# More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize

**Alexander Wei** [1] **Wei Hu** [1] **Jacob Steinhardt** [1]

## Abstract

Of theories for why large-scale machine learning models generalize despite being vastly overparameterized, which of their assumptions are needed to capture the qualitative phenomena of generalization in the real world? On one hand, we find that most theoretical analyses fall short of capturing these qualitative phenomena even for kernel regression, when applied to kernels derived from large-scale neural networks (e.g., ResNet-50) and real data (e.g., CIFAR-100). On the other hand, we find that the classical GCV estimator (Craven and Wahba, 1978) accurately predicts generalization risk even in such overparameterized settings. To bolster this empirical finding, we prove that the GCV estimator converges to the generalization risk whenever a local random matrix law holds. Finally, we apply this random matrix theory lens to explain why pretrained representations generalize better as well as what factors govern scaling laws for kernel regression. Our findings suggest that random matrix theory, rather than just being a toy model, may be central to understanding the properties of neural representations in practice.

## 1. Introduction

The fact that deep neural networks trained with many more parameters than data points can generalize well contradicts conventional statistical wisdom (Zhang et al., 2017). This observation has inspired much theoretical work, with one of the goals being to explain the generalization and scaling behavior of such models. In this paper, we study how these theoretical perspectives map onto reality. What assumptions are necessary (or sufficient) to capture the qualitative phenomena (e.g., pretraining vs. random initialization, scaling laws) of large-scale models? And what do they reveal about generalization in the real world?

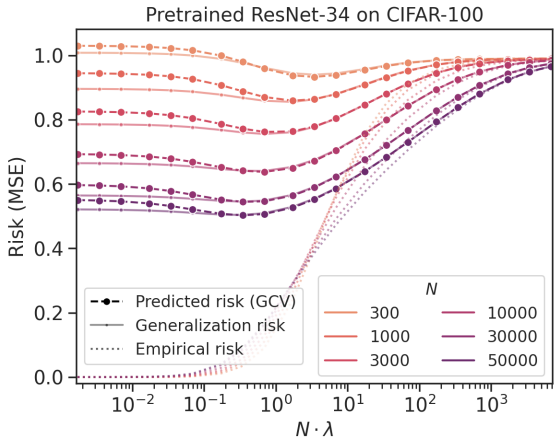


*Figure 1.* Predicted vs. actual generalization risk of a pretrained ResNet-34 empirical NTK on CIFAR-100 over dataset sizes $N$ and ridge regularizations $\lambda$. Corresponding training risks are plotted in the background. The fit achieving the lowest MSE has $19.9\%$ test error on CIFAR-100 (vs. $15.9\%$ from finetuning the ResNet).

An adequate theoretical treatment should at least predict the behavior of high-dimensional *linear* models. To assess this, we focus on linear models derived from neural representations (e.g., final layer activations or empirical neural tangent kernels) of large-scale networks on vision data. We test whether different theories can predict how kernel ridge regression on these representations generalizes, given only the training data.

In this setting of regression on realistic kernels, we find that most theoretical analyses already face severe challenges. A major difficulty is that the ground truth function has large—effectively infinite—kernel norm, which we verify empirically on several datasets. Consequently, norm-based generalization bounds are vacuous or even increase with dataset size, echoing concerns raised by Belkin et al. (2018) and Nagarajan & Kolter (2019). Other challenges for estimating generalization include the slow convergence of the empirical covariance matrix and the fact that noise and signal are indistinguishable in high-dimensional settings.

However, not all is lost. We find that the *generalized cross-validation (GCV) estimator* (Craven & Wahba, 1978) does accurately predict the generalization risk, even when typical norm- or spectrum-based formulas struggle. GCV is accu-

[1]UC Berkeley, Berkeley, California, USA. Correspondence to: Alexander Wei <awei@berkeley.edu>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

rate over a wide range of dataset sizes and regularization strengths, for classification tasks of varying complexities, and for representations extracted from residual networks both at random initialization and after pretraining. For instance, Figure 1 compares the GCV estimate against the true generalization risk for an ImageNet-pretrained ResNet-34 representation on CIFAR-100.

To justify the performance of the GCV estimator, we prove that it converges to the true generalization risk whenever a local random matrix law (Knowles & Yin, 2017) holds. Our analysis of this estimator allows for the highly anisotropic covariates and large-norm ground truth functions observed in our empirical setting. Along the way, we also generalize recent random matrix analyses of high-dimensional ridge regression (Hastie et al., 2020; Canatar et al., 2021; Wu & Xu, 2020; Jacot et al., 2020b; Loureiro et al., 2021; Richards et al., 2021; Mel & Ganguli, 2021; Simon et al., 2021) to this setting. Finally, our analysis provides a new perspective on this classical estimator that explains how its form arises in connection to random matrix theory.

We next apply this random matrix theory lens to explore basic questions about neural representations: Why do pretrained models generalize better than randomly initialized ones? And what factors govern the rates observed in neural scaling laws (Kaplan et al., 2020)? We find that alignment—how easy it is to represent the ground truth function in the eigenbasis (Marquardt & Snee, 1975; Caponnetto & Vito, 2007; Canatar et al., 2021)—is necessary to explain the performance of deep learning models. In particular, pretrained representations perform better than random representations due to better alignment, and *despite* worse eigenvalue decay. Finally, we provide sample-efficient methods to estimate the alignment and eigenvalue decay, which circumvent the slow convergence of the sample covariance matrix, and show that these two quantities are sufficient to predict the scaling law rate of ridge regression on natural data.

Our empirical findings and theoretical analysis show that a random matrix theoretic perspective stands apart at capturing the generalization of high-dimensional linear models on real data. More classical approaches, which often boil down to norms and/or eigendecay, do not suffice because generalization typically depends on the specific alignment between a high-norm ground truth function and the population covariance matrix. More broadly, our results suggest that accounting for random matrix effects is necessary to model the qualitative phenomena of deep learning—and in the case of kernel regression, sufficient.

*Remark.* In addition to our scientific contribution, we develop a library for computing large-scale empirical neural tangent kernels (e.g., for all of CIFAR-10 on a ResNet-101): `https://github.com/aw31/empirical-ntks`, filling in a gap in tools for exploring neural tangent kernels at scale.

### 1.1. Related Work

Since Zhang et al. (2017), many researchers have sought to explain why overparameterized models generalize. High-dimensional linear models capture many of the central empirical phenomena and are a natural proving ground for theories of overparameterized models (Mei & Montanari, 2020; Belkin et al., 2020; Bartlett et al., 2020). Recently, a flurry of works has analyzed the generalization risk of high-dimensional ridge regression under various assumptions, typically Gaussian data in the asymptotic limit (Hastie et al., 2020; Canatar et al., 2021; Wu & Xu, 2020; Jacot et al., 2020b; Rosset & Tibshirani, 2020; Loureiro et al., 2021; Richards et al., 2021; Mel & Ganguli, 2021; Simon et al., 2021). Our analysis, like that of Hastie et al. (2020), is based on a local random matrix law (Knowles & Yin, 2017) and produces non-asymptotic bounds for general distributions.

Other, more classical, approaches to generalization include Rademacher complexity (e.g., Bartlett & Mendelson (2001); Bartlett et al. (2002)), norm-based measures (e.g., Bartlett (1996); Neyshabur et al. (2015)), PAC-Bayes approaches for stochastic models (e.g., McAllester (1999); Dziugaite & Roy (2017)), and spectral notions of effective dimension (e.g., Zhang (2005); Dobriban & Wager (2018); Bartlett et al. (2020)). While some of these measures have been studied in large-scale experiments (Jiang et al., 2020; Dziugaite et al., 2020), our evaluations focus on a different perspective: we study whether they capture the basic empirical phenomena of overparameterized models, such as scaling laws and the effect of pretraining.

To estimate generalization risk, we revisit the GCV estimator of Craven & Wahba (1978). GCV was initially studied as an estimate of error over a *fixed* sample (Golub et al., 1979; Li, 1986; Cao & Golubev, 2006). Such analyses, however, do not account for the randomness of the sample and thus fail to capture high-dimensional settings with disparate train and test risks. Recently, the high-dimensional setting has received more attention: Jacot et al. (2020b) analyze GCV for random, Gaussian covariates in the "classical" regime where train risk approximates test risk.[1] And, Hastie et al. (2020), Adlam & Pennington (2020), and Patil et al. (2021) asymptotically analyze GCV when the ratio $P/N$ between the dimension $P$ and the sample size $N$ converges to a fixed limit. In contrast, to study scaling in $N$ (for fixed $P$), we prove *non-asymptotic* bounds on the convergence of GCV that hold: (i) beyond the classical regime, (ii) for a wide range of $N/P$, and (iii) for general covariance structures. Experimentally, GCV has previously been studied by Efron (1986) and Rosset & Tibshirani (2020) in numerical simulations and by Jacot et al. (2020b) for shift-invariant kernels on the MNIST and Higgs datasets. Our experiments take

[1] See Appendix D for a detailed discussion of the classical vs. non-classical regimes of high-dimensional ridge regression.

these investigations to a significantly larger scale and focus on more realistic neural representations.

One phenomenon we study—neural scaling laws—was first observed by Kaplan et al. (2020). Since this observation, Bahri et al. (2021) derive a spectrum-only formula for kernel regression scaling, and Cui et al. (2021) derive precise rates for ridge regression scaling in random matrix regimes. In comparison, we show that alignment (and not just the eigenvalues) is essential for understanding scaling in practice, and we also use random matrix theory to give a more principled way to estimate the decay rates of the population eigenvalues and alignment coefficients.

Finally, the neural representations we study are motivated by the *neural tangent kernel* (NTK) (Jacot et al., 2018). There has been a rich line of theoretical work studying ultra-wide neural networks and their relationship to NTKs (e.g., Arora et al. (2019); Lee et al. (2019); Yang (2019)). In contrast, we work with NTKs extracted from realistic, finite-width networks—including pretrained networks—and use them as a testbed for exploring measures of generalization.

## 2. Preliminaries

### 2.1. High-dimensional Ridge Regression

We study a simple model of linear regression, in which we predict labels $y \in \mathbb{R}$ from data points $x \in \mathbb{R}^P$. Each $x$ is drawn from a distribution $\mathcal{D}$ with *unknown* second moment $\Sigma := \mathbb{E}_{x\sim\mathcal{D}}\left[xx^\mathsf{T}\right]$, and its label $y$ is given by $y = \beta^\mathsf{T} x$ for an *unknown* ground truth function[2] $\beta \in \mathbb{R}^P$. Let $\Sigma$ have eigendecomposition $\sum_{i=1}^P \lambda_i v_i v_i^\mathsf{T}$, with $\lambda_1 \geq \cdots \geq \lambda_P$.

To estimate $\beta$, we assume we have a dataset $\{(x_i, y_i)\}_{i=1}^N$ of $N$ independent samples, with $x_i \sim \mathcal{D}$ and $y_i = \beta^\mathsf{T} x_i$ for all $i$. For notational convenience, we write this dataset as $(X, y)$, where $X \in \mathbb{R}^{N\times P}$ has $i$-th row $x_i$ and $y \in \mathbb{R}^N$ has $i$-th entry $y_i$. Let $\widehat{\Sigma} := \frac{1}{N} X^\mathsf{T} X$ be the empirical second moment matrix, with eigendecomposition $\sum_{i=1}^P \hat{\lambda}_i \hat{v}_i \hat{v}_i^\mathsf{T}$ such that $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_N$.

Given training data $(X, y)$ and an estimator $\hat{\beta} = \hat{\beta}(X, y)$, our goal in this paper is to predict its generalization risk $\mathcal{R}$, defined as $\mathcal{R}(\hat{\beta}) := \mathbb{E}_{x\sim\mathcal{D}}\left[(\beta^\mathsf{T} x - \hat{\beta}^\mathsf{T} x)^2\right]$, *without access to* an independently drawn test dataset.

We focus on the ridge regression estimators $\hat{\beta}_\lambda$ given by

$$\hat{\beta}_\lambda := \arg\min_{\hat{\beta}} \frac{1}{N}\sum_{i=1}^N \left(y_i - \hat{\beta}^\mathsf{T} x_i\right)^2 + \lambda \|\hat{\beta}\|_2^2$$

for $\lambda > 0$, and $\hat{\beta}_0 := \lim_{\lambda\to 0^+} \hat{\beta}_\lambda$.

[2] We assume—for simplicity's sake—that the linear model is well-specified and that labels are noiseless. This holds without loss of generality in high dimensions: both noise and misspecification can be embedded into the model by adding an additional "noise" dimension. See Section 3.3 and Appendix B for details.

Recent theoretical advances (Hastie et al., 2020; Canatar et al., 2021; Wu & Xu, 2020; Jacot et al., 2020b; Loureiro et al., 2021; Richards et al., 2021; Mel & Ganguli, 2021; Simon et al., 2021) have characterized $\mathcal{R}(\hat{\beta}_\lambda)$ under a variety of random matrix assumptions. These works all show that $\mathcal{R}(\hat{\beta}_\lambda)$ can be approximated by the *omniscient risk estimate*

$$\mathcal{R}^\lambda_{\text{omni}} := \frac{\partial \kappa}{\partial \lambda} \cdot \kappa^2 \sum_{i=1}^P \left( \frac{\lambda_i}{(\kappa + \lambda_i)^2} \left(\beta^\mathsf{T} v_i\right)^2 \right), \quad (1)$$

where $\kappa = \kappa(\lambda, N)$ is an *effective regularization* term (see (4) for a definition). We call this expression the omniscient risk estimate because it depends on the *unknown* second moment matrix $\Sigma$ and the *unknown* ground truth $\beta$. Our analysis will approximate (1) using only the empirical second moment matrix $\widehat{\Sigma}$ and the observations $y$, while also yielding a concrete relationship between train and test risk.

### 2.2. Methods for Predicting Generalization Risk

We discuss several baseline approaches for predicting generalization risk and then describe the GCV estimator.

The simplest method uses empirical risk (i.e., training error) $\mathcal{R}_{\text{empirical}}(\hat{\beta}) := \frac{1}{N}\sum_{i=1}^N (y_i - \hat{\beta}^\mathsf{T} x_i)^2$ as a proxy. This is the foundation of uniform convergence approaches in learning theory (e.g., VC-dimension and Rademacher complexity). However, training error is a poor predictor of test error in the overparameterized regime, as seen in Figure 1.

Ridge regression admits more specific analyses. A typical approach takes a bias-variance decomposition over label noise and bounds each term with norm- or spectrum-based quantities. For instance, the recent textbook of Bach (2023) shows, based on matrix concentration inequalities, that

$$\mathcal{R}(\hat{\beta}_\lambda) \leq \underbrace{16\lambda\|\beta\|_2^2}_{\text{norm-based}} + \underbrace{16\frac{\sigma^2}{N}\operatorname{Tr}(\Sigma(\Sigma+\lambda I)^{-1})}_{\text{spectrum-based}} \quad (2)$$

holds when $N\lambda$ is large enough, where $\sigma^2$ upper bounds the variance of the label noise. Such norm- or spectrum-based terms are typical of many theoretical analyses.

**The GCV estimator.** Cross-validation is a third approach to predicting generalization risk. However, cross-validation is not guaranteed to work in high dimensions and can fail in practice (Bates et al., 2021). Craven & Wahba (1978) thus introduce the *generalized cross-validation* (GCV) estimator

$$\text{GCV}_\lambda := \left(\frac{1}{N}\sum_{i=1}^N \frac{\lambda}{\lambda + \hat{\lambda}_i}\right)^{-2} \mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda), \quad (3)$$

which they and Golub et al. (1979) heuristically derive by modifying cross-validation to be rotationally invariant.[3] We

[3] As we did for $\hat{\beta}_0$, we define $\text{GCV}_0 := \lim_{\lambda\to 0^+} \text{GCV}_\lambda$.

| Configuration | *Finetuning* | eNTK | Last layer |
|---|---|---|---|
| CIFAR-10 / ResNet-18 | 4.3% | 6.7% | 14.0% |
| CIFAR-100 / ResNet-34 | 15.9% | 19.0% | 33.9% |
| Flowers-102 / ResNet-50 | 5.6% | 7.0% | 9.7% |
| Food-101 / ResNet-101 | 15.3% | 21.3% | 33.7% |

*Table 1.* Test classification error rates of finetuning with SGD, kernel regression on the eNTK, and linear regression on the last layer activations for various datasets and pretrained models.

will study this estimator empirically and show its form can be understood as a consequence of random matrix theory.

### 2.3. Experimental Setup: Empirical NTKs

To benchmark our risk estimates in realistic settings, we use feature representations derived from large-scale, possibly pretrained neural networks. Specifically, we use the *empirical* neural tangent kernel (eNTK). Given a neural network $f(\cdot;\theta)$ with $P$ parameters ($\theta \in \mathbb{R}^P$) and $C$ output logits ($f(x;\theta) \in \mathbb{R}^C$), the eNTK representation of a data point $x$ at $\theta_0$ is the Jacobian $\varphi_{\text{eNTK}}(x) := \frac{\partial f}{\partial \theta}(x;\theta_0) \in \mathbb{R}^{P\times C}$.

**Models and datasets.** We consider eNTK representations of residual networks on several computer vision datasets, both at random initialization and after pretraining. Specifically, we consider ResNet-{18, 34, 50, 101} applied to the CIFAR-{10, 100} (Krizhevsky, 2009), Fashion-MNIST (Xiao et al., 2017), Flowers-102 (Nilsback & Zisserman, 2008), and Food-101 (Bossard et al., 2014) datasets. All random initialization was done following He et al. (2015); pretrained networks (obtained from PyTorch) were pretrained on ImageNet and had randomly re-initialized output layers.

To verify that pretrained eNTK representations achieve competitive generalization performance, we compare kernel regression on pretrained eNTKs to regression on the last layer activations and to finetuning the full network with SGD (see Table 1). We find that pretrained eNTKs achieve accuracy much closer to that of finetuning than that of regression on the last layer. The eNTKs we consider also have stronger empirical performance than the best-known infinite-width NTKs (Arora et al., 2019; Li et al., 2019; Lee et al., 2020).

**Computational considerations.** For computations with eNTK representations, we apply the kernel trick and instead work with the *eNTK matrix* $\left[\varphi_{\text{eNTK}}(x_i)^\mathsf{T}\varphi_{\text{eNTK}}(x_j)\right]_{i,j=1}^N \in \mathbb{R}^{(N\times C)\times(N\times C)}$. To further speed up computation, we use the fact that, since our models have randomly initialized output layers, the expected eNTK can be written as $I_C \otimes K_0$, for some kernel $K_0 \in \mathbb{R}^{N\times N}$ and the $C \times C$ identity matrix $I_C$ (Lee et al., 2020). The full $NC \times NC$ eNTK can thus be approximated by $I_C \otimes K$, where $K$ is the eNTK with respect to a single randomly initialized output logit. Notice that kernel regression with respect to $I_C \otimes K$ decomposes into $C$ independent kernel regression problems, each with respect to $K$. To reduce compute, we apply this approximation in all of our experiments. For further details, see Appendix E.

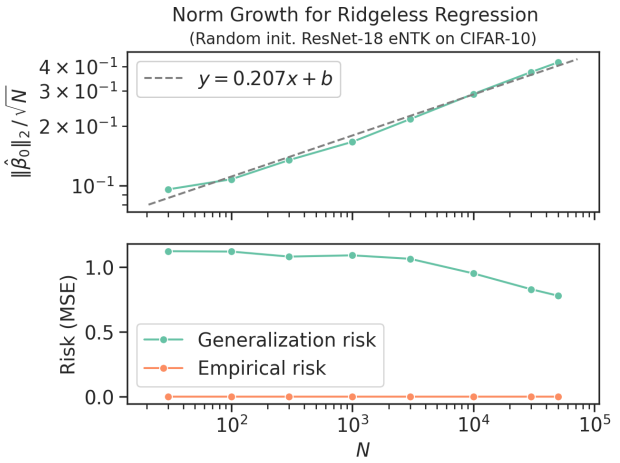


*Figure 2.* The top graph plots the growth of $\frac{\|\hat{\beta}_0\|_2}{\sqrt{N}}$ in $N$ for linear regression on the eNTK of a randomly initialized ResNet-18 on Fashion-MNIST. The bottom graph shows that the generalization risk of $\hat{\beta}_0$ decreases in $N$ under the same setup, despite the growth in $\|\hat{\beta}_0\|_2$, while the empirical risk of $\hat{\beta}_0$ remains 0 throughout.

**Baseline approaches.** To illustrate some of the challenges inherent to this setting, we compare GCV against two norm- and spectrum-based expressions, similar to those of (2). We describe these baselines in detail in Section 4.

## 3. Challenges of High-Dimensional Regression from the Real World

We make several empirical observations that challenge most theoretical analyses: (i) The ground truth $\beta$ has effectively infinite norm, leading $\|\hat{\beta}_\lambda\|_2$ to grow quickly with $N$ and making norm-based bounds vacuous. (ii) When $N \ll P$, the empirical second moment $\widehat{\Sigma}$ is not close to its population mean $\Sigma$. (iii) Many analyses estimate risk in terms of noise in the training set, but noise and signal are interchangeable in high dimensions, making such estimates break down.

### 3.1. Norm-based Bounds Are Vacuous

The norms $\|\beta\|_2$ and $\|\hat{\beta}\|_2$ are often used to measure function complexity in generalization bounds. Here, we examine how these norms behave for kernel regression in practice.

Many theoretical analyses, including Rademacher complexity (Bartlett & Mendelson, 2001), give risk bounds for an estimator $\hat{\beta}$ in terms of the quantity $\|\hat{\beta}\|_2/\sqrt{N}$ (or a monotonic function thereof). However, $\|\hat{\beta}\|_2/\sqrt{N}$ can *increase* as $N$ increases (and the generalization risk decreases): Figure 2 depicts this for $\hat{\beta}_0$ computed on the eNTK of a randomly initialized ResNet-18 on Fashion-MNIST. Moreover,

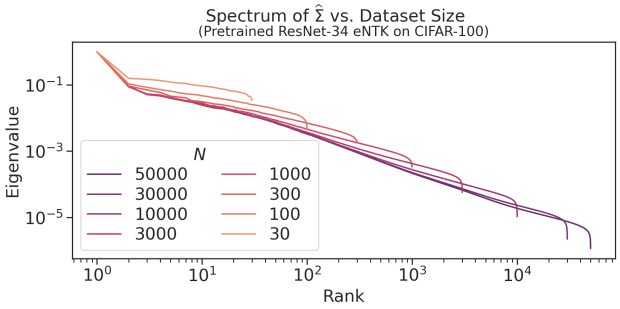


*Figure 3.* Each line plots the pairs $(i, \hat{\lambda}_i)$ for $\widehat{\Sigma}$ from $N$ pretrained ResNet-34 eNTK representations of CIFAR-100 images. The $\widehat{\Sigma}$ eigenvalues converge slowly, and it is not obvious—particularly from considering only a single $N$—what the scaling trend is.

this finding is consistent across models and datasets (see Appendix H). Consequently, norm-based bounds give the wrong qualitative prediction for scaling. This echoes the findings of Nagarajan & Kolter (2019) and shows norm-based bounds can fail even for practical linear models.

Other analyses rely on the norm $\|\beta\|_2$ of the ground truth, either directly in the risk estimate (e.g., Dobriban & Wager (2018)) or as a term in the error bound (e.g., Hastie et al. (2020)). However, Figure 2 suggests that $\beta$ has large norm: for a clean dataset like CIFAR-10, we can assume labels are close to noiseless.In this case, $\hat{\beta}_0$ is the projection of $\beta$ onto $X$, from which it follows that $\|\hat{\beta}_0\|_2 \leq \|\beta\|_2$. Supposing $\|\hat{\beta}_0\|_2$ continues to grow superlinearly in $N$, the norm $\|\beta\|_2$ must be large. It may thus make the most sense to think of $\beta$ as having effectively infinite norm. However, this has the effect of making bounds that rely on $\|\beta\|_2$ vacuous.[4]

### 3.2. $\widehat{\Sigma}$ Converges Slowly to $\Sigma$

The high dimensionality of our setting ($P \gg N$) implies the empirical second moment $\widehat{\Sigma}$ is slow to converge to its expectation $\Sigma$. Figure 3 depicts slow convergence of the spectrum of $\widehat{\Sigma}$ derived from a pretrained ResNet-34 on CIFAR-100. Similar conclusions hold for other models and datasets—see Appendix H. We now discuss the consequences.

