POWER-LAW DATA SPECTRA AS IMPLICIT REGULARIZATION IN LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks often exhibit substantially different learning performance on real-world and synthetic data, suggesting that structural properties of real data may play an important role in learning. Large-scale datasets used in modern neural-network applications, including images and language, commonly exhibit approximately power-law covariance spectra. Such spectral structure may therefore contribute to the learning properties observed on real data. However, how this ubiquitous power-law structure affects learning remains unclear. Here, we show that power-law spectral structure can act as an implicit regularizer of learning. To characterize this effect quantitatively, we analyze double-descent learning curves in ridge regression while systematically varying the power-law exponent. We establish an explicit relation between the power-law exponent and the ridge penalty that reproduces the leading behavior of the double-descent peak in the weak-ridge regime. This relation shows that increasing the power-law exponent suppresses noise amplification in a manner analogous to increasing explicit regularization. Furthermore, we find that the spectral structure affects the dependence of prediction error on sample size beyond the interpolation regime, reshaping the overall learning curve. Establishing the power-law structure of real-world data as an intrinsic source of implicit regularization that shapes learning performance, our findings advance the theoretical understanding of the role of data structure in learning.

1 INTRODUCTION

Understanding how the statistical structure of data and learned representations shapes generalization is an important problem in learning theory. One prominent structural feature is the approximately power-law decay of covariance spectra observed in high-dimensional datasets, including image and language data, and in learned neural representations (Spigler et al., 2020; Canatar et al., 2021; Wei et al., 2022; Li et al., 2025; Liu et al., 2026). This spectral structure provides a direct link between representation geometry and sample-dependent generalization. In particular, the power-law exponent controls how prediction error varies with the number of training samples, as characterized by learning curves (Bordelon et al., 2020; Cui et al., 2021).

Learning curves, however, are not necessarily monotonic. A prominent example is double descent, in which the prediction error increases sharply near the interpolation threshold, where the number of samples matches the number of parameters, before decreasing again (Belkin et al., 2019; Nakkiran et al., 2020). In ridge regression, this peak is driven by the contribution of label noise to the prediction error, which we refer to as the noise risk (Hastie et al., 2022). We call the dependence of the noise risk on the number of samples the noise curve. Thus, the behavior of the noise curve near the interpolation threshold is central to understanding the emergence and suppression of the double-descent peak.

One mechanism that suppresses the noise peak is ridge regularization (Nakkiran et al., 2021). Importantly, ridge regularization is closely related to the covariance spectrum of the input data. Under a rescaling of the input coordinates, spectral anisotropy can be interpreted as direction-dependent ridge regularization (Hoerl & Kennard, 1970; Eldén, 1977; Wu & Xu, 2020). Consistent with this connection, low-variance directions can provide an implicit regularizing effect even in overparameterized interpolation without an explicit penalty (Bartlett et al., 2020; Kobak et al., 2020), while power-law anisotropy has been shown to suppress variance peaks in kernel regression (Rizzi et al., 2026). However, although these studies establish a close connection between spectral anisotropy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

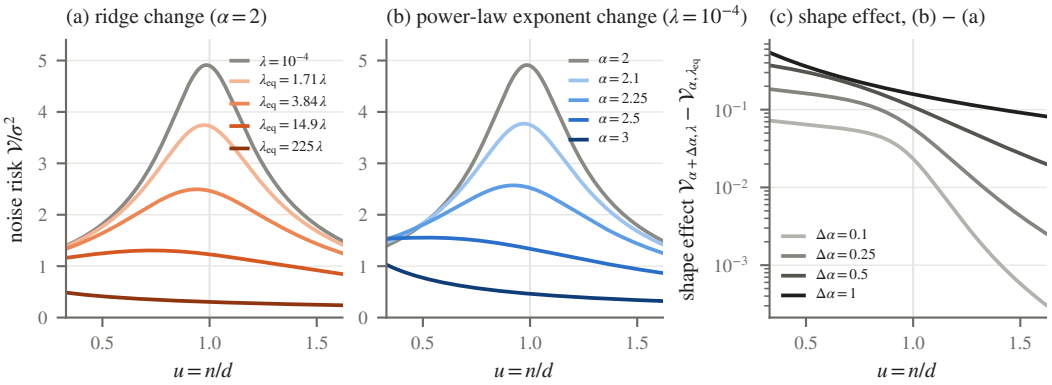

Figure 1: The effect of a power-law exponent change on the noise curve, compared with a ridge change. The covariance eigenvalues are $\mu_k = k^{-\alpha}$ with $d = 300$, and the noise variance is $\sigma^2 = 1$. (a) Noise curves at $\alpha = 2$ for the ridge penalty $\lambda = 10^{-4}$ (gray) and for the larger ridge penalties 1.71λ, 3.84λ, 14.9λ, and 225λ. Each of these is the ridge penalty λ_{eq} of Theorem 1 that gives the same value of λS_1, with $S_1 = \sum_k 1/\mu_k$, as one of the spectra in (b). (b) Noise curves at the fixed ridge penalty $\lambda = 10^{-4}$ for $\alpha = 2$ (gray) and for the larger exponents $\alpha = 2.1$, 2.25, 2.5, and 3, that is, for $\Delta\alpha = 0.1$, 0.25, 0.5, and 1. The peak near $u = 1$ is lowered in the same way in (a) and (b), and this common lowering is the scale effect. (c) The shape effect of Eq. (9), $\mathcal{V}_{\alpha+\Delta\alpha,\lambda} - \mathcal{V}_{\alpha,\lambda_{\text{eq}}}$, the difference between a noise curve in (b) and the corresponding curve in (a), on a logarithmic axis. It is positive at every sample ratio, and near the peak it is much smaller than the scale effect. All curves are computed from the DE in Eqs. (3)–(4).

and regularization, it remains unclear how the power-law exponent of the input spectrum quantitatively controls the height and shape of the noise peak.

To address this question, we study a linear model with a power-law covariance spectrum, $\mu_k = k^{-\alpha}$, using the deterministic equivalent (DE) of the noise risk. The DE provides a deterministic approximation to the noise risk averaged over random inputs and becomes exact when the numbers of samples and parameters grow proportionally. Because it expresses the noise risk explicitly in terms of the covariance eigenvalues, it allows us to determine how varying the power-law exponent modifies the noise peak. We show that this dependence can be decomposed into two distinct contributions, which we call the scale effect and the shape effect. The scale effect arises from the change in the overall spectral scale at fixed shape, whereas the shape effect arises from the change in spectral shape at fixed scale. As shown in Section 4, the scale effect is exactly equivalent to a change in ridge regularization and therefore suppresses the peak, providing an implicit regularization induced by the spectrum. In contrast, the shape effect is small near the peak, where the scale effect dominates, but persists away from the peak. Moreover, in the continuum limit, it cannot be eliminated by any choice of the ridge penalty. Figure 1 illustrates these two effects.

Our main contributions are as follows.

- **An explicit equivalence between the power-law exponent and ridge regularization near interpolation.** Near interpolation and for a small ridge penalty, we derive an explicit leading-order form of the noise curve that depends on the spectrum and the ridge penalty only through a single combination (Section 4, Theorem 1), namely, the product of the ridge penalty and the sum of the inverse eigenvalues. This result establishes an explicit mapping between changes in the power-law exponent of the input spectrum and changes in ridge regularization. Raising the exponent is equivalent, to leading order, to multiplying the ridge penalty by the ratio of this sum after and before the change, a factor that grows exponentially with the change in the exponent.
- **An explicit threshold for the disappearance of the noise peak.** In the continuum limit, where the numbers of samples and parameters grow at a fixed ratio with a suitably rescaled ridge penalty held fixed, and the spectral sums in the DE become integrals, we derive an explicit inequality that determines whether the noise curve has an interior peak (Section 5,

Theorem 2). Beyond this threshold, the noise curve decreases monotonically from its limit at small sample sizes, implying that a sufficiently large exponent eliminates the peak rather than merely suppressing it. In this continuum limit, for exponents at most one, the peak persists for any ridge penalty.

- **Prediction of the noise peak with finite-dimensional and empirical covariance spectra.** For finite Gaussian inputs and principal components of MNIST with a modified power-law exponent, changing the ridge penalty by the theoretically predicted factor reproduces 97–99% of the effect of changing the exponent at interpolation (Section 6). Moreover, when the covariance spectrum of MNIST is adjusted to follow a power law, the peak near interpolation disappears at the theoretically predicted threshold, and the maximum of the finite-dimensional curve moves to small sample ratios (Section 5).

2 Related work

Spectral structure as regularization. Previous work has established several connections between spectral structure and regularization. In overparameterized interpolation, tail eigenvalues can act as an effective ridge (Bartlett et al., 2020; Kobak et al., 2020), while under a surrogate random design, the expected minimum-norm estimator coincides with a population ridge solution (Dereziński et al., 2020, Theorem 2). These results concern either ridgeless estimators or averages over estimators, whereas we study how spectral structure modifies the noise risk curve at positive ridge. Rescaling the input features is exactly equivalent to applying a different ridge penalty along each direction (Hoerl & Kennard, 1970; Eldén, 1977). This form of generalized ridge regularization and its optimal direction-dependent weights were studied by Wu & Xu (2020). In contrast, we ask to what extent the effect of spectral anisotropy can be reproduced by a single scalar ridge penalty.

Spectral risk theory and double descent. Our analysis builds on deterministic equivalents for ridge risk developed in Dobriban & Wager (2018), Hastie et al. (2022), Canatar et al. (2021), Simon et al. (2023), Bach (2024), and Atanasov et al. (2026). Double descent has been studied both as model size varies and as the number of samples increases (Belkin et al., 2019; Nakkiran et al., 2020; Mei & Montanari, 2022), while spectral decay together with target regularity determines learning rates in kernel regression (Caponnetto & De Vito, 2007; Spigler et al., 2020; Cui et al., 2021). In noisy interpolation, power-law spectra can produce a nonzero asymptotic noise floor (Mallinar et al., 2022; Cheng & Montanari, 2024). In Section 5, we use this noise floor to determine whether the noise curve develops an interior peak above it. Relatedly, whitening leaves the ordinary least-squares noise risk unchanged and does not increase the ridgeless noise risk (Holzmüller, 2021, Remark G.1). Appendix G discusses the relation between these results and ours in more detail.

Power-law anisotropy. Most closely related to our work, Rizzi et al. (2026) study kernel ridge regression with power-law Gaussian covariates and characterize the suppression of variance peaks. For a linear kernel and $0 \le \alpha < 1$, their continuum equation agrees with ours after a change of normalization. For $\alpha > 1$, however, the two asymptotic regimes differ: their limit keeps the ridge penalty fixed, whereas ours keeps $r = \lambda d^{\alpha-1}$ fixed, leading to different scalings of the penalty (Appendix F).

3 Model and deterministic equivalent

Let $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n\times d}$ with independent inputs $x_i \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \operatorname{diag}(\mu_1, \dots, \mu_d) \succ 0$, and let $y = Xw^* + \epsilon$ with noise ϵ independent of X, of mean zero and covariance $\sigma^2 I_n$. The ridge estimator with penalty $\lambda > 0$ is

$$\widehat{w} = \arg\min_w \{\|y - Xw\|^2 + \lambda\|w\|^2\} = RX^\top y, \qquad R := (X^\top X + \lambda I_d)^{-1}. \tag{1}$$

Minimizing the average loss $n^{-1}\|y - Xw\|^2 + \lambda'\|w\|^2$ gives the same estimator with $\lambda = n\lambda'$. The prediction error on a fresh input $x \sim \mathcal{N}(0, \Sigma)$, averaged over the label noise, splits into a bias term, which depends on the teacher w^*, and a variance term, which is caused by the label noise,

$$\mathbb{E}_\epsilon \mathbb{E}_x \left[(x^\top \widehat{w} - x^\top w^*)^2\right] = \left\|\Sigma^{1/2}(RX^\top X - I_d)w^*\right\|^2 + \sigma^2 \operatorname{Tr}\left[\Sigma R X^\top X R\right].$$

We call the average of the variance term over the input data X the noise risk,

$$\mathcal{V}_{\text{exact}} = \sigma^2 \mathbb{E}_X \operatorname{Tr}\big[\Sigma R X^\top X R\big]. \tag{2}$$

It is the whole prediction error when $w^* = 0$, and it contains the double-descent peak, as we explain below.

