High Dimensional Linear Interpolation for Structured Data

a thesis presented
by
Kevin Luo
to
The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Bachelor of Arts (Honors)
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts
April 2024

Thesis advisors: Pragya Sur, Yufan Li   Kevin Luo

High Dimensional Linear Interpolation for Structured Data

Abstract

The success of overparameterization in modern machine learning has caused a paradigm shift in statistics. In particular, the phenomenon of double descent, wherein model performance improves with model size far past the interpolation threshold, has upended the classical understanding of the bias-variance tradeoff. Seeking tractable models in which to study this, statisticians have returned to the canonical problem of linear regression, though now under a high dimensional lens. There, examining the effects of overparameterization in these simple models, they recover some hallmarks of double descent. Throughout this literature, the assumption that the rows of the design are independent and identically distributed is ubiquitous – little is known about what may occur in settings of heavy dependence, which arise in, for example, neuroscience and finance. Here, we instead model the design as right-rotationally invariant, a distribution permitting significant row dependence that has received recent attention. Under this model, we derive the out-of-sample risks for minimum norm linear interpolation and ridge regression; furthermore, we prove that Generalized Cross Validation is no longer consistent, and offer a consistent alternative. Lastly, we present partial findings on the random features model with Gaussian inputs and right-rotationally invariant weights, demonstrating numerical support for conjectures underlying our results.

Contents

Listing of figures

To Chunxu Qu and Shengcheng Luo.

Acknowledgments

It is not false to claim that I chose to write a thesis because I wanted to write the acknowledgements.

In my four years at this institution, I've been lucky to meet many good friends and mentors. Foremost thanks to my advisor, Professor Pragya Sur, without whom this project would have never been possible. Though we did not prove the result we wanted, this has been a wonderful first real foray into research. Under you, I feel like I have learned a huge amount about the landscape of statistical literature, much more than is expressed in this thesis. Thank you for the encouragement, the fun stories, and for keeping a backup plan in mind in case it was needed. I think often about what you told me about your work in the mountains during your PhD – hopefully I can have my own moment like that. Thanks also to Yufan, for his words of encouragement and his time and patience, for answering my all my random questions, for the resources on random matrix theory, and for helping me understand early on the dedication necessary to be a researcher.

A huge thanks to Professor Subhabrata Sen: I entered this university 4 years ago not knowing what I wanted to do; I spent my first two years bouncing around statistics, math, computer science, and even physics for a little while, but never found my calling until I took Stat 217. I grew up believing that when I found "that thing" that I was destined for in life, it would happen in a flash, and I would suddenly just *know* from the start. In reality, I think things happened more subtly, and somewhere in the middle I looked back and was suddenly amazed at even the little bit I had seen, astounded by the beautifully sharp structure that emerges in these chaotically random problems. I feel so lucky to have taken the first iteration of the course, and I am sure it will be a source of inspiration for generations of students to come. Thanks also to Professor Mark Sellke for teaching Stat 291, which has served as a magical sequel, and to Professors Lucas Janson and Morgane Austern for giving me glimpses of the bountiful depths of other areas of this field.

Next, thanks to the faceless Leverett administrator who read a certain floater's request to live near one of the two people they knew (COVID wasn't exactly great for meeting people, especially if you're already bad at it) and decided it was a good idea to put him with a group of three that didn't actually really know him. Wittmann, Yiting, and Mincheol – you three are the luckiest thing that has ever happened to me. I cannot imagine having gone through these last three years with anyone else. Thanks too to Sunera, Diran, and Anming for the emotional support and the late night conversations; to Alan Chung, my academic older brother, for his guidance in academia, cooking, and just life, for sharing his perspectives as a graduate student – cheers; to Oliver, for hearing me out about anything, whether it be research or life, and for being willing to eat Joe's at degenerate hours; to Nithin and Skyler for the banter and the Lizzy's runs; to the friendly faces in the Lev dining hall and the other Lev gamblers that never gamble: Kevin, Hanyue, Audrey, Ricky, Joel, John, Ryan, Zev, Molly, Austin – I will forever miss our dinner conversations.

Lastly, to my parents: while I've spent the last four years surrounded by the world's foremost academics, it has become clear to me that the greatest teachers in my life have always been you. Thank you for the limitless patience, for the unwavering faith, for the constant concern, and for dealing with the little kid that often refused (and refuses still) to listen. All of me is due to you.

0 Introduction

The most well studied statistical problem is that of regression – given a training dataset $\{(x_i, y_i)\}_{i=1}^n$ consisting of covariates x_i together with their labels y_i, how can one best predict the label of a new datapoint x_{new}? The standard approach to this problem is that of learning a model. A practitioner first proposes a *model class*, a set of functions $\mathcal{F}$ which serve as candidates for modeling the true labelling process. They then choose a function $\hat{f} \in \mathcal{F}$ from this model class using some prescribed method, often by taking the model minimizing a given error metric on training set (called empirical risk minimization), with the hope that this model will then accurately predict y_{new}, the label of x_{new}.

One of the core tenets of classical statistics which governs this procedure is that of the bias-variance tradeoff. The excess generalization error of the learned model, that is, it's accuracy in predicting y_{new}, is composed of two pieces: its bias, or its inability to reach the ground truth labelling process, and its variance, or how sensitive it is to the noise in the training data. The canonical example of this is the following decomposition of the

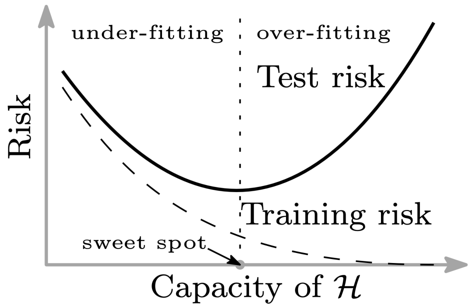


Figure 1: The bias-variance tradeoff in the classical regime. [Bel+19, Figure 1(a)]

conditional squared error:

$$\mathbb{E}[(y_{\mathsf{new}} - \hat{f}(x_{\mathsf{new}}))^2 \mid x_{\mathsf{new}}] = (\underbrace{\mathbb{E}[y_{\mathsf{new}} - \hat{f}(x_{\mathsf{new}}) \mid x_{\mathsf{new}}]}_{\text{bias}})^2 + \underbrace{\mathsf{Var}(\hat{f}_{\mathsf{new}}(x_{\mathsf{new}}) \mid x_{\mathsf{new}})}_{\text{variance}} + \underbrace{\mathsf{Var}(y_{\mathsf{new}} \mid x_{\mathsf{new}})}_{\text{irreducible noise}}. \quad (1)$$

Classical intuition states that as the numbers of parameters of the model class increases, the flexibility of the learned model, and thus its ability to fit the ground truth, increases, and hence the model's bias decreases. However, at the same time, these increased degrees of freedom in turn increase the model's sensitivity to the training data, thereby causing its variance to grow. The result is a delicate balance, where a practitioner choosing a model class must balance these two competing quantities – choosing too few parameters risks failing to capture the ground signal, called underfitting; choosing too many parameters may potentially cause the model to fit to spurious signals in the training data, called overfitting. This is summarized abstractly in Figure 1, and replicated in an experimental setting in Figure 2(a); both plots show test loss first decreasing due to decreasing bias, before increasing due to variance exploding. Three particular examples are shown in figures 2(b, c, d), depicting the range of fitted lines that are produced when $3, 7$, and 10 features are used, respectively. The dashed black line shows the ground truth process; the light blue lines are one of 200 fits where the noise is resampled; the purple line shows the mean of

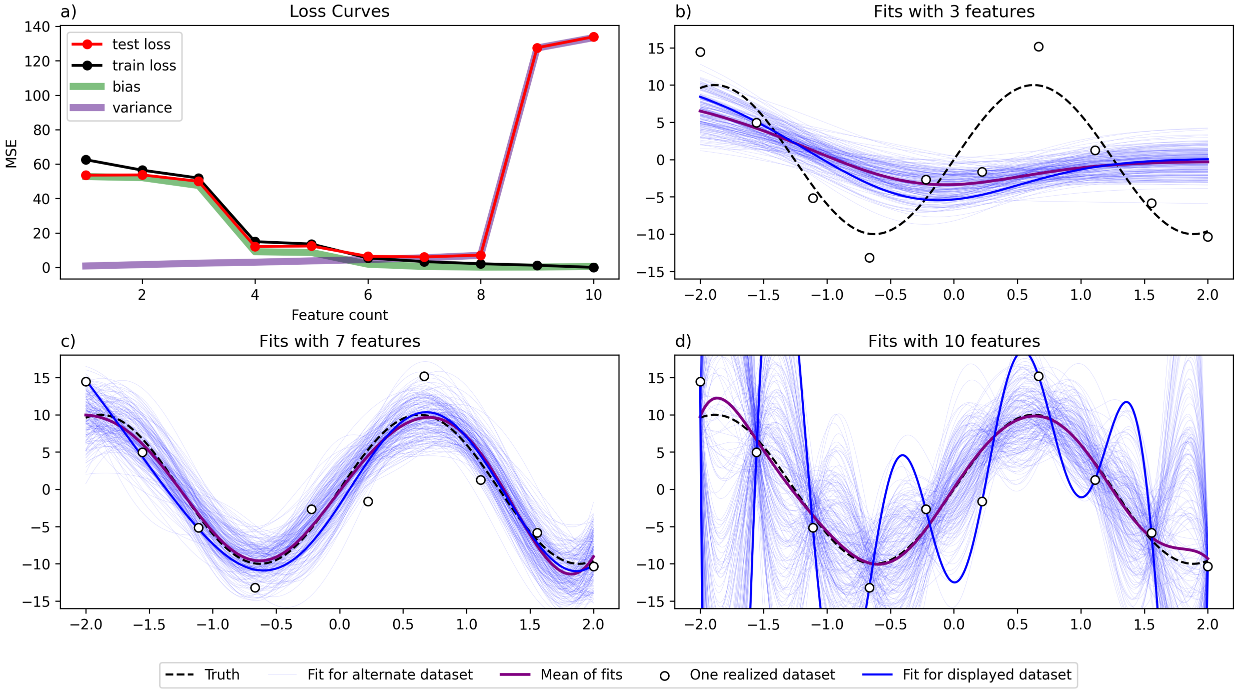


Figure 2: Experimental example of bias-variance tradeoff in the classical regime. (*a*): test and train loss. (*b*-*d*): Light blue lines are one of 200 fits, where the noise is resampled. Purple line shows the mean of the 200 fits. Black dashed line shows the true process. Solid blue line shows the fit for the displayed dataset (the white datapoints). *Setup*: Random Features regression, with features of the form sigmoid$(g_1 x + g_2)$, where $g_1, g_2 \sim$ N$(0, 1)$. See Appendix A.5.1 for full experimental details.

these 200 fits, and the solid blue line shows the fit for the displayed dataset consisting of the white points. In 2(b), with only 3 features, it is clear that the model class does not have enough flexiblity to fit the underlying ground truth – none of the fits are able to capture the trend, much less their average. At the same time, their tight clustering reveals that this model class has low variance. Figure (c) shows the range of fits when 7 features are used – this minimizes the test loss curve. We see that all of the 200 fits capture the signal well, and indeed their mean almost perfectly matches the true signal. Lastly, (d) illustrates the behavior when 10 features are used. While the mean captures the signal almost perfectly, illustrating that this model class has close to no bias, the variance in the 200 fits is extremely large – the model is forced to pass through every point in the training set, making it too sensitive to the noise within the dataset, which leads to the worst

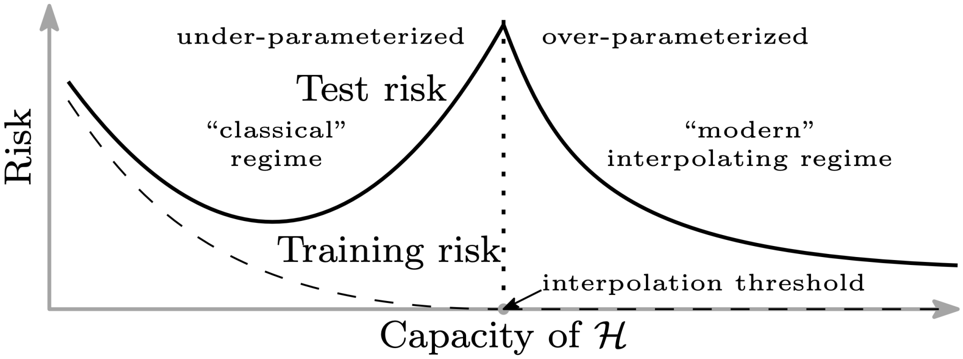


Figure 3: Double descent curve, [Bel+19, Figure 1(b)]

performance of any number of features, as seen in (a).

This problem has been extensively studied [HTF01], with techniques such as regularization developed to placate the effects of overfitting by reducing the variance of the fitted model.

Modern machine learning has illustrated the inadequacy of this picture. With the notable caveat of large language models, which are still underparameterized, state of the art performance in machine learning regression tasks is achieved by overparameterized models (those with more parameters than datapoints) that are trained to nearly zero training loss[1], with performance continuing to improve as models grow in scale (parameter count), in spite of classical wisdom suggesting that having zero training error should be indicative of overfitting. This has contributed to a modern dogma of "scale is all you need," calling into question the usefulness of the classical statistical paradigm. In fact, neural network architectures, at the same sizes at which they generalize well, are expressive enough to memorize entirely random labels, even with regularization [Zha+17]. This illustrates that the model class has the capacity to exhibit very high variance, and yet the models learned in practice do not – indeed, fitting these models to zero error on data with a significant level of noise *still* proves to yield strong predictive performance.

The behavior emerging as models are progressively overparameterized has been studied

[1] see e.g. [For+21], state of the art performance on CIFAR-100

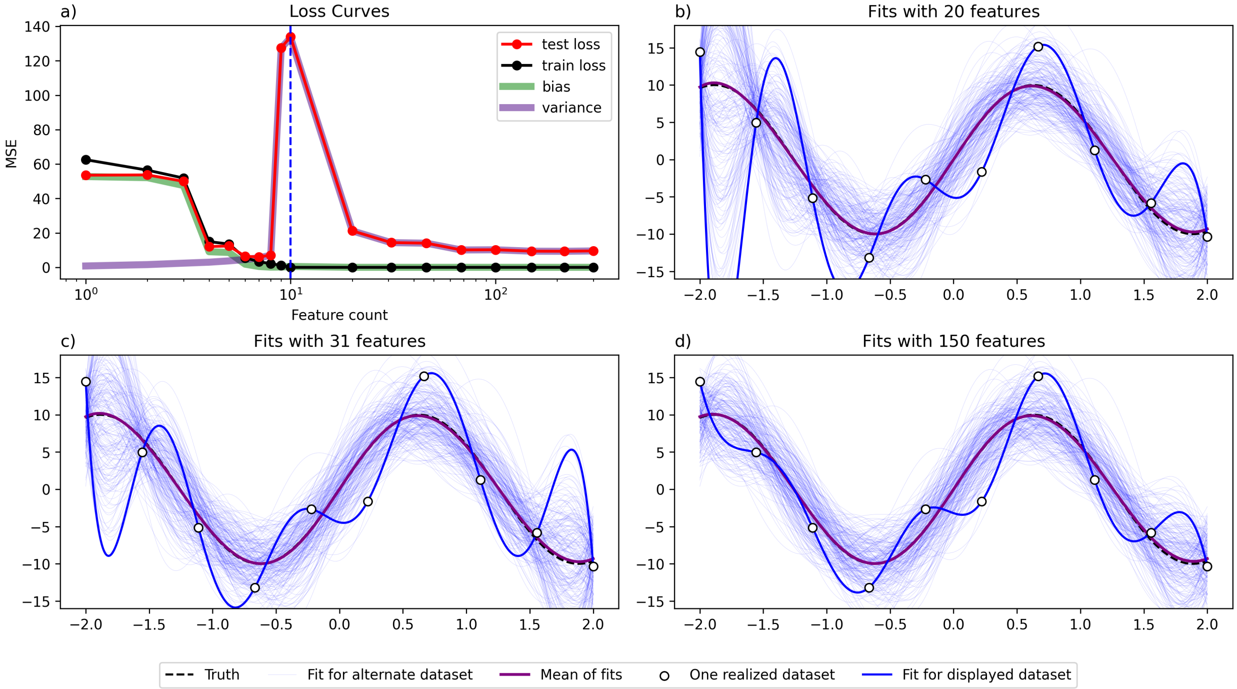


Figure 4: Same setting as in Figure 2. Despite having more parameters, the variance of increasingly large models falls.

extensively; while this began as early as the late 1980s [Loo+20], it’s recognition as a general empirical phenomenon in the machine learning literature began with [Bel+19]. Here, the authors describe how as the number of parameters increases while the training dataset size is fixed, the generalization error does indeed first increase as the classical theory suggests. However, as shown in Figure 3, this increase abruptly stops when the training loss reaches zero. This point, called the interpolation threshold, usually occurs when the number of parameters equals the number of datapoints, since in general it only takes n parameters to fit n datapoints [YSJ19]. Models of this size or larger now all exhibit zero training loss (and are hence referred to as interpolators), marking a phase transition in model behavior. Beyond the interpolation threshold, generalization error now decreases, despite the training loss remaining zero, signifying that the models remain overfit; this is in stark contrast to the classical theory, which suggests that the variance of such models should continue to grow, meaning the error should continue to climb. This non-monotone

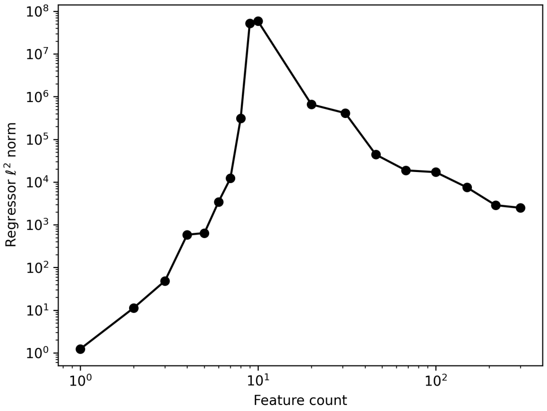


Figure 5: Regressor norm, in the same setting as Figure 2.

behavior of generalization error was dubbed *double descent*, and the fact that these overfit models have good performance was called *benign overfitting*. Figures 4(a-d) illustrate this trend in an experimental setting, as when the feature count continues to grow, the models are now all unbiased, as shown by the mean fits perfectly following the trend of the data. However, the variance of these models with larger feature counts in fact appears to be smaller, especially at the edges of the domain (compare the light blue lines in (d) with (b, c)).

In [Bel+19], the authors empirically reproduce double descent in Random Fourier Features regression, neural networks, and decision trees. They remark that past the interpolation threshold, there exist multiple solutions that perfectly fit the data. The possible key mechanism behind double descent is then that the solution chosen by stochastic gradient descent, the concept behind most empirical risk minimization for neural networks, naturally converges to "simple" solutions, in the sense of having a small ℓ^2 norm. As a result, as the flexibility of the model class increases, its ability to fit the smoothest possible interpolator increases, leading to decreasing risk in the overparameterized regime. See Figure 5 for an illustration of this trend. This implicit preference for simpler solutions, denoted an inductive bias, has motivated the study of the generalization errors of minimum norm interpolators of various kinds [BRT19].

In the time since, double descent has been analyzed in a variety of empirical and theoretical settings. For example, [Nak+21] investigates the phenomenon over a variety of optimization algorithms and network architectures on MNIST and CIFAR-10, concluding that the phenomenon is quite robust to the noise level, network structure, and training specifications. Likewise, [Nea+19] examines neural networks on a variety of datasets and empirically estimates model bias and variance, finding that, contrary to the classical understanding, variance of the fitted model past the interpolation threshold falls with model size; this is shown, for example, in Figure 4(a). As for theoretical results, while it has been hard to make progress in proving double descent in true deep neural networks, extensive analysis of overparameterization has been conducted in the well-worn statistical test bed of linear regression [Bar+20; CM22; HTF01], classification [Mon+23], as well as, more recently, in random features models [AP20b; DL20; MM19]. As our work builds on these results, we discuss them in detail in Chapter 1.

Despite the rich variety of models considered and approaches taken in these works, all of them remain in the relatively idealistic setting where the samples used to train the models of interest are identically and independently distributed (i.i.d.) with certain moment conditions – very little work attempts to understand their behavior when the samples are allowed to be dependent. In this work, we will first empirically examine why considering only this setting is insufficient in Chapter 2, and instead make the case for considering an alternate model for the randomness in the design, that of being right-rotational invariant. This is a rich class of random designs which in particular allow for heavy dependence between samples. Following this, in Chapter 3, we will then consider the behavior of overparameterization in linear regression for these designs, studying both the minimum norm setting as well as ridge regression. In Chapter 4, we will move to understanding how one can choose the optimal ridge regularization parameter; as we will show, the generalized cross validation in these problems is no longer consistent, and we present a provably consistent alternative. Finally in Chapter 5, we return to the i.i.d. setting to analyze random features models; we present partial progress on understanding the risks of these models where the weights are taken to be right-rotationally invariant rather than Gaussian, thereby moving closer to being able to understand deterministic designs. We give conjectures under which our results hold, and provide computational evidence of their validity.

1

The i.i.d. theory

1.1 Overparameterization in Linear Regression

As studying deep neural networks is analytically difficult, theoretical analyses of overparameterization began with studying the out-of-sample risks of well-specified linear regression. Concretely, one observes $(\mathbf{X}, \mathbf{y})$, such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

with $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\epsilon}, \boldsymbol{\beta} \in \mathbb{R}^p$. Hence n is the sample size, p is both the data dimension and the parameter count, $\boldsymbol{\beta}$ is the signal, and $\boldsymbol{\epsilon}$ is generally a component-wise independent noise vector. Under an assumed data generating distribution for $\mathbf{X}$, one then examines some version of out-of-sample risk[1] for a given estimator $\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})$, such as the

[1] The choice of this metric is varied. As our results follow those of [Has+22], we follow their convention and use the design-conditional excess generalization error $\mathbb{E}[\|\mathbf{X}_{\text{new}}\boldsymbol{\beta} - \mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}\|_2^2|\mathbf{X}]$, but also studied is this

usual least-squares estimator, or a ridge regularized variant. In particular, when $p > n$, the least-squares estimator $(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ is not defined – replacing the inverse with the Moore-Penrose inverse results in the minimum norm linear interpolator. For all works surveyed below, $\mathbf{X}$ is assumed to have i.i.d. rows, and all quantities are scaled such that the out-of-sample risk remains order 1 in the limit.

This problem is of course well understood in the classical setting, where p is fixed and n is taken to infinity. Here, it is known that the least-squares estimator is the best linear unbiased estimator, consistent, and has fluctuations $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ distributed as $\mathsf{N}(0, \sigma^2(n^{-1}\mathbf{X}^\top \mathbf{X})^{-1})$.

With the rise of so-called "big data", where the number of parameters is large compared to the number of datapoints, attention has turned to settings where n and p grow together to infinity. This is largely separated into three regimes of consideration. The first is when $p = o(n)$, which was studied as early as in 1988, with [Por88], which reobtains classical guarantees in the setting where $p = \Theta(\sqrt{n})$. If one considers p as the number of directions one must learn, and n as the amount of information on hand, then in this setting, one still has overwhelmingly more information than directions to learn, and hence it makes sense to reobtain classical guarantees. The second regime is where $p = \omega(n)$. This setting has many more directions to learn than datapoints given - as a result, structural assumptions such as sparsity are generally necessary [CT07; Has15], where $\|\boldsymbol{\beta}\|_0$ is generally assumed to grow like $o\left(\frac{n}{\log p}\right)$. Under such sparsity assumptions, the effective dimension is much lower, and one is essentially back in the $p = o(n)$ setting; hence similar results emerge [Büh11]. The final regime is that where $p = \Theta(n)$. This is referred to as "linear" or "proportional" asymptotics, and has received much attention since most classical results indeed break down at this point, and yet it is still possible to derive interesting results without assumptions on sparsity. In particular, these results are often quite different than those in the classical setting (see e.g. [SC19; SCC19]).

