Surprises in High-Dimensional Ridgeless Least Squares Interpolation

Trevor Hastie Andrea Montanari* Saharon Rosset Ryan J. Tibshirani*

Abstract

Interpolators—estimators that achieve zero training error—have attracted growing attention in machine learning, mainly because state-of-the art neural networks appear to be models of this type. In this paper, we study minimum ℓ_2 norm ("ridgeless") interpolation in high-dimensional least squares regression. We consider two different models for the feature distribution: a linear model, where the feature vectors $x_i \in \mathbb{R}^p$ are obtained by applying a linear transform to a vector of i.i.d. entries, $x_i = \Sigma^{1/2} z_i$ (with $z_i \in \mathbb{R}^p$); and a nonlinear model, where the feature vectors are obtained by passing the input through a random one-layer neural network, $x_i = \varphi(W z_i)$ (with $z_i \in \mathbb{R}^d$, $W \in \mathbb{R}^{p \times d}$ a matrix of i.i.d. entries, and φ an activation function acting componentwise on $W z_i$). We recover—in a precise quantitative way—several phenomena that have been observed in large-scale neural networks and kernel machines, including the "double descent" behavior of the prediction risk, and the potential benefits of overparametrization.

1 Introduction

Modern deep learning models involve a huge number of parameters. In nearly all applications of these models, current practice suggests that we should design the network to be sufficiently complex so that the model (as trained, typically, by gradient descent) interpolates the data, i.e., achieves zero training error. Indeed, in a thought-provoking experiment, Zhang et al. (2016) showed that state-of-the-art deep neural network architectures are complex enough that they can be trained to interpolate the data even when the actual labels are replaced by entirely random ones.

Despite their enormous complexity, deep neural networks are frequently observed to generalize well in practice. At first sight, this seems to defy conventional statistical wisdom: interpolation (vanishing training error) is commonly taken to be a proxy for overfitting or poor generalization (large gap between training and test error). In an insightful series of papers, Belkin et al. (2018b,c,a) pointed out that these concepts are in general distinct, and interpolation does not contradict generalization. For example, estimation in reproducing kernel Hilbert spaces (via kernel ridge regression) is a well-understood setting in which interpolation can coexist with good generalization (Liang and Rakhlin, 2018).

In this paper, we examine the prediction risk of minimum ℓ_2 norm or "ridgeless" least squares regression, under different models for the features. A skeptical reader might ask what least squares has to do with neural networks. To motivate our study, we appeal to line of work that draws a concrete connection between the two settings (Jacot et al., 2018; Du et al., 2018b,a; Allen-Zhu et al., 2018). Following Chizat and Bach (2018b), suppose that we have a nonlinear model $\mathbb{E}(y_i|z_i) = f(z_i;\theta)$ that relates responses $y_i \in \mathbb{R}$ to inputs $z_i \in \mathbb{R}^d$, $i = 1, \dots, n$, via a parameter vector $\theta \in \mathbb{R}^p$ (while we have in mind a neural network, the setting here is actually quite general). In some problems, the number of parameters p is so large that training effectively moves each of them just a small amount with respect to some random initialization $\theta_0 \in \mathbb{R}^p$. It thus makes sense to linearize the model around θ_0. Further, supposing that the initialization is such that $f(z;\theta_0) \approx 0$, and letting $\theta = \theta_0 + \beta$, we obtain

$$\mathbb{E}(y_i|z_i) \approx \nabla_\theta f(z_i;\theta_0)^T \beta, \quad i = 1, \dots, n. \tag{1}$$

We are therefore led to consider a linear regression problem, with random features $x_i = \nabla_\theta f(z_i;\theta_0)$, $i = 1, \dots, n$, of high-dimensionality (p much greater than n). Notice that the features are random because of the initialization. In this setting, many vectors β give rise to a model that interpolates the data. However, using gradient descent on the least squares objective for training yields a special interpolating parameter $\hat{\beta}$ (having implicit regularity): the least squares solution with minimum ℓ_2 norm.

We consider two different models for the features.

*Corresponding authors.

- **Linear model.** Here each $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ has i.i.d. entries with zero mean and unit variance and $\Sigma \in \mathbb{R}^{p \times p}$ deterministic and positive definite. This is a standard model in random matrix theory and we leverage known results from that literature to obtain a fairly complete asymptotic characterization of the out-of-sample prediction risk. In particular, we recover a number of phenomena that have been observed in kernel regression and neural networks (summary in the next subsection).
- **Nonlinear model.** Here each $x_i = \varphi(W z_i)$, where $z_i \in \mathbb{R}^d$ has i.i.d. entries from $N(0,1)$, $W \in \mathbb{R}^{p \times d}$ has i.i.d. entries from $N(0, 1/d)$, and φ is an activation function acting componentwise. This corresponds to a two-layer neural network with random first layer weights and second layer weights given by the regression coefficients β. This model was introduced by Rahimi and Recht (2008) as a randomized approach for scaling kernel methods to large datasets, and relative to the linear model described above, it is far less studied in the random matrix theory literature. We prove a new asymptotic result that allows us to characterize the prediction variance in the nonlinear model. This result is remarkably close to the analogous result for the linear model (and in some cases, identical to it), despite the fact that d can be much smaller than p.

Because of its simplicity, the linear model abstracts away some interesting properties of the model (1), specifically, the distinction between the input dimension d and the number of parameters p. On the other hand, also due to this simplicity, we are able to develop a fairly complete picture of the asymptotic risk (elucidating the effect of correlations between the features, signal-to-noise ratio, strength of model misspecification, etc.). The fact that the two models give closely related results (for the variance component of the risk) supports the hope that the linear model can still provide useful insights into the model (1).

1.1 Summary of results

In what follows, we analyze the out-of-sample prediction risk of the minimum ℓ_2 norm (or min-norm, for short) least squares estimator, in an asymptotic setup where both the number of samples and features diverge, $n, p \to \infty$, and their ratio converges to a nonzero constant, $p/n \to \gamma \in (0, \infty)$. When $\gamma < 1$, we call the problem *underparametrized*, and when $\gamma > 1$, we call it *overparametrized*. Below we summarize our main results on the asymptotic risk. Denote by β, Σ the underlying signal and feature covariance matrix, respectively, and by SNR the signal-to-noise ratio (defined precisely in Section 3.3). We note that all but the last three points below pertain to the linear model. Also, see Figure 1 for a supporting plot of the asymptotic risk curves for different cases of interest.

0. In the underparametrized regime ($\gamma < 1$), the risk is purely variance (there is no bias), and does not depend on β, Σ (see Theorem 1). Moreover, the risk diverges as we approach the interpolation boundary (as $\gamma \to 1$).
1. In the overparametrized regime ($\gamma > 1$), the risk is composed of both bias and variance, and generally depends on β, Σ (see Theorem 3). In each of two different models for the feature covariance Σ, we find that the bias is decreasing with the strength of correlation, and the variance is nondecreasing (see Corollary 2, Appendix A.5).
2. When $\mathrm{SNR} \leq 1$, the risk is decreasing for $\gamma \in (1, \infty)$, and approaches the null risk (from above) as $\gamma \to \infty$ (see Section 3.3).
3. When $\mathrm{SNR} > 1$, the risk has a *local* minimum on $\gamma \in (1, \infty)$, is better than the null risk for large enough γ, and approaches the null risk (from below) as $\gamma \to \infty$ (see Section 3.3).
4. For a misspecified model, when $\mathrm{SNR} > 1$, the risk can attain its *global* minimum on $\gamma \in (1, \infty)$ (when there is strong enough approximation bias, see Section 5.3).
5. Optimally-tuned ridge regression dominates the min-norm least squares estimator in risk, across all values of γ and SNR, in both the well-specified and misspecified settings. For a misspecified model, optimally-tuned ridge regression attains its global minimum around $\gamma = 1$ (see Section 6).
6. Tuning ridge regression by minimizing the leave-one-out cross-validation (CV) error is asymptotically optimal, i.e., results in the optimal risk (see Theorem 6).
7. For the nonlinear model, as $n, p, d \to \infty$, with $p/n \to \gamma \in (0, \infty)$ and $d/p \to \psi \in (0, 1)$, the limiting variance is increasing for $\gamma \in (0, 1)$, decreasing for $\gamma \in (1, \infty)$, and diverging as $\gamma \to 1$, confirming our general picture. In fact, in the underparametrized regime ($\gamma < 1$), the variance coincides exactly with the one in the linear model case (see Theorem 7).

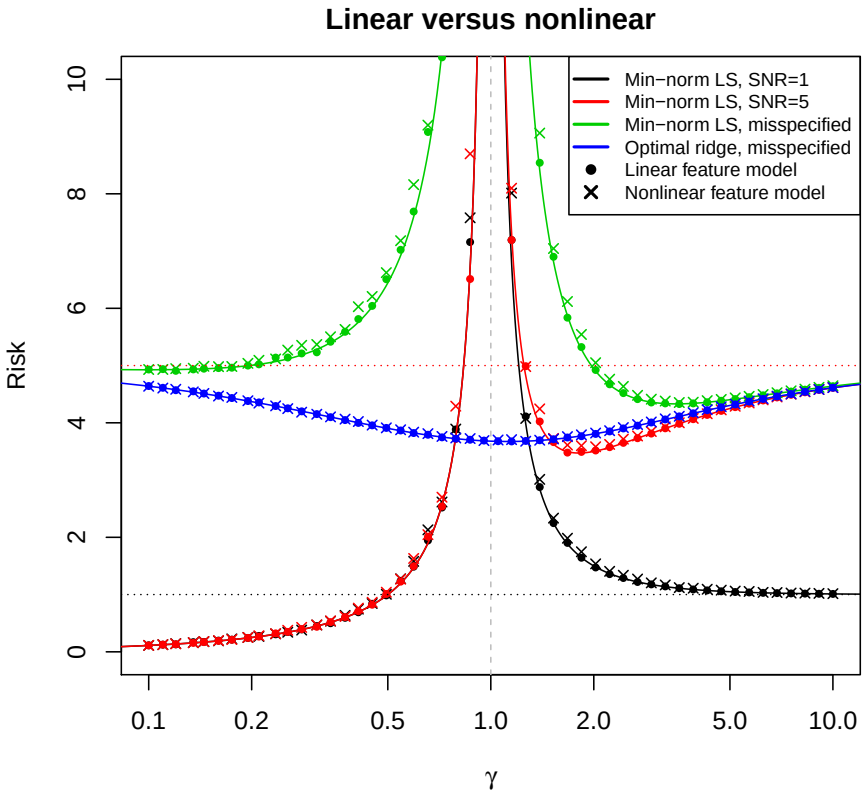

Figure 1: Asymptotic risk curves for the linear feature model, as a function of the limiting aspect ratio γ. The risks for min-norm least squares, when $\mathrm{SNR} = 1$ and $\mathrm{SNR} = 5$, are plotted in black and red, respectively. These two match for $\gamma < 1$ but differ for $\gamma > 1$. The null risks for $\mathrm{SNR} = 1$ and $\mathrm{SNR} = 5$ are marked by the dotted black and red lines, respectively. The risk for the case of a misspecified model (with significant approximation bias, $a = 1.5$ in (13)), when $\mathrm{SNR} = 5$, is plotted in green. Optimally-tuned (equivalently, CV-tuned) ridge regression, in the same misspecified setup, has risk plotted in blue. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries. Meanwhile, the "x" points mark finite-sample risks for a nonlinear feature model, with $n = 200$, $p = [\gamma n]$, $d = 100$, and $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries, W has i.i.d. $N(0, 1/d)$ entries, and $\varphi(t) = a(|t| - b)$ is a "purely nonlinear" activation function, for constants a, b. The theory predicts that this nonlinear risk should converge to the linear risk with p features (regardless of d). The empirical agreement between these two—and the agreement in finite-sample and asymptotic risks—is striking.

8. In the overparametrized regime ($\gamma > 1$), the asymptotic formula for the variance in the nonlinear model depends only on the activation function φ via its "linear component" $\mathbb{E}[G\varphi(G)]$ (where $G \sim N(0, 1)$, and we assume the normalization $\mathbb{E}[\varphi(G)^2] = 1$). This points to a remarkable degree of universality.

9. In the overparametrized regime ($\gamma > 1$), for a "purely nonlinear" activation function φ, meaning $\mathbb{E}[G\varphi(G)] = 0$ (where again $G \sim N(0, 1)$), the asymptotic formula for the variance coincides with the one derived for the linear model case, despite the fact that the input dimension d can be much smaller than the number of parameters p. In this case, the bias also matches to the one for the linear model case (see Corollary 4).

A few remarks are in order. We number the first point above (which provides important context for the points that follow) as 0, because it is a known result that has appeared in various places in the random matrix theory literature. Points 3 through 6 are formally established for isotropic features, $\Sigma = I$, but qualitatively similar behavior holds for general Σ. Several of the arguments in this paper rely on more or less standard results in random matrix theory; though the mathematics is standard, the insights, we believe, are new. An important distinction are the nonlinear model results, which are not covered by existing literature on random matrix theory. In this setting, we derive a new asymptotic result on resolvents of certain block matrices, which may be of independent interest (see Theorem 8). This allows us to get asymptotically *exact* expressions for the prediction risk.

1.2 Intuition and implications

We discuss some intuition behind and implications of our results.

Bias and variance. The shape of the asymptotic risk curve for min-norm least squares is, of course, controlled by its components: bias and variance. In the overparametrized regime, the bias increases with γ, which is intuitive. When $p > n$, the min-norm least squares estimate of β is constrained to lie the row space of X, the training feature matrix. This is a subspace of dimension n lying in a feature space of dimension p. Thus as p increases, so does the bias, since this row space accounts for less and less of the ambient p-dimensional feature space.

Meanwhile, in the overparametrized regime, the variance *decreases* with γ. This may seem counterintuitive at first, because it says, in a sense, that the min-norm least squares estimator becomes *more* regularized as p grows. However, this too can be explained intuitively, as follows. As p grows, the minimum ℓ_2 norm least squares solution—i.e., the minimum ℓ_2 norm solution to the linear system $Xb = y$, for a training feature matrix X and response vector y—will generally have decreasing ℓ_2 norm. Why? Compare two such linear systems: in each, we are asking for the min-norm solution to a linear system with the same y, but in one instance we are given more columns in X, so we can generally decrease the components of b (by distributing them over more columns), and achieve a smaller ℓ_2 norm. This can in fact be formalized asymptotically, see Corollaries 1 and 3.

Double descent. Recently, Belkin et al. (2018a) pointed out a fascinating empirical trend where, for popular methods like neural networks and random forests, we can see a *second* bias-variance tradeoff in the out-of-sample prediction risk beyond the interpolation limit. The risk curve here resembles a traditional U-shape curve before the interpolation limit, and then descends again beyond the interpolation limit, which these authors call "double descent". (A closely related phenomenon was found earlier by Spigler et al. (2018), who studied the "jamming transition" from underparametrized to overparametrized neural networks.) Our results formally verify that this double descent phenomenon occurs even in the simple and fundamental case of least squares regression. The appearance of the second descent in the risk, past the interpolation boundary ($\gamma = 1$), is explained by the fact that the variance decreases as γ grows, as discussed above.

In the misspecified case, the variance still decreases with γ (for the same reasons), but interestingly, the bias can now also decrease with γ, provided γ is not too large (not too far past the interpolation boundary). The intuition here is that in a misspecified model, some part of the true regression function is always unaccounted for, and adding features generally improves our approximation capacity. Our results show that this double descent phenomenon can be even more pronounced in the misspecified case (depending on the strength of the approximation bias), and that the risk can attain its global minimum past the interpolation limit.

In-sample prediction risk. Our focus throughout this paper is out-of-sample prediction risk. It is reasonable to ask how the results would change if we instead look at in-sample prediction risk. In the data model (2), (3) we study, the in-sample prediction risk of the min-norm least squares estimator $\hat{\beta}$ is $\mathbb{E}[\|X\hat{\beta} - X\beta\|_2^2/n \,|\, X] = \sigma^2(p/n \wedge 1)$ (where we abbreviate $a \wedge b = \min\{a, b\}$, and we are assuming that $\mathrm{rank}(X) = n \wedge p$). The asymptotic in-sample prediction risk, as $p/n \to \gamma$, is therefore just $\sigma^2(\gamma \wedge 1)$. Compare this to the much richer and more complex behavior exhibited by the limiting out-of-sample risk (see the curves in Figure 1, or (8), (14) for the precise mathematical forms in the well-specified and misspecified settings, respectively): their behaviors could not be more different. This serves as an important reminder that the former (in-sample prediction risk) is not always a good proxy for the latter (out-of-sample prediction risk). Although much of classical regression theory is based on the former (e.g., optimism, effective degrees of freedom, and covariance penalties), the latter is more broadly relevant to practice.

Interpolation versus regularization. The min-norm least squares estimator can be seen as the limit of ridge regression as the tuning parameter tends to zero. It is also the convergence point of gradient descent run on the least squares loss. We would not in general expect the best-predicting ridge solution to be at the end of its regularization path. Our results, comparing min-norm least squares to optimally-tuned ridge regression, show that (asymptotically) this is never the case, and dramatically so near $\gamma = 1$. It is worth noting that early stopped gradient descent is known to be closely connected to ridge regularization, see, e.g., Friedman and Popescu (2004); Ramsay (2005); Yao et al. (2007); Raskutti et al. (2014); Wei et al. (2017). In fact, a tight coupling between the two has been recently developed for least squares problems in Ali et al. (2019). This coupling implies that that optimally-stopped gradient descent will have risk at most 1.22 times that of optimally-tuned ridge regression, and hence will often have better risk than min-norm least squares. In practice, of course, we would not have access to the optimal tuning parameter for ridge (optimal stopping for gradient

descent), and we would rely on, e.g., cross-validation (CV). Our theory shows that for ridge regression, CV tuning is asymptotically equivalent to optimal tuning (and we would expect the same results to carry over to gradient descent, but have not pursued this formally).

Historically, the debate between interpolation and regularization has been alive for the last 30 or so years. Support vector machines find maximum-margin decision boundaries, which often perform very well for problems where the Bayes error is close to zero. But for less-separated classification tasks, one needs to tune the cost parameter (Hastie et al., 2004). Relatedly, in classification, it is common to run boosting until the training error is zero, and similar to the connection between gradient descent and ℓ_2 regularization, the boosting path is tied to ℓ_1 regularization (Rosset et al., 2004; Tibshirani, 2015). Again, we now know that boosting can overfit, and the number of boosting iterations should be treated as a tuning parameter.

Virtues of nonlinearity. Stochastic gradient descent is the method of choice in the deep learning community, where it is often run until zero training error. If we postulate that the weights do not change much during training, then the linearization (1) could be accurate, and training a deep learning model with squared error loss by gradient descent is akin to finding the min-norm least squares solution with respect to nonlinear random features, which recall we model as $x_i = \varphi(Wz_i)$, where W plays the role of the random weights initialization and z_i is an input feature vector. A priori, it is unclear whether our success in precisely analyzing the linear model setting should carry over to the nonlinear setting. Remarkably, under high-dimensional asymptotics with $n, p, d \to \infty$, $p/n \to \gamma$, and $d/p \to \psi$, this turns out to be the case. Even more surprising, the relevant dimensions-to-samples ratio is given by p/n (not d/n): for "purely nonlinear" activations φ, the results we derived in the linear model setting with p features remain asymptotically exact. In other words, each component $x_{ij} = \varphi((Wx_i)_j)$ of the feature vector behaves "as if" it was independent of the others, even when d is much smaller than p.

Finally, although we believe our study in this paper is certainly relevant to understanding overparametrized neural networks, we note that some caution must be taken in translating between the two problem settings. There is a critical difference to be made clear: in a neural network, the feature representation and the regression function or classifier are learned *simultaneously*. In both our linear and nonlinear model settings, the features X are not learned, but observed. Learning X could significantly change some aspects of the behavior of an interpolator. (See for instance Chapter 9 of Goodfellow et al. (2016), and also Chizat and Bach (2018b); Zhang et al. (2019), which emphasize the importance of learning the representation.)

1.3 Related work

The present work connects to and is motivated by the recent interest in interpolators in machine learning (Belkin et al., 2018b,a; Liang and Rakhlin, 2018; Belkin et al., 2018c; Geiger et al., 2019) Several auhors have argued that minimum ℓ_2 norm least squares regression captures the basic behavior of deep neural networks, at least in early (lazy) training (Jacot et al., 2018; Du et al., 2018b,a; Allen-Zhu et al., 2018; Zou et al., 2018; Chizat and Bach, 2018b; Lee et al., 2019). The connection between neural networks and kernel ridge regression arises when the number of hidden units diverges. The same limit was also studied (beyond the linearized regime) by Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2018); Chizat and Bach (2018a).

For the linear model, our risk result in the general Σ case basically follows by taking a limit (as the ridge tuning parameter tends to zero) in the ridge regression result of Dobriban and Wager (2018). For the $\Sigma = I$ case, our analysis bears similarities to the ridge regression analysis of Dicker (2016) (although we manage to avoid the assumption of Gaussianity of the features, by invoking a generalized Marchenko-Pastur theorem). Furthermore, our discussion of min-norm least squares versus ridge regression is somewhat related to the "regimes of learning" problem studied by Liang and Srebro (2010); Dobriban and Wager (2018).

For the nonlinear model, the random matrix theory literature is much sparser, and focuses on the related model of kernel random matrices, namely, symmetric matrices of the form $K_{ij} = \varphi(z_i^T z_j)$. El Karoui (2010) studied the spectrum of such matrices in a regime in which φ can be approximated by a linear function (for $i \neq j$) and hence the spectrum converges to a rescaled Marchenko-Pastur law. This approximation does not hold for the regime of interest here, which was studied instead by Cheng and Singer (2013) (who determined the limiting spectral distribution) and Fan and Montanari (2015) (who characterized the extreme eigenvalues). The resulting eigenvalue distribution is the free convolution of a semicircle law and a Marchenko-Pastur law. In the current paper, we must consider asymmetric (rectangular) matrices $x_{ij} = \varphi(w_j^T x_i)$, whose singular value distribution was recently computed by Pennington and Worah (2017), using the moment method. Unfortunately, the prediction variance depends on both the singular values

and vectors of this matrix. In order to address this issue, we apply the leave-one out method of Cheng and Singer (2013) to compute the asymptotics of the resolvent of a suitably extended matrix. We then extract the information of interest from this matrix.

After completing this paper, we became aware of the highly-related (earlier) work of Advani and Saxe (2017) and (concurrent) work of Belkin et al. (2019). Both papers study the risk of min-norm least squares regression; the latter is focused on finite-sample analysis in a specalized setting where the response and features are jointly Gaussian, and relies on properties of Wishart matrices; the former is focused on asymptotic analysis in a broader setting, and utilizes random matrix theory, just as in our paper. Advani and Saxe (2017) also consider deep linear networks and shallow nonlinear networks. Many of the conclusions drawn in Advani and Saxe (2017); Belkin et al. (2019) (especially the former) are qualitatively similar to ours, but there are differences in the details and scope. Our analysis of the linear model case is broader than that in Advani and Saxe (2017); Belkin et al. (2019), and we are able to give precise results in the nonlinear case as well, which is not done in these papers.

1.4 Outline

Section 2 provides important background. Sections 3–7 consider the linear model case, focusing on isotropic features, correlated features, misspecified models, ridge regularization, and cross-validation, respectively. Section 8 covers the nonlinear model case. Nearly all proofs are deferred until the appendix.

2 Preliminaries

We describe our setup and gather a number of important preliminary results.

2.1 Data model and risk

Assume we observe training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ from a model

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon, \quad i = 1, \ldots, n, \tag{2}$$

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where the random draws across $i = 1, \ldots, n$ are independent. Here, P_x is a distribution on $\mathbb{R}^p$ such that $\mathbb{E}(x_i) = 0$, $\mathrm{Cov}(x_i) = \Sigma$, and P_ϵ is a distribution on $\mathbb{R}$ such that $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$. We collect the responses in a vector $y \in \mathbb{R}^n$, and the features in a matrix $X \in \mathbb{R}^{n \times p}$ (with rows $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$).

Consider a test point $x_0 \sim P_x$, independent of the training data. For an estimator $\hat{\beta}$ (a function of the training data X, y), we define its out-of-sample prediction risk (or simply, risk) as

$$R_X(\hat{\beta}; \beta) = \mathbb{E}\big[(x_0^T \hat{\beta} - x_0^T \beta)^2 \,|\, X\big] = \mathbb{E}\big[\|\hat{\beta} - \beta\|_\Sigma^2 \,|\, X\big],$$

where $\|x\|_\Sigma^2 = x^T \Sigma x$. Note that our definition of risk is conditional on X (as emphasized by our notation R_X). Note also that we have the bias-variance decomposition

$$R_X(\hat{\beta}; \beta) = \underbrace{\|\mathbb{E}(\hat{\beta}|X) - \beta\|_\Sigma^2}_{B_X(\hat{\beta};\beta)} + \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\beta}|X)\Sigma]}_{V_X(\hat{\beta};\beta)}.$$

2.2 Ridgeless least squares

Consider the minimum ℓ_2 norm (min-norm) least squares regression estimator, of y on X, defined by

$$\hat{\beta} = (X^T X)^+ X^T y, \tag{4}$$

where $(X^T X)^+$ denotes the Moore-Penrose pseudoinverse of $X^T X$. Equivalently, we can write

$$\hat{\beta} = \arg\min \Big\{ \|b\|_2 : b \text{ minimizes } \|y - Xb\|_2^2 \Big\},$$

which justifies its name. An alternative name for (4) is the "ridgeless" least squares estimator, motivated by the fact that $\hat{\beta} = \lim_{\lambda \to 0^+} \hat{\beta}_\lambda$, where $\hat{\beta}_\lambda$ denotes the ridge regression estimator,

$$\hat{\beta}_\lambda = (X^T X + n\lambda I)^{-1} X^T y, \tag{5}$$

which we can equivalently write as

$$\hat{\beta}_\lambda = \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

When X has full column rank (equivalently, when $X^T X$ is invertible), the min-norm least squares estimator reduces to $\hat{\beta} = (X^T X)^{-1} X^T y$, the usual least squares estimator. When X has rank n, importantly, this estimator interpolates the training data: $y_i = x_i^T \hat{\beta}$, for $i = 1, \ldots, n$.

Lastly, the following is a well-known fact that connects the min-norm least squares solution to gradient descent (as referenced in the introduction).

Proposition 1. *Initialize $\beta^{(0)} = 0$, and consider running gradient descent on the least squares loss, yielding iterates*

$$\beta^{(k)} = \beta^{(k-1)} + t X^T (y - X\beta^{(k-1)}), \quad k = 1, 2, 3, \ldots,$$

where we take $0 < t \leq 1/\lambda_{\max}(X^T X)$ (and $\lambda_{\max}(X^T X)$ is the largest eigenvalue of $X^T X$). Then $\lim_{k \to \infty} \beta^{(k)} = \hat{\beta}$, the min-norm least squares solution in (4).

Proof. The choice of step size guarantees that $\beta^{(k)}$ converges to a least squares solution as $k \to \infty$, call it $\tilde{\beta}$. Note that $\beta^{(k)}$, $k = 1, 2, 3, \ldots$ all lie in the row space of X; therefore $\tilde{\beta}$ must also lie in the row space of X; and the min-norm least squares solution $\hat{\beta}$ is the unique least squares solution with this property. □

2.3 Bias and variance

We recall expressions for the bias and variance of the min-norm least squares estimator, which are standard.

Lemma 1. *Under the model (2), (3), the min-norm least squares estimator (4) has bias and variance*

$$B_X(\hat{\beta}; \beta) = \beta^T \Pi \Sigma \Pi \beta \quad \text{and} \quad V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n} \mathrm{tr}(\hat{\Sigma}^+ \Sigma),$$

where $\hat{\Sigma} = X^T X/n$ is the (uncentered) sample covariance of X, and $\Pi = I - \hat{\Sigma}^+ \hat{\Sigma}$ is the projection onto the null space of X.

Proof. As $\mathbb{E}(\hat{\beta}|X) = (X^T X)^+ X^T X \beta = \hat{\Sigma}^+ \hat{\Sigma} \beta$ and $\mathrm{Cov}(\hat{\beta}|X) = \sigma^2 (X^T X)^+ X^T X (X^T X)^+ = \sigma^2 \hat{\Sigma}^+ / n$, the bias and variance expressions follow from plugging these into their respective definitions. □

2.4 Underparametrized asymptotics

We consider an asymptotic setup where $n, p \to \infty$, in such a way that $p/n \to \gamma \in (0, \infty)$. Recall that when $\gamma < 1$, we call the problem underparametrized; when $\gamma > 1$, we call it overparametrized. Here, we recall the risk of the min-norm least squares estimator in the underparametrized case. The rest of this paper focuses on the overparametrized case.

