat{f}_l) \leq c_2 n^{\eta_{\min}} \\
\Rightarrow \quad & n^{\eta(\lambda_{\text{opt}}) - \eta_{\min}} \leq \frac{c_2}{c_1} \quad \Rightarrow \quad \eta(\lambda_{\text{opt}}) - \eta_{\min} \leq \frac{\log(c_2/c_1)}{\log n}.
\end{aligned}$$

To conclude the proof we need to somehow show that there exists $l \in \ell_{\lambda_{\min}}$ sufficiently close to $\ell_{\lambda_{\text{opt}}}$. Formally we need to show that:

$$d^{\ell_{\lambda_{\text{opt}}} - l} \in \Theta(1). \qquad (30)$$

Once we have shown that this is the case, we can simply use the definition of $\ell_{\lambda_{\text{opt}}}$ and compute:

$$\lambda_{\text{opt}} \in \Theta(d^{\ell_{\lambda_{\text{opt}}}}) = \Theta(d^l d^{\ell_{\lambda_{\text{opt}}} - l}) \overset{(i)}{=} \Theta(d^l),$$

where in (i) we use Equation (30). This concludes the proof. Now we need to show that Equation (30) actually holds. To do so, we distinguish between three cases:

1. *($\ell_{\lambda_{opt}} \in \ell_{\lambda_{min}}$):* in this case we simply notice that we can take $l \in \ell_{\lambda_{\min}}$ to be exactly $\ell_{\lambda_{\text{opt}}}$, hence:
$$d^{\ell_{\lambda_{\text{opt}}} - l} = d^0 \in \Theta(1).$$

2. *($\ell_{\lambda_{opt}} < l$, for any $l \in \ell_{\lambda_{min}}$):* let $\underline{l}$ be $\min(\ell_{\lambda_{\min}})$. Then, thanks to Lemma 19, we have that, as long as $\eta(\ell_{\lambda_{\text{opt}}}) - \eta_{\min} \leq c$,
$$\eta(\ell_{\lambda_{\text{opt}}}) = \eta_{\min} + \frac{2}{\ell}(\underline{l} - \ell_{\lambda_{\text{opt}}}).$$
Notice that for n sufficiently large, it is always the case that $\eta(\ell_{\lambda_{\text{opt}}}) - \eta_{\min} \leq c$. Hence, we have that:
$$\underline{l} - \ell_{\lambda_{\text{opt}}} = \frac{\ell}{2}(\eta(\ell_{\lambda_{\text{opt}}}) - \eta_{\min}) \leq \frac{\ell \log(c_2/c_1)}{\log n}.$$
Now, notice that since $\ell_{\lambda_{\text{opt}}} \leq \underline{l}$, $d^{\ell_{\lambda_{\text{opt}}} - \underline{l}} \in \mathcal{O}(1)$, furthermore:
$$d^{\underline{l} - \ell_{\lambda_{\text{opt}}}} \leq d^{\frac{\ell \log(c_2/c_1)}{\log n}} = d^{\frac{\log(c_2/c_1)}{\log d}} = c_2/c_1 \in \mathcal{O}(1),$$
where in (i) we used that $n \in \Theta(d^\ell)$. Since both $d^{\underline{l} - \ell_{\lambda_{\text{opt}}}}$ and its inverse are $\in \mathcal{O}(1)$, and $\underline{l} \in \ell_{\lambda_{\min}}$, we can conclude that there exists a $l \in \ell_{\lambda_{\min}}$, such that:
$$d^{\ell_{\lambda_{\text{opt}}} - l} \in \Theta(1).$$

3. *($\ell_{\lambda_{opt}} > l$, for any $l \in \ell_{\lambda_{min}}$):* let $\bar{l}$ be $\max(\ell_{\lambda_{\min}})$. Then, thanks to Lemma 19, we have that:

$$\begin{aligned}\eta(\ell_{\lambda_{\text{opt}}}) &= -2 - \frac{2}{\ell}(-\ell_{\lambda_{\text{opt}}} - 1 - \beta(L^* - 1)) \\ &= -2 - \frac{2}{\ell}(-\ell_{\lambda_{\text{opt}}} - \bar{l} + \bar{l} - 1 - \beta(L^* - 1)) \\ &= -2 - \frac{2}{\ell}(-\bar{l} - 1 - \beta(L^* - 1)) - \frac{2}{\ell}(\bar{l} - \ell_{\lambda_{\text{opt}}}) \\ &= \eta_{\min} + \frac{2}{\ell}(\ell_{\lambda_{\text{opt}}} - \bar{l})\end{aligned}$$

Hence, we have that:

$$\ell_{\lambda_{\text{opt}}} - \bar{l} = \frac{\ell}{2}(\eta(\ell_{\lambda_{\text{opt}}}) - \eta_{\min}) \leq \frac{\ell \log(c_2/c_1)}{\log n}.$$

Analogously to the previous case, this implies that $d^{\ell_{\lambda_{\text{opt}}} - \bar{l}} \in \mathcal{O}(1)$, and the fact that $\ell_{\lambda_{\text{opt}}} > \bar{l}$ implies that $d^{\ell_{\lambda_{\text{opt}}} - \bar{l}} \in \Omega(1)$. Together, this means that there exists a $l \in \ell_{\lambda_{\min}}$ such that:

$$d^{\ell_{\lambda_{\text{opt}}} - l} \in \Theta(1).$$

□

Proof of Lemma 19. Throughout this proof we will drop the dependencies $\ell, \ell_\sigma, L^*, \beta$ in the notation of η, η_v, η_b, simply writing $\eta(l), \eta_v(l), \eta_b(l)$.

Continuity: the proof start by showing that $\eta_v(l)$ and $\eta_b(l)$, are continuous functions of l. $\eta_b(l)$ is an affine function of l, hence it is also continuous. $\eta_v(l)$, however, has a hidden dependency with l through δ, which is not linear. Hence, to show that $\eta_v(l)$ is continuous, we need to show that $\min\{\delta, 1 - \delta\}$ is continuous. This is equivalent to show that:

$$\varpi(t) \coloneqq \min\{t - \lfloor t \rfloor, 1 - (t - \lfloor t \rfloor)\}$$

is continuous. This is because $\varpi\left(\frac{\ell - l - 1}{\beta}\right) = \min\{\delta, 1 - \delta\}$, and $\frac{\ell - l - 1}{\beta}$ is a linear function of l.

Finally, it is easy to see that ϖ is the triangular wave, and hence, it is indeed continuous. Both $\eta_v(l)$ and $\eta_b(l)$ are then continuous functions, implying that $\eta(l)$ is alsoncontinuous.

