Journal of Machine Learning Research 22 (2021) 1-82 Submitted 10/20; Revised 2/21; Published 4/21

What Causes the Test Error? Going Beyond Bias-Variance via ANOVA

Licong Lin LLC2000@PKU.EDU.CN
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Beijing, China

Edgar Dobriban DOBRIBAN@WHARTON.UPENN.EDU
Departments of Statistics & Computer and Information Science
University of Pennsylvania
Philadelphia, PA, 19104-6340, USA

Editor: Ambuj Tewari

Abstract

Modern machine learning methods are often overparametrized, allowing adaptation to the data at a fine level. This can seem puzzling; in the worst case, such models do not need to generalize. This puzzle inspired a great amount of work, arguing when overparametrization reduces test error, in a phenomenon called "double descent". Recent work aimed to understand in greater depth why overparametrization is helpful for generalization. This lead to discovering the unimodality of variance as a function of the level of parametrization, and to decomposing the variance into that arising from label noise, initialization, and randomness in the training data to understand the sources of the error.

In this work we develop a deeper understanding of this area. Specifically, we propose using *the analysis of variance* (ANOVA) to decompose the variance in the test error in a symmetric way, for studying the generalization performance of certain two-layer linear and non-linear networks. The advantage of the analysis of variance is that it reveals the effects of initialization, label noise, and training data more clearly than prior approaches. Moreover, we also study the monotonicity and unimodality of the variance components. While prior work studied the unimodality of the overall variance, we study the properties of each term in the variance decomposition.

One of our key insights is that often, the *interaction* between training samples and initialization can dominate the variance; surprisingly being larger than their marginal effect. Also, we characterize "phase transitions" where the variance changes from unimodal to monotone. On a technical level, we leverage advanced deterministic equivalent techniques for Haar random matrices, that—to our knowledge—have not yet been used in the area. We verify our results in numerical simulations and on empirical data examples.

Keywords: Test Error, ANOVA, Double Descent, Ridge Regression, Random Matrix Theory

1. Introduction

Modern machine learning methods are often overparametrized, allowing adaptation to the data at a fine level. For instance, competitive methods for image classification—such as WideResNet (Zagoruyko and Komodakis, 2016)—and for text processing—such as GPT-

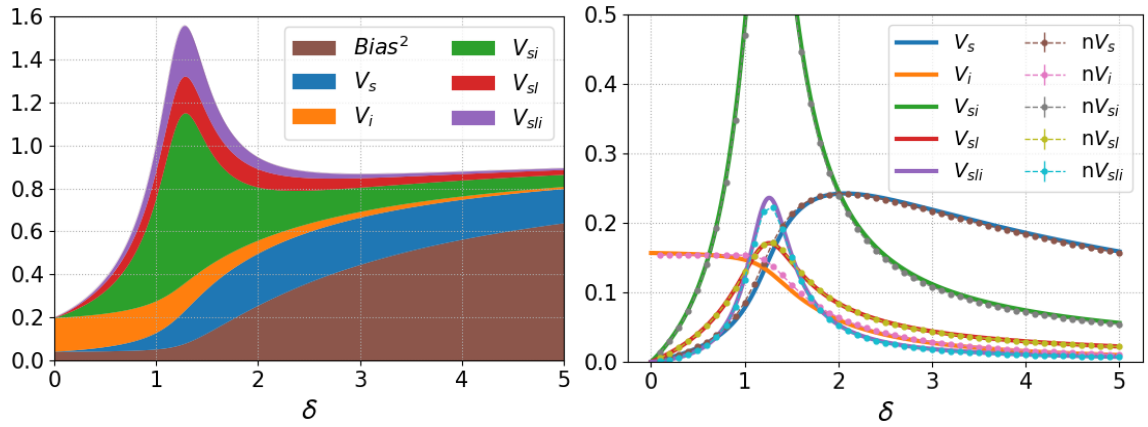


Figure 1: ANOVA decomposition of the variance. The plots show the components of the variance (as well as the bias) in certain two-layer linear networks studied in the paper, as a function of the data aspect ratio $\delta = \lim d/n$, where d is the dimension of features and n is the number of samples. The variance can be decomposed into its contributions from randomness in label noise (l), training data/samples (s), and initialization (i). Namely, the variance is decomposed into the main effects V_a and the interaction effects V_{ab}, V_{abc}, where $a, b, c \in \{l, s, i\}$. We omit V_l, V_{li} in the figures since they equal zero. The key observation is that the *interaction effects* (especially V_{si}) dominate the variance at the interpolation limit where $\lim p/n = 1$ (this turns out to correspond to $\delta = 1.25$ in the Figure) and p is the number of features in the hidden layer. **Left**: Cumulative figure of the bias and variance components. **Right**: Variance components in numerical simulations. ($\star$: theory, $n\star$: numerical, averaged over 5 runs, for $\star = V_s$, etc). Parameters: signal strength $\alpha = 1$, noise level $\sigma = 0.3$, regularization parameter $\lambda = 0.01$, parametrization level $\pi = 0.8$. See Sections 2.2, 4.2 for details.

3 (Brown et al., 2020)—have from millions to billions of explicit optimizable parameters, comparable to the number of datapoints. From a theoretical point of view, this can seem puzzling and perhaps even paradoxical: in the worst case, models with lots of parameters do not need to generalize (i.e., perform similarly on test data as on training data from the same distribution).

This puzzle has inspired a great amount of work. Without being exhaustive, some of the main approaches argue the following. (1) Overparametrization beyond the "interpolation threshold" (number of parameters required to fit the data) can eventually reduce test error (in a phenomenon called "double descent"). (2) The specific algorithms used in the training process have beneficial "implicit regularization" effects which effectively reduce model complexity and help with generalization. These two ideas are naturally connected, as the implicit regularization helps achieve decreasing test error with overparametrization. This area has registered a great deal of progress recently, but its roots can be traced back many years ago. We discuss some of these works in the related work section.

One particular line of work aims to understand in greater depth why overparametrization is helpful for generalization. In this line of work, Yang et al. (2020) has studied the bias-variance decomposition of the mean squared error (and for other losses), and pro-

posed that a key phenomenon is that the variance is *unimodal* as a function of the level of parametrization. This was verified empirically for a wide range of models including modern neural networks, as well as theoretically for certain two-layer linear networks with only the second layer trained. Moreover, d'Ascoli et al. (2020) proposed to decompose the variance in a two-layer non-linear network with only second layer trained (i.e., a random features model) into that arising from label noise, initialization, and randomness in the features of the training data (in this specific order), arguing that—in their particular model—the label and initialization noise dominates the variance.

In this work we develop a set of techniques aiming to improve our understanding of this area; and more broadly of generalization in machine learning. Specifically, we propose to use *the analysis of variance* (ANOVA), a classical tool from statistics and uncertainty quantification (e.g., Box et al., 2005; Owen, 2013), to decompose the variance in the generalization mean squared error into its components stemming from the initialization, label noise, and training data (see Figure 1 for a brief example). The advantage of the analysis of variance is that it reveals the effects of the components in a more clearly interpretable, and perhaps "unequivocal", way than the approach in d'Ascoli et al. (2020). The prior decomposition depends on the specific order in which the conditional expectations are evaluated, while ours does not. We carry out this program in detail in certain two-layer linear and non-linear network models (more specifically, random feature models), which have have been the subject of intense recent study, and are effectively at the frontier of our theoretical understanding.

As is well known in the literature on ANOVA, the variance components form a hierarchy whose first level, the *main effects*, can be interpreted as the effects of varying each variable (here: random initialization, features, label noise) separately, while the higher levels can be interpreted as the interaction effects between them. These are symmetric, which is both elegant and interpretable, and thus provide advantages over the prior approaches. See Figure 2 for an example.

Moreover, we study the monotonicity and unimodality of MSE, bias, variance, and the various variance components in a specific variance decomposition. While Yang et al. (2020) studied the unimodality of the overall variance, we study the properties the components individually. On a technical level, our work is quite involved, and leverages some advanced techniques from random matrix theory, that—to our knowledge—have not yet been used in the area. In particular, we discovered that we can leverage the deterministic equivalent results for Haar random matrices from Couillet et al. (2012). These have been developed for different purposes, for analyzing random beamforming in wireless communications.

After the initial posting of our work, we became aware of the highly related paper Adlam and Pennington (2020b). This was publicly posted on the `arxiv.org` preprint server later than our work, but had been submitted for publication earlier. The conclusions in the two works are similar, but the techniques and setting are different. We discuss this at the end of the next section.

1.1 Related Works

There is an extraordinary amount of related work, as this topic is one of the most exciting and popular ones recently in the theory of machine learning. Due to space limitations, we can only review the most closely related work.

The phenomenon of "double descent", coined in Belkin et al. (2019), states that the limiting test error first increases, then decreases as a function of the parametrization level, having a "double descent", or "w"-shaped behavior. This phenomenon has been studied, in one form or another, in a great number of recent works, see e.g., Advani et al. (2020); Bartlett et al. (2020); Belkin et al. (2019, 2018, 2020b); Dereziński et al. (2019); Geiger et al. (2020); Ghorbani et al. (2021); Hastie et al. (2019); Liang and Rakhlin (2018); Li et al. (2020); Mei and Montanari (2019); Muthukumar et al. (2020); Xie et al. (2020), etc.

Various forms have also appeared in earlier works, see e.g., the discussion on "A brief prehistory of double descent" (Loog et al., 2020) and the reply in Belkin et al. (2020a). This points to the related works Opper (2001); Krämer (2009). The online machine learning community has engaged in a detailed historical reference search, which unearthed the related early works[1] Hertz et al. (1989); Opper et al. (1990); Hansen (1993); Barber et al. (1995); Duin (1995); Opper (1995); Opper and Kinzel (1996); Raudys and Duin (1998). The observations on the "peaking phenomenon" are consistent with empirical results on training neural networks dating back to the 1990s. There it has been suggested that the difficulties captured by the peak in double descent stem from optimization, such as the ill-conditioning of the Hessian (LeCun et al., 1991; Le Cun et al., 1991).

Some works that are especially relevant to us are the following. Hastie et al. (2019) showed that the limiting MSE of ridgeless interpolation in linear regression as a function of the overparametrization ratio, for fixed SNR, has a double descent behavior. Nakkiran et al. (2021) rigorously proved that optimally regularized ridge regression can eliminate double descent in finite samples in a linear regression model. Nakkiran (2019) clearly explained that "more data can hurt", because algorithms do not always adapt well to the additional data. In comparison, the special case of our results pertaining to linear nets allows for certain non-Gaussian data, while only proved asymptotically. Nakkiran et al. (2020) empirically showed a double descent shape for the test risk for various neural network architectures a function of model complexity, number of samples ("sample-wise" double descent), and training epochs.

d'Ascoli et al. (2020) used the (not fully rigorous) replica method to obtain the bias-variance decomposition for two-layer neural networks in the lazy training regime. They further also decomposed the variance in a specific order into that stemming from label noise, initialization, and training features. Compared to this, our work is fully rigorous, and proposes to use the analysis of variance, from which we show that the sequential decompositions like the ones proposed in d'Ascoli et al. (2020) can be recovered. Moreover, we are concerned with a slightly different model (with orthogonal initialization), and some of our results are only proved for linear orthogonal networks (e.g., the forms of the variance components). However, going beyond d'Ascoli et al. (2020), we also obtain rigorous results for the monotonicity and unimodality of the various elements of the variance decomposition.

1. The reader can see the Twitter thread by Dmitry Kobak: `https://twitter.com/hippopedoid/status/1243229021921579010`.

Ba et al. (2020) obtained the generalization error of two-layer neural networks when only training the first or the second layer, and compared the effects of various algorithmic choices involved. Compared with our work, d'Ascoli et al. (2020); Ba et al. (2020) studied more general settings and provided results that involve more complex expressions; the advantage our our simpler expressions is that we can find the variance components and study properties such as their monotonicity. Our results are simpler mainly because we consider orthogonal initialization; and, in several results, consider a linear network. We believe that our results are complementary.

Wu and Xu (2020) calculated the prediction risk of optimally regularized ridge regression under a general covariance assumption of the data. Jacot et al. (2020) argued that random feature models can be close to kernel ridge regression with additional regularization. This is related to the "calculus of deterministic equivalents" for random matrices (Dobriban and Sheng, 2018). Liang et al. (2020) argued that in certain kernel regression problems one may obtain generalization curves with multiple descent points. Chen et al. (2020) studied certain models with provable multiple descent curves.

When the data has general covariance, Kobak et al. (2020) showed that the optimal ridge parameter could be negative. Thus any positive ridge penalty would be sub-optimal if the true parameter vector lies on a direction with high predictor variance. Understanding the implications of this work in our context is a subject of interesting future research.

More broadly viewed, a great deal of effort has been focused on connecting "classical" statistical theory (focusing on low-dimensional models) with "modern" machine learning (focusing on overparametrized models).[2] From this perspective, there are strong analogies with nonparametric statistics (Ibragimov and Has′ Minskii, 2013). Non-parametric estimators such as kernel smoothing have, in effect, infinitely many parameters, yet they can perform well in practice and have strong theoretical guarantees. Nonparameteric statistics already has the same components of the "overparametrize then regularize" principle as in modern machine learning. The same principle also arises in high-dimensional statistics, such as with basis pursuit and Lasso (Chen and Donoho, 1994). Namely, one can get good performance if one considers a large set of potential predictors (overparametrize), and then selects a small, highly-regularized subset.

Even more broadly, our work is connected to the emerging theme in modern statistics and machine learning of studying high-dimensional asymptotic limits, where both the sample size and the dimension of data tend to infinity. This is a powerful framework that allows us to develop new methods, and to uncover phenomena not detectable using classical fixed-dimension asymptotics (see e.g., Couillet and Debbah, 2011; Paul and Aue, 2014; Yao et al., 2015). It also dates back to the 1970s, see e.g., the literature review in Dobriban and Wager (2018), which points to works by Raudys (1967); Deev (1970); Serdobolskii (1980), etc. Some other recent related works include Pennington and Worah (2017); Louart et al. (2018); Liao and Couillet (2018, 2019); Benigni and Péché (2019); Goldt et al. (2019); Fan and Wang (2020); Deng et al. (2019); Gerace et al. (2020); Liao et al. (2020); Adlam et al. (2019); Adlam and Pennington (2020a). See also Geman et al. (1992); Bös and Opper

2. The reader can see e.g., the talks titled "From classical statistics to modern machine learning" by M. Belkin at the Simons Institute (`https://simons.berkeley.edu/talks/tbd-65`) at the Institute of Advanced Studies (`https://video.ias.edu/theorydeeplearning/2019/1016-MikhailBelkin`), and and other venues.

(1997); Neal et al. (2018) for various classical and modern discussions of bias-variance tradeoffs and dynamics of training.

The most closely related work to ours is Adlam and Pennington (2020b), publicly posted later, but submitted for publication earlier. Both works study the generalization error via ANOVA decomposition, and show that the interaction effect can dominate the variance. The conclusions in the two works are similar. For instance, the V_{si} term (interaction between samples and initialization, defined later) dominates the total variance; V_{si} and V_{sli} (interaction between samples, label noise and initialization) diverge as the ridge regularization parameter $\lambda \to 0$. On the other hand, there are many differences between two works. (1). The mathematical settings are different. Their work studies a two-layer nonlinear network with Gaussian initialization, while we study both linear and nonlinear networks with orthogonal initialization. (2). The mathematical tools employed in the two papers are different. They use Gaussian equivalents and the linear pencil representation, while we exploit orthogonal deterministic equivalents. (3). The results are different. Beyond the ANOVA decomposition, they also study the effect of ensemble learning. On our end, we study optimally tuned ridge regression and prove properties of the bias, variance and MSE.

Another related paper, also publicly posted after our work is by Rocks and Mehta (2020). They study generalization error in linear regression and two-layer networks by deriving the formulas for bias and variance. The main techinque they used is the cavity method originating from statistical physics. Similarly, they also show that the generalization error diverges at the interpolation threhold due to the large variance. We provide a more detailed comparison later, after stating our main results.

As already mentioned, our work is related to the one by Yang et al. (2020). The model we consider is related to theirs, with several key differences. One is the orthogonal initialization, in contrast to their Gaussian initialization. Also, they assume that the ratio $d/n \to 0$ while we study the proportional regime where $d/n \to \delta > 0$ (which can be arbitrarily small, so our setting is in a sense effectively more general). As for the results, they prov the unimodality of variance and monotonicity of the bias under their setting. They also make some conjectures on the variance unimodality that we prove (keeping in mind the different settings), see the results section for more details.

1.2 Our Contributions

Our contributions can be summarized as follows:

1. We study a two-layer linear network where the first layer is a fixed partial orthogonal embedding (which determines the latent features) and the second layer is trained with ridge regularization. While the expressive power of this model only captures certain linear functions, training only the second layer already exhibits certain intriguing statistical and generalization phenomena. We study the prediction error of this learning method in a noisy linear model. We consider three sources of randomness that contribute to the error: the random initialization (a random partial orthogonal embedding), the label noise, and the randomness over the training data. We propose to use *the analysis of variance* (ANOVA), a classical tool from statistics and uncertainty quantification (e.g., Box et al., 2005; Owen, 2013) to decompose and understand their contribution.

We study an asymptotic regime where the data dimension, sample size, and number of latent features tends to infinity together, proportionally to each other. In this model, we calculate the limits of the variance components (Theorem 2); in terms of moments of the Marchenko-Pastur distribution (Marchenko and Pastur, 1967). We then show how to recover various sequential variance decompositions, such as the one from d'Ascoli et al. (2020) (albeit only for linear rather than nonlinear networks). We also show that the order in the sequence of decompositions matters. Our work leverages deterministic equivalent results for Haar random matrices from Couillet et al. (2012) that, to our knowledge, have not yet been used in the area. We also leverage recent technical developments such as the *calculus of deterministic equivalents* for random matrices of the sample covariance type (Dobriban and Sheng, 2018, 2020). Proofs are in Appendix B.

2. We then study the bias-variance decomposition in greater detail. As a corollary of the ANOVA results, we study the decomposition of the variance in the order label-sample-initialization, which has some special properties (Theorem 3). When using an optimal ridge regularization, we study the monotonicity and unimodality properties of these components (Theorem 5 and Table 1). With this, we shed further light on phenomena discovered by Yang et al. (2020), who wrote that "The main unexplained mystery is the unimodality of the variance". Specifically, we are able to show that the variance is indeed unimodal in a broad range of settings. This analysis goes beyond prior works e.g., Yang et al. (2020) (who studied setting with a number of inner neurons being much larger than the number of datapoints), or "double descent mitigation" as in Nakkiran et al. (2021), because it studies bias and variance separately.

   We uncover several intriguing properties: for instance, for a fixed parametrization level π, as a function of aspect ratio or "dimensions-per-sample", the variance is monotonically decreasing when $\pi < 0.5$, and unimodal when $\pi \geq 0.5$. We discuss and offer possible explanations.

   We also discuss the special case of linear models, which has received a great deal of prior attention (Proposition 6). We view the results on standard linear models as valuable, as they are both simpler to state and to prove, and moreover they also directly connect to some prior work.

3. We develop some further special properties of the bias, variance, and MSE. We report a seemingly surprising simple relation between the MSE and bias at the optimum (Section 2.3.1). We study the properties of the bias and variance for a *fixed* (as opposed to optimally tuned) ridge regularization parameter (Theorem 7). In particular, we show that the bias decreases as a function of the parametrization, and increases as a function of the data aspect ratio. In contrast to choosing λ optimally, we see that double descent is *not* mitigated, and may occur in our setting when we use a small regularization parameter λ that is fixed across problem sizes (going beyond the models where this was known from prior work). This corroborates that the lack of proper regularization plays a crucial role for the emergence of double descent.

We also give an added noise interpretation of the initial random initialization step (Section 2.3.3). Further, we provide some detailed analysis and intuition of these phenomena, aided by numerical plots of the variance components (Section 2.3.4).

4. The above results are about ridge regression as a heuristic for regularized empirical risk minimization. In some settings, ridge regularization is known to have limitations (Derezinski and Warmuth, 2014), thus it is an importat question to understand its fundamental limitations here. In fact, we can show that ridge regression is an asymptotically optimal estimator, in the sense that it converges to the Bayes optimal estimator in our model (Theorem 8). This provides some justification for studying ridge regression in a two-layer network, which is not covered by standard results.

5. We extend some of our results to two-layer networks with a non-linear activation function with orthogonal initialization. In particular, we provide the limits of the MSE, bias, and variance in the same asymptotic regime (Theorem 9). Furthermore, we provide the monotonicity and unimodality properties of these quantities as a function of parametrization and aspect ratio (Table 2).

6. We provide numerical simulations to check the validity of our theoretical results (Section 4), including the MSE, the bias-variance decomposition, and the variance components. We also show some experiments on empirical data, specifically on the superconductivity data set (Hamidieh, 2018), where we test our predictions for two-layer orthogonal nets. Code associated with the paper is available at `https://github.com/licong-lin/VarianceDecomposition`.

1.3 Highlights and Implications

We discuss some of the highlights and implications of our results.

Beyond bias-variance. Much of the prior work in this area has focused on the fundamental bias-variance decomposition. In this work, we demonstrate that it is possible to go significantly beyond this via the ANOVA decomposition. Specifically, using this methodology, one can understand how the random training data, initialization, and label noise contribute to the test error in more detailed and comprehensive ways than what was previously possible. We carry out this in certain two-layer linear and non-linear networks with only the second layer trained (i.e., random features models), but our approach may be more broadly relevant.

Non-additive test error. A key finding of our work is that in the specific neural net models considered here, the random training data, initialization, and label noise contribute highly non-additively to the test error. Thus, when discussing "the effects of initialization", some care ought to be taken; i.e., to clarify which interaction effects (e.g., with label noise or training data) this includes. The interaction term between the initialization and the training data can be large in our setting.

Beyond double descent: Prevalence of unimodality. While initial work on asymptotic generalization error of one and two-layer neural nets focused on the "double descent" or peaking shape of the test error, our work gives further evidence that the unimodal

shape of the variance is a prevalent phenomenon. Moreover, our work also suggests the the unimodality holds not just for the overall variance, but also for specific and variance components; which was not known in prior work. We show that unimodality with respect to both overparametrization level and data aspect ratio holds in specific parameter settings for the variance and certain other decompositions for the optimal setting of the regularization parameter. In other parameter settings, we obtain monotonicity results for these components. This also underscores that regularization and the associated bias-variance tradeoff plays a key role in determining monotonicity and unimodality.

2. ANOVA for a Two-layer Linear Network

2.1 Setup

In this section, we study the bias-variance tradeoff and ANOVA decomposition for a two-layer linear network model. Suppose that we have a training data set $\mathcal{T}$ containing n data points $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, with features x_i and outcomes y_i. We assume the data is drawn independently from a distribution such that $x_i = (x_{i1}, x_{i2}, ..., x_{id})$, where x_{ij} are i.i.d. random variables satisfying

$$\mathbb{E}x_{ij} = 0, \qquad \mathbb{E}x_{ij}^2 = 1, \qquad \mathbb{E}x_{ij}^{8+\eta} < \infty,$$

where $\eta > 0$ is an arbitrary constant. Also, each (x_i, y_i) are drawn from the model

$$y = f^*(x) + \varepsilon = x^\top \theta + \varepsilon,\ \theta \in \mathbb{R}^d,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the label noise independent of x and $\sigma \geq 0$ is the noise standard deviation. In matrix form, $Y = X\theta + \mathcal{E}$, where $X = (x_1, x_2, ..., x_n)^\top \in \mathbb{R}^{n\times d}$ has input vectors x_i, $i = 1, 2, ..., n$ as its rows and $Y = (y_1, y_2, ..., y_n)^\top \in \mathbb{R}^{n\times 1}$ with output values y_i, $i = 1, 2, ..., n$ as its entries. Our task is to learn the true regression function $f^*(x) = x^\top\theta$ by using a two-layer linear neural network with weights $W \in \mathbb{R}^{p\times d}$, $\beta \in \mathbb{R}^{p\times 1}$, which computes for an input $x \in \mathbb{R}^d$,

$$f(x) = (Wx)^\top \beta. \tag{1}$$

Later in Section 3 we will also study two-layer nonlinear networks. For analytical tractability, we assume that the true parameters θ are random: $\theta \sim \mathcal{N}(0, \alpha^2 I_d/d)$. Here α^2 can be viewed as a signal strength parameter. This assumption corresponds to performing an "average-case" analysis of the difficulty of the problem over random problem instances given by various θ.

We also consider a random *orthogonal* initialization W independent of $\mathcal{T}$, so W is a $p \times d$ matrix uniformly distributed over the set of matrices satisfying $WW^\top = I_p$, also known as the Stiefel manifold. This requires that $p \leqslant d$, so the dimension of the inner representation of the neural net is not larger than the number of input features. To an extent, this can be seen as a random projection model, where a lower-dimensional representation of the high-dimensional input features is obtained by randomly projecting the input features into a subspace. Both training and prediction are based on the lower dimensional representation. In some works studying the orthogonal initialization of neural networks (e.g., Hu et al.,

2020), the first layer weights W_1 satisfy $W_1^\top W_1 = I_1$, while the last layer weights W_L satisfy $W_L W_L^\top = I_n$, so the dimension of the hidden representation is larger than the dimension of the input features. Similarly, in several recent works on wide neural networks, the number of inner neurons is large. However, we think that in many applications, the number of "higher level features" should indeed not be larger than the number of input features. For instance, the number of features in facial image data such as eyes, hair, is expected to be not more than the number of pixels.

The model we consider here is related to the one from Yang et al. (2020), with several key differences. The orthogonal initialization is a key difference, as Yang et al. (2020) assume that W is a random Gaussian matrix. The expressive power of the two models is the same, but orthogonal initialization has some benefits (see Appendix A for more information).

During training, we fix W and estimate β by performing ridge regression:

$$\hat{\beta}_{\lambda,\mathcal{T},W} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|Y - (WX^\top)^\top\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2, \tag{2}$$

where $\lambda > 0$ is the regularization parameter. This has a closed-form solution

$$\hat{\beta}_{\lambda,\mathcal{T},W} = \left(\frac{WX^\top XW^\top}{n} + \lambda I_p\right)^{-1} \frac{WX^\top Y}{n}. \tag{3}$$

We will often use the notation $R = (WX^\top XW^\top/n + \lambda I_p)^{-1}$ for the so-called resolvent matrix of $WX^\top XW^\top$. By plugging it into (1), we obtain our estimated prediction function, for a new datapoint x, projected first via W and thus accessed via Wx:

$$f(x) = (Wx)^\top \hat{\beta}_{\lambda,\mathcal{T},W} = x^\top W^\top R \frac{WX^\top Y}{n}. \tag{4}$$

Ridge regression is equivalent to ℓ_2 weight decay, a popular heuristic. We will later show that ridge has some asymptotic optimality properties in our model, which thus justify its choice. In contrast, if we follow the approach from Hu et al. (2020) and take $p \geqslant d$ with $W^\top W = I_d$, then it is readily verified that we would obtain

$$f(x) = x^\top \left(X^\top X/n + \lambda I_d\right)^{-1} X^\top Y/n.$$

This means that the prediction function reduces to standard ridge regression. Thus we assume instead that $WW^\top = I_p$ and this makes our model resemble the "feature extraction" layers of a neural network.

We will consider the following asymptotic setting. Let $\{p_d, d, n_d\}_{d=1}^\infty$ be a sequence such that $p_d \leq d$ and $p_d, d, n_d \to \infty$ proportionally, i.e.,

$$\lim_{d\to\infty} \frac{p_d}{d} = \pi, \qquad\qquad \lim_{d\to\infty} \frac{d}{n_d} = \delta,$$

where $\pi \in (0, 1]$ and $\delta \in (0, \infty)$. Here $\pi \in (0, 1]$ denotes the parametrization factor, i.e., the number of parameters in β relative to the input dimension. Also, $\delta > 0$ is the data aspect ratio. We will also use $\gamma = \delta\pi = \lim p/n$, the ratio of learned parameters to number of samples. In Yang et al. (2020), the assumption $\lim_{d\to\infty} d/n_d = 0$ implies that the number of samples is much larger than the number of parameters, which is limiting in high dimensional problems. Thus, we study a broader setting, in which the sample size is proportional to the model size, with an arbitrary ratio.

2.2 Bias-Variance Decompositions and ANOVA

2.2.1 Introduction and the Main Result

Now we analyze the mean squared prediction error in our model,

$$\begin{aligned}\mathbf{MSE}(\lambda) &= \mathbb{E}_{\theta,x,\varepsilon,X,\mathcal{E},W}(f_{\lambda,\mathcal{T},W}(x)-y)^2 \\ &= \mathbb{E}_{\theta,x,X,\mathcal{E},W}(f_{\lambda,\mathcal{T},W}(x)-x^\top\theta)^2+\sigma^2.\end{aligned}$$

The expectation is over a random test datapoint (x, y) from the same distribution as the training data, i.e., $x \in \mathbb{R}^d$ has i.i.d. zero mean, unit variance entries with finite $8+\eta$ moment, $y = \theta^\top x + \varepsilon$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. It is also over the random training input X, the random training label noise $\mathcal{E}$, the random initialization W, and the random true parameter θ. Thus, this MSE corresponds to an average-case error over various training data sets, initializations, and true parameters. In this work, we will always average over the random test data point x, as is usual in classical statistical learning. We will also average over the random parameter θ, which corresponds to a Bayesian average-case analysis over various generative models. This is partly for technical reasons; we can also show almost sure convergence over the random θ, but the analysis for general θ is beyond our scope. Thus, we write the MSE as $\mathbb{E}_{\theta,x}\mathbb{E}_{X,\mathcal{E},W}(f_{\lambda,\mathcal{T},W}(x)-x^\top\theta)^2+\sigma^2$, and the outer expectation $\mathbb{E}_{\theta,x}$ is always present in our formulas.

We can study the mean squared error via the standard bias-variance decomposition corresponding to the average prediction function $\mathbb{E}_{X,\mathcal{E},W} f_{\lambda,\mathcal{T},W}(x)$ over the random training set $\mathcal{T}$ $(X, \mathcal{E})$ and initialization W: $\mathbf{MSE}(\lambda) = \mathbf{Bias}^2(\lambda) + \mathbf{Var}(\lambda) + \sigma^2$, where

$$\begin{aligned}\mathbf{Bias}^2(\lambda) &= \mathbb{E}_{\theta,x}(\mathbb{E}_{X,\mathcal{E},W} f_{\lambda,X,\mathcal{E},W}(x)-x^\top\theta)^2, \\ \mathbf{Var}(\lambda) &= \mathrm{Var}[\hat{f}(x)] = \mathbb{E}_{\theta,x}\mathbb{E}_{X,\mathcal{E},W}(f_{\lambda,\mathcal{T},W}(x)-\mathbb{E}_{X,\mathcal{E},W} f_{\lambda,\mathcal{T},W}(x))^2.\end{aligned}$$

A key point is that the variance can be further decomposed into the components due to the randomness in the training data X, label noise $\mathcal{E}$, and initialization W.

We use s, l, i to represent the samples X, label noise $\mathcal{E}$ and initialization W, respectively. Their impact on the variance can be decomposed in a symmetric way into their main and interaction effects via the the analysis of variance (ANOVA) decomposition (e.g., Box et al., 2005; Owen, 2013) as follows:

$$\mathrm{Var}[\hat{f}(x)] = V_s + V_l + V_i + V_{sl} + V_{si} + V_{li} + V_{sli},$$

where

$$\begin{aligned}V_a &= \mathbb{E}_{\theta,x}\mathrm{Var}_a[\mathbb{E}_{-a}(\hat{f}(x)|a)], && a \in \{s,l,i\} \\ V_{ab} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{ab}[\mathbb{E}_{-ab}(\hat{f}(x)|a,b)] - V_a - V_b, && a,b \in \{s,l,i\}, a \neq b. \\ V_{abc} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{abc}[\mathbb{E}_{-abc}(\hat{f}(x)|a,b,c)] - V_a - V_b - V_c - V_{ab} - V_{ac} - V_{bc} \\ &= \mathrm{Var}[\hat{f}(x)] - V_s - V_l - V_i - V_{sl} - V_{si} - V_{li}, && \{a,b,c\} = \{s,l,i\}.\end{aligned}$$

Here, $\mathbb{E}_{-a}$ means taking expectation with respect to all components except for a. Then $V_a \geqslant 0$ can be interpreted as the effect of varying a alone, also referred to as the *main*

effect of a. Also, $V_{ab} = V_{ba} \geqslant 0$ can be interpreted as the second-order interaction effect between a and b beyond their main effects, and $V_{abc} \geqslant 0$ can be seen as the interaction effect among a, b, c, beyond their pairwise interactions. The ANOVA decomposition is symmetric in a, b, c. We will show how to recover some sequential variance decompositions from the ANOVA decomposition later. We mention that this decomposition is sometimes referred to as functional ANOVA (Owen, 2013).

For intuition, consider the noiseless case when the label noise equals zero, so $\mathcal{E} = 0$ and the index l does not contribute to the variance. Then the main effect of the samples/training data is $V_s = \mathbb{E}_{\theta,x}\mathrm{Var}_s[\mathbb{E}_{-s}(\hat{f}(x)|s)] = \mathbb{E}_{\theta,x}\mathrm{Var}_X[\mathbb{E}_W(\hat{f}(x)|X)]$. This can be interpreted as the variance of the ensemble estimator $\tilde{f}(x) := \mathbb{E}_W\hat{f}(x)$, the average of models parametrized by various W-s on the same data set X. Therefore, V_s can be regarded as the expected variance with respect to training data of the ensemble estimator $\tilde{f}$; where the expectation is over the test datapoint x and the true parameter θ. Furthermore, $V_{si} + V_i$ is the variance that can be eliminated by ensembling. This is because $V_{si} + V_s + V_i$ is the total variance of the estimator $\hat{f}(x)$; and V_s is the variance of the ensemble, thus $V_{si} + V_i$ is the remaining variance that can be eliminated.

Our bias-variance decomposition is slightly different from the standard one in statistical learning theory (e.g., Hastie et al., 2009, p. 24), where the variability is introduced only by the random training set $\mathcal{T}$. Here we also consider the variability due to the random initialization W; and decompose the variance due to $\mathcal{T}$ into that due to samples X and label noise $\mathcal{E}$. The motivation is because randomization in the algorithms, such as random initialization, as well as randomness in stochastic gradient descent, are very common in modern machine learning. Our decomposition helps understand such scenarios.

Now, we define the following quantities which are frequently used throughout our paper. These are "resolvent moments" of the well-known Marchenko-Pastur (MP) distribution F_γ (Marchenko and Pastur, 1967). The MP distribution is the limit of the distribution of eigenvalues of sample covariance matrices $n^{-1}Z^\top Z$ of $n \times p$ data matrices Z with iid zero-mean unit-variance entries when $n, p \to \infty$ with $p/n \to \gamma > 0$ (Bai and Silverstein, 2010; Couillet and Debbah, 2011; Anderson et al., 2010; Yao et al., 2015).

Definition 1 (Resolvent moments) *In this paper, we use the first and second resolvent moments*

$$\theta_1(\gamma, \lambda) := \int \frac{1}{x + \lambda} dF_\gamma(x), \qquad \theta_2(\gamma, \lambda) := \int \frac{1}{(x + \lambda)^2} dF_\gamma(x), \tag{5}$$

where $F_\gamma(x)$ is the Marchenko-Pastur distribution with parameter γ. Recall that for us, $\gamma = \delta\pi$. Then $\theta_1 := \theta_1(\gamma, \lambda)$ and $\theta_2 := \theta_2(\gamma, \lambda)$ have explicit expressions (see e.g., Bai and Silverstein, 2010, for the first one; and our proofs also contain the derivations):

$$\theta_1 = \frac{(-\lambda + \gamma - 1) + \sqrt{(-\lambda + \gamma - 1)^2 + 4\lambda\gamma}}{2\lambda\gamma}, \tag{6}$$

$$\theta_2 = -\frac{d}{d\lambda}\theta_1 = \frac{(\gamma - 1)}{2\gamma\lambda^2} + \frac{(\gamma + 1) \cdot \lambda + (\gamma - 1)^2}{2\gamma\lambda^2\sqrt{(-\lambda + \gamma - 1)^2 + 4\lambda\gamma}}. \tag{7}$$

We further define

$$\tilde{\lambda} := \lambda + \frac{1 - \pi}{2\pi}\left[\lambda + 1 - \gamma + \sqrt{(\lambda + \gamma - 1)^2 + 4\lambda}\right],$$

and denote $\tilde{\theta}_1 := \theta_1(\delta, \tilde{\lambda})$, $\tilde{\theta}_2 := \theta_2(\delta, \tilde{\lambda})$.

