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 interpolators. In particular, the authors prove that a stronger inductive bias (small $p \to 1$) not only 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 not only is interpolation harmless for weak inductive biases, but that 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 related lines of work and explain how the theoretical results cannot reflect the phase transition between harmless and harmful interpolation for high-dimensional kernel learning.

Low-dimensional kernel learning: Many recent papers (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 regimes - however, these bounds crucially rely on ridge regularization. In general, one cannot expect harmless interpolation for kernel regression for such kernels in the low-dimensional setting (Rakhlin & Zhai, 2019; Mallinar et al., 2022; Buchholz, 2022) - positive results only exist for very specific adaptive spiked kernels (Belkin et al., 2019b). Furthermore, techniques developed for low dimensional regimes (see e.g., Schölkopf et al. (2018)) usually suffer from a curse of dimensionality; that is, become vacuous in high-dimensional settings when the dimension is allowed to grow with n.

High-dimensional kernel learning: Several papers in the literature (Liang et al., 2020; McRae et al., 2022; Liang & Rakhlin, 2020; Liu et al., 2021) tackle high-dimensional kernel learning and prove non-asymptotic bounds using advanced high-dimensional random matrix concentration tools in (El Karoui, 2010). However, their results heavily rely on a bounded Hilbert norm assumption, that is a natural in the low-dimensional regime. However, as pointed out in (Donhauser et al., 2021b), such an assumption is misleading in the high-dimensional regime.

Another line of research (Ghorbani et al., 2021; 2020; Mei et al., 2021; Ghosh et al., 2022; Misiakiewicz & Mei, 2021; Mei et al., 2022) gives precise asymptotic characterizations of the risk of the kernel regression estimator in specific settings where we have access to the kernel eigenfunctions and eigenvalues. However, to investigate how varying the filter size of the convolution kernel affects the risk of the kernel regression estimator, these asymptotic results are insufficient. In contrast to the above papers, 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.

Related work on overfitting of structured interpolators: Several papers question the generality of harmless interpolation for models which incorporate strong structural assumptions. Examples include structural invariances 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; Sanyal et al., 2021), 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 this line of research and offer a new theoretical perspective that characterizes 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 the filter size (as an example for 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 for optimally ridge-regularized and minimum-norm interpolators as a function of the filter size (Theorem 1). Furthermore, for small filter sizes, we not only prove that harmless interpolation

occurs (Theorem 1, but that fitting the noise (at lest partially) is even necessary to some degree 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, e.g. (Misiakiewicz & Mei, 2021) and assume that the training samples $\{x_i, y_i\}_{i=1}^n$ are i.i.d. draws from a distribution with $x_i \sim \mathcal{U}(\{-1,1\}^d)$ and $y_i = f^*(x_i) + \epsilon_i$ with ground truth f^* and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For simplicity of exposition, we further assume that $f^*(x) = x_1 x_2 \cdots x_{L^*}$, 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 input data distribution is a difficult task. Indeed, existing results that establish precise characterizations for the risk (see Section 2) crucially rely on the assumption that the input data distribution is hypercontractive – so far only shown to hold for few high-dimensional distributions, including the hypersphere (Beckner, 1992) and the discrete hypercube (Beckner, 1975) that we use in this paper. It is essential to tightly control the empirical kernel matrix in Lemma 1 in Section 5. Generalizations beyond this assumption would require the development of new tools in random matrix theory that we consider important future work.

We consider (cyclic) convolutional kernels with filter size $q \in [1, d]$ 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 we assume that $\kappa : [-1,1] \to \mathbb{R}$ satisfies standard regularity assumptions (see detailed Assumption 3 in Appendix B)) that includes for instance the exponential function. By decreasing the filter size q, we encourage the KRR solution 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 minimum-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 by:

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

3.2 MAIN RESULT

We now present tight upper and lower bounds for the prediction error of kernel regression estimators in the setting of 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 refer to Section 5 for the proof.

[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 use the standard notations $\mathcal{O}(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$, as well as $\lesssim, \gtrsim$, which hides constant depending on d, n, λ and q only via the exponents $\ell, \ell_\lambda, \beta$ defined in Theorem 1, and similarly use $c, c_1, \ldots$ for generic positive constants.

Theorem 1 (Non-asymptotic prediction error rates)**.** *Assume a dataset and kernel as described in Section 3.1 with noise variance σ^2 and let $\hat{f}_\lambda$ be the KRR estimate in Eq.* (2)*. Further, let $n \in \Theta(d^\ell)$, filter size $q \in \Theta\left(d^\beta\right)$, $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$ and $\sigma^2 \in \Theta(d^{-\ell_\sigma})$ for some $\ell, \beta, \ell_\lambda, \ell_\sigma \in \mathbb{R}$ and $\ell_\lambda \geq 0$ and define $\delta = \frac{\ell - \ell_\lambda - 1}{\beta} - \lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \rfloor$ and $\bar{\delta} = \frac{\ell - 1}{\beta} - \lfloor \frac{\ell - 1}{\beta} \rfloor$. Then, for some universal constant c, we have with probability at least $1 - cd^{-\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$, that*

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

Further, for a ground truth $f^*(x) = x_1 x_2 \cdots x_{L^*}$, with $L^* \leq 1 + \lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \rfloor$, with probability at least $1 - cd^{-\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$, it also holds that:

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

Finally, the rates for bias and variance are also valid for $\hat{f}_0$ in Eq. (3) *by setting $\ell_\lambda = 0$.*

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

Phase transition as a function of β: In our discussion, we would like to focus on the impact of the filter size $q \in \Theta\left(d^\beta\right)$ on the risk (sum of bias and variance), entering 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 illustrates how the strength of inductive bias affects generalization. Specifically, Figure 1a shows the rates for the minimum-RKHS-norm interpolator $\hat{f}_0$ and the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$, where λ_{opt} is chosen 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 $\hat{f}_0$. The ground truth $f^\star(x) = x_1 x_2$ used for the corresponding experiments is chosen such that the assumption on L^* is satisfied for all β. The following phase transition occurs at the threshold $\beta^* \in (0, 1)$ at which the rates of the statistical bias and variance in Theorem 1 match:

- 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 minimum-RKHS-norm interpolator $\hat{f}_0$ match. Hence, we observe *harmless interpolation*.

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

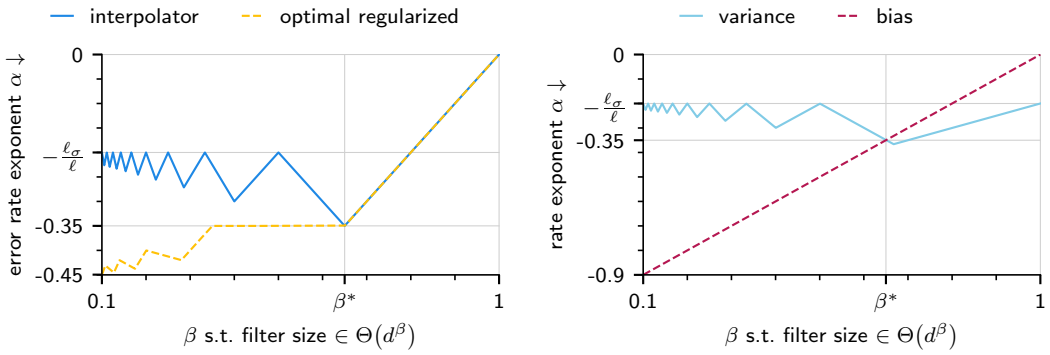
Theorem 2 (Informal - training error)**.** *Let λ_{opt} be such that the expected population risk $\textbf{Risk}(\hat{f}_{\lambda_{opt}})$ is minimal and let β^* be the threshold where the bias and variance bounds in Theorem 1 are of the same order for the interpolator $\hat{f}_0$ (setting $\ell_\lambda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

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

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

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

Bias-variance trade-off: We now discuss how the phase transition arises from a (statistical) bias and variance trade-off of the minimum-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 rates in Theorem 1 for high-dimensional kernel ridge regression: (a) Rate exponents α of the **Risk** $\in \Theta(n^\alpha)$ of the interpolator $\hat{f}_0$ vs. the optimally ridge-regularized estimator $\hat{f}_{\lambda_{\text{opt}}}$ and similarly (b) Rate exponents of the variance and bias for interpolator as a function of β — the rate of the filter size q where $q \in \Theta(d^\beta)$. In the experiments we choose $\hat{f}_0$ with $\ell = 2$, $\ell_\sigma = 0.5$, and the ground truth $f^*(x) = x_1 x_2$. β^* denotes the threshold where the bias and variance terms in Theorem 1 match and we observe a phase transition between harmless and harmful interpolation occurs.

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 a multiple descent curve for the variance has been previously observed in (Liang et al., 2020) that arises with varying input dimension with no connection to structural biases.

4 Experiments

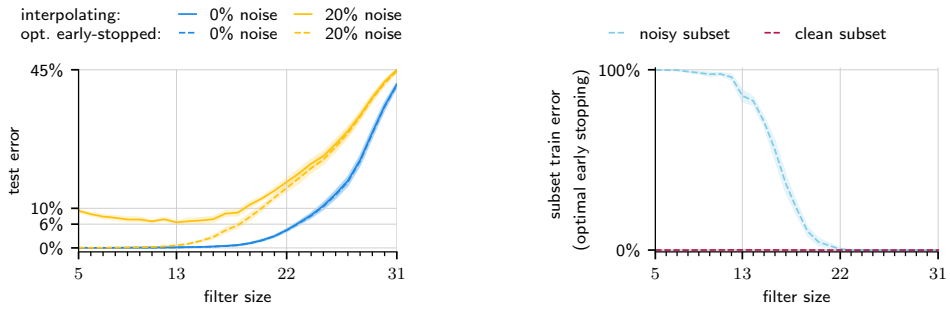
We now empirically study whether the phase transition phenomenon that we prove for kernel regression persists for deep neural networks with feature learning. More precisely, we present controlled experiments to investigate how the strength of a CNN's inductive bias influences if interpolating noisy data 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 neural 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 filter size 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, the more harmful interpolation becomes. These results suggest exciting future work: to prove this trend for models with feature learning.

4.1 Effect of filter size on CNNs for synthetic image dataset

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.[3] Thus, decreasing filter size up to 5×5

[3] See Figure 4 in Appendix E.1 for examples.



(a) Test error

(b) Training error of optimally early-stopped models

Figure 2: Convolutional neural network experiments on synthetic image dataset with varying filter size. (a) For noisy data, small filter sizes (strong inductive bias towards local structures) 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 the training set. Under label noise, for strong inductive biases optimally early-stopped models ignore all noisy samples (100% error on the noisy subset), while models with weak inductive biases 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 more experimental details.

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.

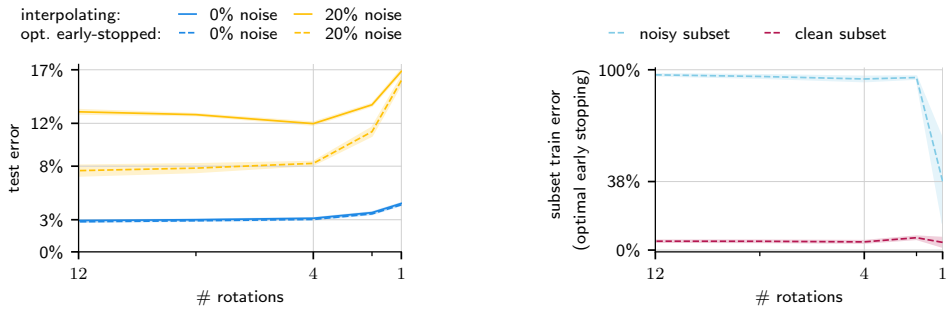
Training setup: In the experiments, we use CNNs with a single convolution layer, followed by global max pooling over the spatial dimensions and two dense layers. We train those CNNs with different filter sizes on 200 training samples (either noiseless or with 20% label flips) to minimize the logistic 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. Lastly, we provide a detailed discussion on the choice of hyperparameters and more experimental details in 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, models with strong inductive biases (small filter sizes) suffer from an increasing gap in test errors when interpolating versus using optimal early stopping as filter size decreases. Figure 2b, on the other hand, reflects the dual perspective of the phase transition as presented in Theorem 2: for small filter size, optimally early-stopped models entirely avoid fitting training noise such that the training error on the noisy training subset equals 100%, while large filter sizes require interpolation of 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 its number of parameters and hence the degree of overparameterization. Hence, double descentpredicts harmless interpolation due to model size for weak inductive biases. Nevertheless, we argue that the phenomenon we observe here is distinct, and discuss two forms of evidence in Appendix E.3: For one, we choose sufficiently large models and tune hyperparameters to ensure that all models interpolate and yield small training loss, even for the minimum filter size and 20% training noise. Additionally, we repeat the experiments with two and four times as many convolutional filters. Even for significantly larger models, overfitting still increases with stronger inductive biases, and the phase transition between harmless and harmful interpolation persists.

