Interpolation Learning

Zitong Yang
Yi Ma
Jacob Steinhardt

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-51
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-51.html

May 12, 2021

Interpolation Learning

by Zitong Yang

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Yi Ma

Professor Yi Ma
Research Advisor

May 10, 2021

(Date)

* * * * * * *

Professor Jacob Steinhardt
Second Reader

05/11/2021

(Date)

Interpolation Learning

by

Zitong Yang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yi Ma, Chair
Professor Jacob Steinhardt

Spring 2021

Abstract

Interpolation Learning

by

Zitong Yang

Master of Science in Computer Science

University of California, Berkeley

Professor Yi Ma, Chair

The classical bias-variance trade-off predicts that bias decreases and variance increases with model complexity, leading to a U-shaped risk curve. Recent work calls this into question for neural networks and other over-parameterized models, for which it is often observed that larger models generalize better. We provide a simple explanation for this by measuring the bias and variance of neural networks: while the bias is *monotonically decreasing* as in the classical theory, the variance is *unimodal* or *bell-shaped*: it increases then decreases with the width of the network. We vary the network architecture, loss function, and choice of dataset and confirm that variance unimodality occurs robustly for all models we considered. The risk curve is the sum of the bias and variance curves and displays different qualitative shapes depending on the relative scale of bias and variance, with the double descent curve observed in recent literature as a special case.

Recent work showed that there could be a large gap between the classical uniform convergence bound and the actual test error of zero-training-error predictors (interpolators) such as deep neural networks. To better understand this gap, we study the uniform convergence in the nonlinear random feature model and perform a precise theoretical analysis on how uniform convergence depends on the sample size and the number of parameters. We derive and prove analytical expressions for three quantities in this model: 1) classical uniform convergence over norm balls, 2) uniform convergence over interpolators in the norm ball, and 3) the risk of minimum norm interpolator. We show that, in the setting where the classical uniform convergence bound is vacuous (diverges to ∞), uniform convergence over the interpolators still gives a non-trivial bound of the test error of interpolating solutions. We also showcase a different setting where classical uniform convergence bound is non-vacuous, but uniform convergence over interpolators can give an improved sample complexity guarantee. Our result provides a first exact comparison between the test errors and uniform convergence bounds for interpolators beyond simple linear models. This thesis is the compilation of the author's two representative work [2] and [1].

Bibliography

[1] Zitong Yang, Yu Bai, and Song Mei. *Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models*. 2021. arXiv: 2103.04554 [cs.LG].

[2] Zitong Yang et al. "Rethinking Bias-Variance Trade-off for Generalization of Neural Networks". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 10767–10777. URL: http://proceedings.mlr.press/v119/yang20j.html.

Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models

1. Introduction

Uniform convergence—the supremum difference between the training and test errors over a certain function class—is a powerful tool in statistical learning theory for understanding the generalization performance of predictors. Bounds on uniform convergence usually take the form of $\sqrt{\text{complexity}/n}$ (Vapnik, 1995), where the numerator represents the complexity of the function class, and n is the sample size. If such a bound is tight, then the predictor is not going to generalize well whenever the function class complexity is too large.

However, it is shown in recent theoretical and empirical work that overparametized models such as deep neural networks could generalize well, even in the interpolating regime in which the model exactly memorizes the data (Zhang et al., 2016; Belkin et al., 2019a). As interpolation (especially for noisy training data) usually requires the predictor to be within a function class with high complexity, this challenges the classical methodology of using uniform convergence to bound generalization. For example, Belkin et al. (2018c) showed that interpolating noisy data with kernel machines requires exponentially large norm in fixed dimensions. The large norm would effectively make the uniform convergence bound $\sqrt{\text{complexity}/n}$ vacuous. Nagarajan & Kolter (2019a) empirically measured the spectral-norm bound in Bartlett et al. (2017) and find that for interpolators, the bound increases with n, and is thus vacuous at large sample size. Towards a more fine-grained understanding, we ask the following

> **Question**: How large is the gap between uniform convergence and the actual generalization errors for interpolators?

In this paper, we study this gap in the random features model from Rahimi & Recht (2007). This model can be interpreted as a linearized version of two-layer neural networks (Jacot et al., 2018) and exhibit some similar properties to deep neural networks such as double descent (Belkin et al., 2019a). We consider two types of uniform convergence in this model:

- $\mathcal{U}$: The classical uniform convergence over a norm ball of radius $\sqrt{A}$.
- $\mathcal{T}$: The modified uniform convergence over the same norm ball of size $\sqrt{A}$ but only include the interpolators, proposed in Zhou et al. (2020).

Our main theoretical result is the exact asymptotic expressions of two versions of uniform convergence $\mathcal{U}$ and $\mathcal{T}$ in terms of the number of features, sample size, as well as other relevant parameters in the random feature model. Under some assumptions, we prove that the actual uniform convergence concentrates to these asymptotic counterparts. To further compare these uniform convergence bounds with the actual generalization error of interpolators, we adopt

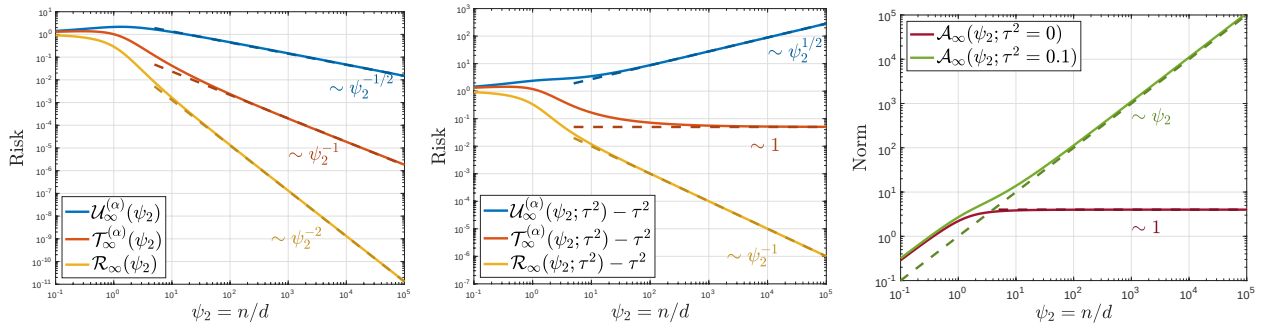
- $\mathcal{R}$: the generalization error (test error) of the minimum norm interpolator.

from Mei & Montanari (2019). To make $\mathcal{U}$, $\mathcal{T}$, $\mathcal{R}$ comparable with each other, we choose the radius of the norm ball $\sqrt{A}$ to be slightly larger than the norm of the minimum norm interpolator. Our limiting $\mathcal{U}$, $\mathcal{T}$ (with norm ball of size $\sqrt{A}$ as chosen above), and $\mathcal{R}$ depend on two main variables: $\psi_1 = \lim_{d\to\infty} N/d$ representing the number of parameters, and $\psi_2 = \lim_{d\to\infty} n/d$ representing the sample size. Our formulae for $\mathcal{U}, \mathcal{T}$ and $\mathcal{R}$ yield three major observations.