Derivative: now we can move our attention to the derivative of $\eta_v(l)$ and $\eta_b(l)$. Again, it is easy to see that:

$$\partial_l \eta_b(l) = \frac{2}{\ell},$$

since η_b is just an affine function. This also implies that $\eta_b(l)$ is a monotonically increasing function.

Again, the reasoning is a bit more complicated for η_v. Here there are two cases:

$$\begin{aligned}\partial_l \eta_v(l) &= \begin{cases} \partial_l \frac{-\ell_\sigma - l - \beta\delta}{\ell} & \delta < 1 - \delta \\ \partial_l \frac{-\ell_\sigma - l - \beta(1-\delta)}{\ell} & \delta \geq 1 - \delta \end{cases} \\ &= \begin{cases} 0 & \delta < 1 - \delta \\ -\frac{1}{\ell} & \delta \geq 1 - \delta \end{cases}\end{aligned}$$

Since η_v is continuous, and the derivative is always ≤ 0, then η_v is monotonically non-increasing function. Concluding the proof of Item 1.

Since η_b is monotonically increasing, and η_v monotonically non-increasing: if there exist $\tilde{l} \in \left[0, \bar{\ell}\right]$ such that for any $\eta_b(\tilde{l}) = \eta_b(\tilde{l})$, then, for any $l > \tilde{l}, \eta_b(l) > \eta_v(l)$, and for any $l < \tilde{l}, \eta_b(l) < \eta_v(l)$.

Notice that such $\tilde{l}$ might not exist since it might be the case that $\eta_b(l) > \eta_v(l)$ for any $l \in \left[0, \bar{\ell}\right]$.

η function decomposition: Notice now that η is defined as the maximum between η_v and η_b. Hence, if $\tilde{l}$ exists, then:

$$\eta(l) = \begin{cases} \eta_v(l) & l < \tilde{l}, \\ \eta_b(l) & \text{otherwise.} \end{cases}$$

If however, $\tilde{l}$ does not exist, it means that $\eta_v(l)$ does not intersect $\eta_b(l)$. Thanks to continuity, either $\eta_b(l) > \eta_v(l)$ for any $l \in \left[0, \bar{\ell}\right]$, or $\eta_v(l) > \eta_b(l)$. A more accurate analysis, however, shows that the latter is not possible. To see this, we simply compute both $\eta_v(\bar{\ell})$ and $\eta_b(\bar{\ell})$:

$$\begin{aligned} \eta_b(\bar{\ell}) &= -2 - \frac{2}{\ell}\left(-\left(\ell - 1 - \beta(L^* - 1)\right) - \beta(L^* - 1)\right) \\ &= 0 \\ \eta_v(\bar{\ell}) &= -\frac{\ell_\sigma + (\ell - 1 - \beta(L^* - 1))}{\ell} - \frac{\beta}{\ell}\min\{\delta, 1 - \delta\} \\ &\overset{(i)}{\geq} \frac{(\ell - 1 - \beta(L^* - 1))}{\ell} \overset{(ii)}{\geq} 0, \end{aligned}$$

where (i) follows from the positivity of ℓ_σ , and the positivity of $\min\{\delta, 1 - \delta\}$. Step (ii) follows from the assumption on L^*. Hence, in the case $\tilde{l}$ does not exist, since the two functions are continuous $\eta_v(l) \leq \eta_b(l), \forall l \in [0, \bar{\ell}]$. We can conclude that $\eta(l)$ can be written as:

$$\eta(l) = \begin{cases} \eta_v(l) & \tilde{l} \text{ exists, and } l < \tilde{l}, \\ \eta_b(l) & \text{otherwise.} \end{cases}$$

Furthermore, since η_v is non-increasing, and η_b is decreasing, we have that, in the case where $\tilde{l}$ exists for any l, $\eta(\tilde{l}) \leq \eta(l)$ for any other $l \in \left[0, \bar{\ell}\right]$. Hence, $\tilde{l} \in \ell_{\lambda_{\text{min}}}$. In the case where $\tilde{l}$ does not exist, we have that $\eta(l) = \eta_b(l)$, which is a monotonically increasing function. Hence, it is minimal in 0. This proves Item 2.

Proof of Item 3 and Item 4: Item 3 follows directly from the fact that $\eta(l) = \eta_b(l)$ if $\tilde{l}$ does not exist, or for any $l \geq \tilde{l}$, and $\tilde{l} \in \ell_{\lambda_{\text{min}}}$.

Item 4 requires some additional steps. In the case where $\tilde{l}$ does not exist, we have that the $\ell_{\lambda_{\text{min}}} = \{0\}$, hence the result follows directly.

Otherwise, let $l_{\min} = \min \ell_{\lambda_{\text{min}}}$: with the same argument as for Item 3, we have that $\eta(l) = \eta_v(l)$ for any $l \leq l_{\min}$.

The function $\eta_v(l)$ alternates steps with derivative 0, with steps with derivative $-2/\ell$. We argue that there exists a left neighborhood of $l_{\min}$ where the derivative is $-2/\ell$. By contradiction, assume that this neighborhood does not exist: since the function $\eta_v(l)$ alternates between intervals of fixed size with derivative $-2/\ell$ and 0, it would imply that there exists a left neighborhood where the derivative is 0. But this would hence mean that within this neighborhood, the function is constant, and all the values $\eta_v(l)$ are equal to $\eta_{\min}$, which contradicts the hypothesis that $l_{\min} = \min \ell_{\lambda_{\text{min}}}$. Hence, there exist such left neighborhood. Let $\varepsilon_{\ell_\sigma}$ be the diameter of such interval, then, for any $l \in [l_{\min} - \varepsilon_{\ell_\sigma}, l_{\min}]$

$$\eta_v(l) = \eta_{\min} - \frac{2}{\ell}(-\varepsilon_{\ell_\sigma}).$$

Finally, since the function is non-decreasing, as long as $\eta_v(l) < \eta_{\min} + \frac{2\varepsilon_{\ell_\sigma}}{\ell}$, we have that l must be larger than $l_{\min} - \varepsilon_{\ell_\sigma}$. Taking $c_{\ell_\sigma} = \frac{2\varepsilon_{\ell_\sigma}}{\ell}$, concludes the proof of Item 4. □

D.2 Proof of Theorem 2: Training error

The formal version Theorem 2 states a bound for the expected training error in two settings:

- The optimally regularized model achieves a better risk rate than the interpolating one.
- The optimally regularized model achieves the same risk rate as the interpolating one.

The quantity β^* is defined in the main text as the intersection between variance and bias. Since regularization benefits the variance and hurts the bias, it is the case that only for values of $\beta > \beta^*$, i.e., where variance dominates bias, that the regularization helps. This how we state the formal version of Theorem 2 as in Theorem 4.