4.2 Rotational invariance of WideResnet 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 set, effectively performing data



(a) Test error

(b) Training error of optimally early-stopped models

Figure 3: Varying degrees of rotational invariance when fitting WideResNet on satellite images. (a) For noisy data, 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 the optimally early-stopped model (optimized for test error) on the noisy and clean subsets of the training set: For the strongest bias towards 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 (smaller than 50% training error on noisy subset) for optimal generalization. All lines show the mean over five optimization runs, shaded areas the standard error; see Section 4.2 for details.

augmentation. [4] 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 by a different 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 k of rotated versions of each training sample of 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.

Results: Similar to the previous section, 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 rotationally invariant models cannot fit noisy samples for optimal generalization.

5 Proof of the main result

The proof of the main result, Theorem 1, is divided into two steps. First, we present with Theorem 3 in Section 5.1 a result for fixed-design that gives tight non-asymptotic matching upper and lower bounds for the prediction error of general kernels. Then, in Section 5.2, we proceed by showing that the assumptions in Theorem 3 hold.

[4] Data augmentation based techniques have been observed to be more efficient to enforce rotational invariance; see e.g. Yang et al. (2019).

Notation Assuming that the inputs are sampled i.i.d. from a data distribution ν, i.e. $x_i \overset{iid}{\sim} \nu$, we can express any continuous, positive semi-definite kernel function as the following sum:

$$\mathcal{K}(x,x') = \sum_{k=0}^{\infty} \lambda_k \psi_k(x)\psi_k(x') = \underbrace{\sum_{k=0}^{m} \lambda_k \psi_k(x)\psi_k(y)}_{:=\mathcal{K}_{\leq m}(x,x')} + \underbrace{\sum_{k=m+1}^{\infty} \lambda_k \psi_k(x)\psi_k(x')}_{:=\mathcal{K}_{>m}(x,x')},$$

where $\{\psi_k\}_{k>0}$ is an orthonormal basis of the RKHS with respect to the inner kernel induced by ν: $\langle f, g\rangle_\nu := \mathbb{E}_{x\sim\nu}[f(x)g(x)]$. Eigenfunction are sorted in descending order $\lambda_k \geq \lambda_{k+1}$. Define the empirical kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}, [\mathbf{K}]_{ij} = \mathcal{K}(x_i, x_j)$, and analogously $\mathbf{K}_{\leq m}$ and $\mathbf{K}_{>m}$. Furthermore, we make use of the matrices $\mathbf{\Psi}_{\leq m} \in \mathbb{R}^{n\times m}, [\mathbf{\Psi}_{\leq m}]_{il} := \psi_l(x_i)$, and $\mathbf{D}_{\leq m} := \mathrm{diag}(\lambda_1, \dots, \lambda_m)$. Define the squared kernel $\bar{\mathcal{S}}_1(x,x') = \mathbb{E}_{z\sim\nu}[\mathcal{K}(x,z)\mathcal{K}(z,x')]$, the truncated versions of it $\mathcal{S}_{1>m}, \mathcal{S}_{1\leq m}$ and the corresponding empirical kernel matrices $\mathbf{S}, \mathbf{S}_{\leq m}$ and $\mathbf{S}_{>m} \in \mathbb{R}^{n\times n}$. We use $\mu_{\min}(\cdot)$ and $\mu_{\max}(\cdot)$ (or $\|\cdot\|$) to indicate respectively the smallest and the largest eigenvalue of a symmetric positive definite matrix. We use $\langle\cdot,\cdot\rangle$ to indicate the euclidean inner product in $\mathbb{R}^d$.

5.1 Generalization bound for fixed-design

We begin with a fixed design bound for the population error based on the following assumptions.

Assumption 1 (on the eigenfunctions)**.** *Assume that* $\left\|\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}/n - \mathbf{I}_m\right\| \leq \frac{1}{2}$.

Assumption 2 (on the ground truth)**.** *The ground truth is contained in the span of the first m eigenfunctions.*

We now present in Theorem 3 the main result of this section, where we make use of the ratio

$$r = \frac{\mu_{\max}(\mathbf{K}_{>m}) + \lambda}{\mu_{\min}(\mathbf{K}_{>m}) + \lambda}. \tag{4}$$

Theorem 3 (Generalization bound for fixed-design)**.** *Given a kernel $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, and a dataset $\{(x_i, y_i)\}_{i=1}^n$, and $y_i = f^*(x_i) + \epsilon_i$ with ground truth f^*. Let $\hat{f}_\lambda$ be the KRR estimate in Eq.* (2) *for any $\lambda \geq 0$. Let $m \in \mathbb{N}$ be such that Assumption 1 holds. Then:*

$$\tau_1^2 \frac{1}{6r^2}\frac{m}{n} + \frac{\mathbf{Tr}(\mathbf{S}_{>m})}{(\mu_{\max}(\mathbf{K}_{>m}) + \lambda)^2} \leq \textbf{\textit{Variance}}(\hat{f}_\lambda)/\sigma_\epsilon^2 \leq 6r^2\frac{m}{n} + \frac{\mathbf{Tr}(\mathbf{S}_{>m})}{(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2}.$$

Furthermore, if we assume that, with the same m, also Assumption 2 is respected. We can write $\hat{f}_\lambda(x) = \sum_{k=1}^m \mathbf{a}_k\psi_k(x)$, where ψ_k are the eigenfunctions of $\mathcal{K}$. Then:

$$\tau_1^2 \frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}\|^2}{n^2} \leq \textbf{\textit{Bias}}^2(\hat{f}_\lambda) \leq C\tau_2(\mu_{\max}(\mathbf{K}_{>m}) + \lambda)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}\|^2}{n^2},$$

where $C = 8\max\{1.5r, 1\}$ and $\tau_1 = \min\left\{\frac{n\lambda_m}{\max\{1,\lambda\}}, 1\right\}$, $\tau_2 = \max\left\{\frac{n\lambda_{m+1}}{\max\{\lambda,1\}}, 1\right\}$.

We refer the reader to Appendix A for the proof, which is mostly algebra and essentially consists of multiple applications of the Woodbury matrix identity. The main ideas are derived from the proofs in (Bartlett et al., 2020; Tsigler & Bartlett, 2020) where the authors establish tight bounds for the minimum-ℓ_2-norm interpolator on independent sub-Gaussian features.

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

5.2 Proof of Theorem 1

Throughout the proof, all statements hold for the setting described in Section 3.1 and under the assumptions in Theorem 1, where we recall that $n \in \Theta(d^\ell)$ and $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$. Furthermore, as in Theorem 1, $\delta, \bar{\delta} > 0$ with $\delta := \frac{\ell-\ell_\lambda-1}{\beta} - \left\lfloor\frac{\ell-\ell_\lambda-1}{\beta}\right\rfloor$ and $\bar{\delta} := \frac{\ell-1}{\beta} - \left\lfloor\frac{\ell-1}{\beta}\right\rfloor$. To prove the Theorem 1, all we need to do is to show that there exists an m such that Theorem 3 applies with high-probability.

Step 1: Bounding the variance. We begin with a lemma showing the concentration of the spectrum of the late part of the kernel matrix $\mathbf{K}_{\leq m}$ and squared kernel matrix $\mathbf{S}_{\leq m}$, where we make heavy use of the data distribution.

Lemma 1 (Tight bound conditions)**.** *There exists m such that $n\lambda_m \in \Theta(\max\{\lambda, 1\})$, and, for such m, with probability at least $1 - cd^{-2\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$, $r \in \Theta(1)$ and*

$$d^{-\ell_\lambda} q^{\delta-1} \lesssim \frac{\mathbf{Tr}(\mathbf{S}_{>m})}{(\mu_{\max}(\mathbf{K}_{>m}) + \lambda)^2} \leq \frac{\mathbf{Tr}(\mathbf{S}_{>m})}{(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2} \lesssim d^{-\ell_\lambda} q^{-\delta} + d^{-\ell_\lambda} q^{\delta-1}.$$

We refer to Appendix C for the proof. Note that for the choice of m made in Lemma 1, also $C, \tau_1, \tau_2 \in \Theta(1)$. Assume for the moment that Assumption 1 holds for the choice of m presented in Lemma 1, then we can apply Theorem 3 to get the following tight high-probability bound for the variance:

$$d^{-\ell_\lambda} q^{-\delta} + d^{-\ell_\lambda} q^{\delta-1} \lesssim \textbf{Variance}(\hat{f}_\lambda)/\sigma_\epsilon^2 \lesssim d^{-\ell_\lambda} q^{-\delta} + d^{-\ell_\lambda} q^{\delta-1}. \tag{5}$$

Step 2: Bounding the bias. To bound the bias, we need to show that Assumption 2 on the ground truth holds and bound the term $\left\| \mathbf{D}_{\leq m}^{-1} \mathbf{a} \right\|$, as done in the following lemma.

Lemma 2 (Ground truth assumption)**.** *Let the ground truth $f^*(x) = \prod_{i=1}^{L^*} x_i$ be as in Theorem 1 with $L^* > 0$. Then, as long as $\lambda_m \in o\left(\frac{1}{d \cdot q^{L^*-1}}\right)$, Assumption 2 is respected. Furthermore, if we write $f^*(x) = \sum_{k=1}^{m} \mathbf{a}_k \psi_k(x)$, then $\left\| \mathbf{D}_{\leq m}^{-1} \mathbf{a} \right\| \in \Theta\left(d \cdot q^{L^*-1}\right)$.*

We refer to Appendix B for the proof. Assume again for the moment that, for the choice of m in Lemma 1, Assumption 1 and Lemma 2 hold. We would then again obtain from Theorem 3 the following tight high-probability bound for the bias:

$$d^{-2(\ell - \ell_\lambda - 1 - \beta(L-1))} \lesssim \textbf{Bias}^2(\hat{f}_\lambda) \lesssim d^{-2(\ell - \ell_\lambda - 1 - \beta(L^*-1))}. \tag{6}$$

Step 3: Proving the Theorem. We begin with the following lemma, which is a direct consequence of well-known matrix concentration results.

Lemma 3 (Corollary of Theorem 5.44 in (Vershynin, 2012))**.** *For any $m \in \mathcal{O}(n \cdot d^{-\beta\delta})$, Assumption 1 holds with probability $1 - cd^{-\beta\delta}$.*

We refer to Appendix C for the proof. As a result, we are only left to show that the m such that $n\lambda_m \in$ in Lemma 1 satisfies the assumptions in Lemma 2 and 3. This requires a careful analysis of the eigenvalues of the convolutional kernel presented in (Misiakiewicz & Mei, 2021) and summarized in Appendix B. We collected what is useful for the purpose of this proof in the following Lemma.

Lemma 4 (Eigendecay)**.** *Let m be such that $n\lambda_m \in \Theta(\max\{\lambda, 1\})$. Then, assuming $L^* \leq \left\lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \right\rfloor + 1$, we have that $\lambda_m \in o\left(\frac{1}{d \cdot q^{L^*-1}}\right)$ and $m \in \mathcal{O}(n \cdot d^{-\delta})$.*

Finally, we can conclude the proof by taking the union bound to show that Equation (5) and (6) hold with probability $1 - cd^{-2\beta \min\{\bar{\delta}, 1-\bar{\delta}\}}$.

6 Summary and future work

In this paper, we highlight the impact of strength of an inductive bias on the generalization performance of the resulting trained model. In particular we study when the gap between optimally ridge-regularized or early-stopped models with their interpolating counterparts is zero (harmless interpolation). In particular, we prove a phase transition for kernel learning with convolutional kernels of different filter sizes: weak inductive bias (large filter size) yields harmless interpolation whereas for strong inductive bias (small filter size), interpolating noise leads to suboptimal rates. Intuitively, this phenomenon can be explained by a bias-variance trade-off captured by our main result, 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 feature learning models, 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(48):30063–30070, 2020.

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

William Beckner. Inequalities in fourier analysis. *Annals of Mathematics*, 102(1):159–182, 1975.

