

Asymptotics of Effective Regularization and Overfitting Regimes in Kernel Ridge Regression with Power-Law Spectra

Timur Kim

April 24, 2026

Capstone Project, Spring 2026

Supervisor: Rustem Takhanov, assistant professor

Second reader: Adilet Otemissov, assistant professor

Department of Mathematics, Nazarbayev University

Abstract

Modern machine learning has witnessed a striking empirical phenomenon: overparameterized neural networks, trained to perfectly interpolate noisy training data, often generalize remarkably well, defying classical statistical intuition that such interpolation should lead to overfitting. Understanding this behavior is one of the relevant and interesting challenges in learning theory.

A fruitful approach has been to study kernel ridge regression (KRR). In the infinite-width limit, training a neural network with gradient descent corresponds to kernel regression with the neural tangent kernel (NTK). Building on recent works that derived convenient closed-form expressions for the expected test risk of KRR in terms of the kernel eigenspectrum, we perform an asymptotic analysis for the case of decaying ridge parameter and the exact power-law spectra and the spectra majorized by a power law. We discuss the assumptions underlying these closed-form expressions and indicate how they can be relaxed to more realistic settings, following the narrative of the relevant works in this direction.

Our main results characterize the asymptotic behavior of the effective regularization parameter and the noise overfitting coefficient (which determines the test risk for pure noise). We find that the relative decay rates of the ridge parameter and the kernel eigenvalues governs the overfitting regime. We then connect these results to the effective dimension of the input domain (the support of the data distribution) with respect to the kernel, introduced by Takhanov [14] via Kolmogorov n -widths. This provides a link between the intrinsic dimension of the data (as seen through the kernel) and the type of overfitting. We also obtain upper bounds for spectra majorized by a power law and discuss open problems for further work.

1 Introduction

1.1 Motivation and Background

The past decade has seen a revolution in machine learning driven by deep neural networks. These models are typically vastly overparameterized: they contain much more parameters than training examples. Classical statistical learning theory, rooted in the bias–variance tradeoff, suggests that such overparameterization should lead to severe overfitting, especially when the model is allowed to interpolate the training data exactly. Yet, empirically, deep neural networks trained to near-zero training error often achieve excellent test performance, even in the presence of label noise [15]. This surprising phenomenon has prompted a re-examination of the foundations of generalization theory.

A powerful and now standard tool for gaining theoretical insight is *kernel ridge regression* (KRR). KRR is a nonparametric regression method that operates in a reproducing kernel Hilbert space (RKHS) associated with a positive-definite kernel. A landmark result by [8] shows that in the limit of infinite width, a randomly initialized neural network trained by gradient descent evolves as a linear model in an RKHS whose kernel is the *neural tangent kernel* (NTK). Consequently, insights gained from KRR directly inform the behavior of wide neural networks in the so-called “lazy training” regime. Thus, a thorough understanding of overfitting in KRR is not only interesting in its own right but also serves as a stepping stone toward a theory of generalization in deep learning.

1.2 Three Regimes of Overfitting

To classify the asymptotic generalization behavior of learning algorithms that interpolate noisy data, recent work [1, 11] has proposed a taxonomy consisting of three distinct regimes: **benign**, **tempered**, and **catastrophic** overfitting. These regimes describe what happens to the test error as the number of training samples n grows to infinity, in the presence of a fixed level of label noise.

- **Benign overfitting**: the excess risk tends to zero, so the predictor asymptotically achieves the Bayes-optimal performance.
- **Tempered overfitting**: the excess risk converges to a finite positive constant, meaning the predictor is suboptimal but does not diverge.
- **Catastrophic overfitting**: the excess risk diverges to infinity, i.e., the predictor becomes arbitrarily bad as more data are added.

1.3 Main Contributions

In this work we focus on KRR, where these regimes can be characterized quantitatively in terms of the kernel eigenvalue decay and the strength of regularization. In particular, we will show that for power-law eigenvalue decay $\lambda_i \sim Ci^{-\alpha}$ ($\alpha > 1$) and a decaying ridge parameter $\delta = n^{-a}$ with fixed exponent $a > 0$, the asymptotic behavior of the test error (in the pure noise setting) is governed by the relation between a and $\alpha - 1$, and that this behavior maps directly to the three regimes above.

1. **Asymptotics of effective regularization κ .** We solve the implicit equation for the effective regularization parameter κ in the limit $n \rightarrow \infty$ and identify three asymptotic regimes depending on the relation between decay of eigenvalues and decay of ridge-parameter.
2. **Asymptotics of overfitting coefficient $\mathcal{E}_0$ for pure noise.** Substituting the asymptotics of κ into the expression for $\mathcal{E}_0$, we find that $\mathcal{E}_0$ converges to a finite limit.
3. **Connection to effective dimension d_K .** For kernels with polynomial spectral decay on regular domains, the effective dimension satisfies $\alpha = 1 + 2/d_K$. In the weakly regularized regime, the asymptotic noise overfitting factor becomes $1 + 2/d_K$, linking the intrinsic dimension of the data to the overfitting type. The case $d_K = 0$ lies outside this analysis.
4. **Upper bounds for spectra majorized by a power law.** For kernels whose eigenvalues satisfy $\lambda_i = O(i^{-\alpha})$, we derive upper bounds on $\mathcal{E}_0$ that depend on whether the ridge parameter decays more slowly or more rapidly than a threshold determined by the spectral exponent. These bounds are not tight but illustrate how faster spectral decay can worsen overfitting.

All our results are derived under the Gaussian design (universality) assumption of [12]. We explicitly note this assumption further in work and mention here the recent works by [2] and [5] that study the generalization in KRR in the context of more realistic setting.

1.4 Relation to Prior Work

This paper builds on a rich literature. The closed-form expressions for the test risk of KRR in terms of the eigenspectrum were derived by [13] in the context of Gaussian process regression and later refined by [3, 4, 9, 12]. The taxonomy of overfitting (benign/tempered/catastrophic) was proposed by [11], building on earlier work by [1]. The effective dimension d_K was introduced by [14] using Kolmogorov n -widths; the relation $\lambda_i = O(i^{-(1+2/d_K)})$ is established there. Our asymptotic analysis of κ and $\mathcal{E}_0$ under power-law spectra and power-law ridge extends results by [6] and [10] by providing explicit constants and add interpretation via d_K . The Gaussian design assumption is common in this line of work; for a discussion of its validity and extensions to sub-Gaussian settings we refer to [2] and [5].

1.5 Organization of the Paper

The remainder of this report is organized as follows. Section 2 presents the necessary preliminaries regarding the KRR, the eigenlearning equations, and the definition of effective dimension d_K . Section 3 defines the three overfitting regimes. Section 4 contains our main asymptotic results for power-law spectra and power-law ridge, with the analysis of κ and $\mathcal{E}_0$. Section 5 interprets these results in terms of d_K and the classification of overfitting. Section 6 extends the analysis to spectra majorized by a power law, providing upper bounds. Section 7 discusses open problems and directions for further research. Section 8 concludes. Proofs are provided in the appendices.

2 Preliminaries

2.1 Notation and Basic Setup

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact input domain and let μ be a Borel probability measure on $\mathcal{X}$. We consider a supervised learning setting with n i.i.d. training samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ drawn from the distribution μ on $\mathcal{X}$ and the conditional distribution of y given x . The labels are generated as

$$y_i = f(x_i) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \epsilon^2), \quad (2.1)$$

where $f : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown target function and η_i are independent Gaussian noise variables with variance ϵ^2 . The goal is to estimate f from the training data.

Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a continuous, positive-definite kernel. The *reproducing kernel Hilbert space* (RKHS) associated with K is denoted by $\mathcal{H}_K$. It is a Hilbert space of functions on $\mathcal{X}$ with the property that for every $x \in \mathcal{X}$, the evaluation functional $f \mapsto f(x)$ is continuous, and the kernel satisfies the reproducing property: $f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}_K}$ for all $f \in \mathcal{H}_K$.