First, the slow convergence of $\widehat{\Sigma}$ makes it hard to empirically estimate quantities that depend on the spectrum of $\Sigma$, such as $\mathcal{R}^{\lambda}_{\text{omni}}$; Loureiro et al. (2021) and Simon (2021) both note this challenge. Moreover, as shown in Figure 3, trends for eigenvalue decay extrapolated from $\widehat{\Sigma}$ may not hold for $\Sigma$. This can be problematic for estimating scaling law rates (Bahri et al., 2021; Cui et al., 2021).

The slow convergence also hurts analyses that rely on the approximation $\widehat{\Sigma} \approx \Sigma$, e.g. those of of Hsu et al. (2014) and Bach (2023) for ridge regression: the assumptions needed to derive $\widehat{\Sigma} \approx \Sigma$ would also imply $\mathcal{R}_{\text{empirical}}(\hat{\beta}) \approx \mathcal{R}(\hat{\beta})$ (see Appendix D), which we know does not hold (see Figure 1). Therefore, we do not have $\widehat{\Sigma} \approx \Sigma$ in the manner needed for such analyses to apply.

### 3.3. Kernel Regression Is Effectively Noiseless

Many works (e.g., Belkin et al. (2018); Bartlett et al. (2020)) have sought to explain the finding that large models generalize despite being able to interpolate random labels (Zhang et al., 2017), and thus focus on overfitting with label noise. Yet high-dimensional phenomena occur on nearly noiseless datasets like CIFAR-10. We now discuss how label noise is unnecessary in a stronger sense: in high dimensions, any noisy instance of linear regression is *indistinguishable* from a noiseless instance with a complex ground truth.

To show this, we embed linear regression with noisy labels into the noiseless model of Section 2.1 by constructing for each noisy instance a sequence of noiseless instances that approximate it. We sketch the construction here, and present it in full in Appendix B. Suppose that $y = \beta^{\mathsf{T}} x + \xi$, where $\xi$ represents mean-zero noise. We rewrite $y$ as $y = \beta'^{\mathsf{T}} x'$, where $x' = \begin{bmatrix} x \\ t^{1/2}\xi \end{bmatrix}$, $\beta' = \begin{bmatrix} \beta \\ t^{-1/2} \end{bmatrix}$, and $t > 0$. As $t \to 0$, ridge regression on the "augmented" covariates $x'$ converges *uniformly* over all $\lambda \geq 0$ to ridge regression on the original covariates $x$. The original, noisy instance is thus the limit of a sequence of noiseless instances.[5]

This discussion suggests noiseless regression (allowing for $\beta$ of large norm) can capture our empirical setting, whereas analyses that require label noise may not directly apply.

## 4. Empirically Evaluating GCV

Having demonstrated some of the challenges that our empirical setting poses for typical theories, we now empirically show that the GCV estimator,

$$\text{GCV}_\lambda = \left( \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda}{\lambda + \hat{\lambda}_i} \right)^{-2} \mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda),$$

accurately predicts the generalization risk. In Section 4.1, we first study the GCV estimator in isolation, following the setup in Section 2.3. We observe excellent agreement between the predicted and actual generalization risks across a wide range of dataset sizes $N$ and regularization strengths $\lambda$. In Section 4.2, we then quantitatively compare the GCV estimator against both norm- and spectrum-based measures of generalization, and find that GCV both has better correlation with the actual generalization risk and better predicts the asymptotic scaling.

[4] Belkin et al. (2018) suggest the perceptron analysis (Novikoff, 1962) as a way to understand generalization in the noiseless setting; however, a large $\|\beta\|_2$ makes this approach ineffective as well.

[5] In Appendix B, we show that the same reduction applies to misspecified problems. As an application, we additionally show how terms for variance from previous works can be read off of (1).

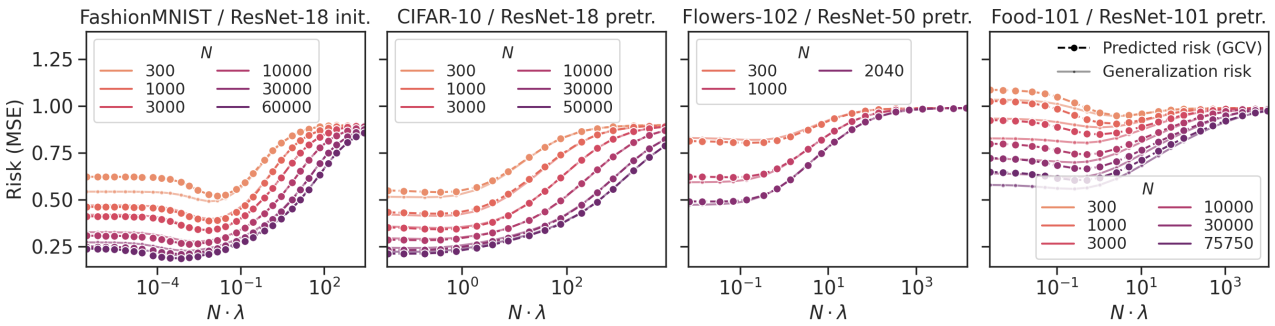


*Figure 4.* Generalization risk vs. the GCV prediction, for various datasets and networks, across sample sizes $N$ and regularization levels $\lambda$.

### 4.1. The Predictive Ability of GCV

To evaluate the GCV estimator, we compute an eNTK for each model-dataset pair listed in Table 2. For each eNTK, we then compare the GCV estimate to the actual generalization risk over a wide range of dataset sizes $N$ and regularization levels $\lambda$. Full details of the experimental setup are given in Appendix E. And in Appendix H, we run the same experiment for ridge regression on last layer activations.

Figures 1 and 4 plot the results of this experiment. All curves demonstrate significant agreement between predicted and actual generalization risks, with over 90% of all predictions having at most 0.09 error in both relative and absolute terms. For most instances, the GCV predictions are nearly perfect for large $N\lambda$ and only diverge slightly for small $N\lambda$. Importantly, the predictions are accurate in two regimes: (i) when mean-squared error is minimized, and (ii) beyond the "classical" regime (i.e., even when the train-test gap is large). Finally, predictions for fixed $N\lambda$ tend to improve as $N$ increases, suggesting convergence in the large $N$ limit.

### 4.2. Comparison to Alternate Approaches

We next use the same setup to compare GCV against two alternative measures, based on the norm of $\hat{\beta}$ and the spectrum of $\widehat{\Sigma}$, respectively.

As discussed in Section 3.1, the norm-based approach gives bounds of the form $\|\hat{\beta}\|_2/\sqrt{N}$. Thus, we consider the estimate $\widehat{\mathcal{R}}^{\lambda}_{\text{norm}} := \|\hat{\beta}_\lambda\|_2/\sqrt{N}$ in our experiments. More general norm-based quantities have been proposed to bound the generalization risk of neural networks (see, e.g., Jiang et al. (2020)); however, when specialized to linear models, these bounds simply become increasing functions of $\|\hat{\beta}\|_2/\sqrt{N}$.

For our spectrum-only estimate, we use a precise estimate of generalization risk in terms of "effective dimension" quantities (Zhang, 2005) when $\beta$ is drawn from an isotropic prior. We consider, for $\hat{\kappa} := \left(\frac{1}{N}\sum_{i=1}^{N}(\lambda+\hat{\lambda}_i)^{-1}\right)^{-1}$, the family

$$\widehat{\mathcal{R}}^{\alpha,\sigma,\lambda}_{\text{spec}} := \hat{\kappa}^2\left(\alpha^2\sum_{i=1}^{N}\frac{\hat{\lambda}_i}{(\lambda+\hat{\lambda}_i)^2}+\frac{\sigma^2}{N}\sum_{i=1}^{N}\frac{1}{(\lambda+\hat{\lambda}_i)^2}\right)$$

of estimates derived from the main theorem of Dobriban & Wager (2018).[6] We fit $\alpha^2$ and $\sigma^2$ so that the predictions best match the observed generalization risks, obtaining an upper bound on the performance of this method over all $\alpha$ and $\sigma$. This family of estimators lets us explore whether naturally-occurring data can be summarized by the two parameters of "signal strength" $\alpha$ and "noise level" $\sigma$.

To evaluate the ability of each predictor to model generalization, we consider two benchmarks. First, we measure the correlation between the predictions and the generalization risk for each dataset on the sets of $(N, \lambda)$ pairs shown in Figures 1 and 4. Correlation lets us equitably compare unscaled predictors, such as $\widehat{\mathcal{R}}_{\text{norm}}$, to more precise estimates, such as GCV and $\widehat{\mathcal{R}}_{\text{spec}}$. Second, we test how well these estimators predict the scaling of optimally tuned ridge regression. For this, we find an optimal ridge parameter $\lambda^*_N$ for each $N$ and then estimate the power law rate (given by $N^{-\alpha}$ for some $\alpha > 0$) of predicted generalization risk with respect to the sample size $N$. (Applied to the ground truth, this would yield the scaling rate of the model.) Full details are provided in Appendix E.

The results of these experiments are displayed in Table 2. Plots of the spectrum- and norm-based predictions are also presented in Appendix H. We find, perhaps unsurprisingly in light of Section 3.1, that the norm-based measure has the *wrong sign* when predicting generalization, both in terms of correlation and in terms of scaling.[7] The spectrum-only approach also struggles to accurately predict generalization risk: it does not predict any scaling on Fashion-MNIST and achieves much lower correlations across the board. Finally, GCV correlates well with the actual generalization risks and accurately predicts scaling behavior on all datasets.

## 5. A Random Matrix Perspective on GCV

We next justify the impressive empirical performance of GCV with a theoretical analysis. We prove a *non-asymptotic*

[6] See Appendix F for a derivation of this estimator.

[7] This cannot be explained by excess regularization reducing the norm while also making performance worse: Figure 2 shows that the trend points the wrong way even when $\lambda = 0$.

| Configuration | *Ground truth* | | GCV | | Spectrum-only | | Norm-based | |
|---|---|---|---|---|---|---|---|---|
| | $r$ | $\alpha$ | $r$ | $\alpha$ | $r$ | $\alpha$ | $r$ | $\alpha$ |
| Fashion-MNIST / ResNet-18 init. | 1.000 | 0.166 | 0.996 | 0.192 | 0.080 | 0.008 | −0.584 | −0.121 |
| CIFAR-10 / ResNet-18 pretr. | 1.000 | 0.162 | 0.999 | 0.182 | 0.977 | 0.134 | −0.641 | −0.044 |
| CIFAR-100 / ResNet-34 pretr. | 1.000 | 0.124 | 0.996 | 0.124 | 0.846 | 0.070 | −0.507 | −0.166 |
| Flowers-102 / ResNet-50 pretr. | 1.000 | — | 0.999 | — | 0.665 | — | −0.786 | — |
| Food-101 / ResNet-101 pretr. | 1.000 | 0.099 | 0.979 | 0.085 | 0.718 | 0.035 | −0.483 | −0.188 |

*Table 2.* The $r$ columns display the correlations of each prediction to generalization risk, and the $\alpha$ columns display the estimated scaling exponents. We do not run the scaling experiment for Flowers-102 because it only consists of 2040 images.

bound on the absolute error $\left|\mathrm{GCV}_\lambda - \mathcal{R}(\hat{\beta}_\lambda)\right|$ of GCV under a random matrix hypothesis.

Our analysis of the GCV estimator has the following features: (i) It holds even for $\beta$ with large norm, requiring only a bound on $\mathbb{E}_{x\sim\mathcal{D}}\left[(\beta^\mathsf{T}x)^2\right] = \beta^\mathsf{T}\Sigma\beta$. This is important for our empirical setting because, while $\|\beta\|_2$ may be large (as discussed in Section 3.1), the fact that our labels are 1-hot implies $(\beta^\mathsf{T}x)^2 \le 1$ always. (ii) It is the first, to our knowledge, non-asymptotic analysis of GCV that applies beyond the "classical" regime, holding even when the train-test gap is large. (iii) It makes *no additional assumptions* beyond a generic random matrix hypothesis and thus makes clear the connection between the GCV estimator and random matrix effects.[8] In particular, we do not make further assumptions about independence, moments, or dimensional ratio.

To illustrate the main technical ideas, we outline our theoretical approach at a high level in the remainder of this section and defer our formal treatment to Appendix A.

### 5.1. The Random Matrix Hypothesis

We assume a local version of the Marchenko-Pastur law as our random matrix hypothesis. To state this hypothesis, we first define $\kappa = \kappa(\lambda, N)$ as the (unique) positive solution to

$$1 = \frac{\lambda}{\kappa} + \frac{1}{N}\sum_{i=1}^{P}\frac{\lambda_i}{\kappa+\lambda_i}, \tag{4}$$

with $\kappa(0, N) \coloneqq \lim_{\lambda\to 0^+}\kappa(\lambda, N)$. We call $\kappa$ the *effective regularization*, as it captures the combined effect of the explicit regularization $\lambda$ and the "implicit regularization" (Neyshabur, 2017; Jacot et al., 2020a) of ridge regression. In terms of $\kappa$, the Marchenko-Pastur law can be roughly thought of as the statement $\lambda(\lambda I + \widehat{\Sigma})^{-1} \approx \kappa(\kappa I + \Sigma)^{-1}$. We assume this approximation holds in the following sense:

**Hypothesis 1** (Marchenko-Pastur law over $\mathbb{R}_{>0}$, informal)**.** The *local Marchenko-Pastur law* holds over $S \subseteq \mathbb{R}_{>0}$ if, for every deterministic $v \in \mathbb{R}^P$ such that $v^\mathsf{T}\Sigma v \le 1$, the following hold uniformly over all $\lambda \in S$:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{\lambda}_i+\lambda} \approx \frac{1}{\kappa} \tag{5}$$

$$v^\mathsf{T}\lambda\big(\lambda I + \widehat{\Sigma}\big)^{-1}v \approx v^\mathsf{T}\kappa\big(\kappa I + \Sigma\big)^{-1}v. \tag{6}$$

Hypothesis 1 is known to hold when $x$ is a linear function of independent (but not necessarily i.i.d.) random variables (Knowles & Yin, 2017), which includes Gaussian covariates as a special case. Hypothesis 1 is expected, in fact, to hold in even greater generality, as an instance of the universality phenomenon for random matrices.

While one cannot verify Hypothesis 1 directly, since it depends on the unknown quantities $\mathcal{D}$ and $\beta$, we present evidence for its empirical validity in Appendix G. Specifically, we verify that (5) and (6) are consistent with each other in our empirical setting, by checking the relationships that they predict between empirically measurable quantities.

### 5.2. The GCV Theorem

We show the following error bound for $\mathrm{GCV}_\lambda$, which states $\mathrm{GCV}_\lambda$ accurately predicts generalization risk under Hypothesis 1 over a wide range of $N$ and $\lambda$. Our bounds are stated under the normalizations $\mathbb{E}\big[y^2\big] \le 1$ and $\mathbb{E}\big[\|x\|_2^2\big] \le 1$.

**Theorem 2** (Informal)**.** *Suppose Hypothesis 1 holds over* $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$*. Then*

$$\left|\mathrm{GCV}_\lambda - \mathcal{R}(\hat{\beta}_\lambda)\right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\lambda}.$$

To prove Theorem 2, we first show $\mathrm{GCV}_\lambda \approx \mathcal{R}^\lambda_{\mathrm{omni}}$. We then prove a sharpened version of the result of Hastie et al. (2020) to show that, if $\mathbb{E}\big[y^2\big] = \beta^\mathsf{T}\Sigma\beta \le 1$, then $\mathcal{R}^\lambda_{\mathrm{omni}} \approx \mathcal{R}(\hat{\beta}_\lambda)$. The first step can be stated as follows.

**Proposition 3** (Informal)**.** *Suppose Hypothesis 1 holds over* $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$*. Then*

$$\left|\mathrm{GCV}_\lambda - \mathcal{R}^\lambda_{\mathrm{omni}}\right| \lesssim N^{-1/2+o(1)} \cdot \big(1 + (N\lambda)^{-3/2}\big).$$

For intuition, we give a heuristic proof of Proposition 3. As a simplification, we use the approximate equalities $\approx$ in Hypothesis 1 instead of precise error bounds. We further

[8] This random matrix hypothesis is known to hold for commonly considered random matrix models (Knowles & Yin, 2017) and is believed to hold even more broadly.

assume that $\approx$ is preserved by differentiation. We justify these approximations in our full analysis in Appendix A.

*Heuristic proof.* By the closed form of $\mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda)$,

$$\text{GCV}_\lambda = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda+\hat{\lambda}_i}\right)^{-2}\beta^\mathsf{T}(\widehat{\Sigma}+\lambda I)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\beta.$$

Hypothesis 1 implies $\left(\frac{1}{N}\sum_{i=1}^{N}(\lambda+\hat{\lambda}_i)^{-1}\right)^{-2} \approx \kappa^2$ and

$$\frac{\partial}{\partial\lambda}(\beta^\mathsf{T}\lambda(\widehat{\Sigma}+\lambda I)^{-1}\beta) \approx \frac{\partial}{\partial\lambda}\left(\beta^\mathsf{T}\kappa(\Sigma+\kappa I)^{-1}\beta\right), \quad (7)$$

assuming we may differentiate through the $\approx$. Hence,

$$\begin{aligned}
&\beta^\mathsf{T}(\widehat{\Sigma}+\lambda I)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\beta \\
&\quad= \frac{\partial}{\partial\lambda}(\beta^\mathsf{T}\lambda(\widehat{\Sigma}+\lambda I)^{-1}\beta) \\
&\quad\approx \frac{\partial}{\partial\lambda}\left(\beta^\mathsf{T}\kappa(\Sigma+\kappa I)^{-1}\beta\right) \\
&\quad= \frac{\partial\kappa}{\partial\lambda}\cdot\beta^\mathsf{T}(\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\beta.
\end{aligned}$$

Substituting into the equation for $\text{GCV}_\lambda$, we obtain

$$\begin{aligned}
\text{GCV}_\lambda &\approx \kappa^2\left(\frac{\partial\kappa}{\partial\lambda}\cdot\beta^\mathsf{T}(\Sigma+\kappa I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\beta\right) \\
&= \frac{\partial\kappa}{\partial\lambda}\cdot\kappa^2\sum_{i=1}^{P}\left(\frac{\lambda_i}{(\kappa+\lambda_i)^2}\left(\beta^\mathsf{T}v_i\right)^2\right) = \mathcal{R}^\lambda_{\text{omni}}. \qquad \square
\end{aligned}$$

## 6. Pretraining and Scaling Laws through a Random Matrix Lens

Having shown that a random matrix approach can fruitfully model generalization risk both in theory and in practice, we apply this theory towards answering: what factors determine whether a neural representation scales well when applied to a downstream task? To answer this question, we revisit $\mathcal{R}^\lambda_{\text{omni}} = \frac{\partial\kappa}{\partial\lambda}\cdot\kappa^2\sum_{i=1}^{P}\left(\frac{\lambda_i}{(\kappa+\lambda_i)^2}\left(\beta^\mathsf{T}v_i\right)^2\right)$, a quantity that depends on the eigenvalues $\lambda_i$ and the *alignment coefficients* $(\beta^\mathsf{T}v_i)^2$ between the eigenvectors and $\beta$.

In Section 6.1, we use eigendecay and alignment to understand why pretrained representations generalize better than randomly initialized ones. We find, perhaps unintuitively, that pretrained representations have *slower* eigenvalue decay (i.e., *higher* effective dimension), but scale better due to better alignment between the eigenvectors and the ground truth. Thus, it is necessary to consider alignment in addition to eigenvalue decay to explain the effectiveness of pretraining.

Motivated by this, in Section 6.2, we study scaling laws for eigendecay and alignment (Caponnetto & Vito, 2007; Cui et al., 2021). We show how to estimate their power law exponents with empirically observable quantities. Combining these yields an empirically accurate estimate of the power law exponent of generalization, suggesting that eigendecay and alignment are *sufficient* statistics for predicting scaling.

### 6.1. Pretraining

A common intuition is that pretraining equips models with simple, "low-dimensional" representations of complex data.

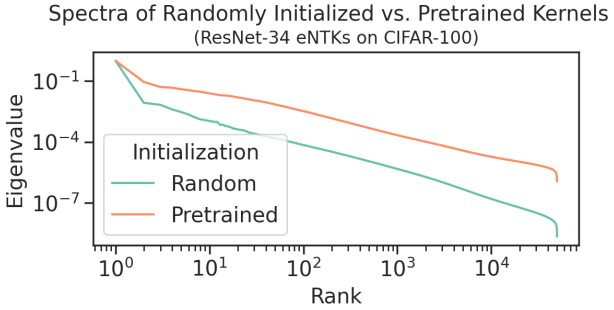


*Figure 5.* The pairs $(i, \hat{\lambda}_i)$ plotted for two ResNet-34 eNTKs: one at random initialization and one after pretraining. Note that the pretrained kernel has *higher* effective dimension.

| Configuration | $\hat{\gamma}$ | $\hat{\delta}$ | $\hat{\alpha}$ | $\alpha$ |
|---|---|---|---|---|
| F-MNIST / ResNet-18 init. | 0.657 | −0.462 | 0.195 | 0.166 |
| F-MNIST / ResNet-18 pretr. | 0.353 | −0.149 | 0.204 | 0.188 |
| CIFAR-10 / ResNet-18 init. | 0.535 | −0.468 | 0.066 | 0.059 |
| CIFAR-10 / ResNet-18 pretr. | 0.270 | −0.089 | 0.181 | 0.162 |
| CIFAR-100 / ResNet-34 init. | 0.482 | −0.466 | 0.016 | 0.014 |
| CIFAR-100 / ResNet-34 pretr. | 0.257 | −0.128 | 0.128 | 0.124 |
| Food-101 / ResNet-101 pretr. | 0.200 | −0.113 | 0.087 | 0.099 |

*Table 3.* The first two columns display the estimated power law rates $\hat{\gamma}$ (of eigendecay) and $\hat{\delta}$ (of alignment). The last two columns compare the estimate $\hat{\alpha} \coloneqq \hat{\gamma} + \hat{\delta}$ for the scaling rate of optimally tuned ridge regression against the actual scaling rate $\alpha$ of $\mathcal{R}(\hat{\beta}_{\lambda^*})$.

Thus, one might expect that pretrained representations have lower effective dimension and that this is the cause of better generalization. Figure 5, however, shows the opposite to be true: on CIFAR-100, a pretrained ResNet-34 eNTK has *slower* eigenvalue decay than a randomly initialized representation (and higher effective dimension). Moreover, this holds consistently across datasets and models, as shown in Appendix H. Thus, dimension alone cannot explain the benefit of pretraining.

The omniscient risk estimate $\mathcal{R}^\lambda_{\text{omni}}$ suggests a possible remedy. While slower eigendecay will increase $\mathcal{R}^\lambda_{\text{omni}}$, the increase can be overcome if the alignment coefficients $(\beta^\mathsf{T}v_i)^2$ decay faster. We will confirm this in Section 6.2 once we develop tools to estimate the decay rates of the eigenvalues and the alignment coefficients: across several models and datasets, pretrained representations exhibit slower eigendecay but better alignment (Table 3). Our finding suggests that the role of pretraining is to make "likely" ground truth functions easily representable and in fact does not reduce data dimensionality. In particular, the covariates cannot be considered in isolation from potential downstream tasks.

### 6.2. Scaling Laws

The omniscient risk estimate $\mathcal{R}^\lambda_{\text{omni}}$ shows that both alignment and eigendecay matter for generalization. To better

understand the behavior of these quantities, which are given in terms of the unobserved $\Sigma$ and $\beta$, we show how the power law rates of these terms can be estimated from empirically observable quantities. We then use these rates to estimate the scaling law rate of the generalization error for optimally regularized ridge regression. We find the estimated rates accurately reflect observed scaling behavior, suggesting that power law models of alignment and eigendecay suffice to capture the scaling behavior of regression on natural data.

We suppose that the population eigenvalues and the alignment coefficients scale as $\lambda_i \asymp i^{-1-\gamma}$ and $(\beta^\mathsf{T} v_i)^2 \asymp i^{-\delta}$, for $\gamma > 0$ and $\delta < 1$. (Note that the latter implies $\|\beta\|_2$ is effectively infinite when $N \ll P$.) Assuming known $\gamma$ and $\delta$, Cui et al. (2021) analyze $\mathcal{R}^{\lambda^*}_{\text{omni}}$, for $\lambda^*$ the optimal ridge regularization, and show in the noiseless regime that

$$\mathcal{R}^{\lambda^*}_{\text{omni}} \asymp N^{-\alpha}, \quad \text{for } \alpha = \gamma + \delta. \tag{8}$$

However, they do not give a satisfactory way to estimate $\gamma$ and $\delta$ from data: they propose simply using the eigenvalues $\widehat{\Sigma}$ as a proxy for those of $\Sigma$. But as we previously observed in Figure 3, convergence of $\widehat{\Sigma}$ to $\Sigma$ can be slow for high-dimensional regression problems.