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

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 1611–1619, 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. *Advances in Neural Information Processing Systems*, 34:18673–18684, 2021.

Simon Buchholz. Kernel interpolation in sobolev spaces is not consistent in low dimensions. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 3410–3440, 2022.

Francesco Cagnetta, Alessandro Favero, and Matthieu Wyart. What can be learnt with wide convolutional networkds?, 2022.

Emmanuel J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Niladri S. Chatterji and Philip M. Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):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*, volume 34, pp. 23465–23477, 2021a.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. *Proceedings of the International Conference on Machine Learning (ICML)*, 139:2804–2814, 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)*, volume 162, pp. 5397–5428, 17–23 Jul 2022.

David L. Donoho and Michael Elad. On the stability of the basis pursuit in the presence of noise. *Signal Processing*, 86(3):511–532, 2006.

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

Alessandro Favero, Francesco Cagnetta, and Matthieu Wyart. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:9456–9467, 2021.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.

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

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 Inc., New York, NY, USA, 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)*, pp. 204–207, 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)*, pp. 27462–27474, 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)*, pp. 9788–9803, 2021.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):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)*, pp. 649–657. PMLR, 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)*, pp. 5853–5875. PMLR, 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)*, pp. 3351–3418, 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.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(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(12):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)*, pp. 8022–8031, 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)*, pp. 2595–2623, 2019.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1): 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)*, pp. 8093–8104, 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. ISBN 0262536579.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression, 2020.

Roman Vershynin. *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 l1-norm interpolation of noisy data. *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. *Advances in Neural Information Processing Systems*, 30, 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)*, volume 32. Curran Associates, Inc., 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12, 2016.

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

Since we have a closed formula for bias and variance, the proof of this result boils down to an accurate application of Woodbury identity.

Proof. Thanks to the representer theorem, we know that the optimization problem in kernel ridge regression defined in Equation 2 is solved by the following estimator:

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

where $\mathbf{y} := [y_1, \dots, y_n]^\intercal$ and $\mathbf{k}(x) := [\mathcal{K}(x_1, x), \dots, \mathcal{K}(x_n, x)]^\intercal$, $\mathbf{H} = \mathbf{K} + \lambda \mathbf{I}_n$. Thus, we can get a closed formula for bias and variance:

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

Where in (i) we use the definition of the $\mathcal{S}_1$ kernel, as well as the following simple consequences of the eigenfunctions orthonormality:

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

Now we decompose the proof of the theorem in two parts, one for variance and one for bias.

Bias We start by applying Lemma 5 to decompose $\mathbf{S}$ in $\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal + \mathbf{S}_{>m}$:

$$
\begin{aligned}
\textbf{Bias}^2(\hat{f}_\lambda) &= \mathbf{a}^\intercal \mathbf{a} - 2\mathbf{a}^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{a} + \mathbf{a}^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{a} \\
&\quad + \mathbf{a}^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{a} \\
&= \underbrace{\left\|\left(\mathbf{I}_m - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\right)\mathbf{a}\right\|^2}_{B1} + \underbrace{\mathbf{a}^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{a}}_{B2}.
\end{aligned}
$$

Now, we first notice that $B1$ can be written as:

$$
\begin{aligned}
B1 &= \left\|\left(\mathbf{I} - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\right)\mathbf{a}_{\leq m}\right\|^2 \\
&= \left\|\left(\mathbf{D}_{\leq m} - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\right)\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\right\|^2 \\
&\overset{(i)}{=} \left\|\left(\mathbf{D}_{\leq m} - \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \left(\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal + \mathbf{H}_{>m}\right)^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\right)\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\right\|^2 \\
&\overset{(ii)}{=} \left\|\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\right\|^2. \qquad (7)
\end{aligned}
$$

Where in (i) we decomposed $\mathbf{H}$ in $\mathbf{K}_{\leq m} + \mathbf{H}_{>m}$, where $\mathbf{H}_{>m} = \mathbf{K}_{>m} + \lambda\mathbf{I}_n$. In (ii) we applied Woodbury identity.

We bound now $B2$:

$$
\begin{aligned}
B2 &\leq \|\mathbf{S}_{>m}\| \mathbf{a}^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-2} \mathbf{\Psi}_{\leq m} \mathbf{a} = \|\mathbf{S}_{>m}\| \left(\mathbf{D}_{\leq m}^{-1}\mathbf{a}\right)^\intercal \mathbf{D}_{\leq m} \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-2} \mathbf{\Psi}_{\leq m} \mathbf{D}_{\leq m} \left(\mathbf{D}_{\leq m}^{-1}\mathbf{a}\right) \\
&\overset{(i)}{=} \|\mathbf{S}_{>m}\| \mathbf{a}^\intercal \mathbf{D}_{\leq m}^{-1} \mathbf{a} \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}\right)^{-1} P m^\intercal \mathbf{H}_{>m}^{-2} \mathbf{\Psi}_{\leq m} \\
&\qquad \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1} \mathbf{a} \\
&\leq \underbrace{\|\mathbf{S}_{>m}\| \left\|\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-2} \mathbf{\Psi}_{\leq m}\right\|}_{C1} \underbrace{\left\|\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1} \mathbf{a}\right\|^2}_{=B1}.
\end{aligned}
$$

Where in (i) we use Lemma 7. Thus, the bias can be bounded by $B1(1 + C1)$. We now start by upper bounding $C1$:

$$
\begin{aligned}
1 + C1 &\leq 1 + n\|\mathbf{S}_{>m}\| \left\|\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}}{n}\right\| \frac{1}{\mu_{\min}\left(\mathbf{H}_{>m}\right)^2}. \\
&\overset{(i)}{\leq} 1 + 1.5 \frac{n\lambda_{m+1}\|\mathbf{K}_{>m}\|}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2} \overset{(ii)}{\leq} 1 + 1.5 \frac{n\lambda_{m+1}\left(\|\mathbf{K}_{>m}\| + \lambda\right)}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2} \\
&\leq 1 + 1.5 \frac{\lambda_{m+1}}{\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda} r \leq 2\max\{1.5r, 1\} \max\left\{\frac{n\lambda_{m+1}}{\max\{\lambda, 1\}}, 1\right\}.
\end{aligned}
$$

Where in (i) we used Assumption 3 to bound $\|\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}/n\|$, and Lemma 6 to bound $\|\mathbf{S}_{>m}\|$.To conclude the bias bound, we need to upper and lower bound $B1$.

Upper bound:

$$
\begin{aligned}
B1 &\leq \left\|\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right\|^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \\
&= \frac{1}{\mu_1\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right)^2} \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \leq \frac{1}{\mu_1\left(\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right)^2} \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \\
&\leq \frac{\|\mathbf{H}_{>m}\|^2}{\mu_1\left(\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}}{n}\right)^2} \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \leq 4\left(\|\mathbf{K}_{>m}\| + \lambda\right)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2}.
\end{aligned}
$$

Lower bound:

$$
\begin{aligned}
B1 &\geq \mu_1\left(\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right)^{-1}\right)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \\
&= \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2\left\|\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right\|^2} \frac{}{n^2} \geq \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2\left(\left\|\frac{\mathbf{D}_{\leq m}^{-1}}{n}\right\| + \left\|\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1} \mathbf{\Psi}_{\leq m}}{n}\right\|\right)^2} \\
&= \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2\left(\frac{1}{n\lambda_m} + \frac{1}{\mu_{\min}(\mathbf{H}_{>m})}\left\|\frac{\mathbf{\Psi}_{\leq m}^\intercal \mathbf{\Psi}_{\leq m}}{n}\right\|\right)^2} \\
&\overset{(i)}{\geq} \left(\frac{\frac{n\lambda_m}{\mu_{\min}(\mathbf{K}_{>m})+\lambda}}{1.5\frac{n\lambda_m}{\mu_{\min}(\mathbf{K}_{>m})+\lambda} + 1}\right)^2 \left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \\
&\geq \frac{1}{3}\min\left\{\frac{n\lambda_m}{\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda}, 1\right\}^2 \left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2}
\end{aligned}
$$

Where (i) follows from Assumption 1. Finally, the bound for the bias follows directly by taking $C_3 = 8\max\{1.5r, 1\}$.

Variance: We start by applying Lemma 5 to decompose $\mathbf{S}$ in $\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^{\intercal} + \mathbf{S}_{>m}$, hence we decompose the variance as:

$$\textbf{Variance}(\hat{f}_\lambda)/\sigma_\epsilon^2 = \underbrace{Tr\left(\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-1}\right)}_{V1} + \underbrace{\mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{S}_{>m}\mathbf{H}^{-1}\right)}_{V2}. \tag{8}$$

We then start by bounding $V1$.

$$\begin{aligned}
V1 &= \mathbf{Tr}\left(\mathbf{H}^{-1}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}^2\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-1}\right) \\
&= \mathbf{Tr}\left(\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\right) \\
&\overset{(i)}{=} \mathbf{Tr}\left(\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1}\right) \\
&= \frac{1}{n}\mathbf{Tr}\left(\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-2}\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}}{n}\right).
\end{aligned}$$

Where in (i) we used Lemma 7. We now use the fact that the trace is the sum of the eigenvalues, hence, it's bounded from above by the size of the matrix times the largest eigenvalue. And from below by the size of the matrix times the smallest eigenvalue. We then get:

$$\begin{aligned}
V1 &\leq \frac{m}{n}\left\|\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-2}\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}}{n}\right\| \\
&\overset{(i)}{\leq} \frac{m}{n}4\left(\|\mathbf{K}_{>m}\| + \lambda\right)^2\frac{\|\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{\Psi}_{\leq m}/n\|}{\mu_{\min}\left(\mathbf{H}_{>m}\right)^2} \leq \frac{m}{n}6\frac{\left(\|\mathbf{K}_{>m}\| + \lambda\right)^2}{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2} \leq \frac{m}{n}6r^2, \\
V2 &\geq \frac{m}{n}\mu_1\left(\left(\frac{\mathbf{D}_{\leq m}^{-1}}{n} + \frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}}{n}\right)^{-2}\frac{\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}}{n}\right) \\
&\overset{(ii)}{\geq} \frac{m}{n}\frac{1}{3}\min\left\{\frac{n\lambda_m}{\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda}, 1\right\}^2\frac{\mu_{\min}\left(\mathbf{\Psi}_{\leq m}^{\intercal}\mathbf{\Psi}_{\leq m}/n\right)}{\|\mathbf{H}_{>m}\|^2} \\
&\geq \frac{m}{n}\frac{1}{6}\min\left\{\frac{n\lambda_m}{\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda}, 1\right\}^2\frac{\left(\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda\right)^2}{\left(\|\mathbf{K}_{>m}\| + \lambda\right)^2} \\
&\geq \frac{m}{n}\frac{1}{6r^2}\min\left\{\frac{n\lambda_m}{\mu_{\min}\left(\mathbf{K}_{>m}\right) + \lambda}, 1\right\}^2.
\end{aligned}$$

Where in (i) and (ii) we bound these terms analogously to the upper and lower bound of $B1$. The rest is algebra and usage of Assumption 1.

Now we move to the bound of $V2$, here we get that:

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

Finally, the bound for the variance follows directly by taking $C_1 = 6r^2, C_2 = 1/(6r^2)$. □

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 $\mathcal{S}_1(x, x') = \mathbb{E}_{z\sim\nu}[\mathcal{K}(x, z)\mathcal{K}(x', z)]$ can be written as:*

$$\mathcal{S}_1(x, x') = \sum_k \lambda_k^2\psi_k(x)\psi_k(x').$$

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

Proof. The proof simply follows from

$$\mathcal{S}_1(x,x') = \sum_{k,k'} \lambda_k \lambda_{k'} \psi_k(x) \underbrace{\mathbb{E}_{z\sim\nu}\left[\psi_k(z)\psi_{k'}(z)\right]}_{\delta_{k,k'}} \psi_{k'}(x') = \sum_k \lambda_k^2 \psi_k(x)\psi_k(x').$$

□

Lemma 6 (Squared kernel tail)**.** *Let $\mathbf{S}_{>m}$ be the kernel matrix of the squared kernel $\mathcal{S}_1$, 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, let us define $\mathbf{\Psi}_k \in \mathbb{R}^{n\times n}, [\mathbf{\Psi}_k]_{ij} = \psi_k(x_i)\psi_k(x_j)$. 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 a decreasing order, we have that $\lambda_k \leq \lambda_{m+1}$ for any $k > m$, hence:

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

Concluding the proof. □

Lemma 7 (Lemma 20 from Bartlett et al. (2020))**.** *Let $\mathbf{\Psi}_{\leq m}, \mathbf{D}_{\leq m}, \mathbf{H}, \mathbf{H}_{\leq m}$ be defined as above, then:*

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

Proof. The proof is simple Woodbury identity with some added algebra.