Remark. From (6), it is readily verified that $\lambda\gamma\theta_1^2+(\lambda-\gamma+1)\theta_1-1=0$. Taking derivatives and noting that $\theta_2 = -d\theta_1/d\lambda$, we get $1+(\gamma-1)\theta_1-2\lambda^2\gamma\theta_1\theta_2-\lambda(\lambda-\gamma+1)\theta_2=0$. These two equations will be useful when simplifying certain formulas. Then we have the following fundamental result on the asymptotic behavior of the variance components. This is our first main result.

Theorem 2 (Variance components) *Under the previous assumptions, consider an* $n\times d$ *feature matrix* X *with i.i.d. entries of zero mean, unit variance and finite* $8+\eta$*-th moment. Take a two-layer linear neural network* $f(x) = (Wx)^\top\beta$*, with* $p \leqslant d$ *intermediate activations, and* $p \times d$ *matrix* W *of first-layer weights chosen uniformly subject to the orthogonality constraint* $WW^\top = I_p$*. Then, the variance components have the following limits as* $n, p, d \to \infty$*, with* $p/d \to \pi \in (0, 1]$ *(parametrization level),* $d/n \to \delta > 0$ *(data aspect ratio). Here* α^2 *is the signal strength,* σ^2 *is the noise level,* λ *is the regularization parameter,* θ_i *are the resolvent moments (and* $\tilde{\lambda}, \tilde{\theta}_i$ *are their adjusted versions).*

$$
\begin{aligned}
\lim_{d\to\infty} V_s &= \alpha^2[1-2\tilde{\lambda}\tilde{\theta}_1+\tilde{\lambda}^2\tilde{\theta}_2-\pi^2(1-\lambda\theta_1)^2]\\
\lim_{d\to\infty} V_l &= 0\\
\lim_{d\to\infty} V_i &= \alpha^2\pi(1-\pi)(1-\lambda\theta_1)^2\\
\lim_{d\to\infty} V_{sl} &= \sigma^2\delta(\tilde{\theta}_1-\tilde{\lambda}\tilde{\theta}_2)
\end{aligned}
$$

$$
\begin{aligned}
\lim_{d\to\infty} V_{li} &= 0\\
\lim_{d\to\infty} V_{si} &= \alpha^2[\pi(1-2\lambda\theta_1+\lambda^2\theta_2+(1-\pi)\delta(\theta_1-\lambda\theta_2))\\
&\qquad -\pi(1-\pi)(1-\lambda\theta_1)^2-1+2\tilde{\lambda}\tilde{\theta}_1-\tilde{\lambda}^2\tilde{\theta}_2]\\
\lim_{d\to\infty} V_{sli} &= \sigma^2\delta[\pi(\theta_1-\lambda\theta_2)-(\tilde{\theta}_1-\tilde{\lambda}\tilde{\theta}_2)].
\end{aligned}
$$

Remark. Theorem 2 shows that the label noise does not contribute to the variance via a main effect (because $\lim_{d\to\infty} V_l = 0$), but instead through its interaction effects with the sample and initialization (the terms V_{sl}, V_{sli}). These can be arbitrarily large if we let $\sigma \to \infty$. In our simulations, we assume a reasonable amount of label noise, e.g., $\sigma = 0.3\alpha$.

2.2.2 Ordered Variance Decompositions

Using Theorem 2, we can calculate all six variance decompositions corresponding to the ordering of the sources of randomness. Namely, suppose that we decompose the variance in the order (a, b, c), where $\{a, b, c\} = \{s, l, i\}$, i.e., we calculate the following three terms:

$$
\begin{aligned}
\Sigma^a_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{a,b,c}[\hat{f}(x)-\mathbb{E}_a\hat{f}(x)]^2\\
\Sigma^b_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{b,c}[\mathbb{E}_a\hat{f}(x)-\mathbb{E}_{a,b}\hat{f}(x)]^2\\
\Sigma^c_{abc} &:= \mathbb{E}_{\theta,x}\mathbb{E}_{c}[\mathbb{E}_{a,b}\hat{f}(x)-\mathbb{E}_{a,b,c}\hat{f}(x)]^2.
\end{aligned}
$$

We can interpret these as follows: (1) Σ^a_{abc} is all the variance related to a. (2) Σ^b_{abc} is all the variance related to b after subtracting all the variance related to a in the total variance. (3) Σ^c_{abc} is the part of the variance that depends only on c. Then, simple calculations show

$$\begin{aligned}\Sigma^a_{abc} &= V_a + V_{ab} + V_{ac} + V_{abc}\\ \Sigma^b_{abc} &= V_{bc} + V_b\\ \Sigma^c_{abc} &= V_c.\end{aligned}$$

In previous work, d'Ascoli et al. (2020) considered the decomposition in the order label - initialization - sample (l, i, s), which becomes in our case (canceling the terms that vanish)

$$\begin{aligned}\Sigma^l_{lis} &= V_{ls} + V_{lsi} + V_l + V_{li} = V_{ls} + V_{lsi}\\ \Sigma^i_{lis} &= V_i + V_{si}\\ \Sigma^s_{lis} &= V_s.\end{aligned}$$

d'Ascoli et al. (2020) argued that the label and initialization noise dominate the variance. However, different decomposition orders can lead to qualitatively different results. We take the decomposition order (l, s, i) as an example. Roughly speaking, in this decomposition, Σ^l_{lsi} and Σ^s_{lsi} can be interpreted as the variance introduced by the data set given a fixed initialization (and model), and Σ^i_{lsi} is the variance of the initialization alone. Figure 2 (left, middle) shows the results under these two different decomposition orders.

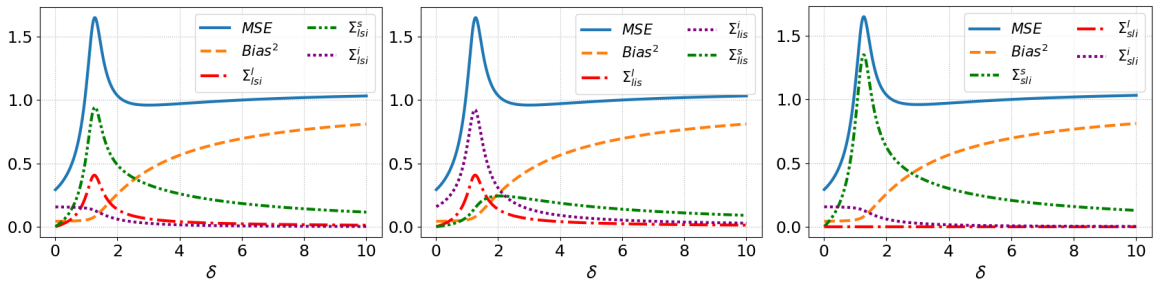


Figure 2: Bias-variance decompositions in three orders. **Left**: Decomposition order: label, sample, initialization (l, s, i); dominated by Σ^s_{lsi} ("samples"). **Middle**: Decomposition order: label, initialization, sample (l, i, s); dominated by Σ^i_{lis} ("initialization"). **Right**: Decomposition order: sample, label, initialization (s, l, i); dominated by Σ^s_{sli} ("samples"). Parameters: signal strength $\alpha = 1$, noise level $\sigma = 0.3$, regularization parameter $\lambda = 0.01$, parametrization level $\pi = 0.8$.

Comparing the left (lsi) and middle (lis) panels of Figure 2, we can see that different decomposition orders indeed lead to qualitatively different results. When $\delta < 2$, in lsi, the variance with respect to samples dominates the total variance, while in lis the variance with respect to initialization dominates. Therefore, to have a better understanding of the limiting MSE of ridge models, it is preferable to decompose the variance in a symmetric and more systematic way using the variance components. In fact, the discrepancy between these two decompositions is due to the term V_{si}, which dominates the variance (as discussed later) and is contained in both Σ^s_{lsi} and Σ^i_{lis}. By identifying this key term V_{si}, which has

not appeared in prior work, we are able to pinpoint the specific reason why the variance is large, namely the interaction between the variation in samples and initialization. Moreover, the variance with respect to samples is even larger in Figure 2 (right), for the (sli) order, since Σ^s_{sli} also contains the interaction effect between the randomness of samples and labels.

2.2.3 A Special Ordered Variance Decomposition

Next, we consider a special case of the variance decomposition, in the order of label-sample-initialization. This is advantageous because it leads to particularly simple formulas, whose monotonicity properties are particularly tractable, as explained below. The variance decomposes as $\mathbf{Var}(\lambda) = \Sigma_{\text{label}} + \Sigma_{\text{sample}} + \Sigma_{\text{init}}$, where

$$\begin{aligned}
\Sigma_{\text{label}} &:= \Sigma^l_{lsi} = \mathbb{E}_{\theta,x}\mathbb{E}_{X,W,\mathcal{E}}(f_{\lambda,\mathcal{T},W}(x) - \mathbb{E}_{\mathcal{E}} f_{\lambda,\mathcal{T},W}(x))^2 \\
\Sigma_{\text{sample}} &:= \Sigma^s_{lsi} = \mathbb{E}_{\theta,x}\mathbb{E}_{X,W}(\mathbb{E}_{\mathcal{E}} f_{\lambda,\mathcal{T},W}(x) - \mathbb{E}_{X,\mathcal{E}} f_{\lambda,\mathcal{T},W}(x))^2 \\
\Sigma_{\text{init}} &:= \Sigma^i_{lsi} = \mathbb{E}_{\theta,x}\mathbb{E}_{W}(\mathbb{E}_{X,\mathcal{E}} f_{\lambda,\mathcal{T},W}(x) - \mathbb{E}_{W,X,\mathcal{E}} f_{\lambda,\mathcal{T},W}(x))^2.
\end{aligned}$$

As above, intuitively Σ_{label} is all variance related to the label noise, Σ_{sample} is the variance related to the samples after subtracting the variance related to the label noise, and Σ_{init} is the variance due only to the initialization. The decomposition "label-init-sample" was studied in (d'Ascoli et al., 2020). Going beyond what was previously known, we can get explicit expressions not only for the variances (using the ANOVA decomposition), but also for the bias, and moreover prove some monotonicity and unimodality properties of these quantities when the ridge parameter λ is optimal.

Yang et al. (2020) observed empirically that the variance when fitting certain neural networks can often be unimodal, and proved this for a 2-layer net similar to our setting, with Gaussian initialization W and assuming $n/d \to \infty$. However, they left open the question of understanding this phenomenon more broadly, writing that "The main unexplained mystery is the unimodality of the variance". Our result sheds further light on this problem.

Corollary 3 (Bias & variance in Two-Layer Orthogonal Net—special ordering) *Under the assumptions from Theorem 2, we have the limits*

$$\lim_{d\to\infty} \mathbf{Bias}^2(\lambda) = \alpha^2(1-\pi+\lambda\pi\theta_1)^2, \tag{8}$$

$$\begin{aligned}
\lim_{d\to\infty} \mathbf{Var}(\lambda) =& \alpha^2\pi[1-\pi+(\pi-1)(2\lambda-\delta)\theta_1 - \pi\lambda^2\theta_1^2 + \lambda(\lambda-\delta+\pi\delta)\theta_2] + \\
& \sigma^2\pi\delta(\theta_1 - \lambda\theta_2).
\end{aligned} \tag{9}$$

More specifically,

$$\lim_{d\to\infty} \Sigma_{\text{label}}(\lambda) = \sigma^2\pi\delta(\theta_1 - \lambda\theta_2), \tag{10}$$

$$\lim_{d\to\infty} \Sigma_{\text{sample}}(\lambda) = \alpha^2\pi\left[-\lambda^2\theta_1^2 + \lambda^2\theta_2 + (1-\pi)\delta(\theta_1 - \lambda\theta_2)\right], \tag{11}$$

$$\lim_{d\to\infty} \Sigma_{\text{init}}(\lambda) = \alpha^2\pi(1-\pi)(1-\lambda\theta_1)^2, \tag{12}$$

where $\theta_i := \theta_i(\pi\delta, \lambda)$, $i = 1, 2$. *Therefore*

$$\lim_{d\to\infty} \mathbf{MSE}(\lambda) = \alpha^2\left\{1-\pi+\pi\delta\left(1-\pi+\sigma^2/\alpha^2\right)\theta_1 + \left[\lambda - \delta\left(1-\pi+\sigma^2/\alpha^2\right)\right]\lambda\pi\theta_2\right\} + \sigma^2. \tag{13}$$

For any fixed δ, π, the asymptotic MSE has a unique minimum at $\lambda^* := \delta(1-\pi+\sigma^2/\alpha^2)$. (except when $\pi = 1, \sigma = 0$).

Remark. Except for the simple formula for the bias, theorem 3 is direct corollary of theorem 2, since $\Sigma_{\text{label}}, \Sigma_{\text{sample}}, \Sigma_{\text{init}}$ and Var are all sums of several variance components.

Almost sure results over random true parameter. Above, we provide average-case results over the true parameters $\theta \sim \mathcal{N}(0, \alpha^2 I_d/d)$. With additional work, we show below a corresponding almost sure result. For the next result, we assume that X has iid Gaussian entries.

Theorem 4 (Almost sure result over true parameter θ) *For each triple (p_d, d, n_d), suppose that the true parameter θ is a sample drawn from $\mathcal{N}(0, \alpha^2 I_d/d)$. Suppose in addition that each entry of X is iid standard normal. Then as $d \to \infty$, Theorems 2 and 3 still hold almost surely over the selection of θ.*

2.2.4 Optimal Regularization Parameter; Monotonicity and Unimodality

In this section, we present some theoretical results about the risks when using an optimal regularization parameter. Moreover, we study the monotonicity and unimodality of certain variance components in that setting.

We can find explicit formulas for the asymptotic bias and variance at the optimal λ^*, by plugging in the expressions of λ^*, θ_1 into equations (8), (9):

$$\begin{aligned}\lim_{d\to\infty} \mathbf{Bias}^2(\lambda^*) &= \alpha^2(1-\pi+\lambda^*\pi\theta_1)^2 \\ &= \alpha^2\left(\frac{\delta(1-\sigma^2/\alpha^2)-1+\sqrt{(\delta(1+\sigma^2/\alpha^2)+1)^2-4\gamma}}{2\delta}\right)^2. \end{aligned} \tag{14}$$

$$\lim_{d\to\infty} \mathbf{Var}(\lambda^*) = -\sigma^2\pi + (\alpha^2+\sigma^2\delta)\left(\frac{\delta(2\pi-1-\sigma^2/\alpha^2)-1+\sqrt{(\delta(1+\sigma^2/\alpha^2)+1)^2-4\gamma}}{2\delta^2}\right).$$

$$\begin{aligned}\lim_{d\to\infty} \mathbf{MSE}(\lambda^*) &= \alpha^2\left[1-\pi+\lambda^*\pi\theta_1\right]+\sigma^2 \\ &= \alpha^2\left(\frac{\delta(1-\sigma^2/\alpha^2)-1+\sqrt{(\delta(1+\sigma^2/\alpha^2)+1)^2-4\gamma}}{2\delta}\right)+\sigma^2. \end{aligned} \tag{15}$$

From Theorem 3, we know that the optimal ridge penalty is $\lambda^* = \delta(1-\pi+\sigma^2/\alpha^2)$. Thus, by plugging the expression of λ^* into (8)—(13), we are able to study the properties of the MSE, bias and variance components at the optimal λ^* as functions of π, δ. Our results are summarized in Table 1. See Figure 3 for illustration.

As for the monotonicity and unimodality properties, we have the following statement, where the properties are summarized in Table 1 for clarity.

Theorem 5 (Monotonicity and unimodality) *Under the assumptions above, the MSE, Bias, and components of the sequential variance decomposition in the $l-i-s$ order have the monotonicity and unimodality properties summarized in Table 1. For instance, the MSE is non-increasing as a function of the parameterization level π, while holding δ fixed.*

Function \ Variable	parametrization $\pi = \lim p/d$	aspect ratio $\delta = \lim d/n$
MSE	$\searrow$	$\nearrow$
Bias^2	$\searrow$	$\nearrow$
Var	$\delta < 2\alpha^2/(\alpha^2+2\sigma^2)$: $\wedge$, max at $[2+\delta(1+2\sigma^2/\alpha^2)]/4$. $\delta \geq 2\alpha^2/(\alpha^2+2\sigma^2)$: $\nearrow$.	$\pi \leq 0.5 : \searrow$. $\pi > 0.5$: $\wedge$, max at $2(2\pi-1)/[1+2\sigma^2/\alpha^2]$.
Σ_{label}	$\nearrow$	$\wedge$: max at $\alpha^2/(\alpha^2+\sigma^2)$
Σ_{init}	$\wedge$	$\searrow$
Σ_{sample}	conjecture: $\nearrow$ or $\wedge$	conjecture: $\wedge$

Table 1: Monotonicity properties of various components of the risk at the optimal λ^*, as a function of π and δ, while holding all other parameters fixed. $\nearrow$: non-decreasing. $\searrow$: non-increasing. $\wedge$: unimodal. Thus, e.g., the MSE is non-increasing as a function of the parameterization level π, while holding δ fixed.

We provide some observations below.

Consistency with prior work. The MSE result is consistent with optimal regularization mitigating double descent, which was shown in finite samples in a certain two-layer Gaussian model in Nakkiran et al. (2021). However, our result holds for more general distributions of data matrices with arbitrary iid entries, while only proven asymptotically.

Variance as a function of δ. For fixed parametrization level $\pi = \lim p/d$, as a function of the "dimensions-per-sample" parameter $\delta = \lim d/n$, the variance is monotonically decreasing when $\pi < 0.5$, and unimodal with a peak at $2(2\pi-1)/[1+2\sigma^2/\alpha^2]]$ when $\pi \geq 0.5$. This prompts the question why $\pi = 1/2$ is special? The special role of this value was also noted in Yang et al. (2020). Recall that d is the original dimension, while p is the number of features in the intermediate layer, and $\pi = \lim p/d$. While the role of $\pi = 1/2$ does not seem straightforward to understand, qualitatively for large π we keep a lot of features in the inner layer. This is close to a "well-specified" model. Thus, when we increase the size of the data set (and thus decrease $\delta = \lim d/n$), it is reasonable that the variance decreases. In contrast, regardless of π, when we severely decrease the size of the data set (and thus increase $\delta = \lim d/n$), the optimal ridge estimator will regularize more strongly, and thus it is possible that its variance may decrease (which is what we indeed observe).

Variance as a function of π. For small δ, the variance is unimodal with respect to π. A possible heuristic is as follows. Recall that $\pi = \lim p/d$ (d is data dimension, p is number of features in inner layer) denotes the amount of "parametrization" we allow. When $\pi \approx 0$, the number of features in the inner layer is very small, which effectively corresponds to a "low signal strength" problem (see also our added noise interpretation below). The optimal ridge estimator thus employs strong regularization, and acts like a constant estimator, thus the variance is almost zero. When $\pi = 1$, we are using the correct number of features to estimate θ, thus the variance is also small. The variance is zero when $\sigma = 0$, and we can plot it when $\sigma > 0$ (see Figure 3). The above reasoning also suggests the variance may be

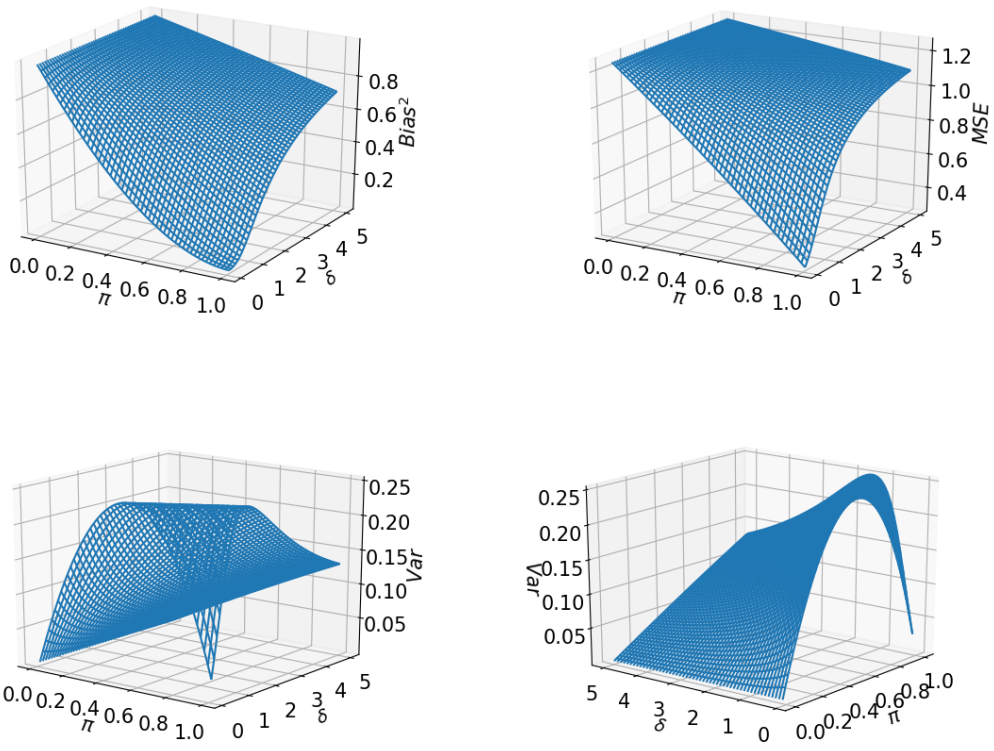


Figure 3: Perspective plots of the performance characteristics. Top row: **Bias**^{2} (left), MSE (right). Bottom row: variance, from two perspectives. As functions of π, δ, at the optimal $\lambda^* = \delta(1 - \pi + \sigma^2/\alpha^2)$, when $\alpha = 1$ and $\sigma = 0.5$.

larger for intermediate values of π. Thus, the unimodality of the variance with respect to π is perhaps reasonable.

Beyond our results on Σ_{label} and Σ_{init}, we conjecture based on numerical experiments that Σ_{sample} is unimodal as a function of δ and can be either unimodal or monotone as a function of π. However, this appears more challenging to establish.

Comparison with Rocks and Mehta (2020). In their paper, they suggest that the training process W should be separated from the sampling of the training data X, ε when studying the variance. Thus, they calculate the variance by fixing θ, W, computing conditional variances (due to X, ε), then taking expectation over θ, W. This can, in principle, still be recovered from our general framework, if we look at the variance components conditioned on θ, W. In constrast, we consider the randomness arising from all components together. Our approach allows us to study some problems that do not easily fall within the scope of the conditional approach. For instance, we can provide intuition for why ensembling works; namely that it can reduce the interaction effect V_{si}.

Multiple descent. It has been argued that other possible shapes of the test error, such as multiple descents, can arise. Liang et al. (2020) study kernel regression under the

limiting regimes $d \sim n^c, c \in (0,1)$. They provide an upper bound for the MSE and show its multiple descent behavior as c increases. Adlam and Pennington (2020a) study the neural tangent kernel under the limiting regimes $p \sim n^c$, $c = 1, 2$ and observe that the MSE has a triple descent shape as a function of p. To conclude, the MSE may exhibit multiple descent when considering different asymptotic regimes. However, since we only consider the proportional limit where $p/d \to \pi, d/n \to \delta$, we have not found evidence of multiple descent in our setting.

Remark: Fully linear regression. The monotonicity of the MSE and bias, and the unimodality of variance at the optimal λ^* also appear in the simpler (one-layer) linear setting. Namely, we consider the usual linear model $Y = X\theta + \mathcal{E}$, where $X \in \mathbb{R}^{n\times d}$ is the data matrix and $Y \in \mathbb{R}^n$ is the response. We fit a linear regression of Y on X, which can be seen as a special case of our two layer setting with $W = I_d$ (and $d = p$). We use the same assumptions (except the assumptions on W) and notations as in the two layer setting. In particular, we assume $n, p \to \infty$ with $p/n \to \gamma > 0$, where the aspect ratio γ is now a measure of the parametrization level. We have the following result.

Proposition 6 (Properties of the limiting MSE, bias & variance in linear setting) *Under the same assumptions as in the two-layer setting, the limiting characteristics of optimally tuned ridge regression ($\lambda^* = \gamma/\alpha^2$) have the following properties as a function of the degree of parameterization γ:*

1. $\mathbf{MSE}(\gamma)$ *is increasing as a function of* γ.
2. $\mathbf{Bias}^2(\gamma)$ *is increasing as a function of* γ.
3. $\mathbf{Var}(\gamma)$ *is unimodal as a function of* γ*, with maximum at* $\gamma = \alpha^2/(\alpha^2+1)$.
4. *At the maximum, the bias equals the variance:* $\mathbf{Bias}^2[\alpha^2/(\alpha^2+1)] = \mathbf{Var}[\alpha^2/(\alpha^2+1)]$.

Figure 6 in Liu and Dobriban (2020) shows the MSE, variance and bias at optimal λ^*. However, that work only studied it visually, and not theoretically. The result on the MSE has appeared before as Proposition 6 of Dobriban and Sheng (2020), in a different context. However, the results on the bias and variance have not been considered in that work.

2.3 Further Properties of the Bias, Variance and MSE

In this section, we report some further properties of the bias, variances and MSE, including but not limited to the optimal ridge parameter setting.

2.3.1 Relation Between MSE and Bias at Optimum

We present a somewhat surprising relation between MSE and bias at the optimal $\lambda^* = \lambda^*(\delta, \pi, \alpha^2, \sigma^2)$: Let $\mathbf{Bias}^2 := \lim_{d\to\infty} \mathbf{Bias}^2(\lambda^*)$, $\mathbf{Bias} = |\sqrt{\mathbf{Bias}^2}|$, and denote $\mathbf{MSE}:= \lim_{d\to\infty}\mathbf{MSE}(\lambda^*)$. In general for all λ, we have that $\mathbf{MSE}(\lambda) = \mathbf{Bias}^2(\lambda) + \mathbf{Var}(\lambda) + \sigma^2$. Also, in prominent problems such as in non-parametric statistics, optimal rates are achieved by balancing bias and variance (e.g., Ibragimov and Has′ Minskii, 2013, etc). Thus we are interested to see if the bias and variance are also balanced at the optimal λ in our case.

However, based on the explicit expressions above, the bias and variance are balanced via the signal strength α^2 via

$$\textbf{Var} = \textbf{Bias} \cdot (\alpha - \textbf{Bias}).$$

This holds for any π, δ and α. Thus, the MSE and bias are linked in a nontrivial way at the optimal λ. We see that the optimal squared bias and variance are in general not equal at the optimum. Instead, we have the above relation, which also balances the bias with the *signal strength* α^2. We think that this explicit relation is remarkable.

2.3.2 Fixed Regularization Parameter

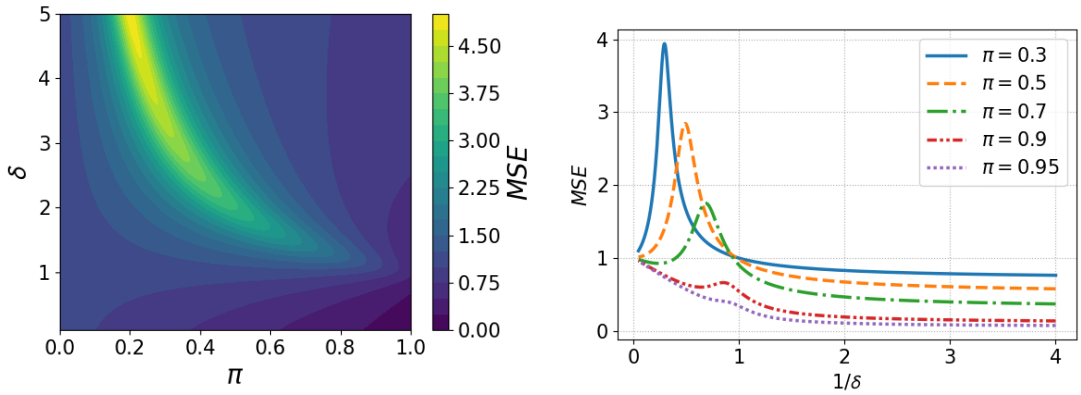


Figure 4: **Left**: Asymptotic MSE of ridge models when $\lambda = 0.01$, $\sigma = 0.3$, $\alpha = 1$. **Right**: Asymptotic MSE as function of $1/\delta$ when $\lambda = 0.01$, $\sigma = 0.3$, $\alpha = 1$. (Note: this figure is plotted as a function of $1/\delta$, instead of δ as before. Increasing $1/\delta$ is equivalent to increasing the number of samples n.)

From Figure 3 and Theorem 5 above, we can see that the MSE is monotone decreasing with respect to the parametrization level $\pi = \lim p/d$ if we choose the optimal λ^*. This is consistent with "double descent being mitigated", as in the results of Nakkiran et al. (2021) for a different problem.

Here we provide additional results for a *suboptimal* choice of λ. Specifically, we consider the simplest case when λ is fixed across problem sizes. In contrast, we find that double descent is *not* mitigated, and may occur when we use a small regularization parameter (also referred to as the ridgeless limit). In Figure 4 (left), we fix $\lambda = 0.01$ and plot a heatmap of the asymptotic MSE as function of the two variables $\pi = \lim p/d$ and $\delta = \lim d/n$. Clearly, the MSE is in general not monotone with respect to π or δ. Note the peak in the MSE around the curve $\gamma = \delta\pi = 1$, or equivalently $\delta = 1/\pi$. This corresponds to the "interpolation threshold" where $\lim p/n = 1$, and the number of learned parameters p is close to the number of samples n. Thus, we fit just enough parameters to interpolate the data.

Besides, we see in Figure 4 (right) that double descent (which we interpret as a change of monotonicity, or a peak in the risk curve) with respect to $1/\delta = \lim n/d$ occurs when π is

suitably large, e.g., $\pi = 0.9$, while the MSE is unimodal when π is small, e.g., $\pi = 0.5$. The intuition is, as in many previous works on double descent (e.g., LeCun et al., 1991; Hastie et al., 2019), that the suboptimal regularization can lead to a somewhat ill-conditioned problem, which increases the error (see also section 2.3.4 for more explanation). Thus, we see that here as in prior works, using a suboptimal penalty λ may lead to non-monotone MSE.

Moreover, we can obtain some quantitative results about the bias and variance with fixed values of the regularization parameter λ.

Theorem 7 (Bias and variance of ridge models given a fixed λ) *Under the assumptions in our two layer setting, we have*

1. *For any fixed $\lambda > 0$, $\lim_{d\to\infty} \mathbf{Bias}^2(\lambda)$ is monotonically decreasing as a function of π and is monotonically increasing as a function of δ.*

2. *$\lim_{\lambda\to 0} \lim_{d\to\infty} \mathbf{Var}(\lambda) = \infty$ on the curve $\delta = 1/\pi$ (the interpolation threshold where $\lim p/d = 1$). More specifically, when $\lambda \to 0$, V_{si}, V_{sli} goes to infinity while other variance components converge to some finite limits on the curve $\delta = 1/\pi$.*

The first part implies that more samples or a larger degree of parameterization can always reduce the prediction bias, which is consistent with our intuition that larger models can, in principle, approximate any function better.

For the variance components, it is natural to expect that some interaction exists, because even the expressions W, X in the prediction function $f(X) = (WX)^\top \beta$ interact non-additively. But we do not fully understand why the interaction terms V_{si}, V_{sli} are large. This can be viewed as a surprising discovery of our paper. One somewhat tautological perspective is that the interaction terms are the part of variance that are most affected by "under-regularization". For instance, the main effect V_i comes from the randomness of initializations. Thus, one has to average—or ensemble—over the choices of initialization W to reduce V_i. However, for V_{si} and V_{sli}, both ensembling *and* optimally tuned ridge regularization can reduce their values significantly. Thus, these components seem to be more affected by the "under-regularization" due to using a sub-optimal ridge parameter. However, this is still a somewhat circular explanation, because the entire reason that they diverge is that they are sensitive to under-regularization.

For any fixed $\lambda > 0$, we conjecture based on numerical results that $\lim_{d\to\infty} \mathbf{Var}(\lambda)$ is unimodal as both a function of π and a function of δ. This appears to be more challenging to show. Here the unimodality would be mainly due to being close to the interpolation threshold $p/n \approx 1$ for $\delta \approx 1/\pi$, which leads to ill-conditioned feature matrices and a large risk.

2.3.3 Added Noise Interpretation

The random projection step in the initialization can be interpreted as creating additional noise. Thus, we can find a ridge model without the projection step (i.e. without the first layer) with larger training set label noise σ'^2 and the same test point label noise σ^2 that has the same asymptotic bias, variance, and MSE as the model with random projection step (i.e. with the first layer). To obtain the "effective noise" level σ'^2, in equation (14), (15),

let us equate the formula determining $\lim_{d\to\infty} \mathbf{MSE}(\lambda^*) = \alpha^2[1-\pi+\lambda^*\pi\theta_1]+\sigma^2$ for two sets of parameters $\alpha^2, \sigma^2, \pi, \delta$ and $\alpha^2, \sigma'^2, \pi = 1, \delta$. This leads to the equation

$$1-\pi+\lambda^*\pi\theta_1 = \lambda'^*\theta_1'.$$

After simple calculation, we obtain

$$\sigma'^2 = \sigma^2+\Delta := \sigma^2+\alpha^2(1-\pi)\frac{\delta(1+\sigma^2/\alpha^2)+1+\sqrt{(\delta(1+\sigma^2/\alpha^2)+1)^2-4\gamma}}{2\gamma}.$$

Note that Variance = MSE − Bias^2 − σ^2, hence the random projection model has the same asymptotic bias, variance and MSE as the ordinary ridge model with additional training set label noise Δ. However, the variance components are specific to the two-layer case, and do not carry over to the one-layer case.

2.3.4 Understanding the Effect of the Optimal Ridge Penalty

In this section, we provide some intuitions for why unimodality and the double descent shape appears in the MSE of ridge models when using a fixed small penalty λ (close to the ridgeless limit), and how the optimal λ^* helps eliminate the non-monotonicity of the MSE. We illustrate this with numerical results.

To qualitatively understand the effect of the optimal penalty λ^*, we plot the variance components, variance, bias and the MSE under two different scenarios. In the first scenario, we use the optimal penaly λ^* for all ridge models (see Figure 5). It is readily verified that V_s and V_i contribute to a large portion of the variance, while the contributions of V_{si} and V_{sli} are relatively small.

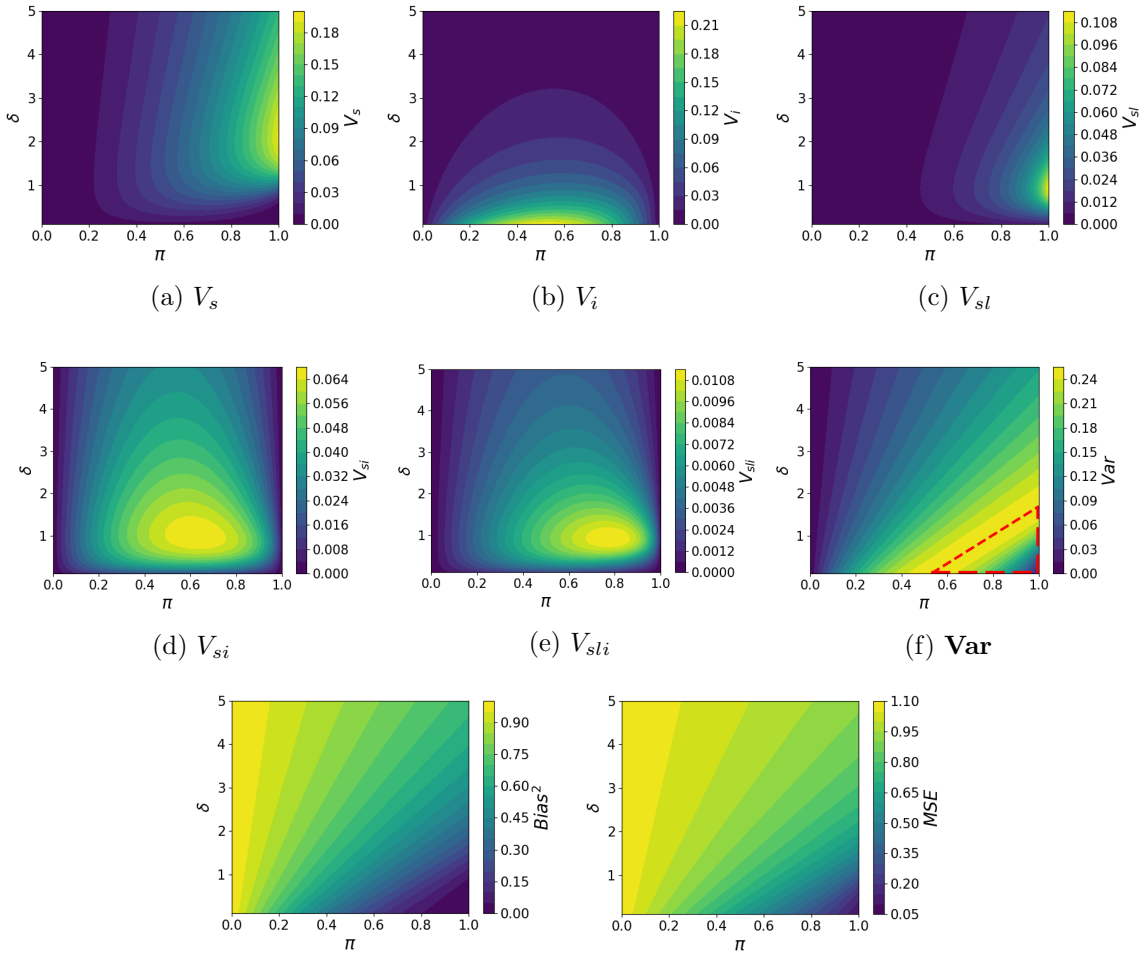
In the second scenario, we choose $\lambda = 0.01$ for all ridge models. From Figure 6, we see that, perhaps surprisingly, it is the *interaction term* V_{si} between sample and initialization that dominates the variance. In particular, we think that it is surprising that this interaction term can be larger than the main effects V_s and V_i of sample and initialization. Also, V_{si} and V_{sli} lead to the modes of the variance on the curve $\delta = 1/\pi$ (the interpolation threshold).