1. **Sample Complexity in the Noisy Regime:** When the training data contains label noise (with variance τ^2), we find that the norm required to interpolate the noisy training set grows linearly with the number of samples ψ_2 (green curve in Figure 1(c)). As a result, the standard uniform convergence bound $\mathcal{U}$ grows with ψ_2 at the rate $\mathcal{U} \sim \psi_2^{1/2}$, leading to a vacuous bound on the generalization error (Figure 1(b)).

   In contrast, in the same setting, we show the uniform convergence over interpolators $\mathcal{T} \sim 1$ is a constant for large ψ_2, and is only order one larger than the actual generalization error $\mathcal{R} \sim 1$. Further, the excess versions scale as $\mathcal{T} - \tau^2 \sim 1$ and $\mathcal{R} - \tau^2 \sim \psi_2^{-1}$.

2. **Sample Complexity in the Noiseless Regime:** When the training set does not contain label noise, the generalization error $\mathcal{R}$ decays faster: $\mathcal{R} \sim \psi_2^{-2}$. In this setting, we find that the classical uniform convergence $\mathcal{U} \sim \psi_2^{-1/2}$ and the uniform convergence over interpolators $\mathcal{T} \sim \psi_2^{-1}$. This shows that, even when the



(a) Noiseless response ($\tau^2 = 0$) (b) Noisy response ($\tau^2 = 0.1$) (c) Minimum norm $\mathcal{A}_\infty(\psi_2)$

Figure 1. Random feature regression with activation function $\sigma(x) = \max(0, x) - 1/\sqrt{2\pi}$, target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}\|_2^2 = 1$, and $\psi_1 = \infty$. The horizontal axes are the number of samples $\psi_2 = \lim_{d\to\infty} n/d$. The solid lines are the the algebraic expressions derived in the main theorem (Theorem 1). The dashed lines are the function ψ_2^p in the log scale. Figure 1(a) and 1(b): Comparison of the classical uniform convergence in the norm ball of size level $\alpha = 1.5$ (Eq. (17), blue curve), the uniform convergence over interpolators in the same norm ball (Eq. (18), red curve), the risk of minimum norm interpolator (Eq. (13), yellow curve). Figure 1(c): Minimum norm required to interpolate the training data (Eq. (12)).

classical uniform convergence already gives a non-vacuous bound, there still exists a sample complexity separation among the classical uniform convergence $\mathcal{U}$, the uniform convergence over interpolators $\mathcal{T}$, and the actual generalization error $\mathcal{R}$.

3. **Dependence on Number of Parameters:** In addition to the results on ψ_2, we find that $\mathcal{U}, \mathcal{T}$ and $\mathcal{R}$ decay to its limiting value at the same rate $1/\psi_1$. This shows that both $\mathcal{U}$ and $\mathcal{T}$ correctly predict that as the number of features ψ_1 grows, the risk $\mathcal{R}$ would decrease.

These results provide a more precise understanding of uniform convergence versus the actual generalization errors, under a natural model that captures a lot of essences of nonlinear overparametrized learning.

1.1. Related work

Classical theory of uniform convergence. Uniform convergence dates back to the empirical process theory of Glivenko (1933) and Cantelli (1933). Application of uniform convergence to the framework of empirical risk minimization usually proceeds through Gaussian and Rademacher complexities (Bartlett & Mendelson, 2003; Bartlett et al., 2005) or VC and fat shattering dimensions (Vapnik, 1995; Bartlett, 1998).

Modern take on uniform convergence. A large volume of recent works showed that overparametrized interpolators could generalize well (Zhang et al., 2016; Belkin et al., 2018b; Neyshabur et al., 2015a; Advani et al., 2020; Bartlett et al., 2020; Belkin et al., 2018a; 2019b; Nakkiran et al., 2020; Yang et al., 2020; Belkin et al., 2019a; Mei & Montanari, 2019; Spigler et al., 2019), suggesting that the classical uniform convergence theory may not be able to explain generalization in these settings (Zhang et al., 2016). Numerous efforts have been made to remedy the original uniform convergence theory using the Rademacher complexity (Neyshabur et al., 2015b; Golowich et al., 2018; Neyshabur et al., 2019; Zhu et al., 2009; Cao & Gu, 2019), the compression approach (Arora et al., 2018), covering numbers (Bartlett et al., 2017), derandomization (Negrea et al., 2020) and PAC-Bayes methods (Dziugaite & Roy, 2017; Neyshabur et al., 2018; Nagarajan & Kolter, 2019b). Despite the progress along this line, Nagarajan & Kolter (2019a); Bartlett & Long (2020) showed that in certain settings "any uniform convergence" bounds cannot explain generalization. Among the pessimistic results, Zhou et al. (2020) proposes that uniform convergence over interpolating norm ball could explain generalization in an overparametrized linear setting. Our results show that in the nonlinear random feature model, there is a sample complexity gap between the excess risk and uniform convergence over interpolators proposed in Zhou et al. (2020).

Random features model and kernel machines. A number of papers studied the generalization error of kernel machines (Caponnetto & De Vito, 2007; Jacot et al., 2020b; Wainwright, 2019) and random features models (Rahimi & Recht, 2009; Rudi & Rosasco, 2017; Bach, 2015; Ma et al., 2020) in the non-asymptotic settings, in which the generalization error bound depends on the RKHS norm. However, these bounds cannot characterize the generalization error for interpolating solutions. In the last three years, a few papers (Belkin et al., 2018c; Liang et al., 2020; 2019) showed that interpolating solutions of kernel ridge regression can also generalize well in high dimensions. Recently, a few papers studied the generalization error of random features model in the proportional asymptotic limit in various settings (Hastie et al., 2019; Louart et al., 2018; Mei & Mon-

tanari, 2019; Montanari et al., 2019; Gerace et al., 2020; d'Ascoli et al., 2020; Yang et al., 2020; Adlam & Pennington, 2020; Dhifallah & Lu, 2020; Hu & Lu, 2020), where they precisely characterized the asymptotic generalization error of interpolating solutions, and showed that double-descent phenomenon (Belkin et al., 2019a; Advani et al., 2020) exists in these models. A few other papers studied the generalization error of random features models in the polynomial scaling limits (Ghorbani et al., 2019; 2020; Mei et al., 2021), where other interesting behaviors were shown.