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

Where in (i) we used Lemma 20 from Bartlett et al. (2020). □

B THE HYPERCUBE

B.1 FUNCTIONS ON THE HYPERCUBE

In this section we focus on the specific hypercube domain $\{-1,1\}^d$ together with the uniform probability density, which has been previously studied in Misiakiewicz & Mei (2021). Define the following inner product:

$$\langle f, g\rangle_{\{-1,1\}^d} := \mathbb{E}_{x\sim\mathcal{U}(\{-1,1\}^d)}[f(x)g(x)].$$

We can now define, for any $S \subseteq \{1,\dots,d\}$:

$$\mathcal{Y}_S^{(d)}(x) := \prod_{i\in S} x_i. \tag{9}$$

It's easy to see that $\{\mathcal{Y}_S\}_{S\subseteq\{1,\dots,d\}}$ is an orthonormal set of functions with respect to $\langle\cdot,\cdot\rangle_{\{-1,1\}^d}$. We will later show that such functions are the eigenfunctions of the kernel of our interest. Furthermore, let us define the polynomials:

$$\mathcal{G}_l^{(d)}\left(\frac{\langle x, x'\rangle}{\sqrt{d}}\right) := \frac{1}{\mathcal{B}(l,d)} \sum_{|S|=l} \mathcal{Y}_S^{(d)}(x)\mathcal{Y}_S^{(d)}(x'), \tag{10}$$

$$\mathcal{B}(l,d) := |\{S \subseteq \{1,\dots,d\} \mid |S| = l\}| = \binom{d}{l}. \tag{11}$$

Notice that $\mathcal{G}_l^{(d)}$ depends only on the inner product of x and y. Furthermore, note that the $\mathcal{G}_l^{(d)}$ are a set of orthonormal polynomial with respect to the distribution of $\langle x, x' \rangle / \sqrt{d}$. Now this general lemma shows how we can decompose any rotational invariant kernel in such eigenbases.

Lemma 8 (Local kernel decomposition)**.** *Let $\kappa : \mathbb{R} \to \mathbb{R}$ be any function, then we can write $\kappa(\langle x, x' \rangle / d)$ as:*

$$\kappa\left(\frac{\langle x, x' \rangle}{d}\right) = \sum_{l=0}^{d} \xi_l^{(d)} \frac{1}{\mathcal{B}(l,d)} \sum_{|S|=l} \mathcal{Y}_S(x)\mathcal{Y}_S(x'). \tag{12}$$

Proof. For the proof of this lemma we are only interested at the evaluation of κ in the values that $\frac{\langle x, x' \rangle}{d}$ can take, given that $x, x \in \{-1, 1\}^d$. Hence we are interestend only in κ computed in $\{-1, -1 + 2/d, \dots, -2/d + 1, 1\}$. Since this set has a cardinality of $d + 1$, we can write κ as a linear combination of $d + 1$ linearly independent functions. This is the case for $\{\mathcal{G}_l^{(d)}\}_{l=0}^d$ since such function are uncorrelated. Hence we can write:

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

Finally, the proof follows by the expansion of $\mathcal{G}_l^{(d)}$ in Equation 10. □

While this lemma is very general, we need some regularity conditions on $\xi_l^{(d)}$, for our theory to go through.

Assumption 3 (Regularity of κ)**.** *In the setting presented in Section 3.1, we say that the κ satisfy the genericity assumption if, writing κ as in Equation 12, there exist constants $c_1, c_2, c_3, c_4 > 0$ such that:*

$$\xi_l^{(q)} \geq c_1 \ \forall l \leq \left\lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \right\rfloor + 1, \qquad \sum_{l > L} \xi_l^{(q)} \geq c_2, \qquad \sum_{l \geq 0} \xi_l^{(q)} \leq c_3,$$

for any $q \geq c_4$.

This condition is quite mild, and it only asks for the tail of the polynomial expansion to not vanish. This allows our model to have the so-called implicit regularization. Furthermore, the expansion of the polynomial degree that we have to learn also to be non-vanishing, this makes the task to learn polynomial of degree L possible.

Now we move to the study of the convolutional kernel to build the tools to prove Lemma 4.

B.2 Convolutional kernel

In this section we propose tool to analyze convolutions in the setting proposed in Section 3.1. Then, we use them to prove Lemma 4 and Lemma 2.

Definition 1 (Diameter of a subset)**.** *Let S be a subset of $\{1, \dots, d\}$, then we call diameter of S:*

$$\gamma(S) = \max\{\min\{\mathrm{mod}\,(j - i, d) + 1, \mathrm{mod}\,(i - j, d) + 1\} \mid i, j \in S\}.$$

Furthermore, we define $\mathcal{B}$ to be:

$$\mathcal{C}(l, q, d) := |\{S \subseteq \{1, \dots, d\} \mid |S| = l, \gamma(S) \leq q\}|. \tag{13}$$

The diameter of S can be understood as the size of the smallest patch that fully contains S.

Proposition 1 (Proposition 1 from Misiakiewicz & Mei (2021))**.** *Let $\mathcal{K}$ be a kernel defined as in Equation 1. Then it can be written as:*

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

The proof of Proposition 1 follows by applying Lemma 8 once on each of the patches, and then sum everything. The factor $q+1-\gamma(S)$ stems from the fact that the eigenfunction $\mathcal{Y}_S$ appears as many times as the number of patches of size q, that cover S. For a formal proof look at Proposition 1 from Misiakiewicz & Mei (2021).

Now we present two lemmas that help us to better characterize the eigenvalues.

Lemma 9 (Number of overlapping sets)**.** *Assume $q < d/2$, then:*

$$\mathcal{C}(l,q,d) = d\binom{q-1}{l-1}.$$

Proof. Let $\tilde{\mathcal{C}}(l,\gamma,d)$ be the number of subset of $\{1,\dots,d\}$, of cardinality l, where the diameter is exactly γ. To compute such value start by computing $\tilde{\mathcal{C}}(l,2,d)$. We can choose the first of the two elements to be in d different position, and the second either. q after it, or before it. Finally, in this way, we count each set exactly twice. Hence:

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

Now, if $l > 2$. We can build all the possible sets by starting with one of the d sets of diameter γ and size 2, and then adding the leftovers $l-2$ elements. Note that if we start with different sets of size 2, we must achieve different sets of size l thanks to the diameter condition. Finally, we need to add those $l-2$ elements in any of the $q-2$ elements between the two in the set of size 2: this can be done in $\binom{\gamma-2}{l-2}$. Finally, combining the two ideas we get that:

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

Hence, we can also compute:

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

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

Lemma 10 (Corollary of Proposition 1)**.** *Assume that $q \in \Theta(d^\beta)$. Let $\{S_d\}_{d\geq l}$ be a sequence of sets such that: $S_d \subseteq \{1,\dots,d\}$, such that $|S_d| = l$ for any $d \geq l$, and such that $\gamma(S_d) \leq q$. Let $\mathcal{K}$ be a kernel respecting Assumption 3 for l. Then, writing our kernel as:*

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

we have that:

$$\frac{1}{d^{1+\beta l.}} \lesssim \lambda_{S_d} \lesssim \frac{1}{d^{1+\beta(l-1)}}.$$

In particular, there exists a sequence $\{S_d\}_{d\geq l}$ such that $\lambda_{S_d} \in \Theta\left(\frac{1}{d^{1+\beta l.}}\right)$.

Proof. Let us define the sequence $\{\mathbf{r}_d\}_{d\geq l}$, where $\mathbf{r}_d = q+1-\gamma(S_d)$. Now, since $l \leq \gamma(S_d) \leq q$, we have that:

$$1 \leq \mathbf{r}_d \leq q+1-l, \qquad \forall d \geq l.$$

In particular, we can take the sequence $\{S_d\}_{d\geq l}$ such that $\gamma(S_d) = q$, for any d: in this case $\mathbf{r}_d = 1$ for all d. Now, thanks to Assumption 3, can write λ_{S_d} as:

$$\lambda_{S_d} = \frac{1}{d}\frac{\xi_l^{(q)}}{\mathcal{B}(l,q)}\mathbf{r}_d \leq \frac{1}{d}(q+1-l)\frac{c_2}{\binom{q}{l}} \in \mathcal{O}\left(\frac{q}{dq^l}\right).$$
$$\geq \frac{1}{d}\frac{c_1}{\binom{q}{l}} \in \Omega\left(\frac{1}{dq^l}\right).$$

In particular, we can take the sequence $\{S_d\}_{d\geq l}$ such that $\gamma(S_d) = q$, for any d: in this case $\mathbf{r}_d = 1$, and thus $\lambda_{S_d} \in \Theta(1/dq^l)$. The proof follows remembering that $q \in \Theta(d^\beta)$. □

Finally, we move to the proof of the two lemmas of interest:

Proof of Lemma 4. We show that m is in $\mathcal{O}(nd^{-\delta})$. Thanks to Proposition 1, we can write the convolutional kernel as:

$$\mathcal{K}(x,x') = \sum_{l=1}^{q} \sum_{\substack{\gamma(S)\leq q \\ |S|=l}} \lambda_S \mathcal{Y}_S(x)\mathcal{Y}_S(x') = \sum_{i>0} \lambda_{S_i}\mathcal{Y}_{S_i}(x)\mathcal{Y}_{S_i}(x').$$

Where S_i is an ordering such that $\lambda_{S_{i+1}} \leq \lambda_{S_i}$. We now show that m is such that $|S_m| = L+1$ for d sufficiently large. By contradiction, assume that $|S_m| = \hat{L} < L+1$, then:

$$n\lambda_m = n\lambda_{S_m} \in \Omega\left(\frac{d^{1+\beta L+\ell_\lambda+\beta\delta}}{d^{1+\beta\hat{L}}}\right) = \Omega\left(\lambda \cdot d^{\beta(L-\hat{L}-\delta)}\right) = \omega(\max\{\lambda,1\}).$$

But this contradicts the assumption that $n\lambda_m \in \Theta(\max\{\lambda,1\})$. Analogously, if we assume that $|S_m| = \hat{L} > L+1$, it is easy to see that $n\lambda_m \in o(\max\{\lambda,1\})$, which again contradicts the same assumption.

Now, thanks to what we have just found and Proposition 1, we can write λ_m as:

$$\begin{aligned} n\lambda_m = n\frac{\xi_{L+1}^{(q)}(q+1-\gamma(\mathbf{S}_m))}{d\mathcal{B}(L+1,q)} &\in \Theta\left(\frac{d\cdot d^{\ell_\lambda}\cdot q^{L+\delta}(q+1-\gamma(\mathbf{S}_m))}{d\cdot q^{L+1}}\right) \\ &= \Theta\left(d^{\ell_\lambda}q^{\delta-1}(q+1-\gamma(\mathbf{S}_m))\right). \end{aligned}$$

Since we want that $n\lambda_m \in \Theta(d^{\ell_\lambda})$, we need $q+1-\gamma(\mathbf{S}_m) \in \Theta(q^{-(\delta-1)})$. Hence, $\gamma(\mathbf{S}_m) \in \Theta(q)$. Finally, we need to find m itself. Hence, we need to count how many S are there such that $\lambda_S < \lambda_m$. Thanks to Lemma 10, we have that, for d sufficiently large, for all S of cardinality smaller than L, $\lambda_S < \lambda_m$; and for all S of cardinality larger than L, $\lambda_S > \lambda_m$.

Finally, if S of cardinality L, than $\lambda_S < \lambda_m$ if and only of $\gamma(S) < \gamma(S_m)$. The number of such S, is exactly $\mathcal{C}(L,\gamma(S_m),d)$. Hence, we can compute m as:

$$\begin{aligned} m &= \mathcal{C}(L,\gamma(S_m),d) + \sum_{l=0}^{L-1}\mathcal{C}(l,q,d) \\ &= d\left(\binom{\gamma(S_m)-1}{L-1} + \sum_{l=0}^{L-1}\binom{q-1}{l-1}\right) \\ &\overset{(i)}{\in} \Theta\left(dq^{L-1}\right). \end{aligned}$$

Where in (i) we used that $\binom{q}{l} \in \Theta(q^l)$, as well as $\gamma(\mathbf{S}_m) \in \Theta(q)$. Finally, we can see n as

$$n \in \Theta(d^\ell) = \Theta\left(d\cdot d^{\ell_\lambda} d^{\beta\frac{\ell-\ell_\lambda-1}{\beta}}\right) = \Theta\left(d\cdot d^{\ell_\lambda}q^{L+\delta}\right). \tag{14}$$

Which, plugging in, yields that $m \in \Theta(n\cdot d^{-\ell_\lambda}\cdot q^{-\delta}) = \mathcal{O}(n\cdot q^{-\delta})$.