Study of linear regression under proportional asymptotics began due to interests independent of understanding double descent, and hence early results did not directly examine the effects risks as a function of overparameterization (that is, how the risk

risk further conditioned on $\mathbf{y}$ and quantities such as $\mathbb{E}[\|\mathbf{y}_{\text{new}} - \mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}\|_2^2 | \mathbf{X}]$. These are all largely equivalent modulo additive noise constants due to concentration. See Appendix A.6.1 for details.

changes as a function of the overparameterization ratio $\gamma := p/n$). While the earliest works in this area, which began in the compressed sensing literature (e.g. [BM12]), were already able to handle a deterministic signal vector $\boldsymbol{\beta}$, the main line of research on which our work builds began with $\boldsymbol{\beta}$ drawn from some delocalized prior, such as an i.i.d. product prior or the uniform distribution on the sphere. In [Dic16], the author study the asymptotic risk of ridge regression in the proportional limit. They assume a linear model, with the covariates x_i drawn from an isotropic Gaussian and the errors independently Gaussian. Extensions to sufficiently regular covariances, as well as under relaxed assumptions, are presented in [DW18], which derives asymptotic risks for predictors of the form $\mathbf{x}_i = \mathbf{\Sigma}^{1/2}\mathbf{z}_i$, with $\mathbf{\Sigma}$ having controlled operator norm and $\mathbf{z}_i$ having independent entries and bounded 12th moment. These results utilize random matrix theory to characterize the high dimensional limit, as the risks of interest fundamentally hinge on the spectrum of the empirical covariance matrix; the need for these results from random matrix theory is partially the reason for the moment conditions.

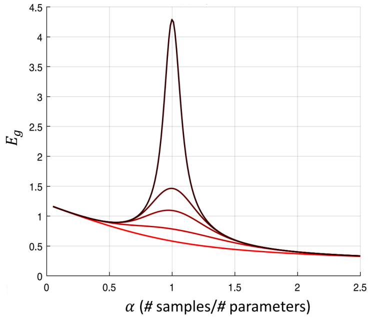


Figure 1.1: [ASS20, Figure 2C] – generalization error at various a variety of training times. $\alpha = 1/\gamma$. Note that this is not equivalent to the risk of minimum norm interpolation, as the training time to convergence differs across α.

Amazingly, even this simple setting can reproduce the hallmarks of double descent. This was noticed as early as in [ASS20] (see Figure 1.1). Here, the authors solve for the training dynamics of gradient descent on the linear model, and in particular analyze the generalization error as a function of γ. They find that when $\gamma = 1$, the left tail of the eigenvalues of the Gram matrix $\mathbf{X}^\top \mathbf{X}$ approaches zero, leading to "catastrophic overtraining" resulting from the eigenvalues of the inverse matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ exploding, causes the model to generalize very poorly. The analysis is also in the high dimensional limit, and again exploits jointly Gaussian data and labels, and furthermore analyzes the effects of early stopping and ridge regularization in preventing the generalization error from exploding near the interpolation threshold. The same model is studied, though with a different framing, in [BHX20], along with a noise-free Fourier series model, where double descent with respect to the parameter count is again observed. A much later work, [MG21], shows that one can actually observe multiple descent peaks under certain settings on the population spectrum.

The study of minimum norm linear regression is intimately related to that of the solution obtained through gradient descent, since gradient descent naturally trains to the minimum norm interpolator when initialized at zero[2] if the learning rate is sufficiently small[3]. Hence, this problem has received much attention over the past few years. Benign overfitting, in this context, refers to when these minimum norm interpolators achieve near optimal risk (among all ridge regularized variants); early works tried to derive conditions under which this occurs. The overall understanding is that a few entities are crucial in the analysis of benign overfitting: the spectra of the population covariance matrix $\mathbf{\Sigma}$ and the empirical covariance matrix $\mathbf{X}^\top\mathbf{X}/n$, along with the alignment of the true signal vector $\boldsymbol{\beta}$ with the top eigenvalues of $\mathbf{\Sigma}$. Finite sample bounds on the test error, along with necessary conditions for benign overfitting on the population covariance were derived in [Bar+20]; here, the authors find that the tail of the eigenvalues of $\mathbf{\Sigma}$ cannot decay too quickly. Another result [KLS20] shows that if $\boldsymbol{\beta}$ is aligned with the top eigenvectors of $\mathbf{\Sigma}$, then it is also possible for the ridgeless limit to be optimal. The authors in this work provide further mechanistic intuition to how this occurs: adding covariates with *zero* signal and small variance to the regression problem can in fact provide implicit ridge regularization when using minimum norm regression – in practice, this is provided by directions of small variance with low correlation with the true signal, which are orthogonal to the signal. As a result of this already present implicit regularization, the optimal ridge parameter is required to be zero or even negative. The effect of this alignment of the signal with the

[2] This result is quite classical, see e.g. [Has+22, Proposition 1] for a proof

[3] The learning rate η must be less than $1/\lambda_{\max}(\mathbf{X}^\top\mathbf{X}^{-1})$.

covariance is further studied in [RMR20]; in fact, having this alignment is quite common in real world datasets – see [LS23] for examples. Lastly, [WX20] also examines this problem where $\boldsymbol{\beta}$ is drawn from an anisotropic prior, as well as the related problem of applying weighted regularization.

A large number of these results were extended under relaxed assumptions in [Has+22], employing truncation arguments to bypass moment conditions. The authors give finite sample bounds for the out-of-sample risk for fixed $\boldsymbol{\beta}$ in a variety of settings, including anisotropy and misspecification (where only a fraction of the relevant predictors are observed). Most significantly, while prior works on linear models had all illustrated the characteristic peaking effect of double descent, generally the optimal value of γ lay in the underparameterized regime, which differs from the behavior of deep learning, wherein the large models reign supreme. In contrast, [Has+22] displays that, under certain misspecification settings, the optimal choice of γ can lie beyond 1, thereby obtaining a setting aligning with modern practice, though the proposed misspecification structure is somewhat ad-hoc. Improved finite sample bounds were then given in [CM22], which furthermore moves away from the proportional asymptotic regime.

1.1.1 Cross-validation in high dimensions

Another issue that emerges in high dimensional inference is cross-validation. Cross-validation is a method for selecting model hyperparameters (such as the ridge regularization parameter) which has improved data efficiency as opposed to a strict train-validate-test split. In practice, k-fold cross validation is usually used because it is computationally light compared to the leave-one-out variant. However, [RM18] and [Wan+18] both experimentally show that k-fold validation induces very large biases in high dimensional settings, thereby rendering it useless. At the same time, both also show that approximate forms of leave-one-out cross-validation (which remain computationally tractable) for their respective problems still succeed in being consistent. This problem for linear regression under proportional asymptotics was examined by [Has+22]; here, it is shown that the usual leave-one-out cross-validation (LOOCV), as well as the related generalized cross-validation (GCV), both are asymptotically (under proportional asymptotics) uniformly consistent on compact intervals. In turn, we will also study GCV and LOOCV in our setting.

1.2 The Random Features Model

While the focus of this work will be on studying the effects of overparameterization in linear regression, studying risk as a function of overparameterization is not entirely natural in this setting. This is because the parameter count is directly tied to the data dimension. In practice, on the other hand, one generally imagines having data of fixed dimension, and then choosing a model of some complexity afterwards. A tractable model in which parameter count can be isolated from data dimension is that of the random features model, which has received much attention as well. Concretely, one considers a one-hidden layer neural network with the first layer weights frozen and learns the optimal last layer weights. That is, one observes $(\mathbf{X}, \mathbf{y})$ generated from some process, with $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^{n}$. Instead of trying to learn a linear model as before, one chooses a weight matrix $\mathbf{F} \in \mathbb{R}^{d \times p}$. Note that the data dimension is now d instead of p, and that p is now used to denote the parameter count. As in the forward pass of a neural network, one now computes the featurization $\sigma(\mathbf{XF}) \in \mathbb{R}^{n \times p}$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is some prescribed activation function applied component-wise. From here, one performes (possible regularized) linear regression of $\mathbf{y}$ onto $\sigma(\mathbf{XF})$, yielding some estimated final layer weights $\mathbf{a}$, and producing the estimator $\hat{f} : \mathbf{x} \mapsto \mathbf{a}^\top \sigma(\mathbf{F}^\top \mathbf{x})$.

This is called the *random* features model because the feature map $\mathbf{F}$ is drawn from some distribution, often having i.i.d. Gaussian entries; note that it is not learned. The samples in $\mathbf{X}$ are usually taken to be i.i.d standard Gaussian. This model, with linear activations, was studied in [ASS20], though only empirically; the Fourier series model studied in [BHX20] is somewhat similar in spirit.

Results for this model again are derived under proportional asymptotics, where one takes $n, p, d \to \infty$, with n/p and p/d converging to constants. The final section of [HTF01] derives the asymptotic variance of this model when the random weights are taken to be i.i.d. Gaussian, and the activation satisfies $\mathbb{E}[\sigma(Z)] = 0$ and $\mathbb{E}[\sigma(Z)^2] = 1$, where $Z \sim \mathsf{N}(0, 1)$. The picture is completed by [MM19], which fully solves the asymptotic risk of these models (though now the weight matrix has columns which are i.i.d. from the sphere). These methods directly leverage the fact that the weights are i.i.d., as they largely hinge on using the leave-one-out method to compute resolvents of a suitably constructed matrix. The double descent curves obtained in [MM19] feature global minima which lie in the overparameterized regime, thus illustrating a setting fully showcasing all hallmarks of double descent without the need of a somewhat arbitrary choice of misspecification. The case of random features kernel regression, which can be viewed as an infinite width limit of this setting, is examined in [AP20b], where the same double descent curves are obtained; the behavior of this model when the inputs have some covariance is examined in [MP22].

Two key observations are made in these works. The first is that the risk of random feature regression is equivalent to that of a linear model with a certain choice of covariance. This inspired the works [DL20; HL20], which explicitly show this equivalence using the Lindeberg method, and use this equivalence to compute the asymptotic risks. Specifically, one decomposes the activation function in the Hermite basis (a basis of polynomials that form an orthonormal basis for $L^2(\mathbb{R}, \Phi)$, where Φ refers to the standard Gaussian measure). We assume the normalization condition $\mathbb{E}_{Z\sim \mathsf{N}(0,1)}[\sigma(Z)] = 0$. One can then write $\sigma(x) = \mu_1 x + \sigma_\perp(x)$, where c_1 is the first order Hermite coefficient and $\sigma_\perp(x)$ contains the remaining terms of the Hermite expansion. From the fact that the Hermite polynomials are an orthonormal basis, $\sigma_\perp$ satisfies $\mathbb{E}[Z \cdot \sigma_\perp(Z)] = \mathbb{E}[\sigma_\perp(Z)] = 0$, and $\mathbb{E}[\sigma_\perp(Z)^2] = (\mathbb{E}[\sigma(Z)^2] - \mu_1^2)^{1/2} =: \mu_2$. Then, miraculously, even though $\sigma_\perp(Z)$ is only *uncorrelated* with Z, in fact, asymptotically, it acts *fully independently*! That is, as $n, p, d \to \infty$, the random features model, which uses features of the form $\sigma(\mathbf{F}^\top \mathbf{x}) = \mu_1 \mathbf{F}^\top \mathbf{x}_i + \sigma_\perp(\mathbf{F}^\top \mathbf{x}_i)$, assumes the same train and test risk as a linear model using features of the form $\mu_1 \mathbf{F}^\top \mathbf{x}_i + \mu_2 \mathbf{z}_i$ (provided $\mathbf{F}$ has i.i.d. suitably normalized Gaussian entries), where $\mathbf{z}_i$ is a fully independent Gaussian. See [HL20] for further details.

The second is that random features models are only capable of learning linear functions of the data [AP20a; MM19], unveiling a fundamental ceiling to how much one can understand by studying this model. Of course, such things are not true in practice, and thus these limitations are consequences of random features models failing to capture the significant mechanism of feature learning, since their weight matrices are fixed. Exciting work in this direction has occurred recently – see [Ba+22; Dan+23; Mon+24], which show that with some gradient steps, the higher degree Hermite portions of a given target function can be learned.

2

When the i.i.d. theory fails

In this section, we empirically examine the behavior of the out-of-sample risk as the design departs from the i.i.d. setting. We begin in totally synthetic settings, where the design $\mathbf{X}$, the signal $\boldsymbol{\beta}$, and the noise $\boldsymbol{\epsilon}$ are generated purely randomly, and observe what can occur as the design $\mathbf{X}$ gains correlation between rows or grows fatter tails. Afterwards, we move to a semi-synthetic setting, where the design is a real dataset, but the signal and noise are still generated. Here, we show that in these settings, the assumption that the rows are i.i.d. can cause the predictions for the behavior of such models to become biased. In the chapters that follow, we present formulas for the risk which capture the behavior of the risk by modeling the design as *right-rotationally invariant* (to be precisely defined in Chapter 3), producing a better characterization of the risk in the settings we examine. That these designs can be used to better characterize real-world datasets has been analyzed in prior work, such as in [LS23]. The details for all experiments below can be found in Appendix A.5.2.

2.1 Preliminaries

In the simulations of this chapter, following existing literature, [CM22; Has+22], we measure the following notion of out-of-sample risk (see Appendix A.6.1 for discussion):

$$R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{n'}\mathbb{E}\left[\left\| \mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}} - \mathbf{X}_{\text{new}}\boldsymbol{\beta} \right\|^2 \middle| \mathbf{X} \right]. \tag{2.1}$$

where $\mathbf{X}_{\mathsf{new}}$ is some independent matrix of dimension $n' \times p$. If $\mathbf{X}_{\mathsf{new}}$ has i.i.d. rows, then all settings of n' are equivalent. In this chapter, $\hat{\boldsymbol{\beta}}$ will always be the usual OLS estimator when $n \leq p$, and the minimum norm estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\mathbf{y}$ when $p > n$.

The extensive literature on the setting where the rows are i.i.d. gives predictions on the out-of-sample risk above for quite general conditions on the covariance $\boldsymbol{\Sigma}$ and signal vector $\boldsymbol{\beta}$. Here, we compare those predictions given for the isotropic setting (wherein the alignment of $\boldsymbol{\beta}$ no longer matters) with the predictions give by our right-rotationally invariant theory.

2.2 Synthetic Settings

Figure 2.1 illustrates multiple synthetic settings in which we plot the predictions given by the i.i.d. theory as well as our own. The last three illustrate cases that are covered by our right-rotationally invariant theory but not that of the i.i.d. theory. We examine the following four settings, and give potential areas in which they could emerge:

1. I.i.d. data: as expected, the predictions from the i.i.d. theory hold, as do ours. The data here are i.i.d. Gaussians, which are provably right-rotationally invariant, but we note that the same results are found for other distributions of covariates, such as suitably scaled uniform and Rademacher random variables (see Appendix A.1).[1]

[1] One may question why our finite sample risks do not match the theoretical predictions as well as those in [Has+22] – the reason for this is somewhat subtle. The first version of the publication focused on $\boldsymbol{\beta}$ drawn from an isotropic prior, and hence the risks they show have additional averaging. The paper was subsequently updated for fixed $\boldsymbol{\beta}$, but the plots were not changed.

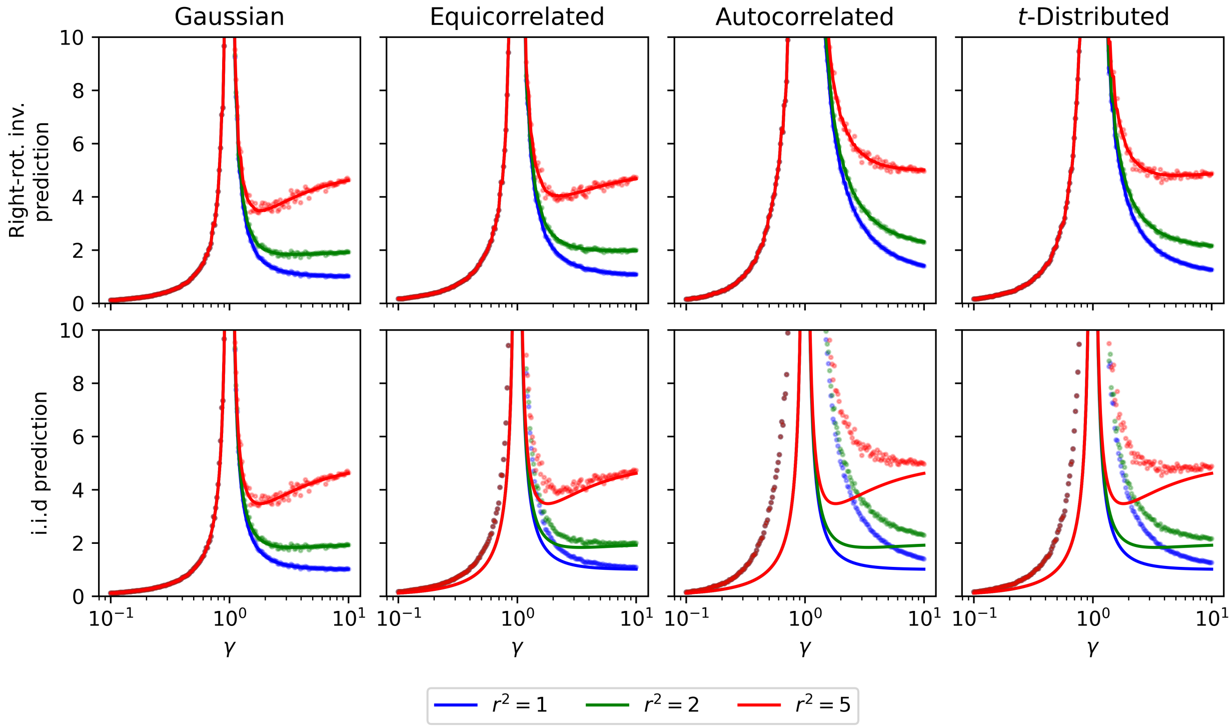


Figure 2.1: Solid lines display the theoretical predictions of the risk; dotted points show finite sample realizations, plotted over a range of the aspect ratio γ. The dimension of all designs is $n = 200$, $p = [\gamma n]$. All settings suitably scaled so that each column has variance 1 and mean 0. For the equicorrelated cases, the common correlation is $\rho = 0.6$. For the autoregressive case, $\rho = 0.8$.

2. Equicorrelated data: $\mathbf{X} \in \mathbb{R}^{n \times p}$ has independent columns, but each column is distributed as a multivariate Gaussian with covariance matrix $\mathbf{\Sigma}$, where $\Sigma_{ij} = \rho$ if $i \neq j$, and $\Sigma_{ii} = 1$. One can imagine that if the design contains data gathered from individual populations, then groups sharing a certain latent factor (such as location) naturally have correlated covariates. If one tries to assume that the rows are still i.i.d., this produces naturally biased predictions for the risk:

3. Autocorrelated data: One could also imagine that the rows of $\mathbf{X}$ are drawn from some time-series. In such settings, it is natural for the rows to be autocorrelated. Here, the rows of $\mathbf{X}$ satisfy $\mathbf{x}_i = \rho \mathbf{x}_{i-1} + \sqrt{1-\rho^2}\epsilon_i$, where ϵ_i are i.i.d. draws from $\mathsf{N}(0, \mathbf{I}_n)$.

4. t-distributed data: Prior results, as discussed in Chapter 1, generically assume some

moment conditions on the covariates $\mathbf{x}$. It is questionable whether certain types of data, such as financial returns, satisfy such conditions, due to having much heavier tails [Gro21]. Here, we examine the behavior when the rows of $\mathbf{X}$ are drawn from a multivariate t distribution, with mean 0 and scale parameter 1/3, so that the variance of any column remains 1.

2.3 Semi-synthetic Settings

It is not surprising that our theory out-performs the i.i.d. theory for the designs constructed above. We now examine semi-synthetic settings, in that the design $\mathbf{X}$ is from a real dataset, but the signal $\boldsymbol{\beta}$ and the noise $\boldsymbol{\epsilon}$ are generated. We then empirically estimate the risk $R_{\mathbf{X}}$ by testing the fitted $\boldsymbol{\beta}$ on a (real) test data set.

2.3.1 Speech data

As studied in [LS23], we examine designs where each row consists of an i-vector of a speech segment (see [IR18]). Full experimental details can be found in A.5.2. Figure 2.2 illustrates that our predictions better match the behavior of these models, primarily in the overparameterized regime. That the gap to the i.i.d. prediction increases as σ^2 is increased points to the fact that treating the design as right-rotationally invariant may better capture the variance in the estimator.

2.3.2 Financial Data

The next type of design we examine is that of minutely residualized returns[2] of various stocks. One can think of trying to predict a linear function of residualized returns as trying to reconstruct the components of a given portfolio given its residualized returns by regressing it onto the residualized returns of its constituents (see Section 3.5.3 for more).

A natural reason for autocorrelated rows emerges in such a setting. If one has, for instance, the return of a given stock over the past 5 minutes in each row, but measures this

[2] See Appendix A.5.2 for details on residualized returns, including our residualization process.

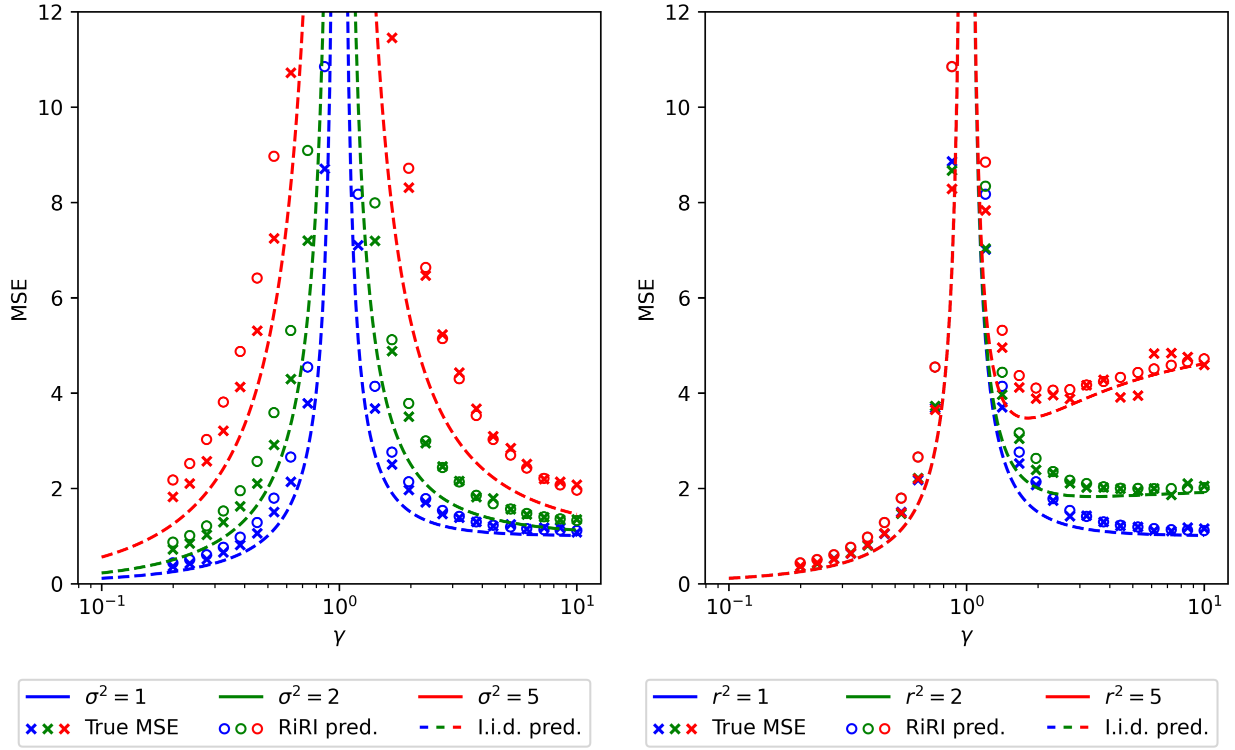


Figure 2.2: Speech data. Here, the design has $p = 400$, and $n = [p/\gamma]$. When varying σ^2, r^2 is held fixed at 1. For varying r^2, σ^2 is held fixed at 1. RiRI pred. refers to the prediction given by our right-rotationally invariant theory; i.i.d. pred refers to the prediction given by the i.i.d. theory.

every 1 minute, then naturally successive rows possess very high correlation, because they look at an overlapping period of 4 minutes. Figure 2.3 presents the behavior of the risk when this occurs. Here, k refers to the number of minutes of returns each row contains. So if $k = 1$, each row contains the returns over the previous minute, and thus there is no overlap, since we observe all returns every minute. If $k = 3$, there a 2 minute overlap between successive rows, producing autocorrelation. Experimental details are in A.5.2.