Comparing Figure 5 and 6, we can see that V_s, V_i and V_{sl} are almost on the same scale in the two scenarios. However, V_{si} and V_{sli} are much larger when $\lambda = 0.01$ than when $\lambda = \lambda^*$. These two terms are the main reason why the variance is significantly larger when $\lambda = 0.01$. Moreover, Figure 5g and 6g show that the bias is even relatively smaller when we use $\lambda = 0.01$ instead of the optimal λ^*. Intuitively, the reason is that the optimal regularization parameter is large, to achieve a better bias-variance tradeoff, and thus makes the bias slightly larger while decreasing the variance a great amount.

Therefore, we may conclude that, under a reasonable assumption on the label noise (e.g., $\sigma = 0.3\alpha$ here),

1. Using a fixed small penalty λ for all ridge models can lead to unimodality/double descent shape in the MSE. The modes of the MSE as a function of δ are close to the interpolation limit curve $\delta = 1/\pi$.

2. The unimodality/double descent shape of the MSE given a fixed small penalty λ is due to the variance. The bias is typically smaller when using a fixed small penalty λ



(a) V_s (b) V_i (c) V_{sl}

(d) V_{si} (e) V_{sli} (f) **Var**

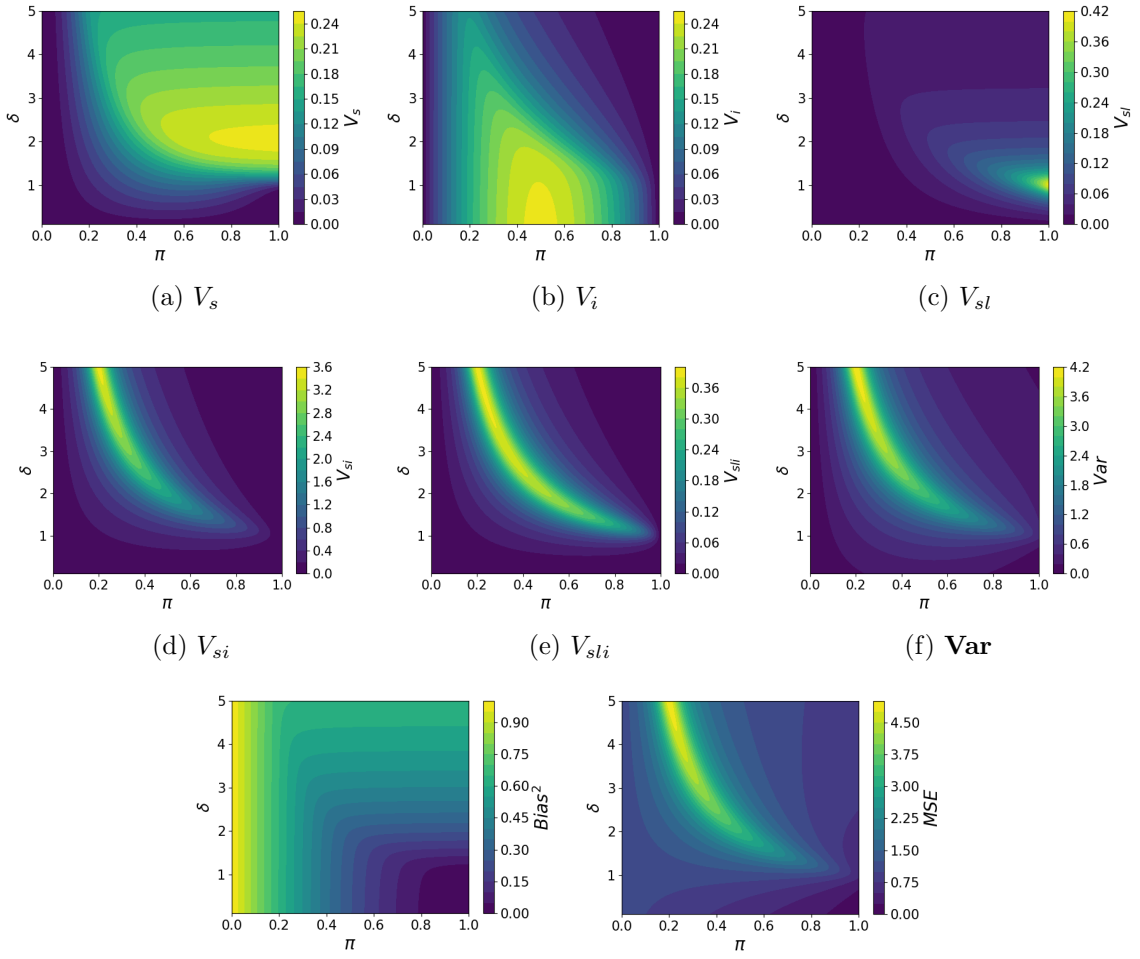
(g) **Bias**^{2} (h) **MSE**

Figure 5: Heatmaps of the performance characteristics for the optimal regularization parameter $\lambda = \lambda^*$. Variance components, variance, bias and the MSE as functions of π and δ when $\alpha = 1, \sigma = 0.3$. ($\mathbf{Var} = V_s + V_i + V_{sl} + V_{si} + V_{sli}$. $\mathbf{MSE} = \mathbf{Bias}^2 + \mathbf{Var} + \sigma^2$.)

instead of the optimal penalty λ^*. As mentioned, this is because the bias and variance are balanced out for the optimal λ^*, and thus we can increase the bias a bit, while significantly decreasing the variance.

3. Compared with choosing the same small ridge penalty for all models, through using the optimal penalty λ^*, one can reduce the variance significantly, especially along the interpolation threshold curve. The unimodality/double decent shape of the MSE will vanish as a result; but the variance itself may still be unimodal.

4. Using the optimal penalty for all ridge models reduces the variance mainly by reducing the interaction component V_{si}. This component is large in an absolute sense, and thus a reduction has a significant effect. The component V_{sli} is also reduced in a relative



(a) V_s (b) V_i (c) V_{sl}

(d) V_{si} (e) V_{sli} (f) **Var**

(g) **Bias**^{2} (h) **MSE**

Figure 6: Heatmaps of the performance characteristics for a fixed parameter $\lambda = 0.01$. Variance components, variance, bias and the MSE as functions of π and δ when $\alpha = 1, \sigma = 0.3$. ($\mathbf{Var} = V_s + V_i + V_{sl} + V_{si} + V_{sli}$. $\mathbf{MSE} = \mathbf{Bias}^2 + \mathbf{Var} + \sigma^2$.)

sense; however, because it is of a smaller magnitude, this reduction has a more limited effect.

5. There is a special region where the bias and variance (for the optimal λ^*) change in the same direction, in the sense that increasing the parametrization $\pi = \lim p/d$ or decreasing the aspect ratio $\delta = \lim d/n$ decrease *both* the bias and the variance. See Figures 5f and 5g.

   This special region is characterized by the "triangle" $0 < \pi \leqslant 1$, $\delta > 0$, with $\delta \leqslant 2(2\pi - 1)/[1 + 2\sigma^2/\alpha^2]$. In finite samples, this is approximated by the inequality $d/n \leqslant 2(2p/d - 1)/[1 + 2\sigma^2/\alpha^2]$ between the sample size n, data dimension d and the number of parameters p. This can be interpreted as saying—for instance—that

the parameter dimension p should be large enough. Thus, in that setting, with more parametrization we can get simultaneously better bias and better variance. In a sense, this can indeed be viewed as a blessing of overparametrization.

6. There is a "hotspot" around $\pi = 1/2$, where in V_i, and V_{si} both take large values (see Figures 5b and 5d). The variance due to initialization is large for intermediate values of the projection dimension p. Roughly speaking, one can consider an analogy with Bernoulli random variables, which have large variance for intermediate values of the success probability.

7. For fixed λ, when $\delta < 1$ ($d < n$), numerical experiments show that the MSE is decreasing as π increases, which means that more parametrization can always give us small MSE when we have enough samples. See Figure 6h. It appears that there may be no double descent for fixed λ when δ is sufficiently small; however investigating this is beyond our current scope.

Recall that, in the noiseless case, V_s can be interpreted as the variance of an ensemble estimator, and $V_{si} + V_i$ is the variance that can be reduced through ensembling. Therefore, the unimodality/double descent shape in the MSE is not intrinsic, and can be removed through regularization techniques such as ensembling, (consistent with d'Ascoli et al. 2020) or optimal ridge penalization.

2.4 Ridge is Optimal

We have obtained precise asymptotic results for optimally tuned ridge regression. However, is ridge regression optimal, or are there other methods that outperform it? In fact, we can prove that the ridge estimator is asymptotically optimal.

Theorem 8 (Ridge is optimal) *Suppose that the samples are drawn from the standard normal distribution, i.e., x and X both have i.i.d. $\mathcal{N}(0,1)$ entries. Given the projection W, projected matrix $XW^\top$ and response Y, we define the optimal regression parameter β_{opt} as the one minimizing the MSE over the posterior distribution $p(\theta|XW^\top, W, Y)$ of the parameter θ,*

$$\beta_{opt} := \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)}\mathbb{E}_{x,\varepsilon}[(Wx)^\top\beta - (x^\top\theta + \varepsilon)]^2, \tag{16}$$

where $x \sim \mathcal{N}(0, I_d)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and x, ε are independent. We will check that this can be expressed in terms of the posterior of θ as

$$\beta_{opt} = W \cdot \mathbb{E}_{p(\theta|XW^\top,W,Y)}\theta. \tag{17}$$

The optimal ridge estimator $\hat{\beta} = (n^{-1}WX^\top XW^\top + \lambda^* I_p)^{-1}WX^\top Y/n$ (Theorem 3) satisfies the almost sure convergence in the mean squared error

$$\lim_{d\to\infty} \mathbb{E}_{XW^\top,W,Y}\|\hat{\beta} - \beta_{opt}\|_2^2 = 0, \tag{18}$$

and is thus asymptotically optimal. Here $d \to \infty$ means $p, d, n \to \infty$ proportionally as in Theorem 3.

Remark. In Theorem 8, the optimal parameter β_{opt} minimizes the mean squared error over the posterior of θ given the projection W, projected matrix $XW^\top$ and response Y. From the proof of Lemmas 12, 13, we know that $\mathbb{E}\|\hat{\beta}\|^2$ converges to some positive constant as $d \to \infty$. Thus, $\hat{\beta}$ has a constant scale as $d \to \infty$, and the result that $\|\hat{\beta} - \beta_{opt}\|^2 \to 0$ of Theorem 8 shows that β_{opt} is indeed non-trivially well approximated. This result implies the asymptotic optimality of ridge regression.

In addition, if we are given the original data matrix X instead of $XW^\top$, then from the optimality of ridge regression in ordinary linear regression with Gaussian prior and noise, we have $\beta_{opt} = W(n^{-1}X^\top X + d\sigma^2 I_p/[n\alpha^2])^{-1}X^\top Y/n$ and ridge regression *over projected data* is not asymptotically optimal. However, in our two-layer model, we only exploit the information of X through $XW^\top$, thus it is reasonable to consider the situation above, in which we are only given $XW^\top$.

3. Nonlinear Activation

It is also possible to consider the bias-variance decomposition for a two-layer neural network with certain scalar nonlinear activation functions $\sigma(x)$ after the first layer. Namely, suppose that the data are generated through the same process, but instead of using a two-layer linear network, we use

$$\hat{f}(x) = \sigma(Wx)^\top \beta \tag{19}$$

as the predictor. Here $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function applied to Wx entrywise. As before, we assume $W \in \mathbb{R}^{p\times d}$ has orthonormal rows, so $p \leqslant d$, and we only train β. This can be viewed as a random features model. We apply ridge regression to estimate β, therefore our prediction function is

$$\hat{f}(x) := \sigma(Wx)^\top \hat{\beta} = \sigma(x^\top W^\top)\left(\frac{\sigma(WX^\top)\sigma(XW^\top)}{n} + \lambda I_p\right)^{-1}\frac{\sigma(WX^\top)Y}{n}. \tag{20}$$

For simplicity, we further assume that $\mathbb{E}\sigma(Z) = 0$, where $Z \sim \mathcal{N}(0,1)$ is a standard normal random variable. The results for activation functions with arbitrary mean can be obtained through similar techniques, but are much more cumbersome. This assumption does not capture the ReLU activation function $\sigma_+(x) = \max(x, 0)$, but it can handle the function $\sigma_+(x) - \mathbb{E}\sigma_+(Z)$, which only differs from the ReLU by a constant. In particular, the mean of our prediction function $\hat{f}(x)$ with the current restriction is always zero, i.e., the prediction function does not have an intercept term. In our model, the true regression function $f^*(x) = \theta^\top x$ does not have an intercept term either; thus we think that the zero-mean restriction may not be significant in the current setting.

Moreover, we suppose that there are constants $c_1, c_2 > 0$ such that σ, σ' grow at most exponentially, i.e., $|\sigma(x)|, |\sigma'(x)| \leq c_1 e^{c_2|x|}$. Define the moments

$$\mu := \mathbb{E}Z\sigma(Z), \qquad\qquad v := \mathbb{E}\sigma^2(Z), \tag{21}$$

where $Z \sim \mathcal{N}(0,1)$. Also, we suppose that the samples are drawn from the standard normal distribution $\mathcal{N}(0, I_d)$, i.e., X and x both have i.i.d. $\mathcal{N}(0,1)$ entries. As before, we can write

down the MSE, bias and variance:

$$\begin{aligned}\mathbf{MSE}(\lambda) &:= \mathbb{E}_{\theta,x,W,X,\mathcal{E}}(\hat{f}(x) - x^\top\theta)^2 + \sigma^2\\ \mathbf{Bias}^2(\lambda) &:= \mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}}\hat{f}(x) - x^\top\theta)^2\\ \mathbf{Var}(\lambda) &:= \mathbb{E}_{\theta,x,W,X,\mathcal{E}}(\hat{f}(x) - \mathbb{E}_{X,W,\mathcal{E}}\hat{f}(x))^2.\end{aligned}$$

Our main result in this section gives asymptotic formulas for their limits.

Theorem 9 (Bias-Variance Decomposition for two-layer nonlinear NN) *Under the previous assumptions (i.e., in the setting of Theorem 2), with the further assumption that the samples are drawn from $\mathcal{N}(0, I_d)$, i.e., x and X have i.i.d. $\mathcal{N}(0,1)$ entries, we have the following limits for the bias, variance, and mean squared error. Recall that we have an $n \times d$ feature matrix X and a two-layer nonlinear neural network $f(x) = \sigma(Wx)^\top\beta$, with p intermediate activations, and $p \times d$ orthogonal matrix W of first-layer weights with $WW^\top = I_p$. Here $n, p, d \to \infty$ and $p/d \to \pi \in (0, 1]$ (parametrization level), $d/n \to \delta > 0$ (data aspect ratio), with α^2 the signal strength, σ^2 the noise level, λ the regularization parameter, θ_i the resolvent moments, and μ, v the Gaussian moments of the activation function σ from* (21)*. Then*

$$\begin{aligned}\lim_{d\to\infty}\mathbf{MSE}(\lambda) = \alpha^2\pi\Bigg[&\frac{1}{\pi} - 1 + \delta(1-\pi)\theta_1 + \frac{\lambda}{v}\left(\frac{\lambda\mu^2}{v^2} - \delta(1-\pi)\right)\theta_2\\ &+(v-\mu^2)\left(\frac{\gamma}{v}\theta_1 + \frac{1}{v} - \frac{\lambda\gamma}{v^2}\theta_2\right)\Bigg] + \sigma^2\gamma\left(\theta_1 - \frac{\lambda}{v}\theta_2\right) + \sigma^2, \qquad (22)\end{aligned}$$

$$\lim_{d\to\infty}\mathbf{Bias}^2(\lambda) = \alpha^2\left[\pi\frac{\mu^2}{v}\left(1 - \frac{\lambda}{v}\theta_1\right) - 1\right]^2, \qquad (23)$$

$$\begin{aligned}\lim_{d\to\infty}\mathbf{Var}(\lambda) = \alpha^2\pi\Bigg[&\frac{2\mu^2}{v} - 1 + \left(-\frac{2\lambda\mu^2}{v^2} + \delta(1-\pi)\right)\theta_1 + \frac{\lambda}{v}\left(\frac{\lambda\mu^2}{v^2} - \delta(1-\pi)\right)\theta_2\\ &-\frac{\pi\mu^4}{v^2}\left(1 - \frac{\lambda}{v}\theta_1\right)^2 + (v-\mu^2)\left(\frac{\gamma}{v}\theta_1 + \frac{1}{v} - \frac{\lambda\gamma}{v^2}\theta_2\right)\Bigg] + \sigma^2\gamma\left(\theta_1 - \frac{\lambda}{v}\theta_2\right), \qquad (24)\end{aligned}$$

where $\theta_1 := \theta_1(\gamma, \lambda/v)$, $\theta_2 := \theta_2(\gamma, \lambda/v)$, $\gamma = \pi\delta$. Similar to the linear case, the limiting MSE has a unique minimum at $\lambda^* := \frac{v^2}{\mu^2}\left[\delta(1 - \pi + \sigma^2/\alpha^2) + \frac{(v-\mu^2)\gamma}{v}\right]$.

Remarks. (1). When expanding the function $\sigma(x)$ in the Hermite polynomial basis, μ is the coefficient of the second basis function x, and $\sqrt{v}$ is $\sigma(x)$'s norm in the Hilbert space. Thus $v \geq \mu^2$ and the equality holds iff $\sigma(x) = kx$. (2). When $\sigma(x) = kx$ (i.e. $v = \mu^2$), the results in theorem 9 reduce to those in theorem 3.

Also, we have monotonicity properties similar to in the linear case (Table 2):

Theorem 10 (Monotonicity and unimodality for non-linear net at optimal λ^*) *Under the assumptions from Theorem 9, for the optimal $\lambda = \lambda^*$, the MSE, Bias, and variance have the monotonicity and unimodality properties summarized in Table 2. The MSE and bias are decreasing as a function of the parametrization level π, and increasing as a function of the data aspect ratio δ. The variance is either monotone or unimodal depending on the setting.*

Function \ Variable	parametrization $\pi = \lim p/d$	aspect ratio $\delta = \lim d/n$
MSE	$\searrow$	$\nearrow$
Bias^2	$\searrow$	$\nearrow$
Var	$\delta < 2\frac{\mu^2}{v}\left(2\frac{\mu^2}{v}-1\right)/\left(1+2\sigma^2/\alpha^2\right)$: $\wedge$, max at $\frac{v}{\mu^2}\left[2+\frac{\delta v}{\mu^2}\left(1+\frac{2\sigma^2}{\alpha^2}\right)\right]/4$. $\delta \geq 2\frac{\mu^2}{v}\left(2\frac{\mu^2}{v}-1\right)/\left(1+2\sigma^2/\alpha^2\right)$: $\nearrow$.	$\pi \leq \frac{v}{2\mu^2}$: $\searrow$. $\pi > \frac{v}{2\mu^2}$: $\wedge$, max at $\frac{2\mu^2(2\pi\mu^2/v-1)}{v(1+2\sigma^2/\alpha^2)}$.

Table 2: Monotonicity properties of various components of the risk for a two-layer network with nonlinear activation at the optimal λ^*, as a function of π and δ, while holding all other parameters fixed. $\nearrow$: non-decreasing. $\searrow$: non-increasing. $\wedge$: unimodal. Thus, e.g., the MSE is non-increasing as a function of the parameterization level π, while holding δ fixed.

Thus, comparing Tables 2 and 1, we see that with stronger Gaussian assumptions on the data distribution, optimal ridge regularization has similar effects in the nonlinear and linear cases. For instance, λ^* can eliminate the "peaking" shape of the MSE.

4. Numerical Simulations

In this section, we perform several numerical experiments, to check the correctness of our theoretical results.

4.1 Verifying the Theoretical Results for the MSE

To check the correctness of the MSE formula, we estimate the MSE from its definition directly. For simplicity, we subtract the test point label noise σ^2 from the MSE formula. We randomly generate $k = 400$ i.i.d. tuples of random variables $(x_i, \theta_i, \varepsilon_i, X_i, W_i)$, $1 \leq i \leq k$, from their assumed distributions (we assume X and x have i.i.d. $\mathcal{N}(0,1)$ entries in numerical simulations), and estimate the MSE by calculating:

$$\widehat{\mathrm{MSE}} = \frac{1}{k}\sum_{i=1}^{k}(\hat{f}_i(x_i) - x_i^\top \theta_i)^2,$$

where $k = 400$ and $\hat{f}_i(x_i) = \sigma(x_i^\top W_i^\top)\left(n^{-1}\sigma(W_i X_i^\top)\sigma(X_i W_i^\top) + \lambda I_p\right)^{-1} n^{-1}\sigma(W_i X_i^\top)(X_i\theta_i + \mathcal{E}_i)$, for $\sigma(x)$ both linear and nonlinear. We repeat this process 20 times and plot the mean and standard error in Figure 7. The regularization parameter λ is set optimally or fixed. We also plot our theoretical MSE formula from (13). The parameters in the experiment are shown in the captions of the figure. We can see from Figure 7 that our theoretical prediction of the MSE is quite accurate.

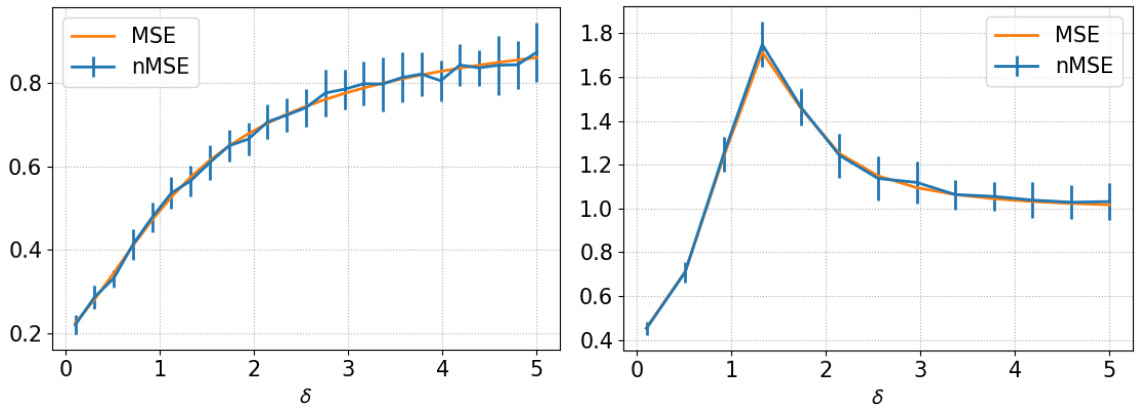


Figure 7: Numerical verification of the theoretical results for MSE. We display, as a function of $\delta = \lim d/n$, the theoretical formula and the numerical mean and standard deviation over 20 repetitions. Parameters: $\alpha = 1, \sigma = 0.3, \pi = 0.8, n = 150, d = \lfloor n\delta \rfloor, p = \lfloor d\pi \rfloor$. **Left**: linear, $\sigma(x) = x$, $\lambda = \lambda^*$(optimal). **Right**: nonlinear, $\sigma(x) = \sigma_+(x) - \mathbb{E}_{x\sim\mathcal{N}(0,1)}\sigma_+(x)$, $\sigma_+(x) = \max(x, 0)$, $\lambda = 0.01$.

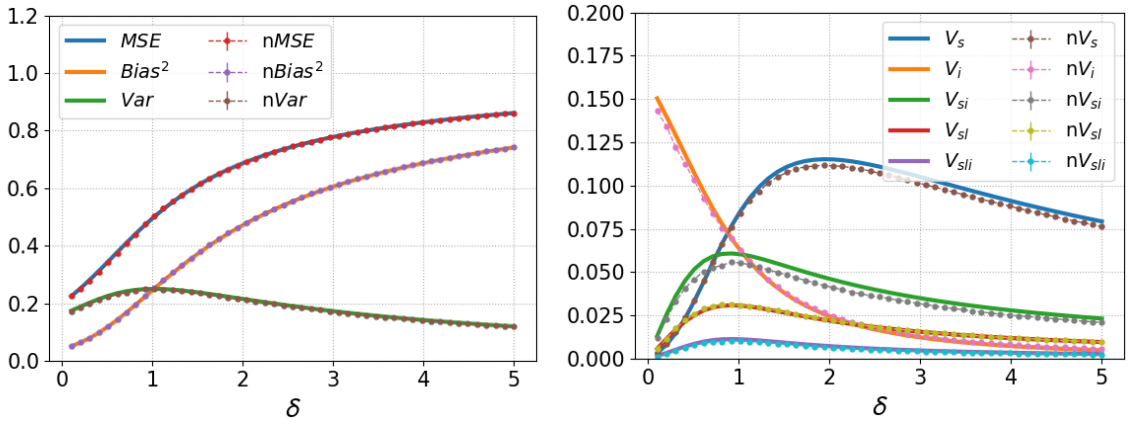


Figure 8: **Left**: numerical simulation verifying the accuracy of the bias, variance and MSE formulas. **Right**: simulations with variance components. For each ANOVA component, symbolized by $\star$, we show two curves: $\star$: theory, $n\star$: numerical (averaged over 5 runs). Parameters: $\alpha = 1, \sigma = 0.3, \pi = 0.8, n = 150, d = \lfloor n\delta \rfloor, p = \lfloor d\pi \rfloor$.

4.2 Bias-variance Decomposition and the Variance Components

We next study the accuracy of the formulas for the bias, variance and the variance components in the linear case. Estimating them directly requires many samples. For example, to estimate the bias based on the defining formula from Section 2.2, we may need to generate, say, 100 pairs of (x, θ), and for each (x, θ) generate 500 triples of i.i.d. $(X, W, \mathcal{E})$. Thus, we may need to simulate $50,000$ samples in total to obtain accurate results. This is beyond our current scope. Therefore, for simplicity, we check instead the formulas that we have derived in Appendix B in equations (25)—(27), (32)—(38). We omit the results for V_l and V_{li} since they converge to 0. In all experiments, we choose $n = 150$. For the bias, variance,

Functional	Estimator
$\mathbb{E}\,\mathrm{tr}(\mathcal{M}\mathcal{M}^\top)$	$\frac{1}{k}\sum_{i=1}^{k}\mathrm{tr}(\mathcal{M}_i\mathcal{M}_i^\top)$
$\|\mathbb{E}\mathcal{M}\|_F^2$	$\left\|\frac{1}{k}\sum_{i=1}^{k}\mathcal{M}_i\right\|_F^2$
$\mathbb{E}_W\|\mathbb{E}_X\mathcal{M}\|_F^2$	$\frac{1}{k}\sum_{j=1}^{k}\left\|\frac{1}{k}\sum_{i=1}^{k}\mathcal{M}_{ij}\right\|_F^2$
$\mathbb{E}_X\|\mathbb{E}_W\mathcal{M}\|_F^2$	$\frac{1}{k}\sum_{i=1}^{k}\left\|\frac{1}{k}\sum_{j=1}^{k}\mathcal{M}_{ij}\right\|_F^2$

Table 3: Empirical estimators of functionals of interest. Here $\mathcal{M}$ is a generic matrix that can be M or $\tilde{M}$. For the bias, variance and the MSE, we take $k = 100$ and $\mathcal{M}_i$ denotes the appropriate matrix $\mathcal{M}$ obtained from the i-th pair of (X, W). For variance components, $k = 20, 50$ and $\mathcal{M}_{ij}$ denotes the appropriate matrix $\mathcal{M}$ obtained from the i-th X and the j-th W. Estimators of the quantities in (25)—(27), (32)—(38) are obtained by combining the above.

and MSE, we generate 100 i.i.d. copies of X, W of certain dimensions (we assume X has i.i.d. $\mathcal{N}(0, 1)$ entries in numerical simulations) and estimate the expectations from the proof of Theorem 3 in Appendix B (25—27) using the Monte Carlo mean. As for the variance components, we randomly generate k i.i.d. X_i-s and W_j-s, and use them to form k^2 pairs of (X_i, W_j), where $k = 20$ in Figure 8 (right) and $k = 50$ in Figure 1 (left). Similarly, we also estimate the expectations from the proof of Theorem 2 in (32)—(38) via the Monte Carlo mean (see the details in Table 3).

Figure 8 shows the results averaged over five runs. We can see that the numerical results are quite close to the theoretical predictions. Moreover, the standard deviations over 5 runs are uniformly less than 0.001 in all settings we considered, which implies that the variance due to the randomness of (X, W) is negligible. The slight discrepancies between the theoretical predictions and experiments are mostly owing to the bias in our estimators (e.g., the second, third and fourth estimators in Table 3 are biased, because they are of the form $g(\mathbb{E}M)$, which is estimated by $g(k^{-1}\sum_{i=1}^{k} M_i)$, and g is nonlinear), and they can be reduced if we have more samples (X, W).

One may wonder: Why is the standard deviation of different runs of the simulation so small (e.g., less than 0.001)? The reason is that the Marchenko-Pastur theorem, on which our theoretical predictions depend, has fast convergence rates (e.g., Bai, 1993; Götze et al., 2004). Since the terms we estimated are expectations of various functionals of the eigenvalue spectrum, we can expect the simulation results to be quite precise.

4.3 General Covariance

Although our theoretical results are proved under an assuming the data distribution is isotropic, we also study the model under general covariance assumptions numerically. Namely, we assume the samples are drawn i.i.d. from $\mathcal{N}(0, \Sigma(r))$, where $\Sigma(r)_{ij} = r^{|i-j|}$ is an AR-1 covariance matrix. We numerically study the bias-variance decomposition and the variance

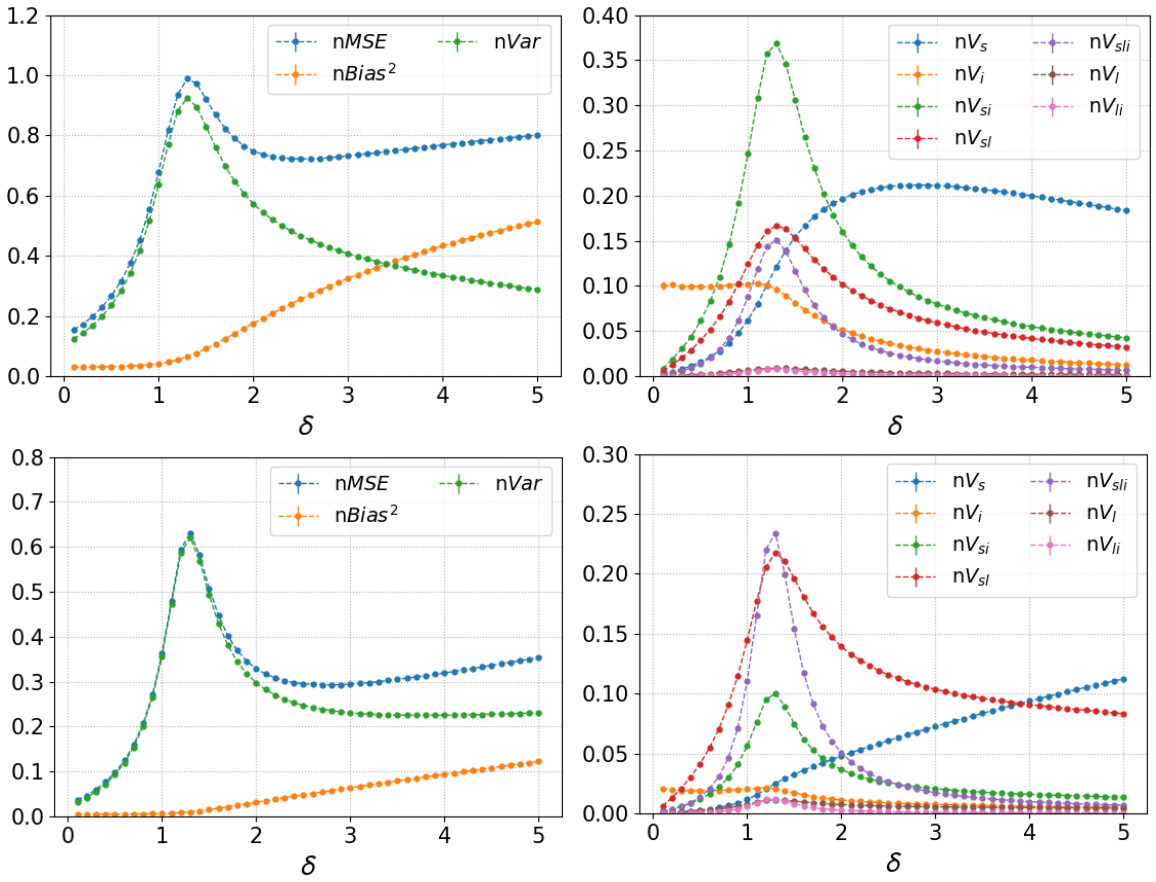


Figure 9: Bias-variance decomposition and the variance components under an AR-1 covariance assumption. **Left**: numerical simulation of the bias, variance and MSE formulas. **Right**: simulations with variance components. Upper: $r = 0.5, \lambda = 0.01$. Down: $r = 0.9, \lambda = 0.001$. For each term, symbolized by $\star$, we show $n\star$: numerical (averaged over 5 runs). Parameters: $\alpha = 1, \sigma = 0.3, \pi = 0.8, n = 150, d = \lfloor n\delta \rfloor, p = \lfloor d\pi \rfloor$.

components the same way as in Section 4.2. The only distinction is that the formulas we estimate are slightly different due to the non-identity covariance. More specifically, one can show that the formulas for the general covariance case are the same as their counterparts in equations (25)—(27), (32)—(38) with $\|\cdot\|_F^2$ replaced by $\mathrm{tr}(\cdot\,\cdot^\top \Sigma(r))$.

In the experiment, we choose fixed small penalties $\lambda = 0.01, 0.001$ to mimic the ridgeless limit. Figure 9 shows the numerical results when $r = 0.5$, 0.9. We see that many observations in the isotropic case (e.g. monotonic bias, non-monotonicity of the MSE and variance, the interaction terms dominating the variance at the interpolation threshold) still hold in the general covariance case. However, the terms contributing the most to the total variance are V_{sl} and V_{sli} when $r = 0.9$, while they are V_{si} when $r = 0.5$ or in the isotropic case. We conjecture that this is because the covariance matrix can implicitly change the ratio between the noise σ and signal α. However, more work is needed in the future for understanding the generalization error under a general covariance assumption.

4.4 Experiments on Empirical Data

In this section, we present an empirical data example to study several phenomena observed in our theoretical analysis. We use the Superconductivity Data Set (Hamidieh, 2018) retrieved from the UC Irvine Machine Learning Repository in our data analysis.