Now we need to find the rate of λ_m. But this is not hard since we know that $n\lambda_m \in \Theta(\max\{\lambda,1\})$:

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

Where (i) follows by the assumption that $\ell_\lambda \leq L+1$. □

Proof of Lemma 2. It is easy to see that $f^* = \mathcal{Y}_{S^*}$, where $S^* = \{1,\dots,L^*\}$. Now, notice that $|S^*| = \gamma(S^*) = L^*$. Hence, we have a formula for $\lambda_{S^*}^{-1}$ in Proposition 1:

$$\lambda_{S^*} = \xi_{L^*}^{(q)}\frac{q+1-\gamma(S^*)}{d\mathcal{B}(L^*,q)} \in \Theta\left(\frac{q}{dq^{L^*}}\right) = \Theta\left(\frac{1}{dq^{L^*-1}}\right).$$

Finally, as long as $\lambda_m \in o\left(1/dq^{L^*-1}\right)$, we have that, since eigenvalues are sorted, the index of the eigenfunction $\mathcal{Y}_{S^*}$ is less than m. Hence, for d sufficiently large, f^* is in the span of the first m eigenfunctions. Finally, if this is the case:

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

Concluding the proof. □

C MATRIX CONCENTRATION

The proof of Lemma 3 is a quite direct consequence of Theorem 5.44 from Vershynin (2012). We state the proof here.

Proof of Lemma 3. Let us use $[\mathbf{\Psi}_{\leq m}]_{i,:}$ to indicate the i-th row of $\mathbf{\Psi}_{\leq m}$. Then we can see that:

$$\|[\mathbf{\Psi}_{\leq m}]_{i,:}\|^2 = \sum_{i=1}^{m} \mathcal{Y}_i(x_i)^2 \overset{(i)}{=} m.$$

Where (i) follows from the fact that $\mathcal{Y}_i(x)^2 = 1$ for any i for any x. To see this, it suffices to look at the eigenfunctions, as presented in Equation 9. Then:

$$\mathcal{Y}_i(x)^2 = \prod_{j \in S_i} x_j^2 = \prod_{j \in S_i} 1 = 1.$$

Note the rows are isotropic since the eigenfunctions form an orthonormal basis. Since the rows of $\mathbf{\Psi}_{\leq m}$ are independent (each row depends on a different sample, and samples are assumed to be i.i.d.), we can apply Theorem 5.44 from Vershynin (2012). This yields with probability $1 - m\exp\{-ct^2\}$:

$$\left\| \frac{\mathbf{\Psi}_{\leq m}^{\intercal} \mathbf{\Psi}_{\leq m}}{n} - \mathbf{I}_m \right\| \leq t\sqrt{\frac{m}{n}}.$$

As long as $t\sqrt{\frac{m}{n}} < 1$. We then take $t = \frac{1}{2}\sqrt{\frac{n}{m}}$ to achieve the desired bound. This yields a statement which is true with probability $1 - m\exp\{-cd^{\delta}\} > 1 - d^{-\delta}$ for d sufficiently large. □

The proof Lemma 1 is quite tricky since we need to combine diverse strategies for different parts of the proof. For this reason, we first introduce some definitions.

Definition 2 (Of useful quantities)**.** *In the same setting as in Section 3.1, let* $n \in \Theta(d^{\ell})$, $q \in \Theta\left(d^{\beta}\right)$, $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$. *Let* m *be such that* $n\lambda_m \in \Theta(\max\{\lambda, 1\})$ *Then we define:*

$$L := \left\lfloor \frac{\ell - \ell_\lambda - 1}{\beta} \right\rfloor, \qquad \delta := \frac{\ell - \ell_\lambda - 1}{\beta} - L,$$
$$\bar{L} := \left\lfloor \frac{\ell - 1}{\beta} \right\rfloor, \qquad \bar{\delta} := \frac{\ell - 1}{\beta} - \bar{L}.$$

We then define:

$$\mathcal{K}_1(x, x') := \sum_{i > m \,|\, |S_i| \leq \bar{L}+1} \lambda_{S_i} \mathcal{Y}_{S_i}(x) \mathcal{Y}_{S_i}(x')$$
$$\mathcal{K}_2(x, x') := \sum_{\substack{\gamma(S) \leq q \\ |S| \geq \bar{L}+2}} \lambda_S \mathcal{Y}_S(x) \mathcal{Y}_S(x')$$

Now, let $\mathcal{S}_1 := \{S_i : i > m \text{ and } |S_i| \leq \bar{L} + 1\}$*, and let us define:*

$$\mathbf{\Psi}_1 \in \mathbb{R}^{n \times |\mathcal{S}_1|},\ [\mathbf{\Psi}_1]_{j,l} = \mathcal{Y}_{S_l}(x_j), \qquad \mathbf{D}_1 = \mathrm{diag}(S_1, \ldots, S_{|\mathcal{S}_1|}), \tag{15}$$

where S_l *is the* l*-th element of* $\mathcal{S}_1$ *in a fixed order. Finally, let us naturally define the respective empirical kernel matrices* $\mathbf{K}_1$ *and* $\mathbf{K}_2$*.*

Lemma 11 ($\mathcal{K}1, \mathcal{K}2$ decomposition)**.** *In the same setting as Definition 2, for* d *sufficiently large, we have that:*

$$\mathcal{K}_m(x, x') = \mathcal{K}_1(x, x') + \mathcal{K}_2(x, x').$$

Proof. While it is clear that every eigenfunction beyond m appears exactly once among $\mathcal{K}_1$ and $\mathcal{K}_2$, we need to guarantee that we are not considering any of the eigenfunctions before m. In $\mathcal{K}_1$ there are

no eigenfunction before m by definition. In $\mathcal{K}_2$ all eigenfunctions are polynomials of degree at least $\bar{L}+2$. Hence, thanks to Lemma 10, we have that all the eigenvalues are such that:

$$n\lambda_S \in \mathcal{O}\left(\frac{n}{d^{1+\beta(\bar{L}+2-1)}}\right) = \mathcal{O}\left(\frac{n}{d^{1+\beta(\bar{L}+\delta)}} d^{-(1-\delta)}\right) = \mathcal{O}(d^{-(1-\delta)}) = o(n\lambda_m).$$

Hence, for d sufficiently large all eigenvalues are smaller than λ_m, which yields the result. □

Lemma 12 (Bound on $\mathbf{K}_1$)**.** *In the setting defined in Section 3.1, we have that:*

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

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

Proof. The proof follows directly from Lemma 3. We only need to show that $|\mathcal{S}_1|$, which is just the number of columns of $\mathbf{\Psi}_1$, is in $\mathcal{O}(d \cdot q^{\bar{L}})$. But this follows simply my noticing that in $\mathcal{S}_1$ we include only polynomials of degree no larger than $\bar{L}+1$. Hence:

$$|\mathcal{S}_1| \leq \sum_{l=1}^{\bar{L}+1} \mathcal{C}(l,q,d) \overset{(i)}{\in} \mathcal{O}(d \cdot q^{\bar{L}}).$$

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

Lemma 13 (Bound on $\mathbf{K}_2$)**.** *In the setting defined in Section 3.1, we have that:*

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

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

Proof. We start by decomposing as:

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

Where we (i) follows from the definition of $\mathcal{Q}_l^{(d,q)}(x,x')$:

$$\mathcal{Q}_l^{(d,q)}(x,x') := \sum_{\substack{\gamma(S)\leq q \\ |S|=l}} \frac{q+1-\gamma(S)}{d\mathcal{B}(l,q)} \mathcal{Y}_S(x)\mathcal{Y}_S(x'). \tag{16}$$

Finally, let us define $\mathbf{Q}_l^{(d,q)}$ to be the kernel matrix with respect to $\mathcal{Q}_l^{(d,q)}$. Now, we can finally bound:

$$\|\mathbf{K}_2\| = \left\| \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \mathbf{Q}_l^{(d,q)} \right\| \leq \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \left\| \mathbf{Q}_l^{(d,q)} \right\| \overset{(i)}{\leq} 1.5 \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \overset{(ii)}{\leq} 1.5c_3. \tag{17}$$

Where (i) follows from Lemma 14 with probability at least $1 - q^{-\bar{\delta}}$, and (ii) follows from Assumption 3. For the lower bound we simply look at this inequality:

$$\mu_{\min}(\mathbf{K}_2) \geq \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \mu_{\min}\left(\mathbf{Q}_l^{(d,q)}\right) \geq \frac{1}{2} \sum_{l=\bar{L}+2}^{q} \xi_l^{(q)} \overset{(i)}{\geq} \frac{c_2}{2}.$$

Where (i) follows from Lemma 14 with probability at least $1 - q^{-\bar{\delta}}$. □

Finally, we get to the proof of Lemma 1.

Proof of Lemma 1. We prove the Lemma in two parts. First, we address the bound on r, and then we cover the bounds on $\mathbf{Tr}(\mathbf{S}_{>m})$.

Bound on r

Step 1: rephrase the problem To upper and lower bound r, we notice that it suffices to show that:

$$\|\mathbf{K}_{>m}\| \leq c_1 \max\{\lambda, 1\}, \qquad \mu_{\min}(\mathbf{K}_{>m}) \geq c_2. \tag{18}$$

From this we would get that:

$$\|\mathbf{K}_{>m}\| + \lambda \leq (1 + c_1)\max\{\lambda, 1\}, \qquad \mu_{\min}(\mathbf{K}_{>m}) + \lambda \geq (1 + c_2)\max\{\lambda, 1\}$$

This then implies that:

$$\frac{1 + c_2}{1 + c_1} \leq r = \frac{\|\mathbf{K}_{>m}\| + \lambda}{\mu_{\min}(\mathbf{K}_{>m}) + \lambda} \leq \frac{1 + c_1}{1 + c_2}.$$

Hence, we must now show that Equation (18) holds with high probability in the setting of interest. Throughout this proof we use multiple times the fact that $\mathbf{K}_{>m} = \mathbf{K}_1 + \mathbf{K}_2$.

Step 2: upper bound

$$\begin{aligned}
\|\mathbf{K}_{>m}\| &\leq \|\mathbf{K}_1\| + \|\mathbf{K}_2\| \\
&\overset{(i)}{=} n\left\|\frac{\mathbf{D}_1 \mathbf{\Psi}_1^t \mathbf{\Psi}_1}{n}\right\| + c \\
&\overset{(ii)}{=} 1.5 n\lambda_{m+1} + c \\
&\overset{(iii)}{\lesssim} \max\{\lambda, 1\}.
\end{aligned}$$

Where in (i) we used Lemma 13, in (ii) Lemma 12, and in (iii) the fact that $n\lambda_m \in \Theta(\max\{\lambda, 1\})$.
Step 3: lower bound

$$\mu_{\min}(\mathbf{K}_{>m}) \overset{(i)}{\geq} \mu_{\min}(\mathbf{K}_2) \overset{(ii)}{\geq} c.$$

Where in (i) we used that $\mathbf{K}_1$ is postive-semidefinite, and in (ii) we used Lemma 13.

Bound on $\mathbf{Tr}(\mathbf{S}_m)$: We start by dividing the squared kernel as:

$$\mathcal{S}_{1>m}(x, x') = \underbrace{\sum_{i>m||S_i|\leq \bar{L}+1} \lambda_{S_i}^2 \mathcal{Y}_{S_i}(x)\mathcal{Y}_{S_i}(x')}_{:=\mathcal{S}_{11}(x,x')} + \underbrace{\sum_{\substack{\gamma(S)\leq q \\ |S|>\bar{L}+1}} \lambda_S^2 \mathcal{Y}_S(x)\mathcal{Y}_S(x')}_{:=\mathcal{S}_{12}(x,x')}.$$

Note that we are dividing the kernel as in Definition 2. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the kernel matrix of respectively $\mathcal{S}_{11}$ and $\mathcal{S}_{12}$.