Note that even when $k = 1$, our characterization better predicts the behavior of the risk, potentially due to the heavy tails that are present in financial data. When $k > 1$ and there is autocorrelation between rows, the risk explodes more quickly on both sides of the interpolation boundary $\gamma = 1$; the predictions from the i.i.d. theory fail to adjust for this, but our predictions are still somewhat accurate.

Thus, from these examples it is clear that the i.i.d. assumption is not always an accurate characterization of designs that arise in real-world datasets. We now shift to a discussion of our theoretical analysis of the risks of right-rotationally invariant designs. We defer discussion of a fully real experiment to Section 3.5.3.

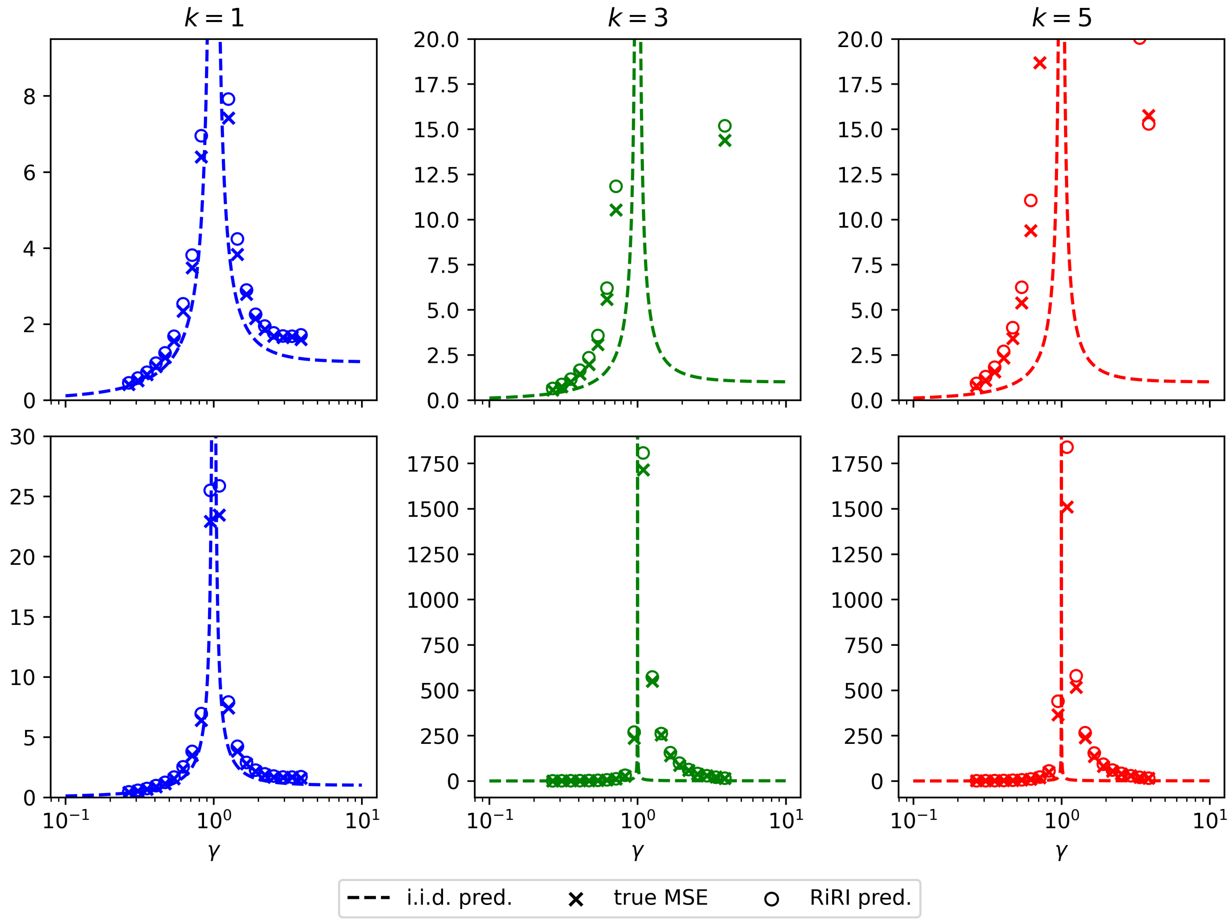


Figure 2.3: Risks for various settings of k. Top plot and bottom plot of each column are the same, just with different y scale to better show the behavior near the top and bottom of the plot. RiRI pred. refers to the prediction given by our right-rotationally invariant theory; i.i.d. pred refers to the prediction given by the i.i.d. theory.

3

Benign overfitting beyond i.i.d.

As discussed earlier, our theory is rooted in modeling the data $\mathbf{X}$ as being *right-rotationally invariant* – this is a rich class of designs which allow for tractable analysis while also capturing potential dependence between datapoints. In this chapter, we study certain out-of-sample risks for linear regression when the design is modeled in this manner.

3.1 Preliminaries

Definition 3.1 (Right rotationally invariant design)**.** A random design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is *right rotationally invariant* if for any $\mathbf{O} \in \mathbb{O}(p)$, one has $\mathbf{X}\mathbf{O} \overset{d}{=} \mathbf{X}$, where $\mathbb{O}(p)$ denotes the group of $p \times p$ orthogonal matrices. Equivalently, let $\mathbf{X} = \mathbf{Q}^\top \mathbf{D}\mathbf{O}$ be the singular value decomposition of $\mathbf{X}$. Then $\mathbf{X}$ is right rotationally invariant if and only if $\mathbf{O}$ is independent of $(\mathbf{Q}, \mathbf{D})$ and $\mathbf{O}$ drawn from the Haar measure on $\mathbb{O}(p)$ - that is, it is uniformly distributed on $\mathbb{O}(p)$. See Lemma A.2 for proof of equivalence.

Note that the isotropic Gaussian is a specific case of a right-rotationally invariant design, wherein $\mathbf{Q}$ is also Haar, independently of $(\mathbf{D}, \mathbf{O})$. As we will see, our analysis allows us to reobtain the results of [Has+22] for isotropic Gaussians, since it is well understood that the entries of $\mathbf{D}$ converge weakly to the square root of the Marchenko-Pastur law. Without loss of generality, we will assume that the entries on the diagonal of $\mathbf{D}$ are sorted such that $D_{11} \geq D_{22} \geq \cdots \geq D_{\min(n,p),\min(n,p)} \geq 0$.

These designs have recently received much attention as alternatives to fully Gaussian designs for theoretical analysis in a multitude of works, notably [DSL23; Fan22; RSF19]. See [LS23] for a more comprehensive list of related works. Crucially, [DSL23] shows that, under certain spectral conditions, the behavior of high dimensional matrices is largely dictated by their spectrum, and in fact will "act" similarly to a right rotationally invariant matrix with the same spectrum. As a result, studying these designs may be suggestive of behavior arising in more general non-Gaussian models.

Some other examples of right-rotationally invariant designs, in addition to those presented in Chapter 2, are as follows:

1. products of Gaussian matrices: $\mathbf{X} = \mathbf{X}_1 \mathbf{X}_2 \cdot \cdots \cdot \mathbf{X}_k$, where $\mathbf{X}_1$ has n rows and $\mathbf{X}_k$ has p columns, while the remaining dimensions are arbitrary.

2. spiked matrices: $\mathbf{X} = \lambda \mathbf{V} \mathbf{W}^\top + \mathbf{G}$, where $\mathbf{V} \in \mathbb{R}^{n \times r}$ and $\mathbf{W} \in \mathbb{R}^{p \times r}$ are the first r columns of two Haar matrices.

See [LS23, Figure 1] for more.

In essence, right rotationally invariant ensembles allow us the additional degree of freedom of allowing some dependence between rows, and allowing us to set the singular value distribution of the design of interest.

At the same time, the fact that $\mathbf{O}$ is Haar does add some restrictions - in particular, the designs we consider are now necessarily isotropic.

Lemma 3.1. *Let* $\mathbf{X} = \mathbf{Q}^\top \mathbf{D} \mathbf{O}$ *be right rotationally invariant. Then* $\mathbb{E}[\mathbf{X}^\top \mathbf{X}] = \mathsf{Tr}(\mathbb{E}[\mathbf{D}^\top \mathbf{D}])/p \cdot \mathbf{I}_p$.

Proof. Let $\mathbf{O}$ have rows $\mathbf{o}_i$. Then

$$\begin{aligned}\mathbb{E}[\mathbf{X}^\top \mathbf{X}] = \mathbb{E}[\mathbf{O}^\top \mathbf{D}^\top \mathbf{D} \mathbf{O}] &= \sum_{i=1}^{n\wedge p} \mathbb{E}[D_{ii}^2 \mathbf{o}_i \mathbf{o}_i^\top] = \sum_{i=1}^{n\wedge p} \mathbb{E}[D_{ii}^2]\mathbb{E}[\mathbf{o}_i \mathbf{o}_i^\top] \\ &\overset{(*)}{=} \frac{1}{p}\mathbf{I}_p \sum_{i=1}^{n\wedge p} \mathbb{E}[D_{ii}^2] = \mathsf{Tr}(\mathbb{E}[\mathbf{D}^\top \mathbf{D}])/p \cdot \mathbf{I}_p\end{aligned}$$

where $(*)$ follows from $\mathbf{I}_p = \mathbb{E}[\mathbf{O}^\top \mathbf{O}] = \sum_{i=1}^{p} \mathbb{E}[\mathbf{o}_i \mathbf{o}_i^\top]$ and the exchangeability of the $\mathbf{o}_i$ (which is derived from the fact that $\mathbb{O}(p)$ contains the permutation matrices). □

In prior analyses, the spectrum of $\mathbf{\Sigma}$ defines the problem geometry. For us, we sacrifice this degree of freedom for the ability to set the spectrum $\mathbf{D}$ to capture distributions beyond those of i.i.d. rows.

3.2 Notation and Setup

We study the linear regression posed in (1.1) when $\mathbf{X}$ is right-rotationally invariant. Specifically, we observe the pair $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a right-rotationally invariant design and $\mathbf{y} \in \mathbb{R}^n$ is generated via

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.1}$$

We take $\boldsymbol{\beta} \in \mathbb{R}^p$ to be fixed and $\boldsymbol{\epsilon}$ to have independent entries with mean 0 and variance σ^2.

The problem we consider will be the expected out-of-sample risk on an independent draw from a new, potentially different right rotationally invariant design, $\mathbf{X}_{\mathsf{new}} = \mathbf{Q}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}} \mathbf{O}_{\mathsf{new}}$, with $\mathbf{X}_{\mathsf{new}} \in \mathbb{R}^{n'\times p}$. This corresponds, for example, to the setting of learning a model on one interdependent population (such as medical data from some population), and using the model on a separate interdependent population. As mentioned in the preceeding chapter, the explicit risk we examine, following [CM22; Has+22], is

$$R_{\mathbf{X}}\left(\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}), \boldsymbol{\beta}\right) = \frac{1}{n'}\mathbb{E}\left[\left\|\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}} - \mathbf{X}_{\text{new}}\boldsymbol{\beta}\right\|^2 \middle| \mathbf{X}\right]. \tag{3.2}$$

This setting is comparable to the isotropic setting in [Has+22] (though see Section 3.2.1 for differences in scaling). Though only the Gaussian setting is explicitly covered by our designs, we believe that they should in fact hold in general for isotropic data, similarly to the universality posed in [DSL23]. See Appendix A.1 for some evidence of this universality.

Most of our results will be stated in terms of finite sample bounds. Occasionally we will consider the asymptotics as $n, p \to \infty$ where $p/n \to \gamma \in (0, \infty)$. In such cases, we may further assume that the entries on the diagonal of $\mathbf{D}$ converge weakly to some distribution μ_{D} in a certain sense, to be detailed later on. A helpful working example is to think of $\mathbf{X}$ as having i.i.d. entries distributed as $\mathsf{N}(0, 1/n)$. Then the entries of $\mathbf{D}$ converge to the square root of the Marchenko-Pastur law with parameter p/n.

Remark 1. One can ask why we do not consider the problem of dependent prediction. In particular, we could try partitioning the rows and columns of $\mathbf{X}$ and $\mathbf{y}$ as $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top]^\top$ and $\mathbf{y} = [\mathbf{y}_1^\top, \mathbf{y}_2^\top]^\top$, where $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{y}_1 \in \mathbb{R}^{n_1}$, while $\mathbf{X}_2 \in \mathbb{R}^{n_2 \times p}, \mathbf{y}_2 \in \mathbb{R}^{n_2}$, with $n_1 + n_2 = n$. We then try using $(\mathbf{X}_1, \mathbf{y}_1)$ to learn $\hat{\boldsymbol{\beta}}_1$, and then calculate the risk

$$\frac{1}{n_2}\mathbb{E}\left[\|\mathbf{X}_2\hat{\boldsymbol{\beta}} - \mathbf{X}_2\boldsymbol{\beta}\|^2 \mid \mathbf{X}_1\right].$$

Note here that $\mathbf{X}_1$ and $\mathbf{X}_2$ would then have some complex dependence due to being drawn jointly from the right rotationally invariant ensemble. This, for example, corresponds to learning some model for predicting a certain quantity based on the first n_1 timesteps of an autocorrelated time series, and then using it to predict on the next n_2 timesteps.

This computation can indeed be done using a certain result for conditioning on Haar matrices that exists in the literature [RSF19, Lemma 4]. The problem is that the resulting form of the risk is quite complicated and depends on $\mathbf{Q}$, for which we have no distributional assumptions. The result is quite uninterpretable and provides no insights, and hence is not pursued – see Appendix A.2.1 for extended discussion.

3.2.1 Nuances on Scaling

Recall that for our setting, we have $\mathbb{E}[\mathbf{X}^\top\mathbf{X}] = \mathcal{O}(1) \cdot \mathbf{I}$, while when the rows of $\mathbf{X}$ are i.i.d. draws from some prescribed distribution with identity covariance, one would instead expect $\mathbb{E}[\mathbf{X}^\top\mathbf{X}/n] = \mathbf{I}$. As such, one should think of $\mathbf{X}$ in our setting as akin to the normalized matrix $\mathbf{X}/\sqrt{n}$ in the i.i.d. setting. Furthermore, $\mathbf{D}^\top\mathbf{D}$ will be a diagonal matrix containing the eigenvectors of $\mathbf{X}^\top\mathbf{X}$. This scaling ensures that, for example, in the setting where the entries are distributed as $\mathsf{N}(0, 1/n)$, that $\|\mathbf{D}^\top\mathbf{D}\|_{\mathsf{op}} = O(1)$. Only in this way is it natural to consider the entries of $\mathbf{D}$ as converging to some limiting distribution μ as $n, p \to \infty$ with $p/n \to \gamma$. In the Gaussian case, this would be the Marchenko-Pastur law with parameter γ.

The consequence of this is that in order to maintain the signal to noise ratio, we must have $\boldsymbol{\beta}$ having norm on the order of $\sqrt{n}$. This will cause some of our results to superficially look different than those in [CM22; Has+22].

3.3 Useful Results

We collect a few useful results before we begin. First, we will often require the concentration of quadratic forms $\mathbf{v}^\top\mathbf{A}\mathbf{v}$, with $\mathbf{v}$ uniform on the sphere and $\mathbf{A}$ fixed or independent. As the uniform distribution on the sphere has dependent coordinates, the standard Hanson-Wright inequality does not apply, and instead we use the following Hanson-Wright inequality for spherical vectors.

Lemma 3.2 (Hanson-Wright for Spherical Vectors)**.** *Let $\mathbf{v} \in \mathbb{R}^n$ be a random vector distributed uniformly on S^{n-1} and let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a fixed matrix. Then*

$$\mathbb{P}(\left|\mathbf{v}^\top\mathbf{A}\mathbf{v} - \mathbb{E}[\mathbf{v}^\top\mathbf{A}\mathbf{v}]\right| \geq t) \leq 2\exp\left(-C\min\left(\frac{n^2t^2}{2K^4\|\mathbf{A}\|_F^2}, \frac{nt}{K^2\|\mathbf{A}\|}\right)\right).$$

for some absolute constants C, K.

Proof. Isoperimetric inequalities imply that vectors uniformly distributed on S^{n-1} satisfy convex concentration (see [Sch14]) with constant $K/\sqrt{n}$, for some absolute constant K, where we note that having sub-Gaussian tails around the median is equivalent to having them around the mean (see e.g. [Cha22]). We can then apply [Ada14, Theorem 2.3] and rescale. Alternatively, one should also be able to view $\mathbf{v}$ as a renormalized Gaussian, then apply standard Hanson-Wright and separately control the norm of the Gaussian using the Bernstein inequality. □

We now present a simple lemma for computing the expectations in Hanson-Wright.

Lemma 3.3 (Expectations of Quadratic Forms)**.** *Let* $\mathbf{v}$ *be uniformly distributed on* S^{n-1}*, and let* $\mathbf{A}$ *be an* $n \times n$ *matrix. Then*

$$\mathbb{E}[\mathbf{v}^\top \mathbf{A} \mathbf{v}] = \mathsf{Tr}(\mathbf{A}\mathbb{E}[\mathbf{v}\mathbf{v}^\top]) = \frac{\mathsf{Tr}(\mathbf{A})}{n} \tag{3.3}$$

Proof. The first equality is clear. The second can be seen to follow from the observation used in the proof of Lemma 3.1 and the fact that columns of a Haar distributed matrix are uniform on the sphere (see e.g. [Mec19]). Alternatively, recall that for $\mathbf{g} \sim \mathsf{N}(0, \mathbf{I}_n)$, $\frac{\mathbf{g}}{\|\mathbf{g}\|} \perp\!\!\!\perp \|\mathbf{g}\|$. Hence

$$\mathbb{E}[\mathbf{v}\mathbf{v}^\top] = \mathbb{E}[\mathbf{g}\mathbf{g}^\top / \|\mathbf{g}\|^2] = \mathbb{E}[\mathbf{g}\mathbf{g}^\top]/\mathbb{E}[\|\mathbf{g}\|^2] = \mathbf{I}_n/n.$$

□

3.4 Some notation

As detailed in Chapter 1, of central importance in the study of this problem is the empirical spectral distribution of the Gram matrix $\mathbf{X}^\top\mathbf{X}$. While in prior works, heavy machinery from random matrix theory is introduced to handle this problem, either explicitly as in [DW18; Has+22], or implicitly using self-contained leave-one-out type arguments, as in [Bar+20; CM22], we have somewhat abstracted out these difficulties by admitting explicit access to the spectrum through $\mathbf{D}$. As one will see in the results ahead, this gives our results a much different flavor than those of [Bar+20; CM22; DW18; Has+22] – in particular, the concentration of the bias and variance, for us, will be stated around empirical quantities involving the spectrum of the realized training data $\mathbf{X}$, while in these previous works, they are around nonrandom, global quantities, such as the covariance matrix $\mathbf{\Sigma}$ and its alignment with the signal vector $\boldsymbol{\beta}$.

The Stieltjes transform is a fundamental object in the study of the spectra of random matrices. While we will never need its full abilities in our analyses, we borrow some notation to simplify some results ahead.

Definition 3.2. The Stieltjes transform of a measure μ on $\mathbb{R}$ is a complex function

$m(\mu, z) : \mathbb{C} \to \mathbb{C}$ defined as

$$m(\mu, z) = \mathbb{E}_{X \sim \mu}\left[\frac{1}{X - z}\right] = \int_{\mathbb{R}} \frac{1}{x - z} \, \mathrm{d}\mu(x). \tag{3.4}$$

Generally, one may restrict the domain to the region outside the support of μ.

In our analysis, we will mostly care when μ is the empirical distribution of the entries of $\mathbf{D}^\top \mathbf{D}$. We define $\mu_{\mathbf{D}^\top \mathbf{D}} = \frac{1}{p} \sum_{i=1}^{n \wedge p} \delta_{D_{ii}^2} + \frac{\max(0, p-n)}{p} \delta_0$, and hence, with some abuse of notation, let

$$m_{\mathbf{D}}(z) = m(\mu_{\mathbf{D}^\top \mathbf{D}}, z) = \frac{1}{p} \sum_{i=1}^{n \wedge p} \frac{1}{D_{ii}^2 - z} - \frac{\max(0, p-n)}{p} \frac{1}{z}. \tag{3.5}$$

Occasionally it will also be useful to consider the *companion Stieltjes transform*, which is the Stieltjes transform of the empirical distribution of $\mathbf{D}\mathbf{D}^\top$. We will write

$$v_{\mathbf{D}}(z) = m(\mu_{\mathbf{D}\mathbf{D}^\top}, z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{D_{ii}^2 - z} - \frac{\max(0, n-p)}{n} \frac{1}{z}. \tag{3.6}$$

These two quantities are directly related. Direct calculation produces, amongst other identities,

$$v_{\mathbf{D}}(z) + \frac{1}{z} = \gamma \left(m_{\mathbf{D}}(z) + \frac{1}{z} \right) \tag{3.7}$$

$$z v_{\mathbf{D}}'(-z) - \frac{1}{z} = \gamma \left(z m_{\mathbf{D}}'(-z) - \frac{1}{z} \right) \tag{3.8}$$

$$v_{\mathbf{D}}(-z) - z v_{\mathbf{D}}'(-z) = \gamma \left(m_{\mathbf{D}}(-z) - z m_{\mathbf{D}}'(-z) \right). \tag{3.9}$$

3.5 Risk on an independent population

We begin with the setting where the risk is evaluated on an independent sample $\mathbf{X}_{\mathsf{new}} = \mathbf{Q}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}} \mathbf{O}_{\mathsf{new}}$. Here, the risk admits a familiar bias-variance decomposition.

Lemma 3.4.

$$R_{\mathbf{X}}\left(\hat{\boldsymbol{\beta}}(\mathbf{X},\mathbf{y}),\boldsymbol{\beta}\right) = \frac{n\mathbb{E}[\mathsf{Tr}(\mathbf{D}_{\mathsf{new}}^{\top}\mathbf{D}_{\mathsf{new}})]}{n'p} \cdot \left(\underbrace{\frac{1}{n}\|\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \boldsymbol{\beta}\|_2^2}_{B_{\mathbf{X}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})} + \underbrace{\frac{1}{n}\mathsf{Tr}(\mathsf{Cov}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}))}_{V_{\mathbf{X}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})}\right).$$

Proof. The proof is standard and deferred to Appendix A.2.2. □

We refer to $B_{\mathbf{X}}$ as the bias and $V_{\mathbf{X}}$ as the variance.

3.5.1 The Risk of Ridgeless Regression

The ridgeless estimator, or minimum norm estimator, takes the form $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\mathbf{X})^{+}\mathbf{X}^{\top}\mathbf{y}$, where $^{+}$ denotes the Moore-Penrose pseudoinverse. We can plug this in directly to Lemma 3.4 to compute the bias and variance.