The original data set contains $N = 21,263$ superconductors as samples and $d = 81$ features for each sample. Our goal is to predict the critical temperatures of the superconductors based on their features. Before doing regression, we preprocess the data set in the following way. We first randomly shuffle the samples. We then separate the data set into a training set containing the first 90% of the samples, and a test set containing the rest. Finally, we normalize the features and responses so that they all have zero mean and unit variance. Since N is quite a bit larger than d, we can estimate the variances of the features quite well; and thus we can standardize new test datapoints from this distribution using the estimated variances. After these steps, we are ready to start our experiments.

Similar to our theoretical setting, for each experiment setting (p, n), we randomly select n samples from the training set to form a data matrix $X \in \mathbb{R}^{n\times d}$, map it into a random p-dimensional subspace multiplying it by a projection matrix W with orthogonal rows and then perform ridge regression on the p-dimensional subspace. For each setting, we generate 50 i.i.d. sample matrices X_i, $1 \le i \le 50$, and 50 i.i.d. random projections W_j, $1 \le j \le 50$, and combine them to form 2500 sample-initialization pairs (X_i, W_j), $1 \le i, j \le 50$. Denote by y_i the response vector of X_i. Then for each (X_i, W_j), we have the ridge estimator

$$\hat{f}_{ij}(x) = x^\top \left(\frac{W_j X_i^\top X_i W_j^\top}{n} + \lambda I_p \right)^{-1} \frac{W_j X_i^\top y_i}{n}, \qquad 1 \le i, j \le 50.$$

We use these estimators to make predictions on the test set and estimate the MSE, bias, and some of the variance components as $\widehat{\mathrm{MSE}} := L^{-1} \sum_{k=1}^{L} \hat{\mathbb{E}}(\hat{f}_{ij}(x_k) - y_k)^2$, and:

$$\widehat{\mathrm{Var}} := \frac{1}{L} \sum_{k=1}^{L} \hat{\mathbb{E}}(\hat{f}_{ij}(x_k) - \hat{\mathbb{E}}\hat{f}_{ij}(x_k))^2, \qquad \widehat{\mathrm{Bias}}^2 := \frac{1}{L} \sum_{k=1}^{L} (\hat{\mathbb{E}}\hat{f}_{ij}(x_k) - y_k)^2,$$

$$\widehat{V_s} := \frac{1}{L n_s} \sum_{k=1}^{L} \sum_{i=1}^{n_s} (\hat{\mathbb{E}}_j \hat{f}_{ij}(x_k) - \hat{\mathbb{E}}\hat{f}_{ij}(x_k))^2, \quad \widehat{V_i} := \frac{1}{L n_i} \sum_{k=1}^{L} \sum_{j=1}^{n_i} (\hat{\mathbb{E}}_i \hat{f}_{ij}(x_k) - \hat{\mathbb{E}}\hat{f}_{ij}(x_k))^2,$$

where $\hat{\mathbb{E}}$ denotes the Monte Carlo mean, $n_i = n_s = 50$, L is the test set size, and x_k, y_k are test features and responses.

Since we have no information about the true noise of the responses, we only study the variance introduced by the choice of the data matrix X and initialization W. Figure 10 shows the empirically estimated bias, variance and MSE as functions of the number of samples n given a fixed amount of parameterization (fixed $\pi = \lim p/d$). From this figure, we observe the following:

1. The MSE is unimodal as a function of number of samples n, which corroborates that increasing the number of training samples can sometimes lower the model's performance when we do not have enough samples and do not regularize well (e.g. use a small $\lambda = 0.01$). This unimodality is quite similar to the unimodality we observed in

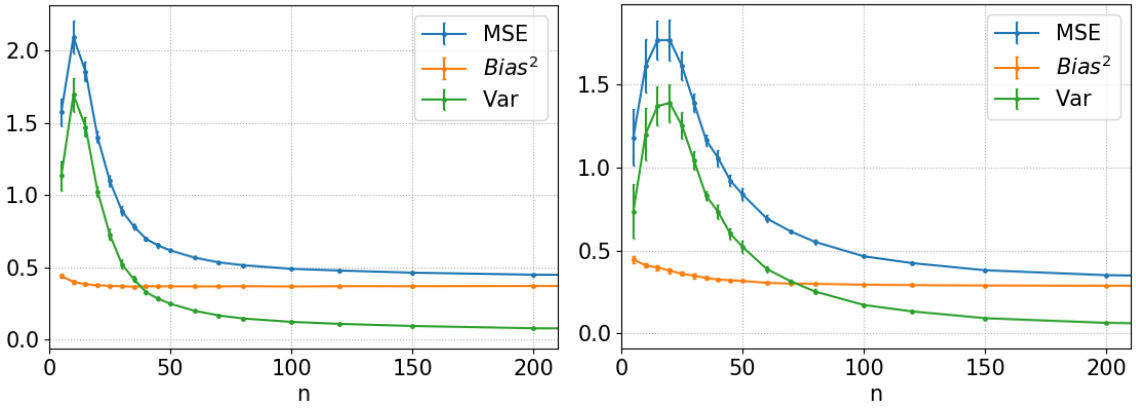


Figure 10: Empirically estimated MSE, variance and bias as functions of number of samples n. We display the mean and one standard deviation of the numerical results over 10 repetitions. **Left**: $\pi = 0.2, \lambda = 0.01$. **Right**: $\pi = 0.9, \lambda = 0.01$. (Both panels are from the same simulation.)

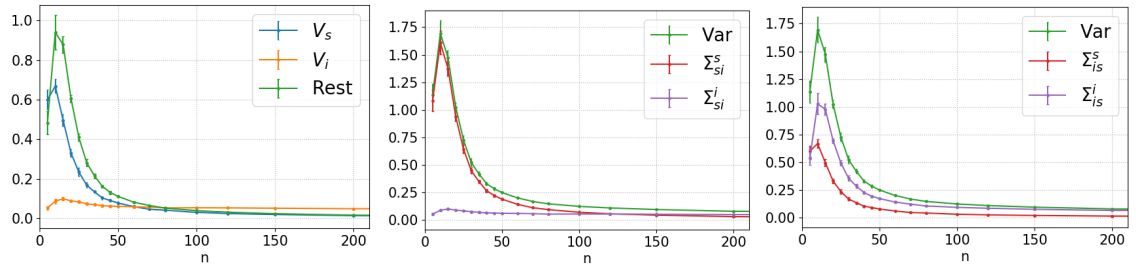


Figure 11: Numerically estimated variance components as a function of the sample size n. **Left**: three components of variance (V_s, V_i, and "Rest": the variance due to interaction and response noise, Rest := Var $- V_s - V_i$). Middle and right: two orders of variance decomposition. Middle: sample, initialization. **Right**: initialization, sample. $\Sigma^a_{ab} :=$ Var $- V_b, \Sigma^b_{ab} := V_b, \{a, b\} = \{s, i\}$. Parameters: $\pi = 0.2, \lambda = 0.01$. We display the mean and one standard deviation over 10 repetitions. (All three panels are from the same simulation.)

Figure 4 in our theoretical setting. It is also consistent with the general phenomenon of sample-wise double descent (Nakkiran, 2019).

2. The bias is decreasing as a function of n, when n is small, and stays roughly constant when n is larger. The reason is that the data does not truly come from a linear model, and thus the linear model that we use has a nonzero approximation bias.

3. The bias is also decreasing as a function of $1/\delta = n/d$. This suggests that more samples can reduce the bias of the ridge estimator. Furthermore, the variance is the main contributor to the unimodality of the MSE. These two observations are also consistent with our theoretical results from Theorem 7, which suggests that the bias is

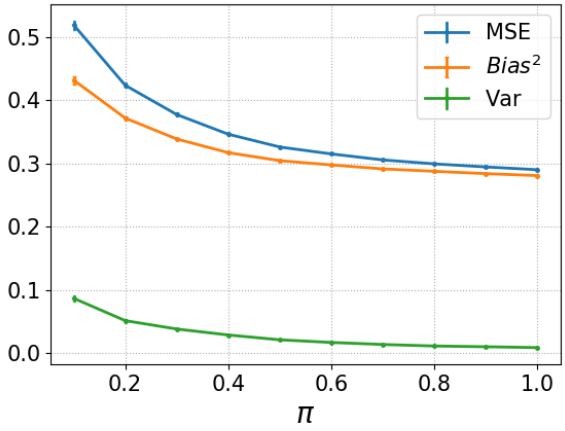


Figure 12: Empirically estimated MSE, variance and bias as functions of degree of parameterization $\pi_d = p/d$. We show the the mean and one standard deviation over 10 repetitions.

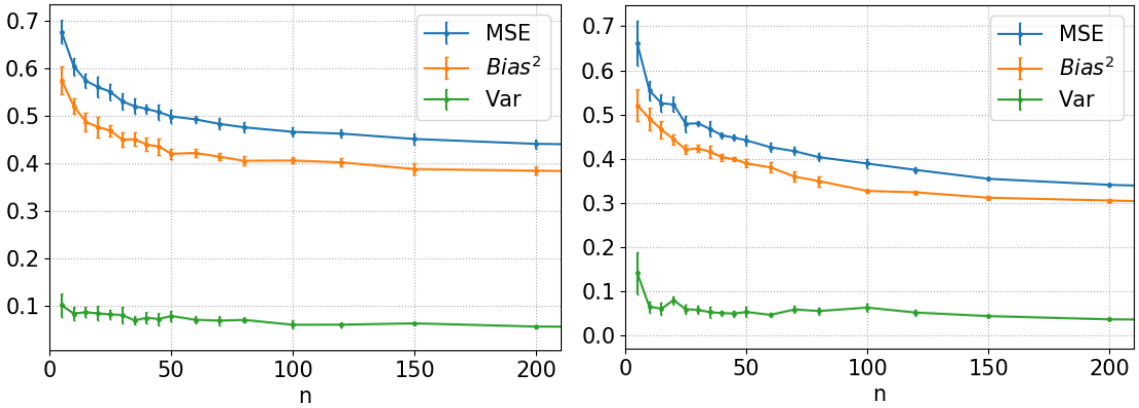


Figure 13: Empirically estimated MSE, variance and bias as functions of number of samples n using the optimal λ^*. We display the mean and one standard deviation of the numerical results over 10 repetitions. **Left**: $\pi = 0.2$. **Right**: $\pi = 0.9$. (Both panels are from the same simulation.)

increasing as a function of δ and the variance can be very large along the interpolation threshold $\delta\pi = 1$ when λ is small.

Figure 11 (left) shows estimates of three components of variance in our data example. In this low parameterization setting ($\pi = 0.2$), when n is small (< 100), the variance V_s due purely to sampling is large, the variance V_i due purely to initialization is small, and the variance V_{is} due to their interaction and also the response noise is large. Thus, combining these variances together, we can see from Figure 11 (middle and right) that different orders of decomposition can indeed lead to different interpretations of the variances introduced by sampling and initialization.

Figure 12 exhibits empirical estimates of the bias, variance, and MSE as functions of the degree of parameterization $\pi = \lim p/d$ when given enough samples, here $n = 1000$, so

that $\delta_d = d/n$ is small. We see that all three terms decrease as π increases, which means more parameters can improve the estimator's performance when we have enough samples (δ is small). This is also close to what we have observed in Figure 6h, i.e., that the MSE is decreasing as π increases when $\delta < 1$.

We also study the effect of optimal ridge penalty on the empirical data. For simplicity, we select the ridge parameter from the set $\{i \times 10^{-j} | i = 1, 2, 5; j = 0, 1, 2, 3\}$ such that it minimizes the empirical MSE. Figure 13 shows the MSE, variance and bias obtained using the optimal ridge penalty. Compared with Figure 10, we see that the optimal ridge penalty can mitigate the non-monotonicity of MSE and keep the bias decreasing as the number of sample n increases. These observations are consistent with what we have shown in our theoretical setting.

To conclude, although our theoretical results are based on quite strong assumptions on the data distribution, many conclusions and insights still carry over to certain problems involving empirical data.

Acknowledgments

We thank the associate editor for handling our paper. We are very grateful for the reviewers for detailed and thorough feedback, which has lead to numerous important improvements. We thank Yi Ma, Song Mei, Zitong Yang, Chong You, Yaodong Yu for helpful discussions. This work was partially supported by a Peking University Summer Research award, and by the NSF-Simons Collaboration on the Mathematical and Scientific Foundations of Deep Learning THEORINET (NSF 2031985). This work was performed when LL was a student at Peking University.

Appendix A. Comparison of Orthogonal and Gaussian Initialization Models

Here we provide a comparison of the orthogonal and Gaussian initialization models for linear networks $f(x) = (Wx)^\top \beta$. The *expressive power* of the two models is the same, as with probability one we can write a $p \times d$ matrix W with iid Gaussian entries, where $p \leqslant d$, via its SVD as $W = UDV$, where U is $p \times p$ orthonormal, D is $p \times p$ diagonal with nonzero entries with probability one, and V is $p \times d$ partial orthonormal with $VV^\top = I_p$. Then $(Wx)^\top \beta = x^\top W^\top \beta = x^\top V^\top D U^\top \beta = x^\top W_o^\top \beta_o$, where $W_o = UV$ is a random partial orthonormal matrix, and $\beta_o = UDU^\top \beta$ is a new regression coefficient. Thus, the two models have the same expressive power.

The orthogonal model we consider has some advantages over the Gaussian model. Indeed, considering the case when $p = d$, in the orthogonal model, we first rotate x orthogonally, then take a linear combination of the coefficients. In contrast, in the Gaussian model we not only rotate x, but also scale it by the singular values of x, which due to the Marchenko-Pastur law (Marchenko and Pastur, 1967) spread out from zero to two. Thus, we induce a significant distortion of the input in the first layer. Then, we can expect that learning may be more challenging due to this additional scaling. Indeed, the regression coefficients corresponding to the directions with near-zero singular values must be scaled up asymptotically by unboundedly large values for accurate prediction. On the other hand, the Gaussian model more closely mimics practical initialization schemes, which can indeed involve iid weights. We also note that recently, some empirical work has argued about the benefits of orthogonal initialization (Hu et al., 2020; Qi et al., 2020). For instance, Qi et al. (2020) argues that orthogonality (or isometry) alone enables training practical >100 layer CNNs on ImageNet without shortcut and BatchNorm, and therefore provides some justification for orthogonality in our theoretical analysis.

Appendix B. Proofs

B.1 Proof of Theorem 3

Different from the order of the theorems, here we first give the proof of theorem 3 and then the proof of theorem 2. This is because the proof of theorem 2 is more complicated and depends on some lemmas in the proof of theorem 3.

In the proofs, we will often refer to the spectral distribution (or measure) a symmetric matrix M, which is simply the discrete distribution placing uniform point masses on each of the (real) eigenvalues of M. When the matrix size grows, we will consider settings where the spectral distribution converges in distribution to a fixed probability distribution.

Let us define

$$\tilde{M}_{X,W}(\lambda) := W^\top (n^{-1} W X^\top X W^\top + \lambda I_p)^{-1} W X^\top / n$$

(a $d \times n$ matrix), and

$$M_{X,W}(\lambda) := \tilde{M}_{X,W}(\lambda) X$$

(a $d \times d$ matrix) and omit their dependence on λ, X, W for simplicity. As we will clearly see below, M can be viewed as a "regularized pseudo-inverse". Also, $\mathbb{E}M - I$ directly controls

the bias, and $M - \mathbb{E}M$ controls part of the variance. The calculations of bias and variance in terms of M follow those of Yang et al. (2020), and we include them here for the reader's convenience. The calculations following them are more novel.

To start, we have $f_{\lambda,\mathcal{T},W}(x) = x^\top \tilde{M} Y$, and similar to the proof of theorem 1 in Yang et al. (2020)

$$\begin{aligned}\textbf{Bias}^2(\lambda) &= \mathbb{E}_{\theta,x}\left[\mathbb{E}_{X,W,\mathcal{E}}\left(x^\top M\theta + x^\top \tilde{M}\mathcal{E}\right) - x^\top\theta\right]^2 \\ &= \mathbb{E}_{\theta,x}\left[x^\top(\mathbb{E}_{X,W,\mathcal{E}}M - I)\theta + \mathbb{E}_{\mathcal{T},W}(x^\top\tilde{M}\mathcal{E})\right]^2.\end{aligned}$$

Since $\mathcal{E}$ has zero mean and is independent of x and $\tilde{M}$, we know $\mathbb{E}_{X,W,\mathcal{E}}(x^\top\tilde{M}\mathcal{E}) = 0$. In what follows, sometimes we omit the subscript when we take expectation over all random variables. Thus, using that for any two vectors and a matrix of conformable sizes, $(a^\top Nb)^2 = a^\top Nbb^\top N^\top a = \operatorname{tr} Nbb^\top N^\top aa^\top$, the above equals

$$\begin{aligned}&\mathbb{E}_{\theta,x}\left[x^\top(\mathbb{E}M - I)\theta\theta^\top(\mathbb{E}M - I)^\top x\right] \\ &= \mathbb{E}_{\theta,x}\operatorname{tr}\left[x^\top(\mathbb{E}M - I)\theta\theta^\top(\mathbb{E}M - I)^\top x\right] \\ &= \operatorname{tr}\left[(\mathbb{E}M - I)\mathbb{E}\left(\theta\theta^\top\right)(\mathbb{E}M - I)^\top\mathbb{E}\left(xx^\top\right)\right] \\ &= \frac{\alpha^2}{d}\|\mathbb{E}M - I\|_F^2. \end{aligned} \tag{25}$$

This shows that the average bias is determined by how well the random matrix M (which depends both on the random data X and the random initialization W) approximates the identity matrix.

Similarly, by grouping terms appropriately, and using again that $\mathbb{E}_{X,W,\mathcal{E}}(x^\top\tilde{M}\mathcal{E}) = 0$,

$$\begin{aligned}\textbf{Var}(\lambda) &= \mathbb{E}_{\theta,x,X,W,\mathcal{E}}\left[x^\top M\theta + x^\top\tilde{M}\mathcal{E} - \mathbb{E}_{X,W,\mathcal{E}}(x^\top M\theta + x^\top\tilde{M}\mathcal{E})\right]^2 \\ &= \mathbb{E}_{\theta,x,X,W,\mathcal{E}}\left[x^\top(M - \mathbb{E}M)\theta + x^\top\tilde{M}\mathcal{E}\right]^2 \\ &= \mathbb{E}_{\theta,x,X,W,\mathcal{E}}\left[x^\top(M - \mathbb{E}M)\theta\right]^2 + (x^\top\tilde{M}\mathcal{E})^2,\end{aligned}$$

where the interaction term is zero because of the independence between $\mathcal{E}$ and other variables. Then, using that $\operatorname{tr} A^\top A = \|A\|_F^2$,

$$\begin{aligned}\textbf{Var}(\lambda) &= \mathbb{E}_{\theta,x,X,W,\mathcal{E}}\left\{\left[x^\top(M - \mathbb{E}M)\theta\theta^\top(M - \mathbb{E}M)^\top x\right] + x^\top\tilde{M}\mathcal{E}\mathcal{E}^\top\tilde{M}^\top x\right\} \\ &= \mathbb{E}_{\theta,x,X,W}\left\{\operatorname{tr}\left[(M - \mathbb{E}M)\theta\theta^\top(M - \mathbb{E}M)^\top xx^\top\right] + \sigma^2\operatorname{tr}[x^\top\tilde{M}\tilde{M}^\top x]\right\} \\ &= \mathbb{E}_M\operatorname{tr}\left[(M - \mathbb{E}M)\mathbb{E}(\theta\theta^\top)(M - \mathbb{E}M)^\top\mathbb{E}(xx^\top)\right] + \sigma^2\mathbb{E}\operatorname{tr}\left[\tilde{M}\tilde{M}^\top\mathbb{E}(xx^\top)\right] \\ &= \frac{\alpha^2}{d}\mathbb{E}\|M - \mathbb{E}M\|_F^2 + \sigma^2\mathbb{E}\|\tilde{M}\|_F^2.\end{aligned} \tag{26}$$

Thus, the variance is determined by how much M varies around its mean, and by how large $\tilde{M}$ is. Combining results for variance and bias, and using that $\mathbb{E}\|M - I\|_F^2 = \mathbb{E}\|M -$

$\mathbb{E}M\|_F^2 + \mathbb{E}\|\mathbb{E}M - I\|_F^2$, we obtain

$$\mathbf{MSE}(\lambda) = \mathbf{Var}(\lambda) + \mathbf{Bias}^2(\lambda) + \sigma^2 = \frac{\alpha^2}{d}\mathbb{E}\|M - I\|_F^2 + \sigma^2\mathbb{E}\|\tilde{M}\|_F^2 + \sigma^2. \tag{27}$$

Similarly, for $\Sigma_{\text{label}}, \Sigma_{\text{sample}}, \Sigma_{\text{init}}$, we have

$$\begin{aligned}\Sigma_{\text{label}} &= \mathbb{E}_{\theta,x}\mathbb{E}_{W,X,\mathcal{E}}[f_{\lambda,\mathcal{T},W}(x) - \mathbb{E}_{\mathcal{E}} f_{\lambda,\mathcal{T},W}(x)]^2 \\ &= \mathbb{E}_{\theta,x,W,X,\mathcal{E}}(x^\top \tilde{M}\mathcal{E})^2 = \sigma^2\mathbb{E}\|\tilde{M}\|_F^2.\end{aligned}$$

Thus, the variance due to label noise is determined by the magnitude of $\tilde{M}$. Also,

$$\begin{aligned}\Sigma_{\text{sample}} &= \mathbb{E}_{\theta,x}\mathbb{E}_{W,X}[\mathbb{E}_{\mathcal{E}} f_{\lambda,\mathcal{T},W}(x) - \mathbb{E}_{X,\mathcal{E}} f_{\lambda,\mathcal{T},W}(x)]^2 \\ &= \mathbb{E}_{\theta,x,W,X}[x^\top M\theta - \mathbb{E}_X(x^\top M\theta)]^2 \\ &= \frac{\alpha^2}{d}\mathbb{E}_{W,X}\left\|M - \mathbb{E}_X M\right\|_F^2.\end{aligned}$$

Finally,

$$\begin{aligned}\Sigma_{\text{init}} &= \mathbb{E}_{\theta,x}\mathbb{E}_W(\mathbb{E}_{X,\mathcal{E}} f_{\lambda,\mathcal{T},W}(x) - \mathbb{E}_{W,X,\mathcal{E}} f_{\lambda,\mathcal{T},W}(x))^2 \\ &= \mathbb{E}_{\theta,x,W}[\mathbb{E}_X(x^\top M\theta) - \mathbb{E}_{W,X}(x^\top M\theta)]^2 \\ &= \frac{\alpha^2}{d}\mathbb{E}_W\left\|\mathbb{E}_X M - \mathbb{E}_{W,X} M\right\|_F^2.\end{aligned}$$

This shows that in the specific decomposition order: label, samples, initialization, the variance due to the randomness in the sample X is determined by the Frobenius variability of M around its mean with respect to X. The respective statement is also true for the variance due to initialization.

Therefore, to prove theorem 3, it suffices to study the limiting behaviours of $\mathbb{E}M$, $\mathbb{E}_X M$, $\tilde{M}$ and M. Under the assumptions in Theorem 3, we have characterize their behavior in the following lemmas.

Lemma 11 (Behavior of $\mathbb{E}M$)

$$\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\,\mathrm{tr}(M) = \pi(1-\lambda\theta_1), \quad \forall i \geq 1. \qquad \lim_{d\to\infty}\frac{1}{d}\|\mathbb{E}M\|_F^2 = \pi^2(1-\lambda\theta_1)^2. \tag{28}$$

Lemma 12 (Behavior of the Frobenius norm of M)

$$\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\|M\|_F^2 = \pi\left[1 - 2\lambda\theta_1 + \lambda^2\theta_2 + (1-\pi)\delta(\theta_1 - \lambda\theta_2)\right]. \tag{29}$$

Lemma 13 (Behavior of the Frobenius norm of $\tilde{M}$)

$$\lim_{d\to\infty}\mathbb{E}\|\tilde{M}\|_F^2 = \pi\delta(\theta_1 - \lambda\theta_2). \tag{30}$$

Lemma 14 (Behavior of the Frobenius norm of $\mathbb{E}_X M$)

$$\lim_{d\to\infty}\frac{1}{d}\mathbb{E}_W\|\mathbb{E}_X M\|_F^2 = \pi(1-\lambda\theta_1)^2. \tag{31}$$

We put the proof of these lemmas after the proof of theorem 3 and 2 for clarity (see Appendix B.3). By using Lemmas 11—14, we are able to complete our proof.

Proof [Proof of Theorem 3] Plugging equation (28) into (25), we have

$$\mathbf{Bias}^2(\lambda) = \frac{\alpha^2}{d}\|\mathbb{E}M - I\|_F^2 \to \alpha^2(\pi(1-\lambda\theta_1)-1)^2 = \alpha^2(1-\pi+\lambda\pi\theta_1)^2.$$

Plugging equations (28), (29), (30) into (26), we have

$$\begin{aligned}
\mathbf{Var}(\lambda) &= \frac{\alpha^2}{d}\mathbb{E}\|M-\mathbb{E}M\|_F^2 + \sigma^2\mathbb{E}\|\tilde{M}\|_F^2 \\
&= \frac{\alpha^2}{d}\left[\mathbb{E}\|M\|_F^2 - \|\mathbb{E}M\|_F^2\right] + \sigma^2\mathbb{E}\|\tilde{M}\|_F^2 \\
&\to \alpha^2\pi\left[1-2\lambda\theta_1+\lambda^2\theta_2+(1-\pi)\delta(\theta_1-\lambda\theta_2)-\pi(1-\lambda\theta_1)^2\right] + \sigma^2\pi\delta(\theta_1-\lambda\theta_2) \\
&= \alpha^2\pi\left[1-\pi+(\pi-1)(2\lambda-\delta)\theta_1-\pi\lambda^2\theta_1^2+\lambda(\lambda-\delta+\pi\delta)\theta_2\right] + \sigma^2\pi\delta(\theta_1-\lambda\theta_2).
\end{aligned}$$

Similarly, by Lemmas 12, 13, 14

$$\begin{aligned}
\Sigma_{\text{label}} &= \sigma^2\mathbb{E}\|\tilde{M}\|_F^2 \to \sigma^2\pi\delta(\theta_1-\lambda\theta_2). \\
\Sigma_{\text{sample}} &= \frac{\alpha^2}{d}\mathbb{E}_{W,X}\|M-\mathbb{E}_X M\|_F^2 = \frac{\alpha^2}{d}[\mathbb{E}\|M\|_F^2 - \mathbb{E}_W\|\mathbb{E}_X M\|_F^2] \\
&\to \alpha^2\pi\left[-\lambda^2\theta_1^2+\lambda^2\theta_2+(1-\pi)\delta(\theta_1-\lambda\theta_2)\right].
\end{aligned}$$

Finally,

$$\begin{aligned}
\Sigma_{\text{init}} &= \mathbf{Var}(\lambda) - \Sigma_{\text{sample}} - \Sigma_{\text{label}} \to \alpha^2\pi(1-\pi)(1-\lambda\theta_1)^2. \\
\mathbf{MSE}(\lambda) &= \mathbf{Var}(\lambda) + \mathbf{Bias}^2(\lambda) + \sigma^2 \\
&\to \alpha^2\left\{1-\pi+\pi\delta\left(1-\pi+\sigma^2/\alpha^2\right)\theta_1 + \left[\lambda-\delta\left(1-\pi+\sigma^2/\alpha^2\right)\right]\lambda\pi\theta_2\right\} + \sigma^2.
\end{aligned}$$

As for the choice of optimal λ^*, denote $\delta(1-\pi+\sigma^2/\alpha^2)$ by c and calculate

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\lim_{d\to\infty}\mathbf{MSE}(\lambda) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\alpha^2\left[1-\pi+\pi c\theta_1+(\lambda-c)\lambda\pi\theta_2\right] + \sigma^2 \\
&= \alpha^2\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\int\frac{\pi c}{x+\lambda}dF_\gamma(x) + \int\frac{(\lambda-c)\lambda\pi}{(x+\lambda)^2}dF_\gamma(x)\right) \\
&= \alpha^2\pi\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\int\frac{\lambda^2+cx}{(x+\lambda)^2}dF_\gamma(x)\right) \\
&= 2\alpha^2\pi\int\frac{(\lambda-c)x}{(x+\lambda)^3}dF_\gamma(x).
\end{aligned}$$

If $\pi = 1$ and $\sigma = 0$, then $c = 0$ and the asymptotic MSE is monotonically increasing since $F_\gamma(x)$ is supported on $[0, +\infty)$. Therefore the optimal ridge $\lambda^* = 0$, which is outside of the range $(0, \infty)$ that we considered here. Otherwise , the derivative is less than zero when $\lambda < c$ and larger than zero when $\lambda > c$. Therefore, the asymptotic MSE as a function of λ has a unique minimum at $c = \delta(1-\pi+\sigma^2/\alpha^2)$.

■

B.2 Proof of Theorem 2

In this proof, we will use the same notations as in the proof of theorem 3. Since the main idea of this proof is quite similar to the proof of theorem 3, we will omit some details in the derivation for simplicity.

Proof [Proof of Theorem 2] By definition, for V_s,

$$\begin{aligned} V_s &= \mathbb{E}_{\theta,x}\mathrm{Var}_X(\mathbb{E}_{\mathcal{E},W}(\hat{f}(x)|X)) = \mathbb{E}_{\theta,x,X}[x^\top(\mathbb{E}_W M - \mathbb{E}M)\theta]^2 \\ &= \frac{\alpha^2}{d}\mathbb{E}_X\|\mathbb{E}_W M - \mathbb{E}M\|_F^2. \end{aligned} \tag{32}$$

For V_l,

$$V_l = \mathbb{E}_{\theta,x}\mathrm{Var}_{\mathcal{E}}(\mathbb{E}_{X,W}(\hat{f}(x)|\mathcal{E})) = \sigma^2\|\mathbb{E}\tilde{M}\|_F^2. \tag{33}$$

For V_i,

$$\begin{aligned} V_i &= \mathbb{E}_{\theta,x}\mathrm{Var}_W(\mathbb{E}_{\mathcal{E},X}(\hat{f}(x)|W)) = \mathbb{E}_{\theta,x,W}[x^\top(\mathbb{E}_X M - \mathbb{E}M)\theta]^2 \\ &= \frac{\alpha^2}{d}\mathbb{E}_W\|\mathbb{E}_X M - \mathbb{E}M\|_F^2. \end{aligned} \tag{34}$$

Similarly,

$$\begin{aligned} V_{sl} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{\mathcal{E},X}(\mathbb{E}_W(\hat{f}(x)|\mathcal{E},X)) - V_s - V_l \\ &= \mathbb{E}_{\theta,x,X,\mathcal{E}}[x^\top(\mathbb{E}_W M - \mathbb{E}M)\theta + x^\top\mathbb{E}_W\tilde{M}\mathcal{E}]^2 - V_s - V_l \\ &= \sigma^2\mathbb{E}_X\|\mathbb{E}_W\tilde{M} - \mathbb{E}\tilde{M}\|_F^2. \end{aligned} \tag{35}$$

$$\begin{aligned} V_{li} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{\mathcal{E},W}(\mathbb{E}_X(\hat{f}(x)|\mathcal{E},W)) - V_i - V_l \\ &= \mathbb{E}_{\theta,x,\mathcal{E},W}[x^\top(\mathbb{E}_X M - \mathbb{E}M)\theta + x^\top\mathbb{E}_X\tilde{M}\mathcal{E}]^2 - V_i - V_l \\ &= \sigma^2\mathbb{E}_W\|\mathbb{E}_X\tilde{M} - \mathbb{E}\tilde{M}\|_F^2 \end{aligned} \tag{36}$$

$$\begin{aligned} V_{si} &= \mathbb{E}_{\theta,x}\mathrm{Var}_{X,W}(\mathbb{E}_\tau(\hat{f}(x)|X,W)) - V_s - V_i \\ &= \mathbb{E}_{\theta,x,X,W}[x^\top(M - \mathbb{E}M)\theta]^2 - V_s - V_i \\ &= \frac{\alpha^2}{d}\left(\mathbb{E}\|M\|_F^2 - \mathbb{E}_X\|\mathbb{E}_W M\|_F^2 - \mathbb{E}_W\|\mathbb{E}_X M\|_F^2 + \|\mathbb{E}M\|_F^2\right). \end{aligned} \tag{37}$$

And

$$\begin{aligned} V_{sli} &= \mathrm{Var}(\hat{f}(x)) - (V_s + V_l + V_i + V_{sl} + V_{si} + V_{li}) \\ &= \sigma^2(\mathbb{E}\|\tilde{M}\|_F^2 - \mathbb{E}_W\|\mathbb{E}_X\tilde{M}\|_F^2 - \mathbb{E}_X\|\mathbb{E}_W\tilde{M}\|_F^2 + \|\mathbb{E}\tilde{M}\|^2). \end{aligned} \tag{38}$$

After obtaining the expressions of the variance components, Theorem 2 follows directly by plugging Lemmas 11—17 into equation (32)—(38). ■

Lemma 15 (Behaviour of the Frobenius norm of $\tilde{M}$)

$$\lim_{d\to\infty} \|\mathbb{E}\tilde{M}\|_F^2 = \lim_{d\to\infty} \mathbb{E}_W \|\mathbb{E}_X \tilde{M}\|_F^2 = 0.$$

When X is symmetric, by switching the sign of X, clearly $\mathbb{E}_X \tilde{M} = 0$. This lemma shows that the same result still holds asymptotically when X is not symmetric, and simply has zero-mean entries with finite sixth moment.

Lemma 16 (Behavior of the Frobenius norm of $\mathbb{E}_W \tilde{M}$)

$$\lim_{d\to\infty} \mathbb{E}_X \|\mathbb{E}_W \tilde{M}\|_F^2 = \delta(\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2).$$

Lemma 17 (Behavior of the Frobenius norm of $\mathbb{E}_W M$)

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}_X \|\mathbb{E}_W M\|_F^2 = (1 - 2\tilde{\lambda}\tilde{\theta}_1 + \tilde{\lambda}^2\tilde{\theta}_2).$$

See Appendix B.4 for the proof of the above lemmas.

B.3 Proof of Lemmas 11—14

We first prove these four lemmas under the assumption that X has i.i.d. standard Gaussian entries in Appendix B.3.1—B.3.4 to obtain some heuristics for the formulas. Then in Appendix B.3.5 we generalize the proof to the non-Gaussian case which only requires X to have i.i.d. zero mean unit variance and finite $8+\eta$ moment entries for any $\eta > 0$.

Next, we denote $R = \left(WX^\top XW^\top/n + \lambda I_p\right)^{-1}$ for simplicity. In the Gaussian case, the proof proceeds by moving to the SVD decomposition, and carefully exploiting serveral properties of the normal distribution and the Marchenko-Pastur law. In the general case, we use deterministic equivalents properties for covariance matrices to show that all terms we are concerned with converge to the same limits as in the Gaussian case.