The analysis uses the DE of ridge regression (Dobriban & Wager, 2018; Hastie et al., 2022; Bordelon et al., 2020; Canatar et al., 2021; Bach, 2024), whose derivation we recall in Appendix A. We define the effective ridge $\kappa > 0$ as the solution of

$$\mathrm{df}_1(\kappa) + \frac{\lambda}{\kappa} = n, \qquad \mathrm{df}_1(\kappa) := \sum_{k=1}^{d} C_k, \qquad C_k := \frac{\mu_k}{\mu_k + \kappa}. \tag{3}$$

The left side of Eq. (3) decreases strictly from infinity to zero in κ, so the solution exists and is unique. By the approximation of $\mathbb{E}_X[R]$ in Appendix A, the estimator averaged over the input data and the label noise shrinks the k-th coordinate of w^* by the factor $C_k \in (0, 1)$. The quantity C_k thus measures how well the k-th direction is learned, and df_1 is the effective degrees of freedom. The same factor arises in ridge regression with penalty κ on the covariance Σ itself, which is why we call κ the effective ridge. Simon et al. (2023) call C_k the learnability of the k-th direction and κ the effective regularization, and Atanasov et al. (2026) call κ the renormalized ridge. With the average-loss penalty $\lambda' = \lambda/n$, Eq. (3) reads $\kappa = \lambda'/(1 - \mathrm{df}_1/n)$, so κ is larger than λ'. Through the factors C_k, the DE then approximates the noise risk in Eq. (2) by

$$\mathcal{V} = \frac{\sigma^2 \mathrm{df}_2}{n - \mathrm{df}_2}, \qquad \mathrm{df}_2 := \sum_{k=1}^{d} C_k^2. \tag{4}$$

Since $\mathrm{df}_2 < \mathrm{df}_1 < n$, the denominator $n - \mathrm{df}_2$ is positive. When λ is small, it becomes small near $n = d$, which results in double descent.

In what follows, we consider a power-law spectrum, $\mu_k = k^{-\alpha}$ with $\alpha \geq 0$, whose largest eigenvalue is one. For given d and σ^2, the noise risk in Eq. (4) then depends only on the exponent α, the ridge penalty λ, and the sample size n, and we write it as $\mathcal{V}_{\alpha,\lambda}(n)$. Our object of study is the noise curve, $\mathcal{V}_{\alpha,\lambda}(n)$ as a function of n, which we plot against the sample-to-parameter ratio $u = n/d$. We compare how this curve responds to two changes of a baseline (α, λ), the *power-law exponent change* $(\alpha + \Delta\alpha, \lambda)$ and the *ridge change* $(\alpha, \lambda_{\text{eq}})$, where λ_{eq}, given in Theorem 1, is chosen once and used at every sample size.

4 Scale effect and the equivalence condition

In Section 3, the noise peak comes from the denominator $n - \mathrm{df}_2$ of Eq. (4), which becomes small near $n = d$. To see how the spectrum affects it, we split it with Eq. (3) into two terms,

$$\begin{aligned} n - \mathrm{df}_2 &= (n - \mathrm{df}_1) + (\mathrm{df}_1 - \mathrm{df}_2) \\ &= \frac{\lambda}{\kappa} + \sum_{k=1}^{d} C_k(1 - C_k). \end{aligned}$$

The first term comes from the ridge penalty. The k-th term of the sum is $C_k(1-C_k) = \kappa\mu_k/(\mu_k+\kappa)^2$, which is close to κ/μ_k when $\kappa \ll \mu_k$. Small eigenvalues therefore make the second term large. Near $n = d$, where the denominator is small, this lowers the noise risk in Eq. (4) in the same way as the first term λ/κ from the ridge penalty does. This suggests that a larger power-law exponent, which makes the small eigenvalues smaller, lowers the noise peak like a larger ridge penalty, as Figure 1(b) shows at fixed λ.

More specifically, we ask how a change of the power-law exponent corresponds to a change of the ridge penalty near the peak. The answer is not obvious, because the spectrum enters Eq. (3) through all d eigenvalues, whereas the ridge penalty is a single number. To state the answer, let

$$S_1 := \sum_{k=1}^{d} \mu_k^{-1}, \qquad \kappa_0 := \sqrt{\lambda/S_1}.$$

When $\kappa \ll \mu_{\min} = \min_k \mu_k$, the fixed point at $n = d$ gives $\lambda/\kappa = \sum_k \kappa/(\mu_k + \kappa) \simeq \kappa S_1$, so $\kappa \simeq \kappa_0$. We say that the ridge is weak when $\kappa_0 \ll \mu_{\min}$, so that $C_k \simeq 1$ for every k at interpolation. In this setting, we can prove the following theorem giving an explicit leading-order form of the noise curve that depends on the spectrum and the ridge penalty only through the single combination λS_1 (Appendix B).

Theorem 1 (Spectral–ridge equivalence)**.** *For a positive spectrum $\mu_1, \dots, \mu_d$ and $\lambda > 0$, and for every fixed $L < \infty$, uniformly on the peak region $|n - d| \le 2L\sqrt{\lambda S_1}$ as $\kappa_0/\mu_{\min} \to 0$,*

$$\frac{\mathcal{V}(n)}{\sigma^2} = \big(1 + O(\kappa_0/\mu_{\min})\big)\frac{d}{\sqrt{(n-d)^2 + 4\lambda S_1}}. \tag{5}$$

Consequently, two spectra $\mu^{(A)}$ and $\mu^{(B)}$ of the same dimension, with ridge penalties λ_A and λ_B and sums $S_1^{(A)}$ and $S_1^{(B)}$, have the same leading curve, the right side of Eq. (5), if and only if

$$\lambda_A S_1^{(A)} = \lambda_B S_1^{(B)}. \tag{6}$$

For $\mu_k = k^{-\alpha}$, write $S_1(\alpha) = \sum_k k^{\alpha}$. Changing the exponent from α to $\alpha + \Delta\alpha$ at ridge penalty λ is therefore equivalent to keeping α and changing the ridge penalty to

$$\lambda_{\mathrm{eq}} = \lambda \frac{S_1(\alpha + \Delta\alpha)}{S_1(\alpha)}, \tag{7}$$

and to no other ridge penalty.

Here $O(\kappa_0/\mu_{\min})$ denotes a term bounded in absolute value by a constant times $\kappa_0/\mu_{\min}$ once $\kappa_0/\mu_{\min}$ is small enough, where the constant and the threshold depend only on L. The leading curve is a peak of height $\sigma^2 d/(2\sqrt{\lambda S_1})$ and full width at half maximum $4\sqrt{3\lambda S_1}$, and the peak region extends over L times the width scale $2\sqrt{\lambda S_1}$ on each side of $n = d$. Since a steeper spectrum has a larger S_1, it lowers and broadens the peak in the same way as a larger ridge penalty. The proof in Appendix B expands the DE in the small ratios κ/μ_k, which reduces the fixed point to a quadratic equation for κ whose coefficients involve the spectrum only through S_1.

The sum S_1 of the inverse eigenvalues already appears in ridge asymptotics at interpolation (Dobriban & Wager, 2018), and for equal eigenvalues Canatar et al. (2021) give the height of the noise peak, which Eq. (5) recovers at $n = d$ at weak ridge. The above theorem extends these results to a general spectrum, gives the equivalence condition, and holds uniformly over the peak region.

Figure 1 shows the noise curves at fixed α for increasing ridge penalties in (a) and at fixed λ for increasing power-law exponents in (b). The ridge penalties in (a) are the values λ_{eq} of Eq. (7) for the exponents in (b), and the corresponding curves nearly coincide near the peak. For power-law spectra, the multiplier in Eq. (7) grows exponentially with the exponent change. Its rate is

$$\frac{\partial}{\partial\Delta\alpha} \log \frac{\lambda_{\mathrm{eq}}}{\lambda}\bigg|_{\Delta\alpha=0} = \frac{\sum_k k^{\alpha} \log k}{\sum_k k^{\alpha}} \simeq \log d - \frac{1}{\alpha + 1} \tag{8}$$

for large d. At $d = 300$ and $\alpha = 2$, raising the exponent by 0.1 corresponds to multiplying the ridge penalty by 1.71.

Multiplying the covariance and the ridge penalty by the same constant $c > 0$ leaves the noise risk in Eq. (2) and its DE unchanged, since X becomes $\sqrt{c}\,X$, R becomes R/c, and κ becomes $c\kappa$. Multiplying the covariance by c therefore has the same effect as dividing λ by c, and a value of λ has a meaning only together with the scale of the spectrum. While we fix the scale by $\mu_1 = 1$, another common choice fixes the total variance, $\operatorname{tr}\Sigma = d$, as feature standardization does. However, the theorems hold under either choice, because they involve λ only through combinations that this rescaling leaves unchanged.

One might suspect that the peak suppression only reflects the normalization $\mu_1 = 1$, under which a larger exponent also lowers the total variance $\operatorname{tr}\Sigma$ and, by the invariance above, acts like a larger ridge penalty. However, Theorem 1 shows that this is not the case. At fixed total variance $\operatorname{tr}\Sigma = d$, the eigenvalues have mean one, so S_1/d, the mean of μ_k^{-1}, is at least one by Jensen's inequality, with equality only when all eigenvalues are equal. By Eq. (5), at fixed total variance and fixed λ, any spectrum with unequal eigenvalues therefore has a lower peak than equal eigenvalues, by the factor $\sqrt{S_1/d}$.

The equivalence given by Theorem 1 holds near the peak. Away from it, the exponent change and the ridge change differ, and we write the effect of an exponent change at a fixed ridge penalty as

$$\mathcal{V}_{\alpha+\Delta\alpha,\lambda} - \mathcal{V}_{\alpha,\lambda} = \underbrace{\left(\mathcal{V}_{\alpha,\lambda_{\mathrm{eq}}} - \mathcal{V}_{\alpha,\lambda}\right)}_{\text{scale effect}} + \underbrace{\left(\mathcal{V}_{\alpha+\Delta\alpha,\lambda} - \mathcal{V}_{\alpha,\lambda_{\mathrm{eq}}}\right)}_{\text{shape effect}}. \tag{9}$$

Here the scale of the spectrum is S_1, and its shape is the normalized inverse spectrum $(1/\mu_k)/S_1$. The first term changes the scale of the inverse spectrum at fixed shape, which is the ridge change to λ_{eq}. The second changes the shape at fixed λS_1, and it is the part that a ridge change does not reproduce. By Theorem 1 it is of lower order than the scale effect near the peak. However, the shape effect remains away from the peak (Figure 1(c)). In the continuum limit of Section 5, no choice of the ridge penalty removes it.

5 Existence of an interior noise peak

Theorem 1 shows that a larger power-law exponent lowers the noise peak near interpolation. A lower peak, however, does not mean that the peak disappears, and we now ask whether a large enough exponent entirely removes the interior peak.

However, the answer depends on the whole noise curve, rather than only on its behavior near the peak. At finite d, the DE gives each point of the curve only through the solution κ of Eq. (3), which involves a sum of d terms, and this makes it hard to decide whether the curve has an interior maximum. We therefore take a continuum limit, in which the dimension d tends to infinity while the sample ratio $u = n/d$ and a suitably rescaled ridge penalty are held fixed. In this limit, the sums over the eigenvalues become integrals, and the whole curve has an explicit parametric form.

To obtain integrals, let κ scale with the smallest eigenvalue $\mu_d = d^{-\alpha}$ and write $\kappa = \tau d^{-\alpha}$. Then $C_k = 1/\{1+\tau(k/d)^\alpha\}$ depends on k only through k/d, so $\mathrm{df}_j/d = d^{-1}\sum_k C_k^j$ becomes a Riemann sum. To keep the ridge term $\lambda/\kappa = rd/\tau$ in Eq. (3) of order d, like df_1, we hold fixed the rescaled ridge penalty

$$r := \frac{\lambda}{d\mu_d} = \lambda d^{\alpha-1}. \tag{10}$$

The ridge penalty $\lambda = rd^{1-\alpha}$ therefore changes with d in this limit. As $d \to \infty$ with $\alpha \geq 0$, $r > 0$, and $u > 0$ fixed, df_j/d tends to $J_j(\tau) := \int_0^1 (1+\tau x^\alpha)^{-j}\,\mathrm{d}x$, and the fixed point in Eq. (3) divided by d and the noise risk in Eq. (4) become

$$u = J_1(\tau) + \frac{r}{\tau}, \qquad \frac{\mathcal{V}_\infty(u)}{\sigma^2} = \frac{J_2(\tau)}{u - J_2(\tau)}, \tag{11}$$

where the positive solution τ is unique (Appendix C). As τ runs over $(0,\infty)$, Eq. (11) traces the whole curve $\mathcal{V}_\infty$. Then we obtain the following theorem, which gives an explicit condition for $\mathcal{V}_\infty$ to have an interior peak, that is, a maximum at some $u \in (0,\infty)$ (Appendix C).