Throughout the paper, we use the following standard asymptotic notation for positive functions $f(n)$ and $g(n)$ as $n \rightarrow \infty$:

- $f(n) = O(g(n))$ means there exists a constant $C > 0$ and n_0 such that $f(n) \leq Cg(n)$ for all $n \geq n_0$ (in another notation, this is $f(n) \ll g(n)$).
- $f(n) = \Omega(g(n))$ means $g(n) = O(f(n))$, i.e. $f(n) \geq cg(n)$ for some $c > 0$ and large n ($f(n) \gg g(n)$).
- $f(n) = \Theta(g(n))$ means both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ hold ($f(n) \asymp g(n)$).

- $f(n) = o(g(n))$ if $\lim_{n \rightarrow \infty} f(n)/g(n) = 0$.
- $f(n) = \omega(g(n))$ if $\lim_{n \rightarrow \infty} f(n)/g(n) = \infty$.
- $f(n) \sim g(n)$ (asymptotic equivalence) if $\lim_{n \rightarrow \infty} f(n)/g(n) = 1$.

2.2 Mercer Decomposition and Eigenfunctions

Under mild regularity conditions (e.g., $\mathcal{X}$ compact and K continuous), Mercer's theorem guarantees the existence of an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ of $L_2(\mathcal{X}, \mu)$ and a sequence of non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ such that

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x'), \quad x, x' \in \mathcal{X}, \quad (2.2)$$

with uniform absolute convergence. The eigenfunctions satisfy

$$\int_{\mathcal{X}} K(x, x') \phi_i(x') d\mu(x') = \lambda_i \phi_i(x), \quad \int_{\mathcal{X}} \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij}. \quad (2.3)$$

Because $\{\phi_i\}$ is a complete orthonormal system, any function $f \in L_2(\mathcal{X}, \mu)$ can be expanded as

$$f(x) = \sum_{i=1}^{\infty} v_i \phi_i(x), \quad v_i = \int_{\mathcal{X}} f(x) \phi_i(x) d\mu(x), \quad (2.4)$$

and its L_2 norm is $\|f\|_{L_2}^2 = \sum_{i=1}^{\infty} v_i^2$. If additionally $f \in \mathcal{H}_K$, then its RKHS norm is $\|f\|_{\mathcal{H}_K}^2 = \sum_{i=1}^{\infty} v_i^2 / \lambda_i$.

2.3 Kernel Ridge Regression (KRR)

Given a regularization parameter $\delta \geq 0$ (the ridge parameter), the KRR estimator is the minimizer of the regularized empirical risk:

$$\hat{f} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \delta \|f\|_{\mathcal{H}_K}^2 \right\}. \quad (2.5)$$

By the representer theorem, the solution admits the closed form

$$\hat{f}(x) = \mathbf{k}_{x\mathcal{D}} (\mathbf{K}_{\mathcal{D}\mathcal{D}} + \delta I_n)^{-1} \mathbf{y}, \quad (2.6)$$

where $\mathbf{k}_{x\mathcal{D}} = [K(x, x_1), \dots, K(x, x_n)]$ is a row vector, $\mathbf{K}_{\mathcal{D}\mathcal{D}}$ is the $n \times n$ kernel matrix with entries $K(x_i, x_j)$, and $\mathbf{y} = (y_1, \dots, y_n)^\top$. The expected test mean squared error (MSE) is defined as

$$\mathcal{E}(f) = \mathbb{E}_{\mathcal{D}, \eta, x \sim \mu} [(\hat{f}(x) - f(x))^2], \quad (2.7)$$

where the expectation is taken over the random training set (including noise) and a fresh test point x independent of $\mathcal{D}$. The irreducible noise variance is ϵ^2 , so the Bayes-optimal risk equals ϵ^2 .

2.4 The Eigenlearning Framework [12]

Working in the eigenbasis greatly simplifies the work. For each eigenmode i , denote its coefficient in the target function by $v_i = \langle f, \phi_i \rangle_{L_2}$. Under the *Gaussian design (universality) assumption* (see Section 2.5), one can derive closed-form expressions for the expected test MSE in terms of the eigenvalues λ_i and the target coefficients v_i .

The central quantity is the **effective regularization** κ , defined implicitly as the unique positive solution of

$$n = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \kappa} + \frac{\delta}{\kappa}. \quad (2.8)$$

Intuitively, κ acts as a threshold that separates eigenmodes into two groups: modes with $\lambda_i \gg \kappa$ are well learned, while modes with $\lambda_i \ll \kappa$ are poorly learned. This behaviour is captured by the **learnability** of the i -th mode:

$$\mathcal{L}_i = \frac{\lambda_i}{\lambda_i + \kappa}. \quad (2.9)$$

Indeed, $\mathcal{L}_i \approx 1$ when $\lambda_i \gg \kappa$ (the mode is nearly fully learned) and $\mathcal{L}_i \approx 0$ when $\lambda_i \ll \kappa$ (the mode is not learned at all). The learnabilities also satisfy a **conservation law**: $\sum_{i=1}^{\infty} \mathcal{L}_i = n$ when $\delta = 0$ [12], which reflects the fact that the total “amount of learning” is bounded by the number of training examples.

The **noise overfitting coefficient** is defined as

$$\mathcal{E}_0 = \frac{n}{n - \sum_{i=1}^{\infty} \mathcal{L}_i^2}. \quad (2.10)$$

This quantity measures how much the algorithm overfits pure noise. Finally, the expected test MSE for a target function with coefficients v_i is

$$\mathcal{E}(f) = \mathcal{E}_0 \left(\sum_{i=1}^{\infty} (1 - \mathcal{L}_i)^2 v_i^2 + \epsilon^2 \right). \quad (2.11)$$

In the **pure noise case** where $f \equiv 0$ (hence all $v_i = 0$) which we will consider, this simplifies to

$$\mathcal{E}(0) = \mathcal{E}_0 \epsilon^2. \quad (2.12)$$

Thus, the asymptotic behavior of $\mathcal{E}_0$ directly determines the overfitting regime: benign if $\mathcal{E}_0 \rightarrow 1$, tempered if $\mathcal{E}_0 \rightarrow \text{constant} > 1$, and catastrophic if $\mathcal{E}_0 \rightarrow \infty$.

2.5 On the Gaussian Design (Universality) Assumption

The closed-form expressions (2.8)–(2.11) are derived under the assumption that the random variables $\phi_i(x_j)$ (the eigenfunctions evaluated at the training points) are independent and identically distributed as standard Gaussian. This is known as the *Gaussian design* or *universality* assumption. While this assumption is not satisfied for most real kernels, it has been shown empirically and theoretically that the resulting formulas accurately predict the generalization performance of KRR for a wide range of kernels, especially in high dimensions and in the limit of large n . The assumption is standard in the statistical physics approach to KRR [3, 4, 12]. Recent work [2, 5] has extended these results to more realistic sub-Gaussian designs, and we will discuss these extensions in Section 7.

In the present paper, we adopt the Gaussian design as a working assumption to obtain sharp asymptotic results. The derived asymptotics are expected to hold for a broader class of kernels that satisfy the same spectral decay conditions.