Theorem 3.1. *When $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\mathbf{X})^{+}\mathbf{X}^{\top}\mathbf{y}$, then*

$$V_{\mathbf{X}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) = \frac{\sigma^2}{n}\mathsf{Tr}((\mathbf{D}^{\top}\mathbf{D})^{+}) = \frac{\sigma^2}{n}\sum_{i=1}^{n\wedge p}\frac{1}{D_{ii}^2} \tag{3.10}$$

If $n \geq p$, then

$$B_{\mathbf{X}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) = \mathsf{B}_n = 0,$$

and when $n < p$, then one has

$$\left|B_{\mathbf{X}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) - \mathsf{B}_n\right| = \frac{\|\boldsymbol{\beta}\|^2}{n}\delta_n \tag{3.11}$$

with

$$\mathbb{P}(|\delta_n| \geq t) \leq 2\exp\left(-Cp\min\left(\frac{p}{n}\cdot\frac{t^2}{2K^4}, \frac{t}{K^2}\right)\right). \tag{3.12}$$

where

$$\mathsf{B}_n = \frac{\|\boldsymbol{\beta}\|^2}{n}\left(1 - \frac{\min(n,p)}{p}\right). \tag{3.13}$$

Proof. We begin with the variance. Pseudoinverse properties yield

$$\begin{aligned} V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \frac{1}{n}\mathsf{Tr}(\mathsf{Cov}(\hat{\boldsymbol{\beta}} \mid \mathbf{X})) = \frac{1}{n}\mathsf{Tr}(\mathsf{Cov}((\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\boldsymbol{\epsilon} \mid \mathbf{X})) \\ &= \frac{1}{n}\mathsf{Tr}((\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\mathsf{Cov}(\boldsymbol{\epsilon})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^+) = \frac{\sigma^2}{n}\mathsf{Tr}\left((\mathbf{X}^\top\mathbf{X})^+\right) \\ &= \frac{\sigma^2}{n}\mathsf{Tr}((\mathbf{D}^\top\mathbf{D})^+) \end{aligned}$$

For the bias,

$$\begin{aligned} B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \frac{1}{n}\|\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \boldsymbol{\beta}\|_2^2 = \frac{1}{n}\boldsymbol{\beta}^\top(\mathbf{I} - (\mathbf{X}^\top\mathbf{X})^+(\mathbf{X}^\top\mathbf{X}))^2\boldsymbol{\beta} \\ &= \frac{1}{n}\boldsymbol{\beta}^\top(\mathbf{I} - (\mathbf{X}^\top\mathbf{X})^+(\mathbf{X}^\top\mathbf{X}))\boldsymbol{\beta} \\ &= \frac{1}{n}\|\boldsymbol{\beta}\|_2^2\left(1 - \frac{1}{\|\boldsymbol{\beta}\|_2^2}(\mathbf{O}\boldsymbol{\beta})^\top(\mathbf{D}^\top\mathbf{D})^+(\mathbf{D}^\top\mathbf{D})(\mathbf{O}\boldsymbol{\beta})\right). \end{aligned}$$

Note now that $\mathbf{b} = (\mathbf{O}\boldsymbol{\beta})/\|\boldsymbol{\beta}\|$ is uniformly distributed on the sphere. Furthermore, let $\mathbf{P} = (\mathbf{D}^\top\mathbf{D})^+(\mathbf{D}^\top\mathbf{D})$, and note that $P_{ii} = \mathbb{1}(i \le \min(n,p))$, so $\|\mathbf{P}\|_F^2 = \mathsf{Tr}(\mathbf{P}) = \min(n,p)$ and $\|\mathbf{P}\| = 1$. Note that if $p \le n$, then $\mathbf{P} = \mathbf{I}$, and this term is exactly zero.

If $p > n$, then we can apply Lemma 3.2 and 3.3 to find

$$\mathbb{P}(\left|\mathbf{b}^\top\mathbf{P}\mathbf{b} - n/p\right| \ge t) \le 2\exp\left(-C\min\left(\frac{p^2t^2}{2K^4n}, \frac{pt}{K^2}\right)\right).$$

□

The results in [Has+22] are stated in terms of rates. We opt for the presentation in (3.12) because this is more illustrative of the quantities governing how the constant factors governing how concentrated the bias is. We can provide rates in the following easy corollary of the above; these are generally stated in the context of proportional asymptotics; hence once should consider γ a fixed constant while $n, p \to \infty$ together.

Corollary 3.5. *If $n < p$, then for any positive integers k and D, for $n = \Omega_{k,D,\gamma}(1)$, with probability at least $1 - n^{-D}$, one has*

$$\left|B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - \mathsf{B}_n\right| \leq \frac{\|\boldsymbol{\beta}\|^2}{n} \cdot n^{-\frac{1}{2}+\frac{1}{k}} \tag{3.14}$$

Proof. The proof is direct calculation. Continuing with the same notation as above,

$$\mathbb{P}(\left|\mathbf{b}^\top \mathbf{P}\mathbf{b} - n/p\right| \geq t) = 2\exp\left(-C\min\left(\gamma^2\frac{nt^2}{2K^4}, \gamma\frac{nt}{K^2}\right)\right),$$

where $\gamma = p/n$. Hence taking $t = n^{-\frac{1}{2}+\frac{1}{k}}$, this becomes

$$\mathbb{P}(\left|\mathbf{b}^\top \mathbf{P}\mathbf{b} - n/p\right| \geq n^{-\frac{1}{2}+\frac{1}{k}}) \leq 2\exp\left(-C\min\left(\gamma^2\frac{n^{2/k}}{2K^4}, \gamma\frac{n^{1/2+1/k}}{K^2}\right)\right).$$

which is bounded above by n^{-D} whenever n is sufficiently large in terms of D, k, γ (hence $\Omega_{D,k,\gamma}(1)$), as desired. □

Remark 2. One of the main improvements of [CM22] over prior work was the derivation of multiplicative bounds, as compared to additive ones in [Has+22]. Manipulating (3.14) produces an entirely uninteresting multiplicative bound of the form

$$\left|B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - \mathsf{B}_n\right| \leq \frac{1}{1 - n/p} n^{-\frac{1}{2}+\frac{1}{k}} \cdot \mathsf{B}_n. \tag{3.15}$$

This reflects the same condition that p/n is bounded away from 1 that is present in the multiplicative bounds of [CM22]. As the bias vanishes (as it does when $\gamma = 1$), our ability to control relative fluctuations deteriorates.

Remark 3. The rate of $n^{-3/2}$ present in (3.14) is maybe at first surprising, but this should not be interpreted as an improvement over the classical $\mathcal{O}(n^{-1/2})$ rates present in the finite ridge regression results of [CM22; HTF01], which is expected to be tight in the i.i.d. setting through a central limit theorem heuristic. The additional factor of n^{-1} present in the statement is simply a consequence of our scaling, as discussed in Section 3.2.1. In short, to compare our results with pre-existing ones in [Bar+20; CM22; Has+22], one should think of $\mathbf{X}$ in our work as corresponding to the normalized matrix $\mathbf{X}/\sqrt{n}$ in prior works, and as a result, one should think of $\|\boldsymbol{\beta}\|$ as having order $\mathcal{O}(\sqrt{n})$, which recovers the original rate.

Additionally, [CM22; Has+22] both derive worse rates for the ridgeless setting than for finite ridge regression. For example [Has+22], which derives an $\mathcal{O}(n^{-1/7})$ rate on the bias for ridgeless regression, while maintaining the likely-tight $\mathcal{O}(n^{-1/2})$ rate for finite ridge regression. This is likely due to the fact that our quantities are in terms of the empirical spectrum of our observed design matrix (hence in some sense more data-dependent), which also dramatically simplifies the analysis; on the other hand, fully analytic forms are derived in these works.

Comparison to the isotropic Gaussian setting

The designs covered by both our theory and that of [CM22; DW18; Has+22] is that of the isotropic Gaussian. As expected, we hence produce the same results (see e.g. [Has+22, Theorem 1]) asymptotically. For our risk, written in Lemma 3.4, one takes $n' = 1$; since our design is scaled down by a factor of $\sqrt{n}$, we have

$$\mathbb{E}[\mathsf{Tr}(\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}})] = \mathsf{Tr}(\mathbb{E}[\mathsf{Tr}(\mathbf{X}_{\mathsf{new}}^\top \mathbf{X}_{\mathsf{new}})]) = \mathsf{Tr}(\mathbb{E}[\mathbf{I}_p/n]) = p/n,$$

and thus the multiplicative factor simplifies to 1, leaving only the bias and variance.

In such cases, the empirical spectral distribution of $\mathbf{D}^\top\mathbf{D}$ converges to the Marchenko-Pastur law. Provided γ is bounded away from 1, the support of the distribution, outside of the spike at zero, is bounded away from zero and compactly supported. Hence the function $x \mapsto 1/x$ is a bounded positive function, and thus Portmanteau implies

$$V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{\sigma^2}{n} \sum_{i=1}^{n \wedge p} \frac{1}{D_{ii}^2} \xrightarrow{a.s.} m(\mu_{\mathsf{MP}(\gamma)}, 0) = \begin{cases} \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ \frac{1}{\gamma-1} & \text{for } \gamma > 1. \end{cases}$$

One of the key improvements of [CM22] was the finding that the rate of concentration of the variance is in fact $\mathcal{O}(1/n)$ rather than $\mathcal{O}(1/\sqrt{n})$. This can be thought of as a consequence of the fact that linear spectral statistics of many classes of random matrices concentrate at the rate $\mathcal{O}(1/n)$ rather than the expected $\mathcal{O}(1/\sqrt{n})$, and this fact indeed gives us the $\mathcal{O}(1/n)$ convergence rate of the variance to its limit – see [BS04; LP09].

3.5.2 Finite Ridge Regularization

Computations for the finite ridge setting are identical to those in Theorem 3.1 and hence we state only the results here and defer the proof to Appendix A.3.1.

Theorem 3.2. *When* $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$, *then*

$$\begin{aligned} V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{\sigma^2}{n} \mathsf{Tr}\left((\mathbf{D}^\top \mathbf{D} + \lambda \mathbf{I})^{-2} \mathbf{D}^\top \mathbf{D}\right) &= \frac{\sigma^2}{n} \sum_{i=1}^{n \wedge p} \frac{D_{ii}^2}{(D_{ii}^2 + \lambda)^2} \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{n} \left[\frac{1}{D_{ii}^2 + \lambda} - \frac{\lambda}{(D_{ii}^2 + \lambda)^2}\right] = \sigma^2 (v_{\mathbf{D}}(-\lambda) - \lambda v_{\mathbf{D}}'(-\lambda)) \\ &= \sigma^2 \gamma (m_{\mathbf{D}}(-\lambda) - \lambda m_{\mathbf{D}}'(-\lambda)), \end{aligned} \tag{3.16}$$

and

$$\left| B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - \mathsf{B}_n \right| = \frac{\|\boldsymbol{\beta}\|^2}{n} \delta_n \text{ where } \mathsf{B}_{\lambda,n} = \lambda^2 m_{\mathbf{D}}'(-\lambda) = \frac{\lambda^2}{\gamma} v_{\mathbf{D}}'(-\lambda) + \frac{\gamma - 1}{\gamma}, \tag{3.17}$$

where

$$\mathbb{P}\left(|\delta_n| \geq t\right) = 2 \exp\left(-Cp \min\left(\frac{t^2}{2K^4 \frac{1}{p} \|\mathbf{P}\|_F^2}, \frac{t}{K^2}\right)\right) \tag{3.18}$$

and

$$\mathbf{P} = \operatorname{diag}\left(\frac{\lambda^2}{(D_{ii}^2 + \lambda)^2}\right)_{i=1}^{p}. \tag{3.19}$$

Note that $\frac{1}{p}\|\mathbf{P}\|_F^2 \leq 1$ always; hence the denominator never explodes, nor does it grow with n. One can derive similar rates as to Corollary 3.5 using the same method – these are stated in Appendix A.3.1.

Comparison to the isotropic Gaussian setting

When the design is taken to be isotropic Gaussian, the forms are again equivalent to the asymptotic forms derived in, for instance, [DW18; Has+22]. As in the ridgeless setting, one has $\mathcal{O}(1/n)$ rate of convergence for the variance towards its limit, and $\mathcal{O}(1/\sqrt{n})$ for the bias.

3.5.3 Empirical Takeaways - An attempt at real data

What does this theory tell us about when a design should be "right rotationally invariant" in practice? Recalling the properties required for results within this chapter, one must have that the features are generally uncorrelated. Furthermore, that $\mathbf{O}$ is Haar somewhat reflects that the design must be agnostic to any "direction" – in particular, this means that $\boldsymbol{\beta}$ must not align well with any of the eigenvectors of the Gram matrix $\mathbf{X}^\top \mathbf{X}$, since these are just the rows of $\mathbf{O}$ in the SVD $\mathbf{X} = \mathbf{Q}^\top \mathbf{D} \mathbf{O}$. This is somewhat similar to what was discussed in Chapter 1, where whether $\boldsymbol{\beta}$ aligns with the top eigenvectors of $\boldsymbol{\Sigma}$ affects the behavior of the model quite a bit. Methods for handling this in a different right-rotationally invariant setting have been presented in [LS23].

In this section, we discuss an experiment on real data. The issue with such experiments is that risk predictions require knowledge of the magnitude of the true signal and noise (though we will discuss doing estimation for these quantities in the well-specified setting in Chapter 4), and in reality, these quantities may also be varying over time. One setting that allows at least explicit computation of the signal is that of regressing an electronically traded fund (ETF) (specifically any one tracking the S&P500 index) onto its constituents. The signal here is publicly available, and moreover, due to the linear nature of PCA, a rescaled version (due to standardization) still holds for the residualized returns.

Due to data quality issues, we could not gather return data for all stocks within the index (only around 70%). The movement of these missing stocks then became the noise in our setting - we can estimate the variance σ^2 by subtracting the return computed on the known stocks from the return of the ETF.

Unfortunately, this experiment did not work well, as show on the right hand side of Figure 3.1. The i.i.d. predictions do equally well or even better than our right-rotationally invariant ones. When one tries to induce autocorrelation as discussed in Chapter 2, the noise itself becomes autocorrelated, and hence the behavior of the model is outside both regimes of theory, leading to curious behavior.

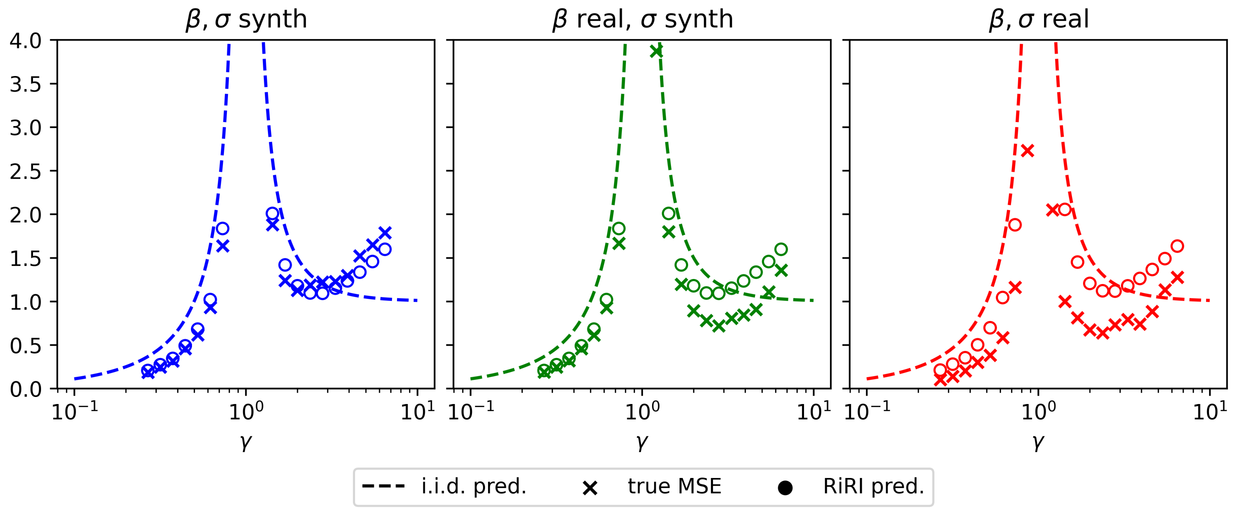


Figure 3.1: Fitting the S&P500. Predicted vs true MSEs. Left plot has synthetically generated signal and noise, where the signal has the same norm as the true signal and the noise has the same norm as true noise. Middle is signal fixed to the true vector, noise is resampled of the same magnitude. Last plot is true signal and true noise in data.

Instead of using the true signal vector, if one tries a semi-synthetic experiment, where the true vector is taken instead to be of the same norm, but drawn uniform on the sphere, and the noise is resampled with the same magnitude, our theory does work, and works better than the i.i.d. predictions. This is illustrated in the left plot of Figure 3.1.

Taking a step back towards reality, if one fixes the signal vector as the truth, but resamples the noise of the same magnitude, the performance begins to deterioriate, as illustrated in the center plot of Figure 3.1. This seems to suggest that at least part of the gap between prediction and theory is this alignment problem discussed earlier on, since if the vector is uniform on the sphere, its alignment with any given eigenvector of the covariance matrix is guaranteed to be quite small (in the literature, this is called being "incoherent"). Part of the reason this true vector has such good alignment is that most of the weight is actually concentrated on a small number of stocks, namely the largest ones, and moreover these largest ones dictate most of the market's movement, hence possibly leading to alignment with the top eigenvectors, even after residualization. As a result, the signal vector is not very "delocalized," compared to when it is drawn from the sphere. Another possible reason why the realized test MSE is so low is that the noise is probably not independent of the signal; the noise is composed of the movement of the stocks we do not observe; these are again correlated with global market movements, which our residualization scheme is not good enough to remove, and hence likely why the realized test losses are so much lower – notably, our right-rotationally invariant scheme does better capture the shape of the loss curve as a function of γ, even though the values are biased.

4
Generalized cross validation

In practice, one wishes to find principled ways in which to choose the optimal ridge regularization parameter λ for the given data distribution. As mentioned in Chapter 1, the standard k-fold cross-validation is often invalid in high dimensional settings [RM18; Wan+18], while it is possible for leave-one-out cross-validation (LOOCV) and the generalized cross-validation (GCV) to be consistent, as in the case of linear regression with i.i.d. rows [Has+22]. We begin by reviewing the form of these two terms; while the LOOCV is not tractable to analyze in our setting, we can show that the GCV is inconsistent, and in turn we propose a provably consistent alternative.

4.1 Preliminaries

Leave-one-out cross-validation is a method for estimating the out-of-sample performance of any estimator – we will focus on the setting where this estimator is $\hat{\boldsymbol{\beta}}_\lambda$, the ridge estimator with regularization parameter λ. Here, the LOOCV error of $\hat{\boldsymbol{\beta}}_\lambda$ is given by

$$\mathsf{LOOCV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda^{-i}\right)^2 \tag{4.1}$$

where $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ denotes the ridge solution where the i-th datapoint is omitted. Note that this visibly appears to be an estimate of $\mathbb{E}(y_{\mathsf{new}} - \mathbf{x}_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}}_\lambda)^2$. While in practice, refitting $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ for every $i \in [n]$ is expensive, within ridge regression, there exists a well-known shortcut formula allowing practicioner to only fit one estimator. Specifically, using the Sherman-Woodbury-Morrison formula, one can show

$$\mathsf{LOOCV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda}{1 - (\mathbf{S}_\lambda)_{ii}}\right)^2 \tag{4.2}$$

where $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top$ is the smoother matrix.

Generalized cross-validation was proposed in [GHW79] as a rotationally invariant alternative to LOOCV.

$$\mathsf{GCV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda}{1 - \mathsf{Tr}(\mathbf{S}_\lambda)/n}\right)^2 = \frac{1}{n - \mathsf{Tr}(\mathbf{S}_\lambda)}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda\right)^2 \tag{4.3}$$

Comparing to (4.2), we have replaced $(\mathbf{S}_\lambda)_{ii}$ with the average $\mathsf{Tr}(\mathbf{S}_\lambda)/n$; note furthermore that because we now take the trace the GCV is invariant to rotations of the data. That is, consider fitting a linear model on $(\mathbf{XR}, \mathbf{y})$ instead of $(\mathbf{X}, \mathbf{y})$, for any $\mathbf{R} \in \mathbb{O}(p)$. Then the ridge estimator becomes $\mathbf{R}^{-1}\hat{\boldsymbol{\beta}}_\lambda$ and thus the second term of the rightmost side of Equation (4.3) does not change, and furthermore, $\mathbf{S}_\lambda$ does not change; hence $\mathsf{GCV}_n(\lambda)$ remains the same.

In [Has+22], it is shown for the designs considered that both LOOCV and GCV are asymptotically (under proportional asymptotics) equal to the out-of-sample error $\mathbb{E}[(y_{\mathsf{new}} - x_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}}_\lambda)^2]$. In our case, LOOCV is not tractable to analyze without additional assumptions on the distribution of $\mathbf{Q}$ – see Appendix A.4.1 for details. We instead focus on GCV.

4.2 Usual GCV with right rotationally invariant design

In this section and those that follow, we will consider the asymptotic setting as $n, p \to \infty$ with $p/n \to \gamma$. We assume that the scaling factor $n\mathbb{E}[\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}}]/p$ which appears as a multiplicative factor in the risk formula 3.4 is of constant order in the limit. This aligns with our scaling on $\mathbf{X}$: our covariates have been scaled down by a factor of $\sqrt{n}$, but we now have n_1 samples; hence $\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}}$ has entries of order n'/n; thus the trace has order $n'p/n$, which exactly cancels. In general, when one's goal is to tune λ to optimize for the out-of-sample risk, this scaling factor is irrelevant, since it is multiplied to the entire risk – we simply require that it does not degenerate or explode in the limit.

Theorem 4.1. *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \ldots$ be a sequence of right rotationally invariant designs where each $\mathbf{X}_i \sim \mathbf{Q}_i^\top \mathbf{D}_i \mathbf{O}_i$ and the following conditions hold:*

1. $\limsup \lambda_{\max}(\mathbf{D}) < C$ *almost surely for some constant C.*

2. *each $\mathbf{X}_n$ has dimensions $n \times p(n)$, with $p(n)/n \xrightarrow{n\to\infty} \gamma \in [0, \infty)$.*

Furthermore, let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n, \ldots$ be a sequence of signal vectors such that for all n, $\beta_n \in \mathbb{R}^p$, and $\|\boldsymbol{\beta}_n\|/\sqrt{n} = r$. Lastly, let $\mathbf{y}_i$ be generated as according to (3.1).

Then if each ϵ_i has finite moment of order $4+\eta$, for some $\eta > 0$,

$$\mathsf{GCV}_n(\lambda) - \frac{r^2(v_{\mathbf{D}}(-\lambda) - \lambda v_{\mathbf{D}}'(-\lambda)) + \sigma^2 \gamma v_{\mathbf{D}}'(-\lambda)}{\gamma v_{\mathbf{D}}(-\lambda)^2} \xrightarrow{a.s.} 0. \tag{4.4}$$

Proof. We will give a sketch of the proof and defer details to Appendix A.4.2.