The following propositions (proven in Appendix C) provide a more principled way to estimate $\gamma$ and $\delta$ in terms of empirically observable quantities, using the same random matrix hypothesis from before.

**Proposition 4.** *Suppose that Hypothesis 1 holds as $\lambda \to 0$ and that $\lambda_i \asymp i^{-1-\gamma}$. Then, $N^{-1} \operatorname{Tr}\big((XX^\mathsf{T})^{-1}\big) \asymp N^{\gamma}$.*

**Proposition 5.** *Suppose that Hypothesis 1 holds at $\lambda > 0$ and that $\lambda_i \asymp i^{-1-\gamma}$ and $(\beta^\mathsf{T} v_i)^2 \asymp i^{-\delta}$. Then,*

$$y^\mathsf{T}\big(XX^\mathsf{T} + N\lambda I\big)^{-1} y \asymp \kappa(\lambda, N)^{-\frac{1-\delta}{1+\gamma}}. \tag{9}$$

Consequently, $\gamma$ can be estimated by fitting the slope of the points $\big(\log N, \log\big(N^{-1}\operatorname{Tr}((XX^\mathsf{T})^{-1})\big)\big) \in \mathbb{R}^2$. And $\delta$ can be estimated by inverting (9) and applying the estimate (5) for $\kappa$ and the preceding estimate for $\gamma$.

To test this approach, we estimate $\gamma$ and $\delta$ as $\hat{\gamma}$ and $\hat{\delta}$ via the quantities in Propositions 4 and 5 and apply these estimates to the datasets listed in Table 3. We also estimate $\hat{\alpha} = \hat{\gamma} + \hat{\delta}$ following (8) and compare $\hat{\alpha}$ to the actual rate $\alpha$ in Table 3.

We find that $\hat{\alpha}$ accurately approximates $\alpha$ for all datasets, suggesting that the power law assumption can be used to model naturally-occurring data. Additionally, we observe for all datasets that $\hat{\delta} < 0$, which suggests that the coefficients $(\beta^\mathsf{T} v_i)^2$ *grow* in $i$, reinforcing our conclusion from Section 3 that $\beta$ has large norm. Finally, for all pairs of randomly initialized and pretrained models in Table 3, note that the pretrained model has smaller $\gamma$ and thus slower eigenvalue decay, but much larger $\delta$. This verifies our hypothesis that pretrained representations scale better due to improved alignment (and despite higher dimension).

## 7. Discussion

In this paper, we identify that the GCV estimator accurately predicts ridge regression generalization on neural representations of large-scale networks and real data, while other more classical approaches fall short. We then elucidate the connection between GCV and random matrix laws, showing that GCV accurately predicts generalization risk whenever a local Marchenko-Pastur law holds. Finally, we apply this perspective to answer basic conceptual questions about neural representations. Our findings suggest several promising directions for future inquiry, which we now discuss.

First, we believe that the random matrix approach has much more to offer towards understanding the statistics of high-dimensional learning: the structure imposed by a random matrix assumption stood apart at capturing the qualitative phenomena of ridge regression. It is thus conceivable that such structure will be *necessary* to understand settings beyond ridge regression, e.g., classification accuracy for logistic regression or modeling natural covariate shifts.

However, there remain open problems even in the setting of ridge regression. For instance, current understanding of random matrix laws does not encompass all the regimes of interest: a natural scaling of regularization is $\lambda \asymp N^{-1}$, but the existing theory (Knowles & Yin, 2017) requires $\lambda$ to be bounded away from 0. Additionally, it would be of interest to achieve a bound on the error of $\mathcal{R}^{\lambda}_{\text{omni}}$ that scales well in the $\lambda \asymp N^{-1}$ limit (like what we have for Proposition 3).

Finally, and most broadly, we hope that the perspective we take towards studying neural representations can inspire more insight towards what is learned by neural networks. We find that eNTK representations reveal much more than the typically considered final-layer activations and serve as a reasonable proxy for understanding finetuning on a pretrained model. Can eNTKs be used as a tractable model to untangle more of the mysteries around large-scale models? For instance, what do eNTKs reveal about features learned via different training procedures? And can the evolution of the eNTK and its associated metrics (e.g., eigendecay and alignment) during training shed light on feature learning?

## Acknowledgements

We thank Yasaman Bahri, Peter Bartlett, Gintare Karolina Dziugaite, Cassidy Laidlaw, Preetum Nakkiran, and Nilesh Tripuraneni for valuable feedback. A. Wei acknowledges support from an NSF Graduate Research Fellowship under grant DGE-2146752. W. Hu acknowledges support from the NSF through grants DMS-2023505 and DMS-2031883 and from the Simons Foundation through award #814639.

## References


Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 74–84, 2020.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32*, pp. 8139–8148, 2019.

Bach, F. *Learning Theory from First Principles*. MIT, 2023.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *arXiv*, abs/2102.06701, 2021.

Bai, Z. and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

Bartlett, P. L. For valid generalization the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems 9*, pp. 134–140, 1996.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. In *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory*, pp. 224–240, 2001.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Localized rademacher complexities. In *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory*, pp. 44–58, 2002.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bates, S., Hastie, T., and Tibshirani, R. Cross-validation: what does it estimate and how well does it do it? *arXiv*, abs/2104.00673, 2021.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 540–548, 2018.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, 2(4): 1167–1180, 2020.

Bloemendal, A., Knowles, A., Yau, H.-T., and Yin, J. On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1):459–552, 2016.

Bossard, L., Guillaumin, M., and Gool, L. V. Food-101 - mining discriminative components with random forests. In *Proceedings of the Thirteenth European Conference on Computer Vision*, pp. 446–461, 2014.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):1–12, 2021.

Cao, Y. and Golubev, Y. On oracle inequalities related to smoothing splines. *Mathematical Methods of Statistics*, 15(4):398–414, 2006.

Caponnetto, A. and Vito, E. D. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7 (3):331–368, 2007.

Craven, P. and Wahba, G. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1978.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborova, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *Advances in Neural Information Processing Systems 34*, 2021.

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

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. In search of robust measures of generalization. In *Advances in Neural Information Processing Systems 33*, 2020.

Efron, B. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

Erdős, L. and Yau, H.-T. *A Dynamical Approach to Random Matrix Theory*. American Mathematical Society, 2017.

Golub, G. H., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv*, abs/1903.08560, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Hsu, D. J., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. *Found. Comput. Math.*, 14 (3):569–600, 2014.

Jacot, A., Hongler, C., and Gabriel, F. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 8580–8589, 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4631–4640, 2020a.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. In *Advances in Neural Information Processing Systems 33*, 2020b.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations*, 2020.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv*, abs/2001.08361, 2020.

Knowles, A. and Yin, J. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169: 257–352, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems 32*, pp. 8570–8581, 2019.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study. In *Advances in Neural Information Processing Systems 33*, 2020.

Li, K.-C. Asymptotic Optimality of $C_L$ and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing. *The Annals of Statistics*, 14(3): 1101–1112, 1986.

Li, Z., Wang, R., Yu, D., Du, S. S., Hu, W., Salakhutdinov, R., and Arora, S. Enhanced convolutional neural tangent kernels. *arXiv*, abs/1911.00809, 2019.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborova, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems 34*, 2021.

Marchenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Marquardt, D. W. and Snee, R. D. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv*, abs/1908.05355, 2020.

Mel, G. and Ganguli, S. A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7578–7587, 2021.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems 32*, pp. 11611–11622, 2019.

Neyshabur, B. Implicit regularization in deep learning. *arXiv*, abs/1709.01953, 2017.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pp. 1376–1401, 2015.

Nilsback, M. and Zisserman, A. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008.

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural tangents: Fast and easy infinite neural networks in python. In *8th International Conference on Learning Representations*, 2020.

Novak, R., Sohl-Dickstein, J., and Schoenholz, S. S. Fast finite width neural tangent kernel. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022.

Novikoff, A. B. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pp. 615–622. Polytechnic Institute of Brooklyn, 1962.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. J. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics*, pp. 3178–3186, 2021.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge(less) regression under general source condition. In *The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 3889–3897, 2021.

Rosset, S. and Tibshirani, R. J. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 115(529):138–151, 2020.

Simon, J. B. A first-principles theory of neural network generalization, October 2021. URL https://bair.berkeley.edu/blog/2021/10/25/eigenlearning/.

Simon, J. B., Dickens, M., and DeWeese, M. R. Neural tangent kernel eigenvalues accurately predict generalization. *arXiv*, abs/2110.03922, 2021.

Steinhardt, J. Robust and nonparametric statistics, April 2021. URL https://jsteinhardt.stat.berkeley.edu/teaching/stat240-spring-2021/notes.pdf.

Wu, D. and Xu, J. On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems 33*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, abs/1708.07747, 2017.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv*, abs/1902.04760, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations*, 2017.

Zhang, T. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9): 2077–2098, 2005.

## A. Analysis of the GCV Estimator (Proofs for Section 5)

In this section, we prove our main theoretical result: that the GCV estimator approximates the generalization risk of ridge regression (Theorem 2). We now give formal statements of Hypothesis 1 and Theorem 2. For Theorem 2, we will assume that Hypothesis 1 holds for $\mathcal{D}$ as well as a family of linear transformations of $\mathcal{D}$.

Before giving formal statements, we make note of a few mathematical conventions that we use throughout this section:

- We say that a family of events $A^N$ indexed by $N$ occurs *with high probability* if, for any (large) constant $D > 0$, there exists a threshold $N_D$ such that $A^N$ occurs with probability at least $1 - N^{-D}$ for all $N \geq N_D$.
- For any two families of functions $f^N, g^N \colon S \to \mathbb{R}_{\geq 0}$ indexed by $N$, we say that $f \lesssim g$ *uniformly over* $S$ if there exists a constant $C > 0$ such that, with high probability, $f^N(z) \leq C \cdot g^N(z)$ uniformly over all $z \in S$. In particular, $\lesssim$ omits constant factors from bounds.
- We let $\mathrm{i}$ (in roman type) denote the imaginary unit and use $i$ (in italic type) as an indexing variable.

With these conventions in mind, Hypothesis 1 is formalized as follows:

**Hypothesis 6** (Local Marchenko-Pastur law over $\mathbb{R}_{>0}$)**.** The *local Marchenko-Pastur law* holds over an open set $S \subseteq \mathbb{R}_{>0}$ if, for every deterministic vector $v \in \mathbb{R}^P$ such that $v^\mathsf{T} \Sigma v \leq 1$, both

$$\left| \frac{1}{\kappa} - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_i + \lambda} \right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\kappa} \sqrt{\frac{\partial \kappa}{\partial \lambda}} \tag{10}$$

and

$$\left| v^\mathsf{T} \big(\kappa(\kappa I + \Sigma)^{-1}\big) v - v^\mathsf{T} \big(\lambda(\lambda I + \widehat{\Sigma})^{-1}\big) v \right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\kappa} \sqrt{\frac{\partial \kappa}{\partial \lambda}} \tag{11}$$

hold uniformly over all $\lambda \in S$.

To analyze the omniscient risk estimate, we will need a slight extension of Hypothesis 6, requiring that Hypothesis 6 hold for a family of linear transformations of the data distribution $\mathcal{D}$:

**Hypothesis 7.** Hypothesis 6 holds for $z = (I + t\Sigma)^{-\frac{1}{2}} x$, where $x \sim \mathcal{D}$, uniformly[9] over all $t \in \{s \in \mathbb{R} : |s| < \frac{1}{2} \|\Sigma\|_{\mathrm{op}}^{-1}\}$.

Theorem 2 can now formally be stated as follows:

**Theorem 8.** *Suppose $\lambda > 0$ is such that Hypothesis 7 holds over $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$. Then,*

$$\left| \mathrm{GCV}_\lambda - \mathcal{R}(\hat{\beta}_\lambda) \right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \beta^\mathsf{T} \Sigma \beta \cdot \left[ \frac{\|\Sigma\|_{\mathrm{op}}}{\lambda} + \left( \frac{\mathrm{Tr}(\Sigma)}{N\lambda} \right)^{3/2} \right].$$

Recall from Section 5 that our analysis of the GCV estimator proceeds in two steps: showing that $\mathrm{GCV}_\lambda \approx \mathcal{R}^\lambda_{\mathrm{omni}}$ and then showing that $\mathcal{R}^\lambda_{\mathrm{omni}} \approx \mathcal{R}(\hat{\beta}_\lambda)$. For the first step, we show the following proposition (formally restating Proposition 3):

**Proposition 9.** *Suppose $\lambda > 0$ is such that Hypothesis 6 holds over $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$. Then,*

$$\left| \mathrm{GCV}_\lambda - \mathcal{R}^\lambda_{\mathrm{omni}} \right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \beta^\mathsf{T} \Sigma \beta \cdot \left( 1 + \frac{\mathrm{Tr}(\Sigma)}{N\lambda} \right)^{3/2}.$$

For the second step, we show the following proposition:

**Proposition 10.** *Suppose $\lambda > 0$ is such that Hypothesis 7 holds over $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$. Then,*

$$\left| \mathcal{R}^\lambda_{\mathrm{omni}} - \mathcal{R}(\hat{\beta}_\lambda) \right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \beta^\mathsf{T} \Sigma \beta \cdot \frac{\|\Sigma\|_{\mathrm{op}}}{\lambda}.$$

In the remainder of this section, we prove Theorem 8. To be self-contained, we briefly recap the setup, precise assumptions, and some background material in Appendix A.1. Next, we prove a general lemma to justify the differentiation step (i.e., (7)) in Appendix A.2. Then, we prove Theorem 8 via Propositions 9 and 10 in Appendices A.3 and A.4.

[9] Since $t$ is 1-dimensional, this uniformity assumption can be relaxed with a standard $\varepsilon$-net argument, which we omit for brevity.

## A.1. Theoretical Preliminaries

### A.1.1. Model

We recall our basic setup from Section 2. We consider a random design model of linear regression, in which covariates $x_i$ are drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathbb{R}^P$ with second moment $\Sigma \in \mathbb{R}^{P\times P}$. Labels are generated by a ground truth $\beta \in \mathbb{R}^p$, with the $i$-th label given by $y_i = \beta^\mathsf{T} x_i$. In this model, the distribution $\mathcal{D}$ (and in particular its second moment $\Sigma$) and the ground truth $\beta$ are *unobserved*. Instead, all we observe are $N$ independent samples $(x_1, y_1), \dots, (x_N, y_N)$.

For our theoretical analysis, we additionally impose the mild assumption that $\lambda \geq N^{-C}$ for some (large) constant $C > 0$.[10] Note that, beyond our random matrix hypothesis, we *do not* assume anything about the dimensional ratio $P/N$, allowing for it to vary widely, and we *do not* assume anything about the covariate distribution $\mathcal{D}$.

For the sake of simplicity, we focus on the case where $\mathbb{E}_{x\sim\mathcal{D}}[x] = 0$. We note that our analysis can be extended to obtain a correction for non-zero means via the Sherman-Morrison rank-1 update formula, but we do not pursue this extension further at this time.

### A.1.2. Ridge Regression

We first recall the notation defined in Section 2. Let $X \in \mathbb{R}^{N\times P}$ be the matrix of covariates and $y \in \mathbb{R}^N$ be the vector of labels. The empirical second moment matrix is denoted by $\widehat{\Sigma} \coloneqq \frac{1}{N} X^\mathsf{T} X$. The eigendecompositions of $\Sigma$ and $\widehat{\Sigma}$ are written as $\Sigma = \sum_{i=1}^P \lambda_i v_i v_i^\mathsf{T}$ and $\widehat{\Sigma} = \sum_{i=1}^N \hat{\lambda}_i \hat{v}_i \hat{v}_i^\mathsf{T}$, respectively, with $\lambda_1 \geq \cdots \geq \lambda_P$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_N$.

Let $\hat{\beta}_\lambda$ be the ridge regression estimator

$$\hat{\beta}_\lambda \coloneqq \arg\min_{\hat{\beta}} \frac{1}{N} \sum_{i=1}^N \left(y_i - \hat{\beta}^\mathsf{T} x_i\right)^2 + \lambda \|\hat{\beta}\|_2^2$$

for $\lambda > 0$, and let $\hat{\beta}_0 \coloneqq \lim_{\lambda\to 0^+} \hat{\beta}_\lambda$. For $\lambda > 0$, one has the closed form $\hat{\beta}_\lambda = \left(\widehat{\Sigma} + \lambda I\right)^{-1} \frac{1}{N} X^\mathsf{T} y = \left(\widehat{\Sigma} + \lambda I\right)^{-1} \widehat{\Sigma} \beta$.

Given an estimator $\hat{\beta} \in \mathbb{R}^P$ for $\beta$, its generalization and empirical risks are

$$\mathcal{R}(\hat{\beta}) \coloneqq \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[(\beta^\mathsf{T} x - \hat{\beta}^\mathsf{T} x)^2] \quad \text{and} \quad \mathcal{R}_{\text{empirical}}(\hat{\beta}) \coloneqq \frac{1}{N} \sum_{i=1}^N (y_i - \hat{\beta}^\mathsf{T} x_i)^2,$$

respectively. For ridge regression when $\lambda > 0$, one has the closed form expressions

$$\mathcal{R}(\hat{\beta}_\lambda) = \lambda^2 \beta^\mathsf{T} \left(\widehat{\Sigma} + \lambda I\right)^{-1} \Sigma \left(\widehat{\Sigma} + \lambda I\right)^{-1} \beta \quad \text{and} \quad \mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda) = \lambda^2 \beta^\mathsf{T} \left(\widehat{\Sigma} + \lambda I\right)^{-1} \widehat{\Sigma} \left(\widehat{\Sigma} + \lambda I\right)^{-1} \beta. \tag{12}$$

### A.1.3. The Asymptotic Stieltjes Transform

To relate our random matrix hypothesis (Hypothesis 1) to the existing random matrix literature, we define the $N$-sample *asymptotic Stieltjes transform* $m$ of $\Sigma$, as $m$ is the more standard object to consider in random matrix theory. We will state a version of Hypothesis 1 in terms of $m$ and later use the properties of $m$ to analyze the GCV estimator.

Before defining $m$, it is helpful to recall the definition of effective regularization $\kappa = \kappa(\lambda, N)$, for $\lambda > 0$, as the (unique) positive solution to

$$1 = \frac{\lambda}{\kappa} + \frac{1}{N} \sum_{i=1}^P \frac{\lambda_i}{\kappa + \lambda_i}. \tag{13}$$

The $N$-sample asymptotic Stieltjes transform $m$ of $\Sigma$ is the analytic continuation of $m(z) = 1/\kappa(-z, N)$ (as a function on the negative reals) to $\mathbb{C} \setminus \mathbb{R}_{\geq 0}$. We define $m$ using an equation similar to (13). Let $\mathbb{H} \coloneqq \{z \in \mathbb{C} : \text{Im}(z) > 0\}$ denote the complex upper half-plane. For each $z \in \mathbb{H}$, one can show that there exists a unique solution in $\mathbb{H}$ to

$$1 = -zm + \frac{1}{N} \sum_{i=1}^P \frac{m\lambda_i}{1 + m\lambda_i},$$

[10] This assumption is made for convenience: relaxing it worsens the bound by only a $\log(1/\lambda)$ factor.

which we take to be $m(z)$. By the Schwarz reflection principle, this function on $\mathbb{H}$ has a unique analytic continuation to $\mathbb{C} \setminus \mathbb{R}_{\geq 0}$. A key property of $m$ is that there exists a unique positive measure $\varrho$ on $[0, \infty)$ such that

$$m(z) = \int \frac{d\varrho(x)}{x - z}. \tag{14}$$

In other words, $m$ is the *Stieltjes transform* of $\varrho$. This measure $\varrho$ is known as the $N$-sample *asymptotic eigenvalue density* of $\Sigma$. For proofs of these claims, we refer the reader to Bai & Silverstein (2010) and Knowles & Yin (2017, Section 2.2).

### A.1.4. The Random Matrix Hypothesis

To make our analysis as general as possible and to make the connection to random matrix theory clear, we give our analysis for any distribution $\mathcal{D}$ that satisfies Hypothesis 6. This hypothesis is a modern interpretation of the Marchenko-Pastur law and formalizes the heuristic random matrix theory identity[11]

$$\lambda\big(\lambda I + \widehat{\Sigma}\big)^{-1} \approx \kappa\big(\kappa I + \Sigma\big)^{-1}.$$

To further connect Hypothesis 6 to the random matrix literature, we state here a stronger version of Hypothesis 6 (in that it implies Hypothesis 6) that has been shown to hold for commonly studied random matrix models (Knowles & Yin, 2017). While this stronger hypothesis provides uniform convergence for *complex*-valued $\lambda$, we will only need uniform convergence for $\lambda$ on the positive real line as in Hypothesis 6.

Let $\Omega \coloneqq \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$. The stronger hypothesis, in terms of the asymptotic Stieltjes transform $m$, is as follows:

**Hypothesis 11** (Local Marchenko-Pastur law over $\Omega \setminus \mathbb{R}$)**.** The *local Marchenko-Pastur law* holds over an open set $S \subseteq \Omega \setminus \mathbb{R}$ if for every deterministic vector $v \in \mathbb{R}^P$ such that $v^\mathsf{T} \Sigma v \leq 1$, both

$$\left| m(z) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_i - z} \right| \lesssim N^{-\frac{1}{2} + o(1)} \sqrt{\frac{\mathrm{Im}(m(z))}{\mathrm{Im}(z)}} \tag{15}$$

and

$$\left| v^\mathsf{T} \big(I + m(z)\Sigma\big)^{-1} v - v^\mathsf{T} \big(I - z^{-1}\widehat{\Sigma}\big)^{-1} v \right| \lesssim N^{-\frac{1}{2} + o(1)} \sqrt{\frac{\mathrm{Im}(m(z))}{\mathrm{Im}(z)}}. \tag{16}$$

hold uniformly over all $z \in S$.

While we do not make further assumptions, we note that Hypothesis 11 is known to hold under general, *non-asymptotic* assumptions, which subsume the typical random matrix theory assumptions of Gaussian covariates and fixed dimensional ratio $P/N$. For instance, Knowles & Yin (2017, Theorem 3.16 and Remark 3.17) show that Hypothesis 11 holds for any open $S \subseteq \Omega \setminus \mathbb{R}$ if the following conditions are satisfied, for an a priori fixed (large) constant $C > 0$:

- *Sufficient independence.* The following two assumptions hold:
  - The covariates $x \sim \mathcal{D}$ are distributed as a linear transformation $Tz$ of independent (but not necessarily identically distributed) random variables $z_1, \ldots, z_P$ such that $\mathbb{E}[z_i] = 0$, and $\mathbb{E}[z_i^2] = 1$ for all $i$.[12,13]
  - At least a $C^{-1}$ fraction of the eigenvalues of $\Sigma$ are at least $C^{-1}$, and $\|\Sigma\|_{\mathrm{op}} \leq C$ (i.e., the spectrum of $\Sigma$ is not concentrated at 0 relative to $\|\Sigma\|_{\mathrm{op}}$).[14]
- *Bounded moments.* The random variables $z_1, \ldots, z_P$ have uniformly bounded $p$-th moments for all $p < \infty$.
- *Bounded domain.* The domain $S$ is such that $C^{-1} \leq |z| \leq C$ for all $z \in S$.

[11] In comparison, the classical Marchenko-Pastur law (Marchenko & Pastur, 1967) derives $\mathrm{Tr}\big((\lambda I + \widehat{\Sigma})^{-1}\big) \approx \frac{\kappa}{\lambda} \mathrm{Tr}\big((\kappa I + \Sigma)^{-1}\big)$ over the complex plane, from which it follows that the spectral measure of $\widehat{\Sigma}$ converges to the measure whose Stieltjes transform is given by the right-hand side.

[12] The assumption $\mathbb{E}[z_i^2] = I$ is without loss: we can absorb any scaling of $z_i$ into $T$.

[13] To see the necessity of this condition, note that if $x = z_1 \cdot (1, 1, \ldots, 1)^\mathsf{T}$, then we would not obtained the desired convergence.

[14] To see the necessity of this condition, note that if we allowed for $T = (1, 1, \ldots, 1)^\mathsf{T} \cdot (1, 0, 0, \ldots, 0)$ (in which case $\Sigma$ would have only one non-zero eigenvalue), then we would again have $x = z_1 \cdot (1, 1, \ldots, 1)^\mathsf{T}$.

- *Log-bounded dimensional ratio*[15]. The dimensions $N$, $P$ satisfy $N^{1/C} \leq P \leq N^C$.