Theorem 2 (Existence of a continuum noise peak)**.** *Consider Eq. (11) with $r > 0$ and $\sigma^2 > 0$. For $\alpha > 1$, an interior noise maximum exists if and only if*

$$(\alpha - 1)r < 1. \tag{12}$$

For $(\alpha-1)r \geq 1$, $\mathcal{V}_\infty$ is strictly decreasing on $u > 0$. For $0 \leq \alpha \leq 1$, an interior maximum exists for every $r > 0$. At $\alpha = 1$, it is unique and occurs at $u = 1$.

The condition in Theorem 2, which we call the peak criterion, can be understood from the behavior of $\mathcal{V}_\infty$ at small sample ratios. For $\alpha > 1$, $\mathcal{V}_\infty$ does not vanish as $u \to 0$. As u decreases, the ridge term r/τ in Eq. (11) becomes negligible compared with $J_1(\tau)$, and $\mathcal{V}_\infty$ approaches $\sigma^2(\alpha-1)$. This value is the limit, as the ridge penalty tends to zero, of the noise risk that Cheng & Montanari (2024, Theorem 5, Eq. (35)) derive for a power-law spectrum. Interpolation with this nonzero noise floor is what Mallinar et al. (2022) call tempered overfitting. This limit increases with α and does not depend on the ridge penalty. The criterion decides whether $\mathcal{V}_\infty$ rises above this limit. When it holds, $\mathcal{V}_\infty$ rises above the limit and has an interior maximum, and when it fails, $\mathcal{V}_\infty$ decreases monotonically from the limit.

In the original parameters, the criterion reads $(\alpha-1)\lambda d^{\alpha-1}<1$, whose left side increases with α. At fixed λ and d, the peak therefore disappears from the threshold exponent α_c on, defined by $(\alpha_c-1)\lambda d^{\alpha_c-1}=1$. Figure 2 shows $\mathcal{V}_\infty$ at fixed λ for exponents below, at, and above α_c in (a), and the region of the (α, r) plane where the criterion holds in (b).

The limit $\sigma^2(\alpha-1)$ also shows that a ridge change cannot reproduce an exponent change over the whole curve. When the ridge is weak, an exponent change is equivalent to a ridge change near the peak (Theorem 1). Far below the peak, the situation is different. For $\alpha>1$, $\mathcal{V}_\infty$ approaches $\sigma^2(\alpha-1)$ as $u\to 0$ whatever the ridge penalty is, because the ridge term r/τ becomes negligible there. Two curves with different exponents $\alpha>1$ therefore approach different values at small sample ratios, and no choice of the ridge penalty makes them coincide (Corollary C.1). In particular, the shape effect of an exponent change $\Delta\alpha$, which is small near the peak, tends to $\sigma^2\Delta\alpha$ as $u\to 0$. The exponent thus also sets the noise risk at small sample ratios, which no ridge penalty can change.

The peak criterion is a statement about the limit $d\to\infty$, that is, about $\mathcal{V}_\infty$. At finite d with $\lambda>0$, the noise risk tends to zero as $u\to 0$ and as $u\to\infty$, so the noise curve always has a maximum. The criterion still describes where this maximum lies. When the criterion holds, the finite curve has its maximum near the interior maximum of $\mathcal{V}_\infty$ (Figure 2(a)). When it fails, the maximum lies far below interpolation. The value at small sample sizes therefore depends on the order of the limits. Taking $d\to\infty$ first and then $u\to 0$ gives $\sigma^2(\alpha-1)$, whereas taking $u\to 0$ at finite d gives zero.

We also test the criterion on real data. We take the leading 300 principal components of the MNIST pixels (LeCun et al., 1998), whiten them, and multiply the k-th component by $k^{-\alpha/2}$. The covariance is then exactly $k^{-\alpha}$, while the eigenvectors and the joint distribution of the whitened components are kept. On these inputs, we compute the trace $\sigma^2\operatorname{Tr}[\Sigma R X^\top X R]$ of Eq. (2) exactly for each subsample and average it over 40 random subsamples. The measured curves follow the finite-sum DE (markers in Figure 2(a)). As the criterion predicts, the peak near interpolation appears at $\alpha=2$, with its maximum at $u=0.83$, and is absent at $\alpha_c\approx 2.18$ and at 2.35, whose measured maxima lie at $u\le 0.08$. The measured values lie 1 to 3% above the DE near interpolation and up to 15% above it at small u.

6 Experiments with Gaussian inputs and real data

We test Theorem 1 first on Gaussian inputs of finite dimension, for which the DE is only an approximation, with the parameters of Figure 1, and then on real data (Figure 3).

In both experiments, we compute the trace $\sigma^2\operatorname{Tr}[\Sigma R X^\top X R]$ in Eq. (2) exactly for each draw of the input data X and average it over draws. The input data with n samples consist of the first n rows of a draw, and the same draws are used for every spectrum and every sample size. This makes the differences between settings paired, which reduces their sampling error. Since the trace already averages over the label noise, no noise is sampled. Grids, seeds, and further details are in Appendix D.

Gaussian inputs. Each draw is $X=Z\Sigma^{1/2}$, where Z has independent standard Gaussian entries. In the setting of Figure 1 ($d=300$, $\alpha=2$, $\lambda=10^{-4}$, $\sigma^2=1$), we compare the baseline with the power-law exponent change and the ridge change to λ_{eq} for $\Delta\alpha=0.1$ and 0.5, using 400 draws (Appendix D.1). At $n=d$, the ridge change reproduces the fraction $(\mathcal{V}_{\alpha,\lambda_{\text{eq}}}-\mathcal{V}_{\alpha,\lambda})/(\mathcal{V}_{\alpha+\Delta\alpha,\lambda}-\mathcal{V}_{\alpha,\lambda})$ of the exponent change, which is 97.9% for $\Delta\alpha=0.1$ and 96.9% for $\Delta\alpha=0.5$. Over the sample grid, the measured noise risks of these five settings differ from the DE by at most 0.86%. The measured shape effects differ from their DE values by about 2%, more than their standard errors, so the finite-size effects are small but detectable.

Real data. To change the exponent of real data while keeping the data themselves, we use the leading 300 principal components of 50,000 MNIST images as the input data and multiply the k-th principal component by $k^{-\Delta\alpha/2}$ for $\Delta\alpha=0.25$, 0.5, and 1. This multiplies the k-th eigenvalue of the empirical covariance by $k^{-\Delta\alpha}$ and leaves the eigenvectors and the non-Gaussian joint distribution of the whitened components unchanged. The exponent of a power-law fit to the spectrum moves from 1.48 to 1.73, 1.98, and 2.48 (Figure 3(a)). We take Σ to be the empirical covariance of the 50,000 images and compute the trace in Eq. (2) exactly for input data X made of n images drawn without replacement. We use 60 draws. The ridge penalty of the baseline satisfies $\lambda S_1=90$.

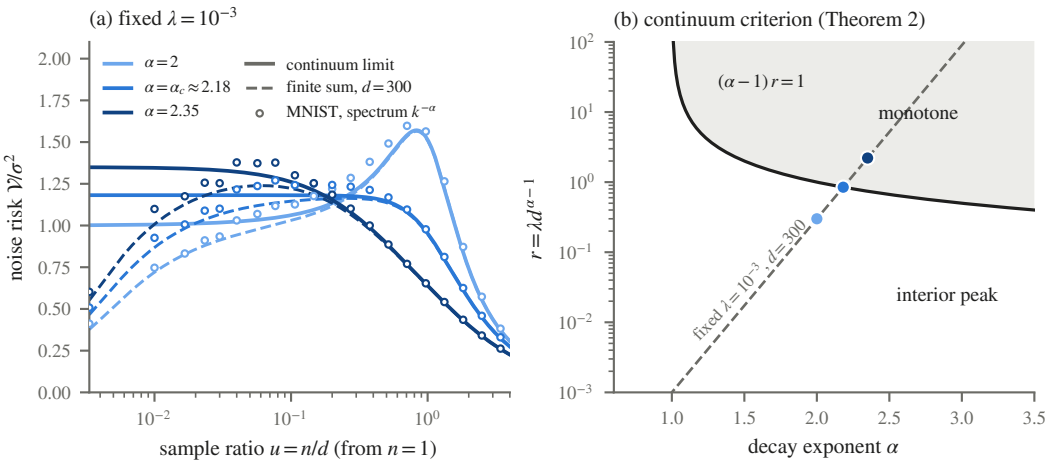

Figure 2: Existence of an interior noise peak (Theorem 2). (a) Noise curves at the ridge penalty $\lambda = 10^{-3}$ for $\alpha = 2$, for the threshold exponent $\alpha_c \approx 2.18$ at which $(\alpha - 1)\lambda d^{\alpha-1} = 1$ with $d = 300$, and for $\alpha = 2.35$. Solid lines are the continuum limit $\mathcal{V}_\infty$ of Eq. (11), dashed lines are the DE of Eqs. (3)–(4) at $d = 300$, and circles are the noise risk measured on the leading 300 principal components of MNIST with the covariance set to $k^{-\alpha}$. The horizontal axis is logarithmic and starts at $n = 1$. As $u \to 0$, $\mathcal{V}_\infty$ tends to $\sigma^2(\alpha - 1)$. It has an interior peak for $\alpha = 2$ and decreases monotonically for α_c and 2.35. At $d = 300$, the DE and the MNIST curves decrease toward zero at small u, so they always have a maximum. This maximum lies near interpolation for $\alpha = 2$ and far below it for the other two exponents. (b) The peak criterion of Theorem 2 in the plane of the exponent α and the rescaled ridge penalty $r = \lambda d^{\alpha-1}$. The solid line is the boundary $(\alpha - 1)r = 1$, and in the shaded region above it $\mathcal{V}_\infty$ decreases monotonically. The dashed line is the path of fixed $\lambda = 10^{-3}$ and $d = 300$ followed in (a), and the three circles mark the exponents of (a).

Figure 3(b) compares the power-law exponent change with the ridge change by the ratio of S_1 after and before the exponent change, as Theorem 1 predicts, with S_1 computed from the empirical spectra. The two curves nearly coincide near the peak and separate below it by the shape effect, which is positive at every sample ratio (Figure 3(c)). At $n = d$, the ridge change reproduces 98.5 to 99.0% of the exponent change, measured by the same fraction as for the Gaussian inputs. Theorem 1 also predicts the factor by which the ridge penalty must be multiplied, namely the ratio of S_1 after and before the exponent change. At $n = d$, the factor that reproduces the exponent change exactly is within 2% and 7% of this ratio for $\Delta\alpha = 0.25$ and 0.5. For $\Delta\alpha = 1$, it is only about two thirds of the ratio (148 instead of 221). This setting violates the weak-ridge condition, with $\kappa_0/\mu_{\min} = 1.8$, and the DE with $\mu_k = k^{-2}$ and the same λS_1 falls short by a similar amount (156 instead of 225). The shortfall therefore comes from the leading-order approximation of Theorem 1, not from the real data. The measured curves of the baseline and the exponent changes lie 6 to 10% above the leading curve of Theorem 1 at $n = d$ and up to 28% above it on the flanks. The exponent change and the ridge change share this deviation, so it largely cancels in their comparison. Section 5 tests Theorem 2 on MNIST with the covariance set to $k^{-\alpha}$, and Appendix D.2 compares four MNIST representations, where the peaks agree under the equivalence condition except for random ReLU features, whose isolated small eigenvalue violates the weak-ridge condition.