Theorem 4 (Formal version of Theorem 2)**.** *In the same setting as Theorem 1, let us fix $\ell > 0, \beta \in (0,1), \ell_\sigma \geq 0, \bar{\ell} = \ell - 1 - \beta(L^* - 1)$. Finally, let λ_{opt} be such that **Risk**$(\hat{f}_{\lambda_{opt}})$ is minimal.*

1. *If $l \in [0, \bar{\ell}]$ that minimizes $\eta(l; \ell, \ell_\sigma, L^*, \beta)$ is such that $\eta(l; \ell, \ell_\sigma, L^*, \beta) < \eta(0; \ell, \ell_\sigma, L^*, \beta)$, then:*

$$\left| \mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_{\lambda_{opt}}(x_i) - y_i)^2 \right] - \sigma^2 \right| \in \mathcal{O}(d^{-l} + \textbf{\textit{Risk}}^2(\hat{f}_{\lambda_{opt}})), \quad \textit{w.p.} \ \geq 1 - cd^{-\beta \min\{\delta, 1-\delta\}}.$$

2. *If, for any $l \in (0, \bar{\ell}]$, $\eta(l; \ell, \ell_\sigma, L^*, \beta) > \eta(0; \ell, \ell_\sigma, L^*, \beta)$, then:*

$$\mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_{\lambda_{opt}}(x_i) - y_i)^2 \right] \leq c\sigma^2 + \mathcal{O}\left(\textbf{\textit{Risk}}^2(\hat{f}_{\lambda_{opt}}) \right), \quad \textit{w.p.} \ \geq 1 - cd^{-\beta \min\{\delta, 1-\delta\}}.$$

For some constant $c < 1$.

This theorem covers the case where either regularization is harmful, or it is beneficial. There are still some edge cases where, for small ℓ_λ, the rate of the risk is the same as for the interpolating model, and for large ℓ_λ, the performance deteriorate. In this case our theory does not allow discerning weather the expected training matches the noise level or not.

Prood of Theorem 4. Throughout the proof we will use with high probability to mean with probability at least $1 - cd^{-\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$. Thanks to Lemma 20, we know that $\max\{\lambda_{\text{opt}}, 1\} \in \Theta(d^l)$, where l minimizes $\eta(l; \ell, \ell_\sigma, L^*, \beta)$. The condition in Item 1 guarantees that the optimal l is positive, and the condition in Item 1 guarantees that the optimal l is 0.

Harmful interpolation setting. In this setting we have that $\lambda_{\text{opt}} \in \Theta(d^l), l > 0$. We start by applying Theorem 1 with $\ell_\lambda = l$ in this setting. This confirms that, with the same m chosen for the proof, Equation (5) holds with high probability, and the ground truth is in the span of the first m eigenfunctions. Hence, we apply Lemma 22:

$$\frac{\lambda_{\text{opt}}^2 \sigma^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_{\lambda_{\text{opt}}}(x_i) - y_i)^2 \right] \leq \frac{\lambda_{\text{opt}}^2 \sigma^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) + 12 \lambda_{\text{opt}}^2 \frac{r_2^2}{r_1^2} \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2}.$$

Then, by subtracting σ^2 and taking the absolute value we get:

$$\left| \mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_{\lambda_{\text{opt}}}(x_i) - y_i)^2 \right] - \sigma^2 \right| \leq \underbrace{\sigma^2 \left| \frac{\lambda_{\text{opt}}^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \right|}_{:= T_1} + \underbrace{12 \left| {\lambda_{\text{opt}}}^2 \frac{\|\mathbf{D}_{\leq m}{}^{-1} a\|^2}{n^2} \right|}_{:= T_2}.$$

To bound T_2 we do:

$$\begin{aligned}
\lambda_{\text{opt}}^2 \frac{\|\mathbf{D}_{\leq m}{}^{-1} a\|^2}{n^2} &\overset{(i)}{\in} \Theta\left(\frac{d^{2l}}{d^{2\ell}} r \|\mathbf{D}_{\leq m}{}^{-1} a\|^2 \right) \\
&\overset{(ii)}{\subseteq} \Theta\left(\frac{d^{2l}}{d^{2\ell}} d^2 \cdot q^{2(L^*-1)} \right) \\
&= \Theta\left(\frac{d^{2l}}{d^2 d^{2l} d^{2(\ell - l - 1)}} d^2 \cdot q^{2(L^*-1)} \right) \\
&= \Theta\left(d^{-2(\ell - \ell_\lambda - 1 - \beta(L^*-1))} \right) \\
&\overset{(iii)}{=} \Theta\left(\mathbf{Bias}^2(\hat{f}_{\lambda_{\text{opt}}}) \right) \in \mathcal{O}\left(\mathbf{Risk}^2(\hat{f}_{\lambda_{\text{opt}}}) \right),
\end{aligned} \tag{31}$$

where in (i) we used that $n \in \Theta(d^\ell), \lambda_{\text{opt}} \in \Theta(d^l)$. In step (ii) we used Lemma 2 to have the rate of $\|\mathbf{D}_{\leq m}{}^{-1}a\|$ and Lemma 3 for $r_2^2, r_1^2 \in \Theta(1)$. In (iii) we see that this rate matches the rate of the bias found in Theorem 1.

To bound T_1 we apply Lemma 21, which yields with high probability:

$$\begin{aligned}
\frac{n - |\mathbb{N} \setminus \mathcal{I}_2|}{(c_2 + \lambda_{\text{opt}})^2} &\leq \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \frac{n}{(c_1 + \lambda_{\text{opt}})^2} \\
\frac{n - |\mathbb{N} \setminus \mathcal{I}_2|}{n} \frac{\lambda_{\text{opt}}^2}{(c_2 + \lambda_{\text{opt}})^2} &\leq \frac{\lambda_{\text{opt}}^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \frac{\lambda_{\text{opt}}^2}{(c_1 + \lambda_{\text{opt}})^2} \\
-\frac{|\mathbb{N} \setminus \mathcal{I}_2|}{n} \frac{\lambda_{\text{opt}}^2}{(c_2 + \lambda_{\text{opt}})^2} + \frac{\lambda_{\text{opt}}^2}{(c_2 + \lambda_{\text{opt}})^2} - 1 &\leq \frac{\lambda_{\text{opt}}^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \leq \frac{\lambda_{\text{opt}}^2}{(c_1 + \lambda_{\text{opt}})^2} - 1 \\
-\frac{|\mathbb{N} \setminus \mathcal{I}_2|}{n} \frac{\lambda_{\text{opt}}^2}{(c_2 + \lambda_{\text{opt}})^2} - \frac{2c_2\lambda_{\text{opt}} + c_2^2}{(c_2 + \lambda_{\text{opt}})^2} &\leq \frac{\lambda_{\text{opt}}^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \leq -\frac{2c_1\lambda_{\text{opt}} + c_1^2}{(c_1 + \lambda_{\text{opt}})^2}.
\end{aligned}$$