The non-asymptotic nature of the dimensional ratio assumption is particularly relevant to us because $N$ varies while $P \gg N$ is fixed when we study scaling in our empirical setting. As a consequence, the dimensional ratio $P/N$ takes on a wide range of values. (In contrast, the classical asymptotic assumptions of $P \to \infty$ and $P/N \to \gamma$ are insufficient for our purposes.)

The following lemma shows that Hypothesis 11 implies Hypothesis 6 (note the change in sign due to $z = -\lambda$):

**Lemma 12.** *Let $S \subseteq \Omega$ be open. If Hypothesis 11 holds on $S \setminus \mathbb{R}$, then Hypothesis 6 holds on $\{\lambda : -\lambda \in S \cap \mathbb{R}\}$.*

*Proof.* Fix $\lambda \in S$. Consider $z = -\lambda + \mathrm{i}\eta$ in the limit $\eta \to 0^+$. Because $S$ is open, Hypothesis 11 holds for $z = -\lambda + \mathrm{i}\eta$ in a (complex) neighborhood of $\lambda$. Since $m$ maps reals to reals, $\lim_{\eta \to 0^+} \operatorname{Im}(m(z))/\operatorname{Im}(z) = \frac{\partial}{\partial \eta} \operatorname{Im}(m(-\lambda)) = m'(-\lambda)$ by the Cauchy-Riemann equations. Moreover, $m'(-\lambda) = \frac{1}{\kappa^2} \frac{\partial \kappa}{\partial \lambda}$. Hence

$$\left| m(-\lambda) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_i + \lambda} \right| = \lim_{\eta \to 0^+} \left| m(z) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_i - z} \right| \lesssim \lim_{\eta \to 0^+} N^{-\frac{1}{2}+o(1)} \sqrt{\frac{\operatorname{Im}(m(z))}{\operatorname{Im}(z)}} = N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\kappa} \sqrt{\frac{\partial \kappa}{\partial \lambda}}$$

for (10) and likewise for (11). □

### A.1.5. The Omniscient Risk Estimate

Recent works (Hastie et al., 2020; Canatar et al., 2021; Wu & Xu, 2020; Jacot et al., 2020b; Loureiro et al., 2021; Richards et al., 2021; Mel & Ganguli, 2021; Simon et al., 2021) have shown under a variety of random matrix assumptions that the generalization risk $\mathcal{R}(\hat{\beta}_\lambda)$ of ridge regression can be approximated by the *omniscient risk estimate* $\mathcal{R}^\lambda_{\text{omni}}$:

$$\mathcal{R}^\lambda_{\text{omni}} \coloneqq \frac{\partial \kappa}{\partial \lambda} \cdot \kappa^2 \sum_{i=1}^{P} \left( \frac{\lambda_i}{(\kappa + \lambda_i)^2} \left(\beta^\mathsf{T} v_i\right)^2 \right) = \frac{\partial \kappa}{\partial \lambda} \kappa^2 \beta^\mathsf{T} (\Sigma + \kappa I)^{-1} \Sigma (\Sigma + \kappa I)^{-1} \beta. \tag{17}$$

The analysis of Hastie et al. (2020) is the most general of these and establishes (17) under a similar set of assumptions as Hypothesis 11, with approximation error proportional to $\|\beta\|_2^2$.

However, in our empirical setting with effectively infinite $\|\beta\|_2$, we need a stronger version of this result than was previously known. Thus, we improve the result of Hastie et al. (2020) so that the error bound scales in the expected size of the label $\beta^\mathsf{T}\Sigma\beta$ rather than the squared norm $\|\beta\|_2^2$ (see Proposition 10). To prove this generalization requires a more careful analysis, as the analysis of Hastie et al. (2020) does not directly extend to large $\|\beta\|_2$.

## A.2. Bounding the Derivative of a Bounded, Real Analytic Function

A key step of our analysis will be arguing that we may differentiate the local random matrix law, as in (7), while preserving the approximate equality. In this section, we show a general lemma that lets us accomplish this. Concretely, we will bound the derivative of a bounded, real analytic function. Our approach here streamlines the argument of Hastie et al. (2020), allowing for sharper bounds while also being easier to apply.

Let $h \colon U \to \mathbb{R}$, for some $U \subseteq \mathbb{R}$. (In applications, $h$ will represent the difference of two "approximately equal" functions.) Suppose $h$ is real analytic at $x_0$ with radius of convergence $R > 0$. Then $h$ has an analytic continuation $\tilde{h}$ to the open ball $V \coloneqq \{z \in \mathbb{C} : |z - x_0| < R\}$. Let $K \subseteq V$ be the closed ball $\{z \in \mathbb{C} : |z - x_0| \leq \frac{1}{2}R\}$. Given that $h$ and $\tilde{h}$ are bounded on $K \cap \mathbb{R}$ and $K$, respectively, the next lemma bounds $h'(x_0)$ with only a *logarithmic* dependence on the bound on $\tilde{h}$. In our applications, this logarithmic dependence will be negligible: the dominant factor will be the ratio $\delta/R$.

**Lemma 13.** *Suppose $M \geq \delta > 0$, and $h \colon U \to \mathbb{R}$ is such that $|h(x)| \leq \delta$ on $K \cap \mathbb{R}$ and $|\tilde{h}(z)| \leq M$ on $K$. Then,*

$$|h'(x_0)| \lesssim \frac{\delta}{R} \left(1 + \log\left(\frac{M}{\delta}\right)\right)^2.$$

[15] Knowles & Yin (2017, Section 2.1) note that their results can be obtained under a relaxed dimensional ratio assumption using the techniques of Bloemendal et al. (2016).

*Proof.* Given the power series expansion $h(x) = \sum_{j=0}^{\infty} c_j(x - x_0)^j$ of $h$ at $x_0$, the Cauchy integral formula tells us that

$$|c_j| = \left| \frac{1}{2\pi i} \int_{\partial K} \frac{\tilde{h}(z)}{(z - x_0)^{j+1}}\, dz \right| \leq \left(\frac{2}{R}\right)^j M.$$

Let $h_k(x) \coloneqq \sum_{j=0}^{k} c_j(x - x_0)^j$ the $k$-th order Taylor expansion of $h$ at $x_0$. If $I \coloneqq \left[x_0 - \frac{1}{4}R, x_0 + \frac{1}{4}R\right]$ and $x \in I$, then

$$|h(x) - h_k(x)| = \left| \sum_{j=k+1}^{\infty} c_j(x - x_0)^j \right| \leq \sum_{j=k+1}^{\infty} |c_j||x - x_0|^j \leq 2^{-k} M.$$

Let $\|\cdot\|_\infty$ denote the sup norm for continuous functions $I \to \mathbb{R}$. Setting $k \coloneqq \lfloor 1 + \log_2(M/\delta) \rfloor$, we have by the triangle inequality that $\|h_k\|_\infty \leq \|h\|_\infty + \|h - h_k\|_\infty \leq 2\delta$. Let $\mathcal{P}_k$ be the vector space of degree $k$ polynomial functions $I \to \mathbb{R}$. The Markov brothers' inequality says that the linear functional $\mathcal{P}_k \to \mathbb{R}$ given by $p \mapsto p'(x_0)$ has operator norm at most $4k^2/R$ with respect to $\|\cdot\|_\infty$. Hence

$$|h'(x_0)| = |h_k'(x_0)| \leq \frac{4k^2}{R} \|h_k\|_\infty \lesssim \frac{\delta}{R}\left(1 + \log\left(\frac{M}{\delta}\right)\right)^2. \qquad \square$$

### A.3. Proof of Proposition 9

To prove Proposition 9, we follow the outline in Section 5. Define

$$f(\lambda) \coloneqq \beta^\mathsf{T}\beta - \beta^\mathsf{T}\lambda\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta \quad \text{and} \quad g(\lambda) \coloneqq \beta^\mathsf{T}\beta - \beta^\mathsf{T}\kappa\big(\Sigma + \kappa I\big)^{-1}\beta.$$

The $\beta^\mathsf{T}\beta$ terms in $f$ and $g$ ensure that $f$ and $g$ can be bounded, so that we may apply Lemma 13. Additionally, define

$$h(\lambda) \coloneqq f(\lambda) - g(\lambda) = -\beta^\mathsf{T}\lambda\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta + \beta^\mathsf{T}\kappa\big(\Sigma + \kappa I\big)^{-1}\beta.$$

Note that $f$, $g$, and $h$ may be analytically continued to take complex arguments $w = \lambda - \mathrm{i}\eta$, since we may take $\kappa = 1/m(-w)$. We will also need these extended functions when applying Lemma 13.

Algebraically, the key drivers of our analysis are the relationships obtained from differentiating $f$ and $g$ with respect to $\lambda$:

$$f'(\lambda) = \beta^\mathsf{T}\big(\widehat{\Sigma} + \lambda I\big)^{-1}\widehat{\Sigma}\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta = \frac{1}{\lambda^2}\mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda) = \left(\sum_{i=1}^{N} \frac{1}{\lambda + \hat{\lambda}_i}\right)^2 \mathrm{GCV}_\lambda$$

$$g'(\lambda) = \frac{\partial \kappa}{\partial \lambda}\beta^\mathsf{T}\big(\Sigma + \kappa I\big)^{-1}\Sigma\big(\Sigma + \kappa I\big)^{-1}\beta = \frac{1}{\kappa^2}\mathcal{R}^{\lambda}_{\text{omni}}.$$

The main technical steps in the analysis will be to bound $|h'(\lambda)| = |f'(\lambda) - g'(\lambda)|$ and $|\kappa^2 f'(\lambda) - \mathrm{GCV}_\lambda|$, so that we may relate $\mathrm{GCV}_\lambda$ and $\mathcal{R}^{\lambda}_{\text{omni}}$. The former we will bound via Lemma 13; the latter we will bound using Hypothesis 6.

#### A.3.1. Auxiliary Lemmas

We now set up the lemmas that let us formalize our heuristic argument from Section 5.

The next three lemmas note some basic properties of the effective regularization $\kappa$:

**Lemma 14.** *For all $\lambda > 0$, $\kappa = \kappa(\lambda, N)$ satisfies*

$$1 \leq \frac{\partial \kappa}{\partial \lambda} \leq \frac{\kappa}{\lambda} \leq 1 + \frac{\mathrm{Tr}(\Sigma)}{N\lambda}.$$

*Proof.* Rearranging (13) gives us

$$\kappa = \lambda + \frac{1}{N}\sum_{i=1}^{P} \lambda_i \left(1 - \frac{\lambda_i}{\kappa + \lambda_i}\right) \leq \lambda + \frac{\mathrm{Tr}(\Sigma)}{N}.$$

Dividing by $\lambda$ immediately yields $\frac{\kappa}{\lambda} \le 1 + \frac{\mathrm{Tr}(\Sigma)}{N\lambda}$. To get the first two inequalities, we compute $\frac{\partial \kappa}{\partial \lambda}$. By the implicit function theorem applied to (13), $\frac{\partial \kappa}{\partial \lambda}$ satisfies

$$\frac{\partial \kappa}{\partial \lambda} = 1 + \frac{\partial \kappa}{\partial \lambda} \cdot \frac{1}{N} \sum_{i=1}^{P} \frac{\lambda_i^2}{(\kappa + \lambda_i)^2}.$$

Solving for $\frac{\partial \kappa}{\partial \lambda}$, we obtain

$$\frac{\partial \kappa}{\partial \lambda} = \frac{1}{1 - \frac{1}{N} \sum_{i=1}^{P} \frac{\lambda_i^2}{(\kappa + \lambda_i)^2}}. \tag{18}$$

From here, it is clear that $\frac{\partial \kappa}{\partial \lambda} \ge 1$. And the upper bound $\frac{\partial \kappa}{\partial \lambda} \le \frac{\kappa}{\lambda}$ follows from the fact that

$$1 - \frac{1}{N} \sum_{i=1}^{P} \frac{\lambda_i^2}{(\kappa + \lambda_i)^2} \ge 1 - \frac{1}{N} \sum_{i=1}^{P} \frac{\lambda_i}{\kappa + \lambda_i} = \frac{\lambda}{\kappa}.$$ □

**Lemma 15.** *Suppose $\kappa = \kappa(\lambda, N)$ and $\tilde{\kappa} = 1/m(-\lambda + \mathrm{i}\eta)$ for $\lambda, \eta > 0$. Then $\mathrm{Re}(\tilde{\kappa}) \ge \kappa$.*

*Proof.* Note that $\mathrm{Re}(\tilde{\kappa})$ satisfies

$$\mathrm{Re}(\tilde{\kappa}) = \lambda + \frac{1}{N} \sum_{i=1}^{P} \lambda_i \left(1 - \mathrm{Re}\left(\frac{\lambda_i}{\tilde{\kappa} + \lambda_i}\right)\right) \ge \lambda + \frac{1}{N} \sum_{i=1}^{P} \lambda_i \left(1 - \frac{\lambda_i}{\mathrm{Re}(\tilde{\kappa}) + \lambda_i}\right).$$

On the other hand, since $\kappa$ is the unique positive solution to (13) and $0 < \lambda$, it holds for all $\kappa' \in [0, \kappa)$ that

$$\kappa' < \lambda + \frac{1}{N} \sum_{i=1}^{P} \lambda_i \left(1 - \frac{\lambda_i}{\kappa' + \lambda_i}\right).$$

Therefore, it must be the case that $\mathrm{Re}(\tilde{\kappa}) \ge \kappa$. □

**Lemma 16.** *Suppose $\lambda > 0$, and let $\kappa = \kappa(\lambda, N)$. If $\lambda' > \frac{1}{2}\lambda$, then*

$$\frac{1}{\kappa(\lambda', N)} \sqrt{\frac{\partial \kappa}{\partial \lambda}(\lambda', N)} \lesssim \frac{1}{\kappa} \sqrt{\frac{\partial \kappa}{\partial \lambda}}.$$

*Proof.* The left- and right-hand sides of the desired inequality are simply $\sqrt{m'(-\lambda')}$ and $\sqrt{m'(-\lambda)}$, respectively. Define $t \coloneqq \lambda'/\lambda$. Then it suffices to show $m'(-t\lambda) \lesssim m'(-\lambda)$ for all $t > \frac{1}{2}$. By the integral representation (14) of $m$,

$$m'(-t\lambda) = \int \frac{d\varrho(x)}{(x + t\lambda)^2} \le \frac{1}{(\min(t, 1))^2} \int \frac{d\varrho(x)}{(x + \lambda)^2} \lesssim m'(-\lambda),$$

where we use the elementary inequality $\min(t, 1)/(x + t\lambda) \le 1/(x + \lambda)$. □

Using properties of $\kappa$, we bound $|f(w)|$ and $|g(w)|$ for complex $w = \lambda - i\eta$ so that we may later apply Lemma 13.

**Lemma 17.** *Suppose $\beta^\mathsf{T} \Sigma \beta \le 1$. Then functions $f$ and $g$ satisfies the bounds*

$$\mathbb{E}\left[\sup_{\mathrm{Re}(w) \ge \lambda_0} |f(w)|\right] \le \frac{1}{\lambda_0} \quad \textit{and} \quad \sup_{\mathrm{Re}(w) \ge \lambda_0} |g(w)| \le \frac{1}{\lambda_0}.$$

*Proof.* We first bound $f(w)$ as follows:

$$|f(w)| = \left|\beta^\mathsf{T} \widehat{\Sigma} \left(\widehat{\Sigma} + wI\right)^{-1} \beta\right| = \left|y^\mathsf{T} \left(XX^\mathsf{T} + N \cdot wI\right)^{-1} y\right| \le \left\|\left(XX^\mathsf{T} + N \cdot wI\right)^{-1}\right\|_{\mathrm{op}} \|y\|_2^2 \le \frac{1}{\lambda_0} \frac{1}{N} \sum_{i=1}^{N} y_i^2.$$

Note that we used the fact that $\left(\frac{1}{N}XX^\mathsf{T} + wI\right)^{-1}$ is normal to bound its operator norm by its spectral radius. Our bound on $|f(w)|$ holds uniformly over all $w$ such that $\mathrm{Re}(w) \geq \lambda_0$. Hence, taking an expectation, we have

$$\mathbb{E}\left[\sup_{\mathrm{Re}(w)\geq\lambda_0} |f(w)|\right] \leq \mathbb{E}\left[\frac{1}{\lambda_0}\frac{1}{N}\sum_{i=1}^{N} y_i^2\right] = \frac{1}{\lambda_0}\beta^\mathsf{T}\Sigma\beta \leq \frac{1}{\lambda_0}.$$

We also have

$$|g(w)| = \left|\beta^\mathsf{T}\Sigma\big(\Sigma+\kappa I\big)^{-1}\right| = \left|\beta^\mathsf{T}\Sigma^{1/2}\big(\Sigma+\kappa I\big)^{-1}\Sigma^{1/2}\beta\right| \leq \beta^\mathsf{T}\Sigma\beta\cdot\left\|\big(\Sigma+\kappa I\big)^{-1}\right\|_{\mathrm{op}} \leq \frac{1}{\mathrm{Re}(\kappa)} \leq \frac{1}{\lambda_0},$$

where the last inequality follows from Lemmas 14 and 15. □

### A.3.2. Proof of Proposition 9

*Proof of Proposition 9.* We first bound $|h'(\lambda)| = |f'(\lambda) - g'(\lambda)|$ by applying Lemma 13 to $h$ with $U := (0, 2\lambda)$. By Hypothesis 6 and Lemma 16, we may take

$$\delta \lesssim N^{-\frac{1}{2}+o(1)}\frac{1}{\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

And by Lemma 17, $|g(w)| \leq 1/\lambda$ when $\mathrm{Re}(w) \geq \frac{1}{2}\lambda$. Setting $M = N^D/\lambda$, Lemma 17 together with Markov's inequality gives us the high probability bound

$$\mathbb{P}\left[\sup_{\mathrm{Re}(w)\geq\frac{1}{2}\lambda} |f(w)| \geq M\right] \leq N^{-D}.$$

Therefore, by Lemma 13,

$$|f'(\lambda) - g'(\lambda)| = |h'(\lambda)| \lesssim \frac{\delta}{\lambda}\log\left(\frac{M}{\delta}\right) \lesssim N^{-\frac{1}{2}+o(1)}\cdot\frac{1}{\lambda\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}}. \tag{19}$$

We now bound the error of $\mathrm{GCV}_\lambda$. Substituting the closed form (12) for $\mathcal{R}_{\mathrm{empirical}}(\hat{\beta}_\lambda)$ into the definition of $\mathrm{GCV}_\lambda$, we have that

$$\mathrm{GCV}_\lambda = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda+\hat{\lambda}_i}\right)^{-2}\beta^\mathsf{T}\big(\widehat{\Sigma}+\lambda I\big)^{-1}\widehat{\Sigma}\big(\widehat{\Sigma}+\lambda I\big)^{-1}\beta.$$

Let $\hat{\kappa} := \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda+\hat{\lambda}_i}\right)^{-1}$. By Hypothesis 6,

$$\left|1-\frac{\kappa}{\hat{\kappa}}\right| \lesssim N^{-\frac{1}{2}+o(1)}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

For sufficiently large $N$, the right-hand side is less than $\frac{1}{2}$, which implies $\kappa \geq \frac{1}{2}\hat{\kappa}$. Therefore,

$$\left|\kappa^2-\hat{\kappa}^2\right| \leq (\kappa+\hat{\kappa})\cdot|\kappa-\hat{\kappa}| \leq 3\kappa\cdot\hat{\kappa}\cdot\left|1-\frac{\kappa}{\hat{\kappa}}\right| \lesssim \kappa^2 N^{-\frac{1}{2}+o(1)}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

This yields the comparison

$$\left|\mathrm{GCV}_\lambda - \kappa^2 f'(\lambda)\right| \lesssim f'(\lambda)\cdot\kappa^2 N^{-\frac{1}{2}+o(1)}\sqrt{\frac{\partial\kappa}{\partial\lambda}} \lesssim g'(\lambda)\cdot\kappa^2 N^{-\frac{1}{2}+o(1)}\sqrt{\frac{\partial\kappa}{\partial\lambda}} \leq N^{-\frac{1}{2}+o(1)}\left(\frac{\partial\kappa}{\partial\lambda}\right)^{3/2}$$

where we applied (19) to get the third expression. We further have from (19) that

$$\left|\kappa^2 f'(\lambda) - \mathcal{R}^\lambda_{\mathrm{omni}}\right| \lesssim N^{-\frac{1}{2}+o(1)}\cdot\frac{\kappa}{\lambda}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

Thus, the triangle inequality followed by Lemma 14 implies

$$\left|\mathrm{GCV}_\lambda - \mathcal{R}^\lambda_{\mathrm{omni}}\right| \lesssim N^{-\frac{1}{2}+o(1)}\left(\left(\frac{\partial\kappa}{\partial\lambda}\right)^{3/2} + \frac{\kappa}{\lambda}\sqrt{\frac{\partial\kappa}{\partial\lambda}}\right) \lesssim N^{-\frac{1}{2}+o(1)}\left(1+\frac{\mathrm{Tr}(\Sigma)}{N\lambda}\right)^{3/2}.$$

□

### A.4. Proof of Proposition 10

As we did for Proposition 9, we first outline a heuristic proof. Let $U \coloneqq (-\frac{1}{2}\|\Sigma\|_{\mathrm{op}}^{-1}, \frac{1}{2}\|\Sigma\|_{\mathrm{op}}^{-1})$. For $t \in U$ and $\lambda > 0$, let $\tilde{\kappa} = \tilde{\kappa}(t, \lambda, N)$ denote the asymptotic Stieltjes transform associated to the covariance matrix $\Sigma(I + t\Sigma)^{-1}$, and define

$$f(t) \coloneqq \beta^{\mathsf{T}}\big(I + t\Sigma\big)^{-1}\beta - \beta^{\mathsf{T}}\lambda\big(\widehat{\Sigma} + \lambda(I + t\Sigma)\big)^{-1}\beta,$$
$$g(t) \coloneqq \beta^{\mathsf{T}}\big(I + t\Sigma\big)^{-1}\beta - \beta^{\mathsf{T}}\tilde{\kappa}\big(\Sigma + \tilde{\kappa}(I + t\Sigma)\big)^{-1}\beta,$$

and

$$h(t) \coloneqq f(t) - g(t) = -\beta^{\mathsf{T}}\lambda\big(\widehat{\Sigma} + \lambda(I + t\Sigma)\big)^{-1}\beta + \beta^{\mathsf{T}}\tilde{\kappa}\big(\Sigma + \tilde{\kappa}(I + t\Sigma)\big)^{-1}\beta.$$

Letting $\widetilde{m} = 1/\tilde{\kappa}$, note that

$$f'(0) = \lambda^2\beta^{\mathsf{T}}\big(\widehat{\Sigma} + \lambda I\big)^{-1}\Sigma\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta = \mathcal{R}(\hat{\beta}_\lambda) \quad \text{and} \quad g'(0) = \left(1 + \frac{\partial\widetilde{m}}{\partial t}\right)\kappa^2\beta^{\mathsf{T}}\big(\Sigma + \kappa I\big)^{-1}\Sigma\big(\Sigma + \kappa I\big)^{-1}\beta.$$

We will show that $1 + \frac{\partial\widetilde{m}}{\partial t} = \frac{\partial\kappa}{\partial\lambda}$ (see Lemma 23), in which case $g'(0) = \mathcal{R}^{\lambda}_{\mathrm{omni}}$. Proposition 10 thus follows, predicated on $h(t) \approx 0$ and differentiation preserving the approximate equality.

#### A.4.1. Auxiliary Lemmas

We now set up the lemmas that let us formalize this heuristic argument. First, we show that $h(t) \approx 0$.