7 Discussion

We have studied how the power-law exponent of the input spectrum acts on the noise peak of ridge regression. The change in noise risk caused by an exponent change splits into a scale effect, which comes from the change of the overall scale of the spectrum, and a shape effect, which comes from the change of its shape. Near interpolation and when the ridge penalty is small, the shape effect is negligible, so the exponent change is equivalent to multiplying the ridge penalty by the ratio of the sums of the inverse eigenvalues (Theorem 1). In this sense, the spectrum acts as an implicit ridge penalty. Away from the peak, the shape effect remains, and in the continuum limit no ridge change

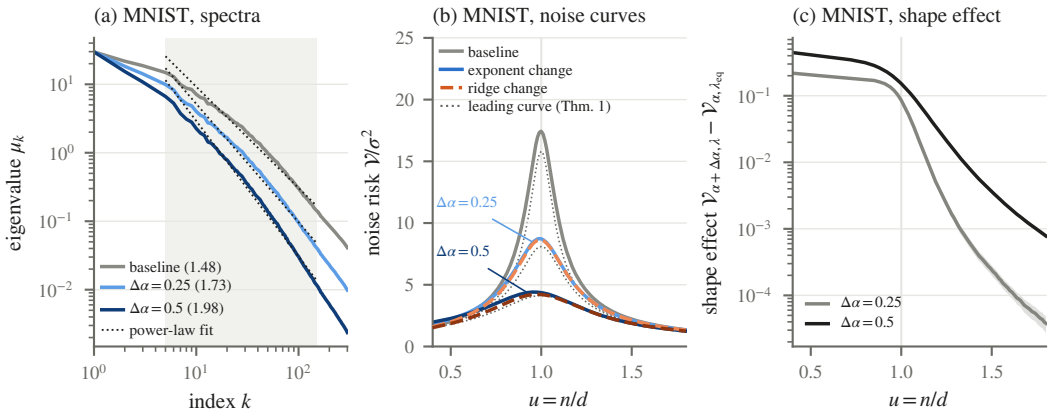

Figure 3: Theorem 1 on real data. The input data are the leading 300 principal components of 50,000 MNIST images, and the power-law exponent is raised by $\Delta\alpha$ by multiplying the k-th component by $k^{-\Delta\alpha/2}$. The ridge penalty of the baseline gives $\lambda S_1 = 90$, with S_1 computed from the empirical spectrum. (a) Eigenvalues μ_k of the empirical covariance for the baseline and for $\Delta\alpha = 0.25$ and 0.5, with the least-squares power-law fits over $k = 5, \ldots, 150$ (dotted, shaded range). The numbers in parentheses are the fitted exponents. (b) Noise curves of the baseline (gray), of the exponent changes (solid), and of the ridge changes by the ratio of S_1 after and before the exponent change, as Theorem 1 predicts (dashed), for $\Delta\alpha = 0.25$ and 0.5. Dotted lines are the leading curves $d/\sqrt{(n-d)^2 + 4\lambda S_1}$ of Theorem 1 for the baseline and for the two changed spectra. Near the peak, the exponent change and the ridge change nearly coincide. (c) The shape effect of Eq. (9), $\mathcal{V}_{\alpha+\Delta\alpha,\lambda} - \mathcal{V}_{\alpha,\lambda_{\mathrm{eq}}}$, the difference between the noise curve of each exponent change and that of the corresponding ridge change in (b), on a logarithmic axis, with bands of ± 1.96 standard errors. It is positive at every sample ratio. All curves are measured on the real inputs at $u = 0.40, 0.41, \ldots, 1.80$ and averaged over 60 draws of the input data shared by all settings.

reproduces an exponent change over the whole noise curve (Corollary C.1). Whether the peak exists at all is decided by the peak criterion of Theorem 2, and in the continuum limit, a large enough exponent removes it.

The noise risk is only one part of the prediction error of Section 3. The other part, the bias term, depends on the teacher w^* and also changes with the exponent, so a lower noise risk does not by itself mean a lower prediction error. Appendix E compares the DE of the prediction error at $n = d$ as α changes, with the signal power $\mu_k (w_k^*)^2$ in each direction held fixed. Raising the exponent lowers the prediction error when the signal power decays as k^{-2}, so that the signal is concentrated on the directions with large eigenvalues, and raises it beyond $\alpha \approx 1.9$ when the signal power is the same in every direction. Whether the implicit regularization by the spectrum lowers the prediction error therefore depends on how the signal power is distributed across directions (Canatar et al., 2021).

Limitations. The theorems are statements about the DE, not about the exact noise risk in Eq. (2). The DE becomes exact when n and d grow with n/d fixed, but we know of no result that bounds its error uniformly near interpolation when λ is small, which is where the peak lies, or at small n/d. Transferring the theorems to finite dimensions therefore remains open, and we compare the DE with Gaussian inputs of finite dimension only numerically (Section 6). Theorem 2 concerns the continuum limit. At finite d the noise curve always has a maximum, and when the peak criterion fails, this maximum lies at very small sample sizes (Section 5). We treat ridge regression with $\lambda > 0$ only. At $\lambda = 0$ and $n < d$, the estimator is the minimum-norm interpolant, and the spectrum changes which interpolant is selected rather than the strength of a penalty (Appendix G). How the scale and shape effects appear for gradient descent with early stopping is left open. The real-data tests use one dataset, MNIST, with fixed linear features. Learning a representation changes not only the spectrum but also the higher-order statistics of the inputs and the alignment of the target with the eigenvectors, and whether the decomposition in Eq. (9) still describes such changes requires further evidence.

REPRODUCIBILITY STATEMENT

Appendices A–G provide the derivations, sampling protocols, and supplementary comparisons. Seeds, sample grids, pairing, and uncertainty estimates are specified in Appendix D. DE evaluations and measurements on Gaussian inputs and on MNIST are identified separately throughout.

AI USE STATEMENT

Generative AI tools were used in the following ways. They helped draft sections of this paper and edit its text, and they were used to search for and summarize related work. They assisted with research planning, hypothesis refinement, the design and implementation of the experiments, and the interpretation of results. They also assisted with the mathematical derivations and with drafting and checking the proofs of the theorems. No data set was generated with AI tools. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, pp. 043404, 2026.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

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.

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

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034, 2020.

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*, 12:2914, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6): 2879–2912, 2024.

Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Michał Dereziński, Feynman Liang, and Michael W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

Lars Eldén. Algorithms for the regularization of ill-conditioned least squares problems. *BIT Numerical Mathematics*, 17:134–145, 1977.

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

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

David Holzmüller. On the universality of the double descent peak in ridgeless regression. In *International Conference on Learning Representations*, 2021.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Adam Santoro, Guillaume Lajoie, and Blake A. Richards. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Andy Zeyi Liu, Elliot Paquette, and John Sous. Spectral lens: Activation and gradient spectra as diagnostics of LLM optimization. *arXiv preprint arXiv:2605.05683*, 2026.

Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Lorenzo Rizzi, Arie Wortsman Zurich, and Bruno Loureiro. Learning between the peaks: Sharp asymptotics for kernel ridge regression under power-law anisotropy. *arXiv preprint arXiv:2608.28564*, 2026.

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. *Transactions on Machine Learning Research*, 2023.

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

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23549–23588, 2022.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A THE DETERMINISTIC EQUIVALENT

This appendix derives the DE in Eqs. (3)–(4). We write the noise risk through the resolvent $R = (X^\top X + \lambda I_d)^{-1}$, approximate $\mathbb{E}_X[R]$ with the leave-one-out argument (Couillet & Liao, 2022, Section 2.2), and insert the approximation, as in Bach (2024). The same equations follow from the replica method (Bordelon et al., 2020; Canatar et al., 2021) and are exact as n and d grow with n/d fixed, under conditions on Σ and λ stated at the end of this appendix (Dobriban & Wager, 2018; Hastie et al., 2022). The setting is that of Section 3: $X = (x_1, \dots, x_n)^\top$ with independent $x_i \sim \mathcal{N}(0, \Sigma)$, $y = Xw^* + \epsilon$, and $\widehat{w} = RX^\top y$.

The estimator error is $\widehat{w} - w^* = (RX^\top X - I_d)w^* + RX^\top \epsilon$. On a fresh input $x \sim \mathcal{N}(0, \Sigma)$, the squared prediction error averaged over the label noise splits into a bias term and a variance term,

$$\begin{aligned}
\mathbb{E}_\epsilon \mathbb{E}_x \big[(x^\top \widehat{w} - x^\top w^*)^2\big] &= \mathbb{E}_\epsilon \big[(\widehat{w} - w^*)^\top \Sigma\, (\widehat{w} - w^*)\big] \\
&= \big\|\Sigma^{1/2}(RX^\top X - I_d)w^*\big\|^2 + 2\big((RX^\top X - I_d)w^*\big)^\top \Sigma R X^\top\, \mathbb{E}_\epsilon[\epsilon] \\
&\quad + \mathbb{E}_\epsilon\big[\epsilon^\top X R \Sigma R X^\top \epsilon\big] \\
&= \big\|\Sigma^{1/2}(RX^\top X - I_d)w^*\big\|^2 + \sigma^2 \operatorname{Tr}\big[\Sigma R X^\top X R\big]. \qquad (13)
\end{aligned}$$

The first line uses $\mathbb{E}_x[xx^\top] = \Sigma$, the second and third insert the error, and the last uses $\mathbb{E}_\epsilon[\epsilon] = 0$, $\mathbb{E}_\epsilon[\epsilon\epsilon^\top] = \sigma^2 I_n$, and the cyclicity of the trace. The noise risk in Eq. (2) is the average of the variance term over X, and it is a derivative of the resolvent,

$$\begin{aligned}
\mathcal{V}_{\text{exact}} &= \sigma^2 \mathbb{E}_X \operatorname{Tr}\big[\Sigma R X^\top X R\big] \\
&= \sigma^2 \mathbb{E}_X \operatorname{Tr}\big[\Sigma (R - \lambda R^2)\big] \\
&= \sigma^2 \frac{\mathrm{d}}{\mathrm{d}\lambda} \mathbb{E}_X \operatorname{Tr}[\Sigma\, \lambda R]. \qquad (14)
\end{aligned}$$

The second line uses $X^\top X R = I_d - \lambda R$, and the third uses $\mathrm{d}R/\,\mathrm{d}\lambda = -R^2$, so that $\mathrm{d}(\lambda R)/\,\mathrm{d}\lambda = R - \lambda R^2$.

To evaluate Eq. (14) we need $\mathbb{E}_X[R]$, which has no closed form because R depends on all rows of X at once. The leave-one-out argument separates one row at a time. Let R_{-i} be the resolvent without the i-th row, which is independent of x_i. Then

$$\begin{aligned}
I_d &= R\,(X^\top X + \lambda I_d) \\
&= \sum_{i=1}^n R x_i x_i^\top + \lambda R \\
&= \sum_{i=1}^n \frac{R_{-i} x_i x_i^\top}{1 + x_i^\top R_{-i} x_i} + \lambda R. \qquad (15)
\end{aligned}$$

The second line writes $X^\top X = \sum_i x_i x_i^\top$, and the third uses the Sherman–Morrison formula $Rx_i = R_{-i}x_i/(1 + x_i^\top R_{-i} x_i)$, which follows from $R^{-1} = R_{-i}^{-1} + x_i x_i^\top$. For large n and d we use two approximations: the quadratic form $x_i^\top R_{-i} x_i$ concentrates around its mean $\operatorname{Tr}(\Sigma R_{-i})$ over x_i, and removing one row changes the resolvent little, so that $\mathbb{E}_X[R_{-i}] \simeq \mathbb{E}_X[R]$ and $\operatorname{Tr}(\Sigma R_{-i}) \simeq \gamma := \operatorname{Tr}(\Sigma\, \mathbb{E}_X[R])$. Averaging Eq. (15) over X,

$$\begin{aligned}
I_d &\simeq \frac{1}{1+\gamma} \sum_{i=1}^n \mathbb{E}_X[R_{-i}]\, \mathbb{E}_X\big[x_i x_i^\top\big] + \lambda\, \mathbb{E}_X[R] \\
&\simeq \mathbb{E}_X[R] \Big(\frac{n}{1+\gamma}\Sigma + \lambda I_d\Big), \qquad (16)
\end{aligned}$$

where the first line also uses that x_i is independent of R_{-i}, and the second uses $\mathbb{E}_X[x_i x_i^\top] = \Sigma$. Solving for $\mathbb{E}_X[R]$,

$$\begin{aligned}\lambda\,\mathbb{E}_X[R] &\simeq \lambda\Big(\frac{n}{1+\gamma}\Sigma + \lambda I_d\Big)^{-1} \\ &= \Big(\frac{\Sigma}{\kappa} + I_d\Big)^{-1} \\ &= \kappa(\Sigma + \kappa I_d)^{-1}, \qquad \kappa := \frac{\lambda(1+\gamma)}{n}. \end{aligned} \tag{17}$$

The random resolvent at ridge penalty λ thus behaves on average like the resolvent of Σ at the effective ridge κ. In particular, since $\mathbb{E}_\epsilon[\widehat{w}] = RX^\top X w^* = (I_d - \lambda R)w^*$, the averaged estimator is $(I_d - \lambda\mathbb{E}_X[R])w^* \simeq \Sigma(\Sigma + \kappa I_d)^{-1}w^*$, whose k-th coordinate is $C_k w_k^*$.