B.3.1 Proof of Lemma 11 (Under Gaussian Assumption)

Proof By definition,

$$\mathbb{E}M = \mathbb{E}M_{X,W}(\lambda) = \mathbb{E}W^\top R \frac{WX^\top X}{n}.$$

Let $V = [W^\top, W_\perp^\top]$ be an orthonormal matrix containing an arbitrary orthogonal complement of $W^\top$. It will be convenient to write this as $W = DV^\top$, where the $p \times d$ matrix $D = \left(\mathrm{I}_{p\times p}, 0_{p\times(d-p)}\right)$ selects the appropriate rows of $V^\top$. Denote $\widetilde{X} := XV$, a matrix of the same size $n \times d$ as the original matrix X. Then,

$$\begin{aligned}
\mathbb{E}M &= \mathbb{E}_{V,X} V D^\top \left(\frac{DV^\top X^\top XVD^\top}{n} + \lambda I_p\right)^{-1} \frac{DV^\top X^\top XVV^\top}{n} \\
&= \mathbb{E}_{V,\widetilde{X}} V D^\top \left(\frac{D\widetilde{X}^\top \widetilde{X} D^\top}{n} + \lambda I_p\right)^{-1} \frac{D\widetilde{X}^\top \widetilde{X} V^\top}{n}.
\end{aligned}$$

By assumption, W is uniformly sampled from the Stiefel manifold, i.e., the manifold of partial orthogonal $p \times d$ ($p \leqslant d$) matrices with orthonormal rows. Thus, we can assume that V is also uniformly distributed over orthogonal matrices (i.e. the Haar measure). Furthermore, since V is orthogonal, we know that $\widetilde{X}$ has independent standard Gaussian entries and is independent of V because XV has the same distribution for any orthogonal matrix V. Noting that $\mathbb{E}_V VAV^\top = \operatorname{tr} A \cdot I_d/d$ when $V \in \mathbb{R}^{d\times d}$ follows the Haar measure, we get

$$\begin{aligned}
\mathbb{E}M &= \mathbb{E}_{\widetilde{X}}\mathbb{E}_V VD^\top \left(\frac{D\widetilde{X}^\top\widetilde{X}D^\top}{n} + \lambda I_p\right)^{-1} \frac{D\widetilde{X}^\top\widetilde{X}V^\top}{n} \\
&= \frac{1}{d}\cdot \mathbb{E}_{\widetilde{X}} \operatorname{tr}\left(D^\top \left(\frac{D\widetilde{X}^\top\widetilde{X}D^\top}{n} + \lambda I_p\right)^{-1} \frac{D\widetilde{X}^\top\widetilde{X}}{n}\right)\cdot I_d \\
&= \frac{1}{d}\cdot \mathbb{E}_{\widetilde{X}} \operatorname{tr}\left(\left(\frac{D\widetilde{X}^\top\widetilde{X}D^\top}{n} + \lambda I_p\right)^{-1} \frac{D\widetilde{X}^\top\widetilde{X}D^\top}{n}\right)\cdot I_d.
\end{aligned}$$

Further defining $\hat{X} := \widetilde{X}D^\top = XW^\top$ which is now of size $n \times p$ (while the original size was $n \times d$), then $\hat{X}$ also has independent standard Gaussian entries. Letting $\hat{X} = \hat{U}\Gamma\hat{V}^\top$ be the SVD decomposition of $\hat{X}$, we have

$$\mathbb{E}M = \frac{1}{d}\cdot \mathbb{E}_{\hat{X}} \operatorname{tr}\left(\left(\frac{\hat{X}^\top\hat{X}}{n} + \lambda I_p\right)^{-1} \frac{\hat{X}^\top\hat{X}}{n}\right)\cdot I_d = \frac{p}{d}\left(1 - \frac{\lambda}{p}\mathbb{E}_\Gamma \operatorname{tr}\left(\frac{\Gamma^\top\Gamma}{n} + \lambda I_p\right)^{-1}\right)\cdot I_d.$$

This is determined by the spectral measure of $n^{-1}\hat{X}^\top\hat{X}$. Since $\hat{X}$ has independent standard normal entries and $\lim_{d\to\infty} p/n = \pi\delta$, applying the Marchenko-Pastur theorem (Marchenko and Pastur, 1967; Silverstein, 1995; Bai and Silverstein, 2010), we get

$$\begin{aligned}
&\frac{1}{d}\mathbb{E}\operatorname{tr}(M) \to \pi\left(1 - \lambda\int \frac{1}{(x+\lambda)} dF_{\pi\delta(x)}\right) = \pi(1 - \lambda\theta_1(\pi\delta, \lambda)), \\
&\frac{1}{d}\|\mathbb{E}M\|_F^2 \to \pi^2(1-\lambda\theta_1)^2.
\end{aligned}$$

This finishes the proof. ■

B.3.2 Proof of Lemma 12 (Under Gaussian Assumption)

Proof By definition of M,

$$\begin{aligned}
\mathbb{E}\|M\|_F^2 &= \mathbb{E}\left\|W^\top R\frac{WX^\top X}{n}\right\|_F^2 \\
&= \mathbb{E}\operatorname{tr}\left(W^\top R\frac{WX^\top X}{n}\frac{X^\top XW^\top}{n}R^\top W\right) \\
&= \mathbb{E}\operatorname{tr}\left(R\frac{WX^\top XX^\top XW^\top}{n^2}R^\top\right).
\end{aligned}$$

Let $W_\perp = f(W) \in \mathbb{R}^{(d-p)\times d}$ be an orthogonal complement of W and define $X_1 := XW^\top$, $X_2 := XW_\perp^\top$. Since X has Gaussian entries, X_1 and X_2 both have Gaussian entries. Since

$$\mathbb{E}X_1^\top X_2 = \mathbb{E}WX^\top XW_\perp^\top = n \cdot \mathbb{E}WW_\perp^\top = 0,$$

it follows that X_1 and X_2 are independent. Noting that $XX^\top = X_1X_1^\top + X_2X_2^\top$, we have

$$\begin{aligned}\mathbb{E}\|M\|_F^2 &= \mathbb{E}\,\mathrm{tr}\left(R\frac{WX^\top X_1X_1^\top XW^\top}{n^2}R^\top\right) + \mathbb{E}\,\mathrm{tr}\left(R\frac{WX^\top X_2X_2^\top XW^\top}{n^2}R^\top\right).\\ &=: \mathrm{C}_1 + \mathrm{C}_2.\end{aligned}$$

For the first term, we have

$$C_1 = \mathbb{E}\,\mathrm{tr}\left(R\frac{WX^\top X_1X_1^\top XW^\top}{n^2}R^\top\right) = \mathbb{E}\,\mathrm{tr}\left(R\frac{X_1^\top X_1X_1^\top X_1}{n^2}R^\top\right).$$

Let $X_1 = U\Gamma_1V^\top$ be the singular value decomposition of X_1. By plugging in the definition of R, we obtain

$$C_1 = \mathbb{E}\,\mathrm{tr}\left[\left(\frac{\Gamma_1^\top\Gamma_1}{n} + \lambda I_p\right)^{-2}\left(\frac{\Gamma_1^\top\Gamma_1}{n}\right)^2\right].$$

Thus, according to the Marchenko-Pastur theorem,

$$\frac{\alpha^2}{d}C_1 \to \alpha^2\pi\int\frac{x^2}{(x+\lambda)^2}dF_{\pi\delta}(x) = \alpha^2\pi(1-2\lambda\theta_1+\lambda^2\theta_2).$$

For the second term, since X_1 and X_2 are indepedent and noting that

$$\mathbb{E}_{X_2}X_2^\top X_2 = \mathbb{E}_{X,W}(XW_\perp^\top W_\perp X^\top) = \mathbb{E}_W(\mathrm{tr}(I_d - W^\top W))I_d = (d-p)I_d,$$

we have

$$\begin{aligned}C_2 &= \mathbb{E}\,\mathrm{tr}\left(R\frac{WX^\top X_2X_2^\top XW^\top}{n^2}R^\top\right) = \mathbb{E}_{X_1}\mathbb{E}_{X_2}\,\mathrm{tr}\left(R\frac{X_1^\top X_2X_2^\top X_1}{n^2}R^\top\right)\\ &= \mathbb{E}_{X_1}\,\mathrm{tr}\left(R\frac{X_1^\top\mathbb{E}_{X_2}(X_2X_2^\top)X_1}{n^2}R^\top\right) = \frac{d-p}{n}\mathbb{E}_{X_1}\,\mathrm{tr}\left(R\frac{X_1^\top X_1}{n}R^\top\right).\end{aligned}$$

Since

$$\mathbb{E}_{X_1}\,\mathrm{tr}\left(R\frac{X_1^\top X_1}{n}R^\top\right) = \mathbb{E}\,\mathrm{tr}\left[\left(\frac{\Gamma_1^\top\Gamma_1}{n} + \lambda I_p\right)^{-2}\frac{\Gamma_1^\top\Gamma_1}{n}\right],$$

by the Marchenko-Pastur theorem, $\frac{\alpha^2}{d}C_2 \to \alpha^2(1-\pi)\pi\delta\int\frac{x}{(x+\lambda)^2}dF_{\pi\delta}(x) = \alpha^2(1-\pi)\pi\delta(\theta_1 - \lambda\theta_2)$. Finally, combining the results for C_1 and C_2 gives

$$\frac{\alpha^2}{d}\mathbb{E}\|M\|_F^2 = \frac{\alpha^2}{d}(C_1+C_2) \to \alpha^2\pi\left[1-2\lambda\theta_1+\lambda^2\theta_2 + (1-\pi)\delta(\theta_1-\lambda\theta_2)\right],$$

and this finishes the proof. ■

B.3.3 Proof of Lemma 13 (Under Gaussian Assumption)

Proof By definition, we have

$$\begin{aligned}\mathbb{E}\|\tilde{M}\|_F^2 &= \mathbb{E}\,\mathrm{tr}\left(W^\top R\frac{WX^\top}{n}\right)\left(W^\top R\frac{WX^\top}{n}\right)^\top \\ &= \mathbb{E}\,\mathrm{tr}\left(W^\top R\frac{WX^\top}{n}\frac{XW^\top}{n}R^\top W\right) = \mathbb{E}\,\mathrm{tr}\left(R\frac{WX^\top XW^\top}{n^2}R^\top\right).\end{aligned}$$

Denote $XW^\top$ by X_1 and write $X_1 = U\Gamma V^\top$ for the SVD of X_1. Then,

$$\begin{aligned}\mathbb{E}\|\tilde{M}\|_F^2 &= \frac{1}{n}\mathbb{E}\,\mathrm{tr}\left(R\frac{X_1^\top X_1}{n}R^\top\right) = \frac{1}{n}\mathbb{E}\,\mathrm{tr}\left(R-\lambda R^2\right) \\ &= \frac{1}{n}\mathbb{E}\,\mathrm{tr}\left[(\frac{\Gamma^\top\Gamma}{n}+\lambda)^{-1} - \lambda(\frac{\Gamma^\top\Gamma}{n}+\lambda)^{-2}\right] \to \pi\delta(\theta_1-\lambda\theta_2),\end{aligned}$$

where the last line follows from the Marchenko-Pastur theorem and that $p/n \to \pi\delta$. ■

B.3.4 Proof of Lemma 14 (Under Gaussian Assumption)

Proof Similarly, let $W_\perp = f(W) \in \mathbb{R}^{(d-p)\times d}$ be an orthogonal complement of W. Denoting $X_1 = XW^\top$, $X_2 = XW_\perp^\top$ and combining the fact that X_1 and X_2 are independent, with $\mathbb{E}X_2 = 0$, we have

$$\begin{aligned}\mathbb{E}_X M &= \mathbb{E}_X W^\top R\frac{WX^\top X}{n} \\ &= W^\top\mathbb{E}_X\left(\frac{X_1^\top X_1}{n}+\lambda I_p\right)^{-1}\frac{X_1^\top(X_1W+X_2W_\perp)}{n} \\ &= W^\top\mathbb{E}_{X_1}\left[\left(\frac{X_1^\top X_1}{n}+\lambda I_p\right)^{-1}\frac{X_1^\top X_1}{n}\right]W \qquad (39)\end{aligned}$$

Write the SVD of X_1 as $X_1 = U\Gamma V^\top$ and note that $V \in \mathbb{R}^{p\times p}$ is uniformly distributed over the set of orthogonal matrices. Then the above equals

$$W^\top\left[I_p - \lambda\mathbb{E}_{\Gamma,V}V\left(\frac{\Gamma^\top\Gamma}{n}+\lambda I_p\right)^{-1}V^\top\right]W = W^\top W\left[1-\frac{\lambda}{p}\mathbb{E}_\Gamma\,\mathrm{tr}\left(\frac{\Gamma^\top\Gamma}{n}+\lambda I_p\right)^{-1}\right].$$

Thus,

$$\begin{aligned}\lim_{d\to\infty}\frac{\alpha^2}{d}\mathbb{E}_W\|\mathbb{E}_X M\|_F^2 &= \lim_{d\to\infty}\frac{\alpha^2}{d}\left[1-\frac{\lambda}{p}\mathbb{E}_\Gamma\,\mathrm{tr}\left(\frac{\Gamma^\top\Gamma}{n}+\lambda I_p\right)^{-1}\right]^2\mathbb{E}_W\,\mathrm{tr}(W^\top W) \\ &= \lim_{d\to\infty}\frac{\alpha^2 p}{d}\left[1-\frac{\lambda}{p}\mathbb{E}_\Gamma\,\mathrm{tr}\left(\frac{\Gamma^\top\Gamma}{n}+\lambda I_p\right)^{-1}\right]^2 \\ &\to \alpha^2\pi(1-\lambda\theta_1)^2,\end{aligned}$$

where the last line follows directly from the Marchenko-Pastur theorem. ■

B.3.5 Proof of Lemma 11—14 (General Case)

In Appendix B.3.1—B.3.4, we have proved Lemma 11—14 under the assumption that the entries of X are i.i.d. standard Gaussian. In this part, we will generalize previous proofs to the non-Gaussian case, i.e., X has i.i.d. zero mean, unit variance entries with finite $8+\eta$ moment, and hence complete the proof of Lemma 11—14.

For simplicity, we only present the proof of Lemmas 12, 13 in non-Gaussian case. Lemmas 11, 14 can be proved using very similar arguments as Lemma 12.

We recall the *calculus of deterministic equivalents* from random matrix theory, which will be used in our proof (Dobriban and Sheng, 2018, 2020). One of the best ways to understand the Marchenko-Pastur law is that *resolvents are asymptotically deterministic*. Let $\widehat{\Sigma} = n^{-1}X^\top X$, where $X = Z\Sigma^{1/2}$ and Z is an $n \times p$ random matrix with iid entries of zero mean and unit variance, and $\Sigma^{1/2}$ is any sequence of $p \times p$ positive semi-definite matrices. We take $n, p, q \to \infty$ proportionally.

We say that the (deterministic or random) not necessarily symmetric matrix sequences A_n, B_n of growing dimensions are *equivalent*, and write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} |\mathrm{tr}\,[C_n(A_n - B_n)]| = 0 \tag{40}$$

almost surely, for any sequence C_n of not necessarily symmetric matrices with bounded trace norm, i.e., such that

$$\limsup \|C_n\|_{tr} < \infty.$$

We call such a sequence C_n a *standard sequence*. Recall here that the trace norm (or nuclear norm) is defined by $\|M\|_{tr} = \mathrm{tr}((M^\top M)^{1/2}) = \sum_i \sigma_i$, where σ_i are the singular values of M.

Moreover, if (40) only holds almost surely for any sequence $C_n \in \mathbb{R}^{d_n \times d_n}$ of positive semidefinite matrices with $O(1/d_n)$ spectral norm, A_n and B_n are said to be weak deterministic equivalents and denoted by $A_n \overset{w}{\asymp} B_n$. It is readily verified that deterministic equivalence implies weak deterministic equivalence.

By the general Marchenko-Pastur (MP) theorem of Rubio and Mestre (Rubio and Mestre, 2011), we have that for any $\lambda > 0$

$$(\widehat{\Sigma} + \lambda I)^{-1} \asymp (q_p\Sigma + \lambda I)^{-1},$$

where q_p is the unique positive solution of the fixed point equation

$$1 - q_p = \frac{q_p}{n}\,\mathrm{tr}\left[\Sigma(q_p\Sigma + \lambda I)^{-1}\right].$$

When $n, p \to \infty$ and the sepctral distribution of Σ converges to H, $q_p \to q$ and q satisfies the equation

$$1 - q = \gamma\left[1 - \lambda\int_0^\infty \frac{dH(t)}{qt + \lambda}\right].$$

We now proceed with the proof.

Proof [Proof of Lemma 12 (general case)]By definition, recalling that $R = (\frac{WX^\top XW^\top}{n} + \lambda I_p)^{-1}$,

$$\mathbb{E}M = \mathbb{E}W^\top R\frac{WX^\top X}{n} = \mathbb{E}W^\top WX^\top \tilde{R}\frac{X}{n}. \tag{41}$$

Therefore, letting $\tilde{R} = (\frac{XW^\top WX^\top}{n} + \lambda I_n)^{-1}$ be the resolvent obtained in the other order,

$$\mathbb{E}\operatorname{tr}(MM^\top) = \mathbb{E}\operatorname{tr}\left[W^\top WX^\top \tilde{R}\frac{XX^\top}{n^2}\tilde{R}XW^\top W\right].$$

Define the regularized resolvent $\tilde{R}_\tau = \left(\frac{X(W^\top W+\tau)X^\top}{n} + \lambda I_n\right)^{-1}$ and

$$M_\tau := W^\top WX^\top \tilde{R}_\tau \frac{X}{n}.$$

Since we have already proved Lemma 12 under the Gaussian assumption, to generalize the results into the non-Gaussian case, we only need to prove the following two steps:

$$\text{(1).}\quad \lim_{d\to\infty}\mathbb{E}\operatorname{tr}(MM^\top)/d = \lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\operatorname{tr}M_\tau M_\tau^\top/d \quad \text{(assuming the limits exist)} \tag{42}$$

$$\text{(2).}\quad \lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\operatorname{tr}M_\tau M_\tau^\top/d \quad \text{exists and is a constant independent of the distribution of } X. \tag{43}$$

(1). Note that

$$\begin{aligned}
\Delta &:= \lim_{\tau\to 0}\lim_{d\to\infty}\frac{1}{d}|\mathbb{E}\operatorname{tr}(MM^\top) - \mathbb{E}\operatorname{tr}M_\tau M_\tau^\top| \\
&\le \lim_{\tau\to 0}\lim_{d\to\infty}\frac{1}{d}|\mathbb{E}\operatorname{tr}(M(M^\top - M_\tau^\top))| + \frac{1}{d}\left|\mathbb{E}\operatorname{tr}[(M-M_\tau)M_\tau^\top]\right| \\
&\le \lim_{\tau\to 0}\lim_{d\to\infty}\frac{1}{d}|\mathbb{E}\operatorname{tr}(M(M^\top - M_\tau^\top))| + \frac{1}{d}\left|\mathbb{E}\operatorname{tr}[M_\tau(M^\top - M_\tau^\top)]\right| \\
&=: \tilde{\Delta}_1 + \tilde{\Delta}_2.
\end{aligned}$$

Therefore it suffices to prove $\tilde{\Delta}_{1,2} \to 0$. For $\tilde{\Delta}_1$, we have the following argument. Denote $X^\top\tilde{R}X/n$ by A_0, $X^\top\tilde{R}_\tau X/n$ by A_τ and $W^\top W$ by P. Note that $M = PA_0$, and

$$\begin{aligned}
\tilde{\Delta}_1 &= \lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\frac{\tau}{d}\left|\operatorname{tr}\left(MA_0A_\tau P\right)\right| \\
&\le \lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\frac{\tau}{2d}[\|M\|_F^2 + \|A_0A_\tau P\|_F^2] \\
&\le \lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\frac{\tau}{2d}[\operatorname{tr}(A_0^2) + \operatorname{tr}(A_0A_\tau^2A_0)] \\
&\le \lim_{\tau\to 0}\lim_{d\to\infty}\tau\mathbb{E}\frac{1}{2d}\left[\frac{1}{\lambda^2}\operatorname{tr}\left(\frac{X^\top X}{n}\right)^2 + \frac{1}{\lambda^4}\operatorname{tr}\left(\frac{X^\top X}{n}\right)^4\right] \\
&\le \lim_{\tau\to 0} O(\tau) = 0,
\end{aligned} \tag{44}$$

where the second line follows from the properties of the Frobenius norm. In the third and fourth line we use $A_0, A_\tau \preceq X^\top X/n\lambda$ and the fact that

$$\mathrm{tr}(M_1M_2M_1) \preceq \mathrm{tr}(M_1M_3M_1) \qquad \forall M_1, M_2, M_3 \quad \text{positive semi-definite and} \quad M_2 \preceq M_3.$$

Finally, the last line is due to the finite $8+\eta$ moment assumption and some direct caclulations. Using the same techinique, it is not hard to show that $\tilde{\Delta}_2$ also converges to zero, and hence we conclude the proof of (1).

(2). We first give an alternative expression for M_τ. Denote $(W^\top W + \tau)$ by S and let $V = U^\top = X/\sqrt{n}$, $C = I_n/\lambda$ and $A = (W^\top W + \tau)^{-1}$ in the Woodbury identity:

$$\begin{aligned} & (A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\ \iff \quad & U(C^{-1} + VA^{-1}U)^{-1}V = A - A^{1/2}(I + A^{-1/2}UCVA^{-1/2})^{-1}A^{1/2}. \end{aligned} \tag{45}$$

We have by left multiplying $W^\top W$ in (45) that, with $R_S := (\lambda I_d + \frac{S^{1/2}X^\top X S^{1/2}}{n})^{-1}$

$$M_\tau = W^\top W S^{-1/2}\left[I_d - \lambda R_S\right] S^{-1/2}. \tag{46}$$

Fix τ and suppose that the spectral distribution of R converges in distribution to H_τ as $d \to \infty$. Since $W^\top W$ has p eigenvalues 1 and $d-p$ eigenvalues 0 and $S = W^\top W + \tau$, it is clear that $H_\tau = \pi\delta_{1+\tau} + (1-\pi)\delta_\tau$. Also, we have by theorem 1 in Rubio and Mestre (2011) and some simple calculations, that

$$R_S \asymp (x_d S + \lambda I_d)^{-1}, \tag{47}$$

where x_d is the unique positive solution of the fixed point equation (where we omit x_d's dependence on τ for notational simplicity)

$$1 - x_d = \frac{x_d}{n}\,\mathrm{tr}[S(x_d S + \lambda I_d)^{-1}].$$

When $n, p, d \to \infty$ proportionally, $x_d \to x$ and x satisfies the equation (again omitting x's dependence on τ for notational simplicity)

$$1 - x = \delta\left(1 - \lambda\int_0^\infty \frac{dH_\tau(t)}{xt+\lambda}\right).$$

Now, we start to calculate $\mathbb{E}\,\mathrm{tr}(M_\tau M_\tau^\top)$. By definition,

$$\begin{aligned} \mathbb{E}\,\mathrm{tr}(M_\tau M_\tau^\top) &= \mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-1}\left(I_d - \lambda R_S\right) S^{-1}\left(I_d - \lambda R_S\right)\right] \\ &=: \Delta_1 + \Delta_2 + \Delta_3, \end{aligned}$$

where

$$\begin{aligned} \Delta_1 &:= \mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-2}\right] \\ \Delta_2 &:= -2\lambda\mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-2} R_S\right] \\ \Delta_3 &:= \lambda^2\mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-1} R_S S^{-1} R_S\right]. \end{aligned}$$

Since $W^\top W$ has p eigenvalues equal to 1 and $d-p$ eigenvalues equal to 0,

$$\lim_{\tau\to 0}\lim_{d\to\infty}\Delta_1/d = \lim_{\tau\to 0}\lim_{d\to\infty}\frac{p}{d(1+\tau)^2} = \pi. \tag{48}$$

Since $\|W^\top W S^{-2}\|_2 \le 1$, using the deterministic equivalent property (47) and the bounded convergence theorem, we get

$$\begin{aligned}
\lim_{\tau\to 0}\lim_{d\to\infty}\Delta_2/d &= -2\lambda\lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-2} R_S\right] \\
&= -2\lambda\lim_{\tau\to 0}\lim_{d\to\infty}\mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-2}\left(x_d S + \lambda I_d\right)^{-1}\right] \\
&= -2\lambda\lim_{\tau\to 0}\lim_{d\to\infty}\frac{p}{d}\frac{1}{[(1+\tau)^2[x_d(1+\tau)+\lambda]} \\
&= -2\lambda\lim_{\tau\to 0}\frac{\pi}{[(1+\tau)^2[x_\tau(1+\tau)+\lambda]} = -2\lambda\lim_{\tau\to 0}\frac{\pi}{x_\tau+\lambda},
\end{aligned} \tag{49}$$

where the third line follows from the fact that W and $S = (W^\top W + \tau)$ are simultaneously diagonizable and $W^\top W$ has p eigenvalues 1 and $d-p$ eigenvalues 0. It can be verified that the limit in (49) exists and is a constant independent of the distribution of X. Now, it remains to prove that $\lim_{\tau\to 0}\lim_{d\to\infty}\Delta_3/d$ converges to a constant limit. Note that

$$\begin{aligned}
&\lim_{\tau\to 0}\lim_{d\to\infty}\Delta_3/d = \lim_{\tau\to 0}\lim_{d\to\infty}\frac{\lambda^2}{d}\mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-1} R_S S^{-1} R_S\right] \\
&= \lim_{\tau\to 0}\lim_{d\to\infty}\frac{\lambda^2}{d}\mathbb{E}\,\mathrm{tr}\left[W^\top W S^{-1/2} R_S S^{-1} R_S S^{-1/2}\right] \\
&= \lim_{\tau\to 0}\lim_{d\to\infty}\frac{\lambda^2}{d}\mathbb{E}\,\mathrm{tr}\left[W^\top W\left(\lambda\tau + \lambda W^\top W + \frac{SX^\top XS}{n}\right)^{-1}\left(\lambda\tau + \lambda W^\top W + \frac{SX^\top XS}{n}\right)^{-1}\right].
\end{aligned} \tag{50}$$

For any $z \in E := \mathbb{C}\backslash\mathbb{R}^+$, we have by Lemma 18 that

$$\left(\lambda W^\top W + \frac{SX^\top XS}{n} - zI_d\right)^{-2} \asymp -(\lambda W^\top W + x_d S^2 - zI_d)^{-2}(x_d'(z)S^2 - I_d), \tag{51}$$

where for any $z \in E$, $x_d(z)$ is the unique solution of certain fixed point equation independent of X, and $x_d'(z) := dx_d(z)/dz$. Furthermore, there exists $x(z)$ such that $x_d(z) \to x(z)$, $x_d'(z) \to x'(z)$. Now, letting $z = -\lambda\tau$ and replacing $(\lambda W^\top W + SX^\top XS/n - zI_d)^{-2}$ by its deterministic equivalent (51), we have from (50) and the bounded convergence theorem that

$$\begin{aligned}
\lim_{\tau\to 0}\lim_{d\to\infty}\Delta_3/d &= \lim_{\tau\to 0}\lim_{d\to\infty}\frac{\lambda^2}{d}\mathbb{E}\,\mathrm{tr}\left[W^\top W\left(\lambda\tau I_d + \lambda W^\top W + \frac{SX^\top XS}{n}\right)^{-2}\right] \\
&= \lim_{\tau\to 0}\lim_{d\to\infty}\frac{-\lambda^2}{d}\mathbb{E}\,\mathrm{tr}\left[W^\top W(\lambda W^\top W + x_d S^2 + \lambda\tau I_d)^{-2}(x_d' S^2 - I_d)\right] \\
&= \lim_{\tau\to 0}\lim_{d\to\infty} -\pi\lambda^2\frac{x_d'(1+\tau)^2 - 1}{(\lambda + x_d(1+\tau)^2 + \lambda\tau)^2} \\
&= \lim_{\tau\to 0}\pi\lambda^2\frac{1 - x'(1+\tau)^2}{(\lambda + x(1+\tau)^2 + \lambda\tau)^2} = \lim_{\tau\to 0}\pi\lambda^2\frac{1-x'}{(\lambda+x)^2},
\end{aligned} \tag{52}$$

where $x := x(-\lambda\tau), x' := x'(-\lambda\tau)$. Since in the Gaussian case the limit of the L.H.S. of (42) exists, by the proof of (42), we know that the limit in (52) also exists, and does not depend of X. This concludes the proof of (2). ■

Lemma 18 (Second order deterministic equivalent) *Suppose that $X \in \mathbb{R}^{n\times d}$ has i.i.d. zero mean, unit variance entries with finite $8+\eta$ moment. Then for any $z \in C\backslash\mathbb{R}^+$,*

$$\left(\lambda W^\top W + \frac{SX^\top XS}{n} - zI_d\right)^{-2} \asymp -(\lambda W^\top W + x_d S^2 - zI_d)^{-2}(x_d'(z)S^2 - I_d), \tag{53}$$

where $x_d(z)$ is the unique solutions of a certain fixed point equation independent of X, and $x_d'(z) := dx_d(z)/dz$. Furthermore, there exists $x(z)$ such that $x_d(z) \to x(z)$, $x_d'(z) \to x'(z)$.

Proof [Sketch of the proof] Since this lemma can be proved following the same steps as the proof of theorem 3.1 (b) in Dobriban and Sheng (2020), here we only provide a sketch of the proof.
Step 1. (First order deterministic equivalent) Denote

$$f(z, W) := (\lambda W^\top W + x_d S^2 - zI_d)^{-1}$$
$$g(z, W, X) := \left(\lambda W^\top W + \frac{SX^\top XS}{n} - zI_d\right)^{-1},$$

where $x_d(z)$ is the unique solution of the fixed point equation

$$1 - x_d = \frac{x_d}{n}\operatorname{tr}[S^2(\lambda W^\top W + x_d S^2 - zI_d)^{-1}].$$

Then we have from theorem 1 in Rubio and Mestre (2011) that (here we need the finite $8+\eta$ moment assumption)

$$f(z, W) \asymp g(z, W, X).$$

Furthermore, for a sequence of W and fixed τ, z, when $n, p, d \to \infty$ proportionally, it can be verified that $x_d(z) \to x(z)$ and $x(z)$ satisfies the fixed point equation

$$1 - x = \delta\left(1 - \int_0^\infty \frac{[\lambda(t-\tau) - z]dH_\tau(t)}{xt^2 + \lambda(t-\tau) - z}\right).$$

Step 2. (Second order deterministic equivalent) Using the same technique as in the proof of theorem 3.1 (b) in Dobriban and Sheng (2020), it can be proved that

$$f'(z, W) \asymp g'(z, W, X).$$

for all $z \in E$.

Step 3. (Explicit expressions for the derivatives) For invertible $A(z)$, we have

$$\frac{dA}{dz} = -A^{-1}\frac{dA}{dz}A^{-1}.$$

Therefore

$$\begin{aligned} f'(z,W) &= -(\lambda W^\top W + x_d S^2 - zI_d)^{-1}(x_d'(z)S^2 - I_d)(\lambda W^\top W + x_d S^2 - zI_d)^{-1} \\ &= -(\lambda W^\top W + x_d S^2 - zI_d)^{-2}(x_d'(z)S^2 - I_d) \end{aligned}$$

and

$$g'(z,W,X) = \left(\lambda W^\top W + \frac{SX^\top XS}{n} - zI_d\right)^{-2}.$$

Also, it can be shown that $x_d'(z) \to x'(z)$ as in the proof of theorem 3.1 (b) in Dobriban and Sheng (2020). ■

Similarly, we can prove Lemma 11 and 14 in the general case via the two steps in (42), (43). Since the proofs are almost the same as Lemma 12 (and in fact even simpler), we omit them for simplicity. Also, here we only mention one difference. When bounding Δ_1 from (44), we need to first replace M by $\mathbb{E}M$ (or $\mathbb{E}_X M$) and move the expectation outside the trace operator, e.g. in the proof of 11,

$$\begin{aligned} \Delta_1 &:= \lim_{\tau\to 0}\lim_{d\to\infty}\frac{1}{d}\left|\mathrm{tr}[\mathbb{E}M(\mathbb{E}M^\top - \mathbb{E}M_\tau^\top)]\right| \\ &\le \lim_{\tau\to 0}\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\left|\mathrm{tr}[\mathbb{E}M(M^\top - M_\tau^\top)]\right|. \end{aligned}$$

Then all results follow the same argument as in (1) of the proof of Lemma 12.

Proof [Proof of Lemma 13 (general case)] By definition, we have

$$\begin{aligned} \mathbb{E}\|\tilde{M}\|_F^2 &= \mathbb{E}\,\mathrm{tr}\left\|W^\top R\frac{WX^\top}{n}\right\|_F^2 \\ &= \mathbb{E}\,\mathrm{tr}\left[R\frac{WX^\top XW^\top}{n}R\right] = \mathbb{E}\frac{1}{n}\,\mathrm{tr}\left[R - \lambda R^2\right]. \end{aligned} \tag{54}$$

Denote $WX^\top XW^\top/n$ by Q_1 and $(W^\top W)^{1/2}X^\top X(W^\top W)^{1/2}/n$ by Q_2. Since Q_1 and Q_2 have the same non-zero eigenvalues, their Limiting Spectral Distributions (LSD) (if one of them exists) only differ from a constant scale and a mass at 0, i.e.,

$$LSD(Q_2) = \pi LSD(Q_1) + (1-\pi)\delta_0.$$

Since the LSD of $W^\top W$ converges to $\pi\delta_1 + (1-\pi)\delta_0$ almost surely, from Silverstein (1995) theorem 1.1, we know that the LSD of Q_2 almost surely weakly converges to a nonrandom distribution. Therefore, the LSD of Q_1 also almost surely weakly converges to a nonrandom distribution and $\frac{1}{n}\,\mathrm{tr}[WX^\top XW^\top/n + \lambda I_p]^{-i}$ for $i = 1, 2$ almost surely converge to some nonrandom limits. Thus, by the bounded convergence theorem, we know (54) converges to a nonrandom limit independent of the exact distribution of X (which only requires X to have i.i.d. zero mean, unit variance entries). Since we have proved Lemma 13 in the Gaussian case, the general case follows directly. ■

B.4 Proof of Lemmas 15—17

B.4.1 Proof of Lemma 15

Proof It is clear that $\|\mathbb{E}\tilde{M}\|_F^2 \leqslant \mathbb{E}_W\|\mathbb{E}_X\tilde{M}\|_F^2$. Thus we only need to prove the second convergence. By definition,

$$\mathbb{E}_X\tilde{M} = \mathbb{E}_X W^\top R\frac{WX^\top}{n}.$$

Since X has i.i.d. rows, by switching the i-th and j-th row of X, it is readily verified that $\mathbb{E}_X\tilde{M} = \mathbb{E}_X W^\top(n^{-1}WX^\top XW^\top + \lambda I_p)^{-1}WX^\top/n$ has identically distributed columns. Note that $\mathbb{E}_X\tilde{M}$ is a $d \times n$ matrix, it is thus enough to prove:

$$n\|\mathbb{E}_X\tilde{M}_{\cdot 1}\|_2^2 = n\left\|\mathbb{E}_X W^\top R\frac{W(X_{1\cdot})^\top}{n}\right\|_2^2 \xrightarrow{u} 0, \tag{55}$$

where $\xrightarrow{u}$ denotes convergence uniformly in W. For notational simplicity, we denote the column vector $W(X_{1\cdot})^\top$ by $\tilde{x}$ (formed by taking the first row $X_{1\cdot}$ of X), $X_{-1\cdot}W^\top$ by $\tilde{X}$ (formed by taking the complement of the first row $X_{1\cdot}$ of X) and $(\widetilde{X}^\top\widetilde{X}/n + \lambda I_p)$ by C. Let also $F = C^{-1}\tilde{x}\tilde{x}^\top C^{-1}/n$. Then

$$\begin{aligned}
(55) &= \frac{1}{n}\left\|\mathbb{E}_X W^\top\left(\lambda I_p + \frac{\widetilde{X}^\top\widetilde{X}}{n} + \frac{\tilde{x}\tilde{x}^\top}{n}\right)^{-1}\tilde{x}\right\|_2^2 \\
&= \frac{1}{n}\left\|\mathbb{E}_X W^\top\left(C^{-1} - \frac{F}{1+\tilde{x}^\top C^{-1}\tilde{x}/n}\right)\tilde{x}\right\|_2^2 \\
&= \frac{1}{n}\left\|\mathbb{E}_X W^\top\left(\frac{F}{1+f}\right)\tilde{x}\right\|_2^2,
\end{aligned} \tag{56}$$

where in the second line we used the Sherman-Morrison formula, and the last line follows from the fact that C^{-1} and $\tilde{x}$ are independent and $\mathbb{E}\tilde{x} = 0$. Let us denote $f = \tilde{x}^\top C^{-1}\tilde{x}/n$. To prove that (56) $\xrightarrow{u} 0$, we only need to prove the following:

$$(1). \qquad \frac{1}{n}\left\|\mathbb{E}_X W^\top\left(\frac{F}{\mathbb{E}_{\tilde{x}}(1+f)}\tilde{x}\right)\right\|_2^2 \xrightarrow{u} 0 \tag{57}$$

$$(2). \qquad \frac{1}{n}\left\|\mathbb{E}_X W^\top\left(\frac{F}{1+f}\tilde{x}\right) - \mathbb{E}_X W^\top\left(\frac{F}{\mathbb{E}_{\tilde{x}}(1+f)}\tilde{x}\right)\right\|_2^2 \xrightarrow{u} 0. \tag{58}$$

(1). For any fixed W, since $\tilde{x}$ and C^{-1} are independent,

$$\mathbb{E}_{\tilde{x}}\frac{\tilde{x}^\top C^{-1}\tilde{x}}{n} = \mathbb{E}_{\tilde{x}}\frac{x^\top(W^\top C^{-1}W)x}{n} = \frac{\mathrm{tr}(W^\top C^{-1}W)}{n} = \frac{\mathrm{tr}(C^{-1})}{n}. \tag{59}$$

Denote $1+\mathrm{tr}(C^{-1})/n$ by c_0, and $W^\top C^{-1}W$ by $\tilde{C}$. Then

$$\begin{aligned}
&\frac{1}{n}\left|\mathbb{E}_X\left(W^\top\frac{F}{\mathbb{E}_{\tilde{x}}(1+f)}\tilde{x}\right)_i\right|^2=\frac{1}{n^3}\left(\mathbb{E}_X\frac{1}{c_0}e_i\tilde{C}xx^\top\tilde{C}x\right)^2\\
&=\frac{1}{n^3}\left(\mathbb{E}_X\sum_{j,k,l=1}^{d}\frac{1}{c_0}e_i\tilde{C}_{ij}x_jx_k\tilde{C}_{kl}x_l\right)^2=\frac{1}{n^3}\left(\mathbb{E}_{\widetilde{X}}\sum_{j=1}^{d}\frac{1}{c_0}\tilde{C}_{ij}\tilde{C}_{jj}\mathbb{E}x_j^3\right)^2\\
&\le\frac{(\mathbb{E}X_{11}^3)^2}{n^3}\left(\mathbb{E}_{\widetilde{X}}\sum_{j=1}^{d}\tilde{C}_{ij}^2\cdot\sum_{j=1}^{d}\tilde{C}_{jj}^2\right),
\end{aligned}$$

where the last line follows from the Jensen inequality, Cauchy-Schwartz inequality and the fact that $c_0\ge 1$. Summing up all coordinates and noting that $0\preceq\tilde{C}\preceq I_d/\lambda$, we get

$$\text{L.H.S. of (57)}\ \le\frac{(\mathbb{E}X_{11}^3)^2}{n^3}\mathbb{E}_{\widetilde{X}}\left(\|\tilde{C}\|_F^2\cdot\sum_{j=1}^{d}\tilde{C}_{j,j}^2\right)\le\frac{(\mathbb{E}X_{11}^3)^2}{n^3}\cdot\frac{d}{\lambda^2}\cdot\frac{d}{\lambda^2}\le\frac{(\mathbb{E}X_{11}^3)^2d^2}{\lambda^4n^3}\xrightarrow{u}0.$$

(2). By definition, theL.H.S. of (58) equals

$$\begin{aligned}
&\frac{1}{n}\left\|\mathbb{E}_XW^\top\left[\left(\frac{F}{1+f}\right)-\left(\frac{F}{\mathbb{E}_{\tilde{x}}(1+f)}\right)\right]\tilde{x}\right\|_2^2=\frac{1}{n^3}\left\|\mathbb{E}_XW^\top\left[\frac{C^{-1}\tilde{x}\tilde{x}^\top C^{-1}(f-\mathbb{E}_{\tilde{x}}f)}{(1+f)(\mathbb{E}_{\tilde{x}}1+f)}\right]\tilde{x}\right\|_2^2\\
&\le\frac{1}{n^3}\mathbb{E}_X\left\|W^\top C^{-1}\tilde{x}\tilde{x}^\top C^{-1}\tilde{x}\right\|_2^2\cdot\mathbb{E}_X\left[\frac{(f-\mathbb{E}_{\tilde{x}}f)}{(1+f)(\mathbb{E}_{\tilde{x}}1+f)}\right]^2\\
&\le\frac{1}{n^3}\mathbb{E}_X\left\|\tilde{C}xx^\top\tilde{C}x\right\|_2^2\cdot\mathbb{E}_{\widetilde{X}}\mathrm{Var}_x(x^\top\tilde{C}x/n)
\end{aligned}$$

For the first term in the last line, note that $0\preceq\tilde{C}\preceq I_d/\lambda$,

$$\mathbb{E}_X\|\tilde{C}xx^\top\tilde{C}x\|_2^2=\mathbb{E}_X(x^\top\tilde{C}xx^\top\tilde{C}^2xx^\top\tilde{C}x)\le\mathbb{E}_x(x^\top x)^3/\lambda^4=O(n^3),$$

where the last equality is due to the fact that there are $O(n^3)$ terms of the form $x_i^2x_j^2x_k^2$, with $1\le i,j,k\le d$ in $(x^\top x)^3$. Now, it is enough to prove $\mathbb{E}_{\widetilde{X}}\mathrm{Var}_x(x^\top\tilde{C}x/n)\xrightarrow{u}0$.