The following is a known result in random matrix theory, and can be found in Chapter 6 of Serdobolskii (2007), where the author traces it back to work by Girko and Serdobolskii from the 1990s through early 2000s. It can also be found in the wireless communications literature (albeit with minor changes in the presentation and conditions), see Chapter 4 of Tulino and Verdu (2004), where it is traced back to work by Verdu, Tse, and others from the late 1990s through early 2000s. Before stating the result, we recall that for a symmetric matrix $A \in \mathbb{R}^{p \times p}$, we define its *spectral distribution* as $F_A(x) = (1/p) \sum_{i=1}^p 1\{\lambda_i(A) \leq x\}$, where $\lambda_i(A)$, $i = 1, \ldots, p$ are the eigenvalues of A.

Theorem 1. *Assume the model (2), (3), and assume $x \sim P_x$ is of the form $x = \Sigma^{1/2} z$, where z is a random vector with i.i.d. entries that have zero mean, unit variance, and a finite 4th moment, and Σ is a deterministic positive definite matrix, such that $\lambda_{\min}(\Sigma) \geq c > 0$, for all n, p and a constant c (here $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ). As $n, p \to \infty$, assume that the spectral distribution F_Σ converges weakly to a measure H. Then as $n, p \to \infty$, such that $p/n \to \gamma < 1$, the risk of the least squares estimator (4) satisfies, almost surely,*

$$R_X(\hat{\beta}; \beta) \to \sigma^2 \frac{\gamma}{1 - \gamma}.$$

Proof. Write $X = Z\Sigma^{1/2}$. Note that

$$\lambda_{\min}(X^TX/n) \geq \lambda_{\min}(Z^TZ/n)\lambda_{\min}(\Sigma) \geq (c/2)(1-\sqrt{\gamma})^2,$$

where the second inequality holds almost surely, following from $\lambda_{\min}(\Sigma) \geq c$, and the Bai-Yin theorem (Bai and Yin, 1993), which implies that the smallest eigenvalue of Z^TZ/n is almost surely larger than $(1-\sqrt{\gamma})^2/2$ for sufficiently large n. As the right-hand side in the above display is strictly positive, we have that X^TX is almost surely invertible. Therefore by the bias and variance results from Lemma 1, we have almost surely $\Pi = 0$ and $B_X(\hat{\beta};\beta) = 0$, and also

$$\begin{aligned} V_X(\hat{\beta};\beta) &= \frac{\sigma^2}{n}\mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) \\ &= \frac{\sigma^2}{n}\mathrm{tr}\bigg(\Sigma^{-1/2}\Big(\frac{Z^TZ}{n}\Big)^{-1}\Sigma^{-1/2}\Sigma\bigg) \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{1}{s_i} \\ &= \frac{\sigma^2 p}{n}\int \frac{1}{s}\, dF_{Z^TZ/n}(s), \end{aligned}$$

where $F_{Z^TZ/n}$ is the spectral measure of Z^TZ/n. Now apply the Marchenko-Pastur theorem (Marchenko and Pastur, 1967; Silverstein, 1995), which says that $F_{Z^TZ/n}$ converges weakly, almost surely, to the Marchenko-Pastur law F_γ (depending only on γ). By the Portmanteau theorem, weak convergence is equivalent to convergence in expectation of all bounded functions h, that are continuous except on a set of zero probability under the limiting measure. Defining $h(s) = 1/s \cdot 1\{s \geq a/2\}$, where we abbreviate $a = (1-\sqrt{\gamma})^2$, it follows that as $n, p \to \infty$, almost surely,

$$\int_{a/2}^{\infty} \frac{1}{s}\, dF_{Z^TZ/n}(s) \to \int_{a/2}^{\infty} \frac{1}{s}\, dF_\gamma(s).$$

Note that we can remove the lower limit of integration on both sides above; for the right-hand side, this follows since support of the Marchenko-Pastur law F_γ is $[a,b]$, where $b = (1+\sqrt{\gamma})^2$; for the left-hand side, this follows again by the Bai-Yin theorem (which as already stated, implies the smallest eigenvalue of Z^TZ/n is almost surely greater than $a/2$ for large enough n). Thus the last display implies that as $n, p \to \infty$, almost surely,

$$V_X(\hat{\beta};\beta) \to \sigma^2\gamma \int \frac{1}{s}\, dF_\gamma(s). \tag{6}$$

It remains to compute the right-hand side above. This can be done in various ways. One approach is to recognize the right-hand side as the evaluation of the Stieltjes transform $m(z)$ of Marchenko-Pastur law at $z = 0$. Fortunately, this has an explicit form (e.g., Lemma 3.11 in Bai and Silverstein 2010), for real $z > 0$:

$$m(-z) = \frac{-(1-\gamma+z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2\gamma z}. \tag{7}$$

Since the limit as $z \to 0^+$ is indeterminate, we can use l'Hopital's rule to calculate:

$$\begin{aligned} \lim_{z\to 0^+} m(-z) &= \lim_{z\to 0^+} \frac{-1 + \frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}}}{2\gamma} \\ &= \frac{-1+\frac{1+\gamma}{1-\gamma}}{2\gamma} = \frac{1}{1-\gamma}. \end{aligned}$$

Plugging this into (6) completes the proof. □

3 Isotropic features

For the next two sections, we focus on the limiting risk of the min-norm least squares estimator when $\gamma > 1$. In the overparametrized case, an important issue that we face is that of bias: $B_X(\hat{\beta};\beta) = \beta^T\Pi\Sigma\Pi\beta$ is generally nonzero,

because Π is. We consider two approaches to analyze the limiting bias. We assume throughout that $x \sim P_x$ takes the form $x = \Sigma^{1/2} z$ for a random vector z with i.i.d. entries that have zero mean and unit variance. In the first approach, considered in this section, we assume $\Sigma = I$, in which case the limiting bias is seen to depend only on $\|\beta\|_2^2$. In the second, considered in Section 4, we allow Σ to be general but place an isotropic prior on β, in which case the limiting bias is seen to depend only on $\mathbb{E}\|\beta\|_2^2$.

3.1 Limiting bias

In the next lemma, we compute the asymptotic bias in for isotropic features, where we will see that it depends only on $r^2 = \|\beta\|_2^2$. To give some intuition as to why this is true, consider the special case where X has i.i.d. entries from $N(0,1)$. By rotational invariance, for any orthogonal $U \in \mathbb{R}^{p \times p}$, the distribution of X and XU is the same. Thus

$$\begin{aligned} B_X(\hat{\beta}; \beta) &= \beta^T \big(I - (X^T X)^+ X^T X\big)\beta \\ &\overset{d}{=} \beta^T \big(I - U^T (X^T X)^+ U U^T X^T X U\big)\beta \\ &= r^2 - (U\beta)^T (X^T X)^+ X^T X (U\beta). \end{aligned}$$

Choosing U so that $U\beta = re_i$, the ith standard basis vector, then averaging over $i = 1, \ldots, p$, yields

$$B_X(\hat{\beta}; \beta) \overset{d}{=} r^2 \big[1 - \mathrm{tr}\big((X^T X)^+ X^T X\big)/p\big] = r^2(1 - n/p).$$

As $n, p \to \infty$ with $p/n \to \gamma > 1$, we see that $B_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma)$, almost surely. As the next result shows, this is still true outside of the Gaussian case, provided the features are isotropic. The intuition is that an isotropic feature distribution, with i.i.d. components, will begin to look rotationally invariant in large samples. This is made precise by a generalized Marchenko-Pastur theorem of Rubio and Mestre (2011), and the proof of the next result is deferred to Appendix A.1.

Lemma 2. *Assume (2), (3), where $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8 + \eta$, for some $\eta > 0$. Assume that $\|\beta\|_2^2 = r^2$ for all n, p. Then for the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, such that $p/n \to \gamma > 1$, its bias satisfies, almost surely,*

$$B_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma).$$

3.2 Limiting variance

The next lemma computes the limiting variance for isotropic features. As in Theorem 1, the calculation is a more or less standard one of random matrix theory (in fact, our proof reduces the calculation to that from Theorem 1).

Lemma 3. *Assume (2), (3), where $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite 4th moment. For the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, with $p/n \to \gamma > 1$, its variance satisfies, almost surely,*

$$V_X(\hat{\beta}; \beta) \to \frac{\sigma^2}{\gamma - 1}.$$

Proof. Recalling the expression for the bias from Lemma 1 (where now $\Sigma = I$), we have

$$V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{1}{s_i},$$

where $s_i = \lambda_i(X^T X/n)$, $i = 1, \ldots, n$ are the nonzero eigenvalues of $X^T X/n$. Let $t_i = \lambda_i(XX^T/p)$, $i = 1, \ldots, p$ denote the eigenvalues of XX^T/p. Then we may write $s_i = (p/n)t_i$, $i = 1, \ldots, n$, and

$$V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{p} \sum_{i=1}^{n} \frac{1}{t_i} = \frac{\sigma^2 n}{p} \int \frac{1}{t}\, dF_{XX^T/p}(t),$$

where $F_{XX^T/p}$ is the spectral measure of XX^T/p. Now as $n/p \to \tau = 1/\gamma < 1$, we are back precisely in the setting of Theorem 1, and by the same arguments, we may conclude that almost surely

$$V_X(\hat{\beta}; \beta) \to \frac{\sigma^2 \tau}{1 - \tau} = \frac{\sigma^2}{\gamma - 1},$$

completing the proof. □

3.3 Limiting risk

Putting together Lemmas 2 and 3 leads to the following result for isotropic features.

Theorem 2. *Assume the model (2), (3), where $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8+\eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 = r^2$ for all n, p. Then for the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, such that $p/n \to \gamma > 1$, it holds almost surely that*

$$R_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma) + \frac{\sigma^2}{\gamma - 1}.$$

Now write $R(\gamma)$ for the asymptotic risk of the min-norm least squares estimator, as a function of the aspect ratio $\gamma \in (0, \infty)$. Putting together Theorems 1 and 2, we have in the isotropic case,

$$R(\gamma) = \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\big(1 - \frac{1}{\gamma}\big) + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases} \tag{8}$$

On $(0, 1)$, there is no bias, and the variance increases with γ; on $(1, \infty)$, the bias increases with γ, and the variance decreases with γ. Below we discuss some further interesting aspects of this curve. Let $\mathrm{SNR} = r^2/\sigma^2$. Observe that the risk of the null estimator $\tilde{\beta} = 0$ is r^2, which we hence call the null risk. The following facts are immediate from the form of the risk curve in (8). See Figure 2 for an accompanying plot when SNR varies from 1 to 5.

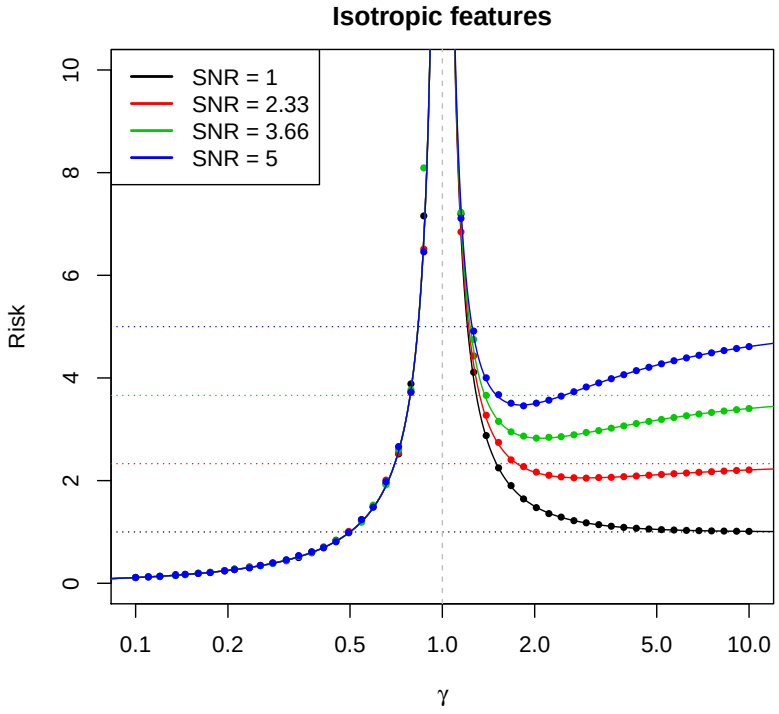

Figure 2: Asymptotic risk curves in (8) for the min-norm least squares estimator, when r^2 varies from 1 to 5, and $\sigma^2 = 1$. For each value of r^2, the null risk is marked as a dotted line, and the points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries.

1. On $(0, 1)$, the least squares risk $R(\gamma)$ is better than the null risk if and only if $\gamma < \frac{\mathrm{SNR}}{\mathrm{SNR}+1}$.
2. On $(1, \infty)$, when $\mathrm{SNR} \leq 1$, the min-norm least squares risk $R(\gamma)$ is always worse than the null risk. Moreover, it is monotonically decreasing, and approaches the null risk (from above) as $\gamma \to \infty$.
3. On $(1, \infty)$, when $\mathrm{SNR} > 1$, the min-norm least squares risk $R(\gamma)$ beats the null risk if and only if $\gamma > \frac{\mathrm{SNR}}{\mathrm{SNR}-1}$. Further, it has a local minimum at $\gamma = \frac{\sqrt{\mathrm{SNR}}}{\sqrt{\mathrm{SNR}}-1}$, and approaches the null risk (from below) as $\gamma \to \infty$.

3.4 Limiting ℓ_2 norm

Calculation of the limiting ℓ_2 norm of the min-norm least squares estimator is quite similar to the study of the limiting risk in Theorem 2, and therefore we state the next result without proof.

Corollary 1. *Assume the conditions of Theorem 2. Then as $n, p \to \infty$, such that $p/n \to \gamma$, the squared ℓ_2 norm of the min-norm least squares estimator (4) satisfies, almost surely,*

$$\mathbb{E}[\|\hat{\beta}\|_2^2 \,|\, X] \to \begin{cases} r^2 + \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2 \frac{1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

We can see that the limiting norm, as a function of γ, has a somewhat similar profile to the limiting risk in (8): it is monotonically increasing on $(0, 1)$, diverges at the interpolation boundary, and is monotonically decreasing on $(1, \infty)$.

4 Correlated features

We broaden the scope of our analysis from the last section, where we examined isotropic features. In this section, we take $x \sim P_x$ to be of the form $x = \Sigma^{1/2} z$, where z is a random vector with i.i.d. entries that have zero mean and unit variance, and Σ is arbitrary (but still deterministic and positive definite). To make the analysis (i.e, the bias calculation) tractable, we introduce a prior

$$\beta \sim P_\beta, \quad \text{where } \mathbb{E}(\beta) = 0, \ \mathrm{Cov}(\beta) = \frac{r^2}{p} I. \tag{9}$$

We consider an integrated or Bayes risk,

$$R_X(\hat{\beta}) = \mathbb{E}[R_X(\hat{\beta}; \beta)],$$

where the expectation is over the prior in (9). We have the bias-variance decomposition

$$R_X(\hat{\beta}) = \underbrace{\mathbb{E}[B_X(\hat{\beta}; \beta)]}_{B_X(\hat{\beta})} + \underbrace{\mathbb{E}[V_X(\hat{\beta}; \beta)]}_{V_X(\hat{\beta})}.$$

For the min-norm least squares estimator (4), its Bayes variance is as before, $V_X(\hat{\beta}) = V_X(\hat{\beta}; \beta) = (\sigma^2/n)\mathrm{tr}(\hat{\Sigma}^+ \Sigma)$, from Lemma (1) (because, as we can see, $V_X(\hat{\beta}; \beta)$ does not actually depend on β). Its Bayes bias is computed next.

4.1 Bayes bias

With the prior (9) in place (in which, note, $r^2 = \mathbb{E}\|\beta\|_2^2$), we have the following result for the Bayes bias

Lemma 4. *Under the prior (9), and data model (2), (3), the min-norm least squares estimator (4) has Bayes bias*

$$B_X(\hat{\beta}) = \frac{r^2}{p} \mathrm{tr}\big((I - \hat{\Sigma}^+ \hat{\Sigma})\Sigma\big).$$

Proof. Using trace rotation, we can rewrite the bias as $B_X(\hat{\beta}; \beta) = \mathrm{tr}(\beta\beta^T \Pi \Sigma \Pi)$. Taking an expectation over β, and using trace rotation again, gives $\mathbb{E}[B_X(\hat{\beta}; \beta)] = (r^2/p)\mathrm{tr}(\Pi\Sigma)$, which is the desired result. □

4.2 Limiting risk

We compute the asymptotic risk for a general feature covariance Σ. Before stating the result, we recall that for a measure G supported on $[0, \infty)$, we define its *Stieltjes transform* m_G, any $z \in \mathbb{C} \setminus \mathrm{supp}(G)$, by

$$m_G(z) = \int \frac{1}{u - z} \, dG(u).$$

Furthermore, the *companion Stieltjes transform* v_G is defined by

$$v_G(z) + 1/z = \gamma\big(m_G(z) + 1/z\big).$$

The proof of the next result is found in Appendix A.2. The main work for calculating for the asymptotic risk here was in fact already done by Dobriban and Wager (2018) (who in turn used a key result on trace functionals involving $\hat{\Sigma}, \Sigma$ from Ledoit and Peche 2011): these authors studied the asymptotic risk of ridge regression for general Σ, and the next result for min-norm least squares can be obtained by taking a limit in their result as the ridge parameter λ tends to zero (though some care is required in exchanging limits as $n, p \to \infty$ and $\lambda \to 0^+$).

Theorem 3. *Assume the prior (9), and data model (2), (3). Assume $x \sim P_x$ is of the form $x = \Sigma^{1/2} z$, where z is a random vector with i.i.d. entries that have zero mean, unit variance, and a finite 12th moment, and Σ is a deterministic positive definite matrix, such that $0 < c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$, for all n, p and constants c, C. As $n, p \to \infty$, assume that F_Σ converges weakly to a measure H. For the min-norm least squares estimator in (4), as $n, p \to \infty$, with $p/n \to \gamma > 1$, we have almost surely*

$$R_X(\hat{\beta}) \to \frac{r^2}{\gamma} \frac{1}{v(0)} + \sigma^2 \left(\frac{v'(0)}{v(0)^2} - 1 \right),$$

where we abbreviate $v = v_{F_{H,\gamma}}$, the companion Stieltjes transform of the empirical spectral distribution $F_{H,\gamma}$ given by the Marchenko-Pastur theorem, and we write v' for its derivative. Also, we write $v(0)$ to denote $v(0) = \lim_{z \to 0^+} v(-z)$, and likewise $v'(0) = \lim_{z \to 0^+} v'(-z)$, which exist under our assumptions above.

It is not always possible to analytically evaluate $v(0)$ or $v'(0)$. But when $\Sigma = I$, the companion Stieltjes transform is available in closed-form (39), and a tedious but straightforward calculation, deferred to Appendix A.3, shows that the asymptotic risk from Theorem 3 reduces to that from Theorem 2 (as it should). The next subsection generalizes this $\Sigma = I$ result, by looking at covariance matrices with constant off-diagonals.

4.3 Equicorrelated features

As a corollary to Theorem 3, we consider a *ρ-equicorrelation* structure for Σ, for a constant $\rho \in [0, 1)$, meaning that $\Sigma_{ii} = 1$ for all i, and $\Sigma_{ij} = \rho$ for all $i \neq j$. Interestingly, we recover the same asymptotic form for the variance as in the $\Sigma = I$ case, but the bias is affected—in fact, helped—by the presence of correlation. In the proof, deferred to Appendix A.4, we leverage the Silverstein equation (Silverstein, 1995) to derive an explicit form for the companion Stieltjes transform when Σ has ρ-equicorrelation structure (by relating it to the transform when $\Sigma = I$).

Corollary 2. *Assume the conditions of Theorem 3, and moreover, assume that Σ has ρ-equicorrelation structure for all n, p, and some $\rho \in [0, 1)$. Then as $n, p \to \infty$, with $p/n \to \gamma > 1$, we have almost surely*

$$R_X(\hat{\beta}) \to r^2(1 - \rho)(1 - 1/\gamma) + \frac{\sigma^2}{\gamma - 1}.$$

Figure 9, deferred until Appendix A.5, displays asymptotic risk curves when Σ has equicorrelation structure, as ρ varies from 0 to 0.75. This same section in the appendix details the computation of the asymptotic risk when we have a *ρ-autoregressive* structure for Σ, for a constant $\rho \in [0, 1)$, meaning that $\Sigma_{ij} = \rho^{|i-j|}$ for all i, j. Figure 10, also in Appendix A.5, displays the asymptotic risk curves in the autoregressive case, as ρ varies from 0 to 0.75.

We make one further point. Inspection of the asymptotic bias and variance curves individually (rather than the risk as a whole) reveals that in the autoregressive setting, *both* the bias and the variance depend on the correlation structure (cf. the equicorrelation setting in Corollary 2, where only the bias did). Figure 11, in Appendix A.5, shows that the bias improves as ρ increases, and the variance worsens with as ρ increases.

4.4 Limiting ℓ_2 norm

Again, as in the isotropic case, analysis of the limiting ℓ_2 norm is similar to analysis of the risk in Theorem 3, and so we give the next result without proof.

Corollary 3. *Assume the conditions of Theorem 3. Then as $n, p \to \infty$, such that $p/n \to \gamma$, the squared ℓ_2 norm of the min-norm least squares estimator (4) satisfies, almost surely,*

$$\mathbb{E}[\|\hat{\beta}\|_2^2 \,|\, X] \to \begin{cases} r^2 + \sigma^2 \gamma m(0) & \text{for } \gamma < 1, \\ r^2 \frac{1}{\gamma} + \sigma^2 \gamma m(0) & \text{for } \gamma > 1, \end{cases}$$

where we abbreviate $m = m_{F_{H,\gamma}}$ for the Stieltjes transform of empirical spectral distribution $F_{H,\gamma}$, and we write $m(0)$ to denote $m(0) = \lim_{z \to 0^+} m(-z)$, which exists under our assumptions.

5 Misspecified model

In this section, we consider a misspecified model, in which the regression function is still linear, but we observe only a subset of the features. Such a setting is more closely aligned with practical interest in interpolation: in many problems, we do not know the form of the regression function, and we generate features in order to improve our approximation capacity. Increasing the number of features past the point of interpolation (increasing γ past 1) can now decrease *both* bias and variance (i.e., not just the variance, as in the well-specified setting considered previously).

As such, the misspecified model setting also yields more interesting asymptotic comparisons between the $\gamma < 1$ and $\gamma > 1$ regimes. Recall that in Section 3.3, assuming isotropic features, we showed that when $\mathrm{SNR} > 1$ the asymptotic risk can have a *local* minimum on $(1, \infty)$. Of course, the risk function in (8) is globally minimized at $\gamma = 0$, which is a consequence of the fact that, in previous sections, we were assuming a well-specified linear model (3) at each γ, and trivially at $\gamma = 0$ there is no bias and no variance, and hence no risk. In a misspecified model, we will see that the story can be quite different, and the asymptotic risk can actually attain its *global* minimum on $(1, \infty)$.

5.1 Data model and risk

Consider, instead of (2), (3), a data model

$$\big((x_i, w_i), \epsilon_i\big) \sim P_{x,w} \times P_\epsilon, \quad i = 1, \dots, n, \tag{10}$$

$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, \quad i = 1, \dots, n, \tag{11}$$

where as before the random draws across $i = 1, \dots, n$ are independent. Here, we partition the features according to $(x_i, w_i) \in \mathbb{R}^{p+d}$, $i = 1, \dots, n$, where the joint distribution $P_{x,w}$ is such that $\mathbb{E}((x_i, w_i)) = 0$ and

$$\mathrm{Cov}\big((x_i, w_i)\big) = \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xw} \\ \Sigma_{xw}^T & \Sigma_w \end{bmatrix}.$$

We collect the features in a block matrix $[\, X \;\; W \,] \in \mathbb{R}^{n \times (p+d)}$ (which has rows $(x_i, w_i) \in \mathbb{R}^{p+d}$, $i = 1, \dots, n$). We presume that X is observed but W is unobserved, and focus on the min-norm least squares estimator exactly as before in (4), from the regression of y on X (not the full feature matrix $[\, X \;\; W \,]$).

Given a test point $(x_0, w_0) \sim P_{x,w}$, and an estimator $\hat{\beta}$ (fit using X, y only, and not W), we define its out-of-sample prediction risk as

$$R_X(\hat{\beta}; \beta, \theta) = \mathbb{E}\big[\big(x_0^T \hat{\beta} - \mathbb{E}(y_0 | x_0, w_0)\big)^2 \,|\, X\big] = \mathbb{E}\big[\big(x_0^T \hat{\beta} - x_0^T \beta - w_0^T \theta\big)^2 \,|\, X\big].$$

Note that this definition is conditional on X, and we are integrating over the randomness not only in ϵ (the training errors), but in the unobserved features W, as well. The next lemma decomposes this notion of risk in a useful way.

Lemma 5. *Under the misspecified model (10), (11), for any estimator $\hat{\beta}$, we have*

$$R_X(\hat{\beta}; \beta, \theta) = \underbrace{\mathbb{E}\big[\big(x_0^T \hat{\beta} - \mathbb{E}(y_0 | x_0)\big)^2 \,|\, X\big]}_{R_X^*(\hat{\beta}; \beta, \theta)} + \underbrace{\mathbb{E}\big[\big(\mathbb{E}(y_0 | x_0) - \mathbb{E}(y_0 | x_0, w_0)\big)^2\big]}_{M(\beta, \theta)}.$$

Proof. Simply add an subtract $\mathbb{E}(y_0 | x_0)$ inside the square in the definition of $R_X(\hat{\beta}; \beta, \theta)$, then expand, and note that the cross term can be written, conditional on x_0, as

$$\mathbb{E}\big[\big(x_0^T \hat{\beta} - \mathbb{E}(y_0 | x_0)\big) \,|\, X, x_0\big] \mathbb{E}\big[\big(\mathbb{E}(y_0 | x_0) - \mathbb{E}(y_0 | x_0, w_0)\big) \,|\, x_0\big] = 0.$$

□

The first term $R_X^*(\hat{\beta}; \beta, \theta)$ in the decomposition in Lemma 5 is the precisely the risk that we studied previously in the well-specified case, except that the response distribution has changed (due to the presence of the middle term in (11)). We call the second term $M(\beta, \theta)$ in Lemma 5 the *misspecification bias*. In general, computing $R_X^*(\hat{\beta}; \beta, \theta)$ and $M(\beta, \theta)$ in finite-sample can be very difficult, owing to the potential complexities created by the middle term in (11). However, in some special cases—for example, when the observed and unobserved features are independent, or jointly Gaussian—we can precisely characterize the contribution of the middle term in (11) to the overall response distribution, and can then essentially leverage our previous results to characterize risk in the misspecified model setting. In what follows, we restrict our attention to the independence setting, for simplicity.

5.2 Isotropic features

When the observed and unobserved features are independent, $P_{x,w} = P_x \times P_w$, the middle term in (11) only adds a constant to the variance, and the analysis of $R^*_X(\hat{\beta}; \beta, \theta)$ and $M(\beta, \theta)$ becomes tractable. Here, we make the additional simplifying assumption that $(x, w) \sim P_{x,w}$ has i.i.d. entries with unit variance, which implies that $\Sigma = I$. (The case of independent features but general covariances Σ_x, Σ_w is similar, and we omit the details.) Therefore, we may write the response distribution in (11) as

$$y_i = x_i^T \beta + \delta_i, \quad i = 1, \ldots, n,$$

where δ_i is independent of x_i, having mean zero and variance $\sigma^2 + \|\theta\|_2^2$, for $i = 1, \ldots, n$. Denote the total signal by $r^2 = \|\beta\|_2^2 + \|\theta\|_2^2$, and the fraction of the signal captured by the observed features by $\kappa = \|\beta\|_2^2 / r^2$. Then $R^*_X(\hat{\beta}; \beta, \theta)$ behaves exactly as we computed previously, for isotropic features in the well-specified setting (Theorem 1 for $\gamma < 1$, and Theorem 2 for $\gamma > 1$), after we make the substitutions:

$$r^2 \mapsto r^2 \kappa \quad \text{and} \quad \sigma^2 \mapsto \sigma^2 + r^2(1 - \kappa). \tag{12}$$

Furthermore, we can easily calculate the misspecification bias:

$$M(\beta, \theta) = \mathbb{E}(w_0^T \theta)^2 = r^2(1 - \kappa).$$

Putting these results together leads to the next conclusion.