The effective ridge is fixed by consistency with the definition of γ. By Eq. (17),

$$\begin{aligned}\frac{n\kappa}{\lambda} - 1 &= \gamma \\ &= \mathrm{Tr}(\Sigma\,\mathbb{E}_X[R]) \\ &\simeq \frac{\kappa}{\lambda}\,\mathrm{Tr}\big[\Sigma(\Sigma + \kappa I_d)^{-1}\big] \\ &= \frac{\kappa}{\lambda}\mathrm{df}_1(\kappa),\end{aligned}$$

and multiplying by λ/κ gives the fixed point in Eq. (3),

$$\mathrm{df}_1(\kappa) + \frac{\lambda}{\kappa} = n. \tag{18}$$

Inserting Eq. (17) into Eq. (14) gives the noise risk,

$$\begin{aligned}\mathcal{V}_{\text{exact}} &= \sigma^2\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\mathrm{Tr}\big[\Sigma\,\lambda\mathbb{E}_X[R]\big] \\ &\simeq \sigma^2\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\mathrm{Tr}\big[\Sigma\,\kappa(\Sigma+\kappa I_d)^{-1}\big] \\ &= \sigma^2\frac{\mathrm{d}}{\mathrm{d}\lambda}\sum_k \frac{\kappa\mu_k}{\mu_k+\kappa} \\ &= \sigma^2\frac{\mathrm{d}\kappa}{\mathrm{d}\lambda}\sum_k \frac{\mu_k^2}{(\mu_k+\kappa)^2} \\ &= \sigma^2\,\mathrm{df}_2\,\frac{\mathrm{d}\kappa}{\mathrm{d}\lambda} \\ &= \frac{\sigma^2\mathrm{df}_2}{n-\mathrm{df}_2},\end{aligned} \tag{19}$$

which is Eq. (4). The first line is Eq. (14) with the expectation moved inside the trace, the second is Eq. (17), and the third uses that Σ is diagonal. The fourth line uses $\partial_\kappa\{\kappa\mu_k/(\mu_k+\kappa)\} = \mu_k^2/(\mu_k+\kappa)^2$, the fifth the definition of df_2, and the last $\mathrm{d}\kappa/\,\mathrm{d}\lambda = 1/(n-\mathrm{df}_2)$.

The derivative $\mathrm{d}\kappa/\,\mathrm{d}\lambda$ follows from the fixed point, because κ is defined only implicitly by Eq. (18). Its left side is constant in λ, and $\partial_\kappa\mathrm{df}_1(\kappa) = -\sum_k \mu_k/(\mu_k+\kappa)^2$, so

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\Big(\mathrm{df}_1(\kappa) + \frac{\lambda}{\kappa}\Big) \\ &= -\frac{\mathrm{d}\kappa}{\mathrm{d}\lambda}\sum_k\frac{\mu_k}{(\mu_k+\kappa)^2} + \frac{1}{\kappa} - \frac{\lambda}{\kappa^2}\,\frac{\mathrm{d}\kappa}{\mathrm{d}\lambda} \\ &= \frac{1}{\kappa}\Big\{1 - \frac{\mathrm{d}\kappa}{\mathrm{d}\lambda}\Big(\kappa\sum_k\frac{\mu_k}{(\mu_k+\kappa)^2} + \frac{\lambda}{\kappa}\Big)\Big\} \\ &= \frac{1}{\kappa}\Big\{1 - \frac{\mathrm{d}\kappa}{\mathrm{d}\lambda}\,(n - \mathrm{df}_2)\Big\}.\end{aligned}$$

The last line uses $\kappa \sum_k \mu_k/(\mu_k+\kappa)^2 = \sum_k C_k(1-C_k) = \mathrm{df}_1 - \mathrm{df}_2$ and $\lambda/\kappa = n - \mathrm{df}_1$ from Eq. (18). Hence $\mathrm{d}\kappa/\,\mathrm{d}\lambda = 1/(n-\mathrm{df}_2)$.

It remains to state when the derivation is accurate. It uses two approximations, the concentration of $x_i^\top R_{-i} x_i$ and the replacement of R_{-i} by R, and the noise risk in Eq. (19) also uses the derivative in λ of the approximation in Eq. (17). Hastie et al. (2022, Theorem 5) prove that the noise risk is close to the DE, including this derivative, under conditions that require, among others, the ratio n/d to be bounded away from one, the average $d^{-1}\sum_k 1/\mu_k$ of the inverse eigenvalues to be bounded, and λ/n or the smallest eigenvalue to be bounded away from zero. The derivative is controlled there because the approximation holds uniformly over complex λ with positive real part, where both sides are analytic. These conditions do not cover the settings of this paper. The eigenvalues of a power-law spectrum accumulate near zero, Theorem 1 concerns a weak ridge near interpolation, and the continuum limit of Section 5 lets $\lambda = rd^{1-\alpha}$ tend to zero for $\alpha > 1$. We therefore take the DE in Eqs. (3)–(4) as the starting point. Theorem 1 is a statement about this DE, and Theorem 2 is a statement about its continuum limit in Eq. (11). Appendix D compares the DE with Gaussian inputs of finite dimension.

B Proof of Theorem 1

This appendix proves Theorem 1. The first part of the theorem states that, on the peak region $|n-d| \le 2L\sqrt{\lambda S_1}$, the noise risk is close to the leading curve $d/\sqrt{(n-d)^2+4\lambda S_1}$, with a relative error of order $\kappa_0/\mu_{\min}$. The second part, the equivalence condition, follows from the form of this leading curve and is proved at the end.

The proof is based on the following observation. The effective ridge κ is the solution of the fixed point in Eq. (3), which is derived in Appendix A. This fixed point suggests that, near the peak and when the ridge is weak, κ is of the order of $\kappa_0 = \sqrt{\lambda/S_1}$ (Section 4). Since κ_0 is much smaller than every eigenvalue, $C_k = \mu_k/(\mu_k+\kappa)$ is then close to one for all k. The proof has three steps. First, we expand df_1 and df_2 in the small ratios κ/μ_k and bound the remainders. Second, we use the fixed point in Eq. (3) to write the noise risk exactly in terms of these remainders. Third, we prove that κ is at most a constant times κ_0 on the peak region, so that the remainders are small.

Throughout, fix $L < \infty$ as in Theorem 1, and write $\delta := n-d$ and $D := \sqrt{\delta^2+4\lambda S_1}$, so that d/D is the leading curve in Eq. (5).

We first expand df_1 and df_2 in the small ratios $x_k := \kappa/\mu_k$. Since $C_k = 1/(1+x_k)$,

$$\begin{aligned}
\mathrm{df}_1 &= \sum_k \frac{1}{1+x_k} \\
&= \sum_k \Big(1 - x_k + \frac{x_k^2}{1+x_k}\Big) \\
&= d - \kappa S_1 + R_1, \\
\mathrm{df}_2 &= \sum_k \frac{1}{(1+x_k)^2} \\
&= \sum_k \Big(1 - 2x_k + \frac{x_k^2(3+2x_k)}{(1+x_k)^2}\Big) \\
&= d - 2\kappa S_1 + R_2,
\end{aligned} \tag{20}$$

where

$$R_1 := \sum_k \frac{x_k^2}{1+x_k}, \qquad R_2 := \sum_k \frac{x_k^2(3+2x_k)}{(1+x_k)^2}. \tag{21}$$

The second equalities in Eq. (20) are exact identities, checked by multiplying out, and the last ones use $\sum_k x_k = \kappa \sum_k \mu_k^{-1} = \kappa S_1$. The remainders are small. Each of their terms is nonnegative, and since $1+x_k \ge 1$ and $3+2x_k \le 3(1+x_k)^2$, the terms of R_1 and R_2 are bounded by x_k^2 and $3x_k^2$.

Since $\mu_k \geq \mu_{\min}$, moreover, $\sum_k x_k^2 = \kappa^2 \sum_k \mu_k^{-2} \leq \kappa^2 \mu_{\min}^{-1} \sum_k \mu_k^{-1}$. Hence

$$0 \leq R_1 \leq \frac{\kappa^2 S_1}{\mu_{\min}}, \qquad 0 \leq R_2 \leq \frac{3\kappa^2 S_1}{\mu_{\min}}. \tag{22}$$

We now write the noise risk in Eq. (4), the label-noise part of the prediction error, exactly in terms of the remainders,

$$\begin{aligned} \frac{\mathcal{V}}{\sigma^2} &= \frac{\mathrm{df}_2}{n - \mathrm{df}_2} \\ &= \frac{d - 2\kappa S_1 + R_2}{(n - \mathrm{df}_1) + (\mathrm{df}_1 - \mathrm{df}_2)} \\ &= \frac{d - 2\kappa S_1 + R_2}{\lambda/\kappa + \kappa S_1 + R_1 - R_2} \\ &= \frac{d - 2\kappa S_1 + R_2}{\sqrt{(\delta - R_1)^2 + 4\lambda S_1} + R_1 - R_2}. \end{aligned} \tag{23}$$

The first line is Eq. (4). The second uses Eq. (20) in the numerator. The third uses the fixed point $\mathrm{df}_1 + \lambda/\kappa = n$ for $n - \mathrm{df}_1$ and Eq. (20) for $\mathrm{df}_1 - \mathrm{df}_2$. The last follows from

$$\begin{aligned} \left(\frac{\lambda}{\kappa} + \kappa S_1\right)^2 &= \left(\frac{\lambda}{\kappa} - \kappa S_1\right)^2 + 4\lambda S_1 \\ &= (\delta - R_1)^2 + 4\lambda S_1, \end{aligned} \tag{24}$$

where the first equality holds because the product of λ/κ and κS_1 is λS_1, and the second because the fixed point with Eq. (20) reads $n = d - \kappa S_1 + R_1 + \lambda/\kappa$, that is, $\lambda/\kappa - \kappa S_1 = \delta - R_1$. We take the positive root because $\lambda/\kappa + \kappa S_1 > 0$. If R_1, R_2, and $\kappa S_1/d$ were zero, the right side of Eq. (23) would be d/D, the leading curve.

Finally we show that R_1, R_2, and κS_1 are small compared with D and d. By Eq. (22) this holds when κ is small, so we first show $\kappa < a\kappa_0$ with $a := 4L + 2$. The left side $\mathrm{df}_1(\kappa) + \lambda/\kappa$ of the fixed point decreases strictly in κ and equals n at the solution, so it suffices to show that it is below n at the trial point $\kappa = a\kappa_0$. Assuming $a\kappa_0 \leq \mu_{\min}/2$, we have

$$\begin{aligned} \mathrm{df}_1(a\kappa_0) + \frac{\lambda}{a\kappa_0} - n &= \frac{\lambda}{a\kappa_0} - a\kappa_0 S_1 + R_1 - \delta \\ &\leq \frac{\lambda}{a\kappa_0} - \frac{a\kappa_0 S_1}{2} - \delta \\ &= \left(\frac{1}{a} - \frac{a}{2}\right)\sqrt{\lambda S_1} - \delta \\ &\leq \left(\frac{1}{a} - \frac{a}{2} + 2L\right)\sqrt{\lambda S_1} \\ &< 0. \end{aligned} \tag{25}$$

The first line uses Eq. (20) and $n - d = \delta$. The second line uses Eq. (22) in the form $R_1 \leq (a\kappa_0)^2 S_1/\mu_{\min} \leq a\kappa_0 S_1/2$. The third line uses $\lambda/\kappa_0 = \kappa_0 S_1 = \sqrt{\lambda S_1}$. The fourth line uses $\delta \geq -2L\sqrt{\lambda S_1}$ on the peak region, and the last holds because $1/a - a/2 + 2L = 1/a - 1 < 0$ for $a = 4L + 2$. Hence $\kappa < a\kappa_0$.