Since c_1 and c_2 are constants, we can bound the absolute value of the inner part as:

$$\begin{aligned}
\left| \frac{\lambda_{\text{opt}}^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \right| &\leq \frac{c + \lambda_{\text{opt}}}{(c + \lambda_{\text{opt}})^2} + \frac{|\mathbb{N} \setminus \mathcal{I}_2|}{n} \frac{\lambda_{\text{opt}}^2}{(c + \lambda_{\text{opt}})^2} \\
&\overset{(i)}{\lesssim} \frac{d^l}{d^{2l}} + \frac{nq^{-\bar{\delta}}}{n} \frac{\lambda_{\text{opt}}^2}{(c + \lambda_{\text{opt}})^2} \in \Theta\left(d^{-l} + q^{-\bar{\delta}}\right) \in \mathcal{O}(d^{-l}),
\end{aligned}$$

where in (i) we use the rate of $|\mathbb{N} \setminus \mathcal{I}_2|$ computed in Lemma 21 itself. Concluding the proof for this case.

Harmless interpolation setting. In this setting we know that the minimizer of the rate is 0, and hence, thanks to Lemma 20, we know that:

$$\lambda_{\text{opt}} \in \Theta(d^l) = \Theta(d^0) = \Theta(1), \quad \text{with high prob.}$$

Like in the other setting, we start by applying Theorem 1 with $\ell_\lambda = l = 0$. Hence, we apply Lemma 22 and look at the upper bound:

$$\mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_{\lambda_{\text{opt}}}(x_i) - y_i)^2 \right] \leq \underbrace{\frac{\lambda_{\text{opt}}^2 \sigma^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right)}_{T_1} + \underbrace{12\lambda_{\text{opt}}^2 \frac{r_2^2}{r_1^2} \frac{\|\mathbf{D}_{\leq m}^{-1} a\|^2}{n^2}}_{T_2}.$$

For T_2 we move analogously to the previous case (Equation (31)) in the specific case where $l = 0$. This yields:

$$T_2 \in \Theta\left(\mathbf{Risk}^2(\hat{f}_{\lambda_{\text{opt}}})\right).$$

To bound T_1 we apply Lemma 21, which yields with high probability:

$$\begin{aligned}
\mathbf{Tr}\left(\mathbf{H}^{-2}\right) &\leq \frac{n}{(c_1 + \lambda_{\text{opt}})^2} \\
\frac{\lambda_{\text{opt}}^2 \sigma^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) &\leq \sigma^2 \frac{\lambda_{\text{opt}}^2}{(c_1 + \lambda_{\text{opt}})^2} \\
&\overset{(i)}{\leq} \sigma^2 \frac{C}{(c_1 + C)^2},
\end{aligned}$$

where in (i) we used that with high probability $\lambda_{\text{opt}} \in \Theta(1)$, hence there exists a $C > 0$ such that $\lambda_{\text{opt}} \leq C$ with high probability. Finally, it is clear that $\frac{C}{(c_1+C)^2} < 1$, Concluding the proof. □

D.3 Technical Lemmas

Lemma 21 (Trace of the inverse)**.** *In the setting defined in Section 3.1. Then, with high probability,*

$$\frac{n - |\mathbb{N} \setminus \mathcal{I}_2|}{(c_2 + \lambda)^2} \leq \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \frac{n}{(c_1 + \lambda)^2}, \tag{32}$$

for two positive constants c_1, c_2. Furthermore, $|\mathbb{N} \setminus \mathcal{I}_2|$, $\mathcal{I}_2, \bar{\delta}$ defined as in Definition 2.

Proof. We start by computing $|\mathbb{N} \setminus \mathcal{I}_2|$. Computing $|\mathbb{N} \setminus \mathcal{I}_2|$ corresponds to counting how many indexes do not appear in $\mathcal{I}_2$. Give the definition of the latter, for an index not to be in $\mathcal{I}_2$, it means that $|S_k| \leq \bar{L} + 1$. Hence, the total number of such index can be at most:

$$|\mathbb{N} \setminus \mathcal{I}_2| \leq \sum_{l=1}^{\bar{L}+1} \mathcal{C}(l, q, d) = \sum_{l=1}^{\bar{L}+1} d \binom{q-1}{l-1} \overset{(i)}{\in} \Theta(d \cdot q^{\bar{m}}),$$

where in (i) we use the bound the binomial coefficient $\binom{q-1}{l-1} \leq \left(e\frac{q-1}{l-1}\right)^{l-1}$. Finally, since $n \in \Theta(d \cdot q^{\bar{L}+\bar{\delta}})$, we have that $|\mathbb{N} \setminus \mathcal{I}_2| \in \mathcal{O}(nq^{-\bar{\delta}})$.

Now we move to the proof of Equation (32). We divide $\mathbf{H} = \mathbf{K}_{\leq m} + \mathbf{K}_1 + \mathbf{K}_2 + \lambda \mathbf{I}$, where $\mathbf{K}_1$ and $\mathbf{K}_2$ are defined as in Definition 2. Now, we define $\mathbf{H}_2 = \mathbf{K}_2 + \lambda \mathbf{I}$. For the lower bound it suffices to see:

$$\mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq n\|\mathbf{H}^{-2}\| = \frac{n}{\mu_{\min}\left(\mathbf{H}\right)^2} \leq \frac{n}{\mu_{\min}\left(\mathbf{H}_2\right)^2} \overset{(i)}{\leq} \frac{n}{(c_1+\lambda)^2}, \quad \text{with prob. } \geq 1 - cd^{-\beta(1-\bar{\delta})},$$

for some constant c_1. In (i) we used Lemma 13 to bound $\mu_{\min}\left(\mathbf{K}_2\right)$.