2.6 Kolmogorov n -Widths and Effective Dimension d_K

To characterize the intrinsic geometric complexity of the domain $\mathcal{X}$ as seen through the kernel K , [14] introduced the notion of *effective dimension* d_K using Kolmogorov n -widths. Recall that the Kolmogorov n -width of a set S in a normed space $\mathcal{H}$ is defined as

$$w_n(S; \mathcal{H}) = \inf_{\substack{L_n \subset \mathcal{H} \\ \dim L_n = n}} \sup_{s \in S} \inf_{h \in L_n} \|s - h\|_{\mathcal{H}}. \quad (2.13)$$

In our context, consider the embedding of $\mathcal{X}$ into the RKHS $\mathcal{H}_K$ given by the map $x \mapsto K(x, \cdot)$. Let $S = \{K(x, \cdot) : x \in \mathcal{X}\} \subset \mathcal{H}_K$. Then the Kolmogorov n -width of S is denoted by $w_K(n) = w_n(S; \mathcal{H}_K)$. It measures how well the set of kernel sections can be approximated by an n -dimensional subspace.

The **effective dimension** is then defined as

$$d_K = \limsup_{n \rightarrow \infty} \frac{\log n}{\log(1/w_K(n))}. \quad (2.14)$$

This parameter captures the rate at which the n -widths decay. For many common kernels on regular domains (e.g., the Laplace kernel on the sphere $\mathbb{S}^{d-1}$), one has $w_K(n) \asymp n^{-1/d_K}$.

Exact Spectral Asymptotics for NTK and Laplace kernel

For certain concrete kernels on regular domains, the asymptotic behaviour of the eigenvalues is known from classical results in harmonic analysis and approximation theory, independently of the concept of effective dimension. In particular, for the **Laplace kernel** $K(x, y) = e^{-\|x-y\|}$ on the sphere $\mathbb{S}^{d-1}$ with the uniform distribution, the eigenvalues satisfy

$$\lambda_i \asymp i^{-(1+\frac{1}{d-1})}.$$

Remarkably, the **ReLU Neural Tangent Kernel (NTK)** on the same domain shares the same eigenfunctions (spherical harmonics) and exhibits the same asymptotic decay of eigenvalues [7]. Consequently, for these kernels we have $\alpha = 1 + 1/(d-1)$.

According to the definition of effective dimension d_K (through Kolmogorov n -widths), for such kernels one finds $d_K = 2(d-1)$; see Table 1 in [14]. A direct calculation then yields $\alpha = 1 + 2/d_K$. This demonstrates how the effective dimension emerges naturally from the known spectral asymptotics of these kernels.

Importantly, this relation is not universal; it holds for kernels with polynomial spectral decay on regular domains.

Majorization via Kolmogorov n -Widths

A key result from [14] (Theorem 3) provides a general upper bound connecting Kolmogorov n -widths and eigenvalues. For any Borel probability measure μ on $\mathcal{X}$,

$$\lambda_{2n}(O_{K,\mu}) \leq \frac{w_K(n)^2}{n}. \quad (2.15)$$

Combining (2.15) with the known asymptotic behaviour $w_K(n) \asymp n^{-1/d_K}$ for the Laplace kernel, ReLU NTK we obtain an upper bound for their eigenvalues:

$$\lambda_n = O\left(n^{-(1+2/d_K)}\right). \quad (2.16)$$

Two Complementary Settings

In this paper, we will focus on two complementary settings:

- **Exact power-law decay** ($\lambda_i = Ci^{-\alpha}$ with $\alpha = 1 + 2/d_K$), which applies to the Laplace kernel, ReLU NTK on the sphere, and other kernels with polynomial spectral decay. For this setting, we obtain sharp asymptotics for κ and $\mathcal{E}_0$ (Section 4).
- **Power-law majorization** ($\lambda_i = O(i^{-\alpha})$), which is a weaker condition satisfied by a broader class of kernels (including those with faster, e.g., exponential, decay). For this setting, we establish universal upper bounds on $\mathcal{E}_0$ (Section 6).

The effective dimension d_K will then appear naturally in the asymptotic expressions for $\mathcal{E}_0$ in both settings.

3 Overfitting Regimes: Benign, Tempered, Catastrophic

Following the taxonomy introduced by [11], we classify the asymptotic generalization behavior of KRR in the pure noise setting ($f \equiv 0$) according to the limit of the noise overfitting coefficient $\mathcal{E}_0$ defined in (2.10). Recall from (2.12) that for pure noise the expected test MSE is $\mathcal{E}(0) = \mathcal{E}_0\epsilon^2$. The three regimes are defined as follows.

- **Benign overfitting** occurs when the excess risk tends to zero, i.e.,

$$\lim_{n \rightarrow \infty} \mathcal{E}_0 = 1. \quad (3.1)$$

In this case the predictor asymptotically achieves the Bayes-optimal risk ϵ^2 . From the definition $\mathcal{E}_0 = n/(n - \sum_i \mathcal{L}_i^2)$, benign overfitting is equivalent to

$$\sum_{i=1}^{\infty} \mathcal{L}_i^2 = o(n) \quad (n \rightarrow \infty),$$

i.e., the total squared learnability grows slower than linearly in n .

- **Tempered overfitting** occurs when the excess risk converges to a finite positive constant greater than the Bayes risk:

$$\lim_{n \rightarrow \infty} \mathcal{E}_0 = C \quad \text{with} \quad 1 < C < \infty. \quad (3.2)$$

This regime is equivalent to

$$n - \sum_{i=1}^{\infty} \mathcal{L}_i^2 = \Theta(n), \quad \text{i.e.,} \quad \sum_{i=1}^{\infty} \mathcal{L}_i^2 = n - \Theta(n).$$

It is characteristic of kernels with polynomial spectral decay (e.g., Laplace kernel, ReLU NTK) when the ridge parameter decays sufficiently slowly or is absent.

- **Catastrophic overfitting** occurs when the excess risk diverges to infinity:

$$\lim_{n \rightarrow \infty} \mathcal{E}_0 = \infty. \quad (3.3)$$

This corresponds to

$$n - \sum_{i=1}^{\infty} \mathcal{L}_i^2 = o(n) \quad \Longleftrightarrow \quad \sum_{i=1}^{\infty} \mathcal{L}_i^2 = n - o(n).$$

Catastrophic overfitting occurs, for example, for the Gaussian kernel in the ridgeless limit ($\delta = 0$).

4 Asymptotic Analysis for Power-Law Spectra

The effective regularization $\kappa = \kappa(n)$ is defined implicitly by equation (2.8). Our goal is to find the leading-order asymptotics of κ and of the noise overfitting coefficient $\mathcal{E}_0$ (given by (2.10)) as $n \rightarrow \infty$. The analysis reveals three distinct regimes depending on the relation between a and $\alpha - 1$.

Asymptotics of κ

Theorem 4.1 (Asymptotics of κ). Let $\lambda_i = Ci^{-\alpha}$ with $\alpha > 1$ and let $\delta = n^{-a}$, $a > 0$. Then as $n \rightarrow \infty$,

$$\kappa \sim \begin{cases} CA^\alpha n^{-\alpha}, & a > \alpha - 1, \\ n^{-a-1}, & a < \alpha - 1, \\ Cc_* n^{-\alpha}, & a = \alpha - 1, \end{cases}$$

where $A = \frac{\pi}{\alpha \sin(\pi/\alpha)}$ and c_* is the unique positive solution of $AC^{1/\alpha}c_*^{-1/\alpha} + 1/c_* = 1$.

Proof idea. The detailed proof of this statement is given in appendix B, here we outline some of the steps. With the substitution $\mu = \kappa/C$ equation (2.8) becomes

$$n = \sum_{i=1}^{\infty} \frac{1}{1 + \mu i^\alpha} + \frac{n^{-a}}{C\mu}. \quad (4.1)$$

Using the integral approximation (see appendix A) we obtain

$$n = A\mu^{-1/\alpha} + O(1) + \frac{n^{-a}}{C\mu}. \quad (4.2)$$

The asymptotic behavior of μ (and hence κ) depends on the competition between the two terms on the right-hand side.