With this bound, the remainders satisfy, for $j = 1, 2$,

$$\begin{aligned} R_j &\leq \frac{3a^2\kappa_0^2 S_1}{\mu_{\min}} \\ &= 3a^2 \frac{\kappa_0}{\mu_{\min}} \sqrt{\lambda S_1} \\ &\leq \frac{3a^2}{2} \frac{\kappa_0}{\mu_{\min}} D. \end{aligned} \tag{26}$$

The first line uses Eq. (22) with $\kappa < a\kappa_0$, the second $\kappa_0 S_1 = \sqrt{\lambda S_1}$, and the third $D \geq 2\sqrt{\lambda S_1}$. The denominator of Eq. (23) then satisfies

$$\begin{aligned}\left|\sqrt{(\delta - R_1)^2 + 4\lambda S_1} + R_1 - R_2 - D\right| &\leq 2R_1 + R_2 \\ &\leq \frac{9a^2}{2}\,\frac{\kappa_0}{\mu_{\min}}\,D,\end{aligned} \tag{27}$$

where the first line holds because $\sqrt{\delta^2 + 4\lambda S_1}$ is 1-Lipschitz in δ and the second uses Eq. (26). The numerator satisfies

$$\begin{aligned}0 &\leq d - \mathrm{df}_2 \\ &\leq 2\kappa S_1 \\ &\leq 2a\kappa_0 S_1 \\ &\leq 2a\frac{\kappa_0}{\mu_{\min}}\,d,\end{aligned} \tag{28}$$

where the first line uses $C_k \leq 1$, the second Eq. (20) with $R_2 \geq 0$, the third $\kappa < a\kappa_0$, and the last $S_1 \leq d/\mu_{\min}$. The denominator is thus D and the numerator d, each up to a relative error $O(\kappa_0/\mu_{\min})$ with constants of order $a^2 \sim L^2$. Dividing,

$$\frac{\mathcal{V}(n)}{\sigma^2} = \frac{d}{\sqrt{(n-d)^2 + 4\lambda S_1}}\Big(1 + O\Big(\frac{\kappa_0}{\mu_{\min}}\Big)\Big) \qquad \text{for } |n - d| \leq 2L\sqrt{\lambda S_1}, \tag{29}$$

which is Eq. (5).

It remains to prove the second part of Theorem 1, which holds for any two spectra, not only for power-law spectra. Consider two covariances of the same dimension d, labeled A and B, with positive eigenvalues $\mu_k^{(A)}$ and $\mu_k^{(B)}$, ridge penalties λ_A and λ_B, and $S_1^{(A)} = \sum_k 1/\mu_k^{(A)}$ and $S_1^{(B)} = \sum_k 1/\mu_k^{(B)}$. The second part states that A and B have the same leading noise curve near interpolation if and only if $\lambda_A S_1^{(A)} = \lambda_B S_1^{(B)}$. By the first part, the leading curve of a spectrum with ridge penalty λ is $d/\sqrt{(n-d)^2 + 4\lambda S_1}$, which depends on the spectrum and the ridge penalty only through the product λS_1. If $\lambda_A S_1^{(A)} = \lambda_B S_1^{(B)}$, the two leading curves are therefore the same function of n. Conversely, the two leading curves are the same only if the products are equal, because their values at $n = d$, $d/(2\sqrt{\lambda_A S_1^{(A)}})$ and $d/(2\sqrt{\lambda_B S_1^{(B)}})$, coincide only when $\lambda_A S_1^{(A)} = \lambda_B S_1^{(B)}$. For power-law spectra, take A to be the spectrum $\mu_k = k^{-\alpha}$ with ridge penalty λ_{eq} and B to be the spectrum $\mu_k = k^{-(\alpha+\Delta\alpha)}$ with ridge penalty λ. The condition then reads $\lambda_{\mathrm{eq}} S_1(\alpha) = \lambda S_1(\alpha + \Delta\alpha)$, whose unique solution is λ_{eq} in Eq. (7).

C Continuum limit and peak criterion

This appendix derives the continuum equations in Eq. (11) and proves Theorem 2. The theorem concerns the whole noise curve, not only the peak region of Theorem 1, and in the continuum limit the whole curve is given by integrals of one variable τ.

We first derive the continuum equations in Eq. (11). Let $\mu_k = k^{-\alpha}$, so that the smallest eigenvalue is $\mu_d = d^{-\alpha}$, and measure the effective ridge in units of it, $\tau := \kappa d^{\alpha}$. This is a change of variables, valid at every d. Then $\kappa/\mu_k = \tau d^{-\alpha} k^{\alpha} = \tau(k/d)^{\alpha}$, so $C_k = 1/\{1 + \tau(k/d)^{\alpha}\}$ and

$$\begin{aligned}\frac{\mathrm{df}_j}{d} &= \frac{1}{d}\sum_{k=1}^{d} C_k^j \\ &= \frac{1}{d}\sum_{k=1}^{d}\big(1 + \tau(k/d)^{\alpha}\big)^{-j} \\ &=: J_{j,d}(\tau).\end{aligned} \tag{30}$$

Dividing the fixed point by d and the risk in Eq. (4) by d in the numerator and the denominator,

$$\begin{aligned} u &= \frac{n}{d} \\ &= \frac{1}{d}\sum_k C_k + \frac{\lambda}{d\kappa} \\ &= J_{1,d}(\tau) + \frac{r}{\tau}, \\ \frac{\mathcal{V}}{\sigma^2} &= \frac{\mathrm{df}_2}{n - \mathrm{df}_2} \\ &= \frac{\mathrm{df}_2/d}{u - \mathrm{df}_2/d} \\ &= \frac{J_{2,d}(\tau)}{u - J_{2,d}(\tau)}. \end{aligned} \tag{31}$$

Both last lines use Eq. (30), and the first chain also uses $\lambda/(d\kappa) = \lambda d^{\alpha-1}/\tau = r/\tau$ with $r = \lambda d^{\alpha-1}$ from Eq. (10). Equation (31) is exact at every d.

The sums $J_{j,d}(\tau)$ become Riemann sums of $J_j(\tau)$ and converge to $J_j(\tau)$ for every $\tau > 0$ as $d \to \infty$. Both $J_{1,d}(\tau) + r/\tau$ and $J_1(\tau) + r/\tau$ decrease strictly and continuously in τ from infinity to zero, so each takes the value $u > 0$ at exactly one τ. Because of this monotonicity, the pointwise convergence implies that, at fixed u and r, the solution τ of Eq. (31) converges to that of Eq. (11), and the noise risk converges with it.

Proof of Theorem 2. The proof compares the curve $\mathcal{V}_\infty$ with its limits $\sigma^2(\alpha - 1)$ as $u \to 0$ (for $\alpha > 1$) and 0 as $u \to \infty$. When the peak criterion in Eq. (12) holds, the curve rises above $\sigma^2(\alpha - 1)$ and therefore has an interior maximum. When it fails, the curve stays below this value and decreases monotonically.

In this proof we write $\mathcal{V}$ for $\mathcal{V}_\infty$ and J_j for $J_j(\tau)$. Equation (11) gives the curve parametrically, with τ running over $(0, \infty)$, and we use three facts about this parametrization. First, $u = J_1 + r/\tau$ decreases strictly in τ, from infinity as $\tau \to 0$ to zero as $\tau \to \infty$, so u and τ correspond one to one, and small u corresponds to large τ. Second, writing $G := u - J_2$ for the denominator of the noise risk in Eq. (11),

$$\frac{\mathcal{V}}{\sigma^2} = \frac{J_2}{G}, \qquad G = J_1 - J_2 + \frac{r}{\tau} > 0,$$

where the expression for G uses $u = J_1 + r/\tau$, and $G > 0$ because $0 < (1 + \tau x^\alpha)^{-1} \le 1$ gives $J_1 \ge J_2$. Third, the curve vanishes at large u, $\mathcal{V} \to 0$ as $u \to \infty$, because then $\tau \to 0$, $J_2 \to 1$, and $G \ge r/\tau \to \infty$.

We now compare the curve with the value $\sigma^2(\alpha - 1)$, which is its limit as $u \to 0$, as this paragraph shows. The comparison uses identities between the integrals. With $\vartheta := (1 + \tau)^{-1}$,

$$\alpha J_2 = \vartheta + (\alpha - 1)J_1, \qquad 2\alpha J_3 = \vartheta^2 + (2\alpha - 1)J_2. \tag{32}$$

These are the cases $j = 1, 2$ of $\vartheta^j = (1 - j\alpha)J_j + j\alpha J_{j+1}$, which follows by integrating

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left\{x(1 + \tau x^\alpha)^{-j}\right\} &= (1 + \tau x^\alpha)^{-j} - j\alpha\, \tau x^\alpha (1 + \tau x^\alpha)^{-j-1} \\ &= (1 - j\alpha)(1 + \tau x^\alpha)^{-j} + j\alpha (1 + \tau x^\alpha)^{-j-1} \end{aligned}$$

over $x \in [0, 1]$. The first line is the product rule, the second writes $\tau x^\alpha = (1 + \tau x^\alpha) - 1$, and the integral of the left side is $(1 + \tau)^{-j} = \vartheta^j$. The second identity is used only for the slope below. The

first gives the difference of the curve from $\sigma^2(\alpha-1)$,

$$\begin{aligned}\frac{\mathcal{V}}{\sigma^2}-(\alpha-1) &= \frac{J_2-(\alpha-1)(J_1-J_2+r/\tau)}{G}\\ &= \frac{\alpha J_2-(\alpha-1)J_1-(\alpha-1)r/\tau}{G}\\ &= \frac{\vartheta-(\alpha-1)r/\tau}{G}\\ &= \frac{\tau\vartheta-(\alpha-1)r}{\tau G}.\end{aligned} \tag{33}$$

The first line inserts $\mathcal{V}/\sigma^2 = J_2/G$ and $G = J_1 - J_2 + r/\tau$, the second collects the terms, the third uses the first identity in Eq. (32), and the last multiplies the numerator and the denominator by τ. The sign of this difference is that of $\tau\vartheta-(\alpha-1)r$, where $\tau\vartheta = \tau/(1+\tau)$ increases from 0 to 1. Since $\tau\vartheta < 1$ for every τ and $\tau\vartheta \to 1$ as $\tau\to\infty$, the difference is positive at large τ when $(\alpha-1)r<1$ and negative for every τ when $(\alpha-1)r \geq 1$. As $\tau\to\infty$, the numerator tends to the finite value $1-(\alpha-1)r$, so the difference tends to zero if the denominator τG tends to infinity. This holds because $\tau G \geq \tau(J_1-J_2)$ and $\tau(J_1-J_2)\to\infty$, which follows from the substitution $y=\tau^{1/\alpha}x$,

$$\begin{aligned}\tau(J_1-J_2) &= \tau\int_0^1 \frac{\tau x^\alpha}{(1+\tau x^\alpha)^2}\,\mathrm{d}x\\ &= \tau^{1-1/\alpha}\int_0^{\tau^{1/\alpha}} \frac{y^\alpha}{(1+y^\alpha)^2}\,\mathrm{d}y\\ &\geq \frac{\tau^{1-1/\alpha}}{4(\alpha+1)} \qquad (\tau\geq 1),\end{aligned}$$

where the bound keeps $0\leq y\leq 1$, on which $(1+y^\alpha)^2\leq 4$, and $\tau^{1-1/\alpha}\to\infty$ requires $\alpha>1$. Hence $\mathcal{V}\to\sigma^2(\alpha-1)$ as $u\to 0$, in agreement with Eq. (35) of Cheng & Montanari (2024) in the limit of zero ridge penalty.

When Eq. (12) holds, the curve therefore has an interior maximum. For all large τ we have $\tau\vartheta > (\alpha-1)r$, so by Eq. (33) there is a sample ratio u_0 with $\mathcal{V}(u_0)>\sigma^2(\alpha-1)$. The limits of $\mathcal{V}$ at both ends, $\sigma^2(\alpha-1)$ as $u\to 0$ and 0 as $u\to\infty$, are smaller than $\mathcal{V}(u_0)$. Hence there are $u_a<u_0<u_b$ with $\mathcal{V}<\mathcal{V}(u_0)$ outside $[u_a,u_b]$. On the closed interval $[u_a,u_b]$ the continuous function $\mathcal{V}$ attains its maximum, which is at least $\mathcal{V}(u_0)$ and therefore the maximum over all $u>0$, attained at an interior point.