We analyze the numerator and denominator separately. The latter is simple:

$$(1 - \mathsf{Tr}(\mathbf{S}_\lambda)/n)^2 = \left(1 - \mathsf{Tr}((\mathbf{D}^\top \mathbf{D} + \lambda \mathbf{I})^{-1} \mathbf{D}^\top \mathbf{D})/n\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n \frac{\lambda}{\lambda + D_{ii}^2}\right)^2 = \gamma^2 \lambda^2 v_{\mathbf{D}}(-\lambda)^2$$

The numerator only requires slightly more work. Recall $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Furthermore, note that the GCV numerator can be rewritten as below:

$$\frac{1}{n}\mathbf{y}^\top(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{y} = \underbrace{\frac{1}{n}\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{X}\boldsymbol{\beta}}_{T_1} + 2\underbrace{\frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{X}\boldsymbol{\beta}}_{T_2} + \underbrace{\frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{I}-\mathbf{S}_\lambda)^2\boldsymbol{\epsilon}}_{T_3}.$$

We handle each term. First, T_1 simplifies easily into a quadratic form which allows us to apply Hanson-Wright (Lemma 3.2):

$$T_1 = \frac{1}{n}(\mathbf{O}\boldsymbol{\beta})^\top\mathbf{D}^\top(\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D}(\mathbf{O}\boldsymbol{\beta})$$

This again concentrates around its expectation, which is

$$\begin{aligned} &= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\mathsf{Tr}(\mathbf{D}^\top(\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D})}{p} \\ &= \frac{\|\boldsymbol{\beta}\|^2}{n}\lambda^2\left(m_\mathbf{D}(-\lambda)-\lambda m'_\mathbf{D}(-\lambda)\right) = \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\lambda^2}{\gamma}\left(v_\mathbf{D}(-\lambda)-\lambda v'_\mathbf{D}(-\lambda)\right) \end{aligned}$$

and the concentration is enough so that we can just apply Borel-Cantelli to obtain almost sure convergence of their difference. The second term has expectation zero, since $\boldsymbol{\epsilon}$ is independently random; we again can control this term using some concentration arguments, and obtain almost sure convergence. For the third term, we use Lemma C.3 of [DW18, SI], which just states that this term again converges almost surely to its expectation (in the sense that the difference converges almost surely to 0), which is

$$\mathbb{E}[T_3] = \frac{\sigma^2}{n}\mathsf{Tr}((\mathbf{I}-\mathbf{S}_\lambda)^2) = \frac{\sigma^2}{n}\mathsf{Tr}\left((\mathbf{I}-\mathbf{D}(\mathbf{D}^\top\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\right) = \sigma^2\lambda^2 v'_\mathbf{D}(-\lambda);$$

this is what requires the moment condition on $\boldsymbol{\epsilon}$.

Plugging everything in produces

$$\mathsf{GCV}_n(\lambda) - \frac{r^2(v_\mathbf{D}(-\lambda)-\lambda v'_\mathbf{D}(-\lambda)) + \sigma^2\gamma v'_\mathbf{D}(-\lambda)}{\gamma v_\mathbf{D}(-\lambda)^2} \xrightarrow{a.s.} 0.$$

□

Remark 4. When compared to the statement of [Has+22, Theorem 7], we no longer require $\boldsymbol{\beta}$ to be random, as for right-rotationally invariant designs, having fixed $\boldsymbol{\beta}$ is equivalent to having $\boldsymbol{\beta}$ drawn independently from $\mathbf{X}$. The differing moment conditions are due to an error in the original paper. The $4+\eta$ moment condition was also required for $\boldsymbol{\epsilon}$; we in general do not require it for $\mathbf{x}_i$, hence allowing for our theory to handle fat-tailed distributions such as the t distribution, as seen earlier; instead, we require that $\mathbf{D}$ has bounded operator norm in the limit.

Remark 5. In fact, one can generalize to having $\|\boldsymbol{\beta}_n\|/\sqrt{n} \to r$, or having a sequence of noise levels $\sigma_n^2 \to \sigma^2$ instead without much difficulty. Furthermore, if we further assume that ϵ_i has entries with uniformly bounded sub-Gaussian norm, then the asymptotic setup above is unnecessary, and in fact one can derive finite sample control by using Hanson-Wright to obtain finite sample concentration of T_3. This analysis is not pursued here, because, as we will discuss, the GCV in this setting is *not* asymptotically consistent for the prediction error.

Expressed in terms of Stieltjes transforms, the above formula for the GCV is equal to the same form as in [Has+22]. The main problem is that the correctness of that form (in terms of being consistent for out-of-sample risk) relies on two miraculous differential equalities satisfied by the Marchenko-Pastur law. There, it is proven that the GCV is asymptotically consistent in the case of isotropic features and i.i.d. rows, where the authors explicitly find

$$\mathsf{GCV}_n(\lambda) - \mathbb{E}[(y_{\mathsf{new}} - x_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2] \xrightarrow{a.s.} 0.$$

In that setting, one is able to write $\mathbb{E}[(y_{\mathsf{new}} - x_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2] = \mathbb{E}[R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})] + \sigma^2$ and then apply results from Theorem 3.2 to write down the explicit form of this expression. Though this is glossed over in the original work [Has+22], these two expressions are not equal in general – in particular, if one compares the result of Theorem 4.1 with that of Theorem 3.2, it is clear that for the two to be equal, one must have the following two identities for the companion Stieltjes transform:

$$v(-\lambda) - \lambda v'(-\lambda) + 1 = \frac{v'(-\lambda)}{v(-\lambda)^2} \qquad \frac{1}{\gamma}\lambda^2 v'(-\lambda) + \frac{\gamma - 1}{\gamma} = \frac{v(-\lambda) - \lambda v'(-\lambda)}{\gamma v(-\lambda)^2}.$$

Both of these identities are consequences of the Silverstein equation [Sil95] for the companion Stieltjes transform, which in this case specializes to

$$\frac{1}{v_{\mathbf{D}}(-\lambda)} = \lambda + \frac{\gamma}{1 + v_{\mathbf{D}}(-\lambda)}.$$

One can derive an equivalent self-consistent equation using the leave-one-out method[1].

4.3 A Modified GCV

Such beautiful identities do not exist for us, and thus the GCV is naturally biased as soon as the spectrum of $\mathbf{X}^\top \mathbf{X}$ departs from being Marchenko-Pastur. We give a possible alternative which is provably consistent. To motivate this method, we begin by re-examining two quantities. First, the GCV numerator takes the form

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda)^2 = \frac{1}{n} \mathbf{y}^\top (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{y} \tag{4.5}$$

and in the course of proving its limit, we show

$$\frac{1}{n} \mathbf{y}^\top (\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{y} - \left[r^2 \frac{\lambda^2}{\gamma} \left(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) \right) + \sigma^2 \lambda^2 v'_{\mathbf{D}}(-\lambda) \right] \xrightarrow{a.s.} 0. \tag{4.6}$$

The risk on an independent test set, as derived in Theorem 3.2, satisfies[2]

$$R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - \left[r^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathbf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + \sigma^2 \left(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) \right) \right] \xrightarrow{a.s.} 0. \tag{4.7}$$

The root purpose of GCV is to accurately tune the parameter λ so that it performs as best possible on the test sample. Hence, it is sufficient to find a way to transform the expression in (4.6) into a monotonic function of (4.7). This is exactly what occurs in the i.i.d. setting,

[1] The key identity necessary is $\frac{1}{v(z)} = -z(v(z) + 1) + (\gamma - 1) = -z + \frac{\gamma}{1+v(z)}$, which can be derived by manipulating the defining quadratic of the Stieltjes transform.

[2] While almost sure convergence was not shown explicitly, for sequences of $n, p, \mathbf{D}$ as in the context of Theorem 4.1, Borel-Cantelli applied to our bounds gives almost sure convergence.

where the correction factor in the denominator changes the expression in (4.6) into the one in (4.7) plus the additive noise factor σ^2. We want to try to do something similar.

We view the GCV and the ridge limit as having two components, one the coefficient of r^2 and the other the coefficient of σ^2. In our original analysis, we sought a new denominator $\mathfrak{d}(\lambda)$ such that a result of the sort

$$R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - \frac{1}{\mathfrak{d}(\lambda)}\left[\frac{1}{n}\mathbf{y}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{y}\right] \xrightarrow{a.s.} ar^2 + b\sigma^2$$

where a and b are free of λ. A result of this sort would enable us to thereby tune for the optimal value of λ using this modified GCV objective, since the additive factors of a and b do not affect which value of λ is optimal. Unfortunately, this is rather difficult for us to do in general, since $\mathbf{D}$ can be rather arbitrary.

However, if one could estimate σ^2 or r^2, then producing a consistent estimator for the out-of-sample error is not so difficult. In particular, if provided an estimator for one of the quantities, one can just use (4.6) as an equality to estimate the second, and then compute the risk using Theorem 3.2 (see also Appendix A.4.2 for an alternative motivation). A consistent estimator for σ^2 is in fact known for right-rotationally invariant designs, and is given in [LS23]:

$$\hat{\sigma}^2(\mathbf{X}, \mathbf{y}, \lambda) = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2 - \|(\mathbf{I}_n + \lambda^{-1}\mathbf{X}\mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)\|^2 \sum_{i=1}^p \frac{\lambda^2 D_{ii}^2}{(D_{ii}^2+\lambda)^2}(\sum_{i=1}^p D_{ii}^2)^{-1}}{\sum_{i=1}^p \frac{\lambda^2}{(\lambda+D_{ii}^2)^2} - n\left(\sum_{i=1}^p D_{ii}^2\right)^{-1}\sum_{i=1}^p \frac{D_{ii}^2}{(\lambda+D_{ii}^2)^2} + n - p}. \tag{4.8}$$

Most results in the paper are stated under the assumptions in 4.1 as well as the following additional assumption:

(A*) Almost surely, the diagonal entries of $\mathbf{D}$ converge in the Wasserstein-2 (W_2) distance to a measure D, where D^2 has non-zero mean and compact support $\text{supp}(\mathsf{D}^2) \subseteq [0, \infty)$. See Appendix A.6.2 for details on Wasserstein-2 convergence.

Remark 6. If such an assumption is made, one can, instead of stating results in terms of the difference converging almost surely to zero, as in (4.6, 4.7), instead write them as converging almost surely to the Stieltjes transform taken over the limiting spectral measure D^2.

Note that any possible value of λ produces a consistent estimator. Our proposed modification then proceeds as follows.

1. Compute $\hat{\sigma}^2(\mathbf{X}, \mathbf{y}, \lambda)$ as above.
2. Compute an estimator for $\hat{r}^2$ by using Equation 4.6 as an estimating equation, i.e. set
$$\hat{r}^2(\hat{\sigma}^2, \lambda) = \frac{\frac{1}{n}\mathbf{y}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{y} - \lambda^2 v'_{\mathbf{D}}(-\lambda)\hat{\sigma}^2}{\frac{\lambda^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)\right)}. \tag{4.9}$$
3. Finally, one just plugs in estimators for $\hat{r}^2, \hat{\sigma}^2$ to the result of Theorem 3.2 and produces
$$\hat{r}^2\left(\frac{\lambda^2}{\gamma}v'_{\mathbf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma}\right) + \hat{\sigma}^2(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) \tag{4.10}$$
as the estimate for the out-of-sample risk.

Lemma 4.1. *Under the assumptions of Theorem 4.1, the estimator* $\hat{r}^2$ *proposed above is consistent for any value of* $\lambda > 0$ *and any consistent estimator* $\hat{\sigma}^2$ *provided there exists some constant* c *such that* $v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) > c$ *with probability approaching* 1*. If instead* $\hat{\sigma}^2$ *is strongly consistent and* $\liminf_{n\to\infty} v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) > 0$ *almost surely, then* $\hat{r}^2$ *is also strongly consistent.*

The proof of the above statement is immediate from (4.6), (4.9), and the consistency of $\hat{\sigma}^2$. We instead discuss the assumptions stated, which is essentially that the denominator of (4.9) is bounded away from zero. One has

$$v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) = \frac{1}{n}\sum_{i=1}^n \frac{D_{ii}^2}{(D_{ii}^2 + \lambda)^2} \geq (D_{11}^2 + \lambda)^{-2}\left(\frac{1}{n}\sum_{i=1}^n D_{ii}^2\right) \tag{4.11}$$

Hence the assumption on bounded operator norm of Theorem 4.1, together with the assumption that the limiting spectrum has nonzero mean (Assumption (A*)), is sufficient for this to hold.

Lemma 4.2. *Under the assumptions of Theorem 4.1, for any consistent estimators* $(\hat{r}^2, \hat{\sigma}^2)$ *of* (r^2, σ^2)*, respectively, one has, for any* $\lambda_0 > 0$*,*

$$\sup_{\lambda > \lambda_0} \left| (\hat{r}^2 - r^2) \left(\frac{\lambda^2}{\gamma} v'_{\mathbf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + (\hat{\sigma}^2 - \sigma^2)(v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) \right| \xrightarrow{p} 0. \tag{4.12}$$

If $\hat{r}^2, \hat{\sigma}^2$ *are strongly consistent, then naturally the convergence is almost sure.*

Proof. This is immediate from the fact that

$$\lambda^2 v'_{\mathbf{D}}(-\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{\lambda^2}{(D_{ii}^2 + \lambda)^2} \leq 1$$
$$v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda) = \frac{1}{n} \sum_{i=1}^{n \wedge p} \frac{D_{ii}^2}{(D_{ii}^2 + \lambda)^2} \leq \frac{1}{4\lambda}.$$

□

Remark 7. In the uniform convergence statement of [Has+22], they require compact intervals bounded away from zero. Here, we do not have this issue because things are stated in terms of the finite sample empirical spectrum. If we instead adopt Assumption A* and try to show convergence of this modified GCV to the limiting spectral object

$$r^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathsf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + \sigma^2 \left(v_{\mathsf{D}}(-\lambda) - \lambda v'_{\mathsf{D}}(-\lambda) \right),$$

where $v_{\mathsf{D}}(z)$ refers to $s(\mathsf{D}^2, z)$, the Stieltjes transform taken over the limiting distribution D^2, then one must do the same.

Lemma 4.3. *Under the assumptions of Theorem 4.1 and Assumption (A*), then one has,* *for any λ_1, λ_2 satisfying $0 < \lambda_1 < \lambda_2$ and any two consistent estimators $\hat{r}^2, \hat{\sigma}^2$,*

$$\sup_{\lambda \in [\lambda_1, \lambda_2]} \left| \left(\hat{r}^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathbf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + \hat{\sigma}^2 (v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) \right) - \right.$$
$$\left. \underbrace{\left(r^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathsf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + \sigma^2 \left(v_{\mathsf{D}}(-\lambda) - \lambda v'_{\mathsf{D}}(-\lambda) \right) \right)}_{R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})} \right| \xrightarrow{p} 0.$$

As usual, if the two estimators are instead strongly consistent, then the above convergence is almost sure.

Proof. Applying the triangle inequality and Lemma 4.2 allows us to reduce to showing

$$\sup_{\lambda \in [\lambda_1, \lambda_2]} \left| \left(r^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathbf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + \sigma^2 (v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) \right) - \right.$$
$$\left. \left(r^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathsf{D}}(-\lambda) + \frac{\gamma - 1}{\gamma} \right) + \sigma^2 \left(v_{\mathsf{D}}(-\lambda) - \lambda v'_{\mathsf{D}}(-\lambda) \right) \right) \right| \xrightarrow{p} 0.$$

This is just

$$\sup_{\lambda \in [\lambda_1, \lambda_2]} \left| r^2 \left(\frac{\lambda^2}{\gamma} v'_{\mathbf{D}}(-\lambda) - \frac{\lambda^2}{\gamma} v'_{\mathsf{D}}(-\lambda) \right) + \right.$$
$$\left. \sigma^2 \left((v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) - (v_{\mathsf{D}}(-\lambda) - \lambda v'_{\mathsf{D}}(-\lambda)) \right) \right| \xrightarrow{p} 0. \quad (4.13)$$

Hence it suffices to prove

$$\begin{aligned} \sup_{\lambda \in [\lambda_1, \lambda_2]} \left| \lambda^2 (v'_{\mathbf{D}}(-\lambda) - v'_{\mathsf{D}}(-\lambda)) \right| &\xrightarrow{p} 0 \\ \sup_{\lambda \in [\lambda_1, \lambda_2]} \left| (v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)) - (v_{\mathsf{D}}(-\lambda) - \lambda v'_{\mathsf{D}}(-\lambda)) \right| &\xrightarrow{p} 0. \end{aligned}$$

We proceed by a standard discretizing argument. First, we note that

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\lambda^2 v_{\mathbf{D}}(-\lambda)\right]\right| &= \left|2\lambda v'_{\mathbf{D}}(-\lambda) - \lambda^2 v''_{\mathbf{D}}(-\lambda)\right| \\ &= \left|\frac{2}{\lambda}\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda^2}{(D_{ii}^2+\lambda)^2} - \frac{1}{\lambda}\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda^3}{(D_{ii}^2+\lambda)^3}\right| \le \frac{2}{\lambda} \le \frac{2}{\lambda_1},\end{aligned}$$

and the same holds for v_{D}. Thus all functions in the sequence as well as the limit itself are Lipschitz on the interval $[\lambda_1, \lambda_2]$. Hence one can discretize the interval $[\lambda_1, \lambda_2]$ into a sufficiently fine finite grid; pointwise convergence (from convergence in Wasserstein-2 distance – note that $v(-\lambda)$ is uniformly bounded for $\lambda > \lambda_1$) holds along every point on the grid, and one controls the difference away from the points on the grid using the Lipschitz property and the triangle inequality, yielding uniform convergence.

Similarly, we can likewise bound the derivative of $v_{\mathbf{D}}(-\lambda) - \lambda v'_{\mathbf{D}}(-\lambda)$, and similar arguments apply. The modifications needed if the estimators are strongly convergent are then clear. □

4.3.1 Alternative estimators for r^2, σ^2

While the above procedure was provably consistent, in finite sample settings, having excessive error in estimating either of r^2 and σ^2 can cause the modified GCV procedure to give poor tuning results. We present an alternative scheme for estimating the two quantities which we empirically observe has better finite sample performance (around 20% lower standard deviation for both r^2 and σ^2 at $n = p = 1000$).

We first describe the motivation[3] for the following procedure. Observe that equation (4.6) can be used as an estimating equation, as the first term is estimable – moreover, it is an estimating equation for *every* value of λ. Hence, if one computes it over a range of λ, then one can solve for values of σ^2 and r^2, since the coefficients of the two terms are in terms of the Stieltjes transform of the empirical measure, which can directly be computed. Furthermore, one generically wishes to avoid errors in estimating σ^2 from

[3] The root motivation was that I implemented the previous scheme incorrectly and thought it was too loose.

propagating into those in estimating r^2. Hence by evaluating this in at least three points, one can produce two estimating equations for r^2 and two for σ^2, each of which does not require an estimate of the other quantity. Explicitly, the scheme proceeds as follows.

1. One first specifies a list of regularization strengths $\lambda_1, \ldots, \lambda_L$. Empirically, we observe that taking them to be logarithmically spaced between 1 and $10^{2.5}$ works well, with $L = 6$.

2. For each λ_ℓ, one computes the GCV numerator

$$\mathsf{GCV}^{\mathsf{num}}(\lambda) = \frac{1}{n}\mathbf{y}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{y}$$

   as well as the coefficients for r^2 and σ^2, which are

$$a_\ell = \frac{\lambda_\ell^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda_\ell) - \lambda_\ell v'_{\mathbf{D}}(-\lambda_\ell)\right) \tag{4.14}$$

$$b_\ell = \lambda_\ell^2 v'_{\mathbf{D}}(-\lambda_\ell), \tag{4.15}$$

   respectively. This produces a system of L equations

$$\mathsf{GCV}^{\mathsf{num}}(\lambda_\ell) = a_\ell r^2 + b_\ell \sigma^2.$$

3. To estimate $\hat{r}^2$, one now eliminates σ^2 from all equations, producing $L - 1$ constraints for r^2. One then fits $\hat{r}^2$ using least-squares (without an intercept). The case for $\hat{\sigma}^2$ is analogous.

4.3.2 Performance of Modified GCV

We now compare this modified GCV to the original GCV in the same synthetic and semi-synthetic settings examined in Chapter 2. For the synthetic settings, our benchmark will be the true theoretical risk, while for the semi-synthetic settings, our benchmark will be the average of certain MSE curves taken over a test dataset. Our modified GCV estimates r^2 and σ^2 using the methods described earlier – a useful benchmark to compare this against is an oracle version that is given direct access to r^2 and σ^2 – we refer to this as Oracle GCV.

Fully synthetic settings

The performance of the modified GCV (labeled New GCV in the figures), as well as the original GCV formulation, are illustrated in Figure 4.1. From 4.1a, one can see that even for Gaussian data, our method does similarly well, whereas for Figures 4.1b, 4.1c, 4.1d, even though the performance of the tuned estimators are numerically similar (see numbers below title for tuned estimator risk), it is clear that the modified GCV is actually accurately estimating the true out-of-sample risk, while the original GCV is just managing to produce a tuned value of λ which is not far from optimal. Note furthermore that over the range of λ, the i.i.d. prediction is biased, especially at small λ, while the right-rotationally invariant risk prediction is less so. This again suggests that it provides better finite sample variance quantification.

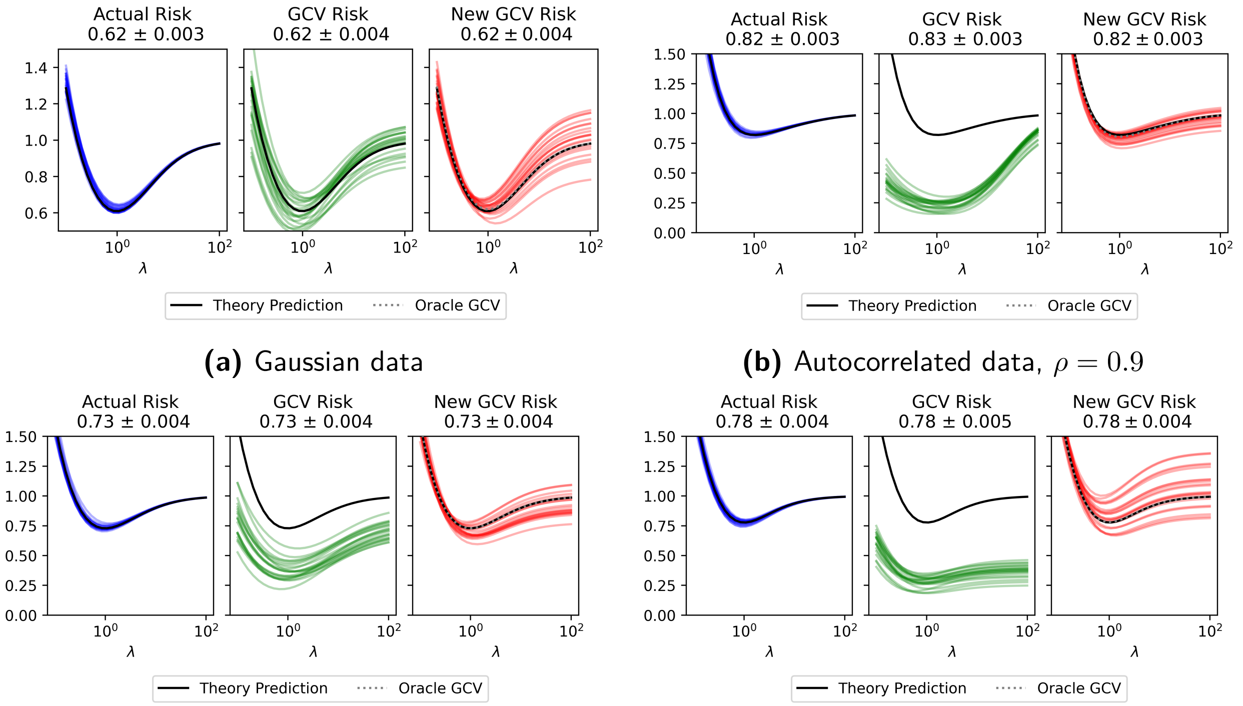
Semi-synthetic settings

We repeat the analysis above, now in the semi-synthetic setting. The modified GCV essentially does as well as the oracle version can – when the oracle version is biased, so too is the modified GCV. Note that the modified GCV is generally good at maintaining the shape of the loss curve, even when it is subject to some bias, such as in 4.2d. Note, however, in all of these cases, the modified GCV still does not outperform the original GCV, even though it more accurately estimates the loss curve itself.

4.3.3 Potential Improvements

A problem that emerges in this scheme for estimating r^2 and σ^2, as well as in the original proposal, is that the resulting estimates for r^2 and σ^2 are negatively correlated. This is undesirable because the value of the optimal choice of λ depends directly on their ratio – hence being negatively correlated distorts this ratio even more than usual.