Precise asymptotics for the Rademacher complexity of some *underparameterized* learning models was calculated using statistical physics heuristics in Abbaras et al. (2020). In our work, we instead focus on the uniform convergence of *overparameterized* random features model.

2. Problem formulation

In this section, we present the background needed to understand the insights from our main result. In Section 2.1 we define the random feature regression task that this paper focuses on. In Section 2.2, we informally present the limiting regime our theory covers.

2.1. Model setup

Consider a dataset $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ with n samples. Assume that the covariates follow $\boldsymbol{x}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and responses satisfy $y_i = f_d(\boldsymbol{x}_i)+\varepsilon_i$, with the noises satisfying $\varepsilon_i \sim_{iid} \mathcal{N}(0,\tau^2)$ which are independent of $(\boldsymbol{x}_i)_{i\in[n]}$. We will consider both the noisy ($\tau^2 > 0$) and noiseless ($\tau^2 = 0$) settings.

We fit the dataset using the random features model. Let $(\boldsymbol{\theta}_j)_{j\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ be the random feature vectors. Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define the random features function class $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ by

$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta}) \equiv \Big\{ f(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \sigma\big(\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle/\sqrt{d}\big) : \boldsymbol{a} \in \mathbb{R}^N \Big\}.$$

Generalization error of the minimum norm interpolator. Denote the population risk and the empirical risk of a predictor $\boldsymbol{a} \in \mathbb{R}^N$ by

$$R(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x},y}\Big(y - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle/\sqrt{d})\Big)^2, \tag{1}$$

$$\widehat{R}_n(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d})\Big)^2, \tag{2}$$

and the regularized empirical risk minimizer with vanishing regularization by

$$\boldsymbol{a}_{\min} = \lim_{\lambda\to 0+} \arg\min_{\boldsymbol{a}} \Big[\widehat{R}_n(\boldsymbol{a}) + \lambda\|\boldsymbol{a}\|_2^2\Big].$$

In the overparameterized regime ($N > n$), under mild conditions, we have $\min_{\boldsymbol{a}} \widehat{R}_n(\boldsymbol{a}) = \widehat{R}_n(\boldsymbol{a}_{\min}) = 0$. In this regime, $\boldsymbol{a}_{\min}$ can be interpreted as the minimum ℓ_2 norm interpolator.

A quantity of interest is the generalization error of this predictor, which gives (with a slight abuse of notation)

$$R(N, n, d) \equiv R(\boldsymbol{a}_{\min}). \tag{3}$$

Uniform convergence bounds. We denote the uniform convergence bound over a norm ball and the uniform convergence over interpolators in the norm ball by

$$U(A, N, n, d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a})\Big), \tag{4}$$

$$T(A, N, n, d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A,\, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}). \tag{5}$$

Here the scaling factor N/d of the norm ball is such that the norm ball converges to a non-trivial RKHS norm ball with size $\sqrt{A}$ as $\psi_1 \to \infty$ (limit taken after $N/d \to \psi_1$). Note that in order for the maximization problem in (5) to have a non-empty feasible region, we need $\widehat{R}_n(\boldsymbol{a}_{\min}) = 0$ and need to take $A \ge (N/d)\|\boldsymbol{a}_{\min}\|_2^2$: we will show that in the region $N > n$ with sufficiently large A, this happens with high probability.

By construction, for any $A \ge (N/d)\|\boldsymbol{a}_{\min}\|_2^2$, we have $U(A) \ge T(A) \ge R(\boldsymbol{a}_{\min})$ (see Figure 2). So a natural problem is to quantify the gap among $U(A)$, $T(A)$, and $R(\boldsymbol{a}_{\min})$, which is our goal in this paper.

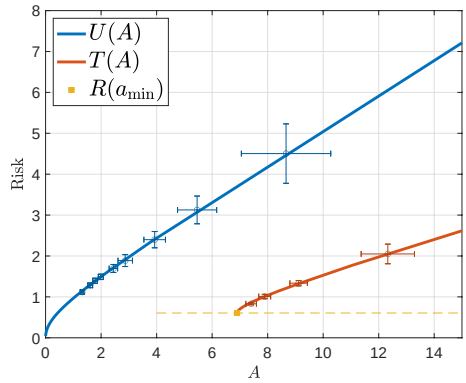


Figure 2. Illustration of uniform convergence U (c.f. eq. (4)), uniform convergence over interpolators T (c.f. eq. (5)), and minimum norm interpolator $R(\boldsymbol{a}_{\min})$. We take $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle$ for some $\|\boldsymbol{\beta}\|_2^2 = 1$, and take the ReLU activation function $\sigma(x) = \max\{x, 0\}$. Solid lines are our theoretical predictions $\mathcal{U}$ and $\mathcal{T}$ (cf. (6) & (7)). Points with error bars are obtained from simulations with the number of features $N = 500$, number of samples $n = 300$, and covariate dimension $d = 200$. The error bar reports $1/\sqrt{20}\times$standard deviation over 20 instances. See Appendix B for details.

2.2. High dimensional regime

We approach this problem in the limit $d \to \infty$ with $N/d \to \psi_1$ and $n/d \to \psi_2$ (c.f. Assumption 3). We further assume the setting of a linear target function f_d and a non-linear activation function σ (c.f. Assumptions 1 and 2). In this regime, our main result Theorem 1 will show that, the uniform convergence U and the uniform convergence over interpolators T will converge to deterministic functions, i.e., writing here informally,

$$U(A, N, n, d) \overset{d\to\infty}{\to} \mathcal{U}(A, \psi_1, \psi_2), \tag{6}$$

$$T(A, N, n, d) \overset{d\to\infty}{\to} \mathcal{T}(A, \psi_1, \psi_2), \tag{7}$$

where $\mathcal{U}$ and $\mathcal{T}$ will be defined in Definition 2 (which depends on the definition of some other quantities that are defined in Appendix A and heuristically presented in Remark 1). In addition to $\mathcal{U}$ and $\mathcal{T}$, Theorem 1 of Mei & Montanari (2019) implies the following convergence

$$(N/d)\|\boldsymbol{a}_{\min}\|_2^2 \overset{d\to\infty}{\to} \mathcal{A}(\psi_1, \psi_2), \tag{8}$$

$$R(\boldsymbol{a}_{\min}) \overset{d\to\infty}{\to} \mathcal{R}(\psi_1, \psi_2). \tag{9}$$

The precise algebraic expression of equation (8) and (9) was given in Definition 1 of Mei & Montanari (2019), and we include in Appendix A for completeness. We will sometimes refer to $\mathcal{U}, \mathcal{T}, \mathcal{A}, \mathcal{R}$ without explicitly mark their dependence on A, ψ_1, ψ_2 for notational simplicity.