Weak regularization ($a > \alpha - 1$). In this regime the ridge term $\delta/\kappa = n^{-a}/(C\mu)$ is negligible compared to n . One can show that $\mu = \Theta(n^{-\alpha})$, and hence $\mu^{-1/\alpha} = \Theta(n)$ so that the ridge term is $O(n^{\alpha-a}) = o(n)$ because $\alpha - a < 1$. Equation 4.2 then reduces to $n \sim A\mu^{-1/\alpha}$, yielding

$$\mu \sim A^\alpha n^{-\alpha} \implies \kappa \sim CA^\alpha n^{-\alpha}. \quad (4.3)$$

Strong regularization ($a < \alpha - 1$). Here the ridge term dominates. We anticipate $\mu \sim n^{-a-1}/C$. With this scaling, $\mu^{-1/\alpha} = O(n^{(a+1)/\alpha}) = o(n)$ because $a + 1 < \alpha$. Thus the sum term is negligible, and 4.2 gives

$$n \sim \frac{n^{-a}}{C\mu} \implies \mu \sim \frac{n^{-a-1}}{C} \implies \kappa \sim n^{-a-1}. \quad (4.4)$$

Critical regularization ($a = \alpha - 1$). Both terms contribute at leading order. Set $\mu = cn^{-\alpha}$ with a constant $c > 0$ to be determined. Then $\mu^{-1/\alpha} = c^{-1/\alpha}n$ and the ridge term becomes $n/(cC)$. Substituting into 4.2 and dividing by n yields

$$1 = Ac^{-1/\alpha} + \frac{1}{cC}. \quad (4.5)$$

This equation has a unique positive solution c_* . Consequently

$$\kappa \sim Cc_* n^{-\alpha}. \quad (4.6)$$

Asymptotics of $\mathcal{E}_0$ for pure noise

Theorem 4.2 (Asymptotics of $\mathcal{E}_0$ for pure noise). Under the same assumptions as in Theorem 4.1, the noise overfitting coefficient satisfies

$$\mathcal{E}_0 \sim \begin{cases} \alpha, & a > \alpha - 1, \\ 1, & a < \alpha - 1, \\ \frac{1}{1 - \frac{\alpha - 1}{\alpha} A c_*^{-1/\alpha}}, & a = \alpha - 1. \end{cases}$$

Proof idea. The detailed proof of this statement is provided in Appendix C, but one obtains this result easily by establishing the 4.1 first and then using asymptotics of κ in each of the three cases to get the corresponding result for $\mathcal{E}_0$.

Discussion

The results show that for power-law spectra:

- **Strong regularization** ($a < \alpha - 1$) leads to **benign overfitting** ($\mathcal{E}_0 \rightarrow 1$).
- **Weak regularization** ($a > \alpha - 1$) leads to **tempered overfitting** ($\mathcal{E}_0 \rightarrow \alpha$).
- **Critical regularization** ($a = \alpha - 1$) also yields tempered overfitting (a finite constant > 1).

The boundary between benign and tempered is precisely $a = \alpha - 1$. When the ridge decays faster than this threshold, the effective regularization κ is dominated by the sum term, and the asymptotic overfitting factor is α , which exceeds 1. When the ridge decays slower, the ridge term dominates and the overfitting is benign. We note that no catastrophic overfitting occurs. In the next section we interpret these results in terms of the effective dimension d_K .

5 Interpretation via Effective Dimension d_K

In Section 4 we derived the asymptotic behaviour of the noise overfitting coefficient $\mathcal{E}_0$ for kernels with exact power-law eigenvalues $\lambda_i = C i^{-\alpha}$. To give a geometric interpretation of these results, we now relate the exponent α to the effective dimension d_K introduced in Section 2.6.

The value of d_K is fixed for a given kernel and domain, but comparing different domains or kernels reveals its role:

- For high-dimensional spheres, as the dimension d grows, $d_K = 2(d - 1)$ tends to infinity, so $\mathcal{E}_0 \rightarrow 1$: overfitting becomes nearly **benign**.
- For low dimensions, d_K is small, giving $\mathcal{E}_0 = 1 + 2/d_K > 1$: **tempered overfitting** with a factor that decreases as d_K increases.
- We cannot give any interpretation for the kernels with exponential spectral decay. For example, for Gaussian kernel we have $d_K = 0$, and $\alpha = 1 + \frac{2}{d_K}$ doesn't apply. It's known that in the ridgeless limit, such kernels exhibit **catastrophic overfitting**; however, we will show that regularisation can restore benign behaviour (see Section 6).

The weakly regularized regime requires $a > \alpha - 1 = 2/d_K$. If $a < 2/d_K$ (stronger regularisation), the ridge term dominates and $\mathcal{E}_0 \rightarrow 1$ independently of d_K . Hence, increasing regularisation can always suppress overfitting, even for kernels with small effective dimension.

6 Upper Bounds for Spectra Majorized by a Power Law

In this section we relax the assumption that the eigenvalues follow an exact power law. Instead, we assume that they are **majorized** by a power law, i.e., $\lambda_i = O(i^{-\alpha})$. We motivated the analysis of this case in 2.6. Our goal is to obtain **upper bounds** for the noise overfitting coefficient $\mathcal{E}_0$ (and hence for the test MSE $\mathcal{E}(0) = \mathcal{E}_0 \epsilon^2$ in the pure noise setting) under this majorization, using the same decaying ridge $\delta = n^{-a}$ with $a > 0$.

The key idea is to derive a universal upper bound for κ using the inequality $\lambda_i \leq C i^{-\alpha}$. Then we combine it with the simple bound $\mathcal{E}_0 \leq n\kappa/\delta$, which follows from $\sum_i \mathcal{L}_i^2 \geq 0$ and the definition of $\mathcal{E}_0$:

$$\mathcal{E}_0 = \frac{n}{n - \sum_i \mathcal{L}_i^2} \leq \frac{n}{n - \sum_i \mathcal{L}_i} = \frac{n}{\delta/\kappa} = \frac{n\kappa}{\delta} = n^{a+1}\kappa. \quad (6.1)$$

Here we used that $\sum_i \mathcal{L}_i = n - \delta/\kappa$ from (2.8).

Lemma 6.1 (Upper bound for κ under majorized spectrum). Let $\lambda_i \leq C i^{-\alpha}$ with $\alpha > 1$ and let $\delta = n^{-a}$, $a > 0$. Then as $n \rightarrow \infty$,

$$\kappa \leq \begin{cases} O(n^{-a-1}), & a < \alpha - 1, \\ O(n^{-\alpha}), & a \geq \alpha - 1. \end{cases}$$

Proof idea. The proof proceeds by contradiction. For $a < \alpha - 1$, assuming κ is asymptotically larger than n^{-a-1} leads to a violation of equation (2.8) because the ridge term becomes too small. For $a \geq \alpha - 1$, assuming κ grows faster than any constant times $n^{-\alpha}$ makes the sum term $A\mu^{-1/\alpha}$ negligible, while the ridge term also becomes negligible, contradicting the fact that their sum must equal n . The detailed rigorous argument is given in Appendix D.

Using Lemma 6.1 together with 6.1 yields the following result.

Theorem 6.2 (Upper bounds for $\mathcal{E}_0$ under majorized spectrum). Assume $\lambda_i \leq C i^{-\alpha}$ with $\alpha > 1$, and let $\delta = n^{-a}$ with $a > 0$. Then for the pure noise case, as $n \rightarrow \infty$,

$$\mathcal{E}_0 \leq \begin{cases} O(1), & a < \alpha - 1, \\ O(n^{a+1-\alpha}), & a \geq \alpha - 1. \end{cases}$$

Consequently, the expected test MSE satisfies $\mathcal{E}(0) = \mathcal{E}_0 \epsilon^2 = O(1)\epsilon^2$ when $a < \alpha - 1$, and $\mathcal{E}(0) = O(n^{a+1-\alpha})\epsilon^2$ when $a \geq \alpha - 1$.