Theorem 4. *Assume the misspecified model (10), (11), and assume $(x, w) \sim P_{x,w}$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8 + \eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 + \|\theta\|_2^2 = r^2$ and $\|\beta\|_2^2 / r^2 = \kappa$ for all n, p. Then for the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, with $p/n \to \gamma$, it holds almost surely that*

$$R_X(\hat{\beta}; \beta, \theta) \to \begin{cases} r^2(1-\kappa) + \big(r^2(1-\kappa) + \sigma^2\big)\frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2(1-\kappa) + r^2\kappa\big(1 - \frac{1}{\gamma}\big) + \big(r^2(1-\kappa) + \sigma^2\big)\frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

We remark that, in the independence setting considered in Theorem 4, the dimension d of the unobserved feature space does not play any role, and the result only depends on the unobserved features via κ. Therefore, we may equally well take $d = \infty$ for all n, p (i.e., infinitely many unobserved features).

The components of the limiting risk from Theorem 4 are intuitive and can be interpreted as follows. The first term $r^2(1 - \kappa)$ is the misspecification bias (irreducible). The second term, which we deem as 0 for $\gamma < 1$ and $r^2\kappa(1 - 1/\gamma)$ for $\gamma > 1$, is the bias. The third term, $r^2(1 - \kappa)\gamma/(1 - \gamma)$ for $\gamma < 1$ and $r^2(1 - \kappa)/(\gamma - 1)$ for $\gamma > 1$, is what we call the *misspecification variance*: the inflation in variance due to unobserved features, when we take $\mathbb{E}(y_0|x_0)$ to be the target of estimation. The last term, $\sigma^2\gamma/(1 - \gamma)$ for $\gamma < 1$ and $\sigma^2/(\gamma - 1)$ for $\gamma > 1$, is the variance itself.

5.3 Polynomial approximation bias

Since adding features should generally improve our approximation capacity, it is reasonable to model $\kappa = \kappa(\gamma)$ as an increasing function of γ. To get an idea of the possible shapes taken by the asymptotic risk curve from Theorem 4, we can inspect different regimes for the approximation bias, i.e., the rate at which $1 - \kappa(\gamma) \to 0$ as $\gamma \to \infty$. For example, we may consider a *polynomial decay* for the approximation bias,

$$1 - \kappa(\gamma) = (1 + \gamma)^{-a}, \tag{13}$$

for some $a > 0$. In this case, the limiting risk in the isotropic setting, from Theorem 4, becomes

$$R_a(\gamma) = \begin{cases} r^2(1+\gamma)^{-a} + \big(r^2(1+\gamma)^{-a} + \sigma^2\big)\frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2(1+\gamma)^{-a} + r^2\big(1 - (1+\gamma)^{-a}\big)\big(1 - \frac{1}{\gamma}\big) + \big(r^2(1+\gamma)^{-a} + \sigma^2\big)\frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases} \tag{14}$$

Compare (14) to the well-specified asymptotic risk (8). By taking $a \to \infty$ in (14), we recover (8). But for small $a > 0$, the misspecified risk curve (14) can have some very different and interesting features. The next points summarize, and Figures 3 and 4 give accompanying plots are given in when $\mathrm{SNR} = 1$ and 5, respectively. Recall that the null risk is r^2, which comes from predicting with the null estimator $\tilde{\beta} = 0$.

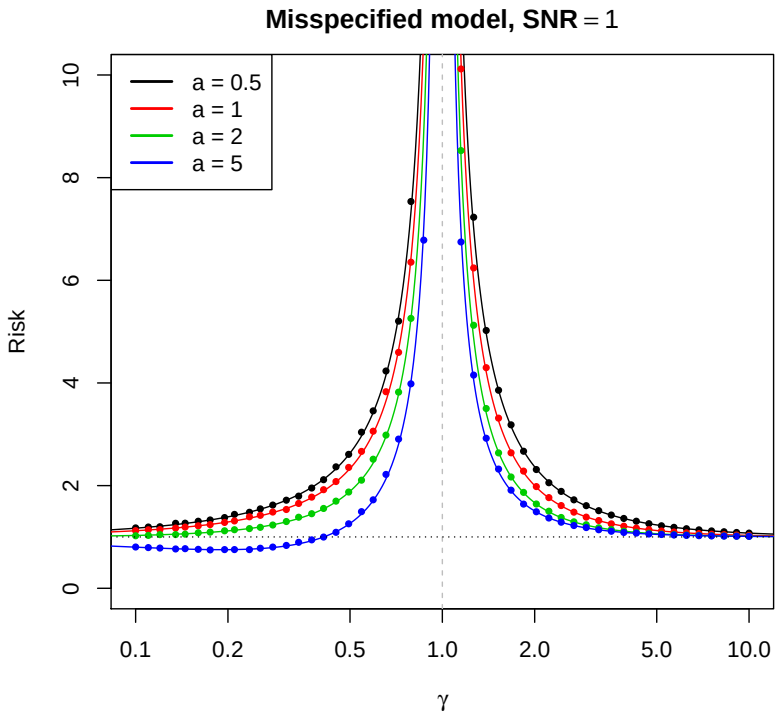

Figure 3: Asymptotic risk curves in (14) for the min-norm least squares estimator in the misspecified case, when the approximation bias has polynomial decay as in (13), as a varies from 0.5 to 5. Here $r^2 = 1$ and $\sigma^2 = 1$, so SNR $= 1$. The null risk $r^2 = 5$ is marked as a dotted black line. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries.

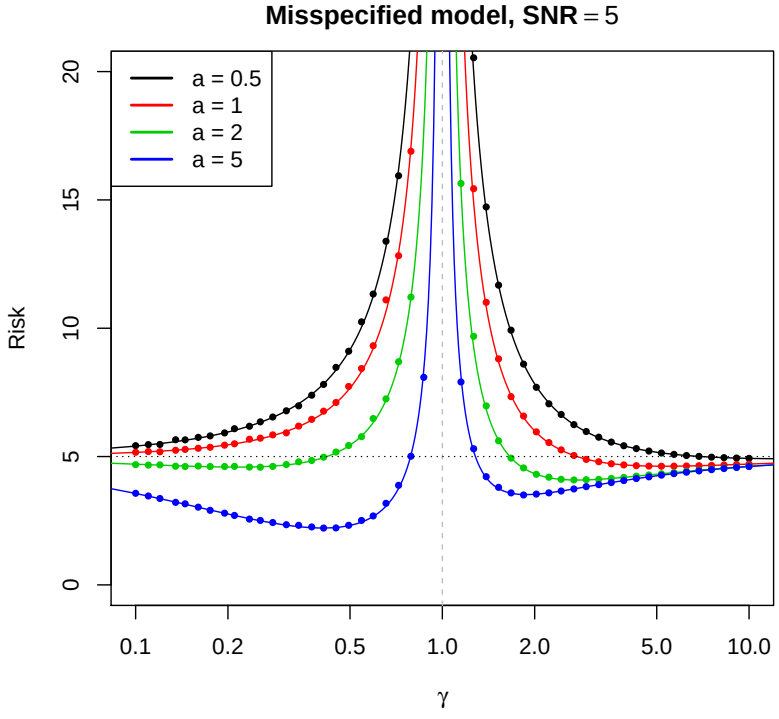

Figure 4: Asymptotic risk curves in (14) for the min-norm least squares estimator in the misspecified case, when the approximation bias has polynomial decay as in (13), as a varies from 0.5 to 5. Here $r^2 = 5$ and $\sigma^2 = 1$, so SNR $= 5$. The null risk $r^2 = 5$ is marked as a dotted black line. The points are again finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ.

1. On $(0,1)$, the least squares risk $R_a(\gamma)$ can only be better than the null risk if $a > 1 + \frac{1}{\mathrm{SNR}}$. Further, in this case, we have $R_a(\gamma) < r^2$ if and only if $\gamma < \gamma_0$, where γ_0 is the unique zero of the function

$$(1+x)^{-a} + \Big(1 + \frac{1}{\mathrm{SNR}}\Big)x - 1$$

that lies in $(0, \frac{\mathrm{SNR}}{\mathrm{SNR}+1})$. Finally, on $(\frac{\mathrm{SNR}}{\mathrm{SNR}+1}, 1)$, the least squares risk $R_a(\gamma)$ is always worse than the null risk, regardless of $a > 0$, and it is monotonically increasing.

2. On $(1,\infty)$, when $\mathrm{SNR} \leq 1$, the min-norm least squares risk $R_a(\gamma)$ is always worse than the null risk. Moreover, it is monotonically decreasing, and approaches the null risk (from above) as $\gamma \to \infty$.

3. On $(1,\infty)$, when $\mathrm{SNR} > 1$, the min-norm least squares risk $R_a(\gamma)$ can be better than the null risk for any $a > 0$, and in particular we have $R_a(\gamma) < r^2$ if and only if $\gamma < \gamma_0$, where γ_0 is the unique zero of the function

$$(1+x)^{-a}(2x-1) + 1 - \Big(1 - \frac{1}{\mathrm{SNR}}\Big)x$$

lying in $(\frac{\mathrm{SNR}}{\mathrm{SNR}-1}, \infty)$. Indeed, on $(1, \frac{\mathrm{SNR}}{\mathrm{SNR}-1})$, the min-norm least squares risk $R_a(\gamma)$ is always worse than the null risk (regardless of $a > 0$), and it is monotonically decreasing.

4. When $\mathrm{SNR} > 1$, for small enough $a > 0$, the global minimum of the min-norm least squares risk $R_a(\gamma)$ occurs after $\gamma = 1$. A sufficient but not necessary condition is $a \leq 1 + \frac{1}{\mathrm{SNR}}$ (because, from points 1 and 3 above, we see that in this case $R_a(\gamma)$ is always worse than null risk for $\gamma < 1$, but will be better than the null risk at some $\gamma > 1$).

6 Ridge regularization

We compare the limiting risks of min-norm least squares and ridge regression. For the case of isotropic features, the limiting risk of ridge regression is yet again a well-known calculation in random matrix theory, and can be found in Chapter 4 of Tulino and Verdu (2004); see also Dicker (2016). A risk comparison for the case of correlated features is also possible, where we would rely on Dobriban and Wager (2018) for the ridge results, but we focus on the isotropic case for simplicity.

We state the next result without proof, as the proof closely follows that of Theorem 2 for the well-specified part, and Theorem 4 for the misspecified part. Very similar (though not identical) results can be found in Dicker (2016); Dobriban and Wager (2018), for the well-specified part.

Theorem 5. *Assume the conditions of Theorem 2 (well-specified model, isotropic features). Then for ridge regression in (5) with $\lambda > 0$, as $n, p \to \infty$, such that $p/n \to \gamma \in (0,\infty)$, it holds almost surely that*

$$R_X(\hat{\beta}_\lambda; \beta) \to \sigma^2\gamma \int \frac{\alpha\lambda^2 + s}{(s+\lambda)^2}\, dF_\gamma,$$

where F_γ is the Marchenko-Pastur law, and $\alpha = r^2/(\sigma^2\gamma)$. The limiting risk can be alternatively written as

$$\sigma^2\gamma\big(m(-\lambda) - \lambda(1-\alpha\lambda)m'(-\lambda)\big).$$

where we abbreviate $m = m_{F_\gamma}$ for the Stieltjes transform of the Marchenko-Pastur law F_γ. Furthermore, the limiting ridge risk is minimized at $\lambda^* = 1/\alpha$, in which case the optimal limiting risk can be written explicitly as

$$\sigma^2\gamma \cdot m(-1/\alpha) = \sigma^2 \frac{-(1-(1+\sigma^2/r^2)\gamma) + \sqrt{(1-(1+\sigma^2/r^2)\gamma)^2 - 4\sigma^2\gamma^2/r^2}}{2\gamma},$$

where we have used the closed-form for the Stieltjes transform of the Marchenko-Pastur law, see (7).

Under the conditions of Theorem 4 (misspecified model, isotropic features), the limiting risk of ridge regression is as in the first two displays, and the optimal limiting risk is as in the third, after we make the substitutions in (12) and add $r^2(1-\kappa)$, to each expression.

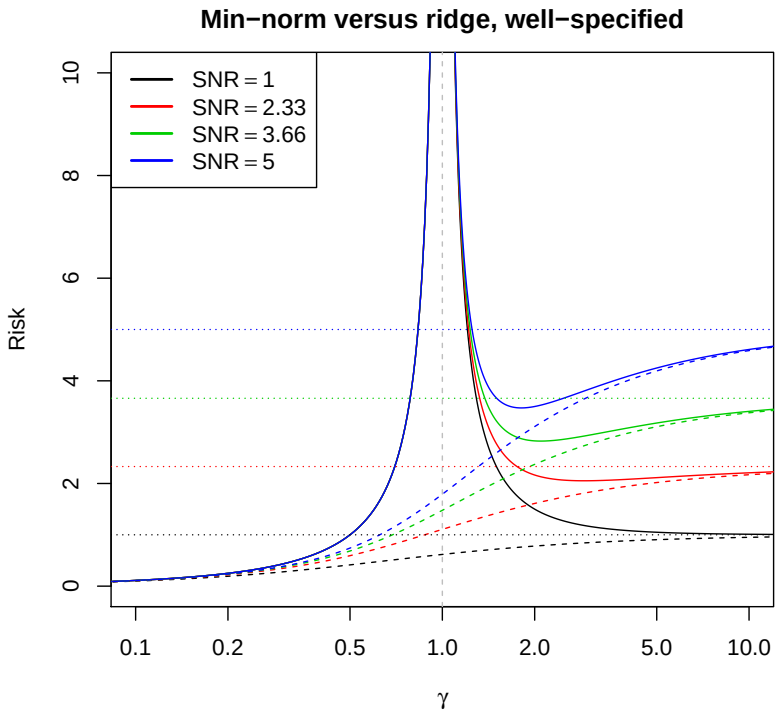

Figure 5: Asymptotic risk curves for the min-norm least squares estimator in (8) as solid lines, and optimally-tuned ridge regression (from Theorem 5) as dashed lines. Here r^2 varies from 1 to 5, and $\sigma^2 = 1$. The null risks are marked by the dotted lines.

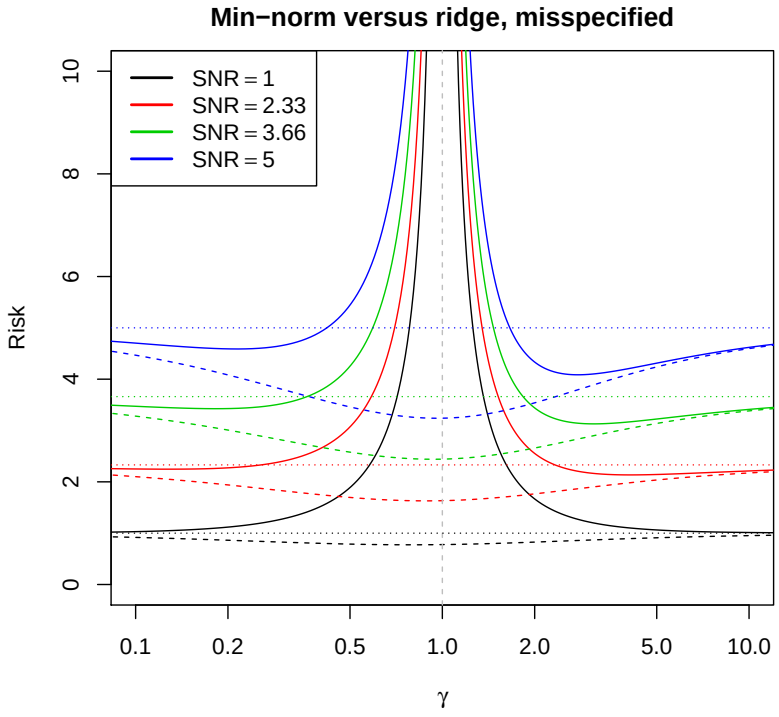

Figure 6: Asymptotic risk curves for the min-norm least squares estimator in (14) as solid lines, and optimally-tuned ridge regression (from Theorem 5) as dashed lines, in the misspecified case, when the approximation bias has polynomial decay as in (13), with $a = 2$. Here r^2 varies from 1 to 5, and $\sigma^2 = 1$. The null risks are marked by the dotted lines.

Figures 5 and 6 compare the risk curves of min-norm least squares to those from optimally-tuned ridge regression, in the well-specified and misspecified settings, respectively. There are two important points to make. The first is that optimally-tuned ridge regression is seen to have strictly better asymptotic risk throughout, regardless of r^2, γ, κ. This should not be a surprise, as by definition optimal tuning should yield better risk than min-norm least squares, which is the special case given by $\lambda \to 0^+$. Moreover, we must note that this is a particularly favorable problem setting for ridge regression. It is not hard to check that the asymptotic risk of ridge regression here, when $\|\beta\|_2^2 = r^2$ for all n, p, is the same as the asymptotic Bayes risk when β is drawn from a spherical prior as in (9), with $\mathbb{E}\|\beta\|_2^2 = r^2$ for all n, p. Under the generic data model (3) and prior (9), optimally-tuned ridge regression has the best asymptotic Bayes risk of *any* linear estimator. To see this, fix any n, p, and observe that for any linear estimator, its Bayes risk only depends on the likelihood (3) and prior (9) and via the parameters σ^2, r^2. If we specialize to the case of a normal-normal pair for the likelihood and prior, then optimally-tuned ridge regression is the (unique) Bayes estimator, so it has better Bayes risk than all estimators, including linear ones. As this holds for all n, p, it must also hold in the limit as $n, p \to \infty$.

The second point is that, in the misspecified case, the limiting risk of optimally-tuned ridge regression appears to have a minimum around $\gamma = 1$, and this occurs closer and closer to $\gamma = 1$ as SNR grows. This behavior is interesting, especially because it is completely antipodal to that of the min-norm least squares risk, and leads us to very different suggestions for practical useage for feature generators: in settings where we apply substantial ℓ_2 regularization (say, using CV tuning to mimic optimal tuning, which the next section shows to be asymptotically equivalent), it seems we want the complexity of the feature space to put us as close to the interpolation boundary ($\gamma = 1$) as possible.

7 Cross-validation

We analyze the effect of using cross-validation to choose the tuning parameter in ridge regression. In short, we find that choosing the ridge tuning parameter to minimize the leave-one-out cross-validation error leads to the same asymptotic risk as the optimally-tuned ridge estimator. The next subsection gives the details; the following subsection presents a new "shortcut formula" for leave-one-out cross-validation in the overparametrized regime, for min-norm least squares, akin to the well-known formula for underparametrized least squares and ridge regression.

7.1 Limiting behavior of CV tuning

Given the ridge regression solution $\hat{\beta}_\lambda$ in (5), trained on (x_i, y_i), $i = 1, \ldots, n$, denote by $\hat{f}_\lambda$ the corresponding ridge predictor, defined as $\hat{f}_\lambda(x) = x^T \hat{\beta}_\lambda$ for $x \in \mathbb{R}^p$. Additionally, for each $i = 1, \ldots, n$, denote by $\hat{f}_\lambda^{-i}$ the ridge predictor trained on all but ith data point (x_i, y_i).[1] Recall that the *leave-one-out cross-validation* (leave-one-out CV, or simply CV) error of the ridge solution at a tuning parameter value λ is

$$\mathrm{CV}_n(\lambda) = \frac{1}{n} \sum_{i=1}^n \big(y_i - \hat{f}_\lambda^{-i}(x_i)\big)^2. \tag{15}$$

We typically view this as an estimate of the out-of-sample prediction error $\mathbb{E}(y_0 - x_0^T \hat{\beta}_\lambda)^2$, where the expectation is taken over everything that is random: the training data (x_i, y_i), $i = 1, \ldots, n$ used to fit $\hat{\beta}_\lambda$, as well as the independent test point (x_0, y_0). Note also that, when we observe training data from the model (2), (3), and when (x_0, y_0) is drawn independently according to the same process, we have the relationship

$$\mathbb{E}(y_0 - x_0^T \hat{\beta}_\lambda)^2 = \sigma^2 + \mathbb{E}(x_0^T \beta - x_0^T \hat{\beta}_\lambda)^2 = \sigma^2 + \mathbb{E}[R_X(\hat{\beta}_\lambda; \beta)],$$

where $R_X(\hat{\beta}_\lambda; \beta)] = \mathbb{E}[(x_0^T \beta - x_0^T \hat{\beta}_\lambda)^2 \,|\, X]$ is the conditional prediction risk, which has been our focus throughout.

Recomputing the leave-one-out predictors $\hat{f}_\lambda^{-i}$, $i = 1, \ldots, n$ can be burdensome, especially for large n. Importantly, there is a well-known "shortcut formula" that allows us to express the leave-one-out CV error (15) as a weighted average of the training errors,

$$\mathrm{CV}_n(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - (S_\lambda)_{ii}} \right)^2, \tag{16}$$

[1] To be precise, this is $\hat{f}^{-i}(x) = x^T (X_{-i}^T X_{-i} + n\lambda I)^{-1} X_{-i}^T y_{-i}$, where X_{-i} denotes X with the ith row removed, and y_{-i} denotes y with the ith component removed. Arguably, it may seem more natural to replace the factor of n here by a factor of $n - 1$; we leave the factor of n as is because it simplifies the presentation in what follows, but we remark that the same asymptotic results would hold with $n - 1$ in place of n.

where $S_\lambda = X(X^TX + n\lambda)^{-1}X^T$ is the ridge smoother matrix. There are several ways to verify (16); one way is to use the Sherman-Morrison-Woodbury formula to relate $(X_{-i}^TX_{-i} + n\lambda I)^{-1}$ to $(X^TX + n\lambda)^{-1}$, where X_{-i} denotes X with the ith row removed. The shortcut formula (16) is valid when $\lambda > 0$, or when $\lambda = 0$ and $\mathrm{rank}(X) = p$. When $\lambda = 0$ and $\mathrm{rank}(X) = n < p$, it is not well-defined, as both the numerator and denominator are zero in each summand. In the next subsection, we give an extension to the case $\lambda = 0$ and $\mathrm{rank}(X) = n$, i.e., to min-norm least squares.

The next result shows that, for isotropic features, the CV error of a ridge estimator converges almost surely to its prediction error. The focus on isotropic features and on the Bayes problem (where β is drawn from the prior in (9)) is only done for simplicity; a more general analysis is possible but is not pursued here. The proof, given in Appendix A.6, relies on the shortcut formula (16). In the proof, we actually first analyze generalized cross-validation (GCV), which turns out to be somewhat of an easier calculation (see the proof for details on the precise form of GCV), and then relate leave-one-out CV to GCV.

Theorem 6. *Assume the prior (9) and data model (2), (3). Assume that $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $4 + \eta$, for some $\eta > 0$. Then for the CV error (15) of the ridge estimator in (5) with tuning parameter $\lambda > 0$, as $n, p \to \infty$, with $p/n \to \gamma \in (0, \infty)$, it holds almost surely that*

$$\mathrm{CV}_n(\lambda) - \sigma^2 \to \sigma^2\gamma\big(m(-\lambda) - \lambda(1 - \alpha\lambda)m'(-\lambda)\big),$$

where $m = m_{F_\gamma}$ denotes the Stieltjes transform of the Marchenko-Pastur law F_γ, and $\alpha = r^2/(\sigma^2\gamma)$. Observe that the right-hand side is the asymptotic risk of ridge regression from Theorem 5. Moreover, the above convergence is uniform over compacts intervals excluding zero. Thus if λ_1, λ_2 are constants with $0 < \lambda_1 \leq \lambda^* \leq \lambda_2 < \infty$, where $\lambda^* = 1/\alpha$ is the asymptotically optimal ridge tuning parameter value, and we define $\lambda_n = \arg\min_{\lambda \in [\lambda_1, \lambda_2]} \mathrm{CV}_n(\lambda)$, then the risk of the CV-tuned ridge estimator $\hat{\beta}_{\lambda_n}$ satisfies, almost surely,

$$R_X(\hat{\beta}_{\lambda_n}) \to \sigma^2\gamma m(-1/\alpha),$$

with the right-hand side above being the asymptotic risk of optimally-tuned ridge regression. Further, the exact same set of results holds for GCV.

Finally, the analogous results also hold in the misspecified model, under the conditions of Theorem 4. Namely, the CV and GCV errors converge almost surely to the asymptotic prediction error of ridge regression (σ^2 plus its asymptotic risk), uniformly over compact intervals in λ excluding zero. Therefore, the CV- or GCV-tuned ridge estimator—where the tuning parameter λ_n is defined to minimize $\mathrm{CV}_n(\lambda)$ or $\mathrm{GCV}_n(\lambda)$ over such an interval containing the asymptotically optimal ridge tuning parameter λ^*—achieves the optimal asymptotic ridge risk.

We remark that the convergence of CV and GCV in Theorem 6 are not really surprising results. In a way, they are also not entirely new; classical theory shows CV and GCV tuning to be both asymptotically optimal for various linear smoothers, including ridge regression; see Li (1986, 1987) (who even allows for the high-dimensional case, where $n, p \to \infty$ together). Our asymptotic results in Theorem 6 are similar in spirit to these older results, but the details differ: owing to our random matrix theory approach, we are able to establish the (stronger) result that the CV and GCV error curves converge *uniformly* to the ridge prediction error curve, by leveraging the fact that they are composed of functionals that have almost sure limits under Marchenko-Pastur asymptotics. We also note that similar results were recently obtained for the lasso in Miolane and Montanari (2018), and for general smooth penalized estimators in Xu et al. (2019). The latter paper covers ridge regression as a special case, and gives more precise results (convergence rates), but assumes more restrictive conditions.

The key implication of Theorem 6, in the context of the current paper and its central focus, is that the CV-tuned or GCV-tuned ridge estimator has the same asymptotic performance as the optimally-tuned ridge estimator, and therefore enjoys the same perfomance gap over min-norm least squares. In other words, the ridge curves in Figures 1, 5, and 6 can be alternatively viewed as the asymptotic risk of ridge under CV tuning or GCV tuning, which suggests that we can still expect to see a significant improvement from using ℓ_2 regularization in these settings, when we use a data-driven rule to choose the tuning parameter. For an empirical comparison of the risks from CV and GCV tuning to the optimal ridge risk, see Figures 12 and 13, given in Appendix A.8.

7.2 Shortcut formula for ridgeless CV

We extend the leave-one-out CV shortcut formula (16) to work when $p > n$ and $\lambda = 0$, i.e., for min-norm least squares. In this case, both the numerator and denominator are zero in each summand of (16). To circumvent this, we can use the

so-called "kernel trick" to rewrite the ridge regression solution (5) with $\lambda > 0$ as

$$\hat{\beta}_\lambda = X^T(XX^T + n\lambda I)^{-1}y. \tag{17}$$

This can be verified using the Woodbury formula (more specifically, the push-through matrix identity, an easy consequence of the Woodbury formula). Under the representation (17), note that the ridge smoother matrix S_λ becomes

$$S_\lambda = XX^T(XX^T + n\lambda I)^{-1} = I - n\lambda(XX^T + n\lambda I)^{-1},$$

hence the shortcut formula for leave-one-out CV in (16) can be rewritten as

$$\mathrm{CV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n \frac{[(XX^T + n\lambda I)^{-1}y]_i}{[(XX^T + n\lambda I)^{-1}]_{ii}},$$

where the common factor of λ in the numerator and denominator cancel. Taking $\lambda \to 0^+$ yields the shortcut formula for leave-one-out CV in min-norm least squares (assuming without a loss of generality that $\mathrm{rank}(X) = n$),

$$\mathrm{CV}_n(0) = \frac{1}{n}\sum_{i=1}^n \frac{[(XX^T)^{-1}y]_i}{[(XX^T)^{-1}]_{ii}}, \tag{18}$$

In fact, the exact same arguments given here still apply when we replace XX^T by a positive definite kernel matrix K (i.e., $K_{ij} = k(x_i, x_j)$ for each $i, j = 1, \ldots, n$, where k is a positive definite kernel function), in which case (18) gives a shortcut formula for leave-one-out CV in kernel ridgeless regression (the limit in kernel ridge regression as $\lambda \to 0^+$). We also remark that, when we include an unpenalized intercept in the model, in either the linear or kernelized setting, the shortcut formula (18) still applies with XX^T or K replaced by their doubly-centered (row- and column-centered) versions, and the matrix inverses replaced by pseudoinverses.

The formula (18) (and the extension which replaces XX^T by a kernel matrix K) is of course practically useful in that it allows us to evaluate the leave-one-out CV error of min-norm least squares (or kernel ridgeless regression) at the computational cost of fitting this estimate just once. Beyond this, it is conceptually interesting that a shortcut formula like (18) is possible at all, for the leave-one-out CV error of an interpolator, because it stems from the representation (16) that is based on reweighting the training errors, which do not contain any information for an interpolator (as they are all zero by definition).

8 Nonlinear model

We consider a nonlinear model for the features, which, as described in the introduction, is motivated by the linearized approximation (1) to neural networks. We observe data as in (2), (3), but now $x_i = \varphi(Wz_i) \in \mathbb{R}^p$, where $z_i \in \mathbb{R}^d$ has i.i.d. entries from $N(0,1)$, for $i = 1, \ldots, n$. Also, $W \in \mathbb{R}^{p\times d}$ has i.i.d. entries from $N(0, 1/d)$, and φ is an activation function acting componentwise.