Kernel regime. Rahimi & Recht (2007) have shown that, as $N \to \infty$, the random feature space $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ (equipped with proper inner product) converges to the RKHS (Reproducing Kernel Hilbert Space) induced by the kernel

$$H(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)\sigma(\langle \boldsymbol{x}', \boldsymbol{w}\rangle)].$$

We expect that, if we take limit $\psi_1 \to \infty$ after $N, d, n \to \infty$, the formula of $\mathcal{U}$ and $\mathcal{T}$ will coincide with the corresponding asymptotic limit of U and T for kernel ridge regression with the kernel H. This intuition has been mentioned in a few papers (Mei & Montanari, 2019; d'Ascoli et al., 2020; Jacot et al., 2020a). In this spirit, we denote

$$\mathcal{U}_\infty(A, \psi_2) \equiv \lim_{\psi_1\to\infty} \mathcal{U}(A, \psi_1, \psi_2), \tag{10}$$

$$\mathcal{T}_\infty(A, \psi_2) \equiv \lim_{\psi_1\to\infty} \mathcal{T}(A, \psi_1, \psi_2), \tag{11}$$

$$\mathcal{A}_\infty(\psi_2) \equiv \lim_{\psi_1\to\infty} \mathcal{A}(\psi_1, \psi_2), \tag{12}$$

$$\mathcal{R}_\infty(\psi_2) \equiv \lim_{\psi_1\to\infty} \mathcal{R}(\psi_1, \psi_2). \tag{13}$$

We will refer to the quantities $\{\mathcal{U}_\infty, \mathcal{T}_\infty, \mathcal{A}_\infty, \mathcal{R}_\infty\}$ as the {uniform convergence in norm ball, uniform convergence over interpolators in norm ball, minimum ℓ_2 norm of interpolators, and generalization error of interpolators} of kernel ridge regression.

Low norm uniform convergence bounds. There is a question of which norm A to choose in $\mathcal{U}$ and $\mathcal{T}$ to compare with $\mathcal{R}$. In order for U and T to serve as proper bounds for $R(\boldsymbol{a}_{\min})$, we need to take at least $A \geq \psi_1\|\boldsymbol{a}_{\min}\|_2^2$. Therefore, we will choose

$$A = \alpha\psi_1\|\boldsymbol{a}_{\min}\|_2^2, \tag{14}$$

for some $\alpha > 1$ (e.g., $\alpha = 1.1$). Note $\psi_1\|\boldsymbol{a}_{\min}\|_2^2 \to \mathcal{A}(\psi_1, \psi_2)$ as $d \to \infty$. So for a fixed $\alpha > 1$, we further define

$$\mathcal{U}^{(\alpha)}(\psi_1, \psi_2) \equiv \mathcal{U}(\alpha\mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2), \tag{15}$$

$$\mathcal{T}^{(\alpha)}(\psi_1, \psi_2) \equiv \mathcal{T}(\alpha\mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2), \tag{16}$$

and their kernel version,

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2) \equiv \lim_{\psi_1\to\infty} \mathcal{U}^{(\alpha)}(\psi_1, \psi_2), \tag{17}$$

$$\mathcal{T}_\infty^{(\alpha)}(\psi_2) \equiv \lim_{\psi_1\to\infty} \mathcal{T}^{(\alpha)}(\psi_1, \psi_2). \tag{18}$$

This definition ensures that $\mathcal{R}(\psi_1, \psi_2) \leq \mathcal{T}^{(\alpha)}(\psi_1, \psi_2) \leq \mathcal{U}^{(\alpha)}(\psi_1, \psi_2)$ and $\mathcal{R}_\infty(\psi_2) \leq \mathcal{T}_\infty^{(\alpha)}(\psi_2) \leq \mathcal{U}_\infty^{(\alpha)}(\psi_2)$.

3. Asymptotic power laws and separations

In this section, we evaluate the algebraic expressions derived in our main result (Theorem 1) as well as the quantities $\mathcal{U}^{(\alpha)}$, $\mathcal{T}^{(\alpha)}$, $\mathcal{A}$, and $\mathcal{R}$, before formally presenting the theorem. We examine their dependence with respect to the noise level τ^2, the number of features $\psi_1 = \lim_{d\to\infty} N/d$, and the sample size $\psi_2 = \lim_{d\to\infty} n/d$, and we further infer their asymptotic power laws for large ψ_1 and ψ_2.

3.1. Norm of the minimum norm interpolator

Since we are considering uniform convergence bounds over the norm ball of size α times $\mathcal{A}_\infty(\psi_2)$ (the norm of the min-norm interpolator), let's first examine how $\mathcal{A}_\infty(\psi_2)$ scale with ψ_2. As we shall see, $\mathcal{A}_\infty(\psi_2)$ behaves differently in the noiseless ($\tau^2 = 0$) and noisy ($\tau^2 > 0$) settings, so here we explicitly mark the dependence on τ^2, i.e. $\mathcal{A}_\infty(\psi_2; \tau^2)$.

The inferred asymptotic power law gives (c.f. Figure 1(c))

$$\mathcal{A}_\infty(\psi_2; \tau^2 > 0) \sim \psi_2,$$
$$\mathcal{A}_\infty(\psi_2; \tau^2 = 0) \sim 1,$$

where $X_1(\psi) \sim X_2(\psi)$ for large ψ means that

$$\lim_{\psi\to\infty} \log(X_1(\psi))/\log(X_2(\psi)) = 1.$$

In words, when there is no label noise ($\tau^2 = 0$), we can interpolate infinite data even with a finite norm. When the responses are noisy ($\tau^2 > 0$), interpolation requires a large norm that is proportional to the number of samples.

On a high level, our statement echoes the finding of Belkin et al. (2018c), where they study a binary classification problem using the kernel machine, and prove that an interpolating classifier requires RKHS norm to grow at least exponentially with $n^{1/d}$ for fixed dimension d. Here instead we consider the high dimensional setting and we show a linear grow in $\psi_2 = \lim_{d\to\infty} n/d$.