**Lemma 18.** *Suppose $\beta^{\mathsf{T}}\Sigma\beta \le 1$ and Hypothesis 7 holds over $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$. Then,*

$$\left|\tilde{\beta}^{\mathsf{T}}\lambda\big(\widehat{\widetilde{\Sigma}} + \lambda I\big)^{-1}\tilde{\beta} - \tilde{\beta}^{\mathsf{T}}\tilde{\kappa}\big(\widetilde{\Sigma} + \tilde{\kappa} I\big)^{-1}\tilde{\beta}\right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\tilde{\kappa}}\sqrt{\frac{\partial\tilde{\kappa}}{\partial\lambda}}.$$

*Proof.* Let $Q \coloneqq I + t\Sigma$. That $t \in U$ implies $Q \succeq \frac{1}{2}I$. Further, define $\widetilde{\Sigma} \coloneqq Q^{-\frac{1}{2}}\Sigma Q^{-\frac{1}{2}}$, $\widetilde{X} \coloneqq XQ^{-\frac{1}{2}}$, $\widehat{\widetilde{\Sigma}} \coloneqq \frac{1}{N}\widetilde{X}^{\mathsf{T}}\widetilde{X}$, and $\tilde{\beta} \coloneqq Q^{-\frac{1}{2}}\beta$. Note that

$$\begin{aligned} h(t) &= -\beta^{\mathsf{T}}Q^{-\frac{1}{2}}\lambda\big(Q^{-\frac{1}{2}}\widehat{\Sigma}Q^{-\frac{1}{2}} + \lambda I\big)^{-1}Q^{-\frac{1}{2}}\beta + \beta^{\mathsf{T}}Q^{-\frac{1}{2}}\tilde{\kappa}\big(Q^{-\frac{1}{2}}\Sigma Q^{-\frac{1}{2}} + \tilde{\kappa} I\big)^{-1}Q^{-\frac{1}{2}}\beta \\ &= -\tilde{\beta}^{\mathsf{T}}\lambda\big(\widehat{\widetilde{\Sigma}} + \lambda I\big)^{-1}\tilde{\beta} + \tilde{\beta}^{\mathsf{T}}\tilde{\kappa}\big(\widetilde{\Sigma} + \tilde{\kappa} I\big)^{-1}\tilde{\beta}. \end{aligned}$$

Because $Q$ and $\Sigma$ commute, $\tilde{\beta}^{\mathsf{T}}\widetilde{\Sigma}\tilde{\beta} = \beta^{\mathsf{T}}\Sigma^{\frac{1}{2}}Q^{-2}\Sigma^{\frac{1}{2}}\beta \le \|Q^{-1}\|_{\mathrm{op}}^2 \le 4$. By Hypothesis 7, since $\widetilde{\Sigma} = \Sigma(I + t\Sigma)^{-1}$,

$$\left|\tilde{\beta}^{\mathsf{T}}\lambda\big(\widehat{\widetilde{\Sigma}} + \lambda I\big)^{-1}\tilde{\beta} - \tilde{\beta}^{\mathsf{T}}\tilde{\kappa}\big(\widetilde{\Sigma} + \tilde{\kappa} I\big)^{-1}\tilde{\beta}\right| \lesssim \tilde{\beta}^{\mathsf{T}}\widetilde{\Sigma}\tilde{\beta} \cdot N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\tilde{\kappa}}\sqrt{\frac{\partial\tilde{\kappa}}{\partial\lambda}} \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\tilde{\kappa}}\sqrt{\frac{\partial\tilde{\kappa}}{\partial\lambda}}. \qquad \square$$

The next two lemmas verify that the conditions for applying Lemma 13 hold. Verifying these conditions turns out to be the most technically challenging part of our analysis. Lemma 19 shows that we can analytically continue $\tilde{\kappa}$ (which we only defined for $t \in U \subseteq \mathbb{R}$) to the complex plane. It follows from Lemma 19 that $f$ and $g$ can be analytically continued over the same domain. We then check in Lemma 20 that this analytic continuation is bounded with high probability.

Our analysis for Lemma 19 extends $\tilde{\kappa}$ using a fixed point definition of effective regularization. This argument proceeds in three steps: (i) we show for each $w = t - \mathrm{i}\eta$ that a fixed point exists using the Brouwer fixed point theorem; (ii) we argue that this fixed point is unique via the Schwarz lemma; (iii) we verify that the set of fixed points defined by these $w$ give rise to a holomorphic function using the implicit function theorem and the Schwarz reflection principle.

Proving Lemma 20 in the case of $f$ requires a more involved analysis than its analog Lemma 17. The previous approach based on diagonalizing the positive semidefinite matrix $\widehat{\Sigma}$ fails because $\widehat{\Sigma} + \lambda(I + w\Sigma)$ is no longer normal when $w$ is complex. (The failure of normality arises because $\Sigma$ and $\widehat{\Sigma}$ do not commute.) While the same $1/\lambda$ bound still holds, proving it is much more difficult; our argument makes careful use of the properties of symmetric matrices $A + \mathrm{i}B$ with positive definite real part $A \succ 0$.

**Lemma 19.** *The effective regularization $\tilde{\kappa}(t, \lambda, N)$ has an analytic continuation in $t$ to the strip $\{z \in \mathbb{C} : \mathrm{Re}(z) \in U\}$.*

*Proof of Lemma 19.* For fixed $\lambda > 0$ and $t \in U$, define

$$\varphi_{\lambda,t}(z) \coloneqq \lambda + \frac{1}{N}\sum_{i=1}^{P}\left(\frac{1}{z} + \frac{1}{\tilde{\lambda}_i}\right)^{-1},$$

where $\tilde{\lambda}_i \coloneqq \lambda_i/(1+t\lambda_i)$ is the $i$-th eigenvalue of $\widetilde{\Sigma}$.[16] Note that (13) for $\tilde{\kappa} = \tilde{\kappa}(t,\lambda,N)$ can be rearranged to $\tilde{\kappa} = \varphi_{\lambda,t}(\tilde{\kappa})$. That is, we can define $\tilde{\kappa}$ as the unique fixed point of $\varphi_{\lambda,t}$ on $\mathbb{R}_{>0}$.

We extend this definition from $t \in U$ to $w$ in the complex plane. Suppose $w = t - \mathrm{i}\eta$ satisfies $t \in U$ and $\eta > 0$. (We will handle $\eta < 0$ via the Schwarz reflection principle.) Define $\tilde{\lambda}_i \coloneqq \lambda_i/(1+w\lambda_i)$ and $\varphi_{\lambda,w}(z)$ as above. Since $t \in U$ and $\eta > 0$, we have $\mathrm{Re}(\tilde{\lambda}_i) > 0$ and $\mathrm{Im}(\tilde{\lambda}_i) > 0$ for all $i$. Let $\tilde{\kappa}(w,\lambda,N)$ be the unique fixed point of $\varphi_{\lambda,w}$ in $\mathbb{H}$. We validate that $\tilde{\kappa}$ is well-defined as a holomorphic function in $w$ through the three steps outlined above.

We show the existence of $\tilde{\kappa}$ by applying the Brouwer fixed point theorem to $\varphi_{\lambda,w}$ acting on the compact, convex set

$$K \coloneqq \{z \in \mathbb{C} : \mathrm{Re}(z) \geq \lambda, \mathrm{Im}(z) \geq 0, |z| \leq M\},$$

where $M \coloneqq \lambda + \sum_{i=1}^{P}\big(\mathrm{Re}(1/\tilde{\lambda}_i)\big)^{-1}$. We first verify that $\varphi_{\lambda,w}$ maps $K$ into $K$. Let $z \in K$ and $q_i \coloneqq 1/z + 1/\tilde{\lambda}_i$. Then $\mathrm{Re}(q_i) > 0$ and $\mathrm{Im}(q_i) < 0$, which in turn implies $\mathrm{Re}(q_i^{-1}) > 0$ and $\mathrm{Im}(q_i^{-1}) > 0$. Hence,

$$\mathrm{Re}(\varphi_{\lambda,w}(z)) = \lambda + \sum_{i=1}^{P}\mathrm{Re}(q_i^{-1}) > \lambda \quad \text{and} \quad \mathrm{Im}(\varphi_{\lambda,w}(z)) = \sum_{i=1}^{P}\mathrm{Im}(q_i^{-1}) > 0.$$

And by the triangle inequality,

$$|\varphi_{\lambda,w}(z)| \leq \lambda + \sum_{i=1}^{P}\frac{1}{|q_i|} < \lambda + \sum_{i=1}^{P}\frac{1}{\mathrm{Re}(1/\tilde{\lambda}_i)} = M.$$

These bounds show that $\varphi_{\lambda,w}$ maps $K$ into the interior of $K$. By the Brouwer fixed point theorem, $\varphi_{\lambda,w}$ has a fixed point $\tilde{\kappa}$ in the interior of $K$. In particular, this fixed point satisfies $\tilde{\kappa} \in \mathbb{H}$.

We now argue that this fixed point $\tilde{\kappa}$ is unique over all $z \in \mathbb{H}$. Following the above argument, one sees that $\varphi_{\lambda,w}$ maps $\mathbb{H}$ to $\mathbb{H}$. Moreover, $\varphi_{\lambda,w}$ is not the identity map. It is then a standard consequence of the Schwarz lemma that $\varphi_{\lambda,w}$ has at most one fixed point: We may identify $\mathbb{H}$ with the unit disk using a biholomorphic map that sends $\tilde{\kappa}$ to 0. (Such a map exists by the Riemann mapping theorem.) The induced automorphism on the unit disk cannot fix any other point—otherwise the Schwarz lemma would imply that it is the identity. Thus, $\varphi_{\lambda,w}$ has at most one fixed point.

Having shown that $\tilde{\kappa}$ is well-defined for each $w = t - \mathrm{i}\eta$, we now verify that it defines a holomorphic function over the set of such $w$. By the (holomorphic) implicit function theorem, if $\frac{\partial}{\partial z}(z - \varphi_{\lambda,w}(z)) \neq 0$ at $z = \tilde{\kappa}$, then we can extend $\tilde{\kappa}$ to a holomorphic function such that $\tilde{\kappa}(z) = \varphi_{\lambda,z}(\tilde{\kappa}(z))$ in a neighborhood of $w$. By continuity, $\mathrm{Im}(\tilde{\kappa}(z)) > 0$ in a neighborhood of $w$. Uniqueness then implies that this function coincides with our definition of $\tilde{\kappa}$ in this neighborhood. In particular, $\tilde{\kappa}$ is holomorphic at $w$. It remains to check that $\frac{\partial}{\partial z}(z - \varphi_{\lambda,w}(z)) \neq 0$ at $z = \tilde{\kappa}$. Substituting in (13),

$$\begin{aligned}
\frac{\partial}{\partial z}(z - \varphi_{\lambda,w}(z))\Big|_{z=\tilde{\kappa}} &= 1 - \frac{1}{N}\sum_{i=1}^{P}\frac{\tilde{\lambda}_i^2}{(\tilde{\kappa}+\tilde{\lambda}_i)^2} \\
&= \frac{\lambda}{\tilde{\kappa}} + \frac{1}{N}\sum_{i=1}^{P}\frac{\tilde{\lambda}_i}{\tilde{\kappa}+\tilde{\lambda}_i} - \frac{1}{N}\sum_{i=1}^{P}\frac{\tilde{\lambda}_i^2}{(\tilde{\kappa}+\tilde{\lambda}_i)^2} \\
&= \frac{\lambda}{\tilde{\kappa}} + \frac{1}{N}\sum_{i=1}^{P}\left(\frac{\tilde{\kappa}}{\tilde{\lambda}_i} + 2 + \frac{\tilde{\lambda}_i}{\tilde{\kappa}}\right)^{-1}.
\end{aligned}$$

Note that $\mathrm{Re}(\tilde{\kappa}/\tilde{\lambda}_i), \mathrm{Re}(\tilde{\lambda}_i/\tilde{\kappa}) > 0$ because both $\tilde{\kappa}$ and $\tilde{\lambda}_i$ have positive real and imaginary parts. Thus, each term in the sum has positive real part. Since $\mathrm{Re}(\lambda/\tilde{\kappa}) > 0$ as well, $\mathrm{Re}\big(\frac{\partial}{\partial z}(z - \varphi_{\lambda,w}(z))\big|_{z=\tilde{\kappa}}\big) > 0$.

[16] Technically, we need to handle zero eigenvalues (in which case the inverse $1/\tilde{\lambda}_i$ becomes undefined). But such eigenvalues do not contribute to the definition (13) and thus may safely be ignored. That is, we assume without loss of generality that $\lambda_i > 0$ for all $i$.

Lastly, we confirm $\tilde{\kappa}$ extends continuously to a map $U \to \mathbb{R}$, which lets us conclude that $\tilde{\kappa}$ extends to $w = t - \mathrm{i}\eta$ with $\eta < 0$ by the Schwarz reflection principle. For $t_0 \in U$ and $\tilde{\kappa} > 0$ such that $\tilde{\kappa} = \varphi_{\lambda,t}(\tilde{\kappa})$, the same implicit function theorem argument shows that $\tilde{\kappa}$ extends to a holomorphic function $\tilde{\kappa}(z)$ in a neighborhood of $t_0$. The fixed point condition implies $\tilde{\kappa}$ decreases in $t$, i.e., $\tilde{\kappa}'(t_0) < 0$. Thus, $\tilde{\kappa}(w) \in \mathbb{H}$ for all $w = t - \mathrm{i}\eta$ with $\eta > 0$ in a neighborhood of $t_0$. Uniqueness then implies this $\tilde{\kappa}(w)$ is consistent with the definition of $\tilde{\kappa}$ above, so our definition extends continuously to $U$. □

**Lemma 20.** *Suppose $\beta^\mathsf{T}\Sigma\beta \le 1$. Then functions $f$ and $g$ satisfy the bounds*

$$\mathbb{E}\left[\sup_{\mathrm{Re}(w)\in U} |f(w)|\right] \lesssim \frac{1}{\lambda} \quad \textit{and} \quad \sup_{\mathrm{Re}(w)\in U} |g(w)| \lesssim \frac{1}{\lambda}.$$

Before proving Lemma 20, we prove a lemma about symmetric matrices with positive definite real part. In analogy to how positive definite matrices generalize positive numbers and how symmetric matrices generalize real numbers, we establish how symmetric matrices with positive definite real part generalize complex numbers in the right half-plane.

**Lemma 21.** *Suppose $Q \in \mathbb{C}^{P\times P}$ is such that $A \coloneqq \mathrm{Re}(Q)$ is positive definite and $B \coloneqq \mathrm{Im}(Q)$ is symmetric. Then:*

*(i) $Q$ is invertible, with its inverse $Q^{-1}$ also being symmetric and having positive definite real part;*

*(ii) the spectrum $\sigma(Q)$ of $Q$ satisfies $\sigma(Q) \subseteq \{z \in \mathbb{C} : \mathrm{Re}(z) \ge \|A^{-1}\|_{\mathrm{op}}^{-1}\}$;*

*(iii) the operator norm of $Q^{-1}$ is bounded as $\|Q^{-1}\|_{\mathrm{op}} \le \|A^{-1}\|_{\mathrm{op}}$.*

*Proof.* For (i), let $T = A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$ and write $Q = A^{\frac{1}{2}}(I + \mathrm{i}T)A^{\frac{1}{2}}$. Note that $T^2 \succeq 0$ and so $I + T^2$ is invertible. Thus, we may compute $(I + \mathrm{i}T)\cdot(I - \mathrm{i}T)(I + T^2)^{-1} = I$ to see that $(I + \mathrm{i}T)^{-1} = (I - \mathrm{i}T)(I + T^2)^{-1}$. It follows that

$$\begin{aligned} Q^{-1} &= A^{-\frac{1}{2}}(I - \mathrm{i}T)(I + T^2)^{-1}A^{-\frac{1}{2}} \\ &= A^{-\frac{1}{2}}(I + T^2)^{-1}A^{-\frac{1}{2}} - \mathrm{i}\cdot A^{-\frac{1}{2}}(I + T^2)^{-\frac{1}{2}}T(I + T^2)^{-\frac{1}{2}}A^{-\frac{1}{2}} \\ &= (A + BA^{-1}B)^{-1} - \mathrm{i}\cdot(A^2 + A^{\frac{1}{2}}BA^{-1}BA^{\frac{1}{2}})^{-\frac{1}{2}}B(A^2 + A^{\frac{1}{2}}BA^{-1}BA^{\frac{1}{2}})^{-\frac{1}{2}}. \end{aligned}$$

For (ii), observe that if $\lambda < \|A^{-1}\|_{\mathrm{op}}^{-1}$, then $A \succ \lambda I$. Applying (i), we have that $Q - \lambda I + \mathrm{i}\eta I$ is invertible for all $\eta \in \mathbb{R}$. It follows that $\lambda - \mathrm{i}\eta \notin \sigma(Q)$ for all such $\lambda$ and $\eta$. In other words, $\sigma(Q) \subseteq \{z \in \mathbb{C} : \mathrm{Re}(z) \ge \|A^{-1}\|_{\mathrm{op}}^{-1}\}$.

For (iii), note that $S \coloneqq \overline{Q}^\mathsf{T}Q$ is normal and $\mathrm{Re}(S) = A^2 + B^2 \succeq A^2$. Hence $S^{-1}$ is normal and its operator norm equals its spectral radius. We thus have

$$\|Q^{-1}\|_{\mathrm{op}}^2 = \|S^{-1}\|_{\mathrm{op}} = \sup_{z\in\sigma(S^{-1})}|z| = \sup_{z\in\sigma(S)}\frac{1}{|z|} \le \sup_{z\in\sigma(S)}\frac{1}{|\mathrm{Re}(z)|} \le \|\mathrm{Re}(S)^{-1}\|_{\mathrm{op}} \le \|A^{-1}\|_{\mathrm{op}}^2,$$

where the penultimate inequality applies (ii) to $S$. □

*Proof of Lemma 20.* We start by bounding $\mathbb{E}\big[\sup_{\mathrm{Re}(w)\in U}|f(w)|\big]$. Let $w = t - \mathrm{i}\eta$, for $t \in U$ and $\eta \in \mathbb{R}$. Let $Q \coloneqq I + w\Sigma$. (Note that $Q$ is a matrix with complex-valued entries.) By the Woodbury matrix identity,

$$f(w) = \beta^\mathsf{T}Q^{-1}\beta - \beta^\mathsf{T}\big(\lambda^{-1}\widehat{\Sigma} + Q\big)^{-1}\beta = \beta^\mathsf{T}Q^{-1}\widehat{\Sigma}^{\frac{1}{2}}\big(\lambda I + \widehat{\Sigma}^{\frac{1}{2}}Q^{-1}\widehat{\Sigma}^{\frac{1}{2}}\big)^{-1}\widehat{\Sigma}^{\frac{1}{2}}Q^{-1}\beta.$$

We first bound the norm of $\widehat{\Sigma}^{\frac{1}{2}}Q^{-1}\beta$ *uniformly* over $w$; then, we bound the operator norm of $\big(\lambda I + \widehat{\Sigma}^{\frac{1}{2}}Q^{-1}\widehat{\Sigma}^{\frac{1}{2}}\big)^{-1}$.

Let $u \coloneqq \widehat{\Sigma}^{\frac{1}{2}}Q^{-1}\beta$. In addition, define $u_0 \coloneqq \widehat{\Sigma}^{\frac{1}{2}}Q_0^{-1}\beta$, where $t_0 = \inf U$ and $Q_0 \coloneqq I + t_0\Sigma$. I claim that $\|u\|_2 \le \|u_0\|_2$, which we will show as Lemma 22, whose proof we defer:

**Lemma 22.** *If $u = \widehat{\Sigma}^{\frac{1}{2}}Q^{-1}\beta$ and $u_0 = \widehat{\Sigma}^{\frac{1}{2}}Q_0^{-1}\beta$, then $\|u\|_2 \le \|u_0\|_2$.*

Supposing Lemma 22, it thus suffices to bound $\|u_0\|_2$ to get a uniform bound over all $w$. We have, since $Q_0 \succeq \frac{1}{2}I$,

$$\mathbb{E}\left[\sup_{\mathrm{Re}(w)\in U}\|u\|_2\right] \le \mathbb{E}\big[\|u_0\|_2\big] = \mathbb{E}\Big[\beta^\mathsf{T}Q_0^{-1}\widehat{\Sigma}Q_0^{-1}\beta\Big] = \beta^\mathsf{T}Q_0^{-1}\Sigma Q_0^{-1}\beta = \beta^\mathsf{T}\Sigma^{\frac{1}{2}}Q_0^{-2}\Sigma^{\frac{1}{2}}\beta \le \|Q_0^{-1}\|_{\mathrm{op}}^2 \le 4.$$

To bound the operator norm of $\big(\lambda I + \widehat{\Sigma}^{\frac{1}{2}} Q^{-1} \widehat{\Sigma}^{\frac{1}{2}}\big)^{-1}$, note that $\lambda I + \widehat{\Sigma}^{\frac{1}{2}} Q^{-1} \widehat{\Sigma}^{\frac{1}{2}}$ can be written as $C + \mathrm{i}D$ with $C \succeq \lambda I$. Thus, by Lemma 21,

$$\Big\| \big(\lambda I + \widehat{\Sigma}^{\frac{1}{2}} Q^{-1} \widehat{\Sigma}^{\frac{1}{2}}\big)^{-1} \Big\|_{\mathrm{op}} \leq \frac{1}{\lambda}.$$

Putting everything together, we obtain

$$\mathbb{E}\left[\sup_{\mathrm{Re}(w) \in U} |f(w)|\right] = \mathbb{E}\left[\sup_{\mathrm{Re}(w) \in U} u^{\mathsf{T}} \big(\lambda I + \widehat{\Sigma}^{\frac{1}{2}} Q^{-1} \widehat{\Sigma}^{\frac{1}{2}}\big)^{-1} u\right] \leq \mathbb{E}\left[\|u_0\|_2^2 \cdot \Big\| \big(\lambda I + \widehat{\Sigma}^{\frac{1}{2}} Q^{-1} \widehat{\Sigma}^{\frac{1}{2}}\big)^{-1} \Big\|_{\mathrm{op}}\right] \leq \frac{4}{\lambda}.$$

We now move to bounding $|g(w)|$. By the Woodbury matrix identity,

$$g(w) = \beta^{\mathsf{T}} Q^{-1} \beta - \beta^{\mathsf{T}} \big(\tilde{\kappa}^{-1} \Sigma + Q\big)^{-1} \beta = \beta^{\mathsf{T}} Q^{-1} \Sigma^{\frac{1}{2}} \big(\tilde{\kappa} I + \Sigma^{\frac{1}{2}} Q^{-1} \Sigma^{\frac{1}{2}}\big)^{-1} \Sigma^{\frac{1}{2}} Q^{-1} \beta.$$

Since $Q$ and $\Sigma$ commute,

$$|g(w)| = \big|\beta^{\mathsf{T}} \Sigma^{\frac{1}{2}} Q^{-1} \big(\tilde{\kappa} I + \Sigma^{\frac{1}{2}} Q^{-1} \Sigma^{\frac{1}{2}}\big)^{-1} Q^{-1} \Sigma^{\frac{1}{2}} \beta\big| \leq \big\| \big(\tilde{\kappa} I + \Sigma^{\frac{1}{2}} Q^{-1} \Sigma^{\frac{1}{2}}\big)^{-1} \big\|_{\mathrm{op}} \cdot \|Q^{-1}\|_{\mathrm{op}}^2 \leq \frac{4}{\mathrm{Re}(\tilde{\kappa})} \leq \frac{4}{\lambda},$$

where for the penultimate inequality we applied Lemma 21 and $\|Q^{-1}\|_{\mathrm{op}} \leq w$. □

*Proof of Lemma 22.* Write $Q^{-1} = A + B\mathrm{i}$ and $Q_0^{-1} = A_0 + B_0\mathrm{i}$ for real matrices $A, A_0 \succ 0$ and $B, B_0$ symmetric, which we can do by Lemma 21. Then,

$$\|u\|_2^2 = \beta^{\mathsf{T}} \overline{Q}^{-1} \widehat{\Sigma} Q^{-1} \beta = \beta^{\mathsf{T}} (A - B\mathrm{i}) \widehat{\Sigma} (A + B\mathrm{i}) \beta = \beta^{\mathsf{T}} \big(A \widehat{\Sigma} A + B \widehat{\Sigma} B\big) \beta.$$

Let $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ denote the Frobenius inner product on $\mathbb{R}^{P \times P}$. And let $A \otimes A$ denote the operator given by $S \mapsto A \cdot \langle A, S \rangle_{\mathrm{F}}$ on $\mathbb{R}^{P \times P}$, with $B \otimes B$ denoting the same for $B$. Then, we may further rewrite

$$\|u\|_2^2 = \beta^{\mathsf{T}} \big(A \widehat{\Sigma} A + B \widehat{\Sigma} B\big) \beta = \sum_{i=1}^{N} \hat{\lambda}_i \big((\beta^{\mathsf{T}} A \hat{v}_i)^2 + (\beta^{\mathsf{T}} B \hat{v}_i)^2\big) = \sum_{i=1}^{N} \hat{\lambda}_i \Big\langle \beta \hat{v}_i^{\mathsf{T}}, \big(A \otimes A + B \otimes B\big)\big(\beta \hat{v}_i^{\mathsf{T}}\big) \Big\rangle_{\mathrm{F}}.$$

We likewise have for $u_0$ that

$$\|u_0\|_2^2 = \sum_{i=1}^{N} \hat{\lambda}_i \Big\langle \beta \hat{v}_i^{\mathsf{T}}, \big(A_0 \otimes A_0 + B_0 \otimes B_0\big)\big(\beta \hat{v}_i^{\mathsf{T}}\big) \Big\rangle_{\mathrm{F}}.$$

To show that $\|u\|_2 \leq \|u_0\|_2$, it therefore suffices to show $A \otimes A + B \otimes B \preceq A_0 \otimes A_0 + B_0 \otimes B_0$ in the Loewner order on operators $\mathbb{R}^{P \times P} \to \mathbb{R}^{P \times P}$.