8.1 Limiting variance

The next result characterizes the limiting prediction variance in the nonlinear setting.

Theorem 7. *Assume the model (2), (3), where each $x_i = \varphi(Wz_i) \in \mathbb{R}^p$, for $z_i \in \mathbb{R}^d$ having i.i.d. entries from $N(0,1)$, $W \in \mathbb{R}^{p\times d}$ having i.i.d. entries from $N(0, 1/d)$ (with W independent of z_i), and for φ an activation function that acts componentwise. Assume that $|\varphi(x)| \le c_0(1 + |x|)^{c_0}$ for a constant $c_0 > 0$. Also, for $G \sim N(0,1)$, assume that the standardization conditions hold: $\mathbb{E}[\varphi(G)] = 0$ and $\mathbb{E}[\varphi(G)^2] = 1$. Define*

$$c_1 = \mathbb{E}[G\varphi(G)]^2.$$

Then for the ridge regression estimator $\hat{\beta}_\lambda$ in (5), as $n, p, d \to \infty$, such that $p/n \to \gamma \in (0, \infty)$, $d/p \to \psi \in (0, 1)$, the following ridgeless limits hold almost surely. For $\gamma < 1$:

$$\lim_{\lambda\to 0^+} \lim_{n,p,d\to\infty} V_X(\hat{\beta}_\lambda; \beta) = \sigma^2 \frac{\gamma}{1-\gamma},$$

which is precisely as in the case of linear features, recall Theorem 1. For $\gamma > 1$:

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} V_X(\hat{\beta}_\lambda; \beta) = \\ - \sigma^2 \frac{c_1 \gamma^3 \chi_0^2 - 2c_1 \gamma^2 \chi_0^2 + c_1 \gamma^2 \chi_0 - 3c_1 \gamma \chi_0 - \gamma^3 \chi_0^2 + 2\gamma^2 \chi_0^2 - 2\gamma^2 \chi_0 + 4\gamma \chi_0 - \gamma + 2}{(\gamma - 1)(c_1 \gamma^2 \chi_0^2 + 2c_1 \gamma \chi_0 - \gamma^2 \chi_0^2 - 2\gamma \chi_0 - 1)},$$

where

$$\chi_0 = \psi \frac{1 - c_1/(\psi\gamma) - \sqrt{(1 - c_1/(\psi\gamma))^2 + 4c_1(1 - c_1)/(\psi\gamma)}}{2c_1(1 - c_1)}.$$

The proof of Theorem 7 is lengthy and will be sketched shortly. We remark that the results that we develop for its proof (in particular, Theorem 8) allow to characterize the limiting prediction variance of the ridge regression estimator $\hat{\beta}_\lambda$ for any fixed $\lambda > 0$. We defer the details to future work.

Figure 7 displays the asymptotic variance curve from Theorem 7 for the activation functions: $\varphi_{\text{tanh}}(x) = a_1 \tanh(x)$, $\varphi_{\text{ReLU}}(x) = a_2(\max(x, 0) - b_2)$, $\varphi_{\text{sign}}(x) = \text{sign}(x)$, and $\varphi_{\text{abs}}(x) = a_3(|t| - b_3)$, for constants a_1, a_2, b_2, a_3, b_3 that are chosen to ensure the standardization conditions ($\mathbb{E}[\varphi(G)] = 0$ and $\mathbb{E}[\varphi(G)^2] = 1$, for $G \sim N(0, 1)$). We remark that Theorem 7 implies a high degree of *universality*, since the asymptotic variance depends on the activation function only through the single parameter $c_1 = \mathbb{E}[G\varphi(G)]^2$. As reflected in the figure, the qualitative behavior with respect to $\gamma > 1$ is quite similar across different values of c_1. Further, for fixed $\gamma > 1$, the variance appears to increase with c_1.

8.2 Pure nonlinearity

A surprisingly simple result is obtained by specializing to the case $c_1 = 0$, which corresponds to a "purely nonlinear" activation, i.e., an activation that has vanishing projection onto the linear function in $L^2(\mathbb{R}, \mu_G)$ (where here and below μ_G denotes the standard Gaussian measure). The proof of the next result is given in Appendix B.5.

Corollary 4. *Assume the conditions of Theorem 7, and moreover, assume that $c_1 = \mathbb{E}[G\varphi(G)]^2 = 0$. Then for $\gamma > 1$, the variance satisfies, almost surely,*

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} V_X(\hat{\beta}_\lambda; \beta) = \frac{\sigma^2}{\gamma - 1},$$

which is precisely as in the case of linear isotropic features, recall Theorem 2. Also, under the prior (9), the Bayes bias satisfies, almost surely

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} B_X(\hat{\beta}_\lambda) = \begin{cases} 0 & \text{for } \gamma < 1, \\ r^2(1 - 1/\gamma) & \text{for } \gamma > 1, \end{cases}$$

which is again as in the case of linear isotropic features, recall Theorems 1 and 2.

In other words, Corollary 4 says that if φ is purely nonlinear, then the feature matrix X behaves "as if" it has i.i.d. entries, in that the asymptotic bias and variance are exactly as in the linear isotropic case, recall (8). This is true despite the fact that the actual dimension d of the input space can be significantly smaller than the number of features p.

Figure 8 compares the asymptotic risk curve from Corollary 4 to that computed by simulation, using an activation function $\varphi_{\text{abs}}(t) = a(|t| - b)$, where $a = \sqrt{\pi/(\pi - 2)}$ and $b = \sqrt{2/\pi}$ are chosen to meet the standardization conditions. This activation function is purely nonlinear, i.e., it satisfies $\mathbb{E}[G\varphi_{\text{abs}}(G)] = 0$ for $G \sim N(0, 1)$, by symmetry. Again, the agreement between finite-sample and asymptotic risks is excellent. Notice in particular that, as predicted by the corollary, the risk depends only on p/n and not on d/n.

8.3 Proof outline for Theorem 7

We denote $\gamma_n = p/n$ and $\psi_n = d/p$. Recall that as $n, p, d \to \infty$, we have $\gamma_n \to \gamma$ and $\psi_n \to \psi$. To reduce notational overhead, we will generally drop the subscripts from γ_n, ψ_n, writing these simply as γ, ψ, since their meanings should be clear from the context. We denote by $Q \in \mathbb{R}^{p \times p}$ the Gram matrix of the weight vectors w_i, $i = 1, \ldots, p$ (rows of $W \in \mathbb{R}^{p \times d}$), with diagonals set to zero. Namely, for each i, j,

$$Q_{ij} = \langle w_i, w_j \rangle 1\{i \neq j\}.$$

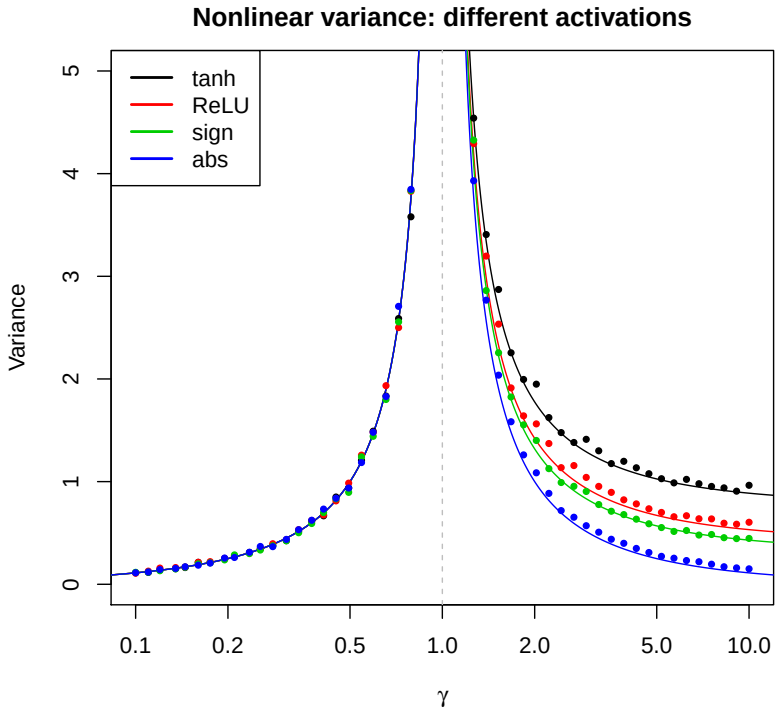

Figure 7: Asymptotic variance curves for the min-norm least squares estimator in the nonlinear feature model (from Theorem 7), for four different activation functions described in the main text: φ_{tanh} in black, φ_{ReLU} in red, φ_{sign} in green, and φ_{abs} in blue. Here $\sigma^2 = 1$, and the points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, and $d = 100$, computed from features $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries and W has i.i.d. $N(0, 1/d)$ entries.

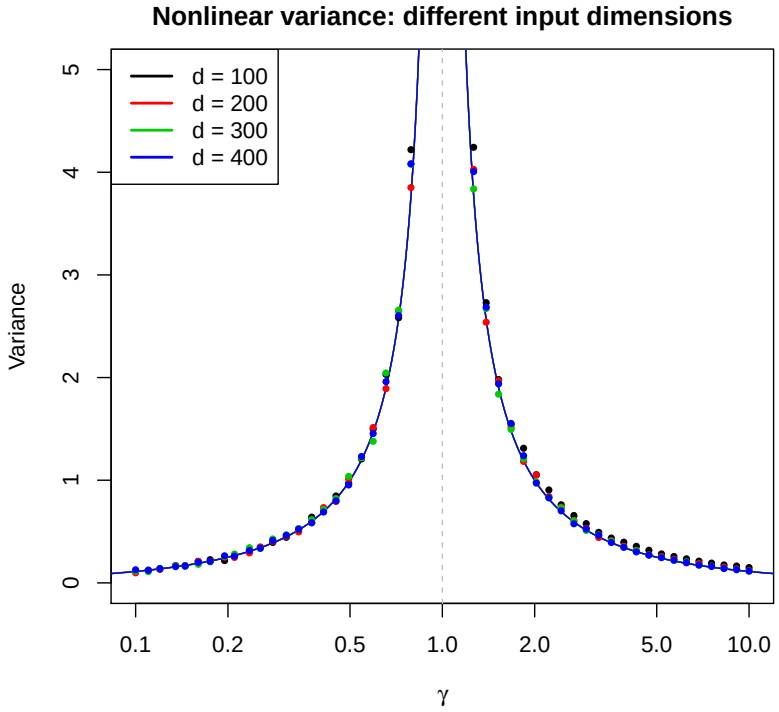

Figure 8: Asymptotic variance curves for the min-norm least squares estimator in the nonlinear feature model (from Corollary 4), for the purely nonlinear activation φ_{abs}. Here $\sigma^2 = 1$, and the points are finite-sample risks, with $n = 200$, $p = [\gamma n]$, over various values of γ, and varying input dimensions: $d = 100$ in black, $d = 200$ in red, $d = 300$ in green, and $d = 400$ in black. As before, the features used for finite-sample calculations are $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries and W has i.i.d. $N(0, 1/d)$ entries.

Let $N = p + n$ and define the symmetric matrix $A(s,t) \in \mathbb{R}^{N \times N}$, for $s \geq t \geq 0$, with the block structure:

$$A(s,t) = \begin{bmatrix} sI_p + tQ & \frac{1}{\sqrt{n}} X^T \\ \frac{1}{\sqrt{n}} X & 0_p \end{bmatrix}, \tag{19}$$

where $I_p, 0_p \in \mathbb{R}^{p \times p}$ are the identity and zero matrix, respectively. We introduce the following resolvents (as usual, these are defined for $\Im(\xi) > 0$ and by analytic continuation, whenever possible, for $\Im(\xi) = 0$):

$$\begin{aligned}
m_{1,n}(\xi, s, t) &= \mathbb{E}\Big\{ \big(A(s,t) - \xi I_N\big)^{-1}_{1,1} \Big\} = \mathbb{E} M_{1,n}(\xi, s, t), \\
M_{1,n}(\xi, s, t) &= \frac{1}{p} \mathrm{tr}_{[1,p]} \Big\{ \big(A(s,t) - \xi I_N\big)^{-1} \Big\}, \\
m_{2,n}(\xi, s, t) &= \mathbb{E}\Big\{ \big(A(s,t) - \xi I_N\big)^{-1}_{p+1,p+1} \Big\} = \mathbb{E} M_{2,n}(\xi, s, t), \\
M_{2,n}(\xi, s, t) &= \frac{1}{n} \mathrm{tr}_{[p+1,p+n]} \Big\{ \big(A(s,t) - \xi I_N\big)^{-1} \Big\}.
\end{aligned}$$

Here and henceforth, we write $[i,j] = \{i+1, \ldots, i+j\}$ for integers i, j. We also write $M^{-1}_{ij} = (M^{-1})_{ij}$ for a matrix M, and $\mathrm{tr}_S(M) = \sum_{i \in S} M_{ii}$ for a subset S. The equalities in the first and third lines above follow by invariance of the distribution of $A(s,t)$ under permutations of $[1,p]$ and $[p+1, p+n]$. Whenever clear from the context, we will omit the arguments from block matrix and resolvents, and write $A = A(s,t)$, $m_{1,n} = m_{1,n}(\xi, s, t)$, and $m_{2,n} = m_{2,n}(\xi, s, t)$.

The next theorem characterizes the asymptotics of $m_{1,n}$, $m_{2,n}$.

Theorem 8. *Assume the conditions of Theorem 7. Consider $\Im(\xi) > 0$ or $\Im(\xi) = 0$, $\Re(\xi) < 0$, with $s \geq t \geq 0$. Let m_1 and m_2 be the unique solutions of the following fourth degree equations:*

$$m_2 = \bigg(-\xi - \gamma m_1 + \frac{\gamma c_1 m_1^2 (c_1 m_2 - t)}{m_1(c_1 m_2 - t) - \psi} \bigg)^{-1}, \tag{20}$$

$$m_1 = \bigg(-\xi - s - \frac{t^2}{\psi} m_1 - m_2 + \frac{t^2 \psi^{-1} m_1^2 (c_1 m_2 - t) - 2 t c_1 m_1 m_2 + c_1^2 m_1 m_2^2}{m_1(c_1 m_2 - \psi) - \psi} \bigg)^{-1}, \tag{21}$$

subject to the condition of being analytic functions for $\Im(z) > 0$, and satisfying $|m_1(z,s,t)|, |m_2(z,s,t)| \leq 1/\Im(z)$ for $\Im(z) > C$ (with C a sufficiently large constant). Then, as $n, p, d \to \infty$, such that $p/n \to \gamma$ and $d/p \to \psi$, we have almost surely (and in L^1),

$$\lim_{n,p,d \to \infty} M_{1,n}(\xi, s, t) = m_1(\xi, s, t), \tag{22}$$

$$\lim_{n,p,d \to \infty} M_{2,n}(\xi, s, t) = m_2(\xi, s, t). \tag{23}$$

The proof of this theorem is given in Appendix B.1. Now define

$$S_n(z) = \frac{1}{p} \mathrm{tr}\big((\hat{\Sigma} - z I_p)^{-1} \big), \tag{24}$$

where recall $\hat{\Sigma} = X^T X / n$. As a corollary of the above, we obtain the asymptotic Stieltjes transform of the eigenvalue distribution of $\hat{\Sigma}$. This confirms a result previously obtained by Pennington and Worah (2017).

Corollary 5. *Assume the conditions of Theorem 7. Consider $\Im(\xi) > 0$. As $n, p, d \to \infty$, with $p/n \to \gamma$ and $d/p \to \psi$, the Stieltjes transform of spectral distribution of $\hat{\Sigma}$ in (24) satisfies almost surely (and in L^1) $S_n(\xi) \to s(\xi)$ where s is a nonrandom function that uniquely solves the following equations (abbreviating $s = s(\xi)$):*

$$-1 - \xi^2 s = \overline{m}_1 \overline{m}_2 - \frac{c_1^2 \overline{m}_1^2 \overline{m}_2^2}{c_1 \overline{m}_1 \overline{m}_2 - \psi}, \tag{25}$$

$$\overline{m}_1 = \xi s, \tag{26}$$

$$\overline{m}_2 = \frac{\gamma - 1}{\xi} + \gamma \xi s, \tag{27}$$

subject to the condition of being analytic for $\Im(z) > 0$, and satisfying $|s(z^2)| < 1/\Im(z^2)$ for $\Im(z) > C$ (where C is a large enough constant). When $c_1 = 0$, the function s is the Stieltjes transform of the Marchenko-Pastur distribution.

We refer to Appendix B.3 for a proof of this corollary. The next lemma connects the above resolvents to the variance of min-norm least squares.

Lemma 6. *Assume the conditions of Theorem 7. Let m_1, m_2 be the asymptotic resolvents given in Theorem 8. Define*

$$m(\xi, s, t) = \gamma m_1(\xi, s, t) + m_2(\xi, s, t).$$

Then for $\gamma \neq 1$, the following Taylor-Laurent expansion holds around $\xi = 0$:

$$-\partial_x m(\xi, x, c_1 x)\big|_{x=0} = \frac{D_{-1}}{\xi^2} + D_0 + O(\xi^2), \tag{28}$$

with each $D_i = D_i(\gamma, \psi, c_1)$. Furthermore, for the ridge regression estimator $\hat{\beta}_\lambda$ in (5), as $n, p, d \to \infty$, such that $p/n \to \gamma \in (0, \infty)$, $d/p \to \psi \in (0, 1)$, the following ridgeless limit holds almost surely:

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} V_X(\hat{\beta}_\lambda; \beta) = D_0.$$

The proof of this lemma can be found in Appendix B.2. Theorem 7 follows by evaluating the formula in Lemma 6, by using the result of Theorem 8. We refer to Appendix B.4 for details.

Acknowledgements

The authors are grateful to Brad Efron, Rob Tibshirani, and Larry Wasserman for inspiring us to work on this in the first place. RT thanks Edgar Dobriban for many helpful conversations about the random matrix theory literature, in particular, the literature trail leading up to Theorem 1, and the reference to Rubio and Mestre (2011) which helped simplify the proof of Lemma 2. TH was partially supported by the grants NSF DMS-1407548, NSF IIS-1837931, and NIH 5R01-EB001988-21. AM was partially supported by NSF DMS-1613091, NSF CCF-1714305, NSF IIS-1741162, and ONR N00014-18-1-2729. RT was partially supported by NSF DMS-1554123.

A Proofs for the linear model

A.1 Proof of Lemma 2

Recall the expression for the bias from Lemma 1 (where now $\Sigma = I$), and note the following key characterization of the pseudoinverse of a rectangular matrix A,

$$(A^T A)^+ A^T = \lim_{z \to 0^+} (A^T A + zI)^{-1} A^T. \tag{29}$$

We can apply this to $A = X/\sqrt{n}$, and rewrite the bias as

$$\begin{aligned} B_X(\hat{\beta}; \beta) &= \lim_{z \to 0^+} \beta^T \big(I - (\hat{\Sigma} + zI)^{-1} \hat{\Sigma}\big) \beta \\ &= \lim_{z \to 0^+} z \beta^T (\hat{\Sigma} + zI)^{-1} \beta, \end{aligned} \tag{30}$$

where in the second line we added and subtracted zI to $\hat{\Sigma}$ and simplified. By Theorem 1 in Rubio and Mestre (2011), which may be seen as a generalized Marchenko-Pastur theorem, we have that for any $z > 0$, and any deterministic sequence of matrices $\Theta_n \in \mathbb{R}^{p \times p}$, $n = 1, 2, 3, \ldots$ with uniformly bounded trace norm, it holds as $n, p \to \infty$, almost surely,

$$\mathrm{tr}\Big(\Theta_n \big((\hat{\Sigma} + zI)^{-1} - c_n(z) I\big)\Big) \to 0, \tag{31}$$

for a deterministic sequence $c_n(z) > 0$, $n = 1, 2, 3, \ldots$ (defined for each n via a certain fixed-point equation). Taking $\Theta_n = I/p$ in the above, note that this reduces to the almost sure convergence of the Stieltjes transform of the specral distribution of $\hat{\Sigma}$, and hence by the (classical) Marchenko-Pastur theorem, we learn that $c_n(z) \to m(-z)$, where m denotes the Stieltjes transform of the Marchenko-Pastur law F_γ. Further, taking $\Theta_n = \beta\beta^T/p$, we see from (31) and $c_n(z) \to m(-z)$ that, almost surely,

$$z \beta^T (\hat{\Sigma} + zI)^{-1} \beta \to z m(-z) r^2. \tag{32}$$

Define $f_n(z) = z\beta^T(\hat{\Sigma} + zI)^{-1}\beta$. Notice that $|f_n(z)| \leq r^2$, and $f_n'(z) = \beta^T(\hat{\Sigma} + zI)^{-2}\hat{\Sigma}\beta$, so

$$|f_n'(z)| \leq r^2 \frac{\lambda_{\max}(\hat{\Sigma})}{(\lambda_{\min}^+(\hat{\Sigma}) + z)^2} \leq 8r^2 \frac{(1+\sqrt{\gamma})^2}{(1-\sqrt{\gamma})^4},$$

where $\lambda_{\max}(\hat{\Sigma})$ and $\lambda_{\min}^+(\hat{\Sigma})$ denote the largest and smallest nonzero eigenvalues, respectively, of $\hat{\Sigma}$, and the second inequality holds almost surely for large enough n, by the Bai-Yin theorem (Bai and Yin, 1993). As its derivatives are bounded, the sequence f_n, $n = 1, 2, 3, \ldots$ is equicontinuous, and by the Arzela-Ascoli theorem, we deduce that f_n converges uniformly to its limit. By the Moore-Osgood theorem, we can exchange limits (as $n, p \to \infty$ and $z \to 0^+$) and conclude from (30), (32) that as $n, p \to \infty$, almost surely,

$$B_X(\hat{\beta}; \beta) \to r^2 \lim_{z \to 0^+} zm(-z).$$

Finally, relying on the fact that the Stieltjes transform of the Marchenko-Pastur law has the explicit form in (7), we can compute the above limit:

$$\begin{aligned} \lim_{z \to 0^+} zm(-z) &= \lim_{z \to 0^+} \frac{-(1-\gamma+z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2\gamma} \\ &= \frac{-(1-\gamma) + (\gamma - 1)}{2\gamma} = 1 - 1/\gamma, \end{aligned}$$

completing the proof.

A.2 Proof of Theorem 3

The asymptotic risk as stated in the theorem can be obtained by taking a limit as the ridge tuning parameter λ tends to zero in Theorem 2.1 in Dobriban and Wager (2018). But some care must be taken in exchanging limits (as $n, p \to \infty$ and $\lambda \to 0^+$) in order to formally conclude a limiting result for min-norm least squares. In what follows, we essentially reproduce the arguments of Dobriban and Wager (2018), just because the way we decompose terms allows us to more easily manage the exchange of limits in the end.

First we give a few notes on conditions. The assumption on a finite 12th moment for the entries of z (where recall, $x \sim P_x$ is of the form $x = \Sigma^{1/2}z$) is needed to invoke a result of Ledoit and Peche 2011 on the convergence of trace functionals involving $\hat{\Sigma}, \Sigma$. The assumption of boundedness of $\lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma)$ is needed to exchange limits in the calculation of the asymptotic bias and variance. In particular, recalling that we have $X = Z\Sigma^{1/2}$ for a matrix Z with i.i.d. entries, the following facts are helpful:

$$\lambda_{\min}^+(\hat{\Sigma}) \geq \lambda_{\min}^+(Z^TZ/n)\lambda_{\min}(\Sigma) \geq (c/2)(1-\sqrt{\gamma})^2, \tag{33}$$

$$\lambda_{\max}(\hat{\Sigma}) \leq \lambda_{\max}(Z^TZ/n)\lambda_{\max}(\Sigma) \leq 2C(1+\sqrt{\gamma})^2, \tag{34}$$

where the second inequality in both lines holds almost surely for large enough n, by the Bai-Yin theorem (Bai and Yin, 1993). The assumption of boundedness of $\lambda_{\min}(\Sigma)$ is also sufficient to prove the existence of the limits $v(0), v'(0)$, via (33), which gives us a lower bound on the support of the density $dF_{H,\gamma}^+/ds$, where $F_{H,\gamma}^+$ denotes the positive part of the empirical spectral distribution (where the point mass at zero has been removed).

Now we analyze the bias from Lemma 4. Applying the key pseudoinverse fact (29), with $A = X/\sqrt{n}$, we have

$$\begin{aligned} B_X(\hat{\beta}) &= \lim_{z \to 0^+} \frac{r^2}{p} \mathrm{tr}\big((I - (\hat{\Sigma} + zI)^{-1}\hat{\Sigma})\Sigma\big) \\ &= \lim_{z \to 0^+} \frac{r^2}{p} z \,\mathrm{tr}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big), \end{aligned}$$

where in the second line we added and subtracted zI to $\hat{\Sigma}$, and in the third line we expanded and simplified. For each $z > 0$, by Lemma 2 in Ledoit and Peche (2011), as $n, p \to \infty$, we have almost surely

$$\frac{r^2}{p} z \,\mathrm{tr}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big) \to r^2\phi(z), \tag{35}$$

where

$$\phi(z) = \frac{1}{\gamma}\bigg(\frac{1}{zv(-z)} - 1\bigg).$$

Recall that we abbreviate $v = v_{F_{H,\gamma}}$ for the companion Stieltjes transform of the empirical spectral distribution $F_{H,\gamma}$ given by the Marchenko-Pastur theorem, and we write v' for its derivative. Let $f_n(z) = (1/p)z\mathrm{tr}((\hat{\Sigma} + zI)^{-1}\Sigma)$, and observe $|f_n(z)| \leq \lambda_{\max}(\Sigma) \leq C$. Also, compute $f_n'(z) = (1/p)\mathrm{tr}((\hat{\Sigma} + zI)^{-2}\hat{\Sigma}\Sigma)$, and note

$$|f_n'(z)| \leq \lambda_{\max}(\Sigma)\frac{\lambda_{\max}(\hat{\Sigma})}{(\lambda^+_{\min}(\hat{\Sigma}) + z)^2} \leq 8(C/c)^2\frac{(1+\sqrt{\gamma})^2}{(1-\sqrt{\gamma})^4},$$

where we have used (33), (34), which hold almost surely for large enough n. Boundedness of derivatives implies that f_n, $n = 1, 2, 3, \ldots$ is equicontinuous, so by the Arzela-Ascoli theorem, it converges uniformly to its limit. Hence, we can take $z \to 0^+$ in (35), and by the Moore-Osgood theorem, we can exchange limits (as $n, p \to \infty$ and $z \to 0^+$) to yield, almost surely

$$B_X(\hat{\beta}) \to r^2 \lim_{z \to 0^+} z\phi(z) = \frac{r^2}{\gamma} \lim_{z \to 0^+} \frac{1}{v(-z)}. \tag{36}$$

This gives the first part of the final result.