3.2. Kernel regime with noiseless data

We first look at the noiseless setting ($\tau^2 = 0$) and present the asymptotic power law for the uniform convergence $\mathcal{U}_\infty^{(\alpha)}$ over the low-norm ball, the uniform convergence over interpolators $\mathcal{T}_\infty^{(\alpha)}$ in the low norm ball, and the minimum norm risk $\mathcal{R}_\infty$ from (17) (18) (13), respectively.

In this setting, the inferred asymptotic power law of $\mathcal{U}_\infty^{(\alpha)}(\psi_2)$, $\mathcal{T}_\infty^{(\alpha)}(\psi_2)$, and $\mathcal{R}_\infty(\psi_2)$ gives (c.f. Figure 1(a))

$$\begin{aligned}\mathcal{U}_\infty^{(\alpha)}(\psi_2;\tau^2=0) &\sim \psi_2^{-1/2},\\ \mathcal{T}_\infty^{(\alpha)}(\psi_2;\tau^2=0) &\sim \psi_2^{-1},\\ \mathcal{R}_\infty^{(\alpha)}(\psi_2;\tau^2=0) &\sim \psi_2^{-2}.\end{aligned}$$

As we can see, all the three quantities converge to 0 in the large sample limit, which indicates that uniform convergence is able to explain generalization in this setting. yet uniform convergence bounds do not correctly capture the convergence rate (in terms of ψ_2) of the generalization error.

3.3. Kernel regime with noisy data

In the noisy setting (fix $\tau^2 > 0$), the Bayes risk (minimal possible risk) is τ^2. We study the excess risk and the excess version of uniform convergence bounds by subtracting the Bayes risk τ^2. The inferred asymptotic power law gives (c.f. Figure 1(b))

$$\begin{aligned}\mathcal{U}_\infty^{(\alpha)}(\psi_2;\tau^2) - \tau^2 &\sim \psi_2^{1/2},\\ \mathcal{T}_\infty^{(\alpha)}(\psi_2;\tau^2) - \tau^2 &\sim 1,\\ \mathcal{R}_\infty(\psi_2;\tau^2) - \tau^2 &\sim \psi_2^{-1}.\end{aligned}$$

In the presence of label noise, the excess risk $\mathcal{R}_\infty - \tau^2$ vanishes in the large sample limit. In contrast, the classical uniform convergence $\mathcal{U}_\infty$ becomes vacuous, whereas the uniform convergence over interpolators $\mathcal{T}_\infty$ converges to a constant, which gives a non-vacuous bound of $\mathcal{R}_\infty$.

The decay of the excess risk of minimum norm interpolators even in the presence of label noise is no longer a surprising phenomenon in high dimensions (Liang et al., 2019; Ghorbani et al., 2019; Bartlett et al., 2020). A simple explanation of this phenomenon is that the nonlinear part of the activation function σ has an implicit regularization effect (Mei & Montanari, 2019).

The divergence of $\mathcal{U}_\infty^{(\alpha)}$ in the presence of response noise is partly due to that $\mathcal{A}_\infty(\psi_2)$ blows up linearly in ψ_2 (c.f. Section 3.1). In fact, we can develop a heuristic intuition that $\mathcal{U}_\infty(A,\psi_2;\tau^2) \sim A/\psi_2^{1/2}$. Then the scaling $\mathcal{U}_\infty^{(\alpha)}(\psi_2;\tau^2>0) \sim \mathcal{A}_\infty(\psi_2;\tau^2>0)/\psi_2^{1/2} \sim \psi_2^{1/2}$ can be explained away by the power law of $\mathcal{A}_\infty(\psi_2;\tau^2>0) \sim \psi_2$. In other words, the complexity of the function space of interpolators grows faster than the sample size n, which leads to the failure of uniform convergence in explaining generalization. This echoes the findings in Nagarajan & Kolter (2019a).

To illustrate the scaling $\mathcal{U}_\infty(A,\psi_2) \sim A/\psi_2^{1/2}$. We fix all other parameters $(\mu_1,\mu_\star,\tau,F_1)$, and examine the dependence of $\mathcal{U}_\infty$ on A and ψ_2. We choose $A = A(\psi_2)$ according to different power laws $A(\psi_2) \sim \psi_2^p$ for $p = 0, 0.25, 0.5, 0.75, 1$. The inferred asymptotic power law gives $\mathcal{U}_\infty(A(\psi_2),\psi_2) \sim \psi_2^{p-0.5}$ (c.f. Figure 3). This provides an evidence for the relation $\mathcal{U}_\infty(A,\psi_2) \sim A/\psi_2^{1/2}$.

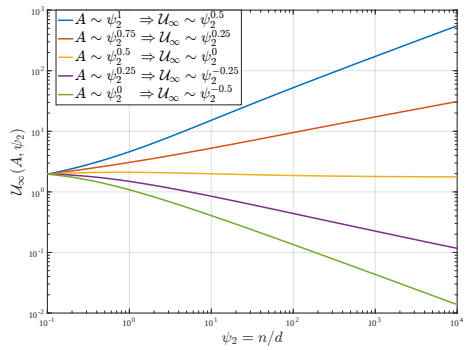


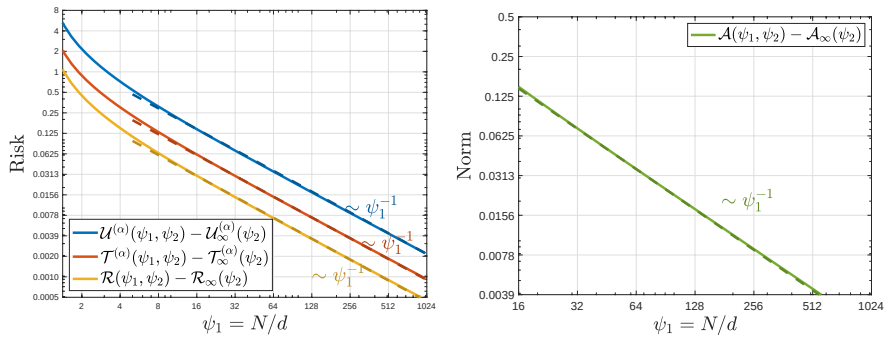
Figure 3. Uniform convergence $\mathcal{U}_\infty(A(\psi_2),\psi_2)$ over the norm ball in the kernel regime $\psi_1 \to \infty$. The size of the norm ball $A = A(\psi_2)$ is chosen according to different power laws as shown in the legend.