For the upper bound we need a more refined analysis. Let us define the following matrix:

$$\mathbf{K}_{-2} \coloneqq \mathbf{K}_{\leq m} + \mathbf{K}_1.$$

Notice that $\mathbf{K} = \mathbf{K}_{-2} + \mathbf{K}_2$, which gives an intuition for the notation. Thanks to Lemma 13, we know that $\mathbf{K}_2$ is non-singular wit probability at least $1 - cd^{-\beta(1-\bar{\delta})}$. Furthermore, since $\mathbf{K}_{-2}$ is still positive semi-definite, also $\mathbf{K}$ and $\mathbf{H}$ and $\mathbf{H}_2$ are non-singular. Thus, we can apply the classical linear algebra formula

$$\mathbf{H}^{-1} = (\mathbf{H}_2 + \mathbf{K}_{-2})^{-1} = \mathbf{H}_2^{-1} - \underbrace{\mathbf{H}_2^{-1}\mathbf{K}_{-2}(\mathbf{H}_2 + \mathbf{K}_{-2})^{-1}}_{A}.$$

Since $\mathbf{K}_{-2}$ is the kernel matrix that stems from a kernel with exactly $|\mathbb{N} \setminus \mathcal{I}_2|$ features, it has rank at most $|\mathbb{N} \setminus \mathcal{I}_2|$. Since A is defined as the product of some matrices including $\mathbf{K}_{-2}$, thus, it also has rank at most $|\mathbb{N} \setminus \mathcal{I}_2|$.

Notice furthermore, that since $|\mathbb{N} \setminus \mathcal{I}_2| \in \mathcal{O}(nd^{-\beta\bar{\delta}}) \subseteq o(n)$, we have that $|\mathbb{N} \setminus \mathcal{I}_2| < n$ for d sufficiently large. Let now $\{v_1, \ldots, v_{|\mathbb{N}\setminus\mathcal{I}_2|}\}$ be an orthonormal basis of $\mathrm{col}(A)$, and $\{v_{|\mathbb{N}\setminus\mathcal{I}_2|+1}, \ldots, v_n\}$ an orthonormal basis of $\mathrm{col}(A)^{\perp}$. This makes $\{v_1, \ldots, v_n\}$ a full orthonormal basis of $\mathbb{R}^n$. Now, since the trace is invariant to similarities, we get that:

$$\begin{aligned}
\mathbf{Tr}(\mathbf{H}^{-2}) &= \sum_{i=1}^{n} v_i^{\intercal}\mathbf{H}^{-2}v_i = \sum_{i=1}^{|\mathbb{N}\setminus\mathcal{I}_2|} v_i^{\intercal}\mathbf{H}^{-2}v_i + \sum_{i=|\mathbb{N}\setminus\mathcal{I}_2|+1}^{n} v_i^{\intercal}\mathbf{H}^{-2}v_i \\
&\geq \sum_{i=|\mathbb{N}\setminus\mathcal{I}_2|+1}^{n} v_i^{\intercal}\left(\mathbf{H}_2{}^{-1} - A\right)^2 v_i \\
&= \sum_{i=|\mathbb{N}\setminus\mathcal{I}_2|+1}^{n} \left(v_i^{\intercal}\mathbf{H}_2{}^{-2}v_i - \underbrace{v_i^{\intercal}A}_{=0}\mathbf{H}_2{}^{-1}v_i - v_i^{\intercal}\mathbf{H}_2{}^{-1}\underbrace{Av_i}_{=0} + \underbrace{v_i^{\intercal}AAv_i}_{=0} \right) \\
&\overset{(i)}{=} \sum_{i=|\mathbb{N}\setminus\mathcal{I}_2|+1}^{n} v_i^{\intercal}\mathbf{H}_2{}^{-2}v_i \\
&\geq (n - |\mathbb{N}\setminus\mathcal{I}_2|)\mu_{\min}\left(\mathbf{H}_2{}^{-2}\right)\|v_i\| \overset{(ii)}{=} (n - |\mathbb{N}\setminus\mathcal{I}_2|)\frac{1}{\|\mathbf{H}_2\|^2} \\
&\overset{(iii)}{\geq} (n - |\mathbb{N}\setminus\mathcal{I}_2|)\frac{1}{(c_2+\lambda)^2}, \quad \text{with prob.} \geq 1 - cd^{-\beta(1-\bar{\delta})},
\end{aligned}$$

where in (i) we use the fact that, as long as $i > |\mathbb{N} \setminus \mathcal{I}_2|$, we have that $Av_i = v_i^{\intercal}A = 0$,since v_i is orthogonal to the column space. In (ii) we used the fact that the vectors v_i in the orthonormal basis have norm 1, and in (iii) we used Lemma 13 to bound $\|\mathbf{K}_2\|$. □

Lemma 22 (Fixed design training error)**.** *In the same setting as Theorem 3, let m be such that Equation* (5) *holds and the ground truth is in the span of the first m eigenfunctions.*

$$\frac{\lambda^2\sigma^2}{n}\mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \mathbb{E}_\epsilon\left[\frac{1}{n}\sum_i(\hat{f}(x_i)-y_i)^2\right] \leq \frac{\lambda^2\sigma^2}{n}\mathbf{Tr}\left(\mathbf{H}^{-2}\right) + 12\lambda^2\frac{r_2^2}{r_1^2}\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2}.$$

Proof. In kernel ridge regression, our estimator is:

$$\hat{f}_\lambda(x^*) = \mathbf{y}^\intercal\mathbf{H}^{-1}\mathbf{k}(x^*),$$

where $\mathbf{y} \coloneqq [y_1,\ldots,y_n]^\intercal$ and $\mathbf{k}(x^*) \coloneqq [\mathcal{K}(x_1,x^*),\ldots,\mathcal{K}(x_n,x^*)]^\intercal$, $\mathbf{H} = \mathbf{K} + \lambda\mathbf{I}_n$. Hence, our estimator, evaluated at $x_1,\ldots,x_n$ is $\mathbf{K}\mathbf{H}^{-1}\mathbf{y}$. Hence, the training error is:

$$\frac{1}{n}\sum_i(\hat{f}_\lambda(x_i)-y_i)^2 = \frac{1}{n}\|\left(\mathbf{I}-\mathbf{H}^{-1}\mathbf{K}\right)\mathbf{y}\|^2 \overset{(i)}{=} \frac{\lambda^2}{n}\|\mathbf{H}^{-1}\mathbf{y}\|^2$$

Where (i) follows from the simple calculation:

$$\mathbf{I}-\mathbf{H}^{-1}\mathbf{K} = \mathbf{I}_n - (\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{K} = \mathbf{I}_n - (\mathbf{K}+\lambda\mathbf{I}_n)^{-1}(\mathbf{K}+\lambda\mathbf{I}_n-\lambda\mathbf{I}_n) = \lambda(\mathbf{K}+\lambda\mathbf{I})^{-1}.$$

Now, $\mathbf{y} = \epsilon + \mathbf{\Psi}_{\leq m}a$. And we are interested in the expectation with regard to noise, hence:

$$\begin{aligned}\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_i(\hat{f}_\lambda(x_i)-y_i)^2\right] &= \frac{\lambda^2}{n}\mathbb{E}_\epsilon\left[(\boldsymbol{\epsilon}+a\mathbf{\Psi}_{\leq m})^\intercal\mathbf{H}^{-2}(\boldsymbol{\epsilon}+a\mathbf{\Psi}_{\leq m})\right] \\ &= \underbrace{\frac{\lambda^2\sigma^2}{n}\mathbf{Tr}\left(\mathbf{H}^{-2}\right)}_{T_1} + \underbrace{\frac{\lambda^2}{n}a^\intercal\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}a}_{T_2}.\end{aligned}$$