Next we work on the variance from Lemma 1. We rewrite this as

$$\begin{aligned}
V_X(\hat{\beta}) &= \frac{\sigma^2}{n}\mathrm{tr}(\hat{\Sigma}^+\hat{\Sigma}\hat{\Sigma}^+\Sigma) \\
&= \frac{\sigma^2 p}{n} \lim_{z \to 0^+} \frac{1}{p}\mathrm{tr}\big((\hat{\Sigma} + zI)^{-1}\hat{\Sigma}(\hat{\Sigma} + zI)^{-1}\Sigma\big) \\
&= \frac{\sigma^2 p}{n} \lim_{z \to 0^+} \bigg[\frac{1}{p}\mathrm{tr}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big) - \frac{1}{p}z\,\mathrm{tr}\big((\hat{\Sigma} + zI)^{-2}\Sigma\big)\bigg].
\end{aligned}$$

In the second line we applied the pseudoinverse fact (29) twice, in the third line we added and subtracted zI to $\hat{\Sigma}$, and in the last we simplified. For fixed $z > 0$, first trace term in the last line above has an asymptotic limit given by the Ledoit-Peche result. Furthermore, we can recognize the second trace term as the derivative of the first:

$$-\mathrm{tr}\big((\hat{\Sigma} + zI)^{-2}\Sigma\big) = \frac{d}{dz}\Big\{\mathrm{tr}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big)\Big\}.$$

As argued in Dobriban and Wager (2018), the function in question here, $g_n(z) = \mathrm{tr}((\hat{\Sigma} + zI)^{-1}\Sigma)$, is bounded and analytic, thus we can use Vitali's theorem, which shows as $n, p \to \infty$, almost surely,

$$\frac{1}{p}\mathrm{tr}\big((\hat{\Sigma} + zI)^{-1}\Sigma\big) - \frac{1}{p}z\,\mathrm{tr}\big((\hat{\Sigma} + zI)^{-2}\Sigma\big) \to \phi(z) + z\phi'(z). \tag{37}$$

Define $h_n(z) = (1/p)\mathrm{tr}((\hat{\Sigma} + zI)^{-2}\hat{\Sigma}\Sigma)$ (which is our earlier, more compact representation for the left-hand side above). Since $g_n = f_n'$, we already have a uniform upper bound for $|g_n(z)|$, as computed previously. Moreover, we compute $g_n'(z) = -(2/p)\mathrm{tr}((\hat{\Sigma} + zI)^{-3}\hat{\Sigma}\Sigma)$, and

$$|g_n'(z)| \leq 2\lambda_{\max}(\Sigma)\frac{\lambda_{\max}(\hat{\Sigma})}{(\lambda^+_{\min}(\hat{\Sigma}) + z)^3} \leq 16(C^2/c^3)\frac{(1+\sqrt{\gamma})^2}{(1-\sqrt{\gamma})^4},$$

where we have again used (33), (34), which hold almost surely for sufficiently enough n. As before, boundedness of derivatives means that f_n, $n = 1, 2, 3, \ldots$ is equicontinuous, and the Arzela-Ascoli theorem shows that this sequence converges uniformly to its limit. Therefore, we can take $z \to 0^+$ in (37), and by the Moore-Osgood theorem, we can exchange limits (as $n, p \to \infty$ and $z \to 0^+$) to yield, almost surely,

$$V_X(\hat{\beta}) \to \sigma^2 \lim_{z \to 0^+} \bigg(\frac{v'(-z)}{v(-z)^2} - 1\bigg). \tag{38}$$

This gives the second part of the final result, and adding together (36), (38) completes the proof.

A.3 Translating Theorem 3 for isotropic features

When $\Sigma = I$, the empirical spectral distribution is denoted F_γ and called Marchenko-Pastur law. Recall that this has a closed-form Stieltjes transform, given in (7). A short calculation therefore leads to

$$v(-z) = \frac{-(\gamma - 1 + z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2z}, \tag{39}$$

The limit of $v(-z)$ as $z \to 0^+$ is indeterminate, so we can use l'Hopital's rule to find

$$\begin{aligned} \lim_{z \to 0^+} v(-z) &= \lim_{z \to 0^+} \frac{-1 + \frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}}}{2} \\ &= \frac{-1 + \frac{1+\gamma}{\gamma-1}}{2} = \frac{1}{\gamma - 1}. \end{aligned} \tag{40}$$

Furthermore,

$$\begin{aligned} v'(-z) &= -\frac{-1 + \frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}}}{2z} + \frac{(\gamma - 1 + z) - \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2z^2} \\ &= -\frac{z\frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}} + (\gamma - 1) - \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2z^2}. \end{aligned}$$

As the limit of $v'(-z)$ as $z \to 0^+$ is again indeterminate, we apply l'Hopital's rule once more, yielding

$$\begin{aligned} \lim_{z \to 0^+} v'(-z) &= \lim_{z \to 0^+} -\frac{\frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}} + z\frac{1}{\sqrt{(1-\gamma+z)^2+4\gamma z}} - z\frac{(1+\gamma+z)^2}{((1-\gamma+z)^2+4\gamma z)^{3/2}} - \frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}}}{4z} \\ &= \lim_{z \to 0^+} -\frac{z\frac{(1-\gamma+z)^2+4\gamma z-(1+\gamma+z)^2}{((1-\gamma+z)^2+4\gamma z)^{3/2}}}{4z} = \frac{\gamma}{(\gamma-1)^3}. \end{aligned} \tag{41}$$

Finally, plugging (40) and (41) into the asymptotic risk expression from Theorem 3 gives

$$\frac{r^2}{\gamma}(\gamma - 1) + \sigma^2\left(\frac{\gamma(\gamma-1)^2}{(\gamma-1)^3} - 1\right) = r^2(1 - 1/\gamma) + \frac{\sigma^2}{\gamma - 1},$$

exactly as in Theorem 2, as claimed.

A.4 Proof of Corollary 2

Let H_ρ denote the weak limit of F_Σ, when Σ has ρ-equicorrelation structure for all n, p. A short calculation shows that such a matrix Σ has one eigenvalue value equal to $1 + (p-1)\rho$, and $p - 1$ eigenvalues equal to $1 - \rho$. Thus the weak limit of its spectral measure is simply $H_\rho = 1_{[1-\rho,\infty)}$, i.e., $dH_\rho = \delta_{1-\rho}$, a point mass at $1 - \rho$ of probability one.

We remark that the present case, strictly speaking, breaks the conditions that we assume in Theorem 3, because $\lambda_{\max}(\Sigma) = 1 + (p-1)\rho$ clearly diverges with p. However, by decomposing $\Sigma = (1-\rho)I + \rho\mathbb{1}\mathbb{1}^T$ (where $\mathbb{1}$ denotes the vector of all 1s), and correspondingly decomposing the functions f_n, g_n, h_n defined in the proof of Theorem 3, to handle the rank one part $\rho\mathbb{1}\mathbb{1}^T$ properly, we can ensure the appropriate boundedness conditions. Thus, the result in the theorem still holds when Σ has ρ-equicorrelation structure.

Now denote by v_ρ the companion Stieltjes transform of the empirical spectral distribution $F_{H_\rho,\gamma}$, to emphasize its dependence on ρ. Recall the Silverstein equation (42), which for the equicorrelation case, as $dH_\rho = \delta_{1-\rho}$, becomes

$$-\frac{1}{v_\rho(z)} = z - \gamma\frac{1-\rho}{1 + (1-\rho)v_\rho(z)},$$

or equivalently,

$$-\frac{1}{(1-\rho)v_\rho(z)} = \frac{z}{1-\rho} - \gamma\frac{1}{1 + (1-\rho)v_\rho(z)}.$$

We can hence recognize the relationship

$$(1-\rho)v_\rho(z) = v_0\big(z/(1-\rho)\big),$$

where v_0 is the companion Stieltjes transform in the $\Sigma = I$ case, the object of study in Appendix A.3. From the results for v_0 from (40) and (41), invoking the relationship in the above display, we have

$$\lim_{z\to 0^+} v_\rho(-z) = \frac{1}{(1-\rho)(\gamma-1)} \quad \text{and} \quad \lim_{z\to 0^+} v_\rho'(-z) = \frac{\gamma}{(1-\rho)^2(\gamma-1)^3}.$$

Plugging these into the asymptotic risk expression from Theorem 3 gives

$$\frac{r^2}{\gamma}(1-\rho)(\gamma-1) + \sigma^2\left(\frac{\gamma(1-\rho)^2(\gamma-1)^2}{(1-\rho)^2(\gamma-1)^3} - 1\right) = r^2(1-\rho)(1-1/\gamma) + \frac{\sigma^2}{\gamma-1}.$$

as claimed.

A.5 Autoregressive features

We consider a *ρ-autoregressive* structure for Σ, for a constant $\rho \in [0,1)$, meaning that $\Sigma_{ij} = \rho^{|i-j|}$ for all i, j. In this case, it is not clear that a closed-form exists for $v(0)$ or $v'(0)$. However, we can compute these numerically. In fact, the strategy we describe below applies to any situation in which we are able to perform numerical integration against dH, where H is the weak limit of the spectral measure F_Σ of Σ.

The critical relationship that we use is the *Silverstein equation* (Silverstein, 1995), which relates the companion Stieltjes transform v to H via

$$-\frac{1}{v(z)} = z - \gamma\int \frac{s}{1+sv(z)}\,dH(s). \tag{42}$$

Taking $z \to 0^+$ yields

$$\frac{1}{v(0)} = \gamma\int \frac{s}{1+sv(0)}\,dH(s). \tag{43}$$

Therefore, we can use a simple univariate root-finding algorithm (like the bisection method) to solve for $v(0)$ in (43). With $v(0)$ computed, we can compute $v'(0)$ by first differentiating (42) with respect to z (see Dobriban 2015), and then taking $z \to 0^+$, to yield

$$\frac{1}{v'(0)} = \frac{1}{v(0)^2} - \gamma\int \frac{s^2}{(1+sv(0))^2}\,dH(s). \tag{44}$$

When Σ is of ρ-autoregressive form, it is known to have eigenvalues (Trench, 1999):

$$s_i = \frac{1-\rho^2}{1-2\rho\cos(\theta_i)+\rho^2}, \quad i = 1,\dots,p,$$

where

$$\frac{(p-i)\pi}{p+1} < \theta_i < \frac{(p-i+1)\pi}{p+1}, \quad i = 1,\dots,p.$$

This allows us to efficiently approximate an integral with respect to dH (e.g., by taking each θ_i to be in the midpoint of its interval given above), solve for $v(0)$ in (43), $v'(0)$ in (44), and evaluate the asymptotic risk as per Theorem 3.

Figure 10 shows the results from using such a numerical scheme to evaluate the asymptotic risk, as ρ varies from 0 to 0.75.

A.6 Proof of Theorem 6

We begin by recalling an alternative to leave-one-out cross-validation, for linear smoothers, called *generalized cross-validation* (GCV) (Craven and Wahba, 1978; Golub et al., 1979). The GCV error of the ridge regression estimator at a tuning parameter value λ defined as

$$\mathrm{GCV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \hat{f}_\lambda(x_i)}{1-\mathrm{tr}(S_\lambda)/n}\right)^2. \tag{45}$$

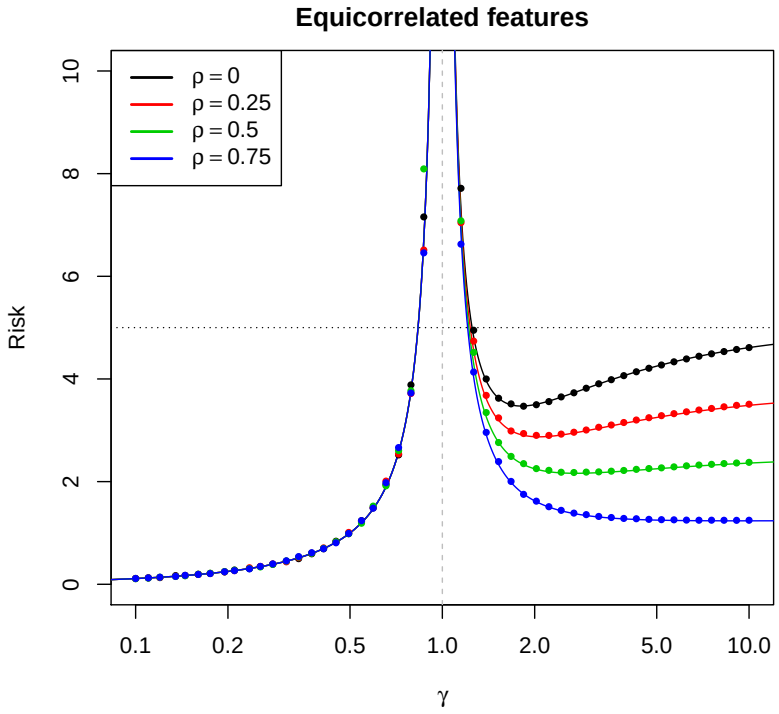

Figure 9: Asymptotic risk curves for the min-norm least squares estimator when Σ has equicorrelation structure (Theorem 1 for $\gamma < 1$, and Corollary 2 for $\gamma > 1$), as ρ varies from 0 to 0.75. Here $r^2 = 5$ and $\sigma^2 = 1$, thus $\mathrm{SNR} = 5$. The null risk $r^2 = 5$ is marked as a dotted black line. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from appropriately constructed Gaussian features.

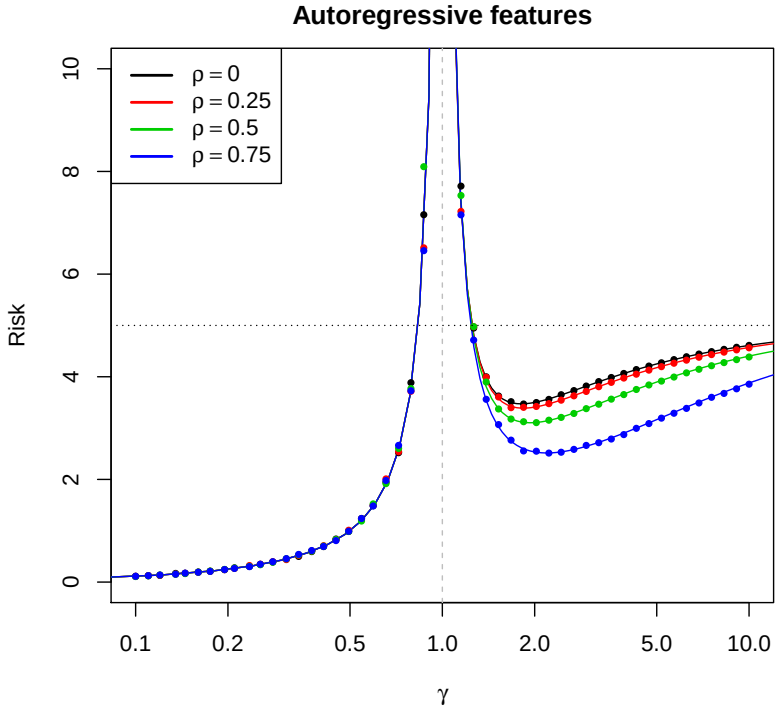

Figure 10: Asymptotic risk curves for the min-norm least squares estimator when Σ has autoregressive structure (Theorem 1 for $\gamma < 1$, and Theorem 3 for $\gamma > 1$, evaluated numerically, as described in Appendix A.5), as ρ varies from 0 to 0.75. Here $r^2 = 5$ and $\sigma^2 = 1$, thus $\mathrm{SNR} = 5$. The null risk $r^2 = 5$ is marked as a dotted black line. The points are again finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from appropriately constructed Gaussian features.

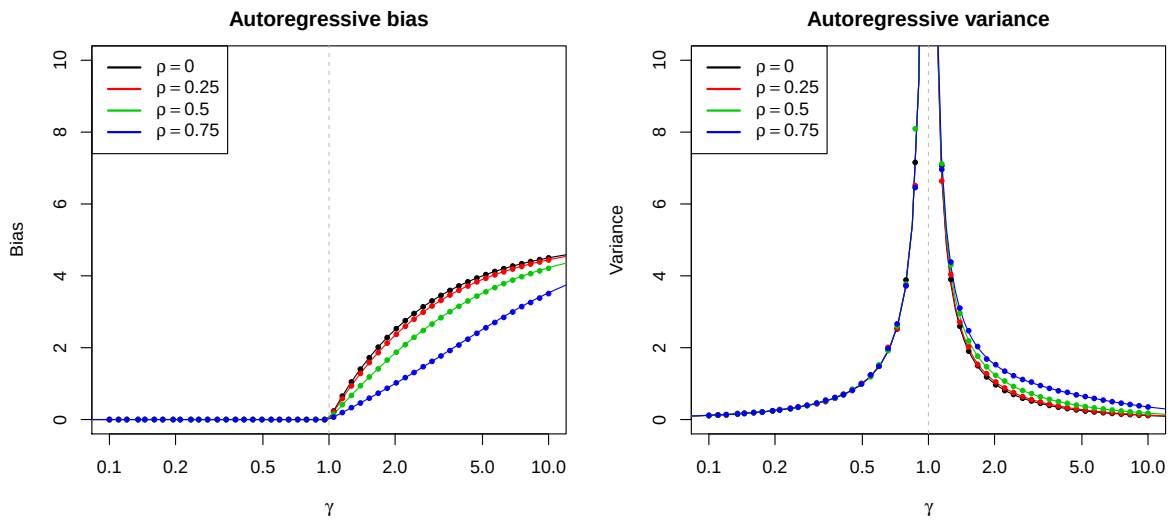

Figure 11: Asymptotic bias (left panel) and variance (right panel) for the min-norm least squares estimator when Σ has autoregressive structure (Theorem 1 for $\gamma < 1$, and Theorem 3 for $\gamma > 1$, evaluated numerically, as described in Appendix A.5), as ρ varies from 0 to 0.75. Here $r^2 = 5$ and $\sigma^2 = 1$, hence $\mathrm{SNR} = 5$. The points mark finite-sample biases and variances, with $n = 200$, $p = [\gamma n]$, computed from appropriately constructed Gaussian features.

Compared to the shortcut formula for leave-one-out CV in (16), we can see that GCV in (45) swaps out the ith diagonal element $(S_\lambda)_{ii}$ in the denominator of each summand with the average diagonal element $\mathrm{tr}(S_\lambda)/n$. This modification makes GCV rotationally invariant (Golub et al., 1979).

It turns out that the GCV error is easier to analyze, compared to the CV error. Thus we proceed by first studying GCV, and then relating CV to GCV. We break up the exposition below into these two parts accordingly.

A.6.1 Analysis of GCV

Let us rewrite the GCV criterion in (45) as

$$\mathrm{GCV}_n(\lambda) = \frac{y^T (I - S_\lambda)^2 y / n}{(1 - \mathrm{tr}(S_\lambda)/n)^2}. \tag{46}$$

We will treat the almost sure convergence of the numerator and denominator separately.

GCV denominator. The denominator is an easier calculation. Denoting $s_i = \lambda_i(X^T X/n)$, $i = 1, \ldots, p$, we have

$$\mathrm{tr}(S_\lambda)/n = \frac{1}{n} \sum_{i=1}^{p} \frac{s_i}{s_i + \lambda} \to \gamma \int \frac{s}{s + \lambda} \, dF_\gamma(s),$$

where this convergence holds almost surely as $n, p \to \infty$, a direct consequence of the Marchenko-Pastur theorem, and F_γ denotes the Marchenko-Pastur law. Meanwhile, we can rewrite this asymptotic limit as

$$\gamma \int \frac{s}{s + \lambda} \, dF_\gamma(s) = \gamma\big(1 - m(-\lambda)\big),$$

where $m = m_{F_\gamma}$ denotes the Stieltjes transform of the Marchenko-Pastur law F_γ, and therefore, almost surely,

$$\big(1 - \mathrm{tr}(S_\lambda)/n\big)^2 \to \Big(1 - \gamma\big(1 - m(-\lambda)\big)\Big)^2. \tag{47}$$

GCV numerator. The numerator requires only a bit more difficult calculation. Let $y = X\beta + \epsilon$ and $c_n = \sqrt{p}(\sigma/r)$. Observe

$$\begin{aligned} y^T(I - S_\lambda)^2 y/n &= (\beta, \epsilon)^T \left(\frac{1}{n} \begin{bmatrix} X \\ I \end{bmatrix}^T (I - S_\lambda)^2 \begin{bmatrix} X \\ I \end{bmatrix} \right) (\beta, \epsilon) \\ &= \underbrace{(\beta, \epsilon/c_n)^T}_{\delta^T} \underbrace{\left(\frac{1}{n} \begin{bmatrix} X \\ c_n I \end{bmatrix}^T (I - S_\lambda)^2 \begin{bmatrix} X \\ c_n I \end{bmatrix} \right)}_{A} \underbrace{(\beta, \epsilon/c_n)}_{\delta}. \end{aligned}$$

Note that δ has independent entries with mean zero and variance r^2/p, and further, note that δ and A are independent. Therefore we can use the almost sure convergence of quadratic forms, from Lemma 7.6 in Dobriban and Wager (2018), which is adapted from Lemma B.26 in Bai and Silverstein (2010).[2] This result asserts that, almost surely,

$$\delta^T A \delta - (r^2/p)\mathrm{tr}(A) \to 0.$$

Now examine

$$\begin{aligned} r^2 \mathrm{tr}(A)/p &= \frac{r^2}{p} \mathrm{tr}\big((I - S_\lambda)^2 (XX^T/n + (c_n^2/n) I\big) \\ &= \underbrace{\frac{r^2}{p} \mathrm{tr}\big(X^T (I - S_\lambda)^2 X/n\big)}_{a} + \underbrace{\frac{\sigma^2}{n} \mathrm{tr}\big((I - S_\lambda)^2\big)}_{b}. \end{aligned}$$

A short calculation and application of the Marchenko-Pastur theorem gives that, almost surely,

$$a = \frac{r^2\lambda^2}{p} \left(\sum_{i=1}^p \frac{1}{s_i + \lambda} - \lambda \sum_{i=1}^p \frac{1}{(s_i + \lambda)^2} \right) \to r^2\lambda^2 \big(m(-\lambda) - \lambda m'(-\lambda)\big),$$

where for the second sum, we used Vitali's theorem to show convergence of the derivative of the Stieltjes transform of the spectral distribution of $X^T X/n$ to the derivative of the Stieltjes transform of F_γ (note that Vitali's theorem applies as the function in question is bounded and analytic). By a similar calculation, we have almost surely,

$$b = \frac{\sigma^2}{n} \left(\sum_{i=1}^p \frac{\lambda^2}{(s_i + \lambda)^2} + (n - p) \right) \to \sigma^2 \gamma \lambda^2 m'(-\lambda) + \sigma^2(1 - \gamma).$$

Hence we have shown that, almost surely,

$$y^T(I - S_\lambda)^2 y/n \to \lambda^2 \Big(r^2 \big(m(-\lambda) - \lambda m'(-\lambda)\big) - \sigma^2 \gamma m'(-\lambda) \Big) + \sigma^2(1 - \gamma). \tag{48}$$

GCV convergence. Putting (47), (48) together with (46), we have, almost surely,

$$\mathrm{GCV}_n(\lambda) \to \frac{\lambda^2(r^2(m(-\lambda) - \lambda m'(-\lambda)) - \sigma^2 \gamma m'(-\lambda)) + \sigma^2(1 - \gamma)}{(1 - \gamma(1 - m(-\lambda)))^2}.$$

To show that this matches to the asymptotic prediction error of ridge regression requires some nontrivial calculations. We start by reparametrizing the above asymptotic limit in terms of the companion Stieltjes transform, abbreviated by $v = v_{F_\gamma}$. This satisfies

$$v(z) + 1/z = \gamma\big(m(z) + 1/z\big),$$

hence

$$zv(-z) - 1 = \gamma\big(zm(-z) - 1\big),$$

[2] As written, Lemma 7.6 of Dobriban and Wager (2018) assumes i.i.d. components for the random vector in question, which is not necessarily true of δ in our case. However, an inspection of their proof shows that they only require independent components with mean zero and common variance, which is precisely as stated in Lemma B.26 of Bai and Silverstein (2010).

and also

$$z^2 v'(-z) - 1 = \gamma\big(z^2 m'(-z) - 1\big).$$

Introducing $\alpha = r^2/(\sigma^2\gamma)$, the almost sure limit of GCV is

$$\begin{aligned}
&\frac{\lambda^2(r^2(m(-\lambda) - \lambda m'(-\lambda)) - \sigma^2\gamma m'(-\lambda)) + \sigma^2(1-\gamma)}{(1-\gamma(1-m(-\lambda)))^2} \\
&\quad = \frac{\sigma^2\lambda(\alpha\gamma(\lambda m(-\lambda) - 1) - \alpha\gamma(\lambda^2 m'(-\lambda) - 1) + (\gamma/\lambda)(\lambda^2 m'(-\lambda) - 1) + \gamma/\lambda + (1-\gamma)/\lambda)}{(1-\gamma(1-m(-\lambda)))^2} \\
&\quad = \frac{\sigma^2\lambda(\alpha(\lambda v(-\lambda) - 1) - \alpha(\lambda^2 v'(-\lambda) - 1) + (1/\lambda)(\lambda^2 v'(-\lambda) - 1) + 1/\lambda)}{\lambda^2 v(-\lambda)^2} \\
&\quad = \frac{\sigma^2(v(-\lambda) + (\alpha\lambda - 1)(v(-\lambda) - \lambda v'(-\lambda))}{\lambda v(-\lambda)^2},
\end{aligned}$$

where in the second line we rearranged, in the third line we applied the companion Stieltjes transform facts, and in the fourth line we simplified. We can now recognize the above as σ^2 plus the asymptotic risk of ridge regression at tuning parameter λ, either from the proof of Theorem 3, or from Theorem 2.1 in Dobriban and Wager (2018). In terms of the Stieltjes transform itself, this is $\sigma^2 + \sigma^2\gamma(m(-\lambda) - \lambda(1-\alpha\lambda)m'(-\lambda))$, which proves the first claimed result.

Uniform convergence. It remains to prove the second claimed result, on the convergence of the GCV-tuned ridge estimator. Denote $f_n(\lambda) = \mathrm{GCV}_n(\lambda)$, and $f(\lambda)$ for its almost sure limit. Notice that $|f_n|$ is almost surely bounded on $[\lambda_1, \lambda_2]$, for large enough n, as

$$\begin{aligned}
|f_n(\lambda)| &\leq \frac{\|y\|_2^2}{n} \frac{\lambda_{\max}(I - S_\lambda)^2}{(1 - \mathrm{tr}(S_\lambda)/n)^2} \\
&\leq \frac{\|y\|_2^2}{n} \frac{(s_{\max} + \lambda)^2}{\lambda^2} \\
&\leq 2(r^2 + \sigma^2)\frac{(2 + \lambda_2)^2}{\lambda_1^2}.
\end{aligned}$$

In the second line, we used $\mathrm{tr}(S_\lambda)/n = (1/n)\sum_{i=1}^p s_i/(s_i + \lambda) \leq s_{\max}/(s_{\max} + \lambda)$, with $s_{\max} = \lambda_{\max}(X^T X/n)$, and in the third line, we used $\|y\|_2^2/n \leq 2(r^2 + \sigma^2)$ almost surely for sufficiently large n, by the strong law of large numbers, and $s_{\max} \leq 2$ almost surely for sufficiently large n, by the Bai-Yin theorem (Bai and Yin, 1993). Furthermore, writing g_n, h_n for the numerator and denominator of f_n, respectively, we have

$$f_n'(\lambda) = \frac{g_n'(\lambda)h_n(\lambda) - g_n(\lambda)h_n'(\lambda)}{h_n(\lambda)^2}.$$

The above argument just showed that $|g_n(\lambda)|$ is upper bounded on $[\lambda_1, \lambda_2]$, and $|h_n(\lambda)|$ is lower bounded on $[\lambda_1, \lambda_2]$; also, clearly $|h_n(\lambda)| \leq 1$; therefore to show that $|f_n'|$ is almost surely bounded on $[\lambda_1, \lambda_2]$, it suffices to show that both $|g_n'|, |h_n'|$ are. Denoting by $u_i, i = 1, \ldots, p$ the eigenvectors of $X^T X/n$ (corresponding to eigenvalues $s_i, i = 1, \ldots, p$), a short calculation shows

$$|g_n'(\lambda)| = \frac{2\lambda}{n}\sum_{i=1}^p (u_i^T y)^2 \frac{s_i}{(s_i + \lambda)^3} \leq \frac{2\lambda}{n}\|y\|_2^2 \leq 4\lambda(r^2 + \sigma^2),$$

the last step holding almost surely for large enough n, by the law of large numbers. Also,

$$|h_n'(\lambda)| = 2\bigg(1 - \frac{1}{n}\sum_{i=1}^n \frac{s_i}{s_i + \lambda}\bigg)\frac{1}{n}\sum_{i=1}^n \frac{s_i}{(s_i + \lambda)^2} \leq \frac{4}{\lambda_1^2},$$

the last step holding almost surely for large enough n, by the Bai-Yin theorem. Thus we have shown that $|f_n'|$ is almost surely bounded on $[\lambda_1, \lambda_2]$, for large enough n, and applying the Arzela-Ascoli theorem, f_n converges uniformly to f. With $\lambda_n = \arg\min_{\lambda \in [\lambda_1, \lambda_2]} f_n(\lambda)$, this means for any for any $\lambda \in [\lambda_1, \lambda_2]$, almost surely

$$\begin{aligned}
f(\lambda_n) - f(\lambda) &= \big(f(\lambda_n) - f_n(\lambda_n)\big) + \big(f_n(\lambda_n) - f_n(\lambda)\big) + \big(f_n(\lambda) - f(\lambda)\big) \\
&\leq \big(f(\lambda_n) - f_n(\lambda_n)\big) + \big(f_n(\lambda) - f(\lambda)\big) \to 0,
\end{aligned}$$

where we used the optimality of λ_n for f_n, and then uniform convergence. In other words, almost surely,

$$f(\lambda_n) \to f(\lambda^*) = \sigma^2 + \sigma^2 \gamma m(-1/\alpha).$$

As the almost sure convergence $R_X(\hat{\beta}_\lambda) + \sigma^2 \to f(\lambda)$ is also uniform for $\lambda \in [\lambda_1, \lambda_2]$ (by similar arguments, where we bound the risk and its derivative in λ), we conclude that almost surely $R_X(\hat{\beta}_\lambda) \to \sigma^2 \gamma m(-1/\alpha)$, completing the proof for GCV.