Upper bound: here we need to upper-bound the Trace of $\mathbf{S}_{>m}$:

$$\mathbf{Tr}(\mathbf{S}_{>m}) = \mathbf{Tr}(\mathbf{S}_1) + \mathbf{Tr}(\mathbf{S}_2).$$

To bound $\mathbf{Tr}(\mathbf{S}_1)$, notice that $\mathbf{S}_1 = \mathbf{\Psi}_1 \mathbf{D}_1^2 \mathbf{\Psi}_1^\intercal$, hence:

$$\begin{aligned}
\mathbf{Tr}(\mathbf{S}_1) = \mathbf{Tr}\left(\mathbf{\Psi}_1 \mathbf{D}_1^2 \mathbf{\Psi}_1^\intercal\right) = n\mathbf{Tr}\left(\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\mathbf{D}_1^2\right) &\leq n\left\|\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\right\| \sum_{i>m||S_i|\leq \bar{L}+1} \lambda_{S_i}^2 \\
&\overset{(i)}{\leq} 1.5n \sum_{i>m||S_i|\leq \bar{L}+1} \lambda_{S_i}^2.
\end{aligned}$$

Where (i) follows from Lemma 12. Now we move our attention to $\mathbf{Tr}(\mathbf{S}_2)$. Let $\bar{m}$ be the largest eigenvalue of a polynomial of degree $\bar{L} + 2$. Then:

$$\mathbf{Tr}(\mathbf{S}_2) \overset{(i)}{\leq} \lambda_{\bar{m}} \mathbf{Tr}(\mathbf{K}_2) \leq n\lambda_{\bar{m}} \|\mathbf{K}_2\| \overset{(ii)}{\leq} 1.5 n\lambda_{\bar{m}} c_3$$

Where (i) follows from Lemma 5, and (ii) from Lemma 13.

Lower bound: For the lower bound we move on similar lines. For $\mathbf{S}_1$:

$$\mathbf{Tr}(\mathbf{S}_1) = n\mathbf{Tr}\left(\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\mathbf{D}_1^2\right) \geq n\mu_{\min}\left(\frac{\mathbf{\Psi}_1^\intercal \mathbf{\Psi}_1}{n}\right) \sum_{i>m||S_i|\leq \bar{L}+1} \lambda_{S_i}^2 \overset{(i)}{\geq} \frac{1}{2}n \sum_{i>m||S_i|\leq \bar{L}+1} \lambda_{S_i}^2.$$

Where (i) follows from Lemma 12. For $\mathbf{S}_2$ we need to be a bit more precise than in the upper bound.

$$\begin{aligned}
\mathcal{S}_{12}(x,x') &= \sum_{\substack{\gamma(S)\leq q\\ |S|\geq \bar{L}+1}} \lambda_S^2 \mathcal{Y}_S(x)\mathcal{Y}_S(x') \geq \sum_{\substack{\gamma(S)\leq q\\ |S|=\bar{L}+2}} \lambda_S^2 \mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&= (\xi_{\bar{L}+2}^{(q)})^2 \sum_{\substack{\gamma(S)\leq q\\ |S|=\bar{L}+2}} \frac{(q+1-\gamma(S))^2}{d^2\mathcal{B}(\bar{L}+2,q)^2}\mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&= \xi_{\bar{L}+2}^{(q)}\lambda_{\bar{m}} \sum_{\substack{\gamma(S)\leq q\\ |S|=\bar{L}+2}} \frac{\xi_{\bar{L}+2}^{(q)}}{\lambda_{\bar{m}}}\frac{(q+1-\gamma(S))^2}{d^2\mathcal{B}(\bar{L}+2,q)^2}\mathcal{Y}_S(x)\mathcal{Y}_S(x') \\
&\overset{(i)}{\geq} \lambda_{\bar{m}}\xi_{\bar{L}+2}^{(q)} \sum_{\substack{\gamma(S)\leq q\\ |S|=\bar{L}+2}} \frac{q+1-\gamma(S)}{d\mathcal{B}(\bar{L}+2,q)}\mathcal{Y}_S(x)\mathcal{Y}_S(x') = \lambda_{\bar{m}}\mathcal{Q}_{\bar{L}+2}^{(d,q)}(x,x').
\end{aligned}$$

Where in (i) we use the fact that $\bar{m}$ is the index of the smallest eigenvalue of a degree $\bar{L}+2$ polynomial. $\mathcal{Q}_{\bar{L}+2}^{(d,q)}$ is defined in Equation 10. Let $\mathbf{Q}_{\bar{L}+2}^{(d,q)}$ be the kernel matrix of $\mathcal{Q}_{\bar{L}+2}^{(d,q)}$. Then, Lemma 14 yields that $\mu_{\min}\left(\mathbf{Q}_{\bar{L}+2}^{(d,q)}\right) \geq 1/2$ with probability at least $1 - cq^{1-\bar{\delta}}$. From this it follows that:

$$\mathbf{Tr}(\mathbf{S}_{>m}) \geq \mathbf{Tr}(\mathbf{S}_2) \geq n\mu_{\min}(\mathbf{S}_2) \gtrsim n\lambda_{\bar{m}}\mu_{\min}\left(\mathbf{Q}_{\bar{L}+2}^{(d,q)}\right) \gtrsim n\lambda_{\bar{m}}\frac{1}{2}.$$

Cases distinction: now, there are two possible cases:

1. $\bar{L} = L$ (Interpolation or mild regularization) In this setting:

$$n\lambda_{\bar{m}} \overset{(i)}{\in} \Theta\left(\frac{d\cdot d^{\ell_\lambda}\cdot q^{L+\delta}}{d\cdot q^{L+1}}\right) = \Theta\left(d^{\ell_\lambda}\cdot q^{-(1-\delta)}\right)$$

   Where in (i) we use Lemma 10. Furthermore,

$$\begin{aligned}
n\sum_{i>m||S_i|\leq L+1} \lambda_{S_i}^2 &\leq \frac{1}{n}\sum_{|S_i|\leq L+1}(n\lambda_{S_i})^2 \\
&\overset{(i)}{\leq} c\frac{\mathcal{C}(L+1,q,d)}{n}\max\{\lambda,1\}^2 \\
&\in \mathcal{O}\left(\frac{d\cdot q^L}{d\cdot d^{\ell_\lambda}\cdot q^{L+\delta}}d^{2\ell_\lambda}\right) = \mathcal{O}\left(q^{-\delta}d^{\ell_\lambda}\right)
\end{aligned}$$

   Where (i) follows from the fact that $n\lambda_{S_i} \leq c\max\{\lambda,1\}$. The proof follows by combining these rates with the upper and lower bounds on $\mathbf{Tr}(\mathbf{S}_{>m})$ found before and by noticing that $\max\{\lambda,1\} \in \Theta(d^{\ell_\lambda})$.

2. $\bar{L} \geq L+1$ (Strong regularization) In this settings, it is easy to see that $n\lambda_{\bar{m}} \in o(q^{-1})$. And the dominating term for the rates will be the one coming from $\mathbf{Tr}(\mathbf{S}_1)$.

$$n\sum_{i>m||S_i|\leq L+1}\lambda_{S_i}^2 = \underbrace{n\sum_{i>m||S_i|\leq L+1}\lambda_{S_i}^2}_{T1} + \underbrace{n\sum_{|S|=L+2}\lambda_S^2}_{T2} + \underbrace{n\sum_{l=L+3}^{\bar{L}+1}\sum_{|S|=l}\lambda_S^2}_{T3}.$$

$T1$ behaves at $\mathbf{Tr}(\mathbf{S}_1)$ in the case where $\mathbf{b}L = L$, hence $T1 \in \mathcal{O}(q^{-\delta} d^{\ell_\lambda})$. It is easy to see that the eigenvalues in $T3$ are small enough that $T3 \in o(q^{-1})$. Hence, the dominating factor will be $T2$. Now we bound it:

$$\begin{aligned} T2 &= n\xi_{L+2}^{(q)} \sum_{|S|=L+2} \left(\frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \right)^2 \\ &\leq \xi_{L+2}^{(q)} \frac{n \cdot q}{d\mathcal{B}(L+2,q)} \sum_{|S|=L+2} \frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \\ &\in \mathcal{O}\left(\frac{d \cdot d^{\ell_\lambda} q^{L+\delta} \cdot q}{d \cdot q^{L+2}} \right) = \mathcal{O}\left(d^{\ell_\lambda} q^{-(1-\delta)} \right) \end{aligned}$$

Where the proof of $\sum_{|S|=L+2} \frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} = 1$ can be found in the beginning of the proof of Lemma 6, when addressing the diagonal elements of $\mathbf{Q}_l^{(d,q)}$. Finally, we show a lower-bound for $T2$:

$$\begin{aligned} T2 &= n\xi_{L+2}^{(q)} \sum_{|S|=L+2} \left(\frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \right)^2 \\ &\geq \xi_{L+w}^{(q)} n \sum_{|S|=L+2, \gamma(S) \leq q/2} \left(\frac{q+1-\gamma(S)}{d\mathcal{B}(L+2,q)} \right)^2 \\ &\geq \xi_{L+2}^{(q)} n \frac{(q/2)^2}{d^2 \mathcal{B}(L+2,q)^2} \mathcal{C}(L+2, q/2, d) \\ &\in \Omega\left(\frac{d \cdot d^{\ell_\lambda} \cdot q^{L+\delta} \cdot q^2}{d^2 \cdot q^2 (L+2)} d \cdot q^{L+1} \right) = \Omega\left(d^{\ell_\lambda} q^{-(1-\delta)} \right). \end{aligned}$$

Finally, we combine the upper and lower bounds with the rate just found. Noticing that $\max\{\lambda, 1\} \in \Theta(d^{\ell_\lambda})$ the proof is concluded.

□

C.1 Random matrix theory lemmas

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

Lemma 14 (Hypercube bound on the tail kernel)**.** *In the same setting as Definition 2, with probability at least* $1 - cq^{(1-\bar{\delta})}$*:*

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

Where $\mathbf{Q}_l^{(d,q)}$ *is defined in Equation 10.*

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 14. First, we show that the diagonal elements of $\mathbf{Q}_l^{(d,q)}$ are just 1. Hence, we compute:

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

Where (i) follows from the fact $\mathcal{Y}_S(x)^2 = 1$, and (ii) follows noticing that we are counting a set S exactly $(q + 1\gamma(S))$ times. And, by summing all $\mathcal{C}(l, \gamma, d)$ over γ, we count S again exactly $(q + 1\gamma(S))$ times. Now, let $\mathbf{W}_l^{(d,q)} := \mathbf{Q}_l^{(d,q)} - \mathbf{I}_n$.

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 $l \leq \mathbf{T}$, and $l > \mathbf{T}$. We then show how to pick $\mathbf{T}$.

Terms $l < \mathbf{T}$: here we use Lemma 15. To successfully apply it, however, we need to find an appropriate $\mathcal{M}$, and show that the assumptions of the theorem 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:

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

Hence we chose $\mathcal{M}^{(l,q)} = d\mathcal{B}(l,q)/lq$. The l factor will be needed for the third assumption to hold. For the third assumption we compute:

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

Where (i) follows from hypercontractivity as in Lemma 16, and (ii) follows from the formula $\binom{q-1}{l-1} = \frac{q}{l}\binom{q}{l}$, and the definition of $\mathcal{C}(l,q,d)$, $\mathcal{B}(l,q)$ and $\mathcal{M}^{(l,q)}$.

Finally, we can apply Lemma 15 and we get that for any $p > 2$:

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

In particular, if we take $p \geq \frac{2}{\beta} + \frac{1}{(1-\bar{\delta})\beta} - (\bar{L} + 2 - \bar{\delta})$, 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 $l = \bar{L}+1, \ldots, \mathbf{T}$, and we get:

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

Where $C_{\bar{L}}$ depends only $\bar{L}$ and $\mathbf{T}$, which are fixed in our setting. We later show how to pick $\mathbf{T}$.