A possible solution may be to exploit a second set of estimating equations. Specifically, one can in fact prove that the *norm* of the estimator $\hat{\boldsymbol{\beta}}_\lambda$ converges to an expression also involving r^2, σ^2, and some terms involving the companion Stieltjes transform. Hence this allows a second avenue for estimating the two quantities, and may help with this anti-correlation issue.



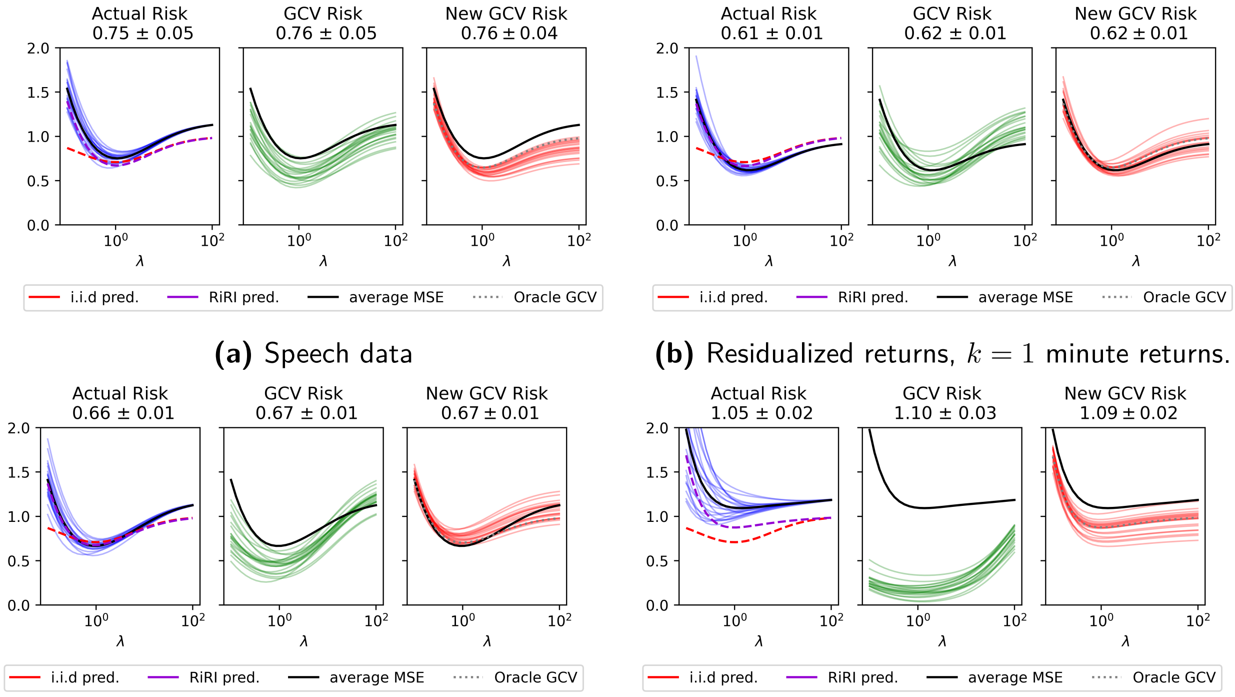
(a) Gaussian data

(b) Autocorrelated data, $\rho = 0.9$

(c) t-distributed data

(d) Equicorrelated data, $\rho = 0.8$

Figure 4.1: Performance of GCV in various synthetic settings. All simulations have $n = p = 1000$ and $r^2 = \sigma^2 = 1$. The x-axis is λ, the regularization parameter. Below each title is the out-of-sample risk obtained when the the given GCV method is used, or, in the case of actual risk, the true minimum, along with its standard error. The colored lines are one of 20 iterations. In each iteration, we compute the GCV metric over a range of λ to produce the line, which reflects the estimated out-of-sample risk. Theoretical risk is plotted in all three as reference, and refers to the exact value of $R_{\mathbf{X}}$ – note that the actual MSE curves differ from this slightly because they are additionally conditional on the instance of $\boldsymbol{\epsilon}$.



(a) Speech data

(b) Residualized returns, $k = 1$ minute returns.

(c) Residualized returns, $k = 3$ minute returns.

(d) Residualized returns, $k = 30$ minute returns.

Figure 4.2: Performance of GCV in various semi-synthetic settings. We again take $r^2 = \sigma^2 = 1$. As in Figure 4.1, below each title is the out-of-sample risk obtained when the the given GCV method is used, or, in the case of actual risk, the true minimum. The colored lines are one of 20 iterations. In each iteration, we compute the GCV metric over a range of λ to produce the line, which reflects the estimated out-of-sample risk. In each plot, the average of the MSE curves over the iterations is plotted as the reference true risk, as opposed to the theoretical risk before. For speech data, $n = p = 400$; for residualized returns, $n = p = 493$. $\gamma = 1$ was chosen for illustrative purposes.

5
Towards the nonlinear model

The original aim of this thesis was to analyze the random features model where the weight matrix was taken to be right rotationally invariant. Unfortunately, some technical barriers prevented us from pushing this through. We give an overview of the methods used to prove this result for the i.i.d. Gaussian setting and where difficulties arise in the right-rotationally invariant setting. We then state some conjectures under which the equivalence holds, include numerical simulations supporting their validity, and outline the remainder of the proof.

5.1 Preliminaries

In the random features model we consider, one is given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where now $\mathbf{x}_i \in \mathbb{R}^d$ (note the change from p to d) generated by some ground truth linear function defined by $\boldsymbol{\xi}$, i.e. $\mathbf{y} = \mathbf{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}$. More generic structures are possible, such as having a nonlinear ground truth as in [MM19], or a generic function applied component-wise to the inner products $\mathbf{X}\boldsymbol{\xi}$ [HL20], but we pursue only this for simplicity. As described at the end of Chapter 1, the random features model is essentially a one-hidden-layer neural network with hidden layer randomly sampled and then fixed, and final layer weights learned. Explicitly, one samples a *random* weight matrix $\mathbf{F} \in \mathbb{R}^{d\times p}$, chooses an activation function σ, and computes the feature map taking $\mathbf{X} \mapsto \sigma(\mathbf{XF}) \in \mathbb{R}^{n\times d}$. One now performs ridge regression of $\mathbf{y}$ onto this feature matrix $\sigma(\mathbf{XF})$, giving final layer weights $\mathbf{w}$ satisfying

$$\mathbf{w} = \arg\min_{\mathbf{w}\in\mathbb{R}} \left\{ \sum_{i=1}^{n} \left(y_i - \frac{1}{\sqrt{p}} \mathbf{w}^\top \sigma(\mathbf{F}^\top \mathbf{x}_i) \right)^2 + \lambda \|\mathbf{w}\|^2 \right\}. \tag{5.1}$$

The resulting estimator is then $\hat{f} : \mathbf{x} \mapsto \mathbf{w}^\top \sigma(\mathbf{F}^\top \mathbf{x})$. We will refer to $\sigma(\mathbf{XF})$ as the *nonlinear features*. Since only the weights of $\mathbf{w}$ are learned, there are a total of p learned parameters in this model.

As in Chapter 3, we wish to understand the out-of-sample risk of this estimator. This problem is analyzed under proportional asymptotics – we take, $n, p, d \to \infty$ together, where $n/d \to \alpha > 0$ and $p/d \to \eta > 0$. The inputs $\mathbf{x}_i$ are drawn from $\mathsf{N}(0, \mathbf{I}_d)$, so the inputs are no longer dependent, but we take the feature matrix $\mathbf{F}$ to be right-rotationally invariant. In contrast, existing literature focuses largely on the setting where the feature matrix is to contain suitably normalized i.i.d. Gaussian – hence showing this result would be a strict improvement over existing theory. We explicitly define the train and test loss which we wish to characterize:

$$\begin{aligned} \mathcal{E}_{\text{train}} &= \frac{1}{p} \left\{ \sum_{i=1}^{n} (\tfrac{1}{\sqrt{p}} \boldsymbol{w}^\top \sigma(\mathbf{F}^\top \mathbf{x}) - y_t)^2 + \sum_{j=1}^{p} h(w_j) \right\} \\ \mathcal{E}_{\text{test}} &= \mathbb{E}\left[\left(y_{\mathsf{new}} - \frac{1}{\sqrt{p}} (\boldsymbol{w}^\top \sigma(\mathbf{F}^\top \mathbf{x}_{\mathsf{new}}) \right)^2 \right]. \end{aligned}$$

5.2 An Overview of Methods and Barriers

5.2.1 Method of [HL20]

[HL20] presents a pathway for proving that a sequence of sufficiently regular random matrices satisfies a Gaussian Equivalence property, namely that, asymptotically, the nonlinear model above in fact possesses the same training and test loss as a Gaussian model with matching first and second moments. Though they only explicitly show their method works for the case of i.i.d. Gaussian weights, their proof is largely deterministic – only a tiny portion uses that $\mathbf{F}$ is random. It in fact applies more generally to various loss functions ℓ and regularizations h – we will discuss only the case were $\ell(x, y) = (x - y)^2$ and $h(w) = \lambda w^2$, so that $\sum_{i=1}^{p} h(w_i) = \lambda \|\mathbf{w}\|$. Additionally, we assume the following regularity conditions:

1. Regularity on the activation: σ is bounded, odd, and admits three derivatives
2. $\boldsymbol{\xi}$ is deterministic, with $\|\boldsymbol{\xi}\| = 1$.

One first defines

$$\mu_1 = \mathbb{E}[Z\sigma(Z)] \qquad \mu_2 = (\mathbb{E}[\sigma^2(Z)] - \mu_0^2 - \mu_1^2)^{1/2}.$$

where Z above is a standard Gaussian, and $\boldsymbol{\Sigma} = \mu_1^2 \mathbf{F}^\top \mathbf{F} + \mu_2^2 \mathbf{I}_p$. Our goal will be to show that the test and train error of the nonlinear features $\mathbf{A} = \sigma(\mathbf{XF})$ is the same as that of the *linearized* features

$$\mathbf{B} = \mu_1 \mathbf{XF} + \mu_2 \mathbf{Z} \tag{5.2}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times p}$ has i.i.d. $\mathsf{N}(0, 1)$ entries independent of everything. Let $\mathbf{a}_i$ be the i-th row of $\mathbf{A}$, and respectively for $\mathbf{b}_i, \mathbf{B}$.

Next, we define, for any set of regressors $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_n]$, the optimization objective

$$\Phi_{\mathbf{R}}(\tau_1, \tau_2) := \inf_{\mathbf{w} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \ell(\tfrac{1}{\sqrt{p}} \mathbf{r}_t^\top \boldsymbol{w}; y_t) + \sum_{j=1}^{p} h(w_j) + \tau_1 \cdot (\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}) + \tau_2 \cdot (\sqrt{p} \mu_1 \boldsymbol{\xi}^\top \mathbf{F} \mathbf{w}) \right\}. \tag{5.3}$$

τ_1 and τ_2 are technical terms needed later, and their range is restricted such that the above problem always remains strongly convex. Note that the training loss exactly corresponds to the setting where $\tau_1 = \tau_2 = 0$, up to rescaling.

The proof is then carried out by the Lindeberg method. The authors construct a sequence of learning problems by repeatedly exchanging the nonlinear and linearized features. On each step of the path, they show that the training loss does not change much. The interpolation path used is

$$L_k(\mathbf{w}) = \sum_{t=1}^{k} \ell(\tfrac{1}{\sqrt{p}}\mathbf{b}_t^\top \mathbf{w}; y_t) + \sum_{t=k+1}^{n} \ell(\tfrac{1}{\sqrt{p}}\mathbf{a}_t^\top \mathbf{w}; y_t) + \underbrace{\tau_1 \cdot (\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}) + \tau_2 \cdot (\sqrt{p}\mu_1 \boldsymbol{\xi}^\top \mathbf{F}\mathbf{w})}_{Q(\mathbf{w})}. \quad (5.4)$$

Essentially, the k-th problem on the path has the first k features as the linearized features, but the remaining $n-k$ are nonlinear. The authors control this difference, and this allows them to prove the equivalence of $\Phi(\mathbf{A})/p$ and $\Phi(\mathbf{B})/p$. To strengthen this to equivalence of test error, they require the following additional assumption:

- There exists a limit function $q^*(\tau_1, \tau_2)$ such that $\Phi_{\boldsymbol{B}}(\tau_1, \tau_2)/p \to q^*(\tau_1, \tau_2)$ for all settings of $\tau_{1,2}$; furthermore, the partial derivatives exist at $\tau_1 = \tau_2 = 0$, with $\partial_{\tau_1} q(0,0) = \rho^*$ and $\partial_{\tau_2} q(0,0) = \pi^*$, and $\rho^* \neq 0$.

Essentially, the terms multiplied to τ_1 and τ_2 are in fact the covariance of the linearized model with the inner product present in the true model. That is, when testing on a new point $\mathbf{x}_{\mathsf{new}}$, the error is characterized by the joint distribution of the linearized feature $\mathbf{b}_{\mathsf{new}} = \mu_1 \mathbf{F}^\top \mathbf{x}_{\mathsf{new}} + \mu_2 \mathbf{z}_{\mathsf{new}}$ and the true signal $\xi^\top \mathbf{x}_{\mathsf{new}}$ – it turns out this covariance information is exactly stored in the coefficients of τ_1 and τ_2, and can thus be extracted by taking these partial derivatives.

As mentioned earlier, most of the proof of [HL20] holds for deterministic weight matrices $\mathbf{F}$. The only remaining steps to be checked for when the right-rotationally invariant model are as follows:

1. An approximate orthogonality condition: Let $\mathbf{f}_i$ be the ith column of $\mathbf{F}$, and $\mathbf{f}_0 = \boldsymbol{\xi}$. Then
$$\max_{0 \le i \le j \le p} \left| \mathbf{f}_i^\top \mathbf{f}_j - \delta_{ij} \right| \le (\log p)^2/p \quad (5.5)$$

with probability at least $1 - (\text{polylog } p)/\sqrt{p}$, where $\delta_{ij} = \mathbb{1}(i = j)$.

2. A bounded ℓ_∞ property, which in their setting is provided by [HL20, Lemma 23 and Proposition 2]. Explicitly, let $\mathbf{w}_k^*$ be the optimal solution to the optimization problem L_k. There exists some $c_\infty > 0$ such that for every p and $0 \leq k \leq n$,

$$\mathbb{P}(\|\boldsymbol{w}_k^*\|_\infty \geq \text{polylog } p) \leq c_\infty \exp\left[-c_\infty^{-1}(\log p)^2\right]. \tag{5.6}$$

In fact the orthogonality condition is not hard to check. One way to see that this should hold is that this amounts to controlling $\|\mathbf{F}^\top \mathbf{F} - \mathbf{I}_p\|_\infty$, and as shown in Lemma 3.1, we know $\mathbb{E}[\mathbf{F}^\top \mathbf{F}] = \frac{\mathsf{Tr}(\mathbf{D}^\top \mathbf{D})}{p} \mathbf{I}_p$. Hence if this trace is fixed to be p, one just needs to prove some concentration. To do this, one can rewrite the inner product $\mathbf{f}_i^\top \mathbf{f}_j$ in terms of $(\mathbf{o}_i, \mathbf{o}_j)$, the ith and jth columns of $\mathbf{O}$. These are nothing but a pair of orthogonal vectors on the sphere, and thus can be represented as two normalized Gaussians, where the second is orthogonalized (via Gram-Schmidt) from the first; these terms can then be controlled through standard techniques, such as the Bernstein inequality.

The true difficulty lies in this bounded ℓ_∞ property. We now give a brief overview of how this is proven in [HL20] and illustrate the barrier that arises – this is the proof of [HL20, Lemma 23].

By exchangeability, it suffices to control the size of a single coordinate of the weight vector – without loss of generality, let it be the last one. The strong convexity of the loss function (given by the presence of the regularizer) allows one to then essentially bound the norm of the last coordinate in terms of three other terms:

1. $\sqrt{p}\boldsymbol{\xi}^\top \boldsymbol{f}_p$,

2. $\mathbf{f}_p^\top \mathbf{F}_{-p} \mathbf{w}_k^*$,

3. $\frac{1}{\sqrt{p}} \sum_{t=1}^n \ell(\frac{1}{\sqrt{p}} \mathbf{r}_t^\top \mathbf{w}_k^*)(\mathbf{r}_t)_p$

where $\mathbf{F}_{-p}$ refers to the first $p-1$ columns of $\mathbf{F}$ and $\mathbf{w}_k^* \in \mathbb{R}^{p-1}$ is the solution to L_k above, but with the p-th coordinate of $\mathbf{w}$ already set to 0 (hence making it an optimization over $p-1$ parameters). Also, here $\mathbf{r}_t$ refers to $\mathbf{b}_t$ for $t \leq k$, and $\mathbf{a}_t$ for $t > k$.

For right-rotationally invariant design, the first term is already controlled in the course of the proof of the orthogonality condition, and the third can be dealt with in the manner shown in [HL20], where instead of using Gaussian Lipschitz concentration, we employ the convex concentration property for spherical vectors shown in [Sch14]. Hence all that remains is the second term. This is the core barrier in the proof, and furthermore highlights the distinction between i.i.d. Gaussian and right-rotationally invariant design. In the Gaussian setting, bounding this term is quite easy. One can show that, for fixed $\mathbf{F}$, the norm of $\mathbf{w}_k^*$ is roughly of order $\sqrt{p}$. Furthermore, since it is only fit on the first $p-1$ features, it, along with $\mathbf{F}_{-p}$ are independent of $\mathbf{f}_p$ – hence one can control the norm of $\mathbf{F}_{-p}\mathbf{w}_k^*$, and then apply standard tail bounds to control the product.

Our issue is that the columns of $\mathbf{F}$ are *not* independent anymore – they are weakly dependent. Standard lemmas for conditioning on columns of Haar matrices, such as [RSF19, Lemma 4], which is restated in Lemma A.1, are insufficient for overcoming this, because $\mathbf{w}_k^*$ is fundamentally dependent on all of $\mathbf{F}$. If one tries conditioning on $\mathbf{f}_{p+1}$ by conditioning on $\mathbf{o}_{p+1}$, the conditional distribution of $\mathbf{F}_{-p}$ can be derived using the aforementioned lemma, but not that of $\mathbf{w}_k^*$. Different techniques must be used to control this term.

5.2.2 Method of [Has+22]

An alternate pathway to computing the out-of-sample risk is presented in [Has+22]. Here, the authors utilize techniques from random matrix theory. In particular, one can extract the bias and variance of the model by taking suitable derivatives of the resolvent of a certain matrix. This amounts to computing the limiting Stieltjes transform of a the block matrix using the leave-one-out technique. This technique arrives at the same barrier as in the previous work, where any given row or column of $\mathbf{F}$ is no longer independent of the rest.

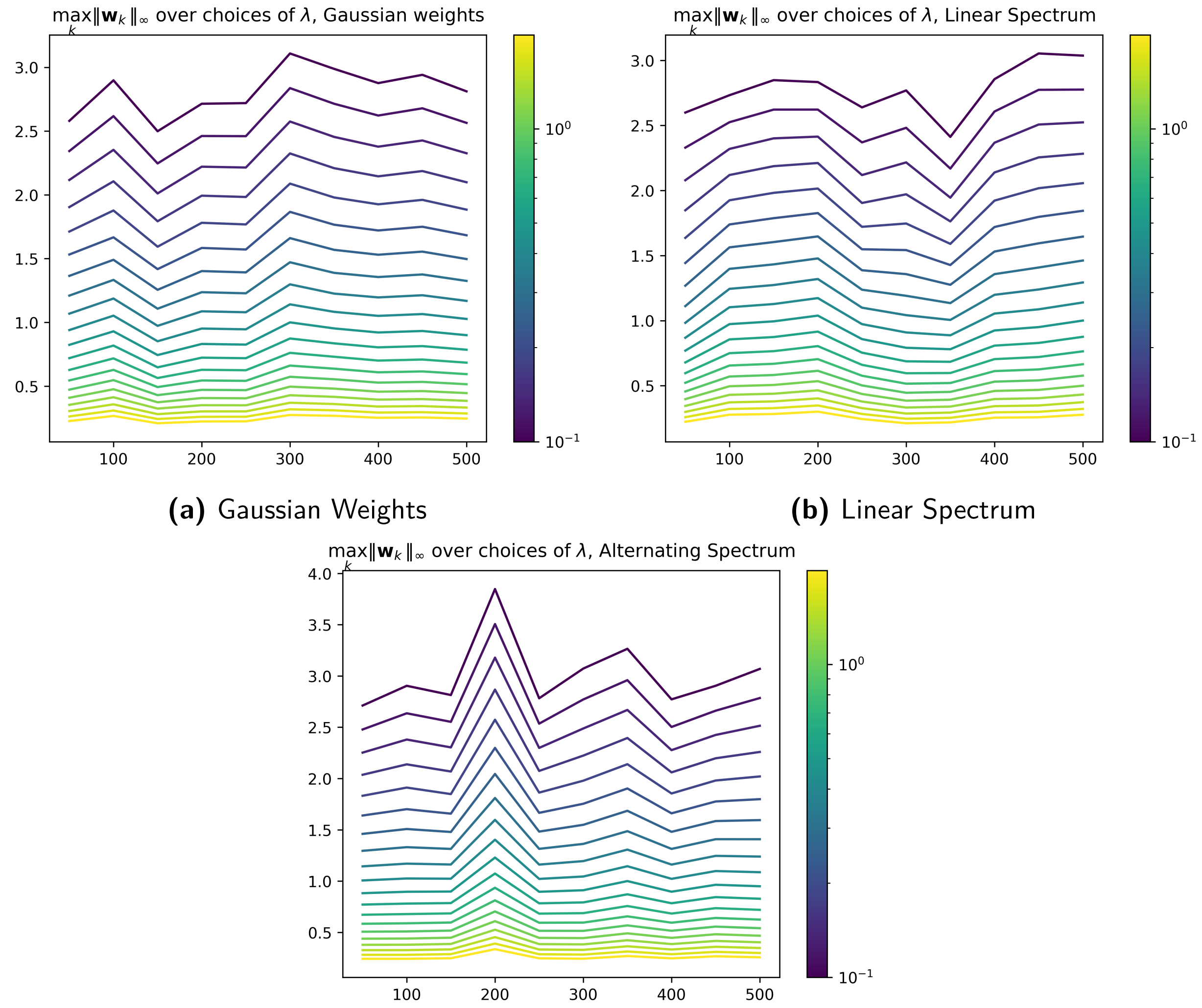
5.3 A Sufficient Conjecture and Numerical Evidence

Hence, the only remaining barrier to proving the Gaussian equivalence is that this quadratic form $\mathbf{f}_p^\top \mathbf{F}_{-p}\mathbf{w}_k^*$ is bounded. We believe this should be true for most settings of $\mathbf{D}$. In this section, we present numerical evidence that this is the case. See Figure 5.1, which displays the maximum ℓ_∞ norm over all k for the Gaussian setting as well as two right-rotationally invariant designs. For the linear spectrum example, the entries of $\mathbf{D}$ are linearly spaced, and further normalized such that $\mathsf{Tr}(\mathbf{D}^\top \mathbf{D}) = p$. For the alternating spectrum example, half of the singular values are 1, and the other half are 2; the spectrum is again normalized.

These examples show that for these settings of the spectrum, the growth of the ℓ_∞ norm is very moderate - we need it to be polylogarithmic in p, and admittedly simulating a large range of $\log p$ is quite difficult. However, all examples seem to suggest very slow growth of this quantity, though its possible one may need to set λ sufficiently large in some settings.

5.3.1 What Remains

Not much remains to be shown after proving the above conjecture. The equivalent Gaussian model has already been extensively studied in works such as [DL20; Has+22], and one then simply needs to analyze its behavior under the covariance structure in the linearized model that is induced by the right-rotationally invariant distribution of $\mathbf{F}$. Hence, one should be able to calculate the limiting risk in terms of the limiting spectrum of $\mathbf{F}$, thus completing the picture.