A.6.2 Analysis of CV

Let us rewrite the CV criterion, starting in its shortcut form (16), as

$$\mathrm{CV}_n(\lambda) = y^T (I - S_\lambda) D_\lambda^{-2} (I - S_\lambda) y/n, \tag{49}$$

where D_λ is a diagonal matrix that has diagonal elements $(D_\lambda)_{ii} = 1 - (S_\lambda)_{ii}$, $i = 1, \ldots, n$.

CV denominators. First, fixing an arbitrary $i = 1, \ldots, n$, we will study the limiting behavior of

$$1 - (S_\lambda)_{ii} = 1 - x_i^T (X^T X/n + \lambda I)^{-1} x_i/n.$$

Since x_i and $X^T X$ are not independent, we cannot immediately apply the almost sure convergence of quadratic forms lemma, as we did in the previous analysis of GCV. But, letting X_{-i} denote the matrix X with the ith row removed, we can write $(X^T X/n + \lambda I)^{-1} = (X_{-i}^T X_{-i}/n + \lambda I + x_i x_i^T/n)^{-1}$, and use the Sherman-Morrison-Woodbury formula to separate out the dependent and independent parts, as follows. Letting $\delta_i = x_i/\sqrt{n}$, $A_i = (X_{-i}^T X_{-i}/n + \lambda I)^{-1}$, and $A = (X^T X/n + \lambda I)$, we have

$$\begin{aligned}
1 - (S_\lambda)_{ii} &= 1 - \delta_i^T A \delta_i \\
&= 1 - \delta_i^T \bigg(A_i - \frac{A_i \delta_i \delta_i^T A_i}{1 + \delta_i^T A_i \delta} \bigg) \delta_i \\
&= \frac{1}{1 + \delta_i^T A_i \delta_i}.
\end{aligned}$$

Note δ_i and A_i are independent (i.e., x_i and $X_{-i}^T X_{-i}$ are independent), so we can now use the almost sure convergence of quadratic forms, from Lemma 7.6 in Dobriban and Wager (2018), adapted from Lemma B.26 in Bai and Silverstein (2010), to get that, almost surely

$$\delta_i^T A_i \delta_i - \mathrm{tr}(A_i)/n \to 0. \tag{50}$$

Further, as $\mathrm{tr}(A_i)/n \to \gamma m(-\lambda)$ almost surely by the Marchenko-Pastur theorem, we have, almost surely,

$$1 - (S_\lambda)_{ii} \to \frac{1}{1 + \gamma m(-\lambda)}. \tag{51}$$

Replacing denominators, controlling remainders. The strategy henceforth, based on the result in (51), is to replace the denominators $1 - (S_\lambda)_{ii}$, $i = 1, \ldots, n$ in the summands of the CV error by their asymptotic limits, and then control the remainder terms. More precisely, we define $\bar{D}_\lambda = (1 + \gamma m(-\lambda))^{-1} I$, and then write, from (49),

$$\mathrm{CV}_n(\lambda) = \underbrace{y^T (I - S_\lambda) \bar{D}_\lambda^{-2} (I - S_\lambda) y/n}_{a} + \underbrace{y^T (I - S_\lambda)(D_\lambda^{-2} - \bar{D}_\lambda^{-2})(I - S_\lambda) y/n}_{b}. \tag{52}$$

We will first show that almost surely $b \to 0$. Observe, by the Cauchy-Schwartz inequality,

$$\begin{aligned}
b &\leq \frac{1}{n} \|(I - S_\lambda) y\|_2^2 \, \lambda_{\max}(D_\lambda^{-2} - \bar{D}_\lambda^{-2}) \\
&\leq 2(r^2 + \sigma^2) \underbrace{\max_{i=1,\ldots,n} \Big| (1 + \delta_i^T A_i \delta_i)^2 - \big(1 + \gamma m(-\lambda)\big)^2 \Big|}_{c},
\end{aligned}$$

where in the second step, we used $\|(I - S_\lambda)y\|_2^2/n \leq \|y\|_2^2/n \leq 2(r^2 + \sigma^2)$, which holds almost surely for large enough n, by the strong law of large numbers. Meanwhile,

$$c \leq \underbrace{\max_{i=1,\ldots,n} \left|(1 + \delta_i^T A_i \delta_i)^2 - \big(1 + \mathrm{tr}(A_i)/n\big)^2\right|}_{d_1} + \underbrace{\max_{i=1,\ldots,n} \left|\big(1 + \mathrm{tr}(A_i)/n\big)^2 - \big(1 + \mathrm{tr}(A)/n\big)^2\right|}_{d_2} + \underbrace{\left|\big(1 + \mathrm{tr}(A)/n\big)^2 - \big(1 + \gamma m(-\lambda)\big)^2\right|}_{d_3}.$$

By the Marchenko-Pastur theorem, we have $d_3 \to 0$ almost surely. Using $u^2 - v^2 = (u - v)(u + v)$ on the maximands in d_2,

$$\begin{aligned}
d_2 &= \max_{i=1,\ldots,n} \big|\mathrm{tr}(A_i)/n - \mathrm{tr}(A)/n\big|\big|2 + \mathrm{tr}(A_i)/n + \mathrm{tr}(A)/n\big| \\
&\leq \frac{2(1+1\lambda)}{n} \max_{i=1,\ldots,n} |\mathrm{tr}(A_i - A)| \\
&\leq \frac{2(1+1\lambda)}{n} \max_{i=1,\ldots,n} \frac{|\mathrm{tr}(A\delta_i\delta_i^T A)|}{|1 - \delta_i^T A \delta_i|} \\
&\leq \frac{2(1+1\lambda)}{n} \frac{s_{\max} + \lambda}{\lambda^3} \max_{i=1,\ldots,n} \|\delta_i\|_2^2 \\
&\leq \frac{2(1+1\lambda)}{n} \frac{(s_{\max} + \lambda) s_{\max}}{\lambda^3} \\
&\leq \frac{4(1+1\lambda)(2+\lambda)}{n\lambda^3} \to 0.
\end{aligned}$$

In the second line above, we used $\mathrm{tr}(A_i)/n \leq \lambda_{\max}(A_i) \leq 1/\lambda$, $i = 1, \ldots, n$, and also $\mathrm{tr}(A)/n \leq 1/\lambda$. In the third line, we used the Sherman-Morrison-Woodbury formula. In the fourth line, for each summand $i = 1, \ldots, n$, we upper bounded the numerator by $\lambda_{\max}(A)^2 \|\delta_i\|_2^2 \leq \|\delta_i\|_2^2/\lambda^2$, and lower bounded the denominator by $\lambda/(s_{\max} + \lambda)$, with $s_{\max} = \lambda_{\max}(X^T X/n)$ (which follows from a short calculation using the eigendecomposition of $X^T X/n$). In the fifth line, we used $\|\delta_i\|_2^2 = e_i^T (XX^T/n) e_i \leq s_{\max}$, and in the sixth line, we used $s_{\max} \leq 2$ almost surely for sufficiently large n, by the Bai-Yin theorem (Bai and Yin, 1993).

Now we will show that $d_1 \to 0$ almost surely by showing that the convergence in (50) is rapid enough. As before, using $u^2 - v^2 = (u - v)(u + v)$ on the maximands in d_1, we have

$$\begin{aligned}
d_1 &= \max_{i=1,\ldots,n} \big|\delta_i^T A_i \delta_i - \mathrm{tr}(A_i)/n\big|\big|2 + \delta_i^T A_i \delta_i + \mathrm{tr}(A_i)/n\big|^2 \\
&\leq (2 + 3/\lambda) \max_{i=1,\ldots,n} \underbrace{\big|\delta_i^T A_i \delta_i - \mathrm{tr}(A_i)/n\big|}_{\Delta_i}.
\end{aligned}$$

Here we used that for $i = 1, \ldots, n$, we have $\mathrm{tr}(A_i)/n \leq 1/\lambda$, and similarly, $\delta_i^T A_i \delta_i \leq \|\delta_i\|_2^2/\lambda \leq s_{\max}/\lambda \leq 2/\lambda$, with the last inequality holding almost surely for sufficiently large n, by the Bai-Yin theorem (Bai and Yin, 1993). To show $\max_{i=1,\ldots,n} \Delta_i \to 0$ almost surely, we start with the union bound and Markov's inequality, where $q = 2 + \eta/2$, and t_n is be specified later,

$$\mathbb{P}\bigg(\Big(\max_{i=1,\ldots,n} \Delta_i\Big) > t_n\bigg) \leq \sum_{i=1}^n \mathbb{E}(\Delta_i^{2q}) t_n^{-2q} \leq C\lambda^{-q} n p^{-q} t_n^{-2q}.$$

In the last step, we used that for $i = 1, \ldots, n$, we have $\mathbb{E}(\Delta_i^{2q}) \leq C\lambda_{\max}(A_i)^q p^{-q} \leq C\lambda^{-q} p^{-q}$ for a constant $C > 0$ that depends only on q and the assumed moment bound, of order $2q = 4 + \eta$, on the entries of P_x. This expectation bound is a consequence of the trace lemma in Lemma B.26 of Bai and Silverstein (2010), as explained in the proof of Lemma 7.6 in Dobriban and Wager (2018). Hence choosing $t_n^{-2q} = p^{\eta/4}$, from the last display we have

$$\mathbb{P}\bigg(\Big(\max_{i=1,\ldots,n} \Delta_i\Big) > t_n\bigg) \leq C\lambda^{-q/2} n p^{-2-\eta/2} p^{\eta/4} \leq \frac{2C\lambda^{-q/2}}{\gamma} p^{-1-\eta/4},$$

where in the last step we used $n/p \leq 2/\gamma$ for large enough n. Since the right-hand side above is summable in p, we conclude by the Borel-Cantelli lemma that $\max_{i=1,\ldots,n} \Delta_i \to 0$ almost surely, and therefore also $d_1 \to 0$ almost surely. This finishes the proof that $b \to 0$.

Relating back to GCV. Returning to (52), observe

$$a = y^T (I - S_\lambda)^2 y/n \cdot \big(1 + \gamma m(-\lambda)\big)^2.$$

The first term in the product on the right-hand side above is precisely the numerator in GCV, whose convergence was established in (48). Therefore, to show that the limit of a, i.e., of $\mathrm{CV}_n(\lambda)$, is the same as it was for $\mathrm{GCV}_n(\lambda)$, we will need to show that the second term in the product above matches the limit for the GCV denominator, in (47). That is, we will need to show that

$$\big(1 + \gamma m(-\lambda)\big)^2 = \frac{1}{(1 - \gamma(1 - m(-\lambda)))^2}, \tag{53}$$

or equivalently,

$$1 - \gamma\big(1 - m(-\lambda)\big) = \frac{1}{1 + \gamma m(-\lambda)}.$$

As before, it helps to reparametrize in terms of the companion Stieltjes transform, giving

$$v(-\lambda) = \frac{1}{\lambda + \lambda v(-\lambda) + \gamma - 1},$$

of equivalently,

$$\big(\lambda + \lambda v(-\lambda) + \gamma - 1\big) v(-\lambda) = 1.$$

Adding $v(-\lambda)$ to both sides, then dividing both sides by $1 + v(-\lambda)$, yields

$$\frac{(\lambda + \lambda v(-\lambda) + \gamma) v(-\lambda)}{1 + v(-\lambda)} = 1,$$

and dividing through by $v(-\lambda)$, then rearranging, gives

$$\frac{1}{v(-\lambda)} = \lambda + \frac{\gamma}{1 + v(-\lambda)}.$$

This is precisely the *Silverstein equation* (Silverstein, 1995), which relates the companion Stieltjes transform v to the weak limit H of the spectral measure of the feature covariance matrix Σ, which in the current isotropic case $\Sigma = I$, is just $H = \delta_1$ (a point mass at 1). Hence, we have established the desired claim (53), and from the limiting analysis of GCV completed previously, we have that, almost surely, $a \to \sigma^2 + \sigma^2 \gamma (m(-\lambda) - \lambda(1 - \alpha\lambda) m'(-\lambda))$. This is indeed also the almost sure limit of $\mathrm{CV}_n(\lambda)$, from (52) and $b \to 0$ almost surely. The proof of uniform convergence, and thus convergence of the CV-tuned risk, follows similar arguments to the GCV case, and is omitted.

A.7 Misspecified model results

These results follow because, as argued in Section 5.2, the misspecified case can be reduced to a well-specified case after we make the substitutions in (12).

A.8 CV and GCV simulations

Figures 12 and 13 investigate the effect of using CV and GCV tuning for ridge regression, under the same simulation setup as Figures 5 and 6, respectively. It is worth noting that for these figures, there is a difference in how we compute finite-sample risks, compared to what is done for all of the other figures in the paper. In all others, we can compute the finite-sample risks exactly, because, recall, our notion of risk is conditional on X,

$$R_X(\hat{\beta}; \beta) = \mathbb{E}\big[\|\hat{\beta} - \beta\|_\Sigma^2 \,|\, X\big],$$

and this quantity can be computed analytically for linear estimators like min-norm least squares and ridge regression. When we tune ridge by minimizing CV or GCV, however, we can no longer compute its risk analytically, and so in our experiments we approximate the above expectation by an average over 20 repetitions. Therefore, the finite-sample risks in Figures 12 and 13 are (only a little bit) farther from their asymptotic counterparts compared to all other figures in this paper, and we have included the finite-sample risk of the optimally-tuned ridge estimator in these figures, when the risk is again computed in the same way (approximated by an average over 20 repetitions), as a reference. All this being said, we still see excellent agreement between the risks under CV, GCV, and optimal tuning, and these all lie close to their (common) asymptotic risk curves, throughout.

B Proofs for the nonlinear model

B.1 Proof of Theorem 8

The proof follows the approach developed by Cheng and Singer (2013) to determine the asymptotic spectral measure of symmetric kernel random matrices. The latter is in turn inspired to the classical resolvent proof of the semicircle law for Wigner matrices, see Anderson et al. (2009). The present calculation is somewhat longer because of the block structure of the matrix A, but does not present technical difficulties. We will therefore provide a proof outline, referring to Cheng and Singer (2013) for further detail.

Let μ_d be the distribution of $\langle w, \boldsymbol{x}\rangle$ when $w \sim N(0, I_d/d)$, $\boldsymbol{x} \sim N(0, I_d)$. By the central limit theorem μ_d converges weakly to μ_G (the standard Gaussian measure) as $d \to \infty$. Further $\mathbb{E}\{f(\langle w, \boldsymbol{x}\rangle)\} \to \int f(z)\, \mu_G(\mathrm{d}z)$ for any continuous function f, with $|f(z)| \le C(1 + |z|^C)$ for some constant C. As shown in Cheng and Singer (2013), we can construct the following orthogonal decomposition of the activation function φ in $L_2(\mathbb{R}, \mu_d)$:

$$\varphi(x) = a_{1,d}\, x + \varphi_\perp(x)\,. \tag{54}$$

This decomposition satisfies the following properties:

1. As mentioned, it is an orthogonal decomposition: $\int x\varphi_\perp(x)\, \mu_d(\mathrm{d}x) = 0$. Further, by symmetry and the normalization assumption, $\int x\, \mu_d(\mathrm{d}x) = \int \varphi_\perp(x)\, \mu_d(\mathrm{d}x) = 0$.
2. $a_{1,d}^2 \to c_1$ as $d \to \infty$.
3. $\int \varphi_\perp(x)^2\, \mu_d(\mathrm{d}x) \to 1 - c_1$, $\int \varphi_\perp(x)^2\, \mu_G(\mathrm{d}x) \to 1 - c_1$, as $d \to \infty$.

Finally, recall the following definitions for the random resolvents:

$$M_{1,n}(\xi, s, t) = \frac{1}{p}\mathrm{tr}_{[1,p]}\big[(A(n) - \xi I_N)^{-1}\big]\,, \qquad M_{2,n}(\xi, s, t) = \frac{1}{n}\mathrm{tr}_{[p+1,p+n]}\big[(A(n) - \xi I_N)^{-1}\big]\,. \tag{55}$$

In the next section, we will summarize some basic facts about the resolvent $m_{1,n}$, $m_{2,n}$ and their analyticity, and some concentration properties of $M_{1,n}, M_{2,n}$. We will then derive Eqs. (20) and (21). Thanks to the analyticity properties, we will assume throughout these derivations $\Im(\xi) \ge C$ for C a large enough constant. Also, we will assume γ, ψ, φ to be fixed throughout, and will not explicitly point out the dependence with respect to these arguments.

B.1.1 Preliminaries

The functions $m_{1,n}$, $m_{2,n}$ are Stieltjes transform of suitably defined probability measures on $\mathbb{R}$, and therefore enjoy some important properties.

Lemma 7. *Let $\mathbb{C}_+ = \{z : \Im(z) > 0\}$ be the upper half of the complex plane. The functions $\xi \mapsto m_{1,n}(\xi)$, $\xi \mapsto m_{2,n}(\xi)$ have the following properties:*

(a) *$\xi \in \mathbb{C}_+$, then $|m_{1,n}|$, $|m_{2,n}| \le 1/\Im(\xi)$.*

(b) *$m_{1,n}$, $m_{2,n}$ are analytic on $\mathbb{C}_+$ and map $\mathbb{C}_+$ into $\mathbb{C}_+$.*

(c) *Let $\Omega \subseteq \mathbb{C}_+$ be a set with an accumulation point. If $m_{a,n}(\xi) \to m_a(\xi)$ for all $\xi \in \Omega$, then $m_a(\xi)$ has a unique analytic continuation to $\mathbb{C}_+$ and $m_{a,n}(\xi) \to m_a(\xi)$ for all $\xi \in \mathbb{C}_+$.*

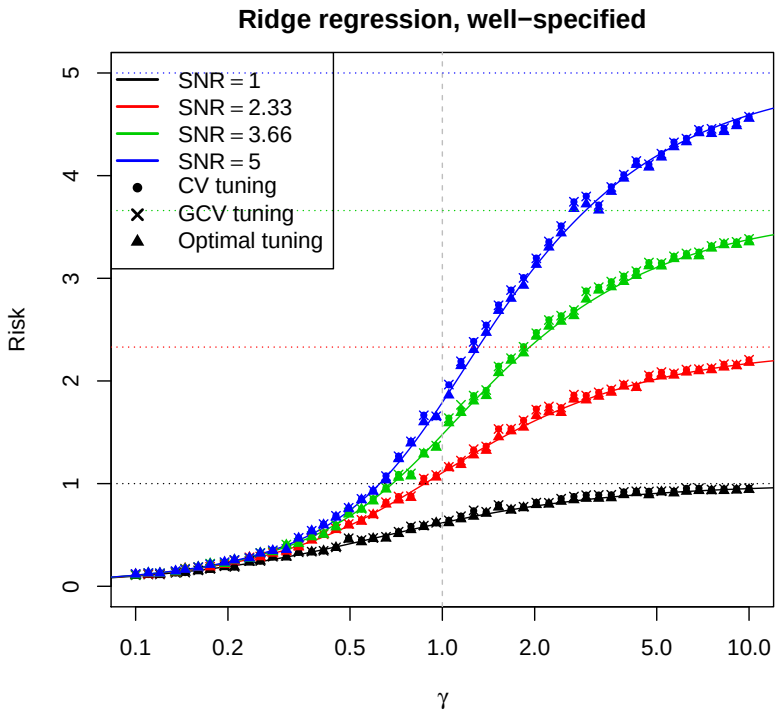

Figure 12: Asymptotic risk curves for the optimally-tuned ridge regression estimator (from Theorem 5), under the same setup as Figure 5 (well-specified model). Finite-sample risks for ridge regression under CV, GCV, and optimal tuning are plotted as circles, "x" marks, and triangles, respectively. These are computed with $n = 200$, $p = [\gamma n]$, across various values of γ, from features X having i.i.d. $N(0, 1)$ entries.

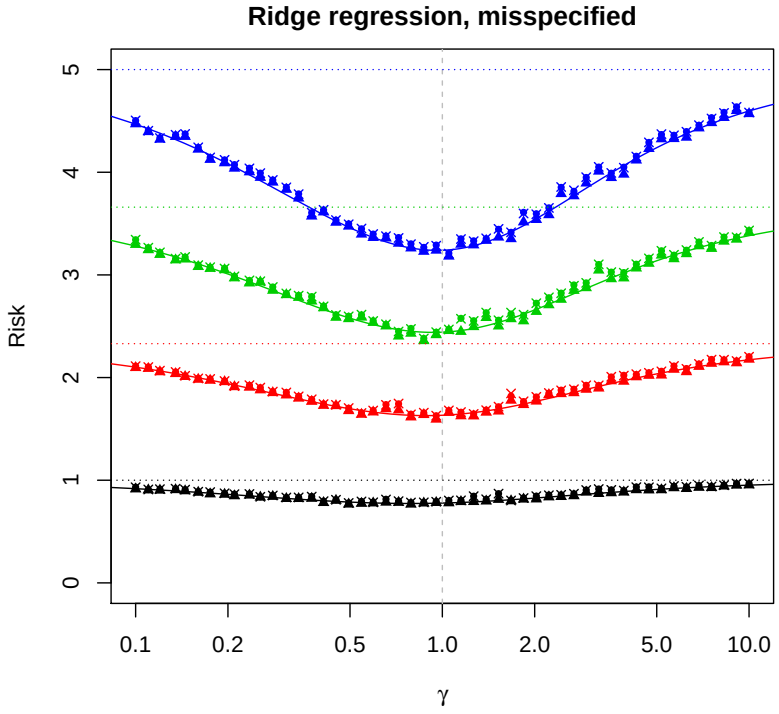

Figure 13: Asymptotic risk curves for the optimally-tuned ridge regression estimator (from Theorem 5), under the same setup as Figure 6 (misspecified model). Finite-sample risks for ridge regression under CV, GCV, and optimal tuning are again plotted as circles, "x" marks, and triangles, respectively. These are again computed with $n = 200$, $p = [\gamma n]$, across various values of γ, from features X having i.i.d. $N(0, 1)$ entries.

Proof. Consider, to be definite $m_{1,n}$. Denoting by $(\lambda_i)_{i\le N}$, $(\boldsymbol{v}_i)_{i\le N}$, the eigenvalues and eigenvectors of $A(n)$, we have

$$\begin{aligned} m_{1,n}(\xi) &= \mathbb{E}\frac{1}{p}\mathrm{tr}_{[1,p]}\big[(A(n)-\xi I_N)^{-1}\big] \\ &= \mathbb{E}\sum_{i=1}^{N}\frac{1}{\lambda_i-\xi}\,\frac{1}{p}\|\mathsf{P}_{[1,p]}\boldsymbol{v}_i\|^2 \\ &= \int\frac{1}{\lambda-\xi}\,\mu_{1,n}(\mathrm{d}\lambda)\,. \end{aligned}$$

Where we defined the probability measure

$$\mu_{1,n} \equiv \mathbb{E}\frac{1}{p}\sum_{i=1}^{N}\|\mathsf{P}_{[1,p]}\boldsymbol{v}_i\|^2\delta_{\lambda_i}\,.$$

Properties (a)-(c) are then standard consequences of $m_{1,n}$ being a Stieltjes transform (see, for instance, Anderson et al. (2009)) □

Lemma 8. *Let $W\in\mathbb{R}^{N\times N}$ be a symmetric positive-semidefinite matrix $W\succeq\mathbf{0}$, and denote by w_i its i-th column, with the i-th entry set to 0. Let $W^{(i)}\equiv W-w_i\boldsymbol{e}_i^T-\boldsymbol{e}_iw_i^T$, where $\boldsymbol{e}_i$ is the i-th element of the canonical basis (in other words, $W^{(i)}$ is obtained from W by zeroing all elements in the i-th row and column except on the diagonal). Finally, let $\xi\in\mathbb{C}$ with $\Im(\xi)\ge\xi_0>0$ or $\Im(\xi)=0$ and $\Re(\xi)\le-\xi_0<0$.*

Then

$$\Big|\mathrm{tr}_S\big[(W-\xi I_N)^{-1}\big]-\mathrm{tr}_S\big[(W^{(i)}-\xi I_N)^{-1}\big]\Big|\le\frac{3}{\xi_0}\,. \tag{56}$$

Proof. Without loss of generality, we will assume $i=N$, and write $W_*=(W_{ij})_{i,j\le N-1}$ $w=(w_{N,i})_{i\le N-1}$, $\omega=W_{NN}$. Define

$$\begin{bmatrix}\boldsymbol{R} & \boldsymbol{r}\\ \boldsymbol{r}^T & \rho\end{bmatrix}\equiv(W-\xi I_N)^{-1}\,, \tag{57}$$

and $\boldsymbol{R}^{(N)}$, $rho^{(N)}$ the same quantities when W is replaced by $W^{(N)}$. Then, by Schur complement formula, it is immediate to get the formulae

$$\boldsymbol{R}=(\tilde{W}_*-\xi I_{N-1})^{-1}\,,\quad \tilde{W}_*\equiv W-\frac{ww^T}{\omega-\xi}\,, \tag{58}$$

$$\boldsymbol{R}^{(N)}=(W_*-\xi I_{N-1})^{-1}\,, \tag{59}$$

$$\rho=\frac{1}{\omega-\xi-w^T\boldsymbol{R}^{(N)}w}\,, \tag{60}$$

$$\rho^{(N)}=\frac{1}{\omega-\xi}\,. \tag{61}$$

Notice that since $W\succeq 0$. we must have $\omega\ge w^TW_*^{-1}w$. For $\xi\in\mathbb{R}$, $\xi\le-\xi_0$, this implies $\omega\ge w^T\boldsymbol{R}^{(N)}w$, and therefore $\rho\le 1/\xi_0$. For $\Im(\xi)\ge\xi_0$, we get $\Im(w^T\boldsymbol{R}^{(N)}w)\ge 0$, and therefore $\Im(\omega-\xi-w^T\boldsymbol{R}^{(N)}w)\le-\xi_0$. We thus conclude that, in either case

$$|\rho|\le\frac{1}{\xi_0}\,. \tag{62}$$

Let $S_0\equiv S\setminus\{N\}$ and $S_1\equiv S\cap\{N\}$ (i.e. $S_1=\{N\}$ or $S_1=\emptyset$ depending whether $N\in S$ or not). Define $\Delta_A\equiv\big|\mathrm{tr}_A\big[(W-\xi I_N)^{-1}\big]-\mathrm{tr}_A\big[(W^{(N)}-\xi I_N)^{-1}\big]\big|$. Then, using the bounds given above,

$$\Delta_{S_1}\le\Big|\rho-\rho^{(N)}\Big|\le|\rho|+|\rho^{(N)}|\le\frac{2}{\xi_0}\,. \tag{63}$$

Next consider Δ_{S_0}. Notice that

$$\boldsymbol{R} - \boldsymbol{R}^{(N)} = \rho\, \boldsymbol{R}^{(N)} w (\boldsymbol{R}^{(N)} w)^T \tag{64}$$

$$= \frac{(W_* - \xi I)^{-1} w [(W_* - \xi I)^{-1} w]^T}{\omega - \xi - w^T (W_* - \xi I_{N-1}) w}\,. \tag{65}$$

We will distinguish two cases.