Lemma 19 *Suppose that $x=(x_1,...,x_d)$ has i.i.d. entries satisfying $\mathbb{E}x_i=0$, $\mathbb{E}x_i^2=1$. Let $A\in\mathbb{R}^{d\times d}$. Then we have (see e.g. Bai and Silverstein, 2010; Couillet and Debbah, 2011 and Mei and Montanari, 2019 Lemma B.6.)*

$$\mathrm{Var}(x^\top Ax)=\sum_{i=1}^{d}A_{ii}^2(\mathbb{E}x_1^4-3)+\|A\|_F^2+\mathrm{tr}(A^2).$$

Using Lemma 19 above and recalling that $0\preceq\tilde{C}\preceq I_d/\lambda$, we have

$$\begin{aligned}
\mathbb{E}_{\widetilde{X}}\mathrm{Var}_{\tilde{x}}(x^\top\tilde{C}x)&=\mathbb{E}_{\widetilde{X}}\sum_{i=1}^{d}\tilde{C}_{ii}^2(\mathbb{E}x_1^4-3)+\|\tilde{C}\|_F^2+\mathrm{tr}(\tilde{C}^2)\\
&=\mathbb{E}_{\widetilde{X}}\sum_{i=1}^{d}\tilde{C}_{ii}^2(\mathbb{E}x_1^4-3)+2\|\tilde{C}\|_F^2\le\frac{d|\mathbb{E}x_1^4-3|}{\lambda^2}+\frac{2d}{\lambda^2}=O(n).
\end{aligned}$$

Therefore $\mathbb{E}_{\widetilde{X}}\mathrm{Var}_x(x^\top \tilde{C}x/n) = O(1/n) \xrightarrow{u} 0$ and we finished the proof. ■

B.4.2 Proof of Lemmas 16, 17

The results in these two lemmas follow from using matrix identities (e.g., the Woodbury identity) for the target matrices and deterministic equivalent results for orthogonal projection (Haar) matrices.

Proof By definition,

$$\mathbb{E}_W \tilde{M} = \mathbb{E}_W W^\top \left(\frac{WX^\top XW^\top}{n} + \lambda \right)^{-1} \frac{WX^\top}{n}.$$

Let $A = X^\top X/n + \lambda_1$, where $0 < \lambda_1 < \lambda$ is an arbitrary value. The Woodbury matrix identity states that

$$\left(A^{-1} + UCV\right)^{-1} = A - AU\left(C^{-1} + VAU\right)^{-1} VA.$$

Define $\lambda_2 := \lambda - \lambda_1$ and take $V = U^\top = W$, $C = I/\lambda_2$, to get

$$\left(A^{-1} + W^\top W/\lambda_2\right)^{-1} = A - AW^\top \left(\lambda_2 + WAW^\top\right)^{-1} WA.$$

Therefore

$$\begin{aligned} C :=& \mathbb{E}_W W^\top \left(\frac{WX^\top XW^\top}{n} + \lambda \right)^{-1} W = \mathbb{E}_W W^\top \left(\lambda_2 + WAW^\top\right)^{-1} W \\ &= A^{-1} - A^{-1}\mathbb{E}_W \left(A^{-1} + W^\top W/\lambda_2\right)^{-1} A^{-1} \\ &= A^{-1/2} \left[I_d - \mathbb{E}_W \left(I_d + \frac{A^{1/2}W^\top W A^{1/2}}{\lambda_2} \right)^{-1} \right] A^{-1/2}. \end{aligned} \tag{60}$$

Also,

$$\mathbb{E}_X \|\mathbb{E}_W \tilde{M}\|_F^2 = \mathbb{E}_X \operatorname{tr}\left(C \frac{X^\top X}{n^2} C^\top \right) = \mathbb{E}_X \frac{1}{n} \operatorname{tr} C(A - \lambda_1) C^\top. \tag{61}$$

$$\frac{1}{d}\mathbb{E}_X \|\mathbb{E}_W M\|_F^2 = \frac{1}{d}\mathbb{E}_X \operatorname{tr}\left(C \frac{X^\top X X^\top X}{n^2} C^\top \right) = \frac{1}{d}\mathbb{E}_X \operatorname{tr} C(A - \lambda_1)^2 C^\top. \tag{62}$$

Now, we define

$$\begin{aligned} C_1 :&= A^{-1} - A^{-1/2}\left(I_p + \frac{\bar{e}_d A}{\lambda_2} \right)^{-1} A^{-1/2} \\ &= A^{-1}\left[I_d - \left(I_p + \frac{\bar{e}_d A}{\lambda_2} \right)^{-1} \right] = \left(A + \frac{\lambda_2}{\bar{e}_d} I_d \right)^{-1}, \end{aligned}$$

where $\bar{e}_d$ is defined in Lemma 20. Then from Lemmas 20, 21, we know that C can be replaced by C_1 when calculating the limits of (61) and (62). Hence

$$\begin{aligned}\lim_{d\to\infty} \mathbb{E}_X \|\mathbb{E}_W \tilde{M}\|_F^2 &= \lim_{d\to\infty} \frac{1}{n}\mathbb{E}_X \operatorname{tr}\left(C_1(A-\lambda_1)C_1^\top\right)\\ &= \lim_{d\to\infty} \frac{1}{n}\mathbb{E}_X \operatorname{tr}\left[(A-\lambda_1)\left(A+\frac{\lambda_2}{\bar{e}_d}I_d\right)^{-2}\right]\\ &\to \delta[\theta_1(\delta,\tilde{\lambda}) - \tilde{\lambda}\theta_2(\delta,\tilde{\lambda})] = \delta(\tilde{\theta}_1 - \tilde{\lambda}\tilde{\theta}_2),\end{aligned}$$

where $\tilde{\lambda} := \lambda_1 + \lambda_2/\bar{e}_0$ is a constant independent of the choice of λ_1, λ_2. The last line follows from Lemma 23, the definition that $A = X^\top X/n + \lambda_1$ and the Marchenko-Pastur theorem. Similarly,

$$\begin{aligned}\lim_{d\to\infty} \frac{1}{d}\mathbb{E}_X \|\mathbb{E}_W M\|_F^2 &= \lim_{d\to\infty} \mathbb{E}_X \frac{1}{d}\operatorname{tr}\left(C_1(A-\lambda_1)^2C_1^\top\right)\\ &= \lim_{d\to\infty} \mathbb{E}_X \frac{1}{d}\operatorname{tr}\left[(A-\lambda_1)^2\left(A+\frac{\lambda_2}{\bar{e}_d}I_d\right)^{-2}\right]\\ &\to [1 - 2\tilde{\lambda}\theta_1(\delta,\tilde{\lambda}) + \tilde{\lambda}^2\theta_2(\delta,\tilde{\lambda})] = 1 - 2\tilde{\lambda}\tilde{\theta}_1 + \tilde{\lambda}\tilde{\theta}_2.\end{aligned}$$

This finishes the proof. ■

Lemma 20 (Weak deterministic equivalent for Haar matrices) *Under the above assumptions, we have*

$$\left(I_d + \frac{A^{1/2}W^\top W A^{1/2}}{\lambda_2}\right)^{-1} \overset{w}{\asymp} \left(I_d + \frac{\bar{e}_d A}{\lambda_2}\right)^{-1}, \tag{63}$$

where $(\bar{e}_d, e_d)$ is the unique solution of the system of equations

$$\begin{aligned}\bar{e}_d &= \frac{p}{d}\left(e_d + 1 - e_d\bar{e}_d\right)^{-1}\\ e_d &= \frac{1}{d}\operatorname{tr} A\left(\bar{e}_d A + \lambda_2 I_d\right)^{-1}.\end{aligned}$$

Proof [Proof of Lemma 20] From properties of sample covariance matrices, we know that the largest eigenvalue of $A_d := X^\top X/n + \lambda_1 \in \mathbb{R}^{d\times d}$ converges to $\lambda_1 + (1+\sqrt{\delta})^2$ almost surely as $d \to \infty$. Therefore, the sequence of values $\|A_d\|_2$ is bounded almost surely. For a fixed sequence of non-negative symmetric A_d with bounded 2-norm, (63) follows from the proof of theorem 7 (see the sketch of the proof) in Couillet et al. (2012). Since A is independent of W and the sequence $\|A_d\|_2$ is bounded almost surely, (63) holds generally. ■

Lemma 21 *(Replacing C by C_1) Under the previous assumptions, we have*

$$\lim_{d\to\infty} \mathbb{E}_X \frac{1}{n}\operatorname{tr}(C(A-\lambda_1)C^\top) = \lim_{d\to\infty} \mathbb{E}_X \frac{1}{n}\operatorname{tr}(C_1(A-\lambda_1)C_1^\top) \tag{64}$$

$$\lim_{d\to\infty} \mathbb{E}_X \frac{1}{d}\operatorname{tr}(C(A-\lambda_1)^2C^\top) = \lim_{d\to\infty} \mathbb{E}_X \frac{1}{d}\operatorname{tr}(C_1(A-\lambda_1)^2C_1^\top). \tag{65}$$

Proof [Proof of Lemma 21] Since the proof of two claims is almost the same, for simplicity, we only present the proof of (65).

For (65), we define (here $\tilde{\Delta}, \tilde{\Delta}_1, \tilde{\Delta}_2$ are different from those in Appendix B.3.5)

$$\begin{aligned}\tilde{\Delta} &:= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_X \operatorname{tr} C(A-\lambda_1)^2 C^\top - \frac{1}{d}\mathbb{E}_X \operatorname{tr} C_1(A-\lambda_1)^2 C_1^\top \\ &= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_X \left\{ \operatorname{tr}\left[(C-C_1)(A-\lambda_1)^2 C^\top\right] + \operatorname{tr}\left[C_1(A-\lambda_1)^2(C-C_1)^\top\right]\right\} \\ &=: \tilde{\Delta}_1 + \tilde{\Delta}_2.\end{aligned}$$

Also, denote

$$\begin{aligned}G_0 &:= (I_d + A^{1/2}W^\top W A^{1/2}/\lambda_2)^{-1} \\ G_1 &:= (I_d + \bar{e}_d A/\lambda_2)^{-1} \\ G &:= \mathbb{E}_W G_0.\end{aligned}$$

Therefore, from Lemma 20, we know that $G_0 \overset{w}{\asymp} G_1$. Substituting the definition of C, C_1 into $\tilde{\Delta}_1$, we have

$$\begin{aligned}\tilde{\Delta}_1 &= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_X \operatorname{tr}\left[A^{-1/2}(G-G_1)A^{-1/2}(A-\lambda_1)^2 A^{-1/2}(I_d-G)^\top A^{-1/2}\right] \\ &= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{X,W} \operatorname{tr}\left[A^{-1/2}(A-\lambda_1)^2 A^{-1/2}(I_d-G)^\top A^{-1}(G_0-G_1)\right].\end{aligned}$$

By Lemma 22, we know A and G are simultaneously diagonalizable. Moreover, it is readily verified that $A^{-1/2}(A-\lambda_1)^2 A^{-1/2}(I_d-G)^\top A^{-1}$ is symmetric and non-negative. Since

$$\begin{aligned}\|A^{-1/2}(A-\lambda_1)^2 A^{-1/2}(I_d-G)^\top A^{-1}\|_2 &= \|(I_d-G)^\top A^{-3/2}(A-\lambda_1)^2 A^{-1/2}\|_2 \\ &\le \|A^{-2}(A-\lambda)^2\|_2 \\ &\le \lambda^2\|A^{-1}\|_2^2 + 2\lambda\|A^{-1}\|_2 + 1 \le 4,\end{aligned}$$

we have by Lemma 20 and the bounded convergence theorem that $\tilde{\Delta}_1 \to 0$. Similarly, we can prove that $\tilde{\Delta}_2 \to 0$ and these conclude the proof of (65).

■

Lemma 22 (Commutativity of A and G) *Under the previous definitions, A and G are simultaneously diagonalizable, and therefore there are commutative.*

Proof [Proof of Lemma 22] Let $A = U\Gamma U^\top$ be the spectral decomposition. By definition and equation (60),

$$\begin{aligned}G &= \mathbb{E}_W(I_d + A^{1/2}W^\top W A^{1/2}/\lambda_2)^{-1} \\ &= I_d - A^{1/2}\mathbb{E}_W W^\top\left(\lambda_2 + WAW^\top\right)^{-1} WA^{1/2} \\ &= I_d - U\Gamma^{1/2}\mathbb{E}_W(WU)^\top\left[\lambda_2 + WU\Gamma(WU)^\top\right]^{-1} WUD^{1/2}U^\top \\ &= I_d - U\Gamma^{1/2}\left[\mathbb{E}_W W^\top\left(\lambda_2 + W\Gamma^\top W^\top\right)^{-1} W\right]\Gamma^{1/2}U^\top,\end{aligned}$$

where the last line is due to $W \overset{d}{=} WU$. Now it suffices to show that $\mathbb{E}_W W^\top \left(\lambda_2 + W\Gamma^\top W^\top\right)^{-1} W$ is a diagonal matrix.

Write $W = (w_1, w_2, .., w_d)$, where w_i, $1 \le i \le d$ are the columns of W. Denote the i-th $(1 \le i \le d)$ diagonal entry of Γ by γ_i and define $W_{-i} := (w_1, ..., w_{i-1}, w_{i+1}..., w_d)$ to be the matrix obtained by removing the i-th column from W. Then, for any $1 \le i \ne j \le d$

$$\mathbb{E}_W(W^\top(\lambda_2 + W\Gamma W^\top)^{-1}W)_{ij} = \mathbb{E}_W w_i^\top \left(\lambda_2 + \sum_{k=1}^d \gamma_k w_k w_k^\top\right)^{-1} w_j$$

$$= \mathbb{E}_{W_{-j}}\mathbb{E}_{w_j|W_{-j}} w_i^\top \left(\lambda_2 + \sum_{k=1}^d \gamma_k w_k w_k^\top\right)^{-1} w_j = 0, \tag{66}$$

where the last line follows from the symmetry of w_j. Therefore, $\mathbb{E}_W(W^\top(\lambda_2 + W\Gamma W^\top)^{-1}W)$ is a diagonal matrix and hence A and G are simultaneously diagonalizable. ■

Lemma 23 (Convergence of $\bar{e}_d$) *Under the previous assumptions and notations, suppose that $(\bar{e}_d(A), e_d(A))$ is the unique solution of*

$$\bar{e}_d = \frac{p}{d}\left(e_d + 1 - e_d\bar{e}_d\right)^{-1} \tag{67}$$

$$e_d = \frac{1}{d}\operatorname{tr} A\left(\bar{e}_d A + \lambda_2 I_d\right)^{-1}. \tag{68}$$

Then $(\bar{e}_d(A), e_d(A)) \to (\bar{e}_0, e_0)$ almost surely, where $(\bar{e}_0, e_0)$ is the unique solution of

$$\bar{e}_0 = \pi\left(e_0 + 1 - e_0\bar{e}_0\right)^{-1} \tag{69}$$

$$e_0 = \frac{1}{\bar{e}_0}\left[1 - \frac{\lambda_2}{\bar{e}_0}\theta_1\left(\delta, \lambda_1 + \frac{\lambda_2}{\bar{e}_0}\right)\right]. \tag{70}$$

Furthermore, for any decomposition $\lambda = \lambda_1 + \lambda_2$, we have

$$\lambda_1 + \lambda_2/\bar{e}_0 = \lambda + \frac{1-\pi}{2\pi}\left[\lambda + 1 - \gamma + \sqrt{(\lambda + \gamma - 1)^2 + 4\lambda}\right]. \tag{71}$$

We introduce λ_1, λ_2 only to ensure the invertibility of A and the uniform boundedness of $\|A^{-1}\|_2$. Equation (71) shows that different decompositions of λ do not affect the results.

Proof Plugging (68) into (67), we obtain

$$\frac{1}{d}\operatorname{tr}\left(\frac{A}{A + \lambda_2/\bar{e}_d}\right) = \frac{p/d - \bar{e}_d}{1 - \bar{e}_d}. \tag{72}$$

The uniqueness of the solution is guaranteed by theorem 7 in Couillet et al. (2012). Now, define for $x \in \mathbb{R}$,

$$g_d(x) := \frac{1}{d}\operatorname{tr}\left(\frac{A}{A + \lambda_2/x}\right) - \frac{p/d - x}{1 - x}.$$

$$g(x) := \left[1 - \frac{\lambda_2}{x}\theta_1\left(\delta, \lambda_1 + \frac{\lambda_2}{x}\right)\right] - \frac{\pi - x}{1 - x}.$$

First, we consider the case when $0 < \pi < 1$. Noting that $g_d(0+) = -p/d$, $g_d(1-) > 0$ and $g_d(x)$ is increasing on $(0,1)$, it follows that $g_d(x)$ has a unique zero $\bar{e}_d$ on $(0,1)$. Similarly, we can conclude that $g(x)$ has a unique zero $\bar{e}_0$ on $(0,1)$.

Since $A = X^\top X/n + \lambda_1$ and $p/d \to \pi$, by applying the Marchenko-Pastur theorem, we know that $g_d(x) \to g(x)$ for any $x \in (0,1)$, almost surely. Since $g_d(x)$ are increasing functions on $(0,1)$, we further have, on any closed interval in $(0,1)$, $g_d(x)$ converges uniformly to $g(x)$, almost surely.

Due to the uniformly convergence of $g_d(x)$ on any closed interval and the fact that the zeros of $g_d(x), g(x)$ are in $(0,1)$, we conclude that the zeros of $g_d(x)$ converge to the zero of $g(x)$ almost surely, i.e., $\bar{e}_d \to \bar{e}_0$ almost surely, and hence $(\bar{e}_d, e_d) \to (\bar{e}_0, e_0)$ almost surely.

Plugging (69) into (70), we obtain

$$\left[1 - \frac{\lambda_2}{\bar{e}_0}\theta_1\left(\delta, \lambda_1 + \frac{\lambda_2}{\bar{e}_0}\right)\right] = \frac{\pi - \bar{e}_0}{1 - \bar{e}_0},$$

Thus,

$$\theta_1\left(\delta, \lambda_1 + \frac{\lambda_2}{\bar{e}_0}\right) = \frac{1-\pi}{\lambda_2(1/\bar{e}_0 - 1)}. \tag{73}$$

Plugging the explicit expression of $\theta_1(\delta, \lambda_1 + \lambda_2/\bar{e}_0)$ into (73) and reorganizing the result, we obtain

$$\pi\left(\lambda_1 + \frac{\lambda_2}{\bar{e}_0}\right)^2 - [(1+\pi)\lambda + (1-\pi)(1-\gamma)]\left(\lambda_1 + \frac{\lambda_2}{\bar{e}_0}\right) + \lambda[\lambda + (1-\delta)(1-\pi)] = 0.$$

Thus, defining

$$h(x) := \pi x^2 - [(1+\pi)\lambda + (1-\pi)(1-\gamma)]x + \lambda[\lambda + (1-\delta)(1-\pi)],$$

we get that $\lambda_1 + \lambda_2/\bar{e}_0$ is a solution of $h(x) = 0$. Since $\bar{e}_0 \in (0,1)$, $\lambda_1 + \lambda_2/\bar{e}_0 > \lambda$. Note that $h(\lambda) = -\delta(1-\pi)^2\lambda \le 0$. From the properties of quadratic functions, we conclude that $\lambda_1 + \lambda_2/\bar{e}_0$ is the larger solution of $h(x) = 0$. Therefore,

$$\lambda_1 + \lambda_2/\bar{e}_0 = \lambda + \frac{1-\pi}{2\pi}\left[\lambda + 1 - \gamma + \sqrt{(\lambda + \gamma - 1)^2 + 4\lambda}\right] =: \tilde{\lambda}.$$

If $\pi = 1$, it can be readily verified that the unique solution of (69), (70) is

$$(\bar{e}_0, e_0) = (1, 1 - \lambda_2\theta_1(\delta, \lambda_1 + \lambda_2)).$$

Thus, equation (71) follows directly. As for the convergence of $(\bar{e}_d, \mathbb{E}_d)$, since $g_d(x) < 0$ for $x < p/d$, the zero of $g_d(x)$ (i.e. $\bar{e}_d$) is larger than p/d and hence converges to 1 as $p/d \to \pi = 1$. Therefore, we have shown that $\bar{e}_d \to \bar{e}_0$, and hence $e_d \to e_0$ follows from equation (68) and the Marchenko-Pastur theorem. This finishes the proof. ■

B.5 Proof of Theorem 5

Proof In the following proof, we will calculate the derivatives of variances and bias separately and obtain the results in Table 1 based on them. Throughout the proof, we denote $c := \delta(1+\sigma^2/\alpha^2)+1$ for simplicity.

(1) MSE. Plugging the explicit expressions of $\theta_1, \theta_2, \lambda^*$ into equation (8) and taking derivatives with respect to π and δ separately

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\pi}\lim_{d\to\infty}\mathbf{MSE}(\lambda^*) &= \frac{\mathrm{d}}{\mathrm{d}\pi}[\alpha^2(1-\pi+\lambda^*\pi\theta_1)+\sigma^2] \\ &= \alpha^2\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{2\delta-c+\sqrt{c^2-4\gamma}}{2\delta}\right) = -\frac{\alpha^2}{\sqrt{c^2-4\gamma}} \le 0, \end{aligned} \tag{74}$$

Thus, $\lim_{d\to\infty}\mathbf{MSE}(\lambda^*)$ is monotonically decreasing as a function of π. Similarly,

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\lim_{d\to\infty}\mathbf{MSE}(\lambda^*) = \frac{\mathrm{d}}{\mathrm{d}\delta}[\alpha^2(1-\pi+\lambda^*\pi\theta_1)+\sigma^2] = \frac{\alpha^2}{2\delta^2}\left(1+\frac{2\gamma-c}{\sqrt{c^2-4\gamma}}\right). \tag{75}$$

Since $c^2-4\gamma \ge c^2-4\gamma(c-\gamma) = (c-2\gamma)^2$, the derivative is larger than 0. Therefore, $\lim_{d\to\infty}\mathbf{MSE}(\lambda^*)$ is increasing as a function of δ.

(2) Bias^2. Since $(\lim_{d\to\infty}\mathbf{MSE}(\lambda^*)-\sigma^2)^2 = \alpha^2\lim_{d\to\infty}\mathbf{Bias}^2(\lambda^*)$, the monotonicity of the MSE implies the monotonicity of Bias^2.

(3) Var. Note that $\mathbf{Var}(\lambda^*) = \mathbf{MSE}(\lambda^*) - \mathbf{Bias}^2(\lambda^*) - \sigma^2$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\pi}\lim_{d\to\infty}\mathbf{Var}(\lambda^*) &= \frac{\mathrm{d}}{\mathrm{d}\pi}\left(\lim_{d\to\infty}\mathbf{MSE}(\lambda^*) - \lim_{d\to\infty}\mathbf{Bias}^2(\lambda^*) - \sigma^2\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}\pi}\left[\lim_{d\to\infty}(\mathbf{MSE}(\lambda^*)-\sigma^2)\left(1-\frac{1}{\alpha^2}\lim_{d\to\infty}(\mathbf{MSE}(\lambda^*)-\sigma^2)\right)\right] \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}\pi}\lim_{d\to\infty}\mathbf{MSE}(\lambda^*)\right)\left(1-\frac{2}{\alpha^2}\lim_{d\to\infty}(\mathbf{MSE}(\lambda^*)-\sigma^2)\right) \\ &= -\frac{\alpha^2}{\sqrt{c^2-4\gamma}}\cdot\left(\frac{\delta\sigma^2/\alpha^2+1-\sqrt{c^2-4\gamma}}{\delta}\right).\end{aligned}$$

Since $\sqrt{c^2-4\gamma}$ is decreasing as π increases, simple calculations reveal that, as a function of π, $\lim_{d\to\infty}\mathbf{Var}(\lambda^*)$ is monotonically increasing when $\delta \ge 2\alpha^2/(\alpha^2+2\sigma^2)$, while it is increasing on $(0, [2+\delta(1+2\sigma^2/\alpha^2)]/4]$ and decreasing on $([2+\delta(1+2\sigma^2/\alpha^2)]/4, 1]$ when $\delta < 2\alpha^2/(\alpha^2+2\sigma^2)$. Similarly,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta}\lim_{d\to\infty}\mathbf{Var}(\lambda^*) &= \left(\frac{\mathrm{d}}{\mathrm{d}\delta}\lim_{d\to\infty}\mathbf{MSE}(\lambda^*)\right)\left(1-\frac{2}{\alpha^2}\lim_{d\to\infty}(\mathbf{MSE}(\lambda^*)-\sigma^2)\right) \\ &= \frac{\alpha^2}{2\delta^2}\left(1+\frac{2\gamma-c}{\sqrt{c^2-4\gamma}}\right)\cdot\left(\frac{\delta\sigma^2/\alpha^2+1-\sqrt{c^2-4\gamma}}{\delta}\right).\end{aligned}$$

From (75), we know the first term in the last line is non-negative. Plugging in the expression of c, it follows from some simple calculations that, as a function of δ, $\lim_{d\to\infty}\mathbf{Var}(\lambda^*)$ is

monotonically decreasing when $\pi < 0.5$, while it is increasing on $(0, 2(2\pi-1)/[1+2\sigma^2/\alpha^2]]$ and decreasing on $(2(2\pi-1)/[1+2\sigma^2/\alpha^2], +\infty)$ when $\pi > 0.5$.

(4) Σ_{label}. Plugging the definition of θ_1, θ_2 and λ^* into $\Sigma_{\text{label}}(\lambda^*)$, we obtain

$$\begin{aligned}\lim_{d\to\infty} \Sigma_{\text{label}}(\lambda^*) &= \sigma^2\pi\delta(\theta_1 - \lambda^*\theta_2) \\ &= \frac{\sigma^2}{2\lambda^*}\left(\sqrt{c^2-4\gamma} - \frac{(\gamma+1)\lambda^* + (\gamma-1)^2}{\sqrt{c^2-4\gamma}}\right) = \frac{\sigma^2 c}{\sqrt{c^2-4\gamma}}.\end{aligned}$$

Therefore, as a function of π, $\lim_{d\to\infty}\Sigma_{\text{label}}(\lambda^*)$ is monotonically increasing on $(0,1]$. Since

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\lim_{d\to\infty}\Sigma_{\text{label}}(\lambda^*) = \frac{2\pi\sigma^2(1-\delta(1+\sigma^2/\alpha^2))}{\sqrt{c^2-4\gamma}},$$

it follows that $\lim_{d\to\infty}\Sigma_{\text{label}}(\lambda^*)$ is increasing on $(0, \alpha^2/(\sigma^2+\alpha^2)]$ and decreasing on $(\alpha^2/[\sigma^2+\alpha^2], +\infty)$ as a function of δ.

(5) Σ_{init}. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\pi}\lim_{d\to\infty}\Sigma_{\text{init}}(\lambda^*) &= \frac{\mathrm{d}}{\mathrm{d}\pi}\alpha^2\pi(1-\pi)(1-\lambda^*\theta_1)^2 \\ &= \alpha^2(1-\lambda^*\theta_1)[(1-2\pi)(1-\lambda^*\theta_1) - 2\pi(1-\pi)(\lambda^*\theta_1)'] \\ &= 2\alpha^2(1-\lambda^*\theta_1)\left(\frac{4\delta\pi^3 - 2c^2\pi + c^2}{(c^2-4\gamma\pi + c\sqrt{c^2-4\gamma})\sqrt{c^2-4\gamma}}\right).\end{aligned}$$

Since $f(t) := 4\delta t^3 - 2c^2 t + c^2$ satisfies the following properties (a). $f(0) = c^2 > 0$, (b). $f(1) = 4\delta - c^2 \le -(\delta(1+\sigma^2/\alpha^2)-1)^2 \le 0$, (c). $f(t)$ has a unique extremum on $(0,+\infty)$; we know that f is decreasing and has a unique zero on $(0,1]$. Therefore, $\lim_{d\to\infty}\Sigma_{\text{init}}(\lambda^*)$ is unimodal as a function of λ_1. Now, taking derivatives with respect to δ,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta}\lim_{d\to\infty}\Sigma_{\text{init}}(\lambda^*) &= \frac{\mathrm{d}}{\mathrm{d}\pi}\alpha^2\pi(1-\pi)(1-\lambda^*\theta_1)^2 \\ &= -2\alpha^2\pi(1-\pi)(1-\lambda^*\theta_1)(\lambda^*\theta_1)' \\ &= -\frac{\alpha^2}{\delta^2}(1-\pi)(1-\lambda^*\theta_1)\left(\frac{2\gamma-c}{\sqrt{c^2-4\gamma}}+1\right) \le 0,\end{aligned}$$

where the last inequality follows from the fact that $c^2 - 4\gamma \ge (2\gamma - c)^2$. Thus, $\lim_{d\to\infty}\Sigma_{\text{init}}(\lambda^*)$ is monotonically decreasing as a function of δ.

■

B.6 Proof of Proposition 6

Proof [Proof of Proposition 6] It is known that if the design matrix X has independent entries of zero mean and unit variance, then as $n, p \to \infty$ proportionally, i.e., $p/n \to \gamma > 0$,

the MSE converges almost surely to the expression (Tulino and Verdú, 2004; Couillet and Debbah, 2011; Dobriban and Wager, 2018)

$$\mathbf{MSE}(\gamma) = \gamma m_\gamma(-\lambda^*) = \gamma\theta_1(\gamma, \lambda^*) + \sigma^2, \tag{76}$$

where $\lambda^* = \gamma/\alpha^2$ is the limit of the optimal regularization parameters, and m_γ is the Stieltjes transform of the limiting eigenvalue distribution F_γ of $\widehat{\Sigma} = n^{-1}X^\top X$, i.e., the Stieltjes transform of the standard Marchenko-Pastur distribution with aspect ratio γ.

Furthermore, as shown in the proof of theorem 2.2 in Liu and Dobriban (2020), the specific forms of the bias and variance are: with $\theta_i := \theta_i(\gamma, \lambda)$, $i = 1, 2$

$$(a).\, \mathbf{Bias}^2 = \alpha^2 \int \frac{\lambda^2}{(x+\lambda)^2} dF_\gamma(x) = \alpha^2\lambda^2\theta_2, \quad (b).\, \mathbf{Var} = \gamma \int \frac{x}{(x+\lambda)^2} dF_\gamma(x) = \gamma(\theta_1 - \lambda\theta_2). \tag{77}$$

Therefore, we can obtain the explicit formulas of the bias, variance and MSE by plugging $\lambda^* = \lambda/\alpha^2$, equation (6), (7) into equation (76), (77). All results in Proposition 6 can be derived by calculating the derivatives as we have done in the proof of theorem 5. However, since the proofs are simpler in this special case, we present them here for the convenience of readers. Throughout this proof, we denote $1/\alpha^2$ by c.

MSE: Let $\tau := (1/\alpha^2 - 1)\gamma - 1$, substitute (6) into (76) and take derivatives:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbf{MSE}(\gamma) = \frac{\mathrm{d}}{\mathrm{d}\gamma}[\gamma\theta_1(\gamma, \lambda^*) + \sigma^2] &= \frac{\mathrm{d}}{\mathrm{d}\gamma}\frac{\tau + \sqrt{\tau^2 + 4c\gamma^2}}{2c\gamma} \\
&= \frac{\tau + 1 + [\tau(\tau+1) + 4c\gamma^2]/\sqrt{\tau^2 + 4c\gamma^2} - (\tau + \sqrt{\tau^2 + 4c\gamma^2})}{2c\gamma^2} \\
&= \frac{\tau + \sqrt{\tau^2 + 4c\gamma^2}}{2c\gamma^2\sqrt{\tau^2 + 4c\gamma^2}} \geq 0.
\end{aligned}$$

Thus, the MSE is strictly increasing as γ increases.