When Eq. (12) fails, $(\alpha-1)r\geq 1>\tau\vartheta$, so Eq. (33) is negative for every τ and the curve stays below its limit $\sigma^2(\alpha-1)$ as $u\to 0$. This alone does not exclude a local maximum, so we show that the curve decreases monotonically, by computing its slope through τ. The integrals satisfy

$$\tau\frac{\mathrm{d}J_j}{\mathrm{d}\tau} = -j\,(J_j - J_{j+1}), \tag{34}$$

which follows by integrating over $x\in[0,1]$

$$\begin{aligned}\tau\frac{\partial}{\partial\tau}(1+\tau x^\alpha)^{-j} &= -j\,\tau x^\alpha(1+\tau x^\alpha)^{-j-1}\\ &= -j\big\{(1+\tau x^\alpha)^{-j}-(1+\tau x^\alpha)^{-j-1}\big\},\end{aligned}$$

where the second line writes $\tau x^\alpha = (1+\tau x^\alpha)-1$. With primes denoting derivatives in τ,

$$\begin{aligned}\frac{\mathrm{d}(\mathcal{V}/\sigma^2)}{\mathrm{d}u} &= \frac{\tau\,\mathrm{d}(J_2/G)/\,\mathrm{d}\tau}{\tau\,\mathrm{d}u/\,\mathrm{d}\tau}\\ &= \frac{\tau J_2' G - J_2\,\tau G'}{G^2\cdot\tau u'}\\ &= \frac{-2(J_2-J_3)(u-J_2)-J_2(-u+3J_2-2J_3)}{-G^3}\\ &= -\frac{u(2J_3-J_2)-J_2^2}{G^3}\\ &=: -\frac{F}{G^3}.\end{aligned} \tag{35}$$

The first line is the chain rule and the second the quotient rule. The third uses Eq. (34) in the forms $\tau J_2' = -2(J_2 - J_3)$, $\tau u' = -(J_1 - J_2) - r/\tau = -G$, and $\tau G' = \tau u' - \tau J_2' = -u + 3J_2 - 2J_3$. The fourth expands the numerator, and the last defines F. The curve decreases where $F > 0$, and when Eq. (12) fails, F is positive everywhere,

$$\begin{aligned}
\alpha F &= \Big(J_1 + \frac{r}{\tau}\Big)\big\{\vartheta^2 + (\alpha-1)J_2\big\} - J_2\big\{\vartheta + (\alpha-1)J_1\big\} \\
&= \frac{r}{\tau}\big\{\vartheta^2 + (\alpha-1)J_2\big\} - \vartheta J_2 + \vartheta^2 J_1 \\
&\geq \frac{\vartheta^2}{(\alpha-1)\tau} + \Big(\frac{1}{\tau} - \vartheta\Big)J_2 + \vartheta^2 J_1 \\
&> 0.
\end{aligned} \tag{36}$$

The first line uses $u = J_1 + r/\tau$, the second identity in Eq. (32) in the form $\alpha(2J_3 - J_2) = \vartheta^2 + (\alpha-1)J_2$, and the first in the form $\alpha J_2^2 = J_2\{\vartheta + (\alpha-1)J_1\}$. The second line cancels $(\alpha-1)J_1J_2$. The third uses $r \geq 1/(\alpha-1)$, which holds whenever Eq. (12) fails, including the case $(\alpha-1)r = 1$. The last holds because all three terms are positive. The first is positive because $\alpha > 1$, the second because $1/\tau - \vartheta = 1/\{\tau(1+\tau)\} > 0$, and the third because $J_1 > 0$. Hence $\mathcal{V}$ strictly decreases in u and has no interior maximum, which completes the proof for $\alpha > 1$.

It remains to prove the last two statements of Theorem 2, which concern $0 \leq \alpha \leq 1$. For $0 \leq \alpha < 1$, the curve has an interior maximum for every $r > 0$, because it tends to zero at small as well as at large sample ratios. As $u \to \infty$, this was shown above. As $u \to 0$, $\tau \to \infty$, and the integrands of τJ_1 and τJ_2 are bounded by $x^{-\alpha}$, which is integrable on $[0,1]$ for $\alpha < 1$. They converge pointwise to $x^{-\alpha}$ and to zero, so dominated convergence gives $\tau J_1 \to 1/(1-\alpha)$ and $\tau J_2 \to 0$, and hence $\mathcal{V}/\sigma^2 = \tau J_2/(\tau J_1 - \tau J_2 + r) \to 0$. The limits of the curve as $u \to 0$ and as $u \to \infty$ are thus smaller than its value at any $u_0 > 0$, and the argument used when Eq. (12) holds gives an interior maximum.

For $\alpha = 1$, the sign of the slope in Eq. (35) can be read off explicitly, which shows that the maximum is unique and lies at $u = 1$. The integral $J_2 = \int_0^1 (1+\tau x)^{-2}\,\mathrm{d}x = 1/(1+\tau) = \vartheta$ and the second identity in Eq. (32) give $2J_3 - J_2 = \vartheta^2 = J_2^2$. Then $F = (u-1)J_2^2$ by Eq. (35), so $\mathcal{V}$ increases for $u < 1$ and decreases for $u > 1$, and the unique maximum is at $u = 1$. □

The limit $\sigma^2(\alpha-1)$ of $\mathcal{V}_\infty$ as $u \to 0$ depends on α but not on r. Hence no choice of the ridge penalty can make two curves $\mathcal{V}_\infty$ with different exponents coincide.

Corollary C.1 (No ridge change reproduces an exponent change over the whole curve)**.** *Let $\mathcal{V}_{\infty,\alpha,r}(u)$ denote the curve $\mathcal{V}_\infty(u)$ of Eq. (11) with exponent α and rescaled ridge penalty r. Let $\alpha_B > \alpha_A > 1$ and $\sigma^2 > 0$. Then, for all $r_A, r_B > 0$,*

$$\lim_{u\to 0}\big\{\mathcal{V}_{\infty,\alpha_B,r_B}(u) - \mathcal{V}_{\infty,\alpha_A,r_A}(u)\big\} = \sigma^2(\alpha_B - \alpha_A) > 0.$$

In particular, no choice of the ridge penalties makes the two curves coincide for all $u > 0$.

Proof. For every $\alpha > 1$ and $r > 0$, $\mathcal{V}_{\infty,\alpha,r}(u) \to \sigma^2(\alpha-1)$ as $u \to 0$, by Eq. (33) and the bound $\tau G \to \infty$ that follows it in the proof of Theorem 2. Subtracting the two limits gives $\sigma^2(\alpha_B - \alpha_A)$, which is positive because $\alpha_B > \alpha_A$. □

Finite spectra. Theorem 2 and Corollary C.1 concern the continuum limit, and their small-sample behavior does not carry over to a finite spectrum. In the continuum limit, $\mathcal{V}_\infty$ tends to $\sigma^2(\alpha-1)$ as $u \to 0$. For a finite spectrum and $\lambda > 0$, in contrast, the noise risk tends to zero as $n \downarrow 0$. Indeed, the fixed point then gives $\kappa \sim (\sum_k \mu_k + \lambda)/n$, so $C_k \sim \mu_k/\kappa$ and $\mathcal{V} \sim \sigma^2 n \sum_k \mu_k^2/(\sum_k \mu_k + \lambda)^2 \to 0$. The noise risk also tends to zero as $n \to \infty$, since $\kappa \to 0$ and $\mathcal{V}/\sigma^2 = \mathrm{df}_2/(n - \mathrm{df}_2) \leq d/(n-d)$. Since the noise risk is continuous and positive, the noise curve at finite d therefore always has a maximum at some real $n > 0$, as stated in Section 5. This maximum need not lie at an integer sample size and can lie below $n = 1$.

D Details of the experiments

This appendix gives the details of the experiments. Appendix D.1 describes the Gaussian experiment of Section 6, and Appendix D.2 compares four MNIST representations.

In all MNIST experiments, the images are the 784 pixel values scaled to $[0, 1]$, and a pool of $50{,}000$ images is drawn once with seed 0. The exponent of a power-law fit is the least-squares slope of $\log \mu_k$ against $\log k$ for $k = 5, \ldots, 150$. The random subsamples use seed 2 in Section 5 (40 draws) and seed 1 in Section 6 and Appendix D.2 (60 draws).

D.1 Gaussian inputs

The Gaussian experiment repeats the setting of Figure 1 ($d = 300$, $\alpha = 2$, $\lambda = 10^{-4}$, $\sigma^2 = 1$) with five settings: the baseline, and the power-law exponent change and the ridge change for $\Delta\alpha \in \{0.1, 0.5\}$. The sample grid is $u = 0.35, 0.40, \ldots, 1.60$ with integer n. Each of the 400 draws, seeded at 20260916, uses one standard Gaussian matrix for all settings and the first n rows for n samples, so contrasts between settings are paired and values at different sample ratios are correlated. The conditional noise trace in Eq. (2) averages over the label noise exactly. DE values are computed with the fixed-point root tolerance 2×10^{-13}. Intervals are the means plus or minus 1.96 standard errors at each sample ratio. For a difference between settings, the standard error is computed from the differences within each draw. The intervals are pointwise and are not simultaneous confidence bands over the curve.

The measurements agree with the DE. Across the five settings, the largest relative difference between the measured noise risk and the DE is 0.86%. The measured shape effect is positive at all 26 sample ratios for both exponent changes, and every interval lies above zero. For $\Delta\alpha = 0.1$, the exponent change lowers the measured noise risk at $n = d$ by 1.1506, and the paired shape effect is 0.0239. The measured shape effects differ from their DE values by about 2%, more than the sampling intervals. For $\Delta\alpha = 0.5$, the DE curves of the exponent change and the ridge change have their maxima at $u = 0.516$ and 0.732, and the measured curves show the same separation. The weak-ridge parameter $\kappa_0/\mu_{\min}$ is 0.40 and 1.34 for the two changed spectra, so these settings are not deep in the weak-ridge regime of Theorem 1. The experiment therefore tests the equivalence at finite parameters, not the convergence to the leading curve, which Theorem 1 establishes within the DE.

D.2 Different MNIST representations

This experiment tests Theorem 1 on representations that differ in more than the spectrum. We compare four 300-dimensional representations of MNIST, namely the leading principal components of the pixels, random Fourier features at bandwidths 1 and 2, and random ReLU features. The random Fourier features are $\sqrt{2}\cos(w_j^\top x + b_j)$ with $w_j \sim \mathcal{N}(0, I/(784s^2))$ and b_j uniform on $[0, 2\pi]$, which approximate the Gaussian radial basis function (RBF) kernel $\exp(-\|x - x'\|^2/(2 \cdot 784\, s^2))$ of bandwidth s (Rahimi & Recht, 2007). The random ReLU features are $\max(w_j^\top x + b_j, 0)$ with $w_j \sim \mathcal{N}(0, I/m^2)$, where m^2 is the mean of $\|x\|^2$ over the pool, and b_j uniform on $[-1, 1]$. Each representation is centered and scaled over the pool. The exponents of power-law fits to their spectra are 1.48, 1.56, 1.63, and 1.36. Unlike the data of Section 6, whose principal components are rescaled, these representations also differ in their eigenvectors and distributions.

Theorem 1 predicts that representations with equal λS_1 have nearly the same noise peak when the ridge is weak. For the three representations that satisfy the weak-ridge condition ($\kappa_0/\mu_{\min} \le 0.38$), the noise risk at $n = d$ differs by 67% under a common λ and by 6.8% under equal λS_1 (Figure 4). The measured values lie 8 to 15% above the leading curve of Theorem 1, and at equal λS_1 the noise risk increases with the fitted exponent at every $u \le 1$. The random ReLU features have an isolated small eigenvalue, so $\kappa_0/\mu_{\min} = 7.0$ violates the weak-ridge condition, and their curve deviates from the other three.