Case I: $\xi \in \mathbb{R}$**,** $\xi \le -\xi_0 < 0$**.** Notice that

$$\Delta_{S_0} = \left|\mathrm{tr}_{S_0}(\boldsymbol{R} - \boldsymbol{R}^{(N)})\right| = |\rho| \sum_{i \in S_0} (\boldsymbol{R}^{(N)} w)_i^2 \tag{66}$$

$$\le \left|\mathrm{tr}_{[N-1]}(\boldsymbol{R} - \boldsymbol{R}^{(N)})\right|. \tag{67}$$

Denoting by $\{\tilde\lambda_i\}_{i \le N}$ the eigenvalues of $\tilde W_*$, we thus get

$$\Delta_{S_0} \le \left| \sum_{i=1}^N \frac{1}{\tilde\lambda_i - \xi} - \sum_{i=1}^N \frac{1}{\lambda_i - \xi} \right|. \tag{68}$$

Since $\tilde W_*$ is a rank-one deformation of W_*, the eigenvalues $\tilde\lambda_i$ interlace the λ_i's. Since both sets of eigenvalues are nonnegative, the difference above is upper bounded by the total variation of the function $x \mapsto (x-\xi)^{-1}$ on $\mathbb{R}_{\ge 0}$ which is equal to $1/\xi_0$. We thus get

$$\Delta_{S_0} \le \frac{1}{\xi_0}\,. \tag{69}$$

Case II: $\xi \in \mathbb{C}$**,** $\Im(\xi) \ge \xi_0 > 0$**.** Let $\mathsf{P}_{S_0} : \mathbb{R}^{N-1} \times \mathbb{R}^{N-1}$ the projector which zeroes the coordinates outside S_0 (i.e. $\mathsf{P}_{S_0} = \sum_{i \in S_0} \boldsymbol{e}_i \boldsymbol{e}_i^T$). Denote by $(\lambda_i, \boldsymbol{v}_i)$ the eigenpairs of W_*. Using Eq. (65), we get

$$\Delta_{S_0} \equiv \frac{A_1}{A_2}\,, \tag{70}$$

$$A_1 \equiv \left| \sum_{i,j=1}^N \langle \boldsymbol{v}_i, \mathsf{P}_{S_0} \boldsymbol{v}_j \rangle \frac{\langle \boldsymbol{v}_i, w\rangle \langle \boldsymbol{v}_j, w\rangle}{(\lambda_i - \xi)(\lambda_j - \xi)} \right|. \tag{71}$$

$$A_2 \equiv \left| \omega - \xi - \sum_{i=1}^N \frac{\langle \boldsymbol{v}_i, w\rangle^2}{\lambda_i - \xi} \right|. \tag{72}$$

Letting $V_{ij} = \langle \boldsymbol{v}_i, \mathsf{P}_{S_0} \boldsymbol{v}_j \rangle$, we have $\|\boldsymbol{V}\|_{\mathrm{op}} \le \|\mathsf{P}_{S_0}\|_{\mathrm{op}} \le 1$. Therefore, defining $u_i = \langle w, \boldsymbol{v}_i\rangle/(\lambda_i - \xi)$,

$$A_1 = \langle \boldsymbol{u}, \boldsymbol{V}\boldsymbol{u}\rangle \le \|\boldsymbol{u}\|_2^2 \tag{73}$$

$$\le \sum_{i=1}^N \frac{\langle \boldsymbol{v}_i, w\rangle^2}{|\lambda_i - \xi|^2} \tag{74}$$

$$\le \sum_{i=1}^N \frac{\langle \boldsymbol{v}_i, w\rangle^2}{(\lambda_i - \Re(\xi))^2 + \Im(\xi)^2}\,. \tag{75}$$

Further

$$A_2 \ge \left| \Im(\xi) + \sum_{i=1}^N \Im\left(\frac{\langle \boldsymbol{v}_i, w\rangle^2}{\lambda_i - \xi} \right) \right| \tag{76}$$

$$\ge \left| \Im(\xi) + \sum_{i=1}^N \frac{\langle \boldsymbol{v}_i, w\rangle^2 \Im((\xi)}{(\lambda_i - \Re(\xi))^2 + \Im(\xi)^2} \right| \tag{77}$$

$$\ge \xi_0 A_1\,, \tag{78}$$

which implies that Eq. (69) also holds in this case.

Using Eqs. (63) and (69) yields the desired claim. □

Lemma 9. *If $\Im(\xi) \geq \xi_0 > 0$, or $\Re(\xi) \leq -\xi_0 < 0$, with $s \geq t \geq 0$. Then there exists $c_0 = c_0(\xi_0)$ such that, for $i \in \{1, 2\}$,*

$$\mathbb{P}\big(\big|M_{i,n}(\xi, s, t) - m_{i,n}(\xi, s, t)\big| \geq u\big) \leq 2\, e^{-c_0 n u^2}\,. \tag{79}$$

In particular, for $\Im(\xi) > 0$, or $\Re(\xi) < 0$, then $|M_{i,n}(\xi, s, t) - m_{i,n}(\xi, s, t)| \to 0$ almost surely.

Proof. The proof is completely analogous to (Cheng and Singer, 2013, Lemma 2.4), except that we use Azuma-Hoeffding instead of Burkholder's inequality. Consider $M_{1,n}(\xi)$ (we omit the arguments s, t for the sake of simplicity). We construct the martingale

$$Z_\ell = \begin{cases} \mathbb{E}\{M_{1,n}(\xi)|\{w_a\}_{a\leq \ell}\} & \text{if } 1 \leq \ell \leq p, \\ \mathbb{E}\{M_{1,n}(\xi)|\{w_a\}_{a\leq p}, \{z_i\}_{i\leq \ell - p}\} & \text{if } p+1 \leq \ell \leq N, \end{cases} \tag{80}$$

By Lemma 8 Z_ℓ has bounded differences $|Z_\ell - Z_{\ell-1}| \leq 3/\xi_0$, whence the claim (79) follows. The almost sure convergence of $|M_{i,n}(\xi, s, t) - m_{i,n}(\xi, s, t)| \to 0$ is implied by Borel-Cantelli. □

Lemma 10. *Let $F : \mathbb{C}^2 \to \mathbb{C}^2$ be the mapping $(m_1, m_2) \mapsto F(m_1, m_2)$ defined by the right-hand side of Eqs. (21), (20), namely*

$$F_1(m_1, m_2) \equiv \left(-\xi - s - \frac{t^2}{\psi} m_1 - m_2 + \frac{t^2\psi^{-1} m_1^2(c_1 m_2 - t) - 2t c_1 m_1 m_2 + c_1^2 m_1 m_2^2}{m_1(c_1 m_2 - \psi) - \psi}\right)^{-1}, \tag{81}$$

$$F_2(m_1, m_2) \equiv \left(-\xi - \gamma m_1 + \frac{\gamma c_1 m_1^2 (c_1 m_2 - t)}{m_1(c_1 m_2 - t) - \psi}\right)^{-1}. \tag{82}$$

Define $\mathbb{D}(r) = \{z : |z| < r\}$ to be the disk of radius r in the complex plane. Then, there exists $r_0 > 0$ such that, for any $r, \delta > 0$ there exists $\xi_0 = \xi_0(s, t, r, \delta) > 0$ such that, if $\Im(\xi) > \xi_0$, then F maps $\mathbb{D}(r_0) \times \mathbb{D}(r_0)$ into $\mathbb{D}(r) \times \mathbb{D}(r)$ and further is Lipschitz continuous, with Lipschitz constant at most δ on that domain.

In particular Eqs. (21), (20) admit a unique solution with $|m_1|, |m_2| < r_0$ for $\xi > \xi_0$.

Proof. Setting $\boldsymbol{m} \equiv (m_1, m_2)$, we note that $F(\boldsymbol{m}) = (-\xi + G(\boldsymbol{m}))^{-1}$, where $\boldsymbol{m} \mapsto G(\boldsymbol{m})$ is L-Lipschitz continuous in a neighborhood of of $\mathbf{0}$, $\mathbb{D}(r_0) \times \mathbb{D}(r_0)$, with $G(\mathbf{0}) = (s, 0)$. We therefore have, for $|m_i| \leq r_0$,

$$|F_i(\boldsymbol{m})| \leq \frac{1}{|\xi| - |F_i(\boldsymbol{m})|} \leq \frac{1}{\xi_0 - |s| - 2Lr_0}\,. \tag{83}$$

whence $|F_i(\boldsymbol{m})| < r$ by taking ξ_0 large enough. Analogously

$$\|\nabla F_i(\boldsymbol{m})\| \leq \frac{1}{(|\xi| - |F_i(\boldsymbol{m})|)^2}\, \|\nabla G_i(\boldsymbol{m})\| \leq \frac{L}{(\xi_0 - |s| - 2Lr_0)^2}\,. \tag{84}$$

Again, the claim follows by taking ξ_0 large enough. □

The next lemma allow to restrict ourselves to cases in which φ is polynomial with degree independent of n.

Lemma 11. *Let φ_A, φ_B be two activation functions, and denote by $m_{1,n}^A$, $m_{2,n}^A$, $m_{1,n}^B$, $m_{2,n}^B$, denote the corresponding resolvents as defined above. Assume ξ to be such that either $\xi \in \mathbb{R}$, $\xi \leq -\xi_0 < 0$, or $\Im(\xi) \geq \xi_0 > 0$. Then there exists a constant $C(\xi_0)$ such that, for all n large enough*

$$\big|m_{1,n}^A(\xi) - m_{1,n}^B(\xi)\big| \leq C(\xi_0)\mathbb{E}\big\{[\varphi_A(G) - \varphi_B(G)]^2\big\}^{1/2}, \tag{85}$$

$$\big|m_{2,n}^A(\xi) - m_{2,n}^B(\xi)\big| \leq C(\xi_0)\mathbb{E}\big\{[\varphi_A(G) - \varphi_B(G)]^2\big\}^{1/2}. \tag{86}$$

Proof. The proof is essentially the same as for Lemma 4.4 in Cheng and Singer (2013). □

Lemma 12. *Let $m_{1,n,p}(\xi)$ and $m_{2,n,p}(\xi)$ be the resolvent defined above where we made explicit the dependence upon the dimensions n, p. Assume ξ to be such that either $\xi \in \mathbb{R}$, $\xi \le -\xi_0 < 0$, or $\Im(\xi) \ge \xi_0 > 0$. Then, there exist $C = C(\xi_0) < \infty$ such that*

$$\big|m_{1,n,p}(\xi) - m_{1,n-1,p}(\xi)\big| + \big|m_{1,n,p}(\xi) - m_{1,n,p-1}(\xi)\big| \le \frac{C}{n}\,, \tag{87}$$

$$\big|m_{2,n,p}(\xi) - m_{2,n-1,p}(\xi)\big| + \big|m_{2,n,p}(\xi) - m_{2,n,p-1}(\xi)\big| \le \frac{C}{n}\,. \tag{88}$$

(Here d is understood to be the same in each case.)

Proof. This follows immediately from Lemma 8. Denote, with a slight abuse of notation, by $A(n,p)$ the matrix in Eq. (19). Consider for instance the difference

$$\big|m_{1,n,p}(\xi) - m_{1,n,p-1}(\xi)\big| = \frac{1}{p}\left|\mathbb{E}\mathrm{tr}_{[1,p]}\big[(A(n,p)-\xi I)^{-1}\big] - \mathbb{E}\mathrm{tr}_{[1,p]}\big[(A(n,p-1)-\xi I)^{-1}\big]\right| \tag{89}$$

$$= \frac{1}{p}\left|\mathbb{E}\left\{\mathrm{tr}_{[1,p]}\big[(A(n,p)-\xi I)^{-1}\big] - \mathrm{tr}_{[1,p]}\big[(A^*(n,p)-\xi I)^{-1}\big]\right\}\right|\,, \tag{90}$$

where $A^*(n,p)$ is obtained from $A(n,p)$ by zero-ing the last row and column. Lemma 8 then implied $|m_{1,n,p}(\xi) - m_{1,n,p-1}(\xi)| \le 3/\xi_0$. The other terms are treated analogously. □

B.1.2 Derivation of Eqs. (20)

Throughout this appendix, we will assume $\Im(\xi) \ge C$ a large enough constant. This is sufficient by Lemma 7.(c). Also, we can restrict ourselves to φ a fixed finite polynomial (independent of n). This is again sufficient by Lemma 11 (polynomials are dense in $L^2(\mathbb{R}, \mu_G)$).

We denote by $A_{\cdot,m} \in \mathbb{R}^{N-1}$ the m-th column of A, with the m-th entry removed. By the Schur complement formula, we have

$$m_{2,n} = \mathbb{E}\left\{\big(-\xi - A^T_{\cdot,n+p}(A_* - \xi I_{N-1})A_{\cdot,n+p}\big)^{-1}\right\}\,, \tag{91}$$

where $A_* \in \mathbb{R}^{(N-1)\times(N-1)}$ is the submatrix comprising the first $N-1$ rows and columns of A.

We let η_a denote the projection of w_a along z_n, and by $\tilde{\boldsymbol{w}}_a$ the orthogonal component. Namely we write

$$w_a = \eta_a \frac{z_n}{\|z_n\|} + \tilde{\boldsymbol{w}}_a\,, \tag{92}$$

where $\langle \tilde{\boldsymbol{w}}_a, z_n\rangle = 0$. We further let $X_* \in \mathbb{R}^{(n-1)\times p}$ denote the submatrix of X obtained by removing the last row. With these notations, we have

$$A_* = \begin{bmatrix} sI_p + tQ & \frac{1}{\sqrt{n}}X_*^T \\ \frac{1}{\sqrt{n}}X_* & \mathbf{0} \end{bmatrix}\,, \tag{93}$$

$$Q_{ab} = \eta_a\eta_b \mathbf{1}_{a\neq b} + \langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{w}}_b\rangle \mathbf{1}_{a\neq b}\,, \qquad 1 \le a,b \le p, \tag{94}$$

$$X_{ja} = \varphi\left(\langle \tilde{\boldsymbol{w}}_a, z_j\rangle + \eta_a \frac{\langle z_j, z_n\rangle}{\|z_n\|}\right)\,, \qquad 1 \le a \le p,\ 1 \le j \le n-1\,, \tag{95}$$

and

$$A_{\cdot,n+p} = \begin{bmatrix} \frac{1}{\sqrt{n}}\varphi(\|z_n\|\eta_1) \\ \vdots \\ \frac{1}{\sqrt{n}}\varphi(\|z_n\|\eta_p) \\ 0 \\ \vdots \\ 0 \end{bmatrix}\,, \tag{96}$$

We next decompose

$$X_{ja} = \widetilde{X}_{ja} + a_{1,d} u_j \eta_a + \sqrt{n}\, E_{1,ja}\,, \tag{97}$$

$$\widetilde{X}_{ja} \equiv \varphi\left(\langle \tilde{\boldsymbol{w}}_a, z_j\rangle\right)\,, \tag{98}$$

$$u_j \equiv \frac{1}{\sqrt{n}} \frac{\langle z_j, z_n\rangle}{\|z_n\|}\,, \tag{99}$$

$$E_{1,ja} \equiv \frac{1}{\sqrt{n}} \varphi_\perp\left(\langle \tilde{\boldsymbol{w}}_a, z_j\rangle + \eta_a \frac{\langle z_j, z_n\rangle}{\|z_n\|}\right) - \frac{1}{\sqrt{n}} \varphi_\perp\left(\langle \tilde{\boldsymbol{w}}_a, z_j\rangle\right)\,. \tag{100}$$

and

$$Q_{ab} = \widetilde{Q}_{ab} + \eta_a \eta_b + E_{0,ab}\,, \tag{101}$$

$$\widetilde{Q}_{ab} \equiv \langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{w}}_b\rangle\,, \tag{102}$$

$$E_{0,ab} \equiv -\eta_a^2 \mathbf{1}_{a=b}\,. \tag{103}$$

We therefore get

$$A_* = \tilde{A}_* + \boldsymbol{\Delta} + \boldsymbol{E}\,, \tag{104}$$

$$\tilde{A}_* = \begin{bmatrix} s I_p + t\widetilde{\boldsymbol{Q}} & \frac{1}{\sqrt{n}} \tilde{X}_*^T \\ \frac{1}{\sqrt{n}} \tilde{X}_* & \mathbf{0} \end{bmatrix}\,, \qquad \boldsymbol{\Delta} = \begin{bmatrix} t\eta\eta^T & a_{1,d}\eta \boldsymbol{u}^T \\ a_{1,d}\boldsymbol{u}\eta^T & \mathbf{0} \end{bmatrix}\,, \qquad \boldsymbol{E} = \begin{bmatrix} \boldsymbol{E}_0 & \boldsymbol{E}_1^T \\ \boldsymbol{E}_1 & \mathbf{0} \end{bmatrix}\,, \tag{105}$$

It is possible to show that $\|\boldsymbol{E}\|_{\rm op} \le \varepsilon_n \equiv (\log n)^M / n^{1/2}$ with probability at least $1 - O(n^{-1})$, where M is an absolute constant (this can be done as in Section 4.3, Step 2 of Cheng and Singer (2013), using intermediate value theorem). Further $\boldsymbol{\Delta}$ is a rank-2 matrix that can be written as $\boldsymbol{\Delta} = U \boldsymbol{C} U^T$, where $U \in \mathbb{R}^{(N-1)\times 2}$ and $\boldsymbol{C} \in \mathbb{R}^{2\times 2}$ are given by

$$U = \begin{bmatrix} \eta & \mathbf{0} \\ \mathbf{0} & \boldsymbol{u} \end{bmatrix}\,, \qquad \boldsymbol{C} = \begin{bmatrix} t & a_{1,d} \\ a_{1,d} & 0 \end{bmatrix}\,. \tag{106}$$

Using Eq. (91), and Woodbury's formula, we get (writing for simplicity $\boldsymbol{v} = A_{\cdot,n+p}$

$$\begin{aligned} m_{2,n} &= \mathbb{E}\left\{\left(-\xi - \boldsymbol{v}^T(\tilde{A}_* + \boldsymbol{\Delta} + \boldsymbol{E} - \xi I_{N-1})^{-1}\boldsymbol{v}\right)^{-1}\right\} \\ &= \mathbb{E}\left\{\left(-\xi - \boldsymbol{v}^T(\tilde{A}_* + \boldsymbol{\Delta} + \boldsymbol{E} - \xi I_{N-1})^{-1}\boldsymbol{v}\right)^{-1} \mathbf{1}_{\|bE\|_{\rm op}\le \varepsilon_n}\right\} \\ &\quad + \frac{C}{\Im(\xi)} O\left(\mathbb{P}(\|bE\|_{\rm op} > \varepsilon_n)\right) \\ &= \mathbb{E}\left\{\left(-\xi - \boldsymbol{v}^T(\tilde{A}_* + \boldsymbol{\Delta} - \xi I_{N-1})^{-1}\boldsymbol{v}\right)^{-1} \mathbf{1}_{\|bE\|_{\rm op}\le \varepsilon_n}\right\} + O(\varepsilon_n) \\ &= \mathbb{E}\left\{\left(-\xi - \boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1}\boldsymbol{v} + \boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1} U \boldsymbol{S}^{-1} U^T (\tilde{A}_* - \xi I_{N-1})^{-1}\boldsymbol{v}\right)^{-1}\right\} + O(\varepsilon_n) \end{aligned} \tag{107}$$

$$\boldsymbol{S} \equiv \boldsymbol{C}^{-1} + U^T(\tilde{A}_* - \xi I_{N-1})^{-1} U\,. \tag{108}$$

Note that, conditional on $\|z_n\|_n$, $\boldsymbol{v}$ is independent of $\tilde{A}_*$ and has independent entries. Further its entries are independent with $n\mathbb{E}\{v_i^2 | \|z_n\|_2\} = n\mathbb{E}\{\varphi(\|z_n|_2 G/\sqrt{d})^2 | \|z_n\|_2\} = 1 + O(n^{-1/2})$. Hence

$$\mathbb{E}\{\boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1}\boldsymbol{v}\} = \frac{p}{n}\mathbb{E}\left\{\frac{1}{p}\mathrm{tr}[(\tilde{A}_* - \xi I_{N-1})^{-1}]\right\} \tag{109}$$

$$= \gamma\, m_{1,n} + O(\varepsilon_n)\,, \tag{110}$$

By a concentration of measure argument (see, e.g., (Tao, 2012, Section 2.4.3)), the same statement also holds with high probability

$$\boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1}\boldsymbol{v} = \gamma\, m_{1,n} + O_P(\varepsilon_n) \tag{111}$$

We next consider $\boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1}U$. We decompose $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2$, where

$$v_{1,i} = \frac{a_{1,d}\|z_n\|}{\sqrt{n}}\,\eta_i\,, \quad v_{2,i} = \frac{1}{\sqrt{n}}\,\varphi_\perp(\|z_n\|\eta_i)\,, \quad 1 \le i \le p\,, \tag{112}$$

Notice, that conditional on $\|z_n\|$, the pairs $\{(v_{1,i}, v_{2,i})\}_{i\le p}$ are mutually independent. Further $n\mathbb{E}\{v_{1,i}v_{2,i}|\|z_n\|\} = \mathbb{E}\{a_{1,d}G\varphi_\perp(G)\} + O(\varepsilon_n) = O(\varepsilon_n)$. Finally, again conditionally on z_n, η and $\boldsymbol{u}$ are independent. Therefore

$$\mathbb{E}\big\{\boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1}U\big\} = \mathbb{E}\left\{\Big[\boldsymbol{v}_1^T(\tilde{A}_* - \xi I_{N-1})^{-1}_{[1,p],[1,p]}\eta\Big|0\Big]\right\} + O(\varepsilon_n) \tag{113}$$

$$= \frac{a_{1,d}\sqrt{d}}{\sqrt{n}}\left[\frac{1}{d}\mathbb{E}\mathrm{tr}_{[1,p]}\big[(\tilde{A}_* - \xi I_{N-1})^{-1}\big]\Big|0\right] + O(\varepsilon_n) \tag{114}$$

$$= \sqrt{c_1\gamma\psi^{-1}}\big[m_{1,n}\,,\,0\big] + O(\varepsilon_n)\,. \tag{115}$$

By a concentration of measure argument, we also have

$$\boldsymbol{v}^T(\tilde{A}_* - \xi I_{N-1})^{-1}U = \sqrt{c_1\gamma\psi^{-1}}\big[m_{1,n}\,,\,0\big] + O(\varepsilon_n)\,. \tag{116}$$

We next consider $U^T(\tilde{A}_* - \xi I_{N-1})^{-1}U$. Since η and $\boldsymbol{u}$ are independent (and independent of $\tilde{A}_*$) and zero mean, with $d\mathbb{E}\{\eta_a^2\} = 1$, $n\mathbb{E}\{u_j^2\} = 1 + O(\varepsilon_n)$, we have

$$\begin{aligned}
&\mathbb{E}\big\{U^T(\tilde{A}_* - \xi I_{N-1})^{-1}U\big\} = \\
&\qquad = \begin{bmatrix} \frac{1}{d}\mathbb{E}\mathrm{tr}_{[1,p]}\{(\tilde{A}_* - \xi I_{N-1})^{-1}\} & 0 \\ 0 & \frac{1}{n}\mathbb{E}\mathrm{tr}_{[p+1,N-1]}\{(\tilde{A}_* - \xi I_{N-1})^{-1}\} \end{bmatrix} + O(\varepsilon_n) \\
&\qquad = \begin{bmatrix} \psi^{-1}m_1 & 0 \\ 0 & m_2 \end{bmatrix} + O(\varepsilon_n)\,.
\end{aligned}$$

As in the previous case, this estimate also holds for $U^T(\tilde{A}_* - \xi I_{N-1})^{-1}U$ (not just its expectation) with high probability. Substituting this in Eq. (107), we get

$$\boldsymbol{S} = \begin{bmatrix} \psi^{-1}m_1 & 1/a_{1,d} \\ 1/a_{1,d} & m_2 - (t/a_{1,d}^2) \end{bmatrix} + O(\varepsilon_n) \tag{117}$$

By using this together with Eqs. (111) and (116), we get

$$m_{2,n} = \left(-\xi - \gamma m_{1,n} + \frac{\gamma c_1 m_{1,n}^2(c_1 m_{2,n} - t)}{m_1(c_1 m_{2,n} - t) - \psi}\right)^{-1} + O(\varepsilon_n)\,. \tag{118}$$

B.1.3 Derivation of Eqs. (21)

The derivation si analogous to the one in the previous section (notice that we will redefine some of the notations used in the last section), and hence we will only outline the main steps.

Let $A_{;m} \in \mathbb{R}^{N-1}$ be the m-th column of A, with the m-th entry removed, and $B_* \in \mathbb{R}^{(N-1)\times(N-1)}$ be the submatrix obtained by removing the p-th column and p-th row from A. By the Schur complement formula, we have

$$m_{1,n} = \mathbb{E}\left\{\big(-\xi - s - A_{\cdot,p}^T(B_* - \xi I_{N-1})A_{\cdot,p}\big)^{-1}\right\}\,, \tag{119}$$

where $A_* \in \mathbb{R}^{(N-1)\times(N-1)}$ is the submatrix comprising the first $N-1$ rows and columns of A.

We decompose w_a, $1 \le a \le p-1$, and z_i, $1 \le i \le n$ into their components along w_p, and the orthogonal complement:

$$w_a = \eta_a\frac{w_p}{\|w_p\|} + \tilde{\boldsymbol{w}}_a\,, \qquad z_i = \sqrt{n}\,u_i\frac{w_p}{\|w_p\|} + \tilde{\boldsymbol{z}}_i\,, \tag{120}$$

where $\langle \tilde{\boldsymbol{w}}_a, w_p\rangle = \langle \tilde{\boldsymbol{z}}_i, w_p\rangle = 0$ (the factor $\sqrt{n}$ in the second expression is introduced for future convenience). With these notations, we have

$$B_* = \begin{bmatrix} sI_p + tQ_* & \frac{1}{\sqrt{n}}X^T \\ \frac{1}{\sqrt{n}}X & \mathbf{0} \end{bmatrix}, \tag{121}$$

$$Q_{ab} = \eta_a\eta_b \mathbf{1}_{a\neq b} + \langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{w}}_b\rangle \mathbf{1}_{a\neq b}\,, \qquad 1\le a,b\le p-1, \tag{122}$$

$$X_{ja} = \frac{1}{\sqrt{n}}\varphi\left(\langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{z}}_j\rangle + \sqrt{n}\, u_j\eta_a\right), \qquad 1\le a\le p,\ 1\le j\le n\,, \tag{123}$$

and

$$A_{\cdot,p} = \begin{bmatrix} t\eta_1\|w_p\| \\ \vdots \\ t\eta_{p-1}\|w_p\| \\ \frac{1}{\sqrt{n}}\varphi(\sqrt{n}u_1\|w_p\|) \\ \vdots \\ \frac{1}{\sqrt{n}}\varphi(\sqrt{n}u_n\|w_p\|) \end{bmatrix} \equiv h\,. \tag{124}$$

We next decompose B_* as follows

$$B_* = \tilde{\boldsymbol{B}}_* + \boldsymbol{\Delta} + \boldsymbol{E}\,, \tag{125}$$

$$\tilde{\boldsymbol{B}}_* \equiv \begin{bmatrix} sI_p + t\widetilde{\boldsymbol{Q}}_* & \frac{1}{\sqrt{n}}\tilde{X}^T \\ \frac{1}{\sqrt{n}}\tilde{X} & \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{\Delta} \equiv \begin{bmatrix} t\eta\eta^T & a_{1,d}\eta\boldsymbol{u}^T \\ a_{1,d}\boldsymbol{u}\eta^T & \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{E} \equiv \begin{bmatrix} \boldsymbol{E}_0 & \boldsymbol{E}_1^T \\ \boldsymbol{E}_1 & \mathbf{0} \end{bmatrix}, \tag{126}$$

where we defined matrices $\widetilde{\boldsymbol{Q}}_*$, $\tilde{X}$, $\boldsymbol{E}_0$, $\boldsymbol{E}_1$ with the following entries:

$$\widetilde{Q}_{*,ab} = \langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{w}}_b\rangle \mathbf{1}_{a\neq b}\,, \qquad 1\le a,b\le p-1\,, \tag{127}$$

$$\widetilde{X}_{ja} = \varphi(\langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{z}}_j\rangle)\,, \qquad 1\le a\le p-1,\ 1\le j\le n\,, \tag{128}$$

$$E_{0,ab} = -t\eta_a^2\mathbf{1}_{a=b}\,, \qquad 1\le a,b\le p-1\,, \tag{129}$$

$$E_{1,ja} = \frac{1}{\sqrt{n}}\varphi_\perp\big(\langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{z}}_j\rangle + \sqrt{n}\,\eta_a u_j\big) - \frac{1}{\sqrt{n}}\varphi_\perp\big(\langle \tilde{\boldsymbol{w}}_a, \tilde{\boldsymbol{z}}_j\rangle\big)\,, \quad 1\le a\le p-1,\ 1\le j\le n\,. \tag{130}$$

As in the previous section, it can be shown that $\|\boldsymbol{E}\|_{\text{op}} \le \varepsilon_n \equiv (\log n)^c/\sqrt{n}$ (with c an absolute constant), and therefore Eq. (119) yields

$$m_{1,n} = \mathbb{E}\left\{\big(-\xi - s - A_{\cdot,p}^T(\tilde{\boldsymbol{B}}_* + \boldsymbol{\Delta} - \xi I_{N-1})A_{\cdot,p}\big)^{-1}\right\} + O(\varepsilon_n)\,. \tag{131}$$

Note that $\boldsymbol{\Delta} = U\boldsymbol{C}U^T$, where $U\in\mathbb{R}^{(N-1)\times 2}$ and $\boldsymbol{C}\in\mathbb{R}^{2\times 2}$ take the same form as in Eq. (106). Hence, by Woodbury's formula and recalling the notation $h = A_{\cdot,p}$ w

$$\begin{aligned} m_{1,n} =& \mathbb{E}\left\{\left(-\xi - h^T(\tilde{\boldsymbol{B}}_* - \xi I_{N-1})^{-1}h + h^T(\tilde{\boldsymbol{B}}_* - \xi I_{N-1})^{-1}U\boldsymbol{S}^{-1}U^T(\tilde{\boldsymbol{B}}_* - \xi I_{N-1})^{-1}h\right)^{-1}\right\} \\ &+ O(\varepsilon_n) \\ \boldsymbol{S} \equiv& \boldsymbol{C}^{-1} + U^T(\tilde{\boldsymbol{B}}_* - \xi I_{N-1})^{-1}U\,,. \end{aligned} \tag{132}$$

We then proceed to compute the various terms on the right-hand side. The calculation is very similar to the one in the previous section, and we limit ourselves to reporting the results:

$$h^T(\widetilde{B}_* - \xi I_{N-1})^{-1}h = t^2\psi^{-1}m_{1,n} + m_{2,n} + O_P(\varepsilon_n)\,, \tag{133}$$

$$h^T(\widetilde{B}_* - \xi I_{N-1})^{-1}U = \big[t\psi^{-1}\,m_{1,n},\, a_{1,d}m_{2,n}\big] + O_P(\varepsilon_n)\,, \tag{134}$$

$$U^T(\widetilde{B}_* - \xi I_{N-1})^{-1}U = \begin{bmatrix} \psi^{-1}\,m_{1,n} & 0 \\ 0 & m_{2,n} \end{bmatrix} + O_P(\varepsilon_n)\,, \tag{135}$$

whence

$$\boldsymbol{S} = \begin{bmatrix} \psi^{-1}\, m_{1,n} & a_{1,d}^{-1} \\ a_{1,d}^{-1} & -t a_{1,d}^{-2} + m_{2,n} \end{bmatrix} + O_P(\varepsilon_n)\,. \tag{136}$$

Substituting Eqs. (133) to (136) into Eq. (132), we get

$$m_{1,n} = \left(-\xi - s - \frac{t^2}{\psi} m_{1,n} - m_{2,n} + \frac{t^2\psi^{-1} m_{1,n}^2 (c_1 m_{2,n} - t) - 2t c_1 m_{1,n} m_{2,n} + c_1^2 m_{1,n} m_{2,n}^2}{m_{1,n}(c_1 m_{2,n} - \psi) - \psi} \right)^{-1} + O(\varepsilon_n)\,. \tag{137}$$

B.1.4 Completing the proof

Let φ_L be a degree $L = L(\varepsilon)$ polynomial such that $\mathbb{E}\{|\varphi(G) - \varphi_L(G)|^2\} \le \varepsilon^2$, for $G \sim N(0,1)$. We will denote by $\boldsymbol{m}_n^L = (m_{1,n}^L, m_{2,n}^L)$ the corresponding expected resolvents.