Bias: Plugging equation (7) into (77)(a) and denoting $1/\gamma$ by x , we have

$$\begin{aligned}
\mathbf{Bias}^2(\gamma) = \alpha^2\lambda^{*2}\theta_2(\gamma, \lambda^*) &= \frac{\alpha^2}{2}\left(1 - \frac{1}{\gamma} + \frac{c\gamma(\gamma+1) + (\gamma-1)^2}{\gamma\sqrt{(c+1)^2\gamma^2 + 2(c-1)\gamma + 1}}\right) \\
&= \frac{\alpha^2}{2}\left(1 - x + \frac{c(x+1) + (x-1)^2}{\sqrt{(c+1)^2 + 2(c-1)x + x^2}}\right) =: \frac{\alpha^2}{2}(1 + f(x, c)).
\end{aligned}$$

Thus it is enough to show that $f(x, c)$ is decreasing for $x > 0$. Taking derivatives with respect to x, we have after some calculations that

$$\frac{d^2}{dx^2} f(x, c) = \frac{6(c+1)^3 + 12c^2x}{((c+1)^2 + 2(c-1)x + x^2)^{\frac{5}{2}}} > 0, \forall x > 0.$$

Thus, for any fixed $c > 0$, $f(x, c)$ is a strictly convex function with respect to x on $(0, +\infty)$. Furthermore, fixing $c > 0$ and letting $x \to \infty$, we get

$$\begin{aligned}
\lim_{x\to\infty} f(x,c) &= \lim_{x\to\infty} -x + \frac{c(x+1)+(x-1)^2}{\sqrt{(c+1)^2+2(c-1)x+x^2}} \\
&= \lim_{x\to\infty} \frac{c(x+1)+(x-1)^2 - x\sqrt{(c+1)^2+2(c-1)x+x^2}}{\sqrt{(c+1)^2+2(c-1)x+x^2}} \\
&= \lim_{x\to\infty} \frac{(x^2+(c-2)x+c+1) - x(x+c-1)\sqrt{1+\frac{4c}{(x+c-1)^2}}}{x} \\
&= \lim_{x\to\infty} \frac{(x^2+(c-2)x+c+1) - x(x+c-1)(1+O(\frac{1}{x^2}))}{x} \\
&= \lim_{x\to\infty} \frac{-x+O(1)}{x} = -1.
\end{aligned}$$

Combining the results above and using the fact that a strictly convex function with a finite limit is strictly decreasing, it follows that $f(x, c)$ is both strictly decreasing and convex. Therefore, we have proved that $\mathbf{Bias}^2(\gamma)$ is strictly increasing on $(0, +\infty)$.

Variance: Similarly, plugging (6), (7) into (77) (b), we get

$$\begin{aligned}
\mathbf{Var}(\gamma) &= -\frac{1}{2} + \frac{(c+1)\gamma+1}{2\sqrt{(\gamma(1-c)-1)^2+4\gamma^2 c}} \\
&= -\frac{1}{2} + \frac{(c+1)\gamma+1}{2\sqrt{(c+1)^2\gamma^2+2(c-1)\gamma+1}}.
\end{aligned}$$

Let $x = (c+1)\gamma$ and $t = \frac{c-1}{c+1}$. Then it suffices to prove the unimodality of

$$g_t(x) := \frac{x+1}{\sqrt{x^2+2tx+1}}, \qquad x \in (0,+\infty), t \in (-1,1).$$

Differentiating $g_t(x)$ with respect to x gives

$$\frac{d}{dx} g_t(x) = \frac{(1-t)(1-x)}{(x^2+2tx+1)^{\frac{3}{2}}}.$$

Since $g_t'(x) > 0$ when $x < 1$, and $g_t'(x) < 0$ when $x > 1$, we see that $g_t(x)$ is strictly increasing on $(0, 1)$ and strictly decreasing on $(1, +\infty)$. Correspondingly, by changing x back to γ, c back to $1/\alpha^2$, we have shown that $\mathbf{Var}(\gamma)$ is strictly increasing on $(0, \alpha^2/(\alpha^2+1))$ and strictly decreasing on $[\alpha^2/(\alpha^2+1), +\infty)$. Moreover, noting that

$$\begin{aligned}
\lim_{\gamma\to+\infty} \mathbf{Var}(\gamma) &= \lim_{\gamma\to+\infty} -\frac{1}{2} + \frac{(c+1)\gamma+1}{2\sqrt{(c+1)^2\gamma^2+2(c-1)\gamma+1}} \\
&= \lim_{\gamma\to+\infty} -\frac{1}{2} + \frac{(c+1)\gamma}{2(c+1)\gamma} = 0.
\end{aligned}$$

Therefore, we conclude that $\mathbf{Var}(\gamma)$ is a unimodal function converges to zero at infinity with a unique maximum point at $\alpha^2/(\alpha^2+1)$. Finally, proposition 6(4) can be obtained by evaluating $\mathbf{Var}(\gamma)$ and $\mathbf{Bias}^2(\gamma)$ at $\alpha^2/(\alpha^2+1)$.

■

B.7 Proof of Theorem 7

Proof (1). By plugging the expression of θ_1 into equation (8) and taking derivatives with respect to δ in equation (8), we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\pi}\lim_{d\to\infty}\mathbf{Bias}^2(\lambda) &= \frac{\mathrm{d}}{\mathrm{d}\pi}\alpha^2(1-\pi+\lambda\pi\theta_1)^2 \\ &= \alpha^2(1-\pi+\lambda\pi\theta_1)\left(\frac{\lambda+\gamma-1}{\sqrt{(-\lambda+\gamma-1)^2+4\lambda\gamma}}-1\right)<0,\end{aligned}$$

where the inequality follows from the facts that $1-\pi+\lambda\pi\theta_1 = \int[\lambda+(1-\pi)x]/(x+\lambda)dF_\gamma(x) \geq 0$ and $(\lambda+\gamma-1)^2-[(-\lambda+\gamma-1)^2+4\lambda\gamma] = -4\lambda < 0$. Similarly,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta}\lim_{d\to\infty}\mathbf{Bias}^2(\lambda) &= \frac{\mathrm{d}}{\mathrm{d}\delta}\alpha^2(1-\pi+\lambda\pi\theta_1)^2 \\ &= \frac{\alpha^2}{\delta^2}(1-\pi+\lambda\pi\theta_1)\left(\lambda+1-\frac{\lambda^2+(\gamma+2)\lambda+(1-\gamma)}{\sqrt{(-\lambda+\gamma-1)^2+4\lambda\gamma}}\right)>0.\end{aligned}$$

Therefore, the limiting $\mathbf{Bias}^2$ is monotonically increasing as a function of δ and monotonically decreasing as a function of π.

(2). Denote $\lambda^* = \delta(1-\pi+\sigma^2/\alpha^2)$ as before. From (9), we have

$$\begin{aligned}\lim_{d\to\infty}\mathbf{Var}(\lambda) &= \alpha^2\pi\left\{1-\pi+\frac{\lambda}{\delta}+\left[(\pi-1)(2\lambda-\delta)+\frac{\lambda(\lambda-\gamma+1)}{\delta}+\frac{\delta\sigma^2}{\alpha^2}\right]\theta_1+\right. \\ &\quad \left.\lambda\left[\lambda-\delta\left(1-\pi+\sigma^2/\alpha^2\right)\right]\theta_2\right\} \\ &= \alpha^2\pi\left\{1-\pi+\frac{\lambda}{\delta}+\left[\left(2\pi-2+\frac{\lambda-\gamma+1}{\delta}\right)\lambda+\lambda^*\right]\theta_1(1,\lambda)+\lambda(\lambda-\lambda^*)\theta_2\right\}\end{aligned}$$

Now suppose $\delta = 1/\pi$ and let $\lambda \to 0+$,

$$\begin{aligned}\lim_{\lambda\to 0}\lim_{d\to\infty}\mathbf{Var}(\lambda) &= \lim_{\lambda\to 0}\alpha^2\pi\left\{1-\pi+\frac{\lambda}{\delta}+\left[\left(2\pi-2+\frac{\lambda-\gamma+1}{\delta}\right)\lambda+\lambda^*\right]\theta_1+\lambda(\lambda-\lambda^*)\theta_2\right\} \\ &= \lim_{\lambda\to 0}\alpha^2\pi[1-\pi+\lambda^*(\theta_1(1,\lambda)-\lambda\theta_2(1,\lambda))] \\ &= \lim_{\lambda\to 0}\alpha^2\pi\left[1-\pi+\lambda^*\left(\frac{4\lambda}{(\sqrt{\lambda^2+4\lambda}+\lambda)^2\sqrt{\lambda^2+4\lambda}}\right)\right] = \lim_{\lambda\to 0}O(\frac{1}{\lambda^{1/2}}) = \infty.\end{aligned}$$

Finally, letting $\lambda \to 0$ in Theorem 2, we obtain after some similar calculations that V_{si} and V_{sli} go to infinity while V_s, V_i and V_{sl} converge to some finite limits as $d \to \infty, \lambda \to 0$. ■

B.8 Proof of Theorem 8

Proof [Proof of Theorem 8] For notational simplicity, we sometimes denote the 2-norm of vectors and the Frobenius norm of matrices by $\|\cdot\|$ in this proof. From definition (16), we

have

$$\begin{aligned}\beta_{opt} :&= \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)}\mathbb{E}_{x,\varepsilon}[(Wx)^\top\beta - (x^\top\theta+\varepsilon)]^2\\ &= \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)}\mathbb{E}_{x}[(Wx)^\top\beta - x^\top\theta]^2\\ &= \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)}\mathbb{E}_{x}[(W^\top\beta-\theta)^\top xx^\top(W^\top\beta-\theta)]^2\\ &= \operatorname{argmin}_\beta \mathbb{E}_{p(\theta|XW^\top,W,Y)}\|W^\top\beta-\theta\|_2^2\\ &= W\mathbb{E}_{p(\theta|XW^\top,W,Y)}\theta.\end{aligned}$$

Thus, we have proved equation (17).

Now, to prove (18), it suffices to do the following:

(1). Calculate the posterior $p(\theta|XW^\top,W,Y)$.

(2). Bound the difference between β_{opt} and $\hat{\beta}$.

(1). Let $W_\perp = f(W) \in \mathbb{R}^{(d-p)\times d}$ be a deterministic orthogonal complement of W, such that $W_\perp^\top W_\perp + W^\top W = I_d$. Then we have

$$\begin{aligned}&p(\theta|XW^\top,W,Y) \propto p(\theta)p(XW^\top,W,Y|\theta) = p(\theta)p(XW^\top,W|\theta)p(Y|XW^\top,W,\theta)\\ &\propto \exp\left(-\frac{\|\theta\|_2^2}{2\alpha^2/d} - \frac{\|XW^\top\|_F^2}{2}\right)\cdot p(Y|XW^\top,W,\theta) \qquad (78)\\ &\propto \exp\left(-\frac{\|\theta\|_2^2}{2\alpha^2/d}\right)\cdot\int p(Y|XW_\perp^\top,XW^\top,W,\theta)\cdot p(XW_\perp^\top|XW^\top,W,\theta)d(XW_\perp^\top)\\ &\propto \exp\left(-\frac{\|\theta\|_2^2}{2\alpha^2/d}\right)\cdot\int \exp\left(-\frac{\|Y-X\theta\|_2^2}{2\sigma^2}\right)\cdot p(XW_\perp^\top|XW^\top,W,\theta)d(XW_\perp^\top),\end{aligned}$$

where in the second line we used the facts that $XW^\top$ and W are independent conditioned on θ, $XW^\top$ has i.i.d. $\mathcal{N}(0,1)$ entries, and W is uniformly distributed over partial orthogonal matrices. Denote $XW^\top$ by A and $XW_\perp^\top$ by A_1. Then using the fact that A, A_1 have i.i.d. $\mathcal{N}(0,1)$ entries and A, A_1 are independent conditioned on W and θ, we obtain

$$p(XW_\perp^\top|XW^\top,W,\theta) \propto \exp\left(-\frac{\|XW_\perp^\top\|_F^2}{2}\right) = \exp\left(-\frac{\|A_1\|_F^2}{2}\right).$$

Therefore,

$$\begin{aligned}&p(Y|XW^\top,W,\theta) = \int p(Y|X,\theta)\cdot p(XW_\perp^\top|XW^\top,W,\theta)d(XW_\perp^\top)\\ &= \int \exp\left(-\frac{\|(Y-AW\theta)-A_1W_\perp\theta\|_2^2+\sigma^2\|A_1\|_F^2}{2\sigma^2}\right)dA_1\\ &= \exp\left(-\frac{\|Y-AW\theta\|_2^2}{2\sigma^2}\right)\int \exp\left\{-\frac{\sigma^2\|A_1\|_F^2+\|A_1W_\perp\theta\|_F^2 - 2\operatorname{tr}[A_1W_\perp\theta(Y-AW\theta)^\top]}{2\sigma^2}\right\}dA_1\end{aligned}$$

Further denote $W\theta$ by $\tilde{\theta}$, $W_\perp\theta$ by $\tilde{\theta}_1$ and $(Y-A\tilde{\theta})\tilde{\theta}_1^\top(\tilde{\theta}_1\tilde{\theta}_1^\top+\sigma^2)^{-1}$ by B. For a fixed Y, conditioned on A, W, θ, by separating A_1 into n rows, applying Fubini's theorem and using

properties of the p.d.f. of a normal distribution, we get

$$
\begin{aligned}
&p(Y|XW^\top, W, \theta) \propto \exp\left(-\frac{\|Y - A\tilde{\theta}\|_2^2}{2\sigma^2}\right) \cdot \\
&\int \exp\left\{-\frac{\mathrm{tr}[(A_1 - B)(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)(A_1 - B)^\top]}{2\sigma^2} + \frac{\mathrm{tr}[B(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)B^\top]}{2\sigma^2}\right\} dA_1 \qquad (79)\\
&\propto \exp\left(-\frac{\|Y - A\tilde{\theta}\|^2}{2\sigma^2} + \frac{\mathrm{tr}[B(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)B^\top]}{2\sigma^2}\right) \cdot \det(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)^{-\frac{n}{2}} \\
&= \exp\left(-\frac{\|Y - A\tilde{\theta}\|^2}{2\sigma^2} + \frac{\mathrm{tr}[(Y - A\tilde{\theta})\tilde{\theta}_1^\top(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)^{-1}\tilde{\theta}_1(Y - A\tilde{\theta})^\top]}{2\sigma^2}\right) \cdot \det(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)^{-\frac{n}{2}}.
\end{aligned}
$$

Let $\tilde{\theta}_1 = UDV^\top$ be the SVD of $\tilde{\theta}_1$. Then

$$
\begin{aligned}
\det(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2) = \det(DD^\top + \sigma^2) = (\|\tilde{\theta}_1\|^2 + \sigma^2)\prod_{i=2}^{d-p}\sigma^2 \propto \|\tilde{\theta}_1\|^2 + \sigma^2. \\
\tilde{\theta}_1^\top(\tilde{\theta}_1\tilde{\theta}_1^\top + \sigma^2)^{-1}\tilde{\theta}_1 = D^\top(DD^\top + \sigma^2)^{-1}D = \|\tilde{\theta}_1\|^2(\|\tilde{\theta}_1\|^2 + \sigma^2)^{-1}.
\end{aligned}
$$

Thus,

$$
p(Y|XW^\top, W, \theta) \propto \exp\left\{[-1 + \|\tilde{\theta}_1\|^2(\|\tilde{\theta}_1\|^2 + \sigma^2)^{-1}]\frac{\|Y - A\tilde{\theta}\|^2}{2\sigma^2}\right\} \cdot (\|\tilde{\theta}_1\|^2 + \sigma^2)^{-\frac{n}{2}}. \quad (80)
$$

Finally, by substituting (80) into (78), we obtain the posterior:

$$
\begin{aligned}
p(\theta|XW^\top, W, Y) &\propto \exp\left(-\frac{\|\theta\|_2^2}{2\alpha^2/d} - \frac{\|Y - A\tilde{\theta}\|^2}{2(\|\tilde{\theta}_1\|^2 + \sigma^2)}\right) \cdot (\|\tilde{\theta}_1\|^2 + \sigma^2)^{-\frac{n}{2}} \\
&= \exp\left(-\frac{\|\tilde{\theta}\|^2}{2\alpha^2/d} - \frac{\|Y - A\tilde{\theta}\|^2}{2(\|\tilde{\theta}_1\|^2 + \sigma^2)}\right) \cdot \exp(-\frac{\|\tilde{\theta}_1\|^2}{2\alpha^2/d}) \cdot (\|\tilde{\theta}_1\|^2 + \sigma^2)^{-\frac{n}{2}}. \qquad (81)
\end{aligned}
$$

(2). Since $\beta_{opt} = W\mathbb{E}_{p(\theta|XW^\top, W, Y)}\theta = \mathbb{E}_{p(\theta|XW^\top, W, Y)}\tilde{\theta}$, it suffices to calculate the posterior mean of $\tilde{\theta}$. In equation (81), we can see that conditioned on $\tilde{\theta}_1$, $\tilde{\theta}$ follows a normal distribution. Moreover, using the same technique as in equation (79), it is not hard to verify that the expectation of $\tilde{\theta}$ conditioned on $\tilde{\theta}_1$ is

$$
\left(\frac{d}{\alpha^2} + \frac{A^\top A}{\|\tilde{\theta}_1\|^2 + \sigma^2}\right)^{-1}\frac{A^\top Y}{\|\tilde{\theta}_1\|^2 + \sigma^2}. \qquad (82)
$$

Now changing A, $\tilde{\theta}$ back to X, W, we obtain

$$
\mathbb{E}_{p(\theta|\tilde{\theta}_1, XW^\top, W, Y)}\tilde{\theta} = \left[\frac{WX^\top XW^\top}{n} + \frac{(\|\tilde{\theta}_1\|^2 + \sigma^2)d}{n\alpha^2}\right]^{-1}\frac{WX^\top Y}{n}. \qquad (83)
$$

Therefore, the conditional expectation of $\tilde{\theta}$ is the ridge estimator with $\lambda = \lambda(\tilde{\theta}_1) := (\|\tilde{\theta}_1\|^2 + \sigma^2)d/[n\alpha^2]$. Thus β_{opt} is in fact a weighted ridge estimator. Besides, note that $W_\perp\theta$ has i.i.d. $\mathcal{N}(0, \alpha^2/d)$ entries. Hence letting $\chi^2(k)$ be a chi squared random variable with k degrees of freedom,

$$\|\tilde{\theta}_1\|^2 = \|W_\perp\theta\|^2 = \frac{\alpha^2}{d}\sum_{i=1}^{d-p}\left(\frac{\sqrt{d}W_\perp\theta}{\alpha}\right)_i^2 \overset{d}{=} \frac{\alpha^2}{d}\chi^2(d-p) \xrightarrow{w} \alpha^2(1-\pi). \tag{84}$$

Denote $\tilde{R}^* = (\frac{WX^\top XW^\top}{n} + \lambda^*)^{-1}$. Thus, we may guess that the posterior of $\|\tilde{\theta}_1\|$ is close to $\alpha^2(1-\pi)$ with high probability and

$$\beta_{opt} \approx \left[\frac{WX^\top XW^\top}{n} + \delta(1-\pi+\sigma^2/\alpha^2)\right]^{-1}\frac{WX^\top Y}{n} = \tilde{R}^*\frac{WX^\top Y}{n} = \hat{\beta},$$

which is the optimal ridge estimator.

We formalize this idea by bounding the difference between β_{opt} and $\hat{\beta}$. Denote the conditional mean (83) of $\tilde{\theta}$ given $\|\tilde{\theta}_1\|^2 = c$ by $R(c)$. Let also $\tilde{R}_1 = (\frac{WX^\top XW^\top}{n} + \lambda(\tilde{\theta}_1))^{-1}$. Then the optimal ridge estimator is $\hat{\beta} = R(\alpha^2(1-\pi))$ and we have

$$\begin{aligned}
&\mathbb{E}_{XW^\top,W,Y}\|\hat{\beta} - \beta_{opt}\|_2^2 = \mathbb{E}_{XW^\top,W,Y}\left\|\int[R(\|\tilde{\theta}_1\|^2) - R(\alpha^2(1-\pi))]\cdot p(\tilde{\theta}_1|XW^\top,W,Y)d\tilde{\theta}_1\right\|^2 \\
&\le \mathbb{E}_{XW^\top,W,Y,\tilde{\theta}_1}\left\|[R(\|\tilde{\theta}_1\|^2) - R(\alpha^2(1-\pi))\right\|^2 = \mathbb{E}_{XW^\top,W,Y,\tilde{\theta}_1}\left\|\left[\tilde{R}^* - \tilde{R}_1\right]\frac{WX^\top Y}{n}\right\|^2,
\end{aligned}$$

where we used the Jensen inequality in the second line. Note that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ and omit the subscripts of the expectation. The above equals

$$\mathbb{E}\left\|\left[\tilde{R}^*\left(\lambda^* - \lambda(\tilde{\theta}_1)\right)\tilde{R}_1\right]\frac{WX^\top Y}{n}\right\|^2. \tag{85}$$

Furthermore, by Cauchy-Schwartz inequality and the fact that $\|AB\|_F \le \|A\|_2\|B\|_F$, the square of this quantity is upper bounded by

$$\begin{aligned}
&\mathbb{E}\left(\lambda^* - \lambda(\tilde{\theta}_1)\right)^4\cdot\mathbb{E}\left\|\left[\tilde{R}^*\cdot\tilde{R}_1\right]\frac{WX^\top Y}{n}\right\|^4. \\
&\le \mathbb{E}\left(\lambda^* - \lambda(\tilde{\theta}_1)\right)^4\cdot\mathbb{E}\left\|\frac{1}{\lambda(\tilde{\theta}_1)\lambda^*}\frac{WX^\top Y}{n}\right\|^4 \\
&\le \frac{1}{\lambda^{*4}}\mathbb{E}\left(\lambda^* - \lambda(\tilde{\theta}_1)\right)^4\cdot\left(\mathbb{E}\frac{1}{\lambda(\tilde{\theta}_1)^8}\right)^{1/2}\cdot\left(\mathbb{E}\left\|\frac{WX^\top Y}{n}\right\|^8\right)^{1/2}.
\end{aligned} \tag{86}$$

Recall that $Y = X\theta + \mathcal{E}$ and note that $\mathbb{E}\|WX^\top X/n\|_F^k$, $\mathbb{E}\|WX^\top/\sqrt{n}\|_F^k$, $\mathbb{E}\|\theta\|_2^k$, $\mathbb{E}\|\mathcal{E}/\sqrt{n}\|_2^k$ are all uniformly bounded (as $d \to \infty$) for any non-negative integer k because of the Gaussian asusmption and the boundeness of moments of Wishart matrices (Muirhead, 2009; Bai and Silverstein, 2010) etc. It follows directly by several applications of the Cauchy-Schwartz

inequality that $\mathbb{E}\|WX^\top Y/n\|^8$ is bounded as $d \to \infty$, i.e., the third term in the R.H.S. of (86) is bounded.

If $\sigma > 0$, then $\lambda(\tilde{\theta}_1) \geq \lambda(0) = \sigma^2 d/[n\alpha^2] > 0$, and hence the second term in (86) is bounded. If $\sigma = 0$ and $\pi < 1$, by the definition of $\lambda(\tilde{\theta}_1)$ and integration in the polar system, it is readily verified that

$$\mathbb{E}\frac{1}{\lambda(\tilde{\theta}_1)^8} = \mathbb{E}\frac{n^8}{(x_1^2 + \cdots + x_{d-p}^2)^8} = \frac{n^8}{\prod_{k=1}^8 (d-p-2k)} \overset{d\to\infty}{\longrightarrow} C_1 < \infty.$$

Therefore, the second term in the R.H.S of (86) is also bounded. Also

$$\begin{aligned}(86) &\leq C \cdot \lim_{d\to\infty} \mathbb{E}_{XW^\top, W, Y, \theta} \left(\lambda^* - \lambda(\tilde{\theta}_1)\right)^4 \\ &\leq C_2 \cdot \lim_{d\to\infty} \left(\left[\lambda^* - \frac{d}{n}(1 - \frac{p}{d} + \sigma^2/\alpha^2)\right]^4 + \mathbb{E}[\lambda(\tilde{\theta}_1) - \mathbb{E}\lambda(\tilde{\theta}_1)]^4 \right) = 0,\end{aligned}$$

where C, C_2 are some finite constants and the last equality follows from equation (84) and properties of the chi-square distribution.

Therefore, we have proved that $\mathbb{E}\|\hat{\beta} - \beta_{opt}\|^2 \to 0$ and the optimal ridge estimator $\hat{\beta}$ is asymptotically optimal. ■

B.9 Proof of Theorem 4

Proof [Proof of Theorem 4] We only need to make small changes in the proof of theorems 2 and 3 to prove theorem 4. We first take $\lim_{d\to\infty} \mathbf{Bias}^2(\lambda)$ as an example. Similar to (25)

$$\begin{aligned}\mathbf{Bias}^2(\lambda) &= \mathbb{E}_x \left[\mathbb{E}_{X,W,\mathcal{E}} \left(x^\top M\theta + x^\top \tilde{M}\mathcal{E}\right) - x^\top\theta\right]^2 \\ &= \mathbb{E}_x \left[x^\top(\mathbb{E}_{X,W} M - I)\theta\right]^2 = \mathbb{E}_x \operatorname{tr}((\mathbb{E}M^\top - I)xx^\top(\mathbb{E}M - I)\theta\theta^\top) \\ &= \operatorname{tr}((\mathbb{E}M^\top - I)(\mathbb{E}M - I)\theta\theta^\top). \end{aligned} \tag{87}$$

Note that we have shown $\mathbb{E}M$ is a multiple of identity in Lemma 11 (under Gaussian assumption), therefore

$$(87) = \left[(\mathbb{E}M^\top - I)(\mathbb{E}M - I)\right] \cdot \operatorname{tr}(\theta\theta^\top) = \frac{1}{d}\operatorname{tr}\left[(\mathbb{E}M^\top - I)(\mathbb{E}M - I)\right] \cdot \operatorname{tr}(\theta\theta^\top). \tag{88}$$

From Lemma 11, we know the first term in the R.H.S. of (88) converges to $\pi^2(1-\lambda\theta_1(\pi\delta,\lambda))^2$. Note that $\operatorname{tr}(\theta\theta^\top) \overset{d}{=} \alpha^2(\sum_{i=1}^d x_i^2)/d$, where $x_i \sim \mathcal{N}(0,1)$. By the Borel-Cantelli lemma and the concentration inequality for χ^2-variables, we have $\operatorname{tr}(\theta\theta^\top) \overset{a.s.}{\to} \alpha^2$.

Thus, (88) almost surely converges to $\alpha^2\pi^2(1 - \theta_1(\pi\delta,\lambda))^2$ and the same asymptotic result for bias as in theorem 3 holds almost surely over the randomness in θ.

From this example, we can see that the results in theorem 2 and 3 will automatically hold in the non-random setting if we can separate θ from other variables (e.g. $\mathbb{E}M$) in (25) —(27), (32)—(38) by showing that the matrices which are multiplied by $\theta\theta^\top$ are in fact a

multiple of identity. For instance, in (25), since $\theta\theta^\top$ is multiplied by a multiple of identity $(\mathbb{E}M^\top - I)(\mathbb{E}M - I)$, the same result follows.

To generalize other results in theorem 2 and 3 to the almost sure setting, from (25)—(27), (32)—(38), we can see that it is enough to show the following matrices are all multiples of the identity:

$$(a).\ \mathbb{E}M^\top M, \qquad (b).\ \mathbb{E}_W(\mathbb{E}_X M^\top \mathbb{E}_X M), \qquad (c).\ \mathbb{E}_X(\mathbb{E}_W M^\top \mathbb{E}_W M).$$

(a). $\mathbb{E}M^\top M$. We will denote $R = \left(WX^\top XW^\top/n + \lambda I_p\right)^{-1}$ in what follows. By the definition of M and the fact that $XW^\top$ and $X(I - W^\top W)$ (denoted by X_2) are two independent matrices with Gaussian entries for any fixed W with orthogonal rows

$$\begin{aligned}\mathbb{E}M^\top M &= \mathbb{E}\frac{X^\top XW^\top}{n} R^2 \frac{WX^\top X}{n} \\ &= \mathbb{E}\frac{W^\top WX^\top XW^\top}{n} R^2 \frac{WX^\top XW^\top W}{n} \qquad (89) \\ &+ \mathbb{E}\frac{X_2^\top XW^\top}{n} R^2 \frac{WX^\top X_2}{n}. \qquad (90)\end{aligned}$$

For (89), note that W and $XW^\top$ are independent, so

$$(89) = \mathbb{E}_W W^\top \left[\mathbb{E}_{XW^\top} \frac{WX^\top XW^\top}{n} R^2 \frac{WX^\top XW^\top}{n}\right] W.$$

Since $\mathbb{E}_W W^\top AW = \operatorname{tr}(A) \cdot I_d/d$ for any constant matrix A, it follows directly that (89) is a multiple of identity. For (90), note that $XW^\top$ and $X(I - W^\top W)$ are independent conditioned on W, thus

$$(90) = \mathbb{E}_W \mathbb{E}_{X_2|W} X_2^\top \left[\mathbb{E}_{XW^\top} \frac{XW^\top}{n} R^2 \frac{WX^\top}{n}\right] X_2. \qquad (91)$$

Let $XW^\top = UDV^\top$ be the SVD. Since $XW^\top$ has i.i.d. $N(0,1)$ entries, we can assume U follows the Haar measure and is independent of $DV^\top$. Therefore, with $X_2 = X(I - W^\top W)$

$$\begin{aligned}(91) &= \mathbb{E}_W \mathbb{E}_{X_2|W} X_2^\top \left[\mathbb{E}_{U,D,V} U \frac{DV^\top}{n} \left(\frac{VD^\top DV^\top}{n} + \lambda I_p\right)^{-2} \frac{VD^\top}{n} U^\top\right] X_2 \\ &= \mathbb{E}_W \mathbb{E}_{X_2|W} X_2^\top \left\{\frac{1}{n}\operatorname{tr}\left[\mathbb{E}_{D,V} \frac{DV^\top}{n} \left(\frac{VD^\top DV^\top}{n} + \lambda I_p\right)^{-2} \frac{VD^\top}{n}\right]\right\} X_2 \\ &= c_0 \cdot \mathbb{E}_W \mathbb{E}_{X_2|W} X_2^\top X_2 = c_1 \cdot \mathbb{E}_W(I - W^\top W) = c_2 \cdot I_d,\end{aligned}$$

where c_0, c_1, c_2 are some constants. Combining (89) and (90), we have proved $\mathbb{E}M^\top M$ is a multiple of identity.

(b). $\mathbb{E}_W(\mathbb{E}_X M^\top \mathbb{E}_X M)$. Since W and $XW^\top$ are independent, similarly

$$\begin{aligned}\mathbb{E}_X M &= \mathbb{E}_X W^\top R \frac{WX^\top XW^\top}{n} W = W^\top \left[\mathbb{E}_{XW^\top|W} R \frac{WX^\top XW^\top}{n}\right] W \\ &= W^\top \mathbb{E}_V V \mathbb{E}_D \left(\frac{D^\top D}{n} + \lambda I_p\right)^{-1} \frac{D^\top D}{n} V^\top W = c_0 W^\top W,\end{aligned}$$

where c_0 is a constant (different from previous constants) and the last line follows from the fact that $\mathbb{E}_V VAV^\top = \mathrm{tr}(A)/p \cdot I_p$ for any constant matrix A. Therefore

$$\mathbb{E}_W(\mathbb{E}_X M^\top \mathbb{E}_X M) = \mathbb{E}_W c_0^2 W^\top W W^\top W = c_0^2 \cdot \mathbb{E}_W W^\top W = c_1 \cdot I_d,$$

where c_1 is a constant.

(c). $\mathbb{E}_X(\mathbb{E}_W M^\top \mathbb{E}_W M)$. Let $X^\top X = U\Gamma U^\top$ be the spectral decomposition of $X^\top X$. By the definition of M, we have

$$\begin{aligned}
\mathbb{E}_W M &= \mathbb{E}_W \left[W^\top R W\right] \frac{X^\top X}{n} \\
&= \mathbb{E}_W U \left[(WU)^\top \left(\frac{(WU)\Gamma(WU)^\top}{n} + \lambda I_p\right)^{-1} WU\right] \frac{\Gamma U^\top}{n} \\
&= \mathbb{E}_W \left[W^\top \left(\frac{W\Gamma W^\top}{n} + \lambda I_p\right)^{-1} W\right] \frac{\Gamma U^\top}{n} = c_0(\Gamma) \cdot I_d \cdot \frac{U\Gamma U^\top}{n},
\end{aligned}$$

where $c_0(\Gamma)$ is a constant depending on Γ, the second line is due to $W \overset{d}{=} WU$ and the last line follows from the proof of Lemma 22 (see (66)). Finally, note that U follows the Haar distribution and is independent of Γ. Thus,

$$\mathbb{E}_X(\mathbb{E}_W M^\top \mathbb{E}_W M) = \mathbb{E}_{\Gamma,U} c_0(\Gamma)^2 \cdot \frac{U\Gamma^2 U^\top}{n} = \mathbb{E}_U U \left(\mathbb{E}_\Gamma c_0(\Gamma)^2 \cdot \frac{\Gamma^2}{n}\right) U^\top = c_1 \cdot I_d,$$

where c_1 is a constant and the last equality follows again from the fact that $\mathbb{E}_U UAU^\top = \mathrm{tr}(A)/d \cdot I_d$ for any constant matrix A.

■

B.10 Proof of Theorem 9

Proof By definition, we have

$$\begin{aligned}
\mathbf{MSE}(\lambda) &:= \mathbb{E}_{\theta,x,W,X,\mathcal{E}}(\hat{f}(x) - x^\top\theta)^2 + \sigma^2 \\
&= \mathbb{E}_{\theta,x,W,X,\mathcal{E}} \hat{f}(x)^2 - 2\mathbb{E}_{\theta,x,W,X,\mathcal{E}} \hat{f}(x) \cdot x^\top\theta + \mathbb{E}_{\theta,x}(x^\top\theta)^2 + \sigma^2. \\
\mathbf{Bias}^2(\lambda) &:= \mathbb{E}_{\theta,x}|\mathbb{E}_{W,X,\mathcal{E}} \hat{f}(x) - x^\top\theta|^2 \\
&= \mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}} \hat{f}(x))^2 - 2\mathbb{E}_{\theta,x,W,X,\mathcal{E}} \hat{f}(x) \cdot x^\top\theta + \mathbb{E}_{\theta,x}(x^\top\theta)^2 \\
\mathbf{Var}(\lambda) &:= \mathbb{E}_{\theta,x,W,X,\mathcal{E}}|\hat{f}(x) - \mathbb{E}_{X,W,\mathcal{E}} \hat{f}(x)|^2 \\
&= \mathbb{E}_{\theta,x,W,X,\mathcal{E}} \hat{f}(x)^2 - \mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}} \hat{f}(x))^2.
\end{aligned}$$

To prove equation (22), (23) and (24), it is thus enough to calculate $\mathbb{E}_{\theta,x,W,X,\mathcal{E}} \hat{f}(x)^2$, $\mathbb{E}_{\theta,x,W,X,\mathcal{E}} \hat{f}(x) \cdot x^\top\theta$ and $\mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}} \hat{f}(x))^2$. In Lemma 24, 25 and 26, we calculate these three terms separately. Equation (22), (23) and (24) follow directly from these three lemmas and the fact that $\mathbb{E}_{\theta,x}(x^\top\theta)^2 = \alpha^2$.

Using the same technique as in the proof of theorem 3, it can be shown that the limiting MSE as a function of λ has a unique minimum at $\lambda^* := \frac{v^2}{\mu^2}\left[\delta(1-\pi+\frac{\sigma^2}{\alpha^2})+\frac{(v-\mu^2)\gamma}{v}\right]$. Furthermore, results in Table 2 can be proved in the same way as results in Table 1. For simplicity, we omit the proof of optimal λ^* and Table 2 here. ■

Lemma 24 (Asymptotic limit of $\mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}}\hat{f}(x))^2$) *Under assumptions in theorem 9*

$$\lim_{d\to\infty}\mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}}\hat{f}(x))^2=\alpha^2\pi^2\frac{\mu^4}{v^2}\left(1-\frac{\lambda}{v}\theta_1\right)^2, \tag{92}$$

where $\theta_1:=\theta_1(\gamma,\lambda/v)$.

Lemma 25 (Asymptotic limit of $\mathbb{E}_{\theta,x,W,X,\mathcal{E}}\hat{f}(x)\cdot x^\top\theta$) *Under assumptions in theorem 9*

$$\lim_{d\to\infty}\mathbb{E}_{\theta,x,W,X,\mathcal{E}}\hat{f}(x)\cdot x^\top\theta=\alpha^2\pi\frac{\mu^2}{v}\left(1-\frac{\lambda}{v}\theta_1\right), \tag{93}$$

where $\theta_1:=\theta_1(\gamma,\lambda/v)$.