3.4. Finite-width regime

Here we shift attention to the dependence of $\mathcal{U}$, $\mathcal{T}$, and $\mathcal{R}$ on the number of features ψ_1. We fix the number of training samples ψ_2, noise level $\tau^2 > 0$, and norm level $\alpha > 1$ similar as before. Since $\mathcal{U}^\alpha \to \mathcal{U}_\infty^\alpha, \mathcal{T}^\alpha \to \mathcal{T}_\infty^\alpha$ and $\mathcal{R} \to \mathcal{R}_\infty$ as $\psi_1 \to \infty$, we look at the dependence of $\mathcal{U}^\alpha - \mathcal{U}_\infty^\alpha, \mathcal{T}^\alpha - \mathcal{T}_\infty^\alpha$ and $\mathcal{R}^\alpha - \mathcal{R}_\infty^\alpha$ with respect to ψ_1. The inferred asymptotic law gives (c.f. Figure 4)

$$\begin{aligned}\mathcal{U}^{(\alpha)}(\psi_1,\psi_2) - \mathcal{U}_\infty^{(\alpha)}(\psi_2) &\sim \psi_1^{-1},\\ \mathcal{T}^{(\alpha)}(\psi_1,\psi_2) - \mathcal{T}_\infty^{(\alpha)}(\psi_2) &\sim \psi_1^{-1},\\ \mathcal{R}(\psi_1,\psi_2) - \mathcal{R}_\infty(\psi_2) &\sim \psi_1^{-1},\\ \mathcal{A}(\psi_1,\psi_2) - \mathcal{A}_\infty(\psi_2) &\sim \psi_1^{-1}.\end{aligned}$$

Note that large ψ_1 should be interpreted as the model being heavily overparametrized (a large width network). This



(a) (b)

Figure 4. Random feature regression with the number of sample $\psi_2 = 1.5$, activation function $\sigma(x) = \max(0, x) - 1/\sqrt{2\pi}$, target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}\|_2^2 = 1$, and noise level $\tau^2 = 0.1$. The horizontal axes are the number of features ψ_1. The solid lines are the the algebraic expressions derived in the main theorem (Theorem 1). The dashed lines are the function ψ_1^p in the log scale. Figure 4(a): Comparison of the classical uniform convergence in the norm ball of size level $\alpha = 1.5$ (Eq. (15), blue curve), the uniform convergence over interpolators in the same norm ball (Eq. (16), red curve), the risk of minimum norm interpolator (Eq. (9), yellow curve). Figure 4(b): Minimum norm required to interpolate the training data (Eq. (8)).

asymptotic power law implies that, both uniform convergence bounds correctly predict the decay of the test error with the increase of the number of features.

Remark on power laws. For the derivation of the power laws in this section, instead of working with the analytical formula, we adopt an empirical approach: we perform linear fits with the inferred slopes, upon the numerical evaluations (of these expressions defined in Definition 2) in the log-log scale. However, these linear fits are for the analytical formulae and do not involve randomness, and thus reliably indicate the true decay rates.

4. Main theorem

In this section, we state the main theorem that presents the asymptotic expressions for the uniform convergence bounds. We will start by stating a few assumptions, which fall into two categories: Assumption 1, 2, and 3, which specify the setup for the learning task; Assumption 4 and 5, which are technical in nature.

4.1. Modeling assumptions

The three assumptions in this subsection specify the target function, the activation function, and the limiting regime.

Assumption 1 (Linear target function)**.** *We assume that* $f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ *with* $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}^{(d)}, \boldsymbol{x} \rangle$*, where* $\boldsymbol{\beta}^{(d)} \in \mathbb{R}^d$ *and*

$$\lim_{d\to\infty} \|\boldsymbol{\beta}^{(d)}\|_2^2 = F_1^2.$$

We remark here that, if we are satisfied with heuristic formulae instead of rigorous results, we are able to deal with non-linear target functions, where the additional nonlinear part is effectively increasing the noise level τ^2. This intuition was first developed in (Mei & Montanari, 2019).

Assumption 2 (Activation function)**.** *Let* $\sigma \in C^2(\mathbb{R})$ *with* $|\sigma(u)|, |\sigma'(u)|, |\sigma''(u)| \le c_0 e^{c_1|u|}$ *for some constant* $c_0, c_1 < \infty$*. Define*

$$\mu_0 \equiv \mathbb{E}[\sigma(G)],\ \mu_1 \equiv \mathbb{E}[G\sigma(G)],\ \mu_\star^2 \equiv \mathbb{E}[\sigma(G)^2] - \mu_0^2 - \mu_1^2,$$

where expectation is with respect to $G \sim \mathcal{N}(0,1)$*. Assume* $\mu_0 = 0$*,* $0 < \mu_1^2, \mu_\star^2 < \infty$*.*

The assumption that $\mu_0 = 0$ is not essential and can be relaxed with a certain amount of additional technical work.

Assumption 3 (Proportional limit)**.** *Let* $N = N(d)$ *and* $n = n(d)$ *be sequences indexed by* d*. We assume that the following limits exist in* $(0, \infty)$*:*

$$\lim_{d\to\infty} N(d)/d = \psi_1, \qquad \lim_{d\to\infty} n(d)/d = \psi_2.$$

4.2. Technical assumptions

We will make some assumptions upon the properties of some random matrices that appear in the proof. These assumptions are technical and we believe they can be proved under more natural assumptions. However, proving them requires substantial technical work, and we defer them to future work. We note here that these assumptions are often implicitly required in papers that present intuitions using heuristic derivations. Instead, we ensure the mathematical rigor by listing them. See Section 5 for more discussions upon these assumptions.

We begin by defining some random matrices which are the key quantities that are used in the proof of our main results.

Definition 1 (Block matrix and log-determinant)**.** *Let* $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\mathsf{T} \in \mathbb{R}^{n\times d}$ *and* $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^\mathsf{T} \in \mathbb{R}^{N\times d}$*,*

where $\boldsymbol{x}_i, \boldsymbol{\theta}_a \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$*, as mentioned in Section 2.1. Define*

$$\boldsymbol{Z} = \frac{1}{\sqrt{d}}\sigma\left(\frac{\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}}{\sqrt{d}}\right), \quad \boldsymbol{Z}_1 = \frac{\mu_1}{d}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T},$$
$$\boldsymbol{Q} = \frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}}{d}, \qquad \boldsymbol{H} = \frac{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}}{d}, \tag{19}$$

and for $\boldsymbol{q} = (s_1, s_2, t_1, t_2, q) \in \mathbb{R}^5$*, we define*

$$\boldsymbol{A}(\boldsymbol{q}) \equiv \begin{bmatrix} s_1\mathbf{I}_N + s_2\boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T} + p\boldsymbol{Z}_1^\mathsf{T} \\ \boldsymbol{Z} + p\boldsymbol{Z}_1 & t_1\mathbf{I}_n + t_2\boldsymbol{H} \end{bmatrix}.$$

Finally, we define the log-deteminant of $\boldsymbol{A}(\boldsymbol{q})$ *by*

$$G_d(\xi; \boldsymbol{q}) \equiv \frac{1}{d}\sum_{i=1}^{N+n} \mathrm{Log}\lambda_i\Big(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_{n+N}\Big).$$

Here Log *is the complex logarithm with branch cut on the negative real axis and* $\{\lambda_i(\boldsymbol{A})\}_{i\in[n+N]}$ *is the set of eigenvalues of* $\boldsymbol{A}$*.*

The following assumption states that for properly chosen λ, some specific random matrices are well-conditioned. As we will see in the next section, this ensures that the dual problems in Eq. (20) and (21) are bounded with high probability.