We show $A \otimes A + B \otimes B \preceq A_0 \otimes A_0 + B_0 \otimes B_0$ by computing $A$ and $B$ explicitly. From Lemma 21 (and using the fact that $I + t\Sigma$ and $\eta\Sigma$ commute),

$$A = (I + t\Sigma)\big((I + t\Sigma)^2 + \eta^2 \Sigma^2\big)^{-1} \quad \text{and} \quad B = \mathrm{i}\eta\Sigma\big((I + t\Sigma)^2 + \eta^2 \Sigma^2\big)^{-1}.$$

Note that $A$, $B$, $A_0$, $B_0$ are all diagonalized in the eigenbasis of $\Sigma$. The operators $A \otimes A + B \otimes B$ and $A_0 \otimes A_0 + B_0 \otimes B_0$ can thus be seen as diagonal $(P \times P) \times (P \times P)$ matrices in this basis. The $v_i v_j^{\mathsf{T}}$ diagonal entry of $A \otimes A + B \otimes B$ is

$$\frac{(1 + t\lambda_i)(1 + t\lambda_j) + \eta^2 \lambda_i \lambda_j}{((1 + t\lambda_i)^2 + \eta^2 \lambda_i^2)((1 + t\lambda_j)^2 + \eta^2 \lambda_j^2)}.$$

We first show that this quantity is decreasing in $\eta$ for all $i, j$ when $\eta > 0$. Thus, for a given $t$, it is maximized at $\eta = 0$. We then show that this quantity, at $\eta = 0$, is decreasing in $t$ for all $i, j$. Taking $t \to t_0^+$, we conclude that

$$A \otimes A + B \otimes B \preceq A_0 \otimes A_0 + B_0 \otimes B_0.$$

We now verify the numerical claims above. We have, for $a_i = \lambda_i^{-1} + t \geq 0$ and $x = \eta^2$, that

$$\frac{(1+t\lambda_i)(1+t\lambda_j)+\eta^2\lambda_i\lambda_j}{((1+t\lambda_i)^2+\eta^2\lambda_i^2)((1+t\lambda_j)^2+\eta^2\lambda_j^2)} = \frac{1}{\lambda_i\lambda_j}\frac{a_ia_j+x}{(a_i^2+x)(a_j^2+x)}.$$

When $x$ increases by $\delta$, the numerator increases by $\delta$ and the denominator increases by $\delta^2 + \delta(a_i^2 + a_j^2 + 2x)$. Since

$$\frac{\delta}{\delta^2+\delta(a_i^2+a_j^2+2x)} \leq \frac{1}{a_i^2+a_j^2+2x} \leq \frac{a_ia_j+x}{(a_i^2+x)(a_j^2+x)},$$

the mediant inequality implies the right-hand side is decreasing in $x$. Thus, for a given $t$, $A \otimes A + B \otimes B$ is maximized (in the Loewner order) at $\eta = 0$. Supposing $\eta = 0$, the $v_iv_j^\mathsf{T}$ diagonal entry becomes $(1+t\lambda_i)^{-1}(1+t\lambda_j)^{-1}$, which is clearly decreasing in $t$. □

The next lemma calculates $\frac{\partial\widetilde{m}}{\partial t}(0)$, which appears in $g'(0)$.

**Lemma 23.** *Let $\widetilde{m}(t) = 1/\tilde{\kappa}(t, \lambda, N)$ and $\kappa = \kappa(\lambda, N)$. Then,*

$$\frac{\partial\widetilde{m}}{\partial t}(0) = \frac{\partial\kappa}{\partial\lambda} - 1.$$

*Proof.* Note that $\widetilde{m} \coloneqq \tilde{\kappa}^{-1}$ satisfies

$$1 = \lambda\widetilde{m} + \frac{1}{N}\sum_{i=1}^{P}\left(1 - \frac{1+t\lambda_i}{1+t\lambda_i+\widetilde{m}\lambda_i}\right).$$

By the implicit function theorem,

$$0 = \lambda\frac{\partial\widetilde{m}}{\partial t} + \frac{1}{N}\sum_{i=1}^{P}\frac{(1+t\lambda_i)(\lambda_i+\lambda_i\frac{\partial\widetilde{m}}{\partial t}) - \lambda_i(1+t\lambda_i+\widetilde{m}\lambda_i)}{(1+t\lambda_i+\widetilde{m}\lambda_i)^2} = \lambda\frac{\partial\widetilde{m}}{\partial t} + \frac{1}{N}\sum_{i=1}^{P}\frac{(1+t\lambda_i)\lambda_i\frac{\partial\widetilde{m}}{\partial t} - \widetilde{m}\lambda_i^2}{(1+t\lambda_i+\widetilde{m}\lambda_i)^2}.$$

Solving for $\frac{\partial\widetilde{m}}{\partial t}$ at $t = 0$, we have that

$$\begin{aligned}
\frac{\partial\widetilde{m}}{\partial t}(0) &= \left(\lambda + \frac{1}{N}\sum_{i=1}^{P}\frac{\lambda_i}{(1+\widetilde{m}\lambda_i)^2}\right)^{-1}\frac{1}{N}\sum_{i=1}^{P}\frac{\widetilde{m}\lambda_i^2}{(1+\widetilde{m}\lambda_i)^2} \\
&= \left(\frac{\lambda}{\kappa} + \frac{1}{N}\sum_{i=1}^{P}\frac{\kappa\lambda_i}{(\kappa+\lambda_i)^2}\right)^{-1}\frac{1}{N}\sum_{i=1}^{P}\frac{\lambda_i^2}{(\kappa+\lambda_i)^2} \\
&= \frac{1}{1-\frac{1}{N}\sum_{i=1}^{P}\frac{\lambda_i^2}{(\kappa+\lambda_i)^2}} - 1 \\
&= \frac{\partial\kappa}{\partial\lambda} - 1
\end{aligned}$$

where the last equality follows from Lemma 14. □

### A.4.2. Proof of Proposition 10

*Proof of Proposition 10.* Recall that $f'(0) = \mathcal{R}(\hat{\beta}_\lambda)$. And by Lemma 23,

$$g'(0) = \left(1 + \frac{\partial\widetilde{m}}{\partial t}\right)\kappa^2\beta^\mathsf{T}\big(\Sigma+\kappa I\big)^{-1}\Sigma\big(\Sigma+\kappa I\big)^{-1}\beta = \frac{\partial\kappa}{\partial\lambda}\kappa^2\beta^\mathsf{T}\big(\Sigma+\kappa I\big)^{-1}\Sigma\big(\Sigma+\kappa I\big)^{-1}\beta = \mathcal{R}^\lambda_{\text{omni}}.$$

To bound $|f'(0) - g'(0)|$, we apply Lemma 13 to $h$ and $U \coloneqq \{t : |t| < \frac{1}{2}\|\Sigma\|_{\text{op}}^{-1}\}$. Note that $h$ extends by Lemma 19 to $\{w \in \mathbb{C} : \text{Re}(w) \in U\}$. We have that

$$|f(0) - g(0)| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\tilde{\kappa}}\sqrt{\frac{\partial\tilde{\kappa}}{\partial\lambda}}$$

by Lemma 18 and $|g(w)| \lesssim 1/\lambda$ uniformly over $\{w \in \mathbb{C} : \mathrm{Re}(w) \in U\}$ by Lemma 20. Setting $M \coloneqq N^D/\lambda$, we get from Markov's inequality and Lemma 20 the high probability bound

$$\mathbb{P}\left[\sup_{\mathrm{Re}(w)\in U} |f(w)| \geq M\right] \leq N^{-D}.$$

Hence, by Lemma 13 applied to $h$,

$$\left|\mathcal{R}(\hat{\beta}_\lambda) - \mathcal{R}^\lambda_{\mathrm{omni}}\right| = |f'(0) - g'(0)| = |h'(0)| \lesssim \frac{\delta}{\|\Sigma\|_{\mathrm{op}}^{-1}} \log\left(\frac{M}{\delta}\right) \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{\|\Sigma\|_{\mathrm{op}}}{\tilde{\kappa}} \sqrt{\frac{\partial \tilde{\kappa}}{\partial \lambda}} \leq N^{-\frac{1}{2}+o(1)} \cdot \frac{\|\Sigma\|_{\mathrm{op}}}{\lambda}.$$

The last inequality above follows Lemma 14. □

## B. Reducing Noise and Misspecification to the Noiseless Case

In this appendix, we elaborate on how noisy (or misspecified) linear regression in high dimensions can be embedded into the *noiseless* model introduced in Section 2, making precise the discussion in Section 3.3. Specifically, we will show that ridge regression on any noisy (or misspecified) instance can be uniformly approximated for all $\lambda \geq 0$ by ridge regression on a noiseless *approximating instance* when $P > N$. The intuition for this approximation is that, when $P > N$, a noisy (or misspecified) problem is indistinguishable from a problem where the ground truth $\beta$ is "complex" and has large norm.

Given this approximation, our subsequent analyses hold whenever the distribution of the approximating instance satisfies Hypothesis 1. In the case of noise, we will in fact show that Hypothesis 1 holds for the approximating instance if it holds for the original covariate distribution. In particular, while the approximating instance may involve a poorly conditioned covariance matrix or a large $\|\beta\|_2$, they need not pose challenges for our random matrix hypothesis (or our subsequent analysis). (On the other hand, as discussed in Section 3, the poor conditioning of the covariance matrix and the large norm of $\beta$ can challenge typical approaches to analyzing ridge regression.)

### B.1. Model

Consider the more general model in which labels $y' \in \mathbb{R}$ are given by $y' = \beta^\mathsf{T} x + \xi$, where the covariate vector $x$ and the linear approximation error $\xi$ are drawn jointly, as $(x, \xi) \sim \mathcal{D}'$, from a distribution $\mathcal{D}'$ over $\mathbb{R}^p \times \mathbb{R}$. We assume that $\beta$ provides the best approximation to $y'$ given $x$ among linear functions $\mathbb{R}^P \to \mathbb{R}$ for $x$ drawn according to $\mathcal{D}'$. This implies the approximation error $\xi$ satisfies

$$\mathop{\mathbb{E}}_{(x,\xi)\sim\mathcal{D}'}[\xi x] = \mathop{\mathbb{E}}_{(x,\xi)\sim\mathcal{D}'}[(y' - \beta^\mathsf{T} x)x] = 0.$$

Finally, let $\sigma^2 \coloneqq \mathbb{E}_{(x,\xi)\sim\mathcal{D}'}[\xi^2]$ be the squared error of the linear approximation.

We highlight two special cases of this model. If $\mathbb{E}[\xi \,|\, x] = 0$, then $\xi$ can be thought of as observation noise on $\beta^\mathsf{T} x$. On the other hand, if $\xi$ is constant conditioned on $x$, then we have a noiseless, but misspecified, linear model. This setup can also capture combinations of these two extremes, involving both observation noise and misspecification.

Slightly abusing notation, we also use $\xi$ to denote the vector $[\,\xi_1\ \xi_2\ \cdots\ \xi_N\,]^\mathsf{T} \in \mathbb{R}^N$ of approximation errors for the dataset $X$. The "type" of $\xi$ will be clear from the context in which it is used.

### B.2. The Approximating Instance

We embed this more general instance of linear regression into our noiseless setup by introducing an extra dimension that captures the contribution of the noise and/or misspecification. Let $t > 0$ be a small constant (which we will consider in the limit $t \to 0^+$). We reparameterize $y'$ as $y' = \beta'^\mathsf{T} x'$, where

$$x' = \begin{bmatrix} x \\ t^{\frac{1}{2}}\xi \end{bmatrix} \qquad \text{and} \qquad \beta' = \begin{bmatrix} \beta \\ t^{-\frac{1}{2}} \end{bmatrix}.$$

Because $\mathbb{E}_{(x,\xi)\sim\mathcal{D}'}[\xi x] = 0$, note that $x'$ has second moment matrix

$$\Sigma' \coloneqq \mathbb{E}\left[x' x'^\mathsf{T}\right] = \begin{bmatrix} \Sigma & 0 \\ 0 & t\sigma^2 \end{bmatrix}.$$

While $\|\beta'\|_2$ does not converge as $t \to 0^+$, note that $\beta'^\intercal \Sigma' \beta' = \beta^\intercal \Sigma \beta + \sigma^2$ has no dependence on $t$.

Let $\hat{\beta}_\lambda$ be the ridge regression estimator for the original problem, and let $\hat{\beta}'_\lambda$ be the ridge regression for the modified problem with parameter $t$. We show the following:

**Proposition 24.** *For each fixed $\lambda > 0$, the ridge regression estimator $\hat{\beta}'_\lambda$ converges to $\left[\begin{smallmatrix}\hat{\beta}_\lambda \\ 0\end{smallmatrix}\right]$ as $t \to 0^+$. If $P > N$ and $\mathcal{D}'$ is non-degenerate[17], then this convergence is uniform over all $\lambda \geq 0$ almost surely.*

*Proof.* Let $\widehat{\Sigma}' \coloneqq \frac{1}{N} X'^\intercal X'$. Recall that the estimators $\hat{\beta}_\lambda$ and $\hat{\beta}'_\lambda$ can be expressed in the closed forms,

$$\hat{\beta}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1} \frac{1}{N} X^\intercal y' \qquad \text{and} \qquad \hat{\beta}'_\lambda = (\widehat{\Sigma}' + \lambda I)^{-1} \frac{1}{N} X'^\intercal y',$$

respectively. It suffices to show that

$$\hat{\beta}'_\lambda - \begin{bmatrix}\hat{\beta}_\lambda \\ 0\end{bmatrix} = \frac{t^{\frac{1}{2}}}{N + t\xi^\intercal (Q + \lambda I)^{-1} \xi} \begin{bmatrix} t^{\frac{1}{2}} \frac{1}{N} X^\intercal (Q + \lambda I)^{-1} \xi \xi^\intercal (Q + \lambda I)^{-1} \\ \xi^\intercal (Q + \lambda I)^{-1} \end{bmatrix} y', \tag{20}$$

where $Q \coloneqq \frac{1}{N} X X^\intercal$ is the normalized kernel matrix: for any fixed $\lambda > 0$, it is clear that taking $t \to 0^+$ makes the difference converge to 0. Moreover, when $P > N$, $Q$ is almost surely non-singular under the non-degeneracy assumption. Hence we may bound the right-hand side in terms of the smallest eigenvalue of $Q$, giving us uniform convergence over all $\lambda \geq 0$.

It remains to show (20). Note that

$$\widehat{\Sigma}' = \begin{bmatrix} \widehat{\Sigma} & t^{\frac{1}{2}} \frac{1}{N} X^\intercal \xi \\ t^{\frac{1}{2}} \frac{1}{N} \xi^\intercal X & t \frac{1}{N} \xi^\intercal \xi \end{bmatrix}$$

The Schur complement of the top-right block of $\widehat{\Sigma}' + \lambda I$ is

$$\lambda + \frac{t}{N} \xi^\intercal \xi - \frac{t^{\frac{1}{2}}}{N} \xi^\intercal X \cdot (\widehat{\Sigma} + \lambda I)^{-1} \cdot \frac{t^{\frac{1}{2}}}{N} X^\intercal \xi = \lambda + \frac{t}{N} \xi^\intercal \xi - \frac{t}{N} \xi^\intercal (Q + \lambda I)^{-1} Q \xi = \lambda \left(1 + \frac{t}{N} \xi^\intercal (Q + \lambda I)^{-1} \xi\right).$$

Therefore, the block matrix inversion formula gives us

$$\begin{aligned} \left(\widehat{\Sigma}' + \lambda I\right)^{-1} - \begin{bmatrix} \left(\widehat{\Sigma} + \lambda I\right)^{-1} & \\ & 0 \end{bmatrix} \\ = \frac{1}{\lambda} \frac{1}{1 + \frac{t}{N} \xi^\intercal (Q + \lambda I)^{-1} \xi} \begin{bmatrix} \frac{t}{N^2} \left(\widehat{\Sigma} + \lambda I\right)^{-1} X^\intercal \xi \xi^\intercal X \left(\widehat{\Sigma} + \lambda I\right)^{-1} & -t^{\frac{1}{2}} \frac{1}{N} \left(\widehat{\Sigma} + \lambda I\right)^{-1} X^\intercal \xi \\ -t^{\frac{1}{2}} \frac{1}{N} \xi^\intercal X \left(\widehat{\Sigma} + \lambda I\right)^{-1} & 1 \end{bmatrix} \\ = \frac{1}{\lambda} \frac{1}{1 + \frac{t}{N} \xi^\intercal (Q + \lambda I)^{-1} \xi} \begin{bmatrix} \frac{t}{N^2} X^\intercal \left(Q + \lambda I\right)^{-1} \xi \xi^\intercal \left(Q + \lambda I\right)^{-1} X & -t^{\frac{1}{2}} \frac{1}{N} X^\intercal \left(Q + \lambda I\right)^{-1} \xi \\ -t^{\frac{1}{2}} \frac{1}{N} \xi^\intercal \left(Q + \lambda I\right)^{-1} X^\intercal & 1 \end{bmatrix}. \end{aligned} \tag{21}$$

Multiplying by $\frac{1}{N} X'^\intercal y'$, we recover (20):

$$\begin{aligned} \hat{\beta}'_\lambda - \begin{bmatrix}\hat{\beta}_\lambda \\ 0\end{bmatrix} &= \frac{1}{\lambda} \frac{1}{1 + \frac{t}{N} \xi^\intercal (Q + \lambda I)^{-1} \xi} \begin{bmatrix} \frac{t}{N^2} X^\intercal \left(Q + \lambda I\right)^{-1} \xi \xi^\intercal \left(Q + \lambda I\right)^{-1} Q - \frac{t}{N^2} X^\intercal \left(Q + \lambda I\right)^{-1} \xi \xi^\intercal \\ -t^{\frac{1}{2}} \frac{1}{N} \xi^\intercal \left(Q + \lambda I\right)^{-1} Q + t^{\frac{1}{2}} \frac{1}{N} \xi^\intercal \end{bmatrix} y' \\ &= \frac{t^{\frac{1}{2}}}{N + t\xi^\intercal (Q + \lambda I)^{-1} \xi} \begin{bmatrix} t^{\frac{1}{2}} \frac{1}{N} X^\intercal \left(Q + \lambda I\right)^{-1} \xi \xi^\intercal \left(Q + \lambda I\right)^{-1} \\ \xi^\intercal \left(Q + \lambda I\right)^{-1} \end{bmatrix} y'. \end{aligned}$$

□

### B.3. The Random Matrix Hypothesis for Noisy Labels

For the theory of Section 5 to apply, the random matrix hypothesis (Hypothesis 1) should hold for the approximating instance of noiseless regression derived from the reduction. Thus, we study when the reduction preserves Hypothesis 1, given that it holds for the marginal distribution $\mathcal{D}$ of $x$. For fully general $\xi$, which may be arbitrarily correlated with $x$, we note that the

[17] It suffices that $\mathbb{P}_{(x,\xi)\sim\mathcal{D}'}[x \in U] = 0$ for any $N$-dimensional subspace $U \subseteq \mathbb{R}^P$. Some assumption is necessary here to rule out "effectively" low-dimensional distributions that lie in a $P'$-dimensional subspace of $\mathbb{R}^P$ for some $P' \leq N$.

error introduced by the reduction can be bounded in $\sigma$ (but this bound does not improve with $N$). We can say more when $\xi$ is *noise* such that $\mathbb{E}[\xi \,|\, x] = 0$ for all $x$. In this case, we show that the reduction preserves the local Marchenko-Pastur law, in the sense that the approximation error increases additively by $\lesssim N^{-\frac{1}{2}}$ (Proposition 25).

For our analysis, we bound the additional error introduced by the reduction to the two approximate equalities posited by Hypothesis 1. Specifically, we compare, as $t \to 0^+$, the errors of these approximations for the original and the approximating instances. It is not hard to see that the "averaged" law (5), given by

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{\lambda}_i + \lambda} \approx \frac{1}{\kappa},$$

is preserved exactly as $t \to 0^+$: this approximate equality relates a continuous function of $\widehat{\Sigma}'$ to a continuous function of $\Sigma'$, and we have the convergences

$$\lim_{t\to 0^+} \widehat{\Sigma}' = \begin{bmatrix} \widehat{\Sigma} & \\ & 0 \end{bmatrix} \quad \text{and} \quad \lim_{t\to 0^+} \Sigma' = \begin{bmatrix} \Sigma & \\ & 0 \end{bmatrix}.$$

Thus, we focus on the "local" law (6), given by

$$v^\mathsf{T}\lambda\big(\lambda I + \widehat{\Sigma}\big)^{-1}v \approx v^\mathsf{T}\kappa\big(\kappa I + \Sigma\big)^{-1}v.$$

The next proposition bounds the approximation error of (6) when moving from the original instance to the approximating instance in the case where $\xi$ is noise. We give our bound assuming the formal version Hypothesis 6 of Hypothesis 1 for the marginal distribution $\mathcal{D}$ of $x$.