Proof idea. From Lemma 6.1, $\kappa = O(n^{-a-1})$ for $a < \alpha - 1$ and $\kappa = O(n^{-\alpha})$ for $a \geq \alpha - 1$. Substituting into $\mathcal{E}_0 \leq n^{a+1}\kappa$ gives the stated bounds. This is done more explicitly in appendix E.

Interpretation

1. **Guarantee against catastrophic overfitting.** For any kernel whose spectrum decays no faster than $i^{-\alpha}$ (i.e. it satisfies the majorisation), the condition $a < \alpha - 1$ ensures that the noise overfitting coefficient $\mathcal{E}_0$ remains bounded (does not grow with n). The theorem provides a *sufficient condition for stability*: the test risk does not diverge.
2. **Universality of the threshold.** The theorem identifies the critical point $a = \alpha - 1$ as a theoretical threshold:
 - For $a < \alpha - 1$ (strong regularization), the model is controlled by the user via the decay rate of δ .

- For $a \geq \alpha - 1$ (weak or critical regularization), the kernel’s spectral tail takes over.

3. **Asymptotic ceiling in the weak regularization regime.** When $a \geq \alpha - 1$, we obtain the upper bound $\kappa = O(n^{-\alpha})$. Hence, even if the external ridge vanishes very quickly, the effective regularization cannot decay slower than $n^{-\alpha}$. The intrinsic smoothness of the kernel (characterized by α) imposes a fundamental limit on how “free” the model can become.

Limitations of the result

These points are crucial for a correct understanding of the majorisation-based approach.

1. **Absence of lower bounds (no lower bounds).** The theorem is proved in terms of O -notation. Thus we know that κ or $\mathcal{E}_0$ does *not* decay slower than the stated rates, but we have *not* proved that it cannot decay faster.
2. **Indistinguishability of infinitely smooth kernels.** The theorem distinguishes between kernels with different polynomial exponents α , but it cannot tell the difference between the Gaussian kernel or any other exponentially decaying spectrum. All such kernels are in the strong regularisation regime (Case 1) for any finite a (for Gaussian, for example, the quantity $d_K = 0$, which means that we can formally perceive $\alpha - 1 = \frac{2}{d_K}$ as infinitely large quantity bigger than any finite a). *Consequence:* The theorem with its current assumptions (specifically the type of ridge decay chosen to be power-law) cannot be used to predict the onset of catastrophic overfitting for the Gaussian kernel.
3. **No direct statement about the total test risk.** The theorem only addresses κ and $\mathcal{E}_0$ (the noise amplification factor). It does not, by itself, determine whether the **total** test MSE tends to zero or to a constant. This would require additional assumptions on the target function.

7 Open Problems and Future Directions

While our analysis has provided a detailed characterization of asymptotic overfitting in KRR for power-law spectra and for spectra majorized by a power law, several important questions remain open. We outline the most promising directions for future research.

7.1 Exponentially Decaying Ridge Parameter

In Section 6 we considered ridge parameters decaying as $\delta = n^{-a}$ with a fixed exponent $a > 0$. A setting of faster decay could be addressed (e.g. $\delta = e^{-na}$ and other exponential regimes). This regime is of practical interest because early stopping in gradient descent corresponds to an effective ridge that decays inversely with training time, which can be exponential in the number of epochs under certain conditions. However, the analysis becomes substantially more delicate because the ridge term δ/κ in equation (2.8) is now exponentially small, and the competition between the sum over eigenvalues and the ridge term depends critically on the precise rate of spectral decay.

For exponentially decaying eigenvalues (e.g., Gaussian kernel), the sum $\sum \lambda_i/(\lambda_i + \kappa)$ itself decays very quickly with κ , and the behaviour of κ is governed by the balance between the two exponentially small quantities. Preliminary observations suggest that:

- If the eigenvalues decay as $e^{-\beta i}$ and $\delta = e^{-na}$ with $a > \beta$, the ridge term is negligible compared to the smallest eigenvalues that contribute to the sum, leading to catastrophic overfitting ($\mathcal{E}_0 \rightarrow \infty$).

- If $a < \beta$, the ridge term dominates and may force κ to be of order δ/n , potentially yielding bounded risk.

7.2 Non-Zero Target Functions (Source Conditions)

Our main results focus on the pure noise case $f \equiv 0$. For a non-zero target function with eigencoefficients v_i , the test MSE formula (2.11) includes an additional term $\mathcal{E}_0 \sum_i (1 - \mathcal{L}_i)^2 v_i^2$. To fully characterise the excess risk, one needs to understand the asymptotic behaviour of this bias term. Under a *source condition* of the form $|v_i| \leq Ci^{-r}$ with $r > 0$, one can bound $\sum (1 - \mathcal{L}_i)^2 v_i^2$ using the asymptotics of $\mathcal{L}_i$. For power-law spectra, one would obtain a trade-off: faster spectral decay (larger α) makes the bias term smaller (since $1 - \mathcal{L}_i$ decays faster), but increases $\mathcal{E}_0$. The optimal regularisation exponent a might be chosen to balance bias and variance, leading to a classical bias-variance trade-off. Extending our analysis to include source conditions is a natural next step.

7.3 Lower Bounds for Majorized Spectra

In Section 6 we derived upper bounds for $\mathcal{E}_0$ under the assumption $\lambda_i \leq Ci^{-\alpha}$. It would be desirable to obtain matching lower bounds (or at least to determine the sharp asymptotic order) for specific classes of spectra that are majorized by a power law. For example, if the eigenvalues decay exactly as $Ci^{-\alpha}$, we already know that $\mathcal{E}_0$ converges to a constant (either 1 or α), which is much smaller than the polynomial upper bound $O(n^{a+1-\alpha})$ when $a \geq \alpha - 1$. The question is: for which faster-than-power-law spectra does $\mathcal{E}_0$ actually diverge, and at what rate?

7.4 Relaxation of the Gaussian Design Assumption

Our entire analysis, following [12], relies on the Gaussian design (universality) assumption. While this assumption is known to give accurate predictions for many kernels, it is not universally valid. Recent work by [2] and [5] has extended similar results to sub-Gaussian designs, but these extensions often come with additional technical conditions. A systematic study of how the overfitting regimes change when moving from Gaussian to sub-Gaussian is an interesting direction for future research.

7.5 Empirical Verification and Real Data

Although this paper is purely theoretical, the results naturally invite empirical validation. Experiments on synthetic data (e.g., power-law kernels on spheres) could confirm the predicted asymptotics of κ and $\mathcal{E}_0$. Testing the effective dimension d_K as a predictor of overfitting behaviour on real datasets (e.g., using empirical NTKs of neural networks) would be valuable. The algorithm proposed by [14] for estimating n -widths from finite samples could be used to compute empirical estimates of d_K and compare with the observed test error curves.

7.6 Connection to Neural Network Training Beyond the Kernel Regime

Finally, while the NTK equivalence holds in the infinite-width limit, real neural networks operate in a *feature learning* regime where the kernel changes during training. Understanding whether the overfitting taxonomy (benign/tempered/catastrophic) extends to this more realistic setting, and how it relates to notions like the neural network’s intrinsic dimension, is a challenging open problem. Our work provides a baseline for the kernel regime; future work could explore how feature learning modifies the effective dimension and the overfitting behaviour.

8 Conclusion

In this work we have carried out a detailed asymptotic analysis of kernel ridge regression under power-law eigenvalue spectra and a decaying ridge parameter $\delta = n^{-a}$. Restricting to the pure-noise setting ($f^* \equiv 0$) and employing the eigenlearning framework [12] (which assumes Gaussian design), we derived sharp asymptotics for the effective regularisation κ and the noise overfitting coefficient $\mathcal{E}_0$ for exact power-law spectra $\lambda_i = i^{-\alpha}$ ($\alpha > 1$). Our results (Section 4) reveal three regimes depending on the relation between the ridge exponent a and the spectral exponent α :

- **Strong regularisation** ($a < \alpha - 1$): $\mathcal{E}_0 \rightarrow 1$ (benign overfitting).
- **Weak regularisation** ($a > \alpha - 1$): $\mathcal{E}_0 \rightarrow \alpha$ (tempered overfitting).
- **Critical regularisation** ($a = \alpha - 1$): $\mathcal{E}_0 \rightarrow c_* > 1$ (still tempered).