Lemma 26 (Asymptotic limit of $\mathbb{E}_{\theta,x,W,X,\mathcal{E}}\hat{f}(x)^2$) *Under assumptions in theorem 9 we have*

$$\begin{aligned}\lim_{d\to\infty}\mathbb{E}_{\theta,x,W,X,\mathcal{E}}\hat{f}(x)^2=\alpha^2\pi\Bigg[1-\frac{2(v-\mu^2)}{v}+\left(\rho(1-\pi)-\frac{2\lambda\mu^2}{v^2}\right)\theta_1+\frac{\lambda}{v}\left(\frac{\lambda\mu^2}{v^2}-\rho(1-\pi)\right)\theta_2\\ +\frac{v-\mu^2}{v}\left(1+\gamma\theta_1-\frac{\lambda\gamma}{v}\theta_2\right)\Bigg]+\sigma^2\gamma\left(\theta_1-\frac{\lambda}{v}\theta_2\right),\end{aligned} \tag{94}$$

where $\theta_1:=\theta_1(\gamma,\lambda/v)$ *and* $\theta_2:=\theta_2(\gamma,\lambda/v)$.

The proofs of these lemmas proceed by applying the leave-one-out technique and the Marchenko-Pastur law (refer to Lemmas 27—29), and by leveraging properties of the orthogonal projection matrix and the normal distribution.

Proof [Proof of Lemma 24] Different from previous sections, we will denote $R=\left(\frac{\sigma(WX^\top)\sigma(XW^\top)}{n}+\lambda\right)^{-1}$ in the proof of Lemmas 24—26. By definition,

$$\begin{aligned}\mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}}\hat{f}(x))^2&=\mathbb{E}_{\theta,x}\left[\mathbb{E}_{W,X,\mathcal{E}}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)Y}{n}\right]^2\\ &=\mathbb{E}_{\theta,x}\left[\mathbb{E}_{W,X}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)X\theta}{n}\right]^2\\ &=\frac{\alpha^2}{d}\mathbb{E}_x\left\|\mathbb{E}_{W,X}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)X}{n}\right\|_F^2=:\frac{\alpha^2}{d}\mathbb{E}_x\|T_1\|_F^2,\end{aligned} \tag{95}$$

For T_1, we continue

$$\begin{aligned} T_1 &= \mathbb{E}_{W,X}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)}{n}[XW^\top W + X(I - W^\top W)] \\ &= \mathbb{E}_{W,X}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)}{n}XW^\top W \\ &= \mathbb{E}_W\sigma(x^\top W^\top)\mathbb{E}_X\left[R\frac{\sigma(WX^\top)}{n}XW^\top\right]W, \end{aligned} \tag{96}$$

where the second line follows from the fact that $XW^\top$ and $X(I - W^\top W)$ are independent and $\mathbb{E}_X X(I - W^\top W)$. Denote

$$D_1 := \mathbb{E}_X\left[R\frac{\sigma(WX^\top)}{n}XW^\top\right],$$

then D_1 is a constant matrix independent of W since WX has i.i.d. $\mathcal{N}(0,1)$ entries for any W with orthonormal rows. We can write the vector x as $x = U(\sqrt{d_1}, 0, ..., 0)^\top$ U is a random orthogonal matrix following the Haar distribution. Denoting the i-th column of W by $W_{\cdot i}$ and substituting x into (96), we get

$$\begin{aligned} T_1 &= \mathbb{E}_W\sigma(x^\top W^\top)D_1 W = \mathbb{E}_W\sigma((\sqrt{d_1}, 0, ..., 0)U^\top W^\top)D_1 WUU^\top \\ &= \mathbb{E}_W\sigma((\sqrt{d_1}, 0, ..., 0)W^\top)D_1 WU^\top \end{aligned}$$

where the last line follows from $W \overset{d}{=} WU$. We can further write this as

$$\begin{aligned} &\mathbb{E}_W\sigma(\sqrt{d_1}W_{\cdot 1}^\top)D_1(W_{\cdot 1}, ..., W_{\cdot d})U^\top \\ &= \left(\mathbb{E}_W\sigma(\sqrt{d_1}W_{\cdot 1}^\top)D_1 W_{\cdot 1}, 0, ..., 0\right)U^\top \\ &= \left(\mathbb{E}_{W_{\cdot 1}}\operatorname{tr}[W_{\cdot 1}\sigma(\sqrt{d_1}W_{\cdot 1}^\top)D_1], 0, ..., 0\right)U^\top \\ &= \left(\mathbb{E}_{W_{11}}[W_{11}\sigma(\sqrt{d_1}W_{11})]\operatorname{tr}(D_1), 0, ..., 0\right)U^\top, \end{aligned} \tag{97}$$

where the second line follows from the symmetry of $W_{\cdot i}(i \geq 2)$ conditional on $W_{\cdot 1}$ and the last line is due to the fact that $\mathbb{E}_{W_{\cdot 1}}W_{\cdot 1}\sigma(\sqrt{d_1}W_{\cdot 1}^\top)$ is a multiple of identity since $W_{i,1}$ is symmetric conditional on $W_{j,1}$, $j \neq i$. Therefore,

$$\begin{aligned} \mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}}\hat{f}(x))^2 &= \frac{\alpha^2}{d}\mathbb{E}_x\|T_1\|_F^2 \\ &= \frac{\alpha^2}{d}\mathbb{E}_{d_1}\left[\mathbb{E}_{W_{11}}W_{11}\sigma(\sqrt{d_1}W_{11})\right]^2\operatorname{tr}(D_1)^2. \end{aligned} \tag{98}$$

Denote $\sqrt{d_1}W_{11}$ by $\tilde{w}$. Noting that $\tilde{w} \xrightarrow{w} \mathcal{N}(0,1)$ and $d_1/d = \|x\|^2/d \xrightarrow{a.s.} 1$, we obtain

$$\begin{aligned} &\lim_{d\to\infty}\mathbb{E}_{\theta,x}(\mathbb{E}_{W,X,\mathcal{E}}\hat{f}(x))^2 = \lim_{d\to\infty}\frac{\alpha^2}{d^2}\operatorname{tr}(D_1)^2\mathbb{E}_{d_1}\left[\mathbb{E}_{\tilde{w}}\tilde{w}\sigma\left(\sqrt{\frac{d_1}{d}}\tilde{w}\right)\right]^2 \\ &= \lim_{d\to\infty}\alpha^2\pi^2\frac{\mu^2}{v^2}\left(1 - \frac{\lambda}{v}\theta_1\right)^2\mathbb{E}_{d_1/d}\left[\mathbb{E}_{\tilde{w}}\tilde{w}\sigma\left(\sqrt{\frac{d_1}{d}}\tilde{w}\right)\right]^2 \\ &= \alpha^2\pi^2\frac{\mu^2}{v^2}\left(1 - \frac{\lambda}{v}\theta_1\right)^2[\mathbb{E}_{a\sim\mathcal{N}(0,1)}\sigma(a)a]^2 = \alpha^2\pi^2\frac{\mu^4}{v^2}\left(1 - \frac{\lambda}{v}\theta_1\right)^2, \end{aligned}$$

where the third line can be rigorously justified using the concentration inequality for $\tilde{w}$ and d_1/d:

$$P(|\tilde{w}| > t) \le 2e^{-(d-2)t^2/d} \text{ (Levy's lemma)}, \quad P(|d_1/d - 1| > t) \le 2e^{-dt^2/8} \text{ (concentration for } \chi^2).$$

the fact that $|\sigma(x)|, |\sigma'(x)| \le c_1 e^{c_2|x|}$, integration by parts and the bounded convergence theorem. Here Levy's lemma refers to usual concentration of the Haar measure (Boucheron et al., 2013). ■

Proof [Proof of Lemma 25] By definition,

$$\begin{aligned}
&\mathbb{E}_{\theta,x}(\mathbb{E}_{X,W,\mathcal{E}}\hat{f}(x)\theta^\top x) = \mathbb{E}_{\theta,x}\left[\mathbb{E}_{W,X,\mathcal{E}}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)Y}{n}\theta^\top x\right] \\
&= \mathbb{E}_{\theta,x}\left[\mathbb{E}_{W,X}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)X\theta}{n}\theta^\top x\right] \\
&= \frac{\alpha^2}{d}\mathbb{E}_x\left[\mathbb{E}_{W,X}\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)Xx}{n}\right] = \frac{\alpha^2}{d}\mathbb{E}_x\,\mathrm{tr}(T_1 x),
\end{aligned}$$

where T_1 is defined in equation (95) in the proof of Lemma 24. Let us again write $x = U(\sqrt{d_1}, 0, ..., 0)^\top$ for an orthogonal U, and denote $\tilde{w} = \sqrt{d}W_{11}$. Then from equation (97) and the fact that $\tilde{w} \xrightarrow{w} \mathcal{N}(0,1)$, $d_1/d \xrightarrow{a.s.} 1$ we have

$$\begin{aligned}
\lim_{d\to\infty}\mathbb{E}_{\theta,x}(\mathbb{E}_{X,W,\mathcal{E}}\hat{f}(x)\theta^\top x) &= \lim_{d\to\infty}\frac{\alpha^2}{d}\mathbb{E}_{d_1}\left(\mathbb{E}_{W_{11}}[W_{11}\sigma(\sqrt{d_1}W_{11})]\,\mathrm{tr}(D_1), 0, ..., 0\right)U^\top U(\sqrt{d_1}, 0, ..., 0)^\top \\
&= \lim_{d\to\infty}\frac{\alpha^2}{d}\,\mathrm{tr}(D_1)\mathbb{E}_{d_1}\mathbb{E}_{W_{11}}[\sqrt{d_1}W_{11}\sigma(\sqrt{d_1}W_{11})] \\
&= \lim_{d\to\infty}\alpha^2\pi\frac{\mu}{v}\left(1 - \frac{\lambda}{v}\theta_1\right)\mathbb{E}_{d_1/d}\mathbb{E}_{W_{11}}\left[\sqrt{\frac{d_1}{d}}\tilde{w}\sigma\left(\sqrt{\frac{d_1}{d}}\tilde{w}\right)\right] \\
&= \alpha^2\pi\frac{\mu}{v}\left(1 - \frac{\lambda}{v}\theta_1\right)\mathbb{E}_{a\sim\mathcal{N}(0,1)}[a\sigma(a)] = \alpha^2\pi\frac{\mu^2}{v}\left(1 - \frac{\lambda}{v}\theta_1\right),
\end{aligned}$$

where the third line follows from similar arguments as in the proof of Lemma 24. ■

Proof [Proof of Lemma 26] By definition,

$$\begin{aligned}
&\mathbb{E}_{\theta,x,W,X,\mathcal{E}}\hat{f}(x)^2 = \mathbb{E}_{\theta,x,W,X,\mathcal{E}}\left|\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)(X\theta + \mathcal{E})}{n}\right|^2 \\
&= \frac{\alpha^2}{d}\mathbb{E}\left\|\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)X}{n}\right\|_2^2 + \sigma^2\mathbb{E}\left\|\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)}{n}\right\|_2^2 =: T_2 + T_3.
\end{aligned}$$

For T_2, we further have that it equals

$$\begin{aligned}
&\frac{\alpha^2}{d}\mathbb{E}_{W,X,x}\left\|\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)XW^\top}{n}\right\|_2^2+\frac{\alpha^2}{d}\mathbb{E}_{W,X,x}\left\|\sigma(x^\top W^\top)R\frac{\sigma(WX^\top)X(I-W^\top W)}{n}\right\|_2^2\\
&=\frac{v\alpha^2}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)XW^\top}{n}\right\|_F^2+\frac{v\alpha^2}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)(I_d-W^\top W)}{n}\right\|_F^2\\
&=\frac{v\alpha^2}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)XW^\top}{n}\right\|_F^2+\frac{v\alpha^2(d-p)}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)}{n}\right\|_F^2=:T_4+T_5,
\end{aligned}$$

where the first and third equations follow from the fact that $XW^\top$ and $X(I-W^\top W)$ are independent conditional on W and $\mathbb{E}_X X(I_d-W^\top W)(I_d-W^\top W)X^\top=\mathrm{tr}(I_d-W^\top W)=d-p$. Also, we used that $Wx\sim N(0,I_p)$ given any orthgonal W, $\mathbb{E}_{a\sim\mathcal{N}(0,1)}\sigma(a)=0$, definition (21) and the independence of Wx, $XW^\top$ and $X(I-W^\top W)$ conditional on W.

Also, due to the independence of Wx, $XW^\top$ conditional on W, we have

$$T_3=\sigma^2 v\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)}{n}\right\|_F^2=\frac{\sigma^2 d}{\alpha^2(d-p)}T_5.$$

Finally, substituting Lemma 28, 29 into T_4, T_5 , we get

$$\begin{aligned}
&\mathbb{E}_{\theta,x,W,X,\mathcal{E}}\hat{f}(x)^2=T_3+T_4+T_5=T_4+\left(1+\frac{\sigma^2 d}{\alpha^2(d-p)}\right)T_5\\
&=\frac{v\alpha^2}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)XW^\top}{n}\right\|_F^2+\left(\frac{v\alpha^2(d-p)}{d}+v\sigma^2\right)\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)}{n}\right\|_F^2\\
&\to\alpha^2\pi\left[1-\frac{2(v-\mu^2)}{v}+\left(\rho(1-\pi)-\frac{2\lambda\mu^2}{v^2}\right)\theta_1+\frac{\lambda}{v}\left(\frac{\lambda\mu^2}{v^2}-\rho(1-\pi)\right)\theta_2\right.\\
&\left.+\frac{v-\mu^2}{v}\left(1+\gamma\theta_1-\frac{\lambda\gamma}{v}\theta_2\right)\right]+\sigma^2\gamma\left(\theta_1-\frac{\lambda}{v}\theta_2\right)
\end{aligned}$$

■

Lemma 27 (Asymptotic behavior of D_1) *Under assumptions in theorem 9, we have*

$$\lim_{d\to\infty}\frac{1}{d}\mathbb{E}_X\,\mathrm{tr}\left[R\frac{\sigma(WX^\top)}{n}XW^\top\right]=\pi\frac{\mu}{v}\left(1-\frac{\lambda}{v}\theta_1\right),\tag{99}$$

where $\theta_1:=\theta_1(\gamma,\lambda/v)$.

Proof [Proof of Lemma 27] Since $WW^\top=I_p$, $WX^\top$ has i.i.d. $\mathcal{N}(0,1)$ entries and the L.H.S. of (99) (if exists) is a constant independent of W. Denote $\widetilde{X}:=XW^\top$ for notational simplicity. Also, let $\widetilde{X}_{\cdot i}$ be the i-th column of $\widetilde{X}$, $\widetilde{X}_{\cdot -i}$ be the matrix obtained by deleting the i-th column of $\widetilde{X}$. By symmetry of $\widetilde{X}$, it suffices to compute the first diagonal entry of

the matrix in the L.H.S. of (99). Namely, we have

$$\begin{aligned}
&\frac{1}{d}\mathbb{E}_X \operatorname{tr}\left[R\frac{\sigma(WX^\top)}{n}XW^\top\right] = \frac{p}{d}\mathbb{E}_{\widetilde{X}}\left[R\frac{\sigma(\widetilde{X}^\top)}{n}\widetilde{X}\right]_{11} \\
&= \frac{p}{d}\mathbb{E}_{\widetilde{X}}\left[\frac{\sigma(\widetilde{X}^\top)}{n}\left(\frac{\sigma(\widetilde{X})\sigma(\widetilde{X}^\top)}{n}+\lambda I_n\right)^{-1}\widetilde{X}\right]_{11} \\
&= \frac{p}{d}\mathbb{E}_{\widetilde{X}}\left[\frac{\sigma(\widetilde{X}_{\cdot 1})^\top}{n}\left(\frac{\sigma(\widetilde{X}_{\cdot 1})\sigma(\widetilde{X}_{\cdot 1}^\top)}{n}+\frac{\sigma(\widetilde{X}_{\cdot -1})\sigma(\widetilde{X}_{\cdot -1}^\top)}{n}+\lambda I_n\right)^{-1}\widetilde{X}_{\cdot 1}\right]. \qquad (100)
\end{aligned}$$

Define $C := [\sigma(\widetilde{X}_{\cdot -1})\sigma(\widetilde{X}_{\cdot -1}^\top)/n+\lambda I_n]$, $u := \widetilde{X}_{\cdot 1}$ and $\tilde{u} := \sigma(\widetilde{X}_{\cdot 1})$. By the Sherman-Morrison formula, the above equals

$$\frac{p}{d}\mathbb{E}\frac{u^\top}{n}\left(C^{-1}-\frac{C^{-1}\tilde{u}\tilde{u}^\top C^{-1}/n}{1+\tilde{u}^\top C^{-1}\tilde{u}/n}\right)u = \frac{p}{d}\left[\mathbb{E}\frac{\tilde{u}^\top C^{-1}u}{n}-\mathbb{E}\left(\frac{u^\top C^{-1}\tilde{u}\tilde{u}C^{-1}u/n}{n+\tilde{u}^\top C^{-1}\tilde{u}}\right)\right]. \qquad (101)$$

Since $\tilde{u} = \sigma(\widetilde{X}_{\cdot 1})$ has i.i.d. zero mean v variance entries, by theorem 1 in Rubio and Mestre (2011), the proof of theorem 2.1 in Liu and Dobriban (2020) , we know that C^{-i} are determinstically equivalent to certain multiples of the identity matrix. Also, the multiples will converge to certain limits, which can be determined by the Marchenko-Pastur law. Thus, we have after some calculations that

$$C^{-1} = \left(\frac{\sigma(\widetilde{X}_{\cdot -1})\sigma(\widetilde{X}_{\cdot -1}^\top)}{n}+\lambda I_n\right)^{-1} \asymp \frac{1}{\gamma v}\theta_1\left(\frac{1}{\gamma},\frac{\lambda}{\gamma v}\right)\cdot I_n = \frac{\gamma}{v}\theta_1\left(\gamma,\frac{\lambda}{v}\right)+\frac{1-\gamma}{\lambda}. \qquad (102)$$

$$C^{-2} = \left(\frac{\sigma(\widetilde{X}_{\cdot -1})\sigma(\widetilde{X}_{\cdot -1}^\top)}{n}+\lambda I_n\right)^{-2} \asymp \frac{1}{\gamma^2 v^2}\theta_2\left(\frac{1}{\gamma},\frac{\lambda}{\gamma v}\right)\cdot I_n = \frac{\gamma}{v^2}\theta_2\left(\gamma,\frac{\lambda}{v}\right)+\frac{1-\gamma}{\lambda^2}. \qquad (103)$$

Therefore, it remains to calculate the limit of (101).

Since $u \sim \mathcal{N}(0, I_n)$, we have by the strong law of large numbers that

$$\limsup\left\|\frac{uu^\top}{n}\right\|_{\mathrm{tr}} = \limsup u^\top u/n <_{a.s.} \infty.$$

Note that C is independent of u, thus we have by (102) that almost surely for a sequence of u_k, $k = 1, 2, ...$

$$\begin{aligned}
&\lim_{d\to\infty}\frac{u^\top C^{-1}u}{n} = \lim_{d\to\infty}\operatorname{tr}\left(\frac{(uu^\top)}{n}C^{-1}\right) \\
&= \lim_{d\to\infty}\left[\frac{\gamma}{v}\theta_1\left(\gamma,\frac{\lambda}{v}\right)+\frac{1-\gamma}{\lambda}\right]\operatorname{tr}\left(\frac{(uu^\top)}{n}\right) = \left[\frac{\gamma}{v}\theta_1\left(\gamma,\frac{\lambda}{v}\right)+\frac{1-\gamma}{\lambda}\right], \text{a.s.}
\end{aligned}$$

Similar results also hold for other terms: $(i = 1, 2)$

$$\frac{u^\top C^{-i} u}{n} \xrightarrow{a.s.} \left[\frac{\gamma}{v^i}\theta_i\left(\gamma, \frac{\lambda}{v}\right) + \frac{1-\gamma}{\lambda^i}\right]. \tag{104}$$

$$\frac{\tilde{u}^\top C^{-i} u}{n} \xrightarrow{a.s.} \mu\left[\frac{\gamma}{v^i}\theta_i\left(\gamma, \frac{\lambda}{v}\right) + \frac{1-\gamma}{\lambda^i}\right]. \tag{105}$$

$$\frac{\tilde{u}^\top C^{-i} \tilde{u}}{n} \xrightarrow{a.s.} v\left[\frac{\gamma}{v^i}\theta_i\left(\gamma, \frac{\lambda}{v}\right) + \frac{1-\gamma}{\lambda^i}\right]. \tag{106}$$

Therefore, simple calculations give

$$\left[\frac{\tilde{u}^\top C^{-1} u}{n} - \left(\frac{u^\top C^{-1}\tilde{u}\tilde{u}C^{-1}u/n^2}{1+\tilde{u}^\top C^{-1}\tilde{u}/n}\right)\right] \to \frac{\mu}{v}\left(1 - \frac{\lambda}{v}\theta_1\right), \text{a.s.} \tag{107}$$

Now, we only need to show that the expectation of the L.H.S. of (107) also converges to its pointwise limit.

To prove this, we start with bounding the mean squared error of (104).

$$\begin{aligned}
&\mathbb{E}\left|\frac{u^\top C^{-1}u}{n} - \left[\frac{\gamma}{v}\theta_1 + \frac{1-\gamma}{\lambda}\right]\right|^2 \le \mathbb{E}\left|\frac{u^\top C^{-1}u}{n} - \mathbb{E}_u\frac{u^\top C^{-1}u}{n}\right|^2 + \mathbb{E}\left|\mathbb{E}_u\frac{u^\top C^{-1}u}{n} - \left[\frac{\gamma}{v}\theta_1 + \frac{1-\gamma}{\lambda}\right]\right|^2 \\
&= \frac{1}{n^2}\mathbb{E}_C\left[\left(\sum_{i=j=k=l} + \sum_{i=j\neq k=l} + \sum_{i=l\neq j=k} + \sum_{i=k\neq j=k}\right) C^{-1}_{ij}C^{-1}_{kl}\mathbb{E}u_iu_ju_ku_l - \mathrm{tr}(C^{-1})^2\right] \\
&+ \mathbb{E}\left|\frac{\mathrm{tr}(C^{-1})}{n} - \left[\frac{\gamma}{v}\theta_1 + \frac{1-\gamma}{\lambda}\right]\right|^2.
\end{aligned}$$

This can be further bounded by

$$\frac{1}{n^2}K_1\|C^{-1}\|_F^2 + \mathbb{E}\left|\frac{\mathrm{tr}(C^{-1})}{n} - \left[\frac{\gamma}{v}\theta_1 + \frac{1-\gamma}{\lambda}\right]\right|^2 \to 0,$$

where K_1 is a constant independent of n. This follows from some simple calculations, and the convergence is due to $C^{-1} \preceq \lambda^{-1}$ and the bounded convergence theorem.

Similarly, we can prove the same results for the other five terms corresponding to equation (104), (105) and (106). With the mean squared error converging to zero, we are now able to show that the expectation of the L.H.S. of (107) also converges to the pointwise constant limit.

For notational simplicity, we further denote $a := u^\top C^{-1}\tilde{u}/n, b := \tilde{u}^\top C^{-1}\tilde{u}/n, c := u^\top C^{-1}u/n$ and define constants $A := \lim_{d\to\infty} a, B := \lim_{d\to\infty} b$. Now, for the second term in the L.H.S. of (107), we have

$$\begin{aligned}
&\lim_{d\to\infty}\mathbb{E}\left|\frac{a^2}{1+b} - \frac{A^2}{1+B}\right| \le \lim_{d\to\infty}\mathbb{E}\left|a\frac{A-a}{1+b}\right| + \left|A\frac{A-a}{1+b}\right| + \left|\frac{A^2(B-b)}{(1+B)(1+b)}\right| \\
&\le \sqrt{\mathbb{E}(A-a)^2\mathbb{E}\left|\frac{a}{1+b}\right|^2} + \sqrt{\mathbb{E}(A-a)^2\mathbb{E}\left|\frac{A}{1+b}\right|^2} + \sqrt{\mathbb{E}(B-b)^2\mathbb{E}\left|\frac{A^2}{(1+b)(1+B)}\right|^2} \\
&\le \sqrt{K_2\mathbb{E}(A-a)^2} + \sqrt{K_3\mathbb{E}(A-a)^2} + \sqrt{K_4\mathbb{E}(B-b)^2} \to 0,
\end{aligned} \tag{108}$$

where K_2, K_3, K_4 are some constants independent of n. The existence of K_2, K_3, K_4 is clear, and here we only take K_2 as an example. Since

$$\mathbb{E}\left|\frac{a}{1+b}\right|^2 \leq \mathbb{E}a^2 \leq \mathbb{E}bc \leq \frac{1}{\lambda^2 n^2}\mathbb{E}u^\top u\tilde{u}^\top\tilde{u} \to \frac{2\mu+v}{\lambda^2},$$

we can choose $K_2 := \sup_n \mathbb{E}u^\top u\tilde{u}^\top\tilde{u}/n^2\lambda^2 < \infty$. (The second inequality can be proved by comparing each term in the expression of a^2 and bc and noting that $C_{ii}^{-1}C_{jj}^{-1} \geq (C_{ij}^{-1})^2$ holds for any i, j.)

Finally, noting $\lim_{d\to\infty}\mathbb{E}a = \lim_{d\to\infty}\mathbb{E}\mu\,\mathrm{tr}(C^{-1})/n \to A$ and using (108), we obtain that the expectation of L.H.S. of (107) satisfies

$$\lim_{d\to\infty}\frac{p}{d}\mathbb{E}\left(a - \frac{a^2}{1+b}\right) = \pi\left(A - \frac{A^2}{1+B}\right) = \pi\frac{\mu}{v}\left[1 - \frac{\lambda}{v}\theta_1\left(\gamma, \frac{\lambda}{v}\right)\right].$$

This finishes the proof. ■

Lemma 28 (Asymptotic behavior of T_4) *Under assumptions in theorem 9, we have*

$$\begin{aligned}&\lim_{d\to\infty}\frac{1}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)XW^\top}{n}\right\|_F^2 \\ &= \frac{\pi}{v}\left[1 - \frac{2(v-\mu^2)}{v} - \frac{2\lambda\mu^2}{v^2}\theta_1 + \frac{\lambda^2\mu^2}{v^3}\theta_2 + \frac{v-\mu^2}{v}\left(1 + \gamma\theta_1 - \frac{\lambda\gamma}{v}\theta_2\right)\right],\end{aligned} \tag{109}$$

where $\theta_1 := \theta_1(\gamma, \lambda/v), \theta_2 := \theta_2(\gamma, \lambda/v)$.

Proof [Proof of Lemma 28] Denote $\tilde{R} = \left(\frac{\sigma(XW^\top)\sigma(WX^\top)}{n} + \lambda\right)^{-1}$. By definition,

$$\begin{aligned}&\frac{1}{d}\mathbb{E}_{W,X}\left\|R\frac{\sigma(WX^\top)XW^\top}{n}\right\|_F^2 = \frac{1}{d}\mathbb{E}_{W,X}\left\|\frac{\sigma(WX^\top)}{\sqrt{n}}\tilde{R}\frac{XW^\top}{\sqrt{n}}\right\|_F^2 \\ &= \frac{1}{nd}\mathbb{E}_{W,X}\,\mathrm{tr}\left[WX^\top\tilde{R}XW^\top\right] - \frac{\lambda}{nd}\mathbb{E}_{W,X}\,\mathrm{tr}\left[WX^\top\tilde{R}^2XW^\top\right] =: M_1 - \lambda M_2.\end{aligned}$$

Using the same notations and techniques as in the proof of Lemma 27, after some similar calculations, we get

$$\lim_{d\to\infty} M_i = \lim_{d\to\infty}\frac{\pi}{n}\mathbb{E}u^\top\left(C^{-1} - \frac{C^{-1}\tilde{u}\tilde{u}^\top C^{-1}}{n + \tilde{u}^\top C^{-1}\tilde{u}}\right)^i u, \qquad i = 1, 2.$$

More specifically,

$$\lim_{d\to\infty} M_1 = \pi\lim_{d\to\infty}\mathbb{E}\left(\frac{u^\top C^{-1}u}{n} - \frac{u^\top C^{-1}\tilde{u}\tilde{u}^\top C^{-1}u/n^2}{1 + \tilde{u}^\top C^{-1}\tilde{u}/n}\right). \tag{110}$$

$$\lim_{d\to\infty} M_2 = \pi\lim_{d\to\infty}\mathbb{E}\left(\frac{u^\top C^{-2}u}{n} - 2\frac{u^\top C^{-2}\tilde{u}\tilde{u}^\top C^{-1}u/n^2}{1 + \tilde{u}^\top C^{-1}\tilde{u}/n} + \frac{u^\top C^{-1}\tilde{u}\tilde{u}^\top C^{-2}\tilde{u}\tilde{u}^\top C^{-1}u/n^3}{(1 + \tilde{u}^\top C^{-1}\tilde{u}/n)^2}\right). \tag{111}$$

Substituting equation (104), (105) and (106) into (110), (111), we can see that the random variables on the R.H.S. of (110), (111) almost surely converge to some constant. Using a similar argument as in the proof of Lemma 27, it can be shown that the expectations on the R.H.S. of equation (110), (111) both converge to their corresponding pointwise constant limits.

Therefore, denoting the R.H.S. of (104) by $k_i (i = 1, 2)$ and replacing the R.H.S of (110), (111) by their pointwise constant limits, we obtain

$$\lim_{d\to\infty} M_1 - \lambda M_2 = \pi \left(k_1 - \frac{\mu^2 k_1^2}{1 + v k_1} \right) - \lambda\pi \left[k_2 - \frac{2\mu^2 k_1 k_2}{1 + v k_1} + \frac{\mu^2 v k_1^2 k_2}{(1 + v k_1)^2} \right]. \tag{112}$$

From the remark after definition 1, it is readily verified that $1/(1 + vk_1) = \lambda\theta_1/v$. Thus (112) is a polynomial function of $\theta_{1,2}$. Finally, canceling the high order (≥ 2) terms in (112) using the equations in the remark after definition 1, we have after some calculations that

$$\lim_{d\to\infty} M_1 - \lambda M_2 = \frac{\pi}{v} \left[1 - \frac{2(v - \mu^2)}{v} - \frac{2\lambda\mu^2}{v^2}\theta_1 + \frac{\lambda^2\mu^2}{v^3}\theta_2 + \frac{v - \mu^2}{v} \left(1 + \gamma\theta_1 - \frac{\lambda\gamma}{v}\theta_2 \right) \right].$$

■

Lemma 29 (Asymptotic behavior of T_5) *Under assumptions in theorem 9, we have*

$$\lim_{d\to\infty} \mathbb{E}_{W,X} \left\| R \frac{\sigma(WX^\top)}{n} \right\|_F^2 = \frac{\gamma}{v} \left(\theta_1 - \frac{\lambda}{v}\theta_2 \right), \tag{113}$$

where $\theta_1 := \theta_1(\gamma, \lambda/v), \theta_2 := \theta_2(\gamma, \lambda/v)$.

Proof [Proof of Lemma 29] By definition,

$$\text{L.H.S. of (113)} = \mathbb{E}_{W,X} \operatorname{tr} \left[R^2 \frac{\sigma(WX^\top)\sigma(XW^\top)}{n^2} \right] = \mathbb{E}_{W,X} \operatorname{tr}[R] - \lambda \mathbb{E}_{W,X} \operatorname{tr}\left[R^2\right].$$

Note that $\sigma(XW^\top)$ has i.i.d. zero mean v variance entries, by the Marchenko-Pastur theorem and similar methods in the proof of theorem 3, we get

$$\text{L.H.S. of (113)} = \frac{\gamma}{v}\theta_1 \left(\gamma, \frac{\lambda}{v} \right) - \frac{\lambda\gamma}{v^2}\theta_2 \left(\gamma, \frac{\lambda}{v} \right).$$

■

References

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84, 2020a.

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, pages 11022–11032, 2020b.

Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning. *arXiv preprint arXiv:1912.00827*, 2019.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Number 118. Cambridge University Press, 2010.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *8th International Conference on Learning Representations, ICLR*, 2020.

Zhidong Bai. Convergence rate of expected spectral distributions of large random matrices. part ii. sample covariance matrices. *The Annals of Probability*, 21(2):649–672, 1993.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, 2nd edition, 2010.

David Barber, David Saad, and Peter Sollich. Finite-size effects and optimal test set size in linear perceptrons. *Journal of Physics A: Mathematical and General*, 28(5):1325, 1995.

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

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.

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

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reply to loog et al.: Looking beyond the peaking phenomenon. *Proceedings of the National Academy of Sciences*, 117 (20):10627–10627, 2020a.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020b.

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019.

Siegfried Bös and Manfred Opper. Dynamics of training. In *Advances in Neural Information Processing Systems*, pages 141–147, 1997.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

EP George Box, J Stuart Hunter, William Gordon Hunter, Roma Bins, Kay Kirlin IV, and Destiny Carroll. *Statistics for experimenters: design, innovation, and discovery*. Wiley New York, 2005.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

Lin Chen, Y. Min, M. Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

Romain Couillet and Merouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Romain Couillet, Jakob Hoydis, and Mérouane Debbah. Random beamforming over quasi-static and fading channels: A deterministic equivalent approach. *IEEE Transactions on Information Theory*, 58(10):6392–6425, 2012.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290, 2020.

AD Deev. Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size. In *Sov. Math. Dokl.*, volume 11, pages 1547–1550, 1970.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Michal Derezinski and Manfred KK Warmuth. The limits of squared euclidean distance regularization. In *Advances in Neural Information Processing Systems*, pages 2807–2815, 2014.

Michał Dereziński, Feynman Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arxiv:1810.00412, to appear in the Annals of Statistics*, 2018.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

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

Robert PW Duin. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964, 1995.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, pages 7710–7721, 2020.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks: the hidden manifold model. *arXiv preprint arXiv:1909.11500*, 2019.

Friedrich Götze, Alexander Tikhomirov, et al. Rate of convergence in probability to the marchenko-pastur law. *Bernoulli*, 10(3):503–548, 2004.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Lars Kai Hansen. Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6(3):393–396, 1993.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer series in statistics, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

JA Hertz, A Krogh, and GI Thorbergsson. Phase transitions in simple learning. *Journal of Physics A: Mathematical and General*, 22(12):2133, 1989.

Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2020.

Il′dar Abdulovich Ibragimov and Rafail Zalmanovich Has′ Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640, 2020.

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

Nicole Krämer. On the peaking phenomenon of the lasso in model selection. *arXiv preprint arXiv:0904.4416*, 2009.

Yann Le Cun, Ido Kanter, and Sara A Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters*, 66(18):2396, 1991.

Yann LeCun, Ido Kanter, and Sara A Solla. Second order properties of error surfaces: Learning time and generalization. In *Advances in Neural Information Processing Systems*, pages 918–924, 1991.

Z. Li, Chuanlong Xie, and Qinwen Wang. Provable more data hurt in high dimensional least squares estimator. *arXiv preprint arXiv:2008.06296*, 2020.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel ridgeless regression can generalize. *arXiv preprint arXiv:1808.00387, to appear in The Annals of Statistics*, 2018.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711, 2020.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, 2018.

Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019.

Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *Advances in Neural Information Processing Systems*, pages 13939–13950, 2020.

Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2020.

Marco Loog, Tom Viering, Alexander Mey, Jesse H Krijthe, and David MJ Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20): 10625–10626, 2020.

Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Vladimir A Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Robb J Muirhead. *Aspects of Multivariate Statistical Theory*, volume 197. John Wiley & Sons, 2009.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.

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

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

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

M Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

Manfred Opper. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pages 922–925, 1995.

Manfred Opper. Learning to generalize. *Frontiers of Life*, 3(part 2):763–775, 2001.

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Debashis Paul and Alexander Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

Haozhi Qi, Chong You, Xiaolong Wang, Yi Ma, and Jitendra Malik. Deep isometric learning for visual recognition. In *International Conference on Machine Learning*, 2020.

Šarūnas Raudys. On determining training sample size of linear classifier. *Comput. Systems (in Russian)*, 28:79–87, 1967.

Sarunas Raudys and Robert PW Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern recognition letters*, 19(5-6):385–392, 1998.

Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models. *arXiv preprint arXiv:2010.13933*, 2020.

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

Vadim Ivanovich Serdobolskii. Discriminant analysis for a large number of variables. In *Dokl. Akad. Nauk SSSR*, volume 22, pages 314–319, 1980.

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

Antonio M Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Communications and Information theory*, 1(1):1–182, 2004.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, pages 10112–10123, 2020.

Yuege Xie, Rachel Ward, Holger Rauhut, and Hung-Hsu Chou. Weighted optimization: better generalization by smoother interpolation. *arXiv preprint arXiv:2006.08495*, 2020.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, 2020.

Jianfeng Yao, Zhidong Bai, and Shurong Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, New York, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.