**Proposition 25.** *Suppose $\mathbb{E}[\xi \,|\, X] = 0$ and $\frac{1}{\sigma}\big(\mathbb{E}[|\xi_i|^p \,|\, x_i]\big)^{\frac{1}{p}} \le C_p < \infty$ almost surely for all $p \in \mathbb{N}$. If $\beta^\mathsf{T}\Sigma\beta + \sigma^2 \le 1$ and $\lambda > N^{-\frac{3}{2}+o(1)}$ is such that Hypothesis 6 holds for the marginal distribution $\mathcal{D}$ of $x$ over $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$, then*

$$\lim_{t\to 0^+}\left|\beta'^\mathsf{T}\lambda\big(\widehat{\Sigma}' + \lambda I\big)^{-1}\beta' - \beta'^\mathsf{T}\kappa\big(\Sigma' + \kappa I\big)^{-1}\beta'\right| \lesssim N^{-\frac{1}{2}+o(1)} \cdot \frac{1}{\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

*Proof.* For the approximating instance, we have that

$$\lim_{t\to 0^+} \beta'^\mathsf{T}\kappa\big(\Sigma' + \kappa I\big)^{-1}\beta' - \frac{1}{t} = \beta^\mathsf{T}\kappa\big(\Sigma + \kappa I\big)^{-1}\beta,$$

and by (21), that

$$\lim_{t\to 0^+} \beta'^\mathsf{T}\lambda\big(\widehat{\Sigma}' + \lambda I\big)^{-1}\beta' - \frac{1}{t} = \beta^\mathsf{T}\lambda\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta - \frac{2}{N}\xi^\mathsf{T}X\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta.$$

The triangle inequality therefore implies that the approximation error increases by at most

$$\lim_{t\to 0^+}\left|\beta'^\mathsf{T}\lambda\big(\widehat{\Sigma}' + \lambda I\big)^{-1}\beta' - \beta'^\mathsf{T}\kappa\big(\Sigma' + \kappa I\big)^{-1}\beta\right| - \left|\beta^\mathsf{T}\lambda\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta - \beta^\mathsf{T}\kappa\big(\Sigma + \kappa I\big)^{-1}\beta\right| \lesssim \frac{1}{N}\left|\xi^\mathsf{T}X\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta\right|. \quad (22)$$

It thus suffices to bound $\frac{1}{N}\left|\xi^\mathsf{T}u\right|$, where $u \coloneqq X\big(\widehat{\Sigma} + \lambda I\big)^{-1}\beta$. This follows from a standard moment bounding argument after conditioning on $X$. Let $\|\cdot\|_p$ denote the $L_p$-norm of a random variable. Conditioning on a fixed $X$, note that the entries of $\xi$ are independent, mean 0 random variables by assumption. Thus, for any deterministic vector $v \in \mathbb{R}^N$ and any $p \in \mathbb{N}$, it follows from the Marcinkiewicz-Zygmund inequality and the triangle inequality that

$$\left\|\xi^\mathsf{T}v\right\|_p = \left\|\sum_{i=1}^{N}\xi_i v_i\right\|_p \lesssim \sqrt{p\cdot\left\|\sum_{i=1}^{N}\xi_i^2 v_i^2\right\|_{\frac{p}{2}}} \le \sqrt{p\sum_{i=1}^{N}\|\xi_i^2\|_{\frac{p}{2}}v_i^2} = \sqrt{p}\,C_p\cdot\sigma\|v\|_2,$$

where all $L_p$ norms are taken conditional on $X$. Thus, by Markov's inequality, conditional on $X$,

$$\mathbb{P}\left[\frac{1}{N}\left|\xi^\mathsf{T}v\right| \ge \frac{t}{\sqrt{N}}\right] \le \left(\frac{\left\|\xi^\mathsf{T}v\right\|_p}{t\sqrt{N}}\right)^p \le \left(\frac{\|v\|_2}{\sqrt{N}}\cdot\frac{\sqrt{p}\,C_p\cdot\sigma}{t}\right)^p. \quad (23)$$

We now set $v = u$ and bound $\frac{1}{\sqrt{N}}\|u\|_2$. By (19) in the argument of Proposition 9 and Lemma 14,

$$\begin{aligned}\frac{1}{N}\|u\|_2^2 &= \beta^\mathsf{T}\big(\widehat{\Sigma}+\lambda I\big)^{-1}\widehat{\Sigma}\big(\widehat{\Sigma}+\lambda I\big)^{-1}\beta \\ &\lesssim \frac{\partial\kappa}{\partial\lambda}\beta^\mathsf{T}\big(\Sigma+\kappa I\big)^{-1}\Sigma\big(\Sigma+\kappa I\big)^{-1}\beta + \beta^\mathsf{T}\Sigma\beta\cdot N^{-\frac{1}{2}+o(1)}\frac{1}{\lambda\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}} \\ &\le \beta^\mathsf{T}\Sigma\beta\cdot\frac{1}{\kappa^2}\frac{\partial\kappa}{\partial\lambda}\left(1+N^{-\frac{1}{2}+o(1)}\frac{\operatorname{Tr}(\Sigma)}{N\lambda}\right) \\ &\lesssim \beta^\mathsf{T}\Sigma\beta\cdot\frac{1}{\kappa^2}\frac{\partial\kappa}{\partial\lambda}.\end{aligned}$$

For any constant $\varepsilon > 0$, we may set $p \coloneqq \lceil D/\varepsilon\rceil$ and

$$t \coloneqq N^\varepsilon\cdot\sqrt{p}\,C_p\cdot\sigma\sqrt{\beta^\mathsf{T}\Sigma\beta}\cdot\frac{1}{\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}} \le N^\varepsilon\cdot\sqrt{p}C_p\cdot\frac{1}{\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

By (23), this implies that $\mathbb{P}\big[\frac{1}{N}\big|\xi^\mathsf{T}u\big| \ge \frac{t}{\sqrt{N}}\big] \lesssim N^{-D}$ over the randomness of $X$. Taking $\varepsilon \to 0^+$ slowly in $N$, we therefore obtain the high probability bound

$$\frac{1}{N}\big|\xi^\mathsf{T}u\big| \lesssim N^{-\frac{1}{2}+o(1)}\cdot\frac{1}{\kappa}\sqrt{\frac{\partial\kappa}{\partial\lambda}}.$$

Combining with Hypothesis 6 now yields the desired result. □

Finally, we note that, with the weaker assumption that $\xi_1 \lesssim \sigma$ and $y_1 \lesssim 1$, equation (22) can also be bounded as

$$\frac{1}{N}\Big|\xi^\mathsf{T}X\big(\widehat{\Sigma}+\lambda I\big)^{-1}\beta\Big| \le \frac{1}{N}\|\xi\|_2\Big\|X\big(\widehat{\Sigma}+\lambda I\big)^{-1}\beta\Big\|_2 \le \frac{1}{\lambda}\cdot\frac{\|\xi\|_2}{\sqrt{N}}\cdot\frac{\|y\|_2}{\sqrt{N}} \lesssim \frac{\sigma}{\lambda}.$$

While this bound limits the error in terms of $\sigma$ for very general misspecification, and thus is useful when $\sigma$ is small, it does not improve as $N$ increases.

### B.4. Theorem 8 and Proposition 10 for Noisy Labels

An immediate consequence of Propositions 24 and 25 is that our analysis of GCV applies to ridge regression with noisy labels, since any instance with noisy labels can be seen as a limit of noiseless approximating instances that preserve the local Marchenko-Pastur law.

As another application of our reduction, we recover without further work the formula for the generalization risk of ridge regression with noisy labels, in greater generality than previously known (Canatar et al., 2021; Hastie et al., 2020).

**Corollary 26.** *Suppose* $\mathbb{E}[\xi\,|\,X] = 0$ *and* $\frac{1}{\sigma}\big(\mathbb{E}[|\xi_i|^p\,|\,x_i]\big)^{\frac{1}{p}} \le C_p < \infty$ *almost surely for all* $p \in \mathbb{N}$. *If* $\lambda > N^{-\frac{3}{2}+o(1)}$ *is such that Hypothesis 7 holds for the marginal distribution* $\mathcal{D}$ *of* $x$ *over* $S = (\frac{1}{2}\lambda, \frac{3}{2}\lambda)$, *then*

$$\Big|\mathcal{R}^{\lambda,\sigma}_{\text{omni}} - \mathcal{R}(\hat{\beta}_\lambda)\Big| \lesssim N^{-\frac{1}{2}+o(1)}\cdot\big(\beta^\mathsf{T}\Sigma\beta+\sigma^2\big)\frac{\|\Sigma\|_{\text{op}}}{\lambda},$$

*where* $\mathcal{R}^{\lambda,\sigma}_{\text{omni}}$ *is defined to be*

$$\mathcal{R}^{\lambda,\sigma}_{\text{omni}} \coloneqq \frac{\partial\kappa}{\partial\lambda}\cdot\kappa^2\sum_{i=1}^{P}\left(\frac{\lambda_i}{(\kappa+\lambda_i)^2}\big(\beta^\mathsf{T}v_i\big)^2\right) + \frac{\partial\kappa}{\partial\lambda}\cdot\sigma^2 = \mathcal{R}^{\lambda}_{\text{omni}} + \frac{\partial\kappa}{\partial\lambda}\cdot\sigma^2.$$

*Proof.* Combining Propositions 24 and 25 with Proposition 10, it suffices to compute the limit as $t \to 0^+$ of $\mathcal{R}^{\lambda}_{\text{omni}}(t)$ for the approximating instance with parameter $t$. Indeed, we have that

$$\lim_{t\to0^+}\mathcal{R}^{\lambda}_{\text{omni}}(t) = \frac{\partial\kappa}{\partial\lambda}\kappa^2\beta^\mathsf{T}\big(\Sigma+\kappa I\big)^{-1}\Sigma\big(\Sigma+\kappa I\big)^{-1}\beta + \lim_{t\to0^+}\frac{\partial\kappa}{\partial\lambda}\cdot\kappa^2\frac{t\sigma^2}{(\kappa+t\sigma^2)^2}\Big(t^{-\frac{1}{2}}\Big)^2 = \mathcal{R}^{\lambda}_{\text{omni}} + \frac{\partial\kappa}{\partial\lambda}\cdot\sigma^2.$$ □

## C. Proofs for Section 6

In this section, we prove Propositions 4 and 5. We also formalize the notation: we write $A \asymp B$ if there exists a constant $C > 0$ (fixed throughout) such that $C^{-1}A \leq B \leq CA$.

*Proof of Proposition 4.* Let $\kappa = \kappa(0, N)$. Applying the Marchenko-Pastur law (5) at $\lambda = 0$, we have that

$$\operatorname{Tr}\big((XX^\mathsf{T})^{-1}\big) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{\lambda}_i} \approx \frac{1}{\kappa}.$$

Moreover, $\kappa$ satisfies $N = \sum_{i=1}^{P}\frac{\lambda_i}{\kappa+\lambda_i}$ by (4). Let $i^*$ be the smallest index $i$ such that $\kappa > \lambda_i$. Then, $\kappa \asymp (i^*)^{-1-\gamma}$ by the eigenvalue decay assumption. Therefore,

$$N = \sum_{i=1}^{P}\frac{\lambda_i}{\kappa+\lambda_i} \asymp i^* + \frac{1}{\kappa}\sum_{i=i^*}^{P}\lambda_i \asymp i^* + \frac{1}{\kappa}\int_{i^*}^{P} x^{-1-\gamma}\,dx \asymp i^* + \frac{1}{\kappa}(i^*)^{-\gamma} \asymp i^*.$$

It follows that $\kappa \asymp N^{-1-\gamma}$ and $N^{-1}\operatorname{Tr}\big((XX^\mathsf{T})^{-1}\big) \asymp \frac{1}{N\kappa} \asymp N^{\gamma}$. □

*Proof of Proposition 5.* Let $\kappa = \kappa(\lambda, N)$. By the fact that $y = X\beta$ and the local Marchenko-Pastur law (6), we have that

$$y^\mathsf{T}\big(XX^\mathsf{T} + N\lambda I\big)^{-1}y = \beta^\mathsf{T}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\beta \approx \beta^\mathsf{T}\Sigma(\Sigma+\kappa I)^{-1}\beta = \sum_{i=1}^{P}\frac{\lambda_i}{\lambda_i+\kappa}(\beta^\mathsf{T}v_i)^2.$$

Let $i^*$ be the smallest index $i$ such that $\kappa > \lambda_i$. Then, $\kappa \asymp (i^*)^{-1-\gamma}$ by the eigenvalue decay assumption. Therefore, we may approximate the right-hand side as

$$\sum_{i=1}^{P}\frac{\lambda_i}{\lambda_i+\kappa}(\beta^\mathsf{T}v_i)^2 \asymp \sum_{i=1}^{i^*}(\beta^\mathsf{T}v_i)^2 + \frac{1}{\kappa}\sum_{i=i^*}^{P}\lambda_i(\beta^\mathsf{T}v_i)^2 \asymp \int_1^{i^*} x^{-\delta}\,dx + \frac{1}{\kappa}\int_{i^*}^{P} x^{-1-\gamma-\delta}\,dx.$$

Using the fact that $\delta < 1$, we further approximate

$$\int_1^{i^*} x^{-\delta}\,dx + \frac{1}{\kappa}\int_{i^*}^{P} x^{-1-\gamma-\delta}\,dx \asymp (i^*)^{1-\delta} + \frac{1}{\kappa}(i^*)^{-\gamma-\delta} \asymp (i^*)^{1-\delta} \asymp \kappa^{-\frac{1-\delta}{1+\gamma}}.$$

Composing the above approximations proves the proposition. □

## D. Characterizing Classical vs. Non-classical Ridge Regression via the Train-Test Gap

Building on our theoretical analysis of Section 5 and Appendix A, we identify a precise and intuitive separation between the "classical" and "non-classical" regimes of ridge regression: we argue that the separation is characterized by the ratio between the generalization and empirical risks of the estimator $\hat{\beta}_\lambda$. We then discuss how our empirical setting belongs to the non-classical regime, whereas many previous non-asymptotic analyses of GCV (and ridge regression) (Golub et al., 1979; Hsu et al., 2014; Jacot et al., 2020b) only apply in the classical regime.

A salient feature of overparameterized machine learning environments is the possibility of a large gap between the empirical and generalization risks. Thus, this gap serves as a natural candidate for characterizing "non-classical" learning problems. For ridge regression, our developments in Section 5 and Appendix A let us precisely discuss this gap. Theorem 2 implies that the ratio between the generalization and empirical risks of $\hat{\beta}_\lambda$ can be approximated as

$$\frac{\mathcal{R}(\hat{\beta}_\lambda)}{\mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda)} \approx \frac{\text{GCV}_\lambda}{\mathcal{R}_{\text{empirical}}(\hat{\beta}_\lambda)} = \left(\sum_{i=1}^{N}\frac{\lambda}{\lambda+\hat{\lambda}_i}\right)^{-2} \approx \left(\frac{\kappa}{\lambda}\right)^2,$$

where the last approximation follows from (5) of Hypothesis 1. In particular, the ratio $\kappa/\lambda$ between the effective and the explicit regularizations determines the (multiplicative) train-test gap.

We say that a ridge regression instance is *non-classical* if $\kappa/\lambda \gg 1$, for $\kappa = \kappa(\lambda, N)$, and *classical* otherwise. (Note that Lemma 14 implies $\kappa/\lambda \geq 1$ always.) Thus, non-classical instances are characterized by having a large train-test gap. The quantity $\kappa/\lambda$ shows up in several places besides the train-test gap: it arises in the definition (13) of $\kappa$, and also in our bound for Proposition 9 relating $\mathrm{GCV}_\lambda$ and $\mathcal{R}^\lambda_{\mathrm{omni}}$[18]. Generally, it appears that problems with a larger $\kappa/\lambda$ are more challenging to understand: this ratio determines the "constant" factor as $N$ grows in our bounds; for other analyses, we will observe that that they in fact *do not apply* once $\kappa/\lambda$ exceeds a constant and thus are limited to the classical regime.

*Remark.* Note that while $P \geq N$ is necessary for a problem to lie in the non-classical regime, it is not sufficient. Even in high dimensions, if we take $\lambda$ to be sufficiently large, we will find ourselves back in the classical regime. However, this can be far from optimal in terms of generalization (see, e.g., Figure 1).

The quantity $\kappa/\lambda$ connects to our empirical setting via the train-test gap. As can be seen from the empirical and generalization risk curves for eNTK regression on pretrained ResNet-34 representations of CIFAR-100 in Figure 1, the optimal regularization is such that ratio between generalization and empirical risk is much larger than 1 (meaning that $\kappa/\lambda$ is large as well), with this trend holding consistently across models and datasets. Thus, for a theoretical analysis to be applicable to our empirical setting, it should work when $\kappa/\lambda$ is large.

We next discuss how this feature of large $\kappa/\lambda$ can be challenging for more "classical" analyses of GCV and ridge regression:

**Fixed design.** The first analyses of GCV (and ridge regression) (Craven & Wahba, 1978; Golub et al., 1979) were for the setting of *fixed design*, where the estimator $\hat{\beta}_\lambda$ is both trained and evaluated on the same dataset $x_1, \ldots, x_N \in \mathbb{R}^P$, but with noisy labels $y_i = \beta^\mathsf{T} x_i + \varepsilon_i$ resampled between train and evaluation time. Without noise, the generalization risk would simply be the empirical risk. Thus, when specialized to the noiseless case, such arguments for the consistency of the GCV estimator would imply the empirical risk approximates the generalization risk, which we know to be false.

To concretely see which assumption fails in such an analysis, we note that Golub et al. (1979) require in their proof of the consistency of GCV that $\frac{1}{N}\operatorname{Tr}\big(\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\big) \to 0$. However, we also have that

$$\frac{1}{N}\operatorname{Tr}\big(\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\big) = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{\lambda}_i}{\hat{\lambda}_i+\lambda} = 1 - \lambda\cdot\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{\lambda}_i+\lambda} \approx 1 - \frac{\lambda}{\kappa},$$

where the last approximation follows from (5) of Hypothesis 1. Thus, their assumption also implies $\kappa/\lambda \to 1$.

**Convergence of $\widehat{\Sigma} \to \Sigma$.** One approach to bounding generalization in the setting of *random design* (i.e., as described in Section 2.1) is to show $\widehat{\Sigma} \approx \Sigma$ in an appropriate sense (Hsu et al., 2014; Steinhardt, 2021; Bach, 2023). Being able to do so, however, often implies that $\mathcal{R}_{\mathrm{empirical}}(\hat{\beta}_\lambda) \approx \mathcal{R}(\hat{\beta}_\lambda)$, since the formulas for empirical and generalization risk can be obtained from each other by swapping a $\widehat{\Sigma}$ for a $\Sigma$.

Concretely, the analyses of Hsu et al. (2014) and Steinhardt (2021) assume $N \geq 2\sum_{i=1}^{P}\frac{\lambda_i}{\lambda+\lambda_i}$. Now, since $\kappa \geq \lambda$, we have by (13) that these analyses apply only when

$$\frac{\kappa}{\lambda} = \left(1 - \frac{1}{N}\sum_{i=1}^{P}\frac{\lambda_i}{\kappa+\lambda_i}\right)^{-1} \leq \left(1 - \frac{1}{N}\sum_{i=1}^{P}\frac{\lambda_i}{\lambda+\lambda_i}\right)^{-1} \leq 2.$$

Similarly, Bach (2023) assumes $N\lambda \geq 2\operatorname{Tr}(\Sigma)$, in which case $\frac{\kappa}{\lambda} \leq 1 + \frac{\operatorname{Tr}(\Sigma)}{N\lambda} < 2$ by Lemma 14.

**Classical random matrix theory.** Finally, we note that more classical random matrix theory techniques, e.g., those used by Jacot et al. (2020b), which were originally developed for asymptotic analyses in the fixed dimensional ratio limit (Marchenko & Pastur, 1967), can also struggle in the $\kappa/\lambda \gg 1$ regime. For instance, the bounds of Jacot et al. (2020b) are only non-vacuous when $\frac{\operatorname{Tr}(\Sigma)}{N\lambda} \leq 1$, in which case $\frac{\kappa}{\lambda} \leq 2$ by Lemma 14. (In contrast, Hypothesis 1 is motivated by recent developments in random matrix theory (Erdős & Yau, 2017; Knowles & Yin, 2017) that provide fine-grained control over the resolvent via fluctuation averaging arguments.)

[18] Note that the multiplier on $N^{-\frac{1}{2}+o(1)}$ in the error bound can also be bounded by $(\kappa/\lambda)^{3/2}$.

# E. Details of the Experimental Setup

## E.1. Computing eNTKs

To compute eNTKs, we pursue a very simple high-level strategy: we compute the $N_0 \times P$ Jacobian matrix, where $N_0$ is the dataset size and $P$ is the number of model parameters and multiply it with its transpose to obtain the kernel $K \in \mathbb{R}^{N_0 \times N_0}$ described in Section 2.3. Naively, this approach is infeasible for large datasets and models. (E.g., for a ResNet-101 (44M parameters) over the Food-101 dataset (75750 images), storing the Jacobian matrix would require over 10TB.) Our approach thus computes this Jacobian in chunks that fit into RAM, performing compute intensive operations on the GPU.

Empirically, the bottleneck for computation time comes from multiplying the Jacobian with itself, which has complexity $O(N_0^2 P)$. With GPU acceleration and optimized data transfer, this approach is nonetheless relatively efficient: on a machine with four A100 GPUs and 755GB RAM, we can compute the $60000 \times 50000$ eNTK of a ResNet-18 over CIFAR-10 at `float32` precision in 43 minutes, at a rate of less than $10^{-6}$ seconds per NTK entry. This performance compares favorably to existing approaches for computing eNTKs (Novak et al., 2020; 2022), despite being algorithmically simple: for instance, the recent work of Novak et al. (2022) achieves a rate $\sim 3 \cdot 10^{-6}$ seconds per NTK for the same task on TPU v4.

For further implementation details, refer to the code released at `https://github.com/aw31/empirical-ntks`.

## E.2. Evaluating GCV

Recall from Section 2.3 that, for each model-dataset pair, we compute a kernel $K \in \mathbb{R}^{N_0 \times N_0}$, where $N_0$ is the dataset size, from the model's eNTK representations of the dataset, and that we approximate the full eNTK by $I \otimes K \in \mathbb{R}^{(N_0 \times C) \times (N_0 \times C)}$. To solve our classification tasks, we perform kernel regression on the one-hot labels $y_i \in \mathbb{R}^C$ corresponding to each data point $x_i$, after normalizing each label to have mean 0. Using our approximation, we have the decomposition of this task into $C$ independent kernel regression problems, one for each class.

To aggregate risk, we simply sum the mean squared error over the $C$ output dimensions. Observe that the normalization is such that predicting 0 trivially obtains risk $\leq 1$. To implement GCV for $C$-dimensional output, we do the same, summing independent estimates of generalization risk for each of the $C$ output dimensions.

For consistent comparisons across dataset sizes, we evaluate for each dataset size $N$ the $\lambda$ values $\{\lambda_0/N : \lambda_0 \in \Lambda_0\}$ for each $N$, where $\Lambda_0 \subseteq \mathbb{R}_{\geq 0}$ is a set of base values chosen in proportion to $\|\widehat{\Sigma}\|_{\text{op}}$. The range of $\Lambda_0$ is chosen to be the smallest one so that the generalization risk approximately converges at both extremes across all dataset sizes.

To solve the kernel regression problems for many regularization levels $\lambda$, we first diagonalize the kernel matrix. Doing so also allows for efficient computation of $\text{GCV}_\lambda$ over multiple values of $\lambda$. The largest kernel matrices that we work with are obtained from the Food-101 dataset and have size $75750 \times 75750$. We note that, while these matrices are substantial in size, they are much smaller than the eNTK representations before applying the kernel trick: a ResNet-101 has 44 million parameters, and thus, the matrix of eNTK representations would be of size approximately $75750 \times 44 \cdot 10^6$.

## E.3. Comparing GCV to Alternate Approaches

To estimate $\alpha$ and $\sigma$ for $\widehat{\mathcal{R}}_{\text{spec}}$, we first note that the risk estimate is linear in $\alpha^2$ and $\sigma^2$. Thus, we fit $\alpha^2$ and $\sigma^2$ to minimize the mean squared error of the estimates over the set of $(N, \lambda)$ pairs considered. We use these estimated $\alpha$ and $\sigma$ for all downstream evaluations.

For the correlation benchmark, we simply compute the Pearson correlation coefficient between each set of predictions over all pairs $(N, \lambda)$ and corresponding values observed for generalization risk. Observe that correlation is (up to sign) invariant under affine transformations of the predictions.

For the scaling law benchmark, we first find for each $N$ the $\lambda_N^*$ that minimizes the generalization risk of ridge regression. Given a predictor, let $\widehat{\mathcal{R}}_N^*$ be the risk prediction corresponding to $N$ and $\lambda^*$. To estimate the rate $\hat{\alpha}$ of optimal scaling from each predictor, we fit the slope of the pairs $(N, \widehat{\mathcal{R}}^*)$ on a log-log plot. To estimate the true scaling rate, we apply the same procedure to the observed generalization risks $\mathcal{R}(\hat{\beta}_{\lambda_N^*})$.

## F. Deriving the Spectrum-only Estimate

The result of Dobriban & Wager (2018) can be recovered from Corollary 26 by assuming an isotropic prior $\mathcal{N}(0, \alpha^2 I)$ over $\beta$. Indeed, we have that

$$\mathop{\mathbb{E}}_{\beta\sim\mathcal{N}(0,\alpha^2 I)}\left[\mathcal{R}^{\lambda,\sigma}_{\text{omni}}\right] = \alpha^2 \cdot \frac{\partial\kappa}{\partial\lambda}\kappa^2 \sum_{i=1}^{P} \frac{\lambda_i}{(\kappa+\lambda_i)^2} + \sigma^2 \cdot \frac{\partial\kappa}{\partial\lambda}.$$

To obtain an estimate for the first term, by Theorem 8, we can use the GCV estimate for the noiseless case:

$$\begin{aligned}
\alpha^2 \cdot \frac{\partial\kappa}{\partial\lambda}\kappa^2 \sum_{i=1}^{P} \frac{\lambda_i}{(\kappa+\lambda_i)^2} &= \alpha^2 \cdot \kappa^2 \frac{\partial}{\partial\lambda}\Big(-\operatorname{Tr}\Big(\Sigma\big(\Sigma+\kappa I\big)^{-1}\Big)\Big) \\
&\approx \alpha^2 \cdot \hat{\kappa}^2 \frac{\partial}{\partial\lambda}\Big(-\operatorname{Tr}\Big(\widehat{\Sigma}\big(\widehat{\Sigma}+\lambda I\big)^{-1}\Big)\Big) \\
&= \alpha^2 \cdot \hat{\kappa}^2 \sum_{i=1}^{N} \frac{\hat{\lambda}_i}{(\lambda+\hat{\lambda}_i)^2}.
\end{aligned}$$

And for the second term, using the fact that $\kappa \approx \hat{\kappa}$, we have

$$\sigma^2 \cdot \frac{\partial\kappa}{\partial\lambda} \approx \sigma^2 \cdot \frac{\partial\hat{\kappa}}{\partial\lambda} = \frac{\sigma^2}{N} \cdot \hat{\kappa}^2 \sum_{i=1}^{N} \frac{1}{(\lambda+\hat{\lambda}_i)^2}.$$

This recovers the expressions used in Section 4.2.