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 a fixed amount of it. Hence, we present another technique for the rest of the terms:

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

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

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

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

Where in (i) we used Markov inequality (aslo known as Chebychev for second moments), and in (i) 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, since we assume $l > \mathbf{T}$, we can compute the probability that $\max_{i,j}[\mathbf{W}_l^{(d,q)}]_{ij}^2 \leq 1/4n^2$ for all $l > \mathbf{T}$, which indeed implies that $\|\mathbf{W}_l^{(d,q)}\| \leq 1/2$ for all $l > \mathbf{T}$. Thanks to union bound this happens with probability $1 - \triangle$, where:

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

To conclude the proof, we just take $\mathbf{T} > \frac{4}{\beta} + 4(\bar{L}+\bar{\delta})$. Notice that $\mathbf{T}$ is fixed, and it does not depend on d. Finally, it is easy to verify that $\triangle \in \mathcal{O}(d^{-1})$. Hence:

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

□

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

Lemma 15 (Graph argument)**.** *Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a random matrix such that:*

- $\mathbf{W}_{i,i} = 0,\ \forall i \in [n]$.
- *There exists a $\mathcal{M} > 0$, such that*

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

- *There exists a l such that, for any p:*

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

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. Doing the same steps done in the proof of Proposition 3 from Ghorbani et al. (2021), we get that:

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

Note that for Ghorbani, $\mathcal{M} = d^l$, but we move in a slightly more general setting. Furthermore, instead of Lemma 4 from 15 within the proof, we use Lemma 16.

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 16 (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 Equation 10.

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 Proof of Theorem 2: Training error

The formal version of Theorem 2 provides rates for the average training error over noise, in two settings: when variance domiantes, and when bias dominates. The informal statement follows simply noticing that for $\beta < \beta^*$, the variance dominates (hence $v_0 > b_0$ in the notation of Theorem 4), and for $\beta > \beta^*$, the bias dominates $b_0 > v_0$

Theorem 4 (Formal version of Theorem 2)**.** *In the same setting as Theorem 1 assume that for two constants $b_0, v_0 \geq 0$:*

$$\textbf{\textit{Bias}}^2(\hat{f}_0) \in \Theta(d^{-b_0}), \qquad \textbf{\textit{Variance}}(\hat{f}_0) \in \Theta(d^{-v_0})$$

*Furthermore, let λ_{opt} be such that **Risk**$(\hat{f}_{\lambda_{opt}})$ is minimal. Then, if $v_0 > b_0$, with probability at least $1 - cd^{-\beta\min\{\delta,1-\delta\}}$, λ_{opt} is $\Theta(d^{\ell_\lambda})$ and*

$$\left|\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_i(\hat{f}_{\lambda_{opt}}(x_i) - y_i)^2\right] - \sigma^2\right| \in \mathcal{O}(d^{-\ell_{\lambda_{opt}}} + q^{-\bar{\delta}} + d^{-b}).$$

If instead $v_0 < b_0$, with probability at least $1 - cd^{-\beta\min\{\delta,1-\delta\}}$, $\lambda_{opt} \leq \Lambda$ for a constant Λ, and

$$\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_i(\hat{f}_{\lambda_{opt}}(x_i) - y_i)^2\right] \leq c\sigma^2 + \mathcal{O}(d^{-b}).$$

For some constant $c < 1$.

Proof. **Case** $(v_0 > b_0)$**.** For the proof of this part, we just show that, as ℓ_λ increases, the rate of bias monotonically increases (slower rate), and the one of the variance decreases (faster rate). Hence,

there exists a $\ell_{\lambda\text{opt}}$ such that the two rates match. Furthermore, for such $\ell_{\lambda\text{opt}}$ the maximum between the rate of the variance and the bias is minimized, hence, such $\ell_{\lambda\text{opt}}$ is the rate of λ_{opt}.

Looking at the rates from Theorem 1 it is clear that the exponent of bias decreases, with ℓ_λ increasing. For the variance we need to be more careful, since also δ depends on ℓ_λ:

$$\text{exponent of the variance} = \begin{cases} \frac{-\ell_\sigma - \ell_\lambda - \beta\delta}{\ell} & \delta < 1-\delta \\ \frac{-\ell_\sigma - \ell_\lambda - \beta(1-\delta)}{\ell} & \delta \geq 1-\delta \end{cases}$$
$$= \begin{cases} \frac{-\ell_\sigma - \ell + 1}{\ell} & \delta < 1-\delta \\ \frac{-\ell_\sigma - 2\ell_\lambda - 2 + \ell}{\ell} & \delta \geq 1-\delta \end{cases}$$

Finally, taking the derivative with respect to ℓ_λ we get:

$$\partial_{\ell_\lambda} \text{exponent of the variance} = \begin{cases} 0 & \delta < 1-\delta \\ -\frac{2}{\ell} & \delta \geq 1-\delta \end{cases} \leq 0.$$

Finally, since the exponent of the variance is continuous, and the derivative with respect to ℓ_λ is never positive, we have that the rate of the variance is non-increasing with ℓ_λ. Hence, the optimal $\ell_{\lambda\text{opt}}$ is well defined. Further, since v_0 is strictly larger than b_0, it is easy to see that such $\ell_{\lambda\text{opt}}$ is strictly larger than 0.

Now we move to the bound of the training error. We start by using Lemma 18

$$\left| \mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_\lambda(x_i) - y_i)^2 \right] - \sigma^2 \right| \leq \sigma^2 \left| \frac{\lambda^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \right| + 12 \left| \lambda^2 \frac{c_2}{c_2} \frac{\|\mathbf{D}_{>m}{}^{-1}\mathbf{a}\|^2}{n^2} \right|.$$

The second term is proportional to the bias, and hence it vanishes at a rate d^{-b}. We now focus on the first term. For this we use Lemma 17.

$$\frac{n-\bar{m}}{(c_2+\lambda)^2} \leq \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \frac{n}{(c_1+\lambda)^2}$$
$$\frac{n-\bar{m}}{n} \frac{\lambda^2}{(c_2+\lambda)^2} \leq \frac{\lambda^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \frac{\lambda^2}{(c_1+\lambda)^2}$$
$$-\frac{\bar{m}}{n} \frac{\lambda^2}{(c_2+\lambda)^2} + \frac{\lambda^2}{(c_2+\lambda)^2} - 1 \leq \frac{\lambda^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \leq \frac{\lambda^2}{(c_1+\lambda)^2} - 1$$
$$\underbrace{-\frac{\bar{m}}{n} \frac{\lambda^2}{(c_2+\lambda)^2}}_{A} \underbrace{-\frac{c_2\lambda + c_2^2}{(c_2+\lambda)^2}}_{B} \leq \frac{\lambda^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right) - 1 \leq \underbrace{-\frac{c_1\lambda + c_1^2}{(c_1+\lambda)^2}}_{C}.$$

Where $\bar{m}$, defined within Lemma 17 itself, is such that $\bar{m}/n \in \mathcal{O}(q^{-\bar{\delta}})$, and hence $A \in \mathcal{O}(q^{-\delta})$. $B, C \in \mathcal{O}(1/\lambda) = \mathcal{O}(d^{-\ell_{\lambda\text{opt}}})$. Concluding the proof for this case.

Case $(v_0 < b_0)$. The idea of this part of the proof is that regularization benefits variance and hurts bias. However, when variance vanishes at a faster rate, it makes no sense to regularize because we can only reduce an already negligible variance. The reason why we cannot prove a tight bound, hence that $\lambda \to 0$, is that it is not clear that even a small λ hurts the bias.

To formalize this we write:

$$\textbf{Bias}^2(\hat{f}_0) - \textbf{Bias}^2(\hat{f}_\lambda) \leq C_3 \|\mathbf{K}_{>m}\|^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}\|^2}{n^2} - \frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2 \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}\|^2}{n^2}$$
$$\overset{(i)}{\leq} \left(C_3 \|\mathbf{K}_{>m}\|^2 - \frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2 \right) c_1 d^{-b_0}.$$

Where (i) follows on the hypothesis on the bound of the bias of $\hat{f}_0$. Furthermore:

$$\textbf{Variance}(\hat{f}_0) - \textbf{Variance}(\hat{f}_\lambda) \leq \textbf{Variance}(\hat{f}_0) \leq c_2 d^{-v_0}.$$

So finally, we want to check for which value of λ the risk decreases:

$$0 \geq \textbf{Risk}(\hat{f}_\lambda) - \textbf{Risk}(\hat{f}_0)$$
$$= \textbf{Bias}^2(\hat{f}_\lambda) - \textbf{Bias}^2(\hat{f}_0) + \textbf{Variance}(\hat{f}_\lambda) - \textbf{Variance}(\hat{f}_0)$$
$$\geq \left(\frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2 - C_3 \|\mathbf{K}_{>m}\|^2 \right) c_1 d^{-b} - c_2 d^{-v}.$$

Hence, note that if λ is such that $\left(\frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m})+\lambda)^2 - C_3\|\mathbf{K}_{>m}\|^2\right) > c_3 > 0$ with high probability, then we are guaranteed that regularization hurts for d sufficiently large. Hence:

$$\begin{aligned}
0 &\geq \frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m})+\lambda)^2 - C_3\|\mathbf{K}_{>m}\|^2 \\
C_3\|\mathbf{K}_{>m}\|^2 &\geq \frac{1}{3}(\mu_{\min}(\mathbf{K}_{>m})+\lambda)^2 \\
\sqrt{3C_3}\|\mathbf{K}_{>m}\| &\geq \mu_{\min}(\mathbf{K}_{>m})+\lambda \\
\lambda &\leq \sqrt{3C_3}\|\mathbf{K}_{>m}\| - \mu_{\min}(\mathbf{K}_{>m}) = \Lambda.
\end{aligned}$$

Finally, since $\|\mathbf{K}_{>m}\| \leq c_2, \mu_{\min}(\mathbf{K}_{>m}) \geq c_1$ with high probability, we have that the optimal $\lambda \leq \Lambda$ with high probability.

Now we move to the bound of the training error:

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\frac{1}{n}\sum_i (\hat{f}_\lambda(x_i)-y_i)^2\right] &\leq \frac{\sigma^2}{n}\sum_{i=1}^n \frac{\lambda^2}{(\mu_i(\mathbf{K})+\lambda)^2} + 12\lambda^2\frac{c_2}{c_2}\frac{\|\mathbf{D}_{>m}{}^{-1}\mathbf{a}\|^2}{n^2}. \\
&\leq \sigma^2\frac{1}{n}\sum_{i=1}^n \frac{\Lambda^2}{(\mu_{\min}(\mathbf{K}_2)+\Lambda)^2} + \mathcal{O}(d^{-b}). \\
&\leq \sigma^2\frac{\Lambda^2}{(\mu_{\min}(\mathbf{K}_2)+\Lambda)^2} + \mathcal{O}(d^{-b}).
\end{aligned}$$

Where, thanks to Lemma 13, $\mu_{\min}(\mathbf{K}_2) \geq c_1$ with high probability. Hence, $\Lambda^2/(\mu_{\min}(\mathbf{K}_2)+\Lambda)^2 \leq c_2 < 1$. Concluding the proof. □

Here, we present the final two technical lemmas.

Lemma 17 (Trace of the inverse)**.** *In the setting defined in Section 3.1. Let $\bar{m}$ be the number of eigenfunctions which are a polynomial of degree at most $\bar{L}+1$. Then, with probability at least $1-cd^{-\bar{\delta}}$,*

$$\frac{n-\bar{m}}{(c_2+\lambda)^2} \leq \mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq \frac{n}{(c_1+\lambda)^2}. \tag{19}$$

for two positive constants c_1, c_2. Furthermore, $\bar{m} \in \Theta(nq^{-\bar{\delta}})$. $\bar{m}, \bar{L}$ as in Definition 2

Proof. We start by computing $\bar{m}$ using Lemma 9.

$$\bar{m} = \sum_{l=1}^{\bar{L}+1}\mathcal{C}(l,q,d) = \sum_{l=1}^{\bar{L}+1} d\binom{q-1}{l-1} \overset{(i)}{\in} \Theta(d\cdot q^{\bar{m}}).$$

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 $\bar{m} \in \mathcal{O}(nq^{-\bar{\delta}})$.