Consider $\Im(\xi) \ge C_0$ a large enough constant. By Eqs. (118), (118), we have

$$\boldsymbol{m}_n^L = F(\boldsymbol{m}_n^L) + o_n(1)\,. \tag{138}$$

By Lemma 7 and Lemma 10, taking C_0 sufficiently large, we can ensure that $\boldsymbol{m}_n^L \in \mathbb{D}(r_0) \times \mathbb{D}(r_0)$, and that F maps $\mathbb{D}(r_0) \times \mathbb{D}(r_0)$ into $\mathbb{D}(r_0/2) \times \mathbb{D}(r_0/2)$, with Lipschitz constant at most $1/2$. Letting $\boldsymbol{m}^L$ denote the unique solution of $\boldsymbol{m}^L = F(\boldsymbol{m}^L)$, we have

$$\|\boldsymbol{m}_n^L - \boldsymbol{m}^L\| = \|F(\boldsymbol{m}_n^L) + o_n(1) - F(\boldsymbol{m}^L)\| \le \frac{1}{2}\|\boldsymbol{m}_n^L - \boldsymbol{m}^L\| + o_n(1)\,, \tag{139}$$

whence $\boldsymbol{m}_n^L(\xi) \to \boldsymbol{m}^L(\xi)$ for all $\Im(\xi) > C_0$. Since *eps* is arbitrary, and using Lemma 11, we get $m_{1,n}(\xi) \to m_1(\xi)$, $m_{2,n}(\xi) \to m_2(\xi)$ for all $\Im(\xi) > C_0$. By Lemma 7.(c), we have $m_{1,n}(\xi) \to m_1(\xi)$, $m_{2,n}(\xi) \to m_2(\xi)$ for all $\xi \in \mathbb{C}_+$. Finally, by Lemma 9, the almost sure convergence of Eqs. (22), (23) holds as well.

B.2 Proof of Lemma 6

We begin approximating the population covariance $\Sigma = \frac{1}{n}\mathbb{E}\{X^T X|w\}$ by $\Sigma_0 \equiv I_p + c_1 Q \in \mathbb{R}^{p\times p}$.

Lemma 13. *With the above definitions there exists a constant C such that*

$$\mathbb{P}\big\{\|\Sigma - \Sigma_0\|_F \ge (\log n)^C\big\} \ge 1 - e^{-(\log n)^2/C},, \tag{140}$$

$$\mathbb{E}\{\|\Sigma - \Sigma_0\|_F^2\} \le (\log n)^C\,. \tag{141}$$

Proof. First notice that

$$\Sigma_{ij} = \mathbb{E}\{\varphi(\langle w_i, z\rangle)\varphi(\langle w_j, z\rangle)|w_i, w_j\} = \mathbb{E}\{\varphi_{\|w_i\|}(G_1)\varphi_{\|w_j\|}(G_2)\}\,, \tag{142}$$

$$\begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & s_{ij} \\ s_{ij} & 1 \end{bmatrix}\right)\,, \tag{143}$$

$$\varphi_t(x) \equiv \varphi(tx)\,, \tag{144}$$

where $s_{ij} = \langle w_i, w_j\rangle/\|w_i\|\|w_j\|$. Let $\varphi_t(x) = \alpha_1(t)x + \varphi_{t,\perp}(x)$ be the orthogonal decomposition of φ_t in $L_2(\mathbb{R}, \mu_G)$, and notice that $\alpha_1(t) = \alpha_1(1)t$, $\alpha_1(1)^2 = c_1$. On the event $\mathcal{G} = \{|\|w_i\| - 1| \le \varepsilon_n\}$ (with $\varepsilon_n = (\log n)^c/\sqrt{n}$), we obtain

$$\Sigma_{ij} = a_{1,d}(\|w_i\|)a_{1,d}(\|w_j\|)s_{ij} + \mathbb{E}\{\varphi_{\|w_i\|,\perp}(G_1)\varphi_{\|w_j\|,\perp}(G_2)\} \tag{145}$$

$$= c_1\langle w_i, w_j\rangle + O(\langle w_i, w_j\rangle^2)\,. \tag{146}$$

Therefore, on the same event (for a suitable constant C)

$$\left\|\Sigma-\Sigma_0\right\|_F^2=\sum_{i=1}^p(\Sigma_{ii}-\Sigma_{0,ii})^2+2\sum_{i<j}^p(\Sigma_{ii}-\Sigma_{0,ii})^2 \tag{147}$$

$$\leq p\,O(\varepsilon_n^2)+C\sum_{i<j}\langle w_i,w_j\rangle^4\,. \tag{148}$$

This implies Eq. (140) because with the stated probability we have $\|w_i\|\leq 1+\varepsilon_n$ for all i and $|\langle w_i,w_j\rangle|\leq\varepsilon_n$ for all $i\neq j$.

Equation (140) follows by the above, together with the remark that $\mathbb{E}(\|\Sigma-\Sigma_0\|_F^{2+c_0})\leq p^{C_0}$ by the assumption that $\mathbb{E}(x_{ij}^{4+c_0})<\infty$. □

We have

$$\left|V_X(\hat\beta_\lambda;\beta)-\frac{1}{n}\mathrm{tr}\left\{(\hat\Sigma+\lambda I)^{-2}\hat\Sigma\Sigma_0\right\}\right|=\left|\frac{1}{n}\mathrm{tr}\left\{(\hat\Sigma+\lambda I)^{-2}\hat\Sigma(\Sigma-\Sigma_0\right\}\right| \tag{149}$$

$$\leq\frac{1}{n}\sqrt{p}\left\|(\hat\Sigma+\lambda I)^{-2}\hat\Sigma\right\|_{\mathrm{op}}\|\Sigma-\Sigma_0\|_F \tag{150}$$

$$\leq\sqrt{\frac{\gamma}{n}}\frac{1}{\lambda}\,\|\Sigma-\Sigma_0\|_F\,. \tag{151}$$

And therefore, by Lemma 13, we obtain

$$\lim_{n\to\infty}\left|V_X(\hat\beta_\lambda;\beta)-\frac{1}{n}\mathrm{tr}\left\{(\hat\Sigma+\lambda I)^{-2}\hat\Sigma\Sigma_0\right\}\right|=0\,, \tag{152}$$

where the convergence takes place almost surely and in L^1.

Denote by $(\lambda_i)_{i\leq p}$, $(\boldsymbol{v}_i)_{i\leq p}$ the eigenvalues and eigenvectors of $\hat\Sigma$. The following qualitative behavior can be extracted from the asymptotic of the Stieltjes transform as stated in Corollary 5.

Lemma 14. *For any $\gamma\neq 1$, $c_1\in[0,1)$, there exists $\rho_0>0$ such that the following happens. Let $S_+\equiv\{i\in[p]:\ |\lambda_i|>\rho_0\}$, $S_-\equiv\{i\in[p]:\ |\lambda_i|\leq\rho_0\}$. Then, the following limits hold almost surely*

$$\lim_{n\to\infty}\frac{1}{n}|S_+|=(\gamma\vee 1)\,, \tag{153}$$

$$\lim_{n\to\infty}\frac{1}{n}|S_-|=(\gamma-1)_+\,, \tag{154}$$

$$\gamma>1\quad\Rightarrow\quad\frac{1}{|S_-|}\sum_{i\in S_-}\delta_{\lambda_i}\Rightarrow\delta_0\,. \tag{155}$$

(Here $\Rightarrow$ denotes weak convergence of probability measures.)

Note that

$$m_n(\xi,s,t)\equiv\gamma\,m_{1,n}(\xi,s,t)+m_{2,n}(\xi,s,t)=\frac{1}{n}\mathbb{E}\mathrm{tr}\big[(A(n)-\xi I_N)^{-1}\big]\,. \tag{156}$$

Denote by $\tilde{\boldsymbol{\Sigma}}_0$ the $N\times N$ matrix whose principal minor corresponding to the first p rows and columns is given by Σ_0. By simple linear algebra (differentiation inside the integral is allowed for $\Im(\xi)>0$ by dominated convergence and by analyticity elsewhere), we get

$$-\partial_x m_n(\xi,x,c_1x)\big|_{x=0}=\frac{1}{n}\mathbb{E}\mathrm{tr}\big[(A-\xi I)^{-1}\tilde{\boldsymbol{\Sigma}}_0(A-\xi I)^{-1}\big]\Big|_{x=0} \tag{157}$$

$$=\frac{1}{n}\mathbb{E}\mathrm{tr}\left[\left(\xi^2I_p+\hat\Sigma-4\xi^2(\hat\Sigma+\xi^2I_p)^{-1}\hat\Sigma\right)^{-1}\Sigma_0\right] \tag{158}$$

$$=\frac{1}{n}\mathbb{E}\mathrm{tr}\left[\left(\hat\Sigma+\xi^2I_p\right)\left(\hat\Sigma-\xi^2I_p\right)^{-2}\Sigma_0\right] \tag{159}$$

Note that $m_n(\xi, x, c_1 x) \to m(\xi, x, c_1 x)$ as $n \to \infty$. Further, for $\Im(\xi) > 0$ or $\Re(\xi) < 0$, it it is immediate to show tha $\partial_x^2 m_n(\xi, x, c_1 x)$ is bounded in n (in a neighborhood of $x = 0$). Hence

$$\lim_{n\to\infty} \partial_x m_n(\xi, x, c_1 x)\Big|_{x=0} = \partial_x m(\xi, x, c_1 x)|_{x=0} \equiv q(\xi)\,, \tag{160}$$

and therefore

$$q(\xi) = \lim_{n\to\infty} \mathbb{E} Q_n(\xi) \tag{161}$$

$$Q_n(\xi) = \frac{1}{n}\mathrm{tr}\left[\left(\hat{\Sigma} + \xi^2 I_p\right)\left(\hat{\Sigma} - \xi^2 I_p\right)^{-2}\Sigma_0\right] \tag{162}$$

$$= \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i + \xi^2}{(\lambda_i - \xi^2)^2}\,\langle \boldsymbol{v}_i, \Sigma_0 \boldsymbol{v}_i\rangle\,. \tag{163}$$

Since the convergence of $M_n = \gamma M_{n,1} + M_{n,2}$ (cf. Eq. (55)) is almost sure, we also have $Q_n(\xi) \to q(\xi)$ almost surely. Define the probability measure on $\mathbb{R}_{ge0}$

$$\mu_n = \frac{1}{p}\sum_{i=1}^{n} \delta_{\lambda_i}\langle \boldsymbol{v}_i, \Sigma \boldsymbol{v}_i\rangle\,. \tag{164}$$

Since $Q_n(\xi) \to q(\xi)$, a weak convergence argument implies $\mu_n \Rightarrow \mu_\infty$ almost surely. Further, defining $\mu_n^+ = \mathbf{1}_{(\rho_0,\infty)}\mu_n$, $\mu_n^- = \mathbf{1}_{[0,\rho_0]}\mu_n$, Lemma 14 implies $\mu_n^+ \Rightarrow \mu_\infty^+$, $\mu_n^- \Rightarrow c_0\delta_0$, where μ_∞^+ is a measure supported on $[\rho_0, \infty)$, with $\mu_\infty^+([\rho_0, \infty)) = 1 - c$. This in turns implies

$$q(\xi) = \frac{\gamma c}{\xi^2} + q_+(\xi)\,, \qquad q_+(\xi) = \gamma\int_{[\rho_0,\infty)} \frac{x + \xi^2}{(x - \xi^2)^2}\mu_\infty^+(\mathrm{d}x)\,, \tag{165}$$

In particular, q_+ is analytic in a neighborhood of 0. This proves Eq. (28), with $q_+(0) = D_0$, $\gamma c = D_{-1}$.

Further, we have

$$D_0 = q_+(0) = \gamma\int_{[\rho_0,\infty)} \frac{x + \xi^2}{(x - \xi^2)^2}\mu_\infty^+(\mathrm{d}x)\,. \tag{166}$$

On the other hand by Eq. (152)

$$\lim_{n\to\infty} V_X(\hat{\beta}_\lambda; \beta) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + \lambda)^2}\,\langle \boldsymbol{v}_i, \Sigma_0 \boldsymbol{v}_i\rangle \tag{167}$$

$$= \gamma\int \frac{x}{(x + \lambda)^2}\mu_\infty(\mathrm{d}x) = \gamma\int \frac{x}{(x + \lambda)^2}\mu_\infty^+(\mathrm{d}x)\,. \tag{168}$$

Comparing the last two displays, we obtain our claim

$$\lim_{\lambda\to 0^+}\lim_{n\to\infty} V_X(\hat{\beta}_\lambda; \beta) = D_0\,. \tag{169}$$

B.3 Proof of Corollary 5

Throughout this section, set $s = t = 0$ (and drop these arguments for the various functions), and let $M_n(\xi) \equiv \gamma M_{1,n}(\xi) + M_{2,n}(\xi)$, $m_n(\xi) \equiv \gamma m_{1,n}(\xi) + m_{2,n}(\xi)$, $m(\xi) = \gamma m_1(\xi) + m_2(\xi)$. In this case we have

$$A = \begin{bmatrix} \mathbf{0} & \frac{1}{\sqrt{n}}X^T \\ \frac{1}{\sqrt{n}}X & \mathbf{0} \end{bmatrix}\,. \tag{170}$$

and therefore, a simple linear algebra calculation yields

$$M_n(z) = 2z\Big[\gamma\, S_n(z^2) + \frac{1}{2}(\gamma - 1)\frac{1}{z^2}\Big]\,. \tag{171}$$

Therefore, Theorem 7 immediately implies $S_n(z^2) \to s(z^2)$ (almost surely and in L^1), where

$$m(z) = 2z\Big[\gamma\, s(z^2) + \frac{1}{2}(\gamma-1)_+\frac{1}{z^2}\Big]\,. \tag{172}$$

Equations (20) and (21) simplify for the case $s = t = 0$ (setting $\overline{m}_j = m_j(z, s = 0, t = 0)$) to yield

$$-z\overline{m}_1 - \overline{m}_1\overline{m}_2 + \frac{c_1^2\overline{m}_1^2\overline{m}_2^2}{c_1\overline{m}_1\overline{m}_2 - \psi} = 1\,, \tag{173}$$

$$-z\overline{m}_2 - \gamma\overline{m}_1\overline{m}_2 + \frac{\gamma c_1^2\overline{m}_1^2\overline{m}_2^2}{c_1\overline{m}_1\overline{m}_2 - \psi} = 1\,. \tag{174}$$

Taking a linear combination of these two equations, we get

$$-zm - 2\gamma\overline{m}_1\overline{m}_2 + \frac{2\gamma c_1^2\overline{m}_1^2\overline{m}_2^2}{c_1\overline{m}_1\overline{m}_2 - \psi} = 1 + \gamma\,. \tag{175}$$

Comparing this with Eq (172), we get Eq. (25). Substituting the latter in Eqs. (173), (174), we get Eqs. (26), (27).

B.4 Proof of Theorem 7

We apply Lemma 6 with $m_j(\xi, s, t)$ defined as per Theorem 8. We start by noting that

$$\partial_x m(\xi, x, c_1 x)\big|_{x=0} = -\partial_s m(\xi, s, t)\big|_{s=t=0} - c_1\partial_t m(\xi, s, t)\big|_{s=t=0}\,. \tag{176}$$

We will prove the Taylor-Laurent expansions

$$-\partial_s m(\xi, s, t)\big|_{s=t=0} = \frac{D_{-1,s}}{\xi^2} + D_{0,s} + O(\xi^2)\,, \tag{177}$$

$$-\partial_t m(\xi, s, t)\big|_{s=t=0} = \frac{D_{-1,t}}{\xi^2} + D_{0,t} + O(\xi^2)\,, \tag{178}$$

whence Eq. (28) follows with

$$\boxed{D_0 = D_{0,s} + c_1 D_{0,t}\,.} \tag{179}$$

Expressions for $D_{0,s}, D_{0,t}$ will be given below, whence the result for D_0 follows.

B.4.1 Limit $s, t \to 0$

The case $s = t = 0$ was already studied in Section B.3. Letting $m_j^{(0)}(\xi) = m_j(\xi, 0, 0)$, we define $x(\xi) = m_1^{(0)}(\xi)m_2^{(0)}(\xi)$. We need to determine These can be expressed in terms of $s(\xi^2)$ using corollary 5:

$$m_1^{(0)}(\xi) = \xi s(\xi^2)\,, \tag{180}$$

$$m_2^{(0)}(\xi) = \frac{\gamma-1}{\xi} + \gamma\xi s(\xi^2)\,, \tag{181}$$

$$x(\xi) = s(\xi^2)(\gamma - 1 + \gamma\xi^2 s(\xi^2))\,. \tag{182}$$

Equations (25), (26), (27) yield a fourth order algebraic equation. Studying its solution for $\xi \to 0$, yields the following expansions (for $\overline{\gamma} = \gamma \wedge 1$)

$$x(\xi) = x_0 + x_1\xi^2 + O(\xi^4)\,, \tag{183}$$

$$x_0 \equiv \psi\frac{1 - c_1\psi^{-1}\overline{\gamma}^{-1} - \sqrt{(1 - c_1\psi^{-1}\overline{\gamma}^{-1})^2 + 4c_1\psi^{-1}\overline{\gamma}^{-1}(1-c_1)}}{2c_1(1-c_1)}\,, \tag{184}$$

$$x_1 \equiv \frac{x_0}{|\gamma-1|(1 - 2r_0 + (r_0^2/c_1))}\,, \qquad r_0 \equiv 1 + \frac{1}{\overline{\gamma}x_0}\,, \tag{185}$$

as well as

$$m_1^{(0)}(\xi) = \begin{cases} \frac{x_0}{\gamma-1}\xi + O(\xi^3) & \text{if } \gamma < 1, \\ \frac{1-\gamma}{\gamma\xi} + \frac{x_0}{1-\gamma}\xi + O(\xi^3) & \text{if } \gamma > 1. \end{cases} \tag{186}$$

B.4.2 Computation of $\partial_s m$

We obtain

$$\boxed{D_{0,s} = -\frac{\gamma x_0}{\gamma - 1}\,.} \tag{187}$$

B.4.3 Computation of $\partial_t m$

We denote the derivatives of $m_j(\xi, 0, t)$, $j \in \{1, 2\}$, with respect to t by

$$\frac{\partial m_1}{\partial t}(\xi, 0, 0) = m_1^{(0)}(\xi)^2\, u_1\,, \qquad \frac{\partial m_2}{\partial t}(\xi, 0, 0) = u_2\,. \tag{188}$$

By expanding (20), (21) for small t, we get the following expressions for $u_1 = u_1(x(\xi))$, $u_2 = u_2(x(\xi))$:

$$u_1(x) = \frac{-q_3(x)(1+\gamma q_2(x))x^{-2} - q_2(x)q_4(x)}{(1+q_1(x))(1+\gamma q_1(x))x^{-2} - \gamma q_2(x)^2}\,, \tag{189}$$

$$u_2(x) = \frac{-(1+q_1(x))q_4(x) - \gamma q_2(x)q_3(x)}{(1+q_1(x))(1+\gamma q_1(x))x^{-2} - \gamma q_2(x)^2}\,, \tag{190}$$

where we defined the following functions of x

$$q_1(x) \equiv xp(x) - \frac{xp(x)^2}{c_1}\,, \qquad q_2(x) \equiv 1 - 2p(x) + \frac{p(x)^2}{c_1}\,, \tag{191}$$

$$q_3(x) \equiv -\frac{2p(x)}{c_1} + \frac{p(x)^2}{c_1^2}\,, \qquad q_4(x) \equiv -\frac{\gamma p(x)}{c_1 x} + \frac{\gamma p(x)^2}{c_1^2 x}\,, \tag{192}$$

$$p(x) \equiv \frac{c_1^2 x}{c_1 x - \psi}\,. \tag{193}$$

Note that u_1, u_2 are analytic functions of x. Using Eq. (183), we get the Taylor-Laurent expansion

$$u_j(x(\xi)) = u_j(x_0) + \frac{\mathrm{d}u_j}{\mathrm{d}x}(x_0)x_1\xi^2 + O(\xi^4)\,, \tag{194}$$

Using (188), we get

$$\partial_t m(\xi, 0, 0) = \gamma \partial_t m_1(\xi, 0, 0) + \partial_t m_2(\xi, 0, 0) \tag{195}$$

$$= \gamma\, m_1^{(0)}(\xi)^2 u_1(x(\xi)) + u_2(x(\xi))\,. \tag{196}$$

By Eq. (186), for$\gamma < 1$, we get

$$\partial_t m(\xi, 0, 0) = u_2(x_0) + O(\xi^2)\,, \tag{197}$$

On the other hand, for $\gamma > 1$,

$$\partial_t m(\xi, 0, 0) = \frac{(1-\gamma)^2}{\gamma\xi^2}u_1(x_0) + \frac{(1-\gamma)^2}{\gamma}\frac{\mathrm{d}u_1}{\mathrm{d}x}(x_0)x_1 + 2x_0u_1(x_0) + u_2(x_0) + O(\xi^2)\,. \tag{198}$$

We therefore conclude

$$\boxed{D_{0,t} = \begin{cases} -u_2(x_0) & \text{if } \gamma < 1, \\ \frac{(1-\gamma)^2}{\gamma}\frac{\mathrm{d}u_1}{\mathrm{d}x}(x_0)x_1 + 2x_0u_1(x_0) + u_2(x_0) & \text{if } \gamma > 1, \end{cases}} \tag{199}$$

The expression given in Theorem 7 were obtained by simplifying the boxed formulas above.

B.5 Proof of Corollary 4

The variance result follows simply by taking $c_1 \to 0$ in Theorem 7.

For the bias, recall that

$$B_X(\hat{\beta}_\lambda) = \frac{r^2}{p}\mathrm{tr}\left[\lambda^2(\hat{\Sigma}_X + \lambda I_p)^{-2}\Sigma\right] . \tag{200}$$

Define

$$\widetilde{B}_X(\hat{\beta}_\lambda) = \frac{r^2}{p}\mathrm{tr}\left[\lambda^2(\hat{\Sigma}_X + \lambda I_p)^{-2}\right] . \tag{201}$$

By Lemma 13, and using the fact that $\Sigma_0 = I_p$ when $c_1 = 0$, we get

$$\left|B_X(\hat{\beta}_\lambda) - \widetilde{B}_X(\hat{\beta}_\lambda)\right| \le \frac{r^2}{p}\|\lambda^2(\hat{\Sigma}_X + \lambda I_p)^{-2}\|_F\|\Sigma - I_p\|_F \tag{202}$$

$$\le \frac{C}{p}\|\lambda^2(\hat{\Sigma}_X + \lambda I_p)^{-2}\|_{\mathrm{op}}\sqrt{p}(\log n)^C \tag{203}$$

$$\le \frac{C}{\sqrt{n}}(\log n)^C .. \tag{204}$$

with porbability larger than $1 - 1/n^2$. Therefore, by Borel-Cantelli it is sufficient to estabilish the claim for $\widetilde{B}_X(\hat{\beta}_\lambda)$. By Corollary 5, for $c_1 = 0$, the empirical spectral distribution of $\hat{\Sigma}$ converges almost surely (in the weak topology) to the Marchenko-Pastur law μ_{MP}. Hence (for $(\lambda_i)_{i\le p}$ the eigenvalues of $\hat{\Sigma}$)

$$\lim_{n\to\infty}\widetilde{B}_X(\hat{\beta}_\lambda) = r^2 \lim_{n\to\infty}\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda^2}{(\lambda+\lambda_i)^2} \tag{205}$$

$$= r^2\int\frac{\lambda^2}{(\lambda+x)^2}\,\mu_{MP}(\mathrm{d}x) = r^2\lambda^2 s'(-\lambda) . \tag{206}$$

Hence the asymptotic bias is the same as in the linear model (for random isotropic features). The claim hence follows by the results of Section 5.2. Alternatively, we may simply recall that $\mu_{MP}(\{0\}) = (1-\gamma^{-1})_+$ and use dominated convergence.

References

Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. arXiv: 1710.03667, 2017.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares. *International Conference on Artificial Intelligence and Statistics*, 22, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. arXiv: 1811.03962, 2018.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2009.

Zhidong Bai and Jack Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

Zhidong Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294, 1993.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. arXiv: 1812.11118, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. arXiv: 1802.01396, 2018b.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? arXiv: 1806.09471, 2018c.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. arXiv: 1903.07571, 2019.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018a.

Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. arXiv: 1812.07956, 2018b.

Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22 (1):1–37, 2016.

Edgar Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 4(4):1550019, 2015.

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

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. arXiv: 1811.03804, 2018a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. arXiv: 1810.02054, 2018b.

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

Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, pages 1–59, 2015.

Jerome Friedman and Bogdan Popescu. Gradient directed regularization. http://www-stat.stanford.edu/ jhf/ftp/pathlite.pdf, 2004.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stephane d'Ascoli, Giulio Biroli, Clement Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. arXiv: 1901.01608, 2019.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. arXiv: 1902.06720, 2019.

Ker-Chau Li. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics*, 14(3):1101–1112, 1986.

Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics*, 15(3):958–975, 1987.

Percy Liang and Nati Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. *International Conference on Machine Learning*, 27, 2010.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. arXiv: 1808.00387, 2018.

Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

Leo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. arXiv: 1811.01212, 2018.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 21, 2008.

James Ramsay. Parameter flows. Working paper, 2005.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. arXiv: 1805.00915, 2018.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Probability Letters*, 81(5):592–602, 2011.

Vadim I. Serdobolskii. *Multiparametric Statistics*. Elsevier, 2007.

Jack Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks. arXiv: 1805.01053, 2018.

Stefano Spigler, Mario Geiger, Stephane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. arXiv: 1810.09665, 2018.

Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Society, 2012.

Ryan J. Tibshirani. A general framework for fast stagewise algortihms. *Journal of Machine Learning Research*, 16: 2543–2588, 2015.

William F. Trench. Asymptotic distribution of the spectra of a class of generalized Kac-Murdock-Szego matrices. *Linear Algebra and Its Applications*, 294(1–3):181–192, 1999.

Antonia M. Tulino and Sergio Verdu. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1):1–182, 2004.

Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. *Advances in Neural Information Processing Systems*, 30, 2017.

Ji Xu, Arian Maleki, and Kamiar Rahnama Rad. Consistent risk estimation in high-dimensional linear regression. arXiv: 1902.01753, 2019.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. arXiv: 1611.03530, 2016.

Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? arXiv: 1902.01996, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. arXiv: 1811.08888, 2018.