These results are fully consistent with the overfitting taxonomy of [11] and the exact asymptotics of [6], and they refine the picture by characterising the effect of a decaying ridge.

For certain kernels on regular domains (e.g., the Laplace kernel or the ReLU neural tangent kernel on the sphere), the eigenvalues are known to decay exactly as a power law, $\lambda_i \asymp i^{-(1+2/d_K)}$, where d_K is the effective dimension introduced by [14]. This motivated our analysis of exact power-law spectra in Section 4.

For kernels whose spectrum is only known to be *majorised* by a power law ($\lambda_i \leq Ci^{-\alpha}$), Section 6 establishes upper bounds on $\mathcal{E}_0$. In the strong regularization regime ($a < \alpha - 1$) we obtain $\mathcal{E}_0 = O(1)$; this is a sufficient condition for stability - the test risk remains bounded. In the weak regularization regime the bound gives $\mathcal{E}_0 = O(n^{a+1-\alpha})$, which is not sharp. The majorisation approach is therefore conservative and its main value lies in the stability guarantee under strong regularization.

Several limitations and open problems remain. Our analysis is restricted to pure noise; for non-zero target functions the bias term would appear and the asymptotic risk may differ (a source condition would be required). The Gaussian design assumption, while standard, may not hold for all real kernels; relaxing it to more realistic settings (as in [?]) is an important direction. Deriving lower bounds for majorised spectra, analysing exponentially decaying ridge parameters, and empirically testing the predicted rates on real data are natural extensions of this work.

References

- [1] Peter L. Bartlett, Philip M. Long, Gabor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- [2] Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions. *arXiv preprint arXiv:2312.15995*, 2023.
- [3] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. *International Conference on Machine Learning*, pages 1024–1034, 2020.
- [4] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):1–12, 2021.

- [5] Tin Sum Cheng, Aurelien Lucchi, Ivan Dokmanic, Anastasis Kratsios, and David Belius. A theoretical analysis of the test error of finite-rank kernel ridge regression. *Advances in Neural Information Processing Systems*, 2023.
- [6] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.
- [7] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Ronen Basri. On the similarity between the laplace and neural tangent kernels. In *Advances in Neural Information Processing Systems*, volume 33, pages 1451–1461. Curran Associates, Inc., 2020.
- [8] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- [9] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Kernel alignment risk estimator: risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020.
- [10] Cheng Li, Haobo Zhang, and Qian Lin. On the asymptotic learning curves of kernel ridge regression under power-law decay. *arXiv preprint arXiv:2309.13337*, 2023.
- [11] 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.
- [12] James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael R. DeWeese. The eigenlearning framework: A conservation law perspective on kernel ridge regression and wide neural networks. *arXiv preprint arXiv:2110.03922*, 2021.
- [13] Peter Sollich. Gaussian process regression with mismatched models. *Advances in Neural Information Processing Systems*, pages 519–526, 2001.
- [14] Rustem Takhanov. On the intrinsic dimensions of data in kernel learning. *arXiv preprint arXiv:2401.12345*.
- [15] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A Integral Estimates

We provide rigorous bounds for the sums appearing in Section 4. Let $\mu > 0$ and $\alpha > 1$. Define

$$f(x) = \frac{1}{1 + \mu x^\alpha}, \quad x \geq 0.$$

Since f is positive and strictly decreasing on $(0, \infty)$, for any integer $N \geq 1$ we have

$$\int_N^\infty f(x) dx \leq \sum_{i=N}^\infty f(i) \leq f(N) + \int_N^\infty f(x) dx.$$

Setting $N = 1$ gives

$$\int_1^\infty f(x) dx \leq \sum_{i=1}^\infty f(i) \leq f(1) + \int_1^\infty f(x) dx.$$

Adding $\int_0^1 f(x) dx$ to both sides yields

$$\int_0^\infty f(x) dx \leq \sum_{i=1}^\infty f(i) + \int_0^1 f(x) dx \leq f(1) + \int_0^\infty f(x) dx.$$

Because $0 \leq f(x) \leq 1$ on $[0, 1]$, we have $\int_0^1 f(x) dx \leq 1$. Also $f(1) = 1/(1 + \mu) \leq 1$. Therefore

$$\left| \sum_{i=1}^\infty f(i) - \int_0^\infty f(x) dx \right| \leq 1. \quad (\text{A.1})$$

The integral is evaluated via the substitution $t = \mu^{1/\alpha}x$:

$$\int_0^\infty \frac{dx}{1 + \mu x^\alpha} = \mu^{-1/\alpha} \int_0^\infty \frac{dt}{1 + t^\alpha}.$$

This integral is known,

$$A = \int_0^\infty \frac{dt}{1 + t^\alpha} = \frac{\pi}{\alpha \sin(\pi/\alpha)}.$$

Hence

$$\int_0^\infty \frac{dx}{1 + \mu x^\alpha} = A \mu^{-1/\alpha}. \quad (\text{A.2})$$

For the sum of squares, consider $g(x) = 1/(1 + \mu x^\alpha)^2$. The same reasoning yields

$$\left| \sum_{i=1}^\infty g(i) - \int_0^\infty g(x) dx \right| \leq 1. \quad (\text{A.3})$$

The integral is

$$\int_0^\infty \frac{dx}{(1 + \mu x^\alpha)^2} = \mu^{-1/\alpha} \int_0^\infty \frac{dt}{(1 + t^\alpha)^2},$$

which is also known to be equal to

$$\int_0^\infty \frac{dt}{(1 + t^\alpha)^2} = \frac{\alpha - 1}{\alpha} \int_0^\infty \frac{dt}{1 + t^\alpha} = \frac{\alpha - 1}{\alpha} A.$$

Thus

$$\int_0^\infty \frac{dx}{(1 + \mu x^\alpha)^2} = J(\alpha) \mu^{-1/\alpha}, \quad J(\alpha) = \frac{\alpha - 1}{\alpha} A. \quad (\text{A.4})$$

These estimates are used throughout Section 4.

B Derivation of the Asymptotics of κ

We now prove Theorem 4.1 by treating the three regimes separately and following the key steps outlined after the theorem.

Recall $\lambda_i = Ci^{-\alpha}$, $\delta = n^{-a}$. Equation (2.8) reads

$$n = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \kappa} + \frac{\delta}{\kappa}. \quad (\text{B.1})$$

Set $\mu = \kappa/C$. Then $\lambda_i/(\lambda_i + \kappa) = 1/(1 + \mu i^\alpha)$ and $\delta/\kappa = n^{-a}/(C\mu)$. Hence

$$n = \sum_{i=1}^{\infty} \frac{1}{1 + \mu i^\alpha} + \frac{n^{-a}}{C\mu}. \quad (\text{B.2})$$

Denote the sum term $\sum_{i=1}^{\infty} \frac{1}{1 + \mu i^\alpha} = S(\mu)$. For an upper bound of this term we use the standard integral estimate

$$S(\mu) \leq \int_0^{\infty} \frac{dx}{1 + \mu x^\alpha} = A\mu^{-1/\alpha}, \quad A = \frac{\pi}{\alpha \sin(\pi/\alpha)}. \quad (\text{B.3})$$

To obtain the lower bound, choose an integer $N = \lfloor (C/\kappa)^{1/\alpha} \rfloor$. For $i \leq N$ we have $Ci^{-\alpha} \geq \kappa$, hence $\frac{\lambda_i}{\lambda_i + \kappa} \geq \frac{1}{2}$. Therefore

$$S(\mu) \geq \sum_{i=1}^N \frac{1}{2} = \frac{N}{2} \geq \frac{1}{2}((C/\kappa)^{1/\alpha} - 1). \quad (\text{B.4})$$

In particular, for all κ small enough (so that $(C/\kappa)^{1/\alpha} \geq 2$) we have the simple bound

$$S(\mu) \geq \frac{1}{4} C^{1/\alpha} \kappa^{-1/\alpha}. \quad (\text{B.5})$$

(Any positive constant would work, we will only need the existence of such a bound when $\kappa \rightarrow 0$.)