(a) Gaussian Weights

(b) Linear Spectrum

(c) Alternating Spectrum

Figure 5.1: The value of $\max_k \|\mathbf{w}_k^*\|$ as p grows, over a range of λ. Darker λ refers to lower regularization. Plotted value is the maximum value across 20 trials. Activation is a centered sigmoid.

A
Proofs

A.1 Evidence for Universality

See Figure A.1, which shows that the right-rotationally invariant risk prediction still holds despite the designs no longer being actually right-rotationally invariant.

A.2 Proofs for Chapter 3

A.2.1 Discussion of in-sample risks

Recall the problem that we wish to consider is as follows. We partition the rows and columns of $\mathbf{X}$ and $\mathbf{y}$ as $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top]^\top$ and $\mathbf{y} = [\mathbf{y}_1^\top, \mathbf{y}_2^\top]^\top$, where $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{y}_1 \in \mathbb{R}^{n_1}$, while $\mathbf{X}_2 \in \mathbb{R}^{n_2 \times p}, \mathbf{y}_2 \in \mathbb{R}^{n_2}$, with $n_1 + n_2 = n$. We then try using $(\mathbf{X}_1, \mathbf{y}_1)$ to learn $\hat{\boldsymbol{\beta}}_1$, and then calculate the risk

$$\frac{1}{n_2}\mathbb{E}\left[\|\mathbf{X}_2\hat{\boldsymbol{\beta}} - \mathbf{X}_2\boldsymbol{\beta}\|^2 \mid \mathbf{X}_1\right].$$

The particular result necessary for this setting is the following:

Lemma A.1 ([RSF19, Lemma 4], conditional dist. of a Haar matrix)**.** *Let* $\mathbf{O} \sim \mathrm{Haar}(\mathbb{O}(p))$*, and let* G *be the event that* $\mathbf{A} = \mathbf{OB}$*, where* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times s}$ *for some* $s \leq p$

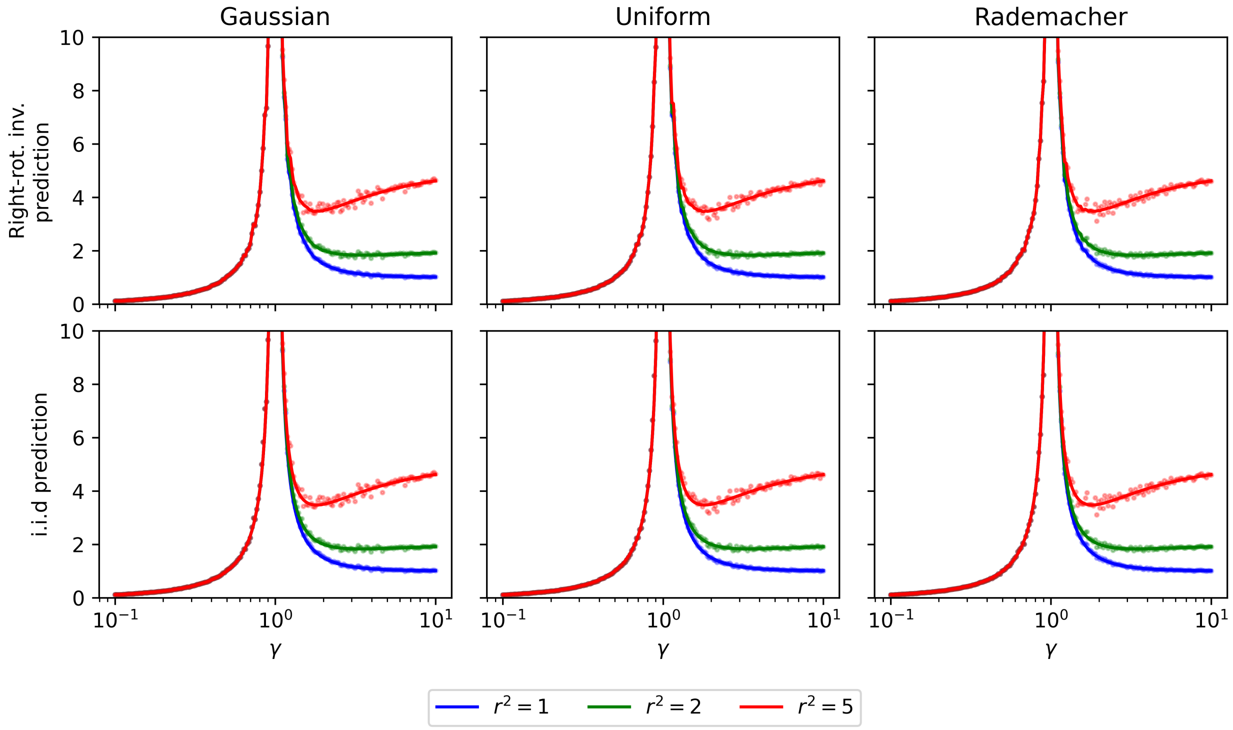


Figure A.1: I.i.d. rows, Gaussian, Uniform, and Rademacher distributions for the entries; scaled so that every row has mean zero and variance 1.

are fixed matrices. Assume $\mathbf{A}, \mathbf{B}$ *are of full column rank, and let* $\mathbf{U}_{\mathbf{A}^\perp}$ *and* $\mathbf{U}_{\mathbf{B}^\perp}$ *be any* $p \times (p - s)$ *matrices whose columns are orthonormal bases for* $\mathrm{Range}(\mathbf{A})^\perp$ *and* $\mathrm{Range}(\mathbf{B})^\perp$*, respectively.*

Then the conditional distribution of $\mathbf{O}$ *given* G *can be written as*

$$\mathbf{O}|_G \overset{d}{=} \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{B}^\top + \mathbf{U}_{\mathbf{A}^\perp}\widetilde{\mathbf{O}}\mathbf{U}_{\mathbf{B}^\perp}^\top \tag{A.1}$$

where $\widetilde{\mathbf{O}} \sim \mathrm{Haar}(\mathbb{O}(p - s))$ *is independent of* G*.*

The standard way to approach this is treat $\mathbf{Q}$, $\mathbf{D}$ as fixed, and then to take a matrix $\mathbf{M}_1 \in \mathbb{R}^{n_1 \times n}$, where $M_{ii} = 1$, for $1 \le i \le n_1$. One conditions on $\mathbf{X}_1$ by conditioning on the event $\mathbf{X}_1 = \mathbf{M}_1\mathbf{X} = \mathbf{M}_1\mathbf{Q}^\top\mathbf{D}\mathbf{O}$, which can now be done using the above Lemma. However, one immediate restriction is now that we must have $n_1 < p$; this is a restriction just from the Lemma, but since we are treating $\mathbf{Q}, \mathbf{D}$ as fixed, in fact $\mathbf{O}$ is already fixed after observing more than n_1 rows of $\mathbf{X}$. If one works under this assumption, then the computation can be carried out:

$$\mathbb{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}_2^\top \mathbf{X}_2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \mid \mathbf{X}_1] = \mathsf{Tr}\left(\mathbb{E}[\mathbf{X}_2^\top \mathbf{X}_2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mid \mathbf{X}_1]\right)$$

One now notes that $\mathbf{X}_2$ and $\hat{\boldsymbol{\beta}}$ are conditionally independent due to the fact that $\boldsymbol{\beta}$ is a function of only $\mathbf{X}_1$ and the independent noise $\boldsymbol{\epsilon}_1$. Hence:

$$= \mathsf{Tr}\left(\mathbb{E}[\mathbf{X}_2^\top \mathbf{X}_2 \mid \mathbf{X}_1]\mathbb{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mid \mathbf{X}_1]\right)$$

Now both terms can be computed directly by employing the Lemma. We write $\mathbf{X}_1 = \mathbf{M}_2\mathbf{Q}^\top\mathbf{D}\mathbf{O}$, where $\mathbf{O}$, conditional on $\mathbf{X}_1$, using the above Lemma, has some closed form representation involving a new independently random matrix $\widetilde{O}$. The result in a complicated expression involving $\mathbf{Q}, \mathbf{D}, \mathbf{X}_1$. As we have no distributional assumptions on $\mathbf{Q}$, this expression is not meaningful – it is perhaps possible to do something when $\mathbf{Q}$ is also Haar, but this analysis is not pursued here.

A.2.2 Bias-Variance Decomposition

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{X}_{\mathsf{new}}\hat{\boldsymbol{\beta}} - \mathbf{X}_{\mathsf{new}}\boldsymbol{\beta}\|^2 \mid \mathbf{X}\right] &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}_{\mathsf{new}}^\top \mathbf{X}_{\mathsf{new}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \mid \mathbf{X}\right] \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbb{E}[\mathbf{X}_{\mathsf{new}}^\top \mathbf{X}_{\mathsf{new}}](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \mid \mathbf{X}\right] \\
&= \frac{\mathbb{E}[\mathsf{Tr}(\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}})]}{p}\mathbb{E}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \mid \mathbf{X}\right] \qquad \text{(Lemma 3.1)} \\
&= \frac{\mathbb{E}[\mathsf{Tr}(\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}})]}{p}\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] + \mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \boldsymbol{\beta}\|^2 \mid \mathbf{X}] \\
&= \frac{\mathbb{E}[\mathsf{Tr}(\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}})]}{p}\left(\|\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \boldsymbol{\beta}\|_2^2 + \mathbb{E}[\|\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}]\|_2^2 \mid \mathbf{X}]\right)
\end{aligned}$$

and thus

$$R_{\mathbf{X}}\left(\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}), \boldsymbol{\beta}\right) = \frac{\mathbb{E}[\mathsf{Tr}(\mathbf{D}_{\mathsf{new}}^\top \mathbf{D}_{\mathsf{new}})]}{p} \cdot \left(\underbrace{\frac{1}{n}\|\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \boldsymbol{\beta}\|_2^2}_{B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})} + \underbrace{\frac{1}{n}\mathsf{Tr}(\mathsf{Cov}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}))}_{V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})}\right).$$

A.3 Ridge(less) regression proofs

A.3.1 Proof: Risk of finite ridge regularization

The variance is direct. One calculates

$$\begin{aligned}
V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= n^{-1}\mathsf{Tr}(\mathsf{Cov}(\hat{\boldsymbol{\beta}} \mid \mathbf{X})) \\
&= n^{-1}\sigma^2\mathsf{Tr}((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}) \\
&= n^{-1}\sigma^2\mathsf{Tr}((\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}) \\
&= n^{-1}\sigma^2\mathsf{Tr}((\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{D}) \\
&= \frac{\sigma^2}{n}\mathsf{Tr}\left((\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-2}\mathbf{D}^\top\mathbf{D}\right) = \sum_{i=1}^{n\wedge p}\frac{D_{ii}^2}{(D_{ii}^2 + \lambda)^2}
\end{aligned}$$

For the bias,

$$\begin{aligned}
B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= n^{-1}\|\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \boldsymbol{\beta}\|_2^2 \\
&= n^{-1}\boldsymbol{\beta}(\mathbf{I} - (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X})^2\boldsymbol{\beta} \\
&= n^{-1}(\mathbf{O}\boldsymbol{\beta})^\top\left[\mathbf{I} - (\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{D}\right]^2(\mathbf{O}\boldsymbol{\beta}) \\
&= n^{-1}(\mathbf{O}\boldsymbol{\beta})^\top\left(\lambda(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\right)^2(\mathbf{O}\boldsymbol{\beta})
\end{aligned}$$

Again let $\mathbf{b} = (\mathbf{O}\boldsymbol{\beta})/\|\boldsymbol{\beta}\|$ and $\mathbf{P} = \left(\lambda(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\right)^2$, and note $\mathsf{Tr}(\mathbf{P}) = \sum_{i=1}^{p}\frac{\lambda^2}{(D_{ii}^2+\lambda)^2}$ and $\|\mathbf{P}\|_F^2 = \sum_{i=1}^{p}\frac{\lambda^4}{(D_{ii}^2+\lambda)^4}$. Furthermore, one has $\mathbb{E}[\mathbf{b}^\top\mathbf{P}\mathbf{b}] = \frac{1}{p}\mathsf{Tr}(\mathbf{P})$ using 3.3, and clearly $\|\mathbf{P}\| \leq 1$. We now apply 3.2 to show concentration:

$$\begin{aligned}
\mathbb{P}\left(\left|\mathbf{b}^\top\mathbf{P}\mathbf{b} - \frac{1}{p}\mathsf{Tr}(\mathbf{P})\right| \geq t\right) &\leq 2\exp\left(-C\min\left(\frac{p^2t^2}{2K^4\|\mathbf{P}\|_F^2}, \frac{pt}{K^2}\right)\right) \\
&= 2\exp\left(-Cp\min\left(\frac{t^2}{2K^4\frac{1}{p}\|\mathbf{P}\|_F^2}, \frac{t}{K^2}\right)\right)
\end{aligned}$$

Note that $\frac{1}{p}\|\mathbf{P}\|_F^2$ is always bounded above by 1, so this bound can never be made vacuous by a certain setting of $\mathbf{D}$.

We can then obtain the same $n^{1/2}$ rates for the bias under proportional asymptotics.

Plugging in $t = n^{-\frac{1}{2}-\frac{1}{k}}$, one obtains

$$\mathbb{P}\left(\left|\mathbf{b}^\top \mathbf{P}\mathbf{b} - \frac{1}{p}\mathsf{Tr}(\mathbf{P})\right| \geq n^{-\frac{1}{2}+\frac{1}{k}}\right) = 2\exp\left(-C\gamma \min\left(\frac{n^{2/k}}{2K^4\frac{1}{p}\|\mathbf{P}\|_F^2}, \frac{n^{1/2+1/k}}{K^2}\right)\right)$$

which again is less than n^{-D}, for any integer D, once n is sufficiently large in terms of k and D. Now note

$$\frac{1}{p}\mathsf{Tr}(\mathbf{P}) = \lambda^2 m_{\mathbf{D}}(-\lambda).$$

A.4 GCV proofs

A.4.1 LOOCV Intractability

The LOOCV in our setting is difficult to analyze for reasons similar to the in-sample prediction problem discussed in Remark 1 – the objective depends in $\mathbf{Q}$, on which we do not place any assumptions. In particular, one can derive one has

$$\begin{aligned}\mathsf{LOOCV}_n(\lambda) &= \mathbf{y}^\top(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{D}_\lambda^{-2}(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}/n \\ &= \mathbf{y}^\top(\mathbf{I} - \mathbf{Q}^\top(\mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)\mathbf{Q})\mathbf{D}_\lambda^{-2}(\mathbf{I} - \mathbf{Q}^\top(\mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)\mathbf{Q})\mathbf{y}/n\end{aligned}$$

where $\mathbf{D}_\lambda = \operatorname{diag}((1 - (\mathbf{S}_\lambda)_{ii})_{i=1}^n)$. We try to simplify the diagonal term first.

$$\begin{aligned}(D_\lambda)_{ii} &= 1 - \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_i \\ &= 1 - \mathbf{q}_i^\top\mathbf{D}\mathbf{O}(\mathbf{O}^\top\mathbf{D}^\top\mathbf{D}\mathbf{O} + \lambda\mathbf{I})^{-1}\mathbf{O}^\top\mathbf{D}^\top\mathbf{q}_i \\ &= 1 - \mathbf{q}_i^\top\mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top\mathbf{q}_i\end{aligned}$$

where $\mathbf{q}_i$ is the i-th row of $\mathbf{Q}$. We analyze

$$\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q}^\top = \sum_{i=1}^n \mathbf{q}_i\mathbf{q}_i^\top(\mathbf{D}_\lambda^{-2})_{ii}.$$

Then LOOCV should contribute two terms, as the cross term should vanish. They are

$$(\mathbf{O}\boldsymbol{\beta})^\top\mathbf{D}^\top\mathbf{Q}(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)(\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q}^\top)(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)\mathbf{O}\boldsymbol{\beta}$$

which should concentrate around its trace.

The second is

$$\boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)(\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q}^\top)(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\boldsymbol{\epsilon}$$

which also concentrates around some multiple of its trace.

Finding this trace now involves numerous terms with $\mathbf{Q}$. Everything except the center term is diagonal, but immediately one arrives at the following difficulty:

$$\begin{aligned}(\mathbf{Q}\mathbf{D}_\lambda^{-2}\mathbf{Q})_{ii} &= \sum_{j=1}^{n}(\mathbf{q}_j\mathbf{q}_j^\top)_{ii}(\mathbf{D}_\lambda^{-2})_{jj} \\ &= \sum_{j=1}^{n}(\mathbf{q}_{ji})^2(\mathbf{D}_\lambda^{-2})_{jj}\end{aligned}$$

Assuming that the rows of $\mathbf{Q}$ are exchangeable conditional on $\mathbf{D}$ is insufficient to make progress. If $\mathbf{Q}$ is assumed to also be Haar, then likely closed forms can be derived, but this is not pursued.

A.4.2 Proof: Standard GCV for right-rotationally invariant designs

We explicitly bound each term and show almost sure convergence.

$$T_1 = \frac{1}{n}(\mathbf{O}\boldsymbol{\beta})^\top\mathbf{D}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D}(\mathbf{O}\boldsymbol{\beta})$$

This again concentrates around its expectation, which is

$$\begin{aligned}&= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\mathsf{Tr}(\mathbf{D}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D})}{p} \\ &= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{1}{p}\sum_{i=1}^{p}D_{ii}^2\left(1 - \frac{D_{ii}^2}{D_{ii}^2 + \lambda}\right)^2 \\ &= \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\lambda^2}{p}\sum_{i=1}^{p}\frac{D_{ii}^2}{(D_{ii}^2 + \lambda)^2} \\ &= \frac{\|\boldsymbol{\beta}\|^2}{n}\lambda^2\left(m_{\mathbf{D}}(-\lambda) - \lambda m_{\mathbf{D}}'(-\lambda)\right) = \frac{\|\boldsymbol{\beta}\|^2}{n}\frac{\lambda^2}{\gamma}\left(v_{\mathbf{D}}(-\lambda) - \lambda v_{\mathbf{D}}'(-\lambda)\right)\end{aligned}$$

Letting $\mathbf{P} = \mathbf{D}^\top(\mathbf{I} - \mathbf{D}(\mathbf{D}^\top\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^\top)^2\mathbf{D}$, Hanson-Wright (Lemma 3.2)implies

$$\mathbb{P}(|T_1 - \mathbb{E}[T_1]| > t) \leq 2\exp\left(-c\min\left(\frac{p^2t^2}{2K^4\|\mathbf{P}\|_F^2}, \frac{pt}{K^2\|\mathbf{P}\|}\right)\right) = 2\exp\left(p\min\left(\frac{t^2}{2K^4\frac{1}{p}\|\mathbf{P}\|_F^2}, \frac{t}{K^2\|\mathbf{P}\|}\right)\right)$$

Note that

$$\mathbf{P} = \operatorname{diag}\left(\frac{\lambda^2 D_{ii}^2}{(D_{ii}^2 + \lambda)^2}\right)_{i=1}^p$$

Hence

$$\begin{aligned}\frac{1}{p}\|\mathbf{P}\|_F^2 &= \frac{1}{p}\sum_{i=1}^p\left(\frac{\lambda^2 D_{ii}^2}{(D_{ii}^2 + \lambda)^2}\right)^2 = \frac{1}{p}\sum_{i=1}^p\left(\frac{\lambda^2}{(D_{ii}^2 + \lambda)^2}\right)^2 D_{ii}^4 \leq \|\mathbf{D}\|_{\mathsf{op}}^4 \\ \|\mathbf{P}\|_{\mathsf{op}} &= \frac{\lambda^2 D_{11}^2}{(D_{11}^2 + \lambda)^2} \leq \|\mathbf{D}\|^2\end{aligned}$$

By assumption, both terms are almost surely bounded in the limit. Recalling that $p(n)/n \to \gamma$, the above bound is summable in n and hence Borel-Cantelli implies almost sure convergence.

To handle the second term, we use a slightly different method.

$$\begin{aligned}T_2 &= \frac{1}{n}\boldsymbol{\epsilon}^\top(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{X}\boldsymbol{\beta} \\ &= \frac{1}{n}(\mathbf{O}\boldsymbol{\beta})^\top\underbrace{\mathbf{D}^\top\mathbf{Q}(\mathbf{I} - \mathbf{S}_\lambda)^2\boldsymbol{\epsilon}}_{\mathbf{T}_4}\end{aligned}$$

Note that $\tilde{\boldsymbol{\beta}} = \mathbf{O}\boldsymbol{\beta}$ is a uniformly random vector of norm $\|\boldsymbol{\beta}\|$. We replace this with $\|\boldsymbol{\beta}\|\frac{\mathbf{g}}{\|\mathbf{g}\|}$ where $\mathbf{g} \sim \mathsf{N}(0, \mathbf{I}_p/p)$. Then conditional on $\mathbf{D}, \boldsymbol{\epsilon}$, we have $\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T}_4 \sim \mathsf{N}(0, \|\mathbf{T}_4\|/(pn))$. Furthermore note $\|\mathbf{T}_4\| \leq \|\mathbf{D}^\top\|\|\mathbf{Q}\|\|\mathbf{I} - \mathbf{S}_\lambda\|^2\|\boldsymbol{\epsilon}\| \leq \|\mathbf{D}^\top\|\|\boldsymbol{\epsilon}\|$. Some standard tail bounds complete this.

$$\frac{1}{n}\|\boldsymbol{\beta}\|\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\mathbf{T}_4 = \frac{\|\boldsymbol{\beta}\|}{\sqrt{n}}\frac{1}{\|\mathbf{g}\|}\left(\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T}_4\right) = r \cdot \frac{1}{\|\mathbf{g}\|}\frac{1}{\sqrt{n}}\mathbf{g}^\top\mathbf{T_4}$$

Hence

$$\mathbb{P}\left(|\frac{1}{\sqrt{n}}\mathbf{g}^\top \mathbf{T_4}| > t \mid \mathbf{D}, \boldsymbol{\epsilon}\right) \leq \sqrt{\frac{2}{\pi}} \frac{\|\mathbf{T}_4\|}{t \cdot \sqrt{pn}} \exp\left(-t^2/2 \cdot \frac{pn}{\|\mathbf{T}_4\|^2}\right).$$

We claim that $\limsup_{n\to\infty} \|\mathbf{T}_4\|/\sqrt{n}$ is bounded. This follows from the bound above, the law of large numbers applied to $\boldsymbol{\epsilon}$, and the assumption on $\mathbf{D}$. Hence the bound above is summable in n once more, so by Borel-Cantelli, the convergence is almost sure. An application of Slutsky finishes, since we know $\|\mathbf{g}\| \xrightarrow{a.s.} 1$ by the strong Law of Large Numbers. The convergence for T_3 follows directly from [DW18, Lemma C.3]

Alternative modified GCV motivation

Suppose we believe that there exists some appropriate denominator for the GCV that provides the correct correction factor. Then it must be true that the ratio of the coefficients of σ^2 and r^2 in the GCV asymptotic risk (after the excess risk σ^2 is removed) must be equal to the ratio of the asymptotic form of the ridge risk. This implies that the denominator $\mathfrak{d}$ must satisfy

$$\begin{aligned}\frac{\sigma^2 \text{ coefficient}}{r^2 \text{ coefficient}} = \frac{\lambda^2 v'(-\lambda) - \mathfrak{d}}{v(-\lambda) - \lambda v'(-\lambda)} &= \frac{m(-\lambda) - \lambda m'(-\lambda)}{m'(-\lambda)} \\ \iff \mathfrak{d} = \lambda^2 v'(-\lambda) - &\frac{(m(-\lambda) - \lambda m'(-\lambda))(v(-\lambda) - \lambda v'(-\lambda))}{m'(-\lambda)}; \quad \text{(A.2)}\end{aligned}$$

If we solve for $\mathfrak{d}$ and use this as the denominator, the result is that GCV is still not correct in general (meaning the modified GCV risk, minus the excess noise, is not equal to the ridge risk) - the error still does not match the ridge asymptotic form, because this will always only guarantee a *scaled* multiple of the asymptotic ridge risk. The scale factor is not λ-free - however, we again know its form, as it is nothing but

$$\frac{\mathsf{GCV}\ r^2 \text{ coefficient}}{\lambda\text{-ridge } r^2 \text{ coefficient}} = \frac{m'(-\lambda)}{v(-\lambda) - \lambda v'(-\lambda)}. \tag{A.3}$$

It turns out after unravelling these steps, this method does exactly what is proposed in the main text.