Assumption 4 (Invertability)**.** *Consider the asymptotic limit as specified in Assumption 3 the activation function as in Assumption 2. We assume the following.*

- *Denote* $\overline{\boldsymbol{U}}(\lambda) = \mu_1^2\boldsymbol{Q} + (\mu_\star^2 - \psi_1\lambda)\mathbf{I}_N - \psi_2^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$*. There exists* $\varepsilon > 0$ *and* $\underline{\lambda}_U = \underline{\lambda}_U(\psi_1, \psi_2, \mu_1^2, \mu_\star^2)$*, such that for any fixed* $\lambda \in (\underline{\lambda}_U, \infty) \equiv \Lambda_U$*, with high probability, we have*
$$\overline{\boldsymbol{U}}(\lambda) \preceq -\varepsilon\mathbf{I}_N.$$
- *Denote* $\overline{\boldsymbol{T}}(\lambda) = \mathsf{P}_{\mathrm{null}}[\mu_1^2\boldsymbol{Q} + (\mu_\star^2 - \psi_1\lambda)\mathbf{I}_N]\mathsf{P}_{\mathrm{null}}$ *where* $\mathsf{P}_{\mathrm{null}} = \mathbf{I}_N - \boldsymbol{Z}^\dagger\boldsymbol{Z}$*. There exists* $\varepsilon > 0$ *and* $\underline{\lambda}_T = \underline{\lambda}_T(\psi_1, \psi_2, \mu_1^2, \mu_\star^2)$*, such that for any fixed* $\lambda \in (\underline{\lambda}_T, \infty) \equiv \Lambda_T$*, with high probability we have*
$$\overline{\boldsymbol{T}}(\lambda) \preceq -\varepsilon\mathsf{P}_{\mathrm{null}},$$
and $\boldsymbol{Z}$ *has full row rank with* $\sigma_{\min}(\boldsymbol{Z}) \geq \varepsilon$ *(which requires* $\psi_1 > \psi_2$*).*

The following assumption states that the order of limits and derivatives regarding G_d can be exchanged.

Assumption 5 (Exchangeability of limits)**.** *We denote*

$$\mathcal{S}_U = \{(\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0; \psi_1, \psi_2) : \lambda \in (\underline{\lambda}_U, \infty)\},$$
$$\mathcal{S}_T = \{(\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0; \psi_1, \psi_2) : \lambda \in (\underline{\lambda}_T, \infty)\},$$

where $\underline{\lambda}_U$ *and* $\underline{\lambda}_T$ *are given in Assumption 4 and depend on* $(\psi_1, \psi_2, \mu_1^2, \mu_\star^2)$*. For any fixed* $(\boldsymbol{q}; \boldsymbol{\psi}) = (s_1, s_2, t_1, t_2, p; \psi_1, \psi_2) \in \mathcal{S}_U \cup \mathcal{S}_T$*, in the asymptotic limit as in Assumption 3, for* $k = 1, 2$*, we have*

$$\lim_{u\to 0_+}\lim_{d\to\infty}\mathbb{E}[\nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q})] = \lim_{u\to 0_+}\nabla_{\boldsymbol{q}}^k\Big(\lim_{d\to\infty}\mathbb{E}[G_d(\boldsymbol{i}u; \boldsymbol{q})]\Big),$$

and

$$\Big\|\nabla_{\boldsymbol{q}}^k G_d(0; \boldsymbol{q}) - \lim_{u\to 0+}\lim_{d\to\infty}\mathbb{E}[\nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q})]\Big\| = o_{d,\mathbb{P}}(1),$$

where $o_{d,\mathbb{P}}(1)$ *stands for convergence to* 0 *in probability.*

4.3. From constrained forms to Lagrangian forms

Before we give the asymptotics of U and T as defined in Eq. (4) and (5), we first consider their dual forms which are more amenable in analysis. These are given by

$$\overline{U}(\lambda, N, n, d) \equiv \sup_{\boldsymbol{a}}\Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda\|\boldsymbol{a}\|_2^2\Big], \tag{20}$$

$$\overline{T}(\lambda, N, n, d) \equiv \sup_{\boldsymbol{a}}\inf_{\boldsymbol{\mu}}\Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle\boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle\Big]. \tag{21}$$

The proposition below shows that the strong duality holds upon the constrained forms and their dual forms.

Proposition 1 (Strong Duality)**.** *For any* $A > 0$*, we have*

$$U(A, N, n, d) = \inf_{\lambda\geq 0}\Big[\overline{U}(\lambda, N, n, d) + \lambda A\Big].$$

Moreover, for any $A > \psi_1\|\boldsymbol{a}_{\min}\|_2^2$*, we have*

$$T(A, N, n, d) = \inf_{\lambda\geq 0}\Big[\overline{T}(\lambda, N, n, d) + \lambda A\Big].$$

The proof of Proposition 1 is based on a classical result which states that strongly duality holds for quadratic programs with single quadratic constraint (Appendix B.1 in Boyd & Vandenberghe (2004)).

4.4. Expressions of $\mathcal{U}$ and $\mathcal{T}$

Proposition 1 transforms our task from computing the asymptotics of U and T to that of $\overline{U}$ and $\overline{T}$. The latter is given by the following proposition.

Proposition 2. *Let the target function* f_d *satisfy Assumption 1, the activation function* σ *satisfy Assumption 2, and* (N, n, d) *satisfy Assumption 3. In addition, let Assumption 4 and 5 hold. Then for* $\lambda \in \Lambda_U$*, with high probability the maximizer in Eq. (20) can be achieved at a unique point* $\overline{\boldsymbol{a}}_U(\lambda)$*, and we have*

$$\overline{U}(\lambda, N, n, d) = \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1),$$
$$\psi_1\|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2 = \mathcal{A}_U(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1).$$

Moreover, for any $\lambda \in \Lambda_T$, with high probability the maximizer in Eq. (21) can be achieved at a unique point $\overline{\boldsymbol{a}}_T(\lambda)$, and we have

$$\overline{T}(\lambda, N, n, d) = \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1),$$
$$\psi_1 \|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2 = \mathcal{A}_T(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1).$$

The functions $\overline{\mathcal{U}}, \overline{\mathcal{T}}, \mathcal{A}_U, \mathcal{A}_T$ are given in Definition 5 in Appendix A.