Case $a > \alpha - 1$ (weak regularization)

Step 1: We claim that $\kappa = \Theta(n^{-\alpha})$.

To show that $\kappa = \Omega(n^{-\alpha})$, for the sake of contradiction assume that $\kappa = o(n^{-\alpha})$. Then $\kappa^{-1/\alpha} = \omega(n)$. Using the lower bound B.5, we obtain

$$S(\mu) \geq \frac{1}{4} C^{1/\alpha} \kappa^{-1/\alpha} = \omega(n). \quad (\text{B.6})$$

But from B.1 we have $S(\mu) = n - n^{-a}/(C\mu) \leq n$. This contradiction shows that there exists a constant $c_1 > 0$ such that $\kappa \geq c_1 n^{-\alpha}$ for all large n , which is by definition $\kappa = \Omega(n^{-\alpha})$.

To get an upper bound, we use

$$n = S(\mu) + \frac{n^{-a}}{\kappa} \leq AC^{1/\alpha} \kappa^{-1/\alpha} + \frac{n^{-a}}{\kappa}$$

and note that since $\kappa \geq c_1 n^{-\alpha}$, we have $\frac{n^{-a}}{\kappa} \leq \frac{n^{-a}}{c_1 n^{-\alpha}} = \frac{1}{c_1} n^{\alpha-a}$. Since $a > \alpha - 1$ implies $\alpha - a - 1 < 0$, this term is $o(n)$. Hence

$$n \leq AC^{1/\alpha} \kappa^{-1/\alpha} + o(n) \implies \kappa^{-1/\alpha} \geq \frac{n - o(n)}{AC^{1/\alpha}}. \quad (\text{B.7})$$

Now we isolate κ . Because $\kappa^{-1/\alpha} \geq \frac{n}{AC^{1/\alpha}}(1 - o(1))$, taking both sides to the power $-\alpha$ yields

$$\kappa \leq \left(\frac{AC^{1/\alpha}}{n(1 - o(1))} \right)^\alpha = A^\alpha C n^{-\alpha} (1 - o(1))^{-\alpha}.$$

For any sequence $\varepsilon(n) = o(1)$, we have $(1 - \varepsilon(n))^{-\alpha} = 1 + o(1)$. Indeed, expanding $(1 - \varepsilon)^{-\alpha} = 1 + \alpha\varepsilon + O(\varepsilon^2)$ shows that it differs from 1 by a term that tends to 0. Therefore

$$\kappa \leq A^\alpha C n^{-\alpha} (1 + o(1)). \quad (\text{B.8})$$

Combined with $\kappa = \Omega(n^{-\alpha})$, we obtain $\kappa = \Theta(n^{-\alpha})$.

Step 2: From the previous step we have $\kappa = \Theta(n^{-\alpha})$. Define the sequence $c_n = \kappa n^\alpha$. Then c_n is bounded and bounded away from zero. We then have $\mu = \kappa/C = c_n n^{-\alpha}/C$.

Using the integral approximation A.1 we have $S(\mu) = A\mu^{-1/\alpha} + O(1)$. Now

$$\mu^{-1/\alpha} = \left(\frac{c_n n^{-\alpha}}{C} \right)^{-1/\alpha} = C^{1/\alpha} c_n^{-1/\alpha} n,$$

and also

$$\frac{n^{-a}}{C\mu} = \frac{n^{-a}}{C \cdot (c_n n^{-\alpha}/C)} = \frac{n^{-a}}{c_n n^{-\alpha}} = \frac{1}{c_n} n^{\alpha-a}.$$

Since $a > \alpha - 1$ is equivalent to $\alpha - a < 1$, we have $n^{\alpha-a} = o(n)$. Hence this term is negligible compared to n .

Substituting into B.2, we obtain

$$n = AC^{1/\alpha} c_n^{-1/\alpha} n + O(1) + o(n).$$

Dividing by n and letting $n \rightarrow \infty$, we find that any limit point c of the sequence $\{c_n\}$ must satisfy

$$1 = AC^{1/\alpha} c^{-1/\alpha}.$$

Thus the only possible limit is $c = A^\alpha C$. Since the sequence is bounded, it converges to this value. Consequently,

$$\kappa \sim CA^\alpha n^{-\alpha}. \quad (\text{B.9})$$

Case $a < \alpha - 1$ (strong regularization)

Step 1: From B.2 we have directly $n \geq n^{-a}/\kappa$, hence

$$\kappa \geq n^{-a-1}. \quad (\text{B.10})$$

We will show that actually $\kappa \sim n^{-a-1}$. Assume for the sake of contradiction that for some $\varepsilon > 0$ there exist arbitrarily large n with $\kappa > (1 + \varepsilon)n^{-a-1}$. Then using B.3,

$$S(\mu) \leq A\mu^{-1/\alpha} = AC^{1/\alpha} \kappa^{-1/\alpha} < AC^{1/\alpha} ((1 + \varepsilon)n^{-a-1})^{-1/\alpha} = AC^{1/\alpha} (1 + \varepsilon)^{-1/\alpha} n^{(a+1)/\alpha}. \quad (\text{B.11})$$

Because $a + 1 < \alpha$, the exponent $(a + 1)/\alpha < 1$, so $S(\mu) = o(n)$. Moreover,

$$\frac{n^{-a}}{\kappa} < \frac{n^{-a}}{(1 + \varepsilon)n^{-a-1}} = \frac{n}{1 + \varepsilon}.$$

Inserting these into A.2:

$$n = S(\mu) + \frac{n^{-a}}{\kappa} < o(n) + \frac{n}{1 + \varepsilon}.$$

For large n the right-hand side is strictly smaller than n (since $\varepsilon > 0$), a contradiction. Hence for any $\varepsilon > 0$ and all sufficiently large n

$$\kappa \leq (1 + \varepsilon)n^{-a-1}. \quad (\text{B.12})$$

Combining B.10 and B.12, we conclude $\kappa \sim n^{-a-1}$.

Case $a = \alpha - 1$ (critical regularization)

Both terms contribute in this case. From the lower bound B.10 we already have $\kappa \geq n^{-\alpha}$. To get an upper bound, assume $\kappa \geq C_0 n^{-\alpha}$ with a large constant C_0 . Then we have

$$n \leq AC^{1/\alpha} \kappa^{-1/\alpha} + \frac{n^{-(\alpha-1)}}{\kappa} \leq AC^{1/\alpha} C_0^{-1/\alpha} n + \frac{n}{C_0}.$$

Dividing by n gives $1 \leq AC^{1/\alpha} C_0^{-1/\alpha} + 1/C_0$. With sufficiently large C_0 the right-hand side becomes less than 1, which is a contradiction. Therefore κ cannot be too large; we have $\kappa = \Theta(n^{-\alpha})$.