While T_1 is already the desired term, for T_2 we do the following:

$$\begin{aligned}T_2 &= \frac{\lambda^2}{n}(\mathbf{D}_{\leq m}^{-1}a)^\intercal\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{D}_{\leq m}^{-1}a) \\ &\overset{(i)}{=} \lambda^2\frac{1}{n}(\mathbf{D}_{\leq m}^{-1}a)^\intercal\left(\mathbf{D}_{\leq m}^{-1}+\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m} \\ &\qquad\left(\mathbf{D}_{\leq m}^{-1}+\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{D}_{\leq m}^{-1}a \\ &= \frac{\lambda^2}{n}\left\|\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\left(\mathbf{D}_{\leq m}^{-1}+\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{D}_{\leq m}^{-1}a\right\|^2 \\ &\leq \lambda^2\frac{\left\|\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}\right\|}{n}\left\|\left(\mathbf{D}_{\leq m}^{-1}+\mathbf{\Psi}_{\leq m}^\intercal\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{D}_{\leq m}^{-1}a\right\|^2 \\ &\overset{(ii)}{\leq} \frac{1.5\lambda^2}{(\mu_{\min}\left(\mathbf{K}_{>m}\right)+\lambda)^2}B_1. \\ &\overset{(iii)}{\leq} 12\lambda^2\frac{(\|\mathbf{K}_{>m}\|+\lambda)^2}{(\mu_{\min}\left(\mathbf{K}_{>m}\right)+\lambda)^2}\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2} \leq 12\lambda^2\frac{r_2^2}{r_1^2}\frac{\|\mathbf{D}_{\leq m}^{-1}a\|^2}{n^2},\end{aligned}$$

where in (i) we use Lemma 6. In (ii) we use Equation (5) as well as the fact that we recognize term B_1 from the proof of Theorem 3. In (iii) we then use the same bound on B_1 achieved in Theorem 3. This concludes the proof. □

E EXPERIMENTAL DETAILS

This section describes our experimental setup and provides additional details. Our submission already contains the code required to replicate all experiments and plots. We will release a documented and cleaned version of the code with the camera-ready version of the paper.

E.1 SETUP FOR FILTER SIZE EXPERIMENTS

The following describes the main filter size experiments presented in Section 4.1.

Network architecture We use a special CNN architecture that amplifies the role of filter size as an inductive bias. Each model of the main filter size experiments in Figure 2 has the following architecture:

1. Convolutional layer with 128 filters of varying size and no padding
2. Global max pooling over the spatial feature dimensions
3. ReLU activation
4. Linear layer with 256 output features
5. ReLU activation
6. Linear layer with 1 output feature

All convolutional and linear layers use a bias term. Since we employ a single convolutional layer before global max pooling, the convolutional filter size directly determines the maximum size of an input patch that can influence the CNN's output. Note that this architecture reduces to an MLP if the filter size equals the input image size.

Optimization procedure We use the same training procedure for all settings in Figure 2. Optimization minimizes the logistic loss for 300 epochs of mini-batch SGD with momentum 0.9 and batch size 100. We linearly increase the learning rate from 10^{-6} to a peak value of 0.2 during the first 50 epochs, and then reduce the learning rate according to an inverse square-root decay every 20 epochs. For peak learning rate γ_0, a decay rate L, the inverse square-root decay schedule at epoch $t \geq 0$ is

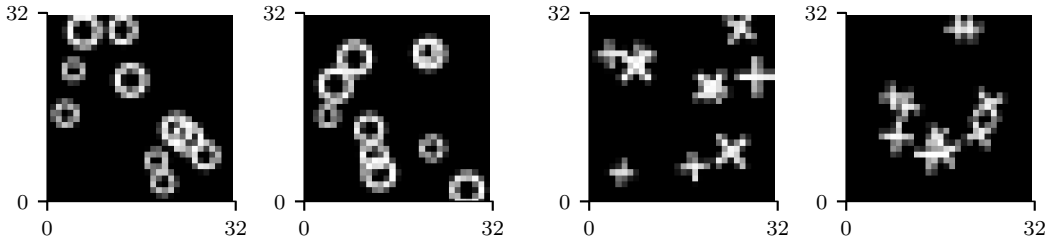
$$\frac{\gamma_0}{\sqrt{1 + \lfloor t/L \rfloor}}. \tag{33}$$

Learning rate warm-up helps to capture the early-stopped test error more precisely. Whenever possible, we use deterministic training algorithms, so that our results are as reproducible as possible. We selected all hyperparameters to minimize the training loss of the strongest inductive bias (filter size 5) on noisy training data, with the constraint that all other settings still converge and interpolate. Note that we do not use data augmentation, dropout, or weight decay.

Evaluation We observed that all models achieved their minimum test error either at the beginning or very end of training. Hence, our experiments evaluate the test error every 2 epochs during the first 150 epochs, and every 10 epochs afterwards to save computation time. We use an oracle, that is, the true test error, to determine the optimal early stopping epoch in retrospective. The optimal early stopping training error is always over the entire training set (including potential noise) for a fixed model, not an average over mini-batches. To mitigate randomness in both the training data and optimization procedure, we average over multiple dataset and training seeds. More precisely, we sample 5 different pairs of training and test datasets. For each dataset, we fit 15 randomly initialized models per filter size on the same dataset, and calculate average metrics. The plots then display the mean and standard error over the 5 datasets.

Dataset All filter size experiments use synthetic images. For a fixed seed, the experiments generate 200 training and 100k test images, both having an equal amount of positive and negative classes. Given a class, the sampling procedure iteratively scatters 10 shapes on an initially black 32×32 image. A single shape is either a circle (negative class) or a cross (positive class), has a uniformly random size in $[3, 5]$, and a uniformly random center such that all shapes end up completely inside the target image. We use a conceptual line width of 0.5 pixels, but discretize the shapes into a grid. See Figure 4 for examples. A single dataset seed fully determines the training data, test data, and all scattered shapes.

Noise model In the noisy case, we select 20% of all training samples uniformly at random without replacement, and flip their label. The noise is deterministic per dataset seed and does not change between different optimization runs. Note that we never apply noise to the test data.



(a) Negative samples

(b) Positive samples

Figure 4: Example synthetic images used in the filter size experiments.

E.2 Setup for rotational invariance experiments

The following describes the rotational invariance experiments presented in Section 4.2.