A.5 Simulation Details

A.5.1 Figures 2,4,5

The true signal is $f(x) = \sin(2.5x) * 10$; the x's are fixed as 10 evenly spaced points between 2 and -2, inclusive (probably should have used 11 points). y_i is computed via $f(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathsf{N}(0, 3^2)$ is independent of everything. Regression features include one bias term, followed by random features of the form $\mathsf{sigmoid}(g_1 x + g_2)$, with g_1, g_2 independent standard normals. The original goal was to do polynomial regression, but this lead to too much numerical instability. Minimum norm regression was used when $p > n$. Each alternative fit was computed by resampling the noise ϵ - the random features were not resampled.

A.5.2 Experiments from Chapter 2

Semi-synthetic: Speech Data

Data was retrieved from OpenML repository with ID 40910 [Le; Van+13]. For a given setting of r^2, σ^2, the experiment was conducted as follows. First, one computes $n = [p/\gamma]$. Then the first n rows of the dataset are taken to be the training set, and all of the remaining rows of the dataset are taken to be a test set. A single $\boldsymbol{\beta}$ is then sampled. Now we repeatedly sample the noise $\boldsymbol{\epsilon}$ (since our risks are only *design* conditional) and measure the squared error on the ground truth of the test set, $\boldsymbol{X}_{\mathsf{new}}\boldsymbol{\beta}$. Note that the noise is also *not* added in the training set. The Gaussian predictions were computed as per [Has+22, Theorem 1], and right-rotationally invariant predictions were just computed according to the formula given in Chapter 3.

Semi-synthetic: residualized returns

When building models for predicting financial returns for a given company, many in the industry working on moderate to long term time horizons are interested in predicting the *residual(ized)* return of a given stock. The residual return refers to the return after market factors, such as the state of the global economy, are removed. One can imagine that when the entire market moves up, that most stocks will also (on average) go up – the individual properties of the actual company many not be predictive of this, and hence it is desirable to have this portion of the return removed. Once all market factors are removed, the result is the residualized return; the residual returns of two companies are then generally taken to be nearly independent, since the underlying factors driving their joint movement are assumed to be factored out.

The standard way in which residualiziation is done by employing factor models, which provide a list of factors, as well as loadings of each stock on each factor (i.e. how much of the stocks movement is dictated by a given factor). The way in this is computed is proprietary, as there are companies entirely dedicated to this, such as MSCI which produces Barra models. One then attempts to construct portfolios which are hedged against all factors, i.e. their total loadings are all zero.

We do not have access to such a factor model, and hence in computing our residualized returns, we use PCA to delete some top left singular vectors of the data matrix $\mathbf{X}$, where each row is an observation of all stocks at a given timepoint, and each column is a stock. Our experiments only remove 8 such components – in comparison, the base Barra model has around 90 factors. In particular, our procedure is not really useful in practice, since we use the test data to help compute the factors to delete. The reason we do this is because PCA is quite crude and weak as a method for removing factors, and hence it is likely more comparable to remove PCs within the sample we will test on.

Experiment Details:

1. We take minutely returns of all symbols within the S&P500, NASDAQ100, and the top 500 largest tech companies using the Polygon API from 02-01 to 03-22 of 2024. The reason for minutely returns was because this data was gathered for the real data test in 3.5.3. The problem there is that the S&P 500 rebalances quarterly, meaning the loadings change over time, and hence recent data was necessary.

2. We remove all symbols which have more than 70 NaN values on a given day, resulting in $p \approx 500$ remaining.

3. We center and standardize all returns before computing the SVD of the matrix $\mathbf{X} = \mathbf{Q}^\top \mathbf{D} \mathbf{O}$. Let $\mathbf{Q}$ have rows $\mathbf{Q}_i$. The matrix deleting the top k factors is then $(\mathbf{I}_n - \mathbf{Q}_{1:k}\mathbf{Q}_{1:k}^\top)$. We apply this to the design, and then restandardize. This produces the residualized returns.

4. For measuring mean-squared error, we have a training pool of the week March 11th to March 15th. The testing pool is then the following week, March 18th to March 22nd.

5. For each setting of $n = [\gamma p]$, we take the first n datapoints of a given day of the week, and test it on the first n datapoints of that same day the next week (if there are that many more datapoints to the end of the week). Hence larger values of n have fewer iterations. For each of these tests, we repeat the same procedure done for the noised example, where we resample the noise 100 times and measure the MSE against the truth.

6. To compute the lagged variants, we simply add consecutive rows of residualized returns together (using the small number approximation $(1+x)(1+y) \sim 1+x+y$).

A.6 Miscellany

Lemma A.2 (Equivalent definitions of right rotationally invariant)**.** *A random matrix* $\mathbf{X}$ *satisfies* $\mathbf{X} \overset{d}{=} \mathbf{X}\mathbf{R}$ *for any* $\mathbf{R} \in \mathrm{Haar}(\mathbb{O}(p))$ *if and only if its SVD* $\mathbf{X} = \mathbf{Q}^\top \mathbf{D}\mathbf{O}$ *can be written to satisfy* $\mathbf{O} \sim \mathrm{Haar}(\mathbb{O}(p))$ *and* $\mathbf{O} \perp\!\!\!\perp \mathbf{Q}, \mathbf{D}$*. We say "can be written" because there exists natural ambguity in the ordering of the right singular vectors, especially when* $\mathbf{D}$ *does not have full column rank.*

Proof. The reverse direction is clear. A matrix $\mathbf{O}$ is distributed according to $\mathrm{Haar}(\mathbb{O}(p))$ if, for any Borel set B and any $\mathbf{R} \in \mathbb{O}(p)$, one has $\mathbb{P}(\mathbf{O} \in B) = \mathbb{P}(\mathbf{O}\mathbf{R} \in B)$. Hence if $\mathbf{X} = \mathbf{Q}^\top \mathbf{D}\mathbf{O}$, with $\mathbf{O} \perp\!\!\!\perp \mathbf{Q}, \mathbf{D}$, then $\mathbf{X}\mathbf{R} = \mathbf{Q}^\top \mathbf{D}\mathbf{O}\mathbf{R} \overset{d}{=} \mathbf{Q}^\top \mathbf{D}\mathbf{O}$, since $\mathbf{O}\mathbf{R} \overset{d}{=} \mathbf{O}$ and this piece is independent from the rest.

For the forward direction, let $\mathbf{O}'$ be Haar independent of $\mathbf{X}$. Then $\mathbf{X}\mathbf{O}' \overset{d}{=} \mathbf{X} \implies \mathbf{Q}^\top \mathbf{D}\mathbf{O} \overset{d}{=} \mathbf{Q}^\top \mathbf{D}\mathbf{O}\mathbf{O}'$. Conditional on $\mathbf{Q}, \mathbf{D}, \mathbf{O}$, note that $\mathbf{O}\mathbf{O}'$ is still Haar; if we now condition on only $\mathbf{Q}, \mathbf{D}$, the product $\mathbf{O}\mathbf{O}'$ is Haar. Hence $\mathbf{O}\mathbf{O}' \perp\!\!\!\perp \mathbf{Q}, \mathbf{D}$. Denote $\widetilde{\mathbf{O}} = \mathbf{O}\mathbf{O}'$. We now have $\mathbf{X} \overset{d}{=} \mathbf{Q}^\top \mathbf{D}\widetilde{\mathbf{O}}$. By Skorokhod, it is then possible to define a coupling $(\mathbf{X}', (\mathbf{Q}', \mathbf{D}', \mathbf{O}'))$ of $\mathbf{X}$ and $(\mathbf{Q}, \mathbf{D}, \widetilde{\mathbf{O}})$ such that $\mathbf{X}' = (\mathbf{Q}')^\top \mathbf{D}'\mathbf{O}'$, and we are done. □

A.6.1 Risk equivalence

We give a quick argument for the equivalence in analyzing the following forms of risk:

1. $\mathbb{E}[(\mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta} - \mathbf{x}_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2]$
2. $\mathbb{E}[(\mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta} - \mathbf{x}_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2 \mid \mathbf{X}]$
3. $\mathbb{E}[(\mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta} - \mathbf{x}_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2 \mid \mathbf{X}, \mathbf{y}]$
4. All of the above but with $(\mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta} - \mathbf{x}_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2$ replaced with $(y_{\mathsf{new}} - \mathbf{x}_{\mathsf{new}}^\top \hat{\boldsymbol{\beta}})^2$

First, to see that analyzing (4) is equivalent to analyze the respective version of (1-3), note that $y_{\mathsf{new}} = \mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta} + \epsilon_{\mathsf{new}}$ and ϵ_{new} is independent of everything. Hence $\mathbb{E}[(y_{\mathsf{new}} - \mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta})^2 \mid \ldots] = \sigma^2 + \mathbb{E}[(\mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta} - \mathbf{x}_{\mathsf{new}}^\top \boldsymbol{\beta})^2 \mid \ldots]$.

To see that (3) concentrates around (2), recall that

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{\beta} + \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}.$$

Now expanding (3) produces three terms: one is a quadratic form involving $\boldsymbol{\beta}$, the second is a cross term involving $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}$, and the last is a quadratic form involving $\boldsymbol{\epsilon}$. Under sufficient conditions on $\boldsymbol{\epsilon}$, one can argue, either using Hanson-Wright or [DW18, Lemma C.3], that it concentrates around its expectation. Likewise, one can show that, because ϵ is independent of $\boldsymbol{\beta}$, using some Law of Large Numbers type argument, that the cross term also concentrates around zero.

Finally, that (2) concentrates around (3) usually involves some limiting behavior on the singular values of $\mathbf{X}$ (or equivalently, the spectrum of $\mathbf{X}^\top \mathbf{X}$) – this is directly illustrated by [CM22; DW18; Has+22], or by the results in this work, which show that this design-conditional risk has a limit independent of $\mathbf{X}$ if the its singular values converge to some sufficiently regular distribution.

A.6.2 Wasserstein-2 Convergence

The following section is taken from [LS23, Definition 2.2]. For a matrix $(\mathbf{v}_1, \ldots, \mathbf{v}_k) = (v_{i,1}, \ldots, v_{i,k})_{i=1}^n \in \mathbb{R}^{n \times k}$ and a random vector $(\mathsf{V}_1, \ldots, \mathsf{V}_k)$, we write

$$(\mathbf{v}_1, \ldots, \mathbf{v}_k) \xrightarrow{W_2} (\mathsf{V}1_, \ldots, \mathsf{V}_k)$$

to mean that the empirical distribution of the columns of $\mathbf{v}_1, \ldots, \mathbf{v}_k)$ converge to $\mathsf{V}_1, \ldots, \mathsf{V}_k$ in Wasserstein-2 distance. This means that for any continuous function $f : \mathbb{R}^k \to \mathbb{R}$ satisfying

$$|f(x, \ldots, x_k)| \le C \left(1 + \|(x_1, \ldots, x_k)\|^2\right) \tag{A.4}$$

for some $C > 0$, we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(v_{i,1}, \ldots, v_{i,k}) = \mathbb{E}\left[f(\mathsf{V}_1, \ldots, \mathsf{V}_k)\right]$$

where $\mathbb{E}\left[\|(\mathsf{V}_1, \ldots, \mathsf{V}_k)\|^2\right] < \infty$.

References

[Ada14] Radosław Adamczak. *A note on the Hanson-Wright inequality for random vectors with dependencies*. 2014. arXiv: 1409.8457 [math.PR].

[AP20a] Ben Adlam and Jeffrey Pennington. "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*. ICML'20. JMLR.org, 2020.

[AP20b] Ben Adlam and Jeffrey Pennington. "Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 11022–11032. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/7d420e2b2939762031eed0447a9be19f-Paper.pdf.

[ASS20] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. "High-dimensional dynamics of generalization error in neural networks". In: *Neural Networks* 132 (2020), pp. 428–446. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2020.08.022. URL: https://www.sciencedirect.com/science/article/pii/S0893608020303117.

[Ba+22] Jimmy Ba et al. "High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 37932–37946. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/f7e7fabd73b3df96c54a320862afcb78-Paper-Conference.pdf.

[Bar+20] Peter L. Bartlett et al. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (Apr. 2020), pp. 30063–30070. ISSN: 1091-6490. DOI: 10.1073/pnas.1907378117. URL: http://dx.doi.org/10.1073/pnas.1907378117.

[Bel+19] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (July 2019), pp. 15849–15854. ISSN: 1091-6490. DOI: 10.1073/pnas.1903070116. URL: http://dx.doi.org/10.1073/pnas.1903070116.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two Models of Double Descent for Weak Features". In: *SIAM Journal on Mathematics of Data Science* 2.4 (Jan. 2020), pp. 1167–1180. ISSN: 2577-0187. DOI: 10.1137/20m1336072. URL: http://dx.doi.org/10.1137/20M1336072.

[BM12] Mohsen Bayati and Andrea Montanari. "The LASSO risk for Gaussian matrices". In: *IEEE Trans. Inform. Theory* 58.4 (2012), pp. 1997–2017. ISSN: 0018-9448,1557-9654. DOI: 10.1109/TIT.2011.2174612. URL: https://doi.org/10.1109/TIT.2011.2174612.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. "Does data interpolation contradict statistical optimality?" In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, Apr. 2019, pp. 1611–1619. URL: https://proceedings.mlr.press/v89/belkin19a.html.

[BS04] Z. D. Bai and Jack W. Silverstein. "CLT for linear spectral statistics of large-dimensional sample covariance matrices". In: *The Annals of Probability* 32.1A (2004), pp. 553–605. DOI: 10.1214/aop/1078415845. URL: https://doi.org/10.1214/aop/1078415845.

[Büh11] Peter Bühlmann. *Statistics for High-Dimensional Data :Methods, Theory and Applications*. eng. 1st ed. 2011. Springer Series in Statistics. Berlin, Heidelberg: Springer Berlin Heidelberg : Imprint: Springer, 2011. ISBN: 3-642-20192-X.

[Cha22] J.-R. Chazottes. *Notes on means, medians and Gaussian tails*. 2022. eprint: https://hal.science/hal-03636138v1/file/Gaussian-concentration-around-the-mean-and-the-median.pdf. URL: https://hal.science/hal-03636138v1/file/Gaussian-concentration-around-the-mean-and-the-median.pdf.

[CM22] Chen Cheng and Andrea Montanari. *Dimension free ridge regression*. 2022. arXiv: 2210.08571 [math.ST].

[CT07] Emmanuel Candes and Terence Tao. "The Dantzig Selector: Statistical Estimation When p Is Much Larger than n". eng. In: *The Annals of statistics* 35.6 (2007), pp. 2313–2351. ISSN: 0090-5364.

[Dan+23] Yatin Dandi et al. *How Two-Layer Neural Networks Learn, One (Giant) Step at a Time*. 2023. arXiv: 2305.18270 [stat.ML].

[Dic16] Lee H. Dicker. “Ridge regression and asymptotic minimax estimation over spheres of growing dimension”. In: *Bernoulli* 22.1 (Feb. 2016). ISSN: 1350-7265. DOI: 10.3150/14-bej609. URL: http://dx.doi.org/10.3150/14-BEJ609.

[DL20] Oussama Dhifallah and Yue M. Lu. *A Precise Performance Analysis of Learning with Random Features*. 2020. arXiv: 2008.11904 [cs.IT].

[DSL23] Rishabh Dudeja, Subhabrata Sen, and Yue M. Lu. *Spectral Universality of Regularized Linear Regression with Nearly Deterministic Sensing Matrices*. 2023. arXiv: 2208.02753 [cs.IT].

[DW18] Edgar Dobriban and Stefan Wager. “High-dimensional asymptotics of prediction: ridge regression and classification”. In: *Ann. Statist.* 46.1 (2018), pp. 247–279. ISSN: 0090-5364,2168-8966. DOI: 10.1214/17-AOS1549. URL: https://doi.org/10.1214/17-AOS1549.

[Fan22] Zhou Fan. “Approximate Message Passing algorithms for rotationally invariant matrices”. eng. In: *The Annals of statistics* 50.1 (2022), p. 197. ISSN: 0090-5364.

[For+21] Pierre Foret et al. *Sharpness-Aware Minimization for Efficiently Improving Generalization*. 2021. arXiv: 2010.01412 [cs.LG].

[GHW79] Gene H. Golub, Michael Heath, and Grace Wahba. “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter”. In: *Technometrics* 21.2 (1979), pp. 215–223. ISSN: 00401706. URL: http://www.jstor.org/stable/1268518 (visited on 03/16/2024).

[Gro21] Klaus Grobys. “What do we know about the second moment of financial markets?” In: *International Review of Financial Analysis* 78 (2021), p. 101891. ISSN: 1057-5219. DOI: https://doi.org/10.1016/j.irfa.2021.101891. URL: https://www.sciencedirect.com/science/article/pii/S1057521921002180.

[Has+22] Trevor Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *Ann. Statist.* 50.2 (2022), pp. 949–986. ISSN: 0090-5364,2168-8966. DOI: 10.1214/21-aos2133. URL: https://doi.org/10.1214/21-aos2133.

[Has15] Trevor Hastie. *Statistical learning with sparsity : the lasso and generalizations*. eng. 1st. Monographs on statistics and applied probability (Series) ; 143. Boca Raton: CRC Press, 2015. ISBN: 0-429-17158-7.

[HL20] Hong Hu and Yue M. Lu. “Universality Laws for High-Dimensional Learning With Random Features”. In: *IEEE Transactions on Information Theory* 69 (2020), pp. 1932–1964. URL: https://api.semanticscholar.org/CorpusID:221738950.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[IR18] Noor Salwani Ibrahim and Dzati Athiar Ramli. “I-vector extraction for speaker recognition based on dimensionality reduction”. In: *Procedia Computer Science* 126 (2018), pp. 1534–1540.

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. “The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization”. In: *J. Mach. Learn. Res.* 21 (2020), Paper No. 169, 16. ISSN: 1532-4435,1533-7928.

[Le] Minh-Anh Le. *OpenML, Dataset ID: 40910.* https://www.openml.org/search?type=data&status=active&id=40910. Accessed: 2023-09-01.

[Loo+20] Marco Loog et al. “A brief prehistory of double descent”. In: *Proceedings of the National Academy of Sciences* 117.20 (2020), pp. 10625–10626. DOI: 10.1073/pnas.2001875117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2001875117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2001875117.

[LP09] A. Lytova and L. Pastur. “Central limit theorem for linear eigenvalue statistics of random matrices with independent entries”. In: *The Annals of Probability* 37.5 (Sept. 2009). ISSN: 0091-1798. DOI: 10.1214/09-aop452. URL: http://dx.doi.org/10.1214/09-AOP452.

[LS23] Yufan Li and Pragya Sur. *Spectrum-Aware Adjustment: A New Debiasing Framework with Applications to Principal Component Regression*. 2023. arXiv: 2309.07810 [math.ST].

[Mec19] Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge Tracts in Mathematics. Cambridge University Press, 2019.

[MG21] Gabriel C. Mel and Surya Ganguli. “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: *International Conference on Machine Learning*. 2021. URL: https://api.semanticscholar.org/CorpusID:235825881.

[MM19] Song Mei and Andrea Montanari. "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve". In: *Communications on Pure and Applied Mathematics* 75 (2019). URL: https://api.semanticscholar.org/CorpusID:199668852.

[Mon+23] Andrea Montanari et al. *The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime*. 2023. arXiv: 1911.01544 [math.ST].

[Mon+24] Behrad Moniri et al. *A Theory of Non-Linear Feature Learning with One Gradient Step in Two-Layer Neural Networks*. 2024. arXiv: 2310.07891 [stat.ML].

[MP22] Gabriel Mel and Jeffrey Pennington. "Anisotropic Random Feature Regression in High Dimensions". In: *International Conference on Learning Representations*. 2022. URL: https://api.semanticscholar.org/CorpusID:251649107.

[Nak+21] Preetum Nakkiran et al. "Deep double descent: where bigger models and more data hurt*". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (Dec. 2021), p. 124003. DOI: 10.1088/1742-5468/ac3a74. URL: https://dx.doi.org/10.1088/1742-5468/ac3a74.

[Nea+19] Brady Neal et al. *A Modern Take on the Bias-Variance Tradeoff in Neural Networks*. 2019. URL: https://openreview.net/forum?id=HkgmzhC5F7.

[Por88] Stephen Portnoy. "Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity". eng. In: *The Annals of statistics* 16.1 (1988), pp. 356–366. ISSN: 0090-5364.

[RM18] Kamiar Rahnama Rad and Arian Maleki. "A scalable estimate of the extra-sample prediction error via approximate leave-one-out". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (Jan. 2018). DOI: 10.1111/rssb.12374.

[RMR20] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. "Asymptotics of Ridge(less) Regression under General Source Condition". In: *International Conference on Artificial Intelligence and Statistics*. 2020. URL: https://api.semanticscholar.org/CorpusID:219573554.

[RSF19] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. "Vector Approximate Message Passing". In: *IEEE Transactions on Information Theory* 65 (2019), pp. 6664–6684. URL: https://api.semanticscholar.org/CorpusID:182554592.

[SC19] Pragya Sur and Emmanuel J. Candès. "A modern maximum-likelihood theory for high-dimensional logistic regression". In: *Proceedings of the National Academy of Sciences* 116.29 (2019), pp. 14516–14525. DOI: 10.1073/pnas.1810420116. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1810420116. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1810420116.

[SCC19] Pragya Sur, Yuxin Chen, and Emmanuel J. Candès. "The likelihood ratio test in high-dimensional logistic regression is asymptotically a *rescaled* chi-square". In: *Probab. Theory Related Fields* 175.1-2 (2019), pp. 487–558. ISSN: 0178-8051,1432-2064. DOI: 10.1007/s00440-018-00896-9. URL: https://doi.org/10.1007/s00440-018-00896-9.

[Sch14] Gideon Schechtman. *Concentration, results and applications*. 2014. eprint: https://www.weizmann.ac.il/math/gideon/sites/math.gideon/files/uploads/concentrationNov19_0.pdf. URL: https://www.weizmann.ac.il/math/gideon/sites/math.gideon/files/uploads/concentrationNov19_0.pdf.

[Sil95] J.W. Silverstein. "Strong Convergence of the Empirical Distribution of Eigenvalues of Large Dimensional Random Matrices". In: *Journal of Multivariate Analysis* 55.2 (1995), pp. 331–339. ISSN: 0047-259X. DOI: https://doi.org/10.1006/jmva.1995.1083. URL: https://www.sciencedirect.com/science/article/pii/S0047259X85710834.

[Van+13] Joaquin Vanschoren et al. "OpenML: Networked Science in Machine Learning". In: *SIGKDD Explorations* 15.2 (2013), pp. 49–60. DOI: 10.1145/2641190.2641198. URL: http://doi.acm.org/10.1145/2641190.2641198.

[Wan+18] Shuaiwen Wang et al. "Approximate Leave-One-Out for High-Dimensional Non-Differentiable Learning Problems". In: *arXiv e-prints*, arXiv:1810.02716 (Oct. 2018), arXiv:1810.02716. DOI: 10.48550/arXiv.1810.02716. arXiv: 1810.02716 [cs.LG].

[WX20] Denny Wu and Ji Xu. *On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression*. 2020. arXiv: 2006.05800 [stat.ML].

[YSJ19] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. "Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[Zha+17] Chiyuan Zhang et al. “Understanding deep learning requires rethinking generalization”. In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy8gdB9xx.