Define the sequence $c_n = \kappa n^\alpha$. Then c_n is bounded and bounded away from zero. Substituting $\kappa = c_n n^{-\alpha}$, we have

$$n = AC^{1/\alpha} c_n^{-1/\alpha} n + O(1) + \frac{n^{-(\alpha-1)}}{c_n n^{-\alpha}} = AC^{1/\alpha} c_n^{-1/\alpha} n + \frac{n}{c_n} + O(1).$$

Dividing by n and letting $n \rightarrow \infty$, we find that any limit point c of the sequence $\{c_n\}$ must satisfy

$$1 = AC^{1/\alpha} c^{-1/\alpha} + \frac{1}{c}. \quad (\text{B.13})$$

The right-hand side is strictly decreasing in c for $c > 0$, tends to $+\infty$ as $c \rightarrow 0^+$ and to 0 as $c \rightarrow \infty$. Hence equation B.13 has a unique positive solution c_* . Consequently, the whole sequence c_n converges to c_* . Therefore

$$\kappa \sim C c_* n^{-\alpha}. \quad (\text{B.14})$$

This completes the proof of Theorem 4.1.

C Derivation of the Asymptotics of $\mathcal{E}_0$

We now prove Theorem 4.2. Using the integral estimates A.3 and A.4 with the same $\mu = \kappa/C$, we have

$$\sum_{i=1}^{\infty} \mathcal{L}_i^2 = J(\alpha) \mu^{-1/\alpha} + O(1).$$

Case $a > \alpha - 1$ (weak regularization)

From Theorem 4.1, $\mu \sim A^\alpha n^{-\alpha}$, so $\mu^{-1/\alpha} \sim A^{-1} n$. Then

$$\sum \mathcal{L}_i^2 = J(\alpha) A^{-1} n + O(1) = \frac{\alpha - 1}{\alpha} n + O(1).$$

Hence

$$n - \sum \mathcal{L}_i^2 = n - \frac{\alpha - 1}{\alpha} n + O(1) = \frac{n}{\alpha} + O(1).$$

Therefore

$$\mathcal{E}_0 = \frac{n}{n - \sum \mathcal{L}_i^2} = \frac{n}{n/\alpha + O(1)} \rightarrow \alpha.$$

Case $a < \alpha - 1$ (strong regularization)

Now $\mu \sim n^{-a-1}/C$, so $\mu^{-1/\alpha} \sim C^{1/\alpha} n^{(a+1)/\alpha}$. Since $a + 1 < \alpha$, this is $o(n)$. Thus

$$\sum \mathcal{L}_i^2 = o(n) + O(1) = o(n), \quad n - \sum \mathcal{L}_i^2 = n - o(n) \sim n,$$

and $\mathcal{E}_0 \rightarrow 1$.

Case $a = \alpha - 1$ (critical regularization)

We have $\mu \sim c_* n^{-\alpha}$, so $\mu^{-1/\alpha} \sim c_*^{-1/\alpha} n$. Then

$$\sum \mathcal{L}_i^2 = J(\alpha) c_*^{-1/\alpha} n + O(1) = \gamma n + O(1), \quad \gamma = \frac{\alpha - 1}{\alpha} A c_*^{-1/\alpha}.$$

Hence

$$n - \sum \mathcal{L}_i^2 = n(1 - \gamma) + O(1), \quad \mathcal{E}_0 = \frac{n}{n(1 - \gamma) + O(1)} \rightarrow \frac{1}{1 - \gamma}.$$

One must check that $\gamma < 1$; this follows from the fact that $\sum \mathcal{L}_i^2 < n$ for all finite n , so $\limsup \sum \mathcal{L}_i^2 / n \leq 1$, and equality would imply $\mathcal{E}_0 \rightarrow \infty$ which is not the case because the ridge term δ/κ is of order n and prevents $\sum \mathcal{L}_i^2$ from approaching n . A direct verification using the equation for c_* shows $\gamma < 1$. This completes the proof of Theorem 4.2.

D Proof of Lemma 6.1

We begin again with the expression

$$n \leq A \mu^{-1/\alpha} + 1 + \frac{n^{-a}}{C \mu}$$

and treat the two regimes separately.

Case $a < \alpha - 1$ (strong regularization)

Assume, for contradiction, that there exists $\varepsilon > 0$ and arbitrarily large n such that $\kappa > (1 + \varepsilon)n^{-a-1}$. Then $\mu = \kappa/C > (1 + \varepsilon)n^{-a-1}/C$. Consequently,

$$\mu^{-1/\alpha} < ((1 + \varepsilon)n^{-a-1}/C)^{-1/\alpha} = \left(\frac{C}{1 + \varepsilon}\right)^{1/\alpha} n^{(a+1)/\alpha}.$$

Because $a + 1 < \alpha$ (since $a < \alpha - 1$), the exponent $(a + 1)/\alpha < 1$, so $n^{(a+1)/\alpha} = o(n)$. Hence the first term satisfies $A \mu^{-1/\alpha} = o(n)$.

For the second term,

$$\frac{n^{-a}}{C \mu} < \frac{n^{-a}}{C \cdot (1 + \varepsilon)n^{-a-1}/C} = \frac{n}{1 + \varepsilon}.$$

Thus for sufficiently large n ,

$$n < o(n) + 1 + \frac{n}{1 + \varepsilon}.$$

Since $1/(1 + \varepsilon) < 1$, the right-hand side is eventually less than n (the $o(n)$ and the constant 1 are negligible), a contradiction. Therefore for all large n we must have $\kappa \leq (1 + \varepsilon)n^{-a-1}$. As ε was arbitrary, $\kappa = O(n^{-a-1})$.

Case $a \geq \alpha - 1$ (weak and critical regularization)

We prove that $\kappa = O(n^{-\alpha})$. Suppose, to the contrary, that $\kappa \neq O(n^{-\alpha})$. Then there exists a sequence $n_k \rightarrow \infty$ such that

$$\frac{\kappa(n_k)}{n_k^{-\alpha}} \rightarrow \infty \quad \text{i.e.} \quad \kappa(n_k) = \omega(n_k^{-\alpha}).$$

Set $\mu_k = \kappa(n_k)/C$. Then $\mu_k = \omega(n_k^{-\alpha})$, and therefore

$$\mu_k^{-1/\alpha} = o(n_k). \tag{D.1}$$

Now estimate the ridge term:

$$\frac{n_k^{-a}}{C\mu_k} = \frac{n_k^{-a}}{C \cdot \omega(n_k^{-\alpha})} = \frac{n_k^{\alpha-a}}{\omega(1)}. \quad (\text{D.2})$$

Because $a \geq \alpha - 1$ implies $\alpha - a \leq 1$, we have $n_k^{\alpha-a} = O(n_k)$. Hence the fraction is $o(n_k)$ (it is $O(n_k)$ divided by a quantity tending to infinity).

Plugging D.1 and D.2 yields

$$n_k \leq o(n_k) + 1 + o(n_k) = o(n_k), \quad (\text{D.3})$$

which for large k is impossible. Therefore $\kappa = O(n^{-\alpha})$. This completes the proof of Lemma 6.1.

E Proof of Theorem 6.2

Recall the simple bound 6.1. We now substitute the upper bounds for κ obtained in Lemma 6.1.

Case $a < \alpha - 1$. From Lemma 6.1 we have $\kappa = O(n^{-a-1})$. Hence

$$\mathcal{E}_0 \leq n^{a+1} \cdot O(n^{-a-1}) = O(1).$$

Case $a \geq \alpha - 1$. In this regime Lemma 6.1 gives $\kappa = O(n^{-\alpha})$. Therefore

$$\mathcal{E}_0 \leq n^{a+1} \cdot O(n^{-\alpha}) = O(n^{a+1-\alpha}).$$

Thus we obtain

$$\mathcal{E}_0 \leq \begin{cases} O(1), & a < \alpha - 1, \\ O(n^{a+1-\alpha}), & a \geq \alpha - 1, \end{cases}$$

which is exactly the statement of Theorem 6.2. The corresponding bound for the test MSE follows immediately from $\mathcal{E}(0) = \mathcal{E}_0 \epsilon^2$.