Dataset We use the EuroSAT (Helber et al., 2018) training split and subsample it into 7680 raw training and 10k raw test samples in a stratified way. For a fixed number of rotations k, we generate a training dataset as follows:

1. In the noisy case, select a random 20% subset of all training samples without replacement; for each, change the label to one of the other 9 classes uniformly at random.
2. For the i-th training sample ($i \in \{1, \ldots, 7680\}$):
   (a) Determine a random offset angle α_i.
   (b) Rotate the original image by each angle in $\{\alpha_i + j \cdot (360°/k) \mid j = 0, \ldots, k-1\}$.
   (c) Crop each of the k rotated 64×64 images to 44×44 so that no network sees black borders from image interpolation.
3. Concatenate all $k \times 7680$ samples into a single training dataset.
4. Shuffle this final dataset (at the beginning of training and every epoch).

To generate the actual test dataset, we apply a random rotation to each raw test sample independently, and crop the rotated images to the same size as the training samples. This procedure rotates every image exactly once, and uses random angle offsets to avoid distribution shift effects from image interpolation. Note that all random rotations are independent of the label noise and the number of training rotations. Hence, all experiments share the same test dataset. Furthermore, since we apply label noise before rotating images, all rotations of an image consistently share the same label.

Network architecture All experiments use a Wide Residual Network (Zagoruyko & Komodakis, 2016) with 16 layers, widen factor 6, and default PyTorch weight initialization. We chose the width and depth such that all networks are sufficiently overparameterized while still being manageable in terms of computational cost.

Optimization procedure We use the same training procedure for all settings in Figure 3. Optimization minimizes the softmax cross-entropy loss using mini-batch SGD with momentum 0.9 and batch size 128. Since the training set size grows in the number of rotations, all experiments fix the number of gradient updates to 144k. This corresponds to 200 epochs over a dataset with 12 rotations. Similar to the filter size experiments, we linearly increase the learning rate from zero to a peak value of 0.15 during the first 4800 steps, and then reduce the learning rate according to an inverse square-root decay (Equation (33)) every 960 steps. Whenever possible, we use deterministic training algorithms, so that our results are as reproducible as possible. We selected all hyperparameters to minimize the training loss of the strongest inductive bias (12 rotations) on noisy training data, with the constraint that all other settings still converge and interpolate. As for all experiments in this paper, we do not use additional data augmentation, dropout, or weight decay.

Evaluation Similar to the filter size experiments, we evaluate the test error more frequently during early training iterations: every 480 steps for the first 9600 steps, every 1920 steps afterwards. The experiments again use the actual test error to determine the best step for early-stopping, and calculate the corresponding training error over the entire training dataset, including all rotations and potential noise. Due to the larger training set size and increased computational costs, we only sample a single training and test dataset, and report the mean and standard error of all metrics over five training seeds.

E.3 Difference to double descent

As mentioned in Section 4.1, our empirical observations resemble the double descent phenomenon. This subsection expands on the discussion and provides additional details about how this paper's phenomenon differs from double descent.

While all models in all experiments interpolate the training data, we observe that both noisy labels and stronger inductive biases increase the final training loss of an interpolating model: Smaller filter size results in a decreasing number of model parameters. Enforcing invariance to more rotations requires a model to interpolate more (correlated) training samples. Thus, in both cases, increasing inductive bias strength decreases a model's overparameterization in relation to the number of training samples — shifting the setting closer to the interpolation threshold.

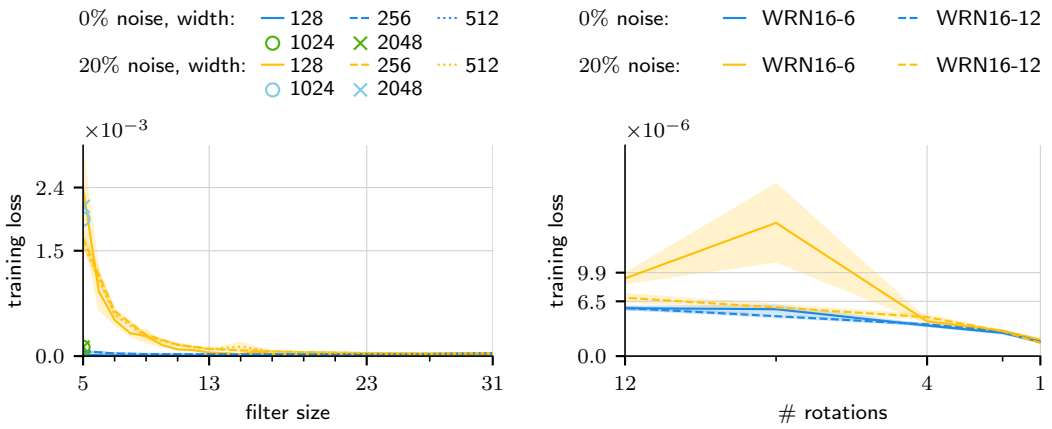
We argue that our choice of architecture and hyperparameter tuning ensures that no model in any experiment is close to the corresponding interpolation threshold. If that is the case, then double descent predicts that increasing the number of model parameters has a negligible effect on whether regularization benefits generalization, and does therefore not explain our observations.

In the following, we first describe how our hyperparameter and model selection procedure ensures that all models in all experiments are sufficiently overparameterized, so that double descent predicts negligible effects from increasing the number of parameters. Then, we provide additional experimental evidence that supports our argument: We repeat a subset of the experiments in Section 4 while upscaling the number of parameters in all models. For a fixed model scale and varying inductive bias, we observe that all phenomena in Section 4 persist. For a fixed inductive bias strength, we further see that the test error of interpolating models saturates at a value that matches our hypothesis. In particular, for strong inductive biases, the gap in test error between interpolating models and their optimally early-stopped version — harmful interpolation — persists.

Hyperparameter tuning We mitigate differences in model complexity for different strengths of inductive biases by tuning all hyperparameters on worst-case settings, that is, maximum inductive bias with noisy training samples. To avoid optimizing on test data, we tune on dataset seeds and network initializations that differ from the ones used in actual experiments. Figure 5 displays the final training loss for all empirical settings in this paper. While models with stronger inductive biases exhibit larger training losses, all values are close to zero, and the numerical difference is small. Finally, we want to stress again that this discussion is only about the training *loss*; all models in all experiments have zero training *error* and perfectly fit the corresponding training data.