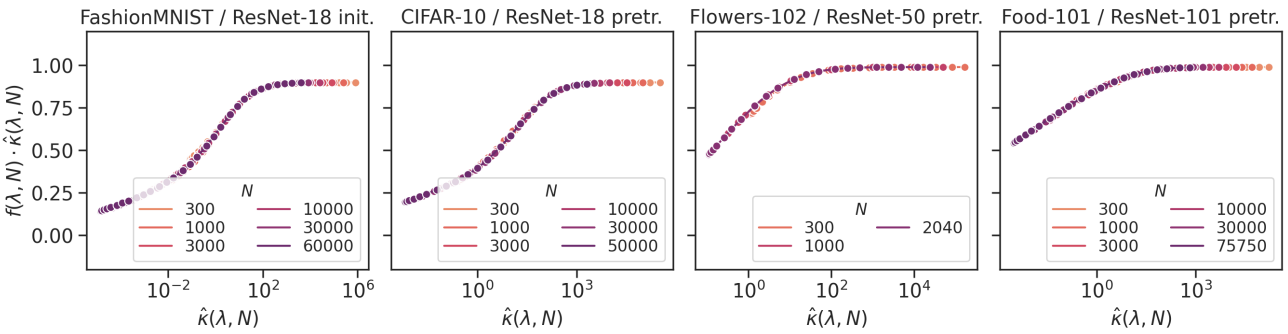


*Figure 6.* Plotting $(\hat{\kappa}(\lambda, N), f(\lambda, N) \cdot \hat{\kappa}(\lambda, N))$ for varying values of $\lambda$ and $N$

## G. Empirical Evidence for the Local Marchenko-Pastur Law

In this section, we present evidence for the validity of Hypothesis 1 in our empirical setting. Our findings here give further support to random matrix effects being a central driver of the phenomena surrounding overparameterized generalization.

While it is impossible to directly verify Hypothesis 1 due to the high dimensionality of our empirical setting, we can still check whether direct consequences of this hypothesis hold. In particular, consider

$$f(\lambda, N) \coloneqq y^{\mathsf{T}}\left(XX^{\mathsf{T}} + N\lambda I\right)^{-1} y = \beta^{\mathsf{T}}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}\beta \approx \beta^{\mathsf{T}}\Sigma(\Sigma + \kappa I)^{-1}\beta,$$

where the approximate equality holds by (6). Thus, if Hypothesis 1 holds, then $f(\lambda, N)$ should be determined by $\kappa(\lambda, N)$. By (5) of Hypothesis 1, we may also estimate $\kappa(\lambda, N)$ as

$$\kappa(\lambda, N) \approx \left(\sum_{i=1}^{N} \frac{1}{\lambda + \hat{\lambda}_i}\right)^{-1} =: \hat{\kappa}(\lambda, N).$$

To check the consistency of Hypothesis 1, we can therefore examine whether the curves traced out by $(\hat{\kappa}(\lambda, N), f(\lambda, N))$ for varying $\lambda$ coincide across values of $N$. We plot a version of this in Figure 6, where we multiply $f$ by $\hat{\kappa}$ for normalization.

Examining Figure 6, we find that the curves traced out for different values of $N$ almost coincide, as predicted by Hypothesis 1, with this holding across a range of models and datasets. Thus, we find support for the local Marchenko-Pastur law being valid in our empirical setting.

# H. Additional Experiments and Figures

## H.1. Growth of $\|\hat{\beta}_0\|_2/\sqrt{N}$ in $N$

In this section, we provide additional examples of when the norm-based estimate $\|\hat{\beta}_0\|_2/\sqrt{N}$ increases as $N$ increases and the generalization risk decreases in Figure 7, showing that this observation is consistent across models and datasets.

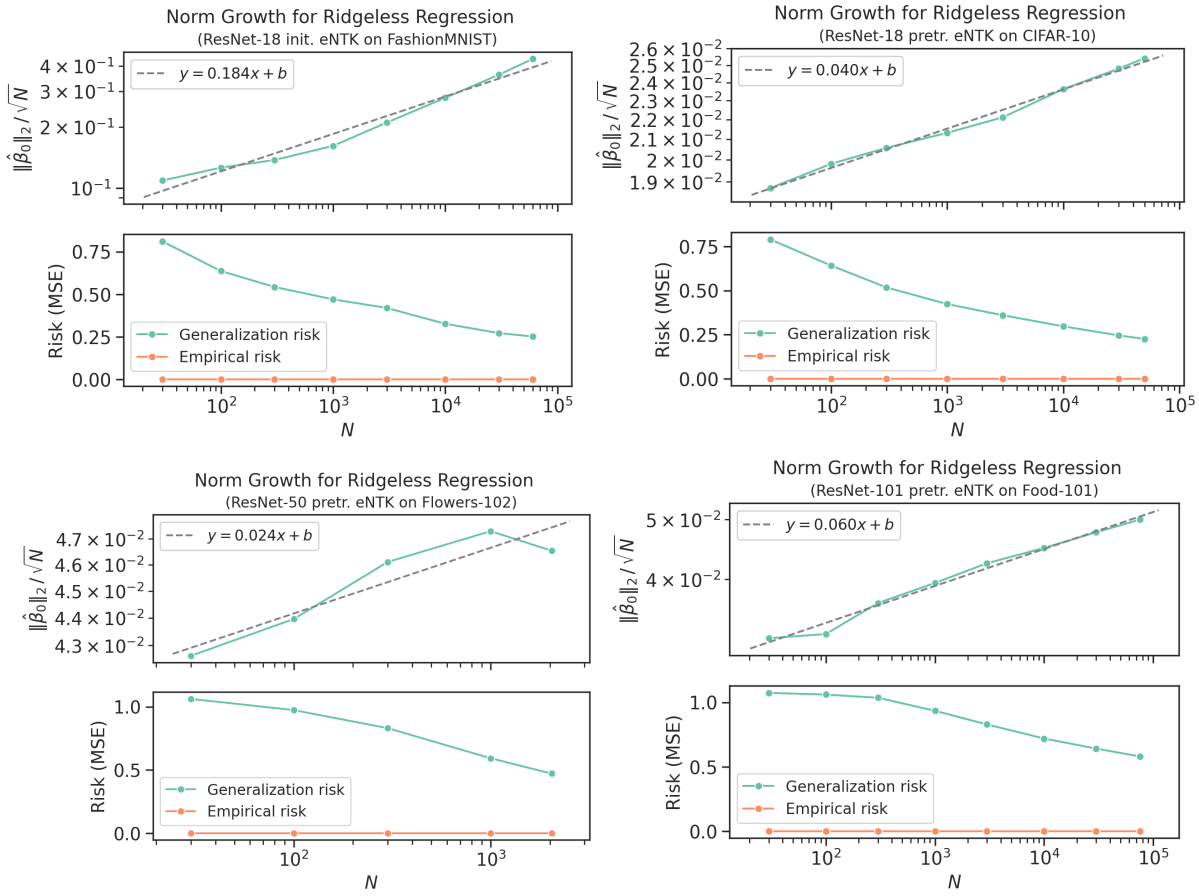


*Figure 7.* Additional plots showing the growth of the norm $\|\hat{\beta}_0\|_2/\sqrt{N}$ for ridge regression on the eNTKs additional models and datasets.

### H.2. Spectrum Comparisons

In this section, we provide additional examples of the slow convergence of the spectrum (Figure 8) and of pretrained models having higher effective dimension (Figure 9), showing that these trends also hold over a variety of datasets and models.

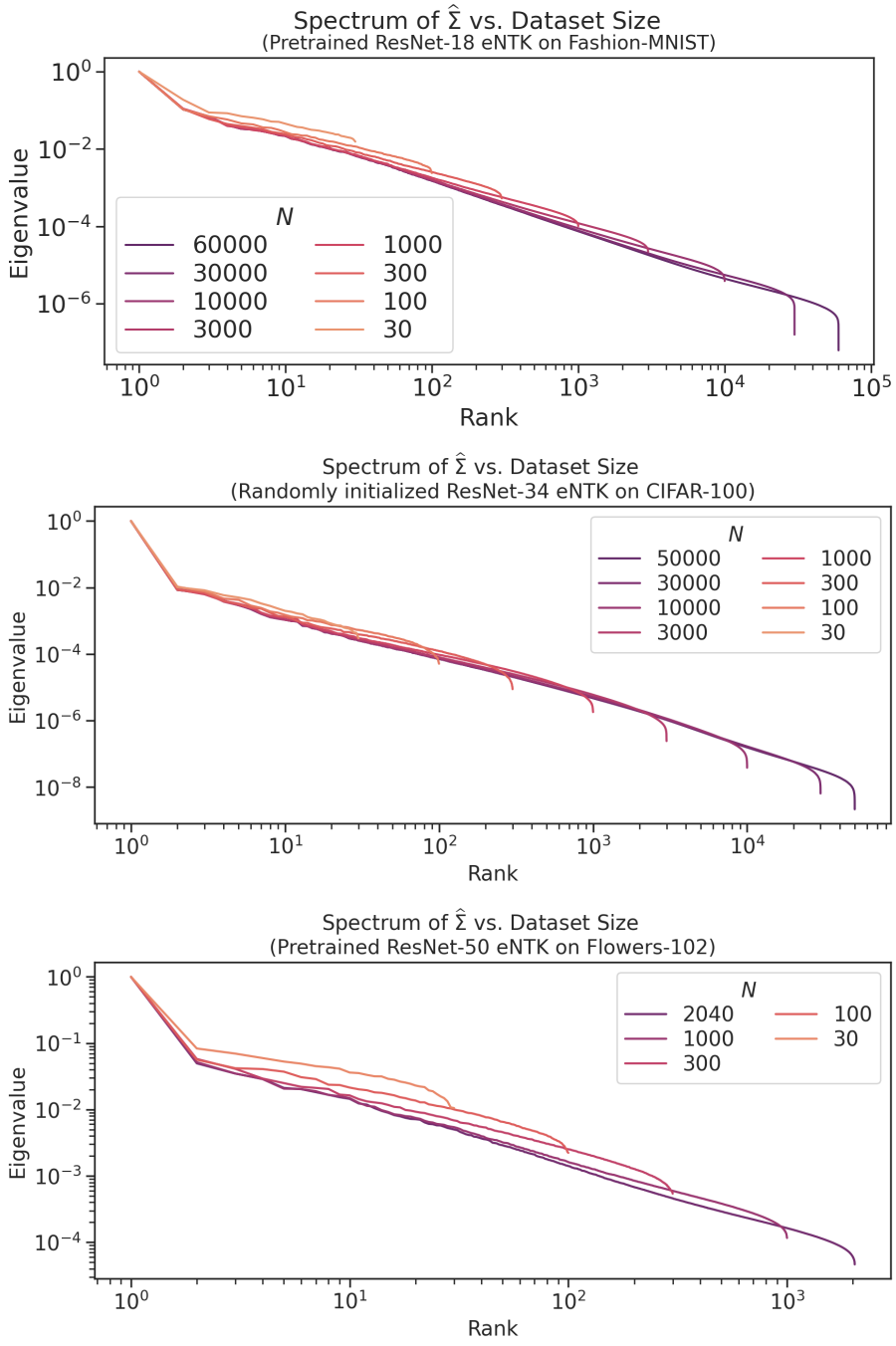


*Figure 8.* Additional plots showing the slow convergence of the empirical eigenvalue spectrum to the population eigenvalue spectrum.

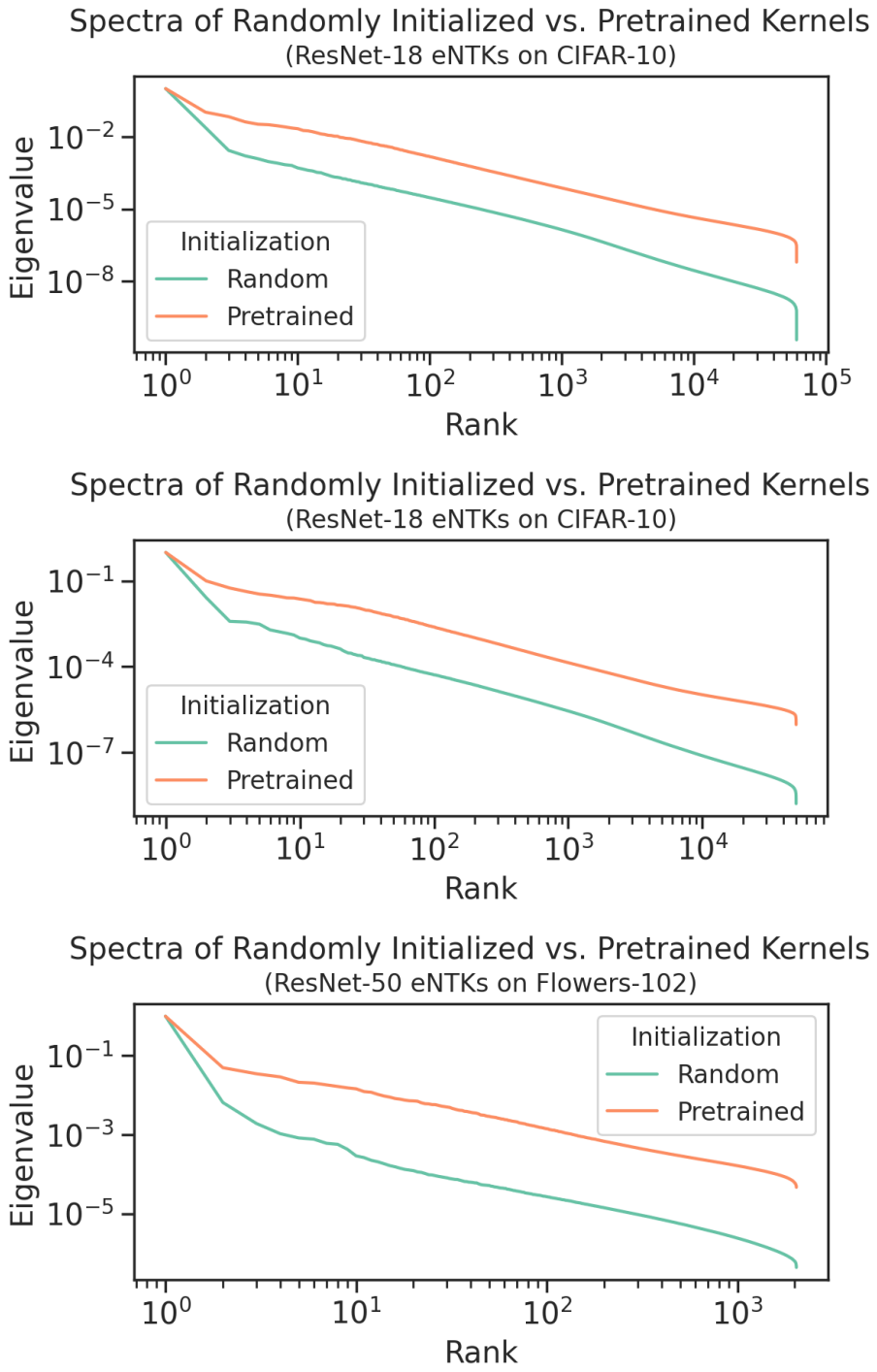


*Figure 9.* Additional plots showing that pretrained representations have slower eigendecay and thus higher effective dimension.

### H.3. Regression on Last Layer Activations

In this section, we consider predicting the generalization risk of ridge regression on the last layer activations of pretrained models. Figure 10 plots the results of these experiments. These plots show that, in this lower-dimensional setting that spans the under- and overparameterized regimes, the GCV estimator continues to perform well.

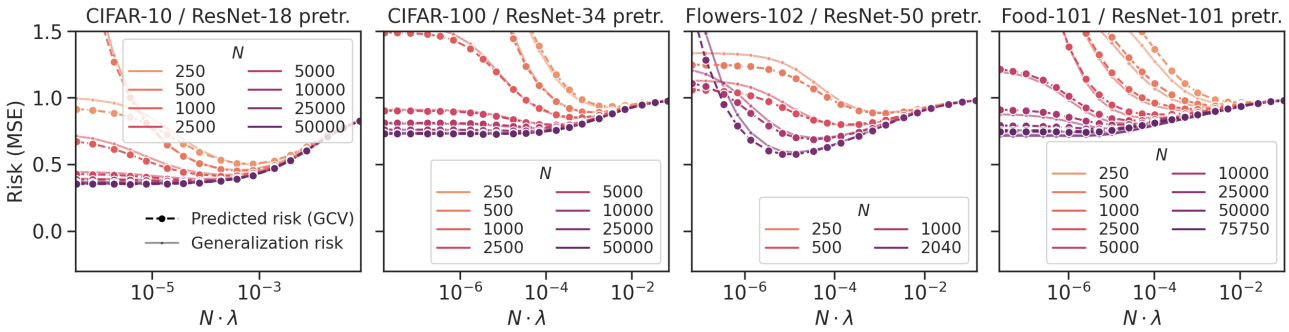


*Figure 10.* Generalization risk vs. the GCV prediction for regression on the *last-layer* activations, for various datasets and networks, across sample sizes $N$ and regularization levels $\lambda$

### H.4. Plots for the Norm- and Spectrum-Based Predictors

To provide further intuition about the predictors $\widehat{\mathcal{R}}_{\mathrm{norm}}$ and $\widehat{\mathcal{R}}_{\mathrm{spec}}$, we provide plots of the predictions that they make for our empirical setting in Figures 11 and 12.

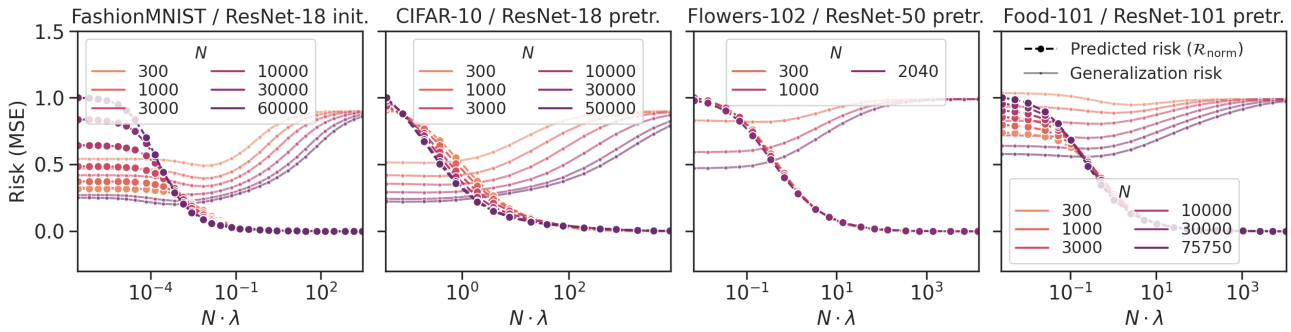


*Figure 11.* Plots of the norm-based predictor $\|\hat{\beta}_\lambda\|_2/\sqrt{N}$ against the generalization risk for various datasets and architectures. We normalize the predictions so that the maximum prediction in any graph is 1. Note that the prediction tends to be negatively correlated with the actual test risk when $N \cdot \lambda$ is small.

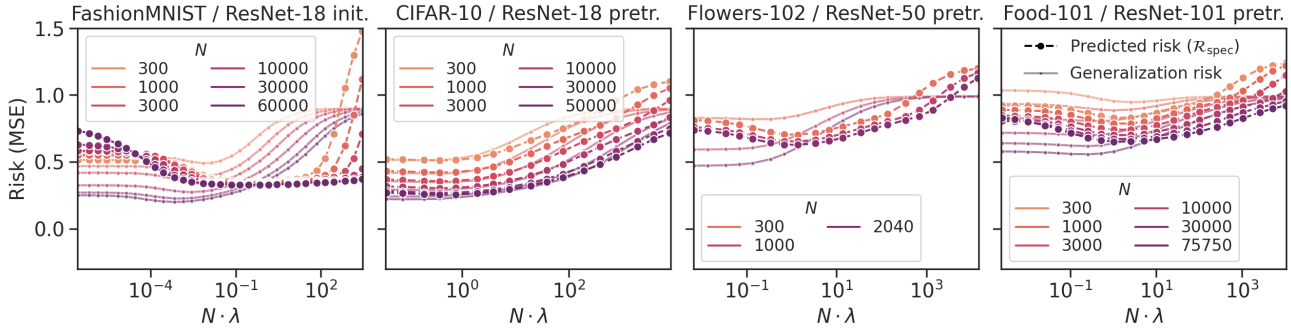


*Figure 12.* Plots of the $\widehat{\mathcal{R}}_{\mathrm{spec}}^{\alpha,\sigma}$ for $\alpha$, $\sigma$ fitted as per Appendix E against the generalization risk for various datasets and architectures. Note that this approach has trouble in particular fitting the randomly initialized setting.

### H.5. Comparing GCV to the Naive Risk Estimate

To show the necessity of the GCV correction, we consider attempting to estimate $\mathcal{R}_{\text{omni}}$ by plugging in $\widehat{\Sigma}$ as an estimate for $\Sigma$, following Loureiro et al. (2021). We plot the result of doing so in Figure 13 for a pretrained ResNet-34 applied over CIFAR-100, where $\widehat{\Sigma}$ is estimated using the full training set of 50000 images. As can be seen from the plot, the estimates of generalization risk obtained from the naive method diverge badly in the regime of small $N \cdot \lambda$ even for moderate values of $N$; thus, the GCV correction is needed to accurately reliably estimate generalization risk.

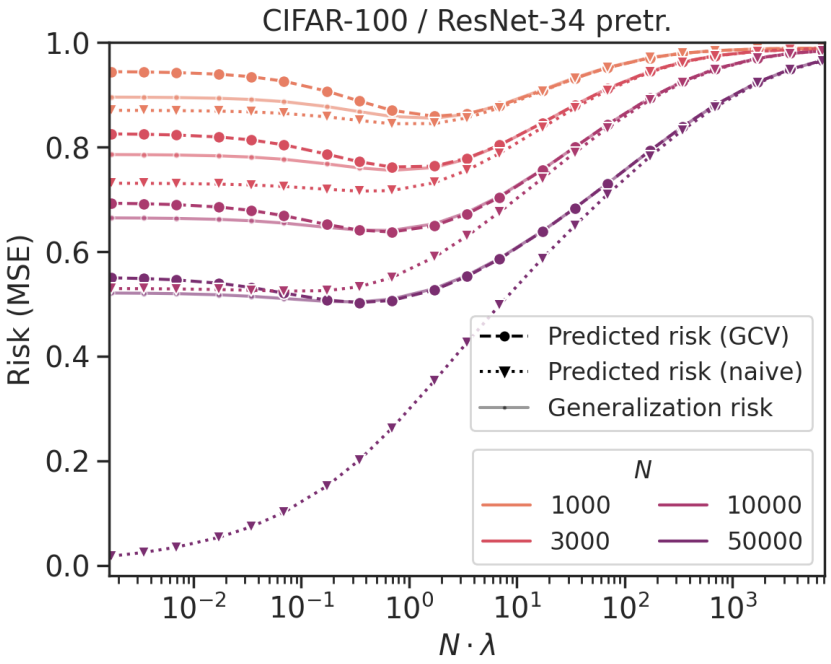


*Figure 13.* Comparing GCV to the naive risk estimate that does not correct for $\Sigma$ vs. $\widehat{\Sigma}$, for $\widehat{\Sigma}$ estimated on 50000 samples.

### H.6. Verifying the Power Law Ansatz

In this section, we verify that a power law—as studied in Sections 4.2 and 6—meaningfully approximates the scaling of generalization risk in our empirical setting. Figure 14 plots the generalization risk of optimally tuned ridge regression for each dataset-model pair from Table 2 against varying values of $N$ on a log-log scale. We find that, for each dataset-model pair, the generalization risk curve becomes roughly linear once $N \gg C$ (for $C$ the number of classes). That is, generalization risk can indeed be approximated as a power law in $N$.

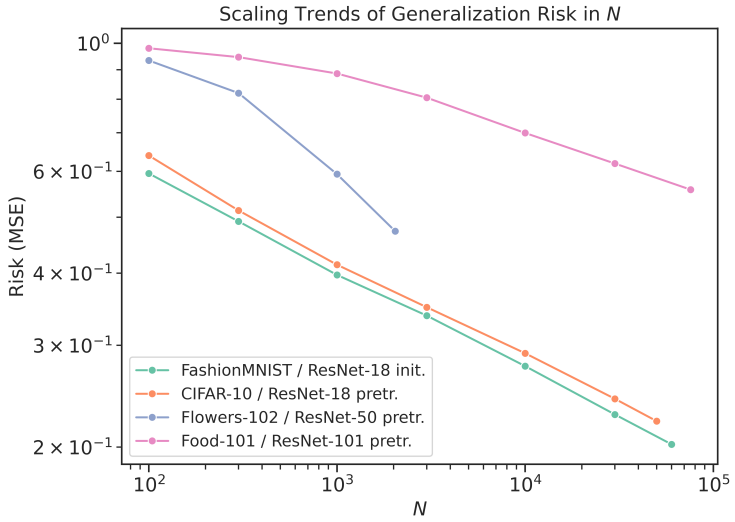


*Figure 14.* Plot of the generalization risk of optimally tuned ridge regression against $N$ for each dataset-model pair in Table 2.