Now we move to the proof of Eq. (19). We divide $\mathbf{H} = \mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal + \mathbf{K}_1 + \mathbf{K}_2 + \lambda\mathbf{I}$ as in Definition 2. Now, we define $\mathbf{H}_2 = \mathbf{K}_2 + \lambda\mathbf{I}$. For the lower bound it sufficies to see:

$$\mathbf{Tr}\left(\mathbf{H}^{-2}\right) \leq n\|\mathbf{H}^{-2}\| = \frac{n}{\mu_{\min}(\mathbf{H})^2} \leq \frac{n}{\mu_{\min}(\mathbf{H}_2)^2} \overset{(i)}{\leq} \frac{n}{(c_1+\lambda)^2}.$$

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

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

$$\bar{\mathbf{K}} := \mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal + \mathbf{K}, \qquad \bar{\mathbf{\Psi}} := [\mathbf{\Psi}_{\leq m}\ \mathbf{\Psi}_1], \qquad \bar{\mathbf{D}} := \begin{bmatrix}\mathbf{D}_{\leq m} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_1\end{bmatrix}.$$

Then it is easy to see that we can write $\bar{\mathbf{K}} = \bar{\mathbf{\Psi}}\bar{\mathbf{D}}\bar{\mathbf{\Psi}}^\intercal$. Finally, note that $\bar{\mathbf{\Psi}} \in \mathbb{R}^{n\times\bar{m}}$. Now, $\mathbf{H} = \bar{\mathbf{K}} + \mathbf{H}_2$, and we apply Woodbury identity:

$$\mathbf{H}^{-1} = \mathbf{H}_2{}^{-1} - \underbrace{\mathbf{H}_2{}^{-1}\bar{\mathbf{\Psi}}\left(\bar{\mathbf{D}}^{-1} + \bar{\mathbf{\Psi}}^\intercal\mathbf{H}_2{}^{-1}\bar{\mathbf{\Psi}}\right)^{-1}\bar{\mathbf{\Psi}}^\intercal\mathbf{H}_2{}^{-1}}_{A}.$$

Now, it's easy to see that A has rank at most $\bar{m}$, this is because $\bar{\mathbf{\Psi}} \in \mathbb{R}^{n\times\bar{m}}$, hence it has rank $\bar{m}$. The product of matrices has rank no bigger than the smallest rank among the multiplicands, hence also A has rank at most $\bar{m}$. Notice furthermore, that since $\bar{m} \in o(n)$, $\bar{m} < n$ for d sufficiently large. Let now $\{v_1, \dots, v_{\bar{m}}\}$ be an orthonormal basis of col(A), and $\{v_{\bar{m}+1}, \dots, v_n\}$ an orthonormal basis of col$(A)^\perp$. This makes $\{v_1, \dots, v_n\}$ a full orthonormal basis of $\mathbb{R}^n$. Now, since the trace is invariant to similarities, we get that:

$$
\begin{aligned}
\mathbf{Tr}(\mathbf{H}^{-2}) &= \sum_{i=1}^{n} v_i^\intercal \mathbf{H}^{-2} v_i = \sum_{i=1}^{\bar{m}} v_i^\intercal \mathbf{H}^{-2} v_i + \sum_{i=\bar{m}+1}^{n} v_i^\intercal \mathbf{H}^{-2} v_i \\
&\geq \sum_{i=\bar{m}+1}^{n} v_i^\intercal \left(\mathbf{H}_2{}^{-1} - A\right)^2 v_i \\
&\overset{(i)}{=} \sum_{i=\bar{m}+1}^{n} \left(v_i^\intercal \mathbf{H}_2{}^{-2} v_i - \underbrace{v_i^\intercal A}_{=0} \mathbf{H}_2{}^{-1} v_i - v_i^\intercal \mathbf{H}_2{}^{-1} \underbrace{A v_i}_{=0} + \underbrace{v_i^\intercal A A v_i}_{=0} \right). \\
&\geq (n-\bar{m}) \mu_{\min}\left(\mathbf{H}_2{}^{-2}\right) = (n-\bar{m}) \frac{1}{\|\mathbf{H}_2\|^2} \overset{(ii)}{\geq} (n-\bar{m}) \frac{1}{(c_2+\Lambda)^2}.
\end{aligned}
$$

Where in (i) we use the fact that, as long as $i > \bar{m}$, we have that $Av_i = v_i^\intercal A = 0$. Since v_i is orthogonal to the column space. In (ii) we used that $\|\mathbf{K}_2\| \leq c_2$ with high probability thanks to Lemma 13. □

Lemma 18 (Fixed design training error)**.** *In the same setting as Theorem 3, let m be such that Assumption 1 and 2 hold.*

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

Proof. In kernel ridge regression, our estimator is:

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

where $\mathbf{y} := [y_1, \dots, y_n]^\intercal$ and $\mathbf{k}(x^*) := [\mathcal{K}(x_1, x^*), \dots, \mathcal{K}(x_n, x^*)]^\intercal$, $\mathbf{H} = \mathbf{K} + \lambda \mathbf{I}_n$. Hence, our estimator, evaluated at $x_1, \dots, x_n$ 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, $\mathbf{y} = \epsilon + \mathbf{\Psi}_{\leq m}\mathbf{a}$. And we are interested in the expectation with regard to noise, hence:

$$
\begin{aligned}
\mathbb{E}_\epsilon \left[\frac{1}{n} \sum_i (\hat{f}_\lambda(x_i) - y_i)^2 \right] &= \frac{\lambda^2}{n} \mathbb{E}_\epsilon \left[(\boldsymbol{\epsilon} + \mathbf{a}\mathbf{\Psi}_{\leq m})^\intercal \mathbf{H}^{-2} (\boldsymbol{\epsilon} + \mathbf{a}\mathbf{\Psi}_{\leq m}) \right] \\
&= \underbrace{\frac{\lambda^2\sigma^2}{n} \mathbf{Tr}\left(\mathbf{H}^{-2}\right)}_{T1} + \underbrace{\frac{\lambda^2}{n} \mathbf{a}^\intercal \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-2} \mathbf{\Psi}_{\leq m} \mathbf{a}}_{T2}.
\end{aligned}
$$

While $T1$ is already the desired term, for $T2$ we do the following:

$$
\begin{aligned}
T2 &= \frac{\lambda^2}{n}(\mathbf{D}_{\leq m}^{-1}\mathbf{a})^\intercal \mathbf{D}_{\leq m}\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}^{-2}\mathbf{\Psi}_{\leq m}\mathbf{D}_{\leq m}\mathbf{D}_{\leq m}^{-1}\mathbf{a}) \\
&\overset{(i)}{=} \lambda^2 \frac{1}{n}(\mathbf{D}_{\leq m}^{-1}\mathbf{a})^\intercal \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m} \\
&\qquad \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1}\mathbf{a} \\
&= \frac{\lambda^2}{n}\left\| \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1}\mathbf{a}\right\|^2 \\
&\leq \lambda^2 \frac{\left\|\mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-2}\mathbf{\Psi}_{\leq m}\right\|}{n} \left\| \left(\mathbf{D}_{\leq m}^{-1} + \mathbf{\Psi}_{\leq m}^\intercal \mathbf{H}_{>m}^{-1}\mathbf{\Psi}_{\leq m}\right)^{-1} \mathbf{D}_{\leq m}^{-1}\mathbf{a}\right\|^2 \\
&\overset{(ii)}{\leq} \frac{1.5\lambda^2}{(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2} B1. \\
&\overset{(iii)}{\leq} 12\lambda^2 \frac{(\|\mathbf{K}_{>m}\| + \lambda)^2}{(\mu_{\min}(\mathbf{K}_{>m}) + \lambda)^2} \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2} \leq 12\lambda^2 r \frac{\|\mathbf{D}_{\leq m}^{-1}\mathbf{a}_{\leq m}\|^2}{n^2}.
\end{aligned}
$$

Where in (i) we use Lemma 7. In (ii) we use Assumption 1 as well as the fact that we recognize term $B1$ from the proof of Theorem 3. In (iii) we then use the same bound on $B1$ 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.

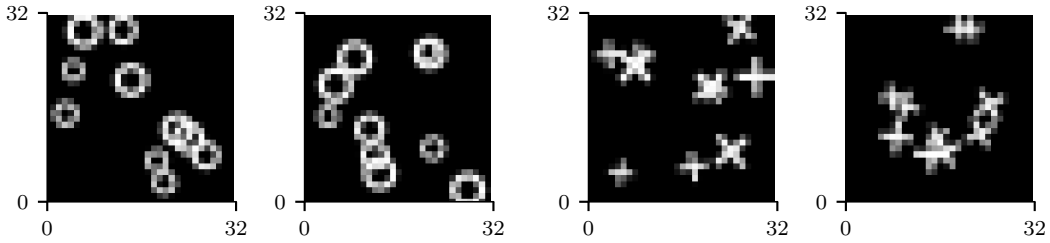
Network architecture We use a special CNN architecture that amplifies the role of filter size as an inductive bias. Each model of the main filter size experiments in Figure 2 has the following architecture:

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

$$
\frac{\gamma_0}{\sqrt{1 + \lfloor t/L \rfloor}}. \tag{20}
$$



(a) Negative samples (b) Positive samples

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

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.

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:
   (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 simply 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 150 epochs over a dataset with 16 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 (20)) 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 (16 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.

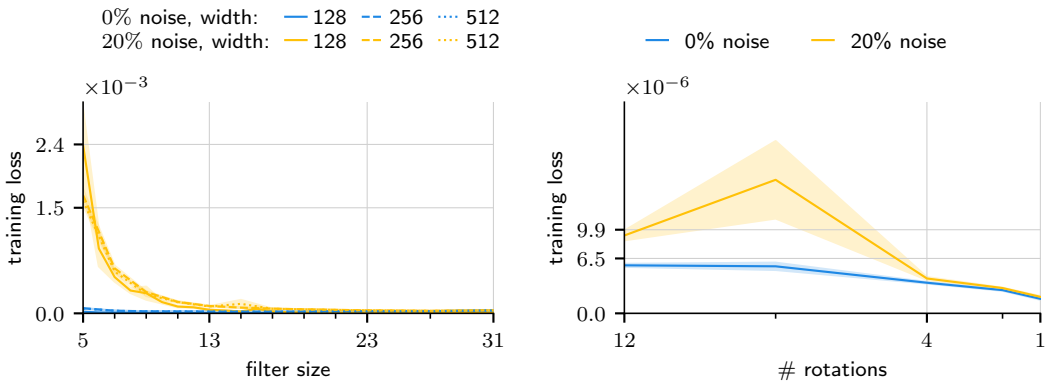
E.3 Difference to double descent

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. Hence, 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.

In the following, we argue that our method of hyperparameter tuning ensures that all models in all experiments lie far in the interpolating regime. Hence, our empirical observations are not a special case of double descent, but an unrelated phenomenon. We further provide experimental evidence by re-running a subset of the filter size experiments using significantly larger models, and observing that the phase transition between harmless and 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



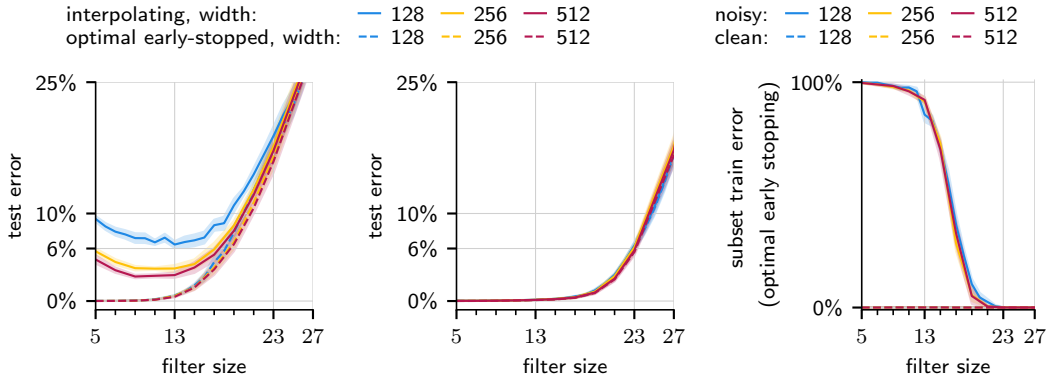
(a) Filter size

(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.

bias with noisy training samples. To avoid optimizing on test data, we tune on dataset seeds that differ from the ones used in actual experiments. Figure 5 displays the final training loss for all empirical settings in this paper. While stronger inductive biases yield increasing 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*, i.e., perfectly fit the corresponding training data.

Increasing model complexity For additional evidence, we further 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 and 512, respectively. We also use a decreased peak learning rate 0.13, since we found the original learning rate 0.2 to be too unstable for the larger model sizes. For computational reasons, we only evaluate a reduced number of filter sizes. 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: overfitting increases with inductive bias strength, and the phase transition between harmless and harmful interpolation persists. 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. Lastly, Figure 6c reveals that the training error of optimally early-stopped models stays the same despite overparameterization.



(a) Test error (20% noise)

(b) Test error (0% noise)

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

Figure 6: An increase in convolution layer width by factors 2 and 4 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 overparameterization, the phase transition between harmless and harmful interpolation persists. Lines show the mean over five random datasets, shaded areas the standard error.