Increasing model complexity for varying filter size As additional evidence, we repeat the main filter size experiments from Figure 2 in Section 4.1 using the same setup as before (see Appendix E.1), but increase the convolutional layer width to 256, 512, 1024, and 2048. For computational reasons, we evaluate a reduced number of filter sizes for widths 256 and 512, and only the smallest filter size 5 for widths 1024 and 2048. Since we found the original learning rate 0.2 to be too unstable for the larger model sizes, we use a decreased peak learning rate 0.13 for widths 256 and 512, and 0.1 for widths 1024 and 2048.

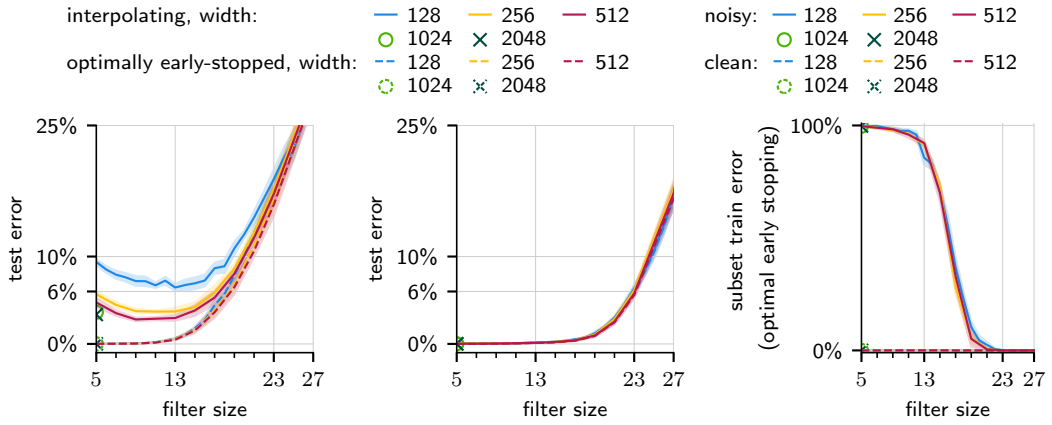
Figures 6a and 6b show the test errors for 20% and 0% training noise, respectively. With noisy training data (Figure 6a), larger interpolating models yield a slightly smaller test error, but the overall trends remain: the gap in test error between converged and optimally early-stopped models increases with inductive bias strength, and the phase transition between harmless and harmful interpolation persists. In particular, Figure 6a shows strong evidence that the number of model parameters does not influence our phenomenon: for example, models with filter size 5 (strong inductive bias) and width 512 (red line) have more parameters than models with filter size 27 (weak inductive bias) and width 128 (blue line). Nevertheless, models with filter size 5 benefit significantly from early stopping,



(a) Filter size

(b) Rotational invariance

Figure 5: Training losses of the models in this paper's experiments as a function of (a) filter size and (b) the number of training set rotations. Models with stronger inductive biases generally exhibit larger losses. However, in all instances, the numerical difference is small. Lines show the mean loss over 5 training set samples in (a) and 5 different optimization runs in (b), shaded areas the corresponding standard error.



(a) Test error (20% noise)

(b) Test error (0% noise)

(c) Opt. early stopping (20% noise)

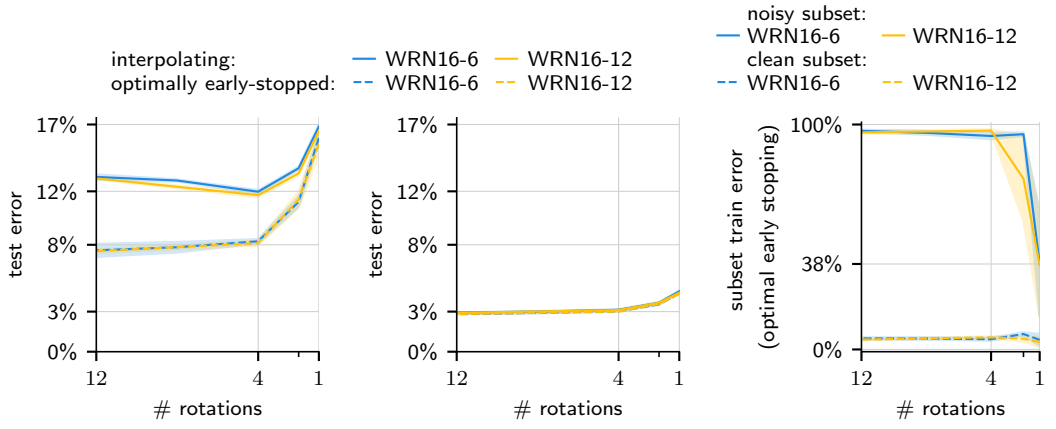
Figure 6: An increase in convolutional layer width by factors 2 to 16 does not significantly alter the behavior of (a) the test error when training on 20% label noise, (b) the test error when training on 0% label noise, and (c) the training error when using optimal early stopping on 20% label noise. Despite significantly larger model size, the phase transition between harmless and harmful interpolation persists. Lines show the mean over five random datasets, shaded areas the standard error.

while interpolation for models with filter size 27 is harmless. In the noiseless case (Figure 6b), increasing model complexity does neither harm nor improve generalization, and all models achieve their optimal performance after interpolating the entire training dataset. Similarly, Figure 6c reveals that the fraction of training noise that optimally early-stopped models fit stays the same for larger models. Finally, for a fixed inductive bias strength, the test errors saturate as model size increases, making a different trend for models with more than 2048 filters unlikely. To increase legibility, we present the numerical results for the largest two filter sizes in Table 1.

Increasing model capacity for varying rotational invariance For completeness, we also repeat the rotation invariance experiments from Figure 3 in Section 4.2 with twice as wide Wide Residual Networks on a reduced number of rotations. More precisely, we increase the network widen-factor from 6 to 12, and otherwise use the same setting as the main experiments (see Appendix E.2). Note that this corresponds to a parameter increase from around 6 million to around 24 million parameters.

Table 1: Test errors for filter size 5 (strongest inductive bias) and very large width under 20% training noise.

	width 1024	width 2048
early-stopped test error	0.0062%	0.0049%
interpolating test error	3.6251%	3.3664%
# parameters	289281	578049



(a) Test error (20% noise) (b) Test error (0% noise) (c) Opt. early stopping (20% noise)

Figure 7: Doubling Wide Residual Network width does not significantly alter the behavior of (a) the test error when training on 20% label noise, (b) the test error when training on 0% label noise, and (c) the training error when using optimal early stopping on 20% label noise. Despite significantly larger model size, the phase transition between harmless and harmful interpolation persists. Lines show the mean over five random network initializations, shaded areas the standard error.

The results in Figure 7 provide additional evidence that our phenomenon is distinct from double descent: both the test error (Figures 7a and 7b) and fraction of fitted noise under optimal early stopping (Figure 7c) exhibit the same trend, despite the significant difference in number of parameters.