E Teacher signal and total risk

This appendix adds the bias term of Section 3 to the noise risk studied in the main text and shows that a lower noise risk need not mean a lower prediction error. The effect is the task–model alignment

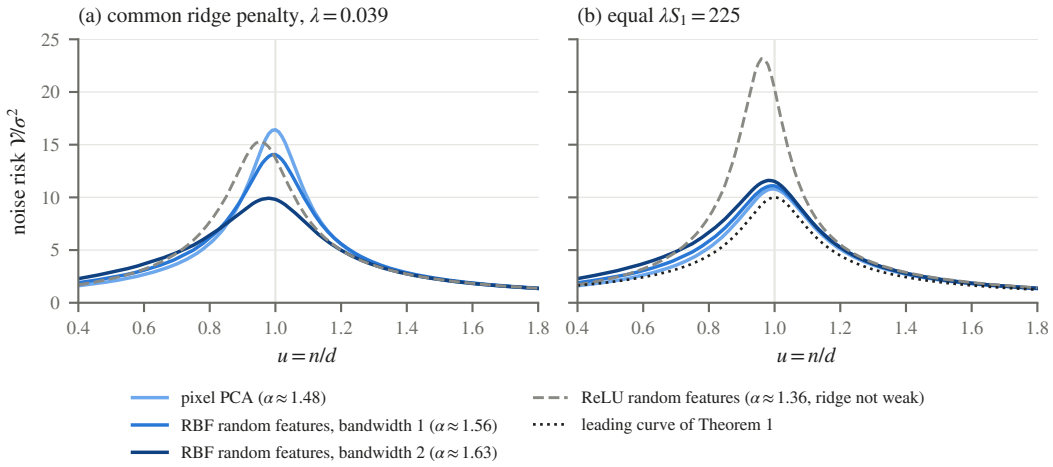

Figure 4: Noise curves of four 300-dimensional MNIST representations. (a) A common ridge penalty λ. (b) Ridge penalties chosen so that λS_1 is the same for all four representations. The black dotted line is the leading curve $d/\sqrt{(n-d)^2 + 4\lambda S_1}$ of Theorem 1. The random ReLU features violate the weak-ridge condition.

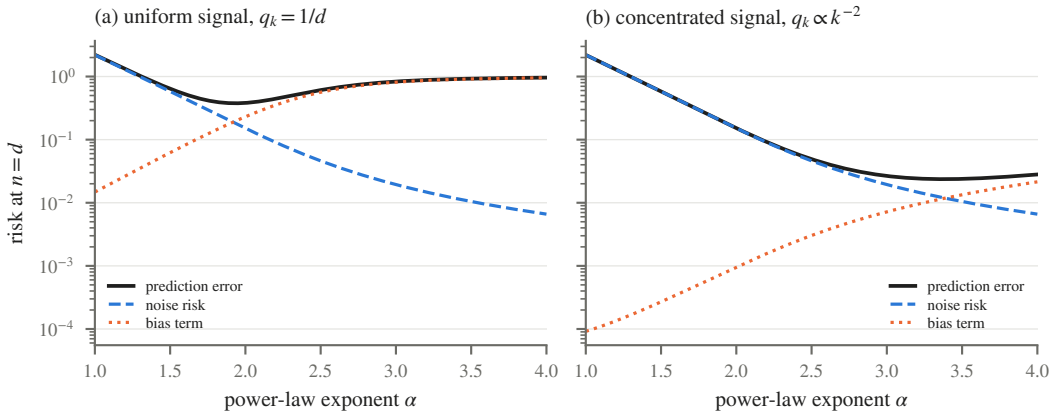

Figure 5: Prediction error at $n = d$ as the power-law exponent changes, with the signal power q_k in each direction held fixed ($d = 300$, $\lambda = 10^{-3}$, $\sigma^2 = 0.1$). Raising the exponent lowers the noise risk (dashed) and raises the bias term (dotted) in both panels. (a) For a signal spread uniformly over the directions, the prediction error (solid) is smallest at $\alpha \approx 1.9$ and rises beyond it. (b) For a signal concentrated on the directions with large eigenvalues, the prediction error falls with the noise risk up to $\alpha \approx 3.4$.

of Canatar et al. (2021), seen here through a change of the power-law exponent. It supports the discussion in Section 7.

With a teacher w^*, the DE of the prediction error is

$$\mathcal{R} = \frac{nB + \sigma^2 \mathrm{df}_2}{n - \mathrm{df}_2}, \qquad B := \sum_k q_k (1 - C_k)^2, \qquad q_k := \mu_k (w_k^*)^2, \tag{37}$$

where q_k is the signal power in direction k (Canatar et al., 2021). The part proportional to σ^2 is the noise risk (4), and the rest, $nB/(n - \mathrm{df}_2)$, is the DE of the bias term.

We change α while keeping the signal power in each direction fixed, with $\sum_k q_k = 1$. Figure 5 compares a signal spread uniformly over the directions, $q_k = 1/d$, with a signal concentrated on the directions with large eigenvalues, $q_k \propto k^{-2}$. Raising the exponent lowers the noise risk in both

cases and raises the bias term, as a larger ridge penalty would. For the uniform signal, most of the signal lies in directions with small eigenvalues, which the estimator shrinks more strongly at larger exponents, so the prediction error is smallest at $\alpha \approx 1.9$ and rises beyond it. For the concentrated signal, the bias term stays small, and the prediction error falls with the noise risk up to $\alpha \approx 3.4$. Whether a larger exponent lowers the prediction error therefore depends on how the signal is aligned with the eigenvectors, as in Canatar et al. (2021).

F Correspondence of regularization conventions

This appendix translates between our convention and that of Rizzi et al. (2026), the closest related work, so that the two sets of results can be compared (Section 2). The two works differ in how they normalize the spectrum and in which ridge penalty they hold fixed as the dimension grows.

Rizzi et al. (2026) normalize the spectrum. For the pure linear kernel $h_1 x^\top x'$ with $h_1 > 0$, their eigenvalues are $(h_1/Z_d)\mu_k$ with $Z_d := \sum_{k=1}^d k^{-\alpha}$, whereas ours are $\mu_k = k^{-\alpha}$. Since multiplying the covariance and the ridge penalty by the same constant leaves the DE unchanged (Section 4), our ridge penalty λ corresponds exactly to their sum-loss penalty $\lambda_{\mathrm{R}} = h_1\lambda/Z_d$. Table 1 lists the penalty λ_{R} that keeps each quantity of this paper fixed: λ in Figures 1 and 2, λS_1 in Theorem 1, and r in Theorem 2.

Table 1: Ridge penalty λ_{R} of Rizzi et al. (2026) that keeps each quantity of this paper fixed, for the pure linear kernel with fixed h_1. Here $S_1 = \sum_{k=1}^d k^\alpha$.

Quantity held fixed	Required penalty λ_{R}
λ	$\lambda_{\mathrm{R}} = h_1\lambda/Z_d$
λS_1	$\lambda_{\mathrm{R}} = h_1\lambda S_1/(Z_d S_1)$
r	$\lambda_{\mathrm{R}} = h_1 r d^{1-\alpha}/Z_d$

The correspondence shows where the two analyses agree and where they differ. For $0 \le \alpha < 1$, the continuum equations of Rizzi et al. (2026) (Lemma 4.1 and Theorem 1, specialized to degree one) coincide with Eq. (11) under $r = \lambda_{\mathrm{R}}/(h_1(1-\alpha))$. The two works thus use the same equations, but holding λ_{R} fixed while changing α changes r, so they follow different paths through them. For $\alpha > 1$, $Z_d \to \zeta(\alpha)$, so a fixed r requires λ_{R} to decrease in proportion to $d^{1-\alpha}$. The positive-ridge case of Proposition 4 in Rizzi et al. (2026) instead keeps λ_{R} of order one as $d \to \infty$, which corresponds to $r \to \infty$. It therefore lies above the threshold of Theorem 2 and does not contain the threshold itself, which occurs at finite r. At $\alpha = 1$, a fixed r requires λ_{R} to behave as $h_1 r/\log d$ for large d.

G Coordinate identities and ridgeless regimes

This appendix collects exact statements on how the spectrum interacts with the ridge penalty, including the ridgeless case $\lambda = 0$, which the main text excludes. They place the results for $\lambda > 0$ in context. For every $\lambda > 0$, changing the exponent is equivalent to replacing the ridge penalty by a penalty with a different weight on each coordinate, and Theorem 1 shows that near the peak this weighted penalty acts like a single ridge penalty. Without a ridge penalty, the spectrum does not act as a penalty at all. Table 2 summarizes these statements, and the paragraphs below derive the identities.

A change of variables moves a change of the power-law exponent from the input data to the ridge penalty, which gives the first and third rows of Table 2. Write $X_\alpha = Z\,\mathrm{diag}(k^{-\alpha/2})$ for a fixed Z, so that $X_{\alpha+\Delta\alpha} = X_\alpha U$ with $U := \mathrm{diag}(k^{-\Delta\alpha/2})$. With the substitution $\theta := Uw$, for any Z and y,

$$\|y - X_{\alpha+\Delta\alpha} w\|^2 + \lambda\|w\|^2 = \|y - X_\alpha\theta\|^2 + \lambda\sum_k k^{\Delta\alpha}\theta_k^2. \tag{38}$$

The exponent change thus turns the ridge penalty λ into the penalty $\lambda k^{\Delta\alpha}$ on coordinate k. For a test input x_α with spectrum α, the corresponding input with spectrum $\alpha + \Delta\alpha$ is Ux_α, and the predictions agree, $(Ux_\alpha)^\top w = x_\alpha^\top\theta$. A nonzero teacher must be transformed in the same way. At $\lambda = 0$ and $n < d$, the estimator is the minimum-norm interpolant, the limit of $\widehat{w}$ as $\lambda \downarrow 0$. The same

Table 2: The effect of the power-law exponent for different ridge penalties λ and sample sizes n. The covariance is positive definite and the label noise is nonzero. The input data are $X = Z\Sigma^{1/2}$, and a change of the exponent changes Σ with Z fixed. The last column states which result shows each entry, and whether it holds exactly for each draw of Z, for the Gaussian expectation, or for the DE.

Setting	Effect of the power-law exponent	Source
$\lambda > 0$	Changing the power-law exponent is equivalent to replacing the ridge penalty λ by the penalty $\lambda k^{\Delta\alpha}$ on the k-th coordinate.	Eq. (38) shows this exactly for each draw of Z.
small λ, n near d	This weighted penalty acts like a single ridge penalty.	Theorem 1 proves this for the DE, to leading order in $\kappa_0/\mu_{\min}$.
$\lambda = 0$, $n < d$	Changing the power-law exponent changes which interpolant the minimum-norm estimator returns, and adds no penalty.	Eq. (38) at $\lambda = 0$ shows this exactly for each draw of Z.
$\lambda = 0$, $n = d$	The noise risk is infinite for every power-law exponent, although the risk for each draw with invertible $X^\top X$ is finite.	For Gaussian inputs, Holzmüller (2021, Theorem 12) shows that the expected noise risk is infinite, and the DE also diverges as $n \uparrow d$.
$\lambda = 0$, $n > d$	The noise risk does not depend on the power-law exponent.	Eq. (39) shows this exactly for each draw of Z with full column rank.

substitution shows that the minimum-norm interpolant for $X_{\alpha+\Delta\alpha}$ corresponds to the interpolant of X_α with the smallest weighted norm $\|U^{-1}\theta\|$. The spectrum then changes which interpolant is selected rather than adding a penalty.

For $\lambda = 0$ and $n > d$, the noise risk does not depend on the spectrum. When X has full column rank, the estimator $\widehat{w} = RX^\top y$ at $\lambda = 0$ is the ordinary least-squares (OLS) estimator $(X^\top X)^{-1}X^\top y$. For $X = Z\Sigma^{1/2}$ with Z of full column rank, $(X^\top X)^{-1} = \Sigma^{-1/2}(Z^\top Z)^{-1}\Sigma^{-1/2}$, and cyclicity of the trace gives

$$\mathcal{V}_{\mathrm{OLS}} = \sigma^2 \mathbb{E}_Z \operatorname{Tr}[(Z^\top Z)^{-1}]. \tag{39}$$

Changing Σ while fixing the law of Z therefore leaves the OLS noise risk unchanged (Holzmüller, 2021, Remark G.1). For standard Gaussian entries, $\mathcal{V}_{\mathrm{OLS}} = \sigma^2 d/(n-d-1)$ for $n > d+1$, and the expectation is infinite for $n = d$ and $n = d+1$ (Holzmüller, 2021, Theorem 12), although each realization has finite conditional risk.

The DE shows the same picture at $\lambda = 0$. For $0 < n < d$, sending $\lambda \downarrow 0$ in Eq. (3) leaves a unique $\kappa(0^+) > 0$ with $n = \sum_k \mu_k/(\mu_k + \kappa(0^+))$. For $n > d$, instead $\kappa \downarrow 0$ and $\lambda/\kappa \to n - d$, which gives $\mathcal{V} = \sigma^2 d/(n-d)$. As $n \uparrow d$ in the ridgeless DE, $d - n \sim \kappa(0^+)S_1$ and $n - \mathrm{df}_2 \sim \kappa(0^+)S_1$, so $\mathcal{V} \sim \sigma^2 d/(d-n)$ diverges, whereas at a positive ridge penalty the DE is finite at $n = d$.