Remark 1. *Here we present the heuristic formulae of $\overline{\mathcal{U}}, \overline{\mathcal{T}}, \mathcal{A}_U, \mathcal{A}_T$, and defer their rigorous definition to the appendix. Define a function $g_0(\boldsymbol{q}; \boldsymbol{\psi})$ by*

$$\begin{aligned} g_0(\boldsymbol{q}; \boldsymbol{\psi}) \equiv\ & \mathrm{ext}_{z_1, z_2}\Big[\log\big((s_2 z_1 + 1)(t_2 z_2 + 1) \\ & - \mu_1^2 (1+p)^2 z_1 z_2\big) - \mu_\star^2 z_1 z_2 + s_1 z_1 + t_1 z_2 \\ & - \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \psi_1 - \psi_2\Big], \end{aligned} \tag{22}$$

where ext *stands for setting z_1 and z_2 to be stationery (which is a common symbol in statistical physics heuristics). We then take*

$$\overline{\mathcal{U}}(\lambda, \boldsymbol{\psi}) = F_1^2(1 - \mu_1^2 \gamma_{s_2} - \gamma_p - \gamma_{t_2}) + \tau^2(1 - \gamma_{t_1}),$$

where $\gamma_a \equiv \partial_a g_0(\boldsymbol{q}; \boldsymbol{\psi})|_{\boldsymbol{q} = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)}$ for the symbol $a \in \{s_1, s_2, t_1, t_2, p\}$, and

$$\overline{\mathcal{T}}(\lambda, \boldsymbol{\psi}) = F_1^2(1 - \mu_1^2 \nu_{s_2} - \nu_p - \nu_{t_2}) + \tau^2(1 - \nu_{t_1}),$$

where we define $\nu_a \equiv \partial_a g_0(\boldsymbol{q}; \boldsymbol{\psi})|_{\boldsymbol{q} = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)}$ for symbols $a \in \{s_1, s_2, t_1, t_2, p\}$. Finally $\mathcal{A}_U = -\partial_\lambda \overline{\mathcal{U}}$, $\mathcal{A}_T = -\partial_\lambda \overline{\mathcal{T}}$. By a further simplification, we can express these formulae to be rational functions of $(\mu_1^2, \mu_\star^2, \lambda, \psi_1, \psi_2, m_1, m_2)$ where (m_1, m_2) is the stationery point of the variational problem in Eq. (22) (c.f. Remark 2).

We next define $\mathcal{U}$ and $\mathcal{T}$ to be dual forms of $\overline{\mathcal{U}}$ and $\overline{\mathcal{T}}$.

Definition 2 (Formula for uniform convergence bounds)**.** *For $A \in \Gamma_U \equiv \{\mathcal{A}_U(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_U\}$, define*

$$\mathcal{U}(A, \psi_1, \psi_2) \equiv \inf_{\lambda \geq 0} \Big[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A\Big].$$

For $A \in \Gamma_T \equiv \{\mathcal{A}_T(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_T\}$, define

$$\mathcal{T}(A, \psi_1, \psi_2) \equiv \inf_{\lambda \geq 0} \Big[\overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) + \lambda A\Big].$$

Finally, we are ready to present the main theorem of this paper, which states that the uniform convergence bounds $U(A, N, n, d)$ and $T(A, N, n, d)$ converge to the formula presented in the definition above.

Theorem 1. *Let the same assumptions in Proposition 2 hold. For any $A \in \Gamma_U$, we have*

$$U(A, N, n, d) = \mathcal{U}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \tag{23}$$

and for $A \in \Gamma_T$ we have

$$T(A, N, n, d) = \mathcal{T}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \tag{24}$$

where functions $\mathcal{U}$ and $\mathcal{T}$ are given in Definition 2.

The proof of this theorem is contained in Section E.

5. Discussions

In this paper, we calculated the uniform convergence bounds for random features models in the proportional scaling regime. Our results exhibit a setting in which standard uniform convergence bound is vacuous while uniform convergence over interpolators gives a non-trivial bound of the actual generalization error.

Modeling assumptions and technical assumptions. We made a few assumptions to prove the main result Theorem 1. Some of these assumptions can be relaxed. Indeed, if we assume a non-linear target function f_d instead of a linear one as in Assumption 1, the non-linear part will behave like additional noises in the proportional scaling limit. However, proving this rigorously requires substantial technical work. Similar issue exists in Mei & Montanari (2019). Moreover, it is not essential to assume vanishing μ_0^2 in Assumption 2.

Assumption 4 and 5 involve some properties of specific random matrices. We believe these assumptions can be proved under more natural assumptions on the activation function σ. However, proving these assumptions requires developing some sophisticated random matrix theory results, which could be of independent interest.

Relationship with non-asymptotic results. We hold the same opinion as in Abbaras et al. (2020): the exact formulae in the asymptotic limit can provide a complementary view to the classical theories of generalization. On the one hand, asymptotic formulae can be used to quantify the tightness of non-asymptotic bounds; on the other hand, the asymptotic formulae in many cases are comparable to non-asymptotic bounds. For example, Lemma 22 in Bartlett & Mendelson (2003) coupled with the bound of Lipschitz constant of the square loss in proper regime implies that $\mathcal{U}_\infty(A, \psi_2)$ have a non-asymptotic bound that scales linearly in A and inverse proportional to $\psi_2^{1/2}$ (c.f. Proposition 6 of E et al. (2020)). This coincides with the intuitions in Section 3.3.

Uniform convergence in other settings. A natural question is whether the power law derived in Section 3 holds for models in more general settings. One can perform a similar analysis to calculate the uniform convergence bounds in a few other settings (Montanari et al., 2019; Dhifallah & Lu, 2020; Hu & Lu, 2020). We believe the power law may be different, but the qualitative properties of uniform convergence bounds will share some similar features.

Relationship with Zhou et al. (2020). The separation of uniform convergence bounds (U and T) is first pointed out

by Zhou et al. (2020), where the authors worked with the linear regression model in the "junk features" setting. We believe random features model are more natural models to illustrate the separation: in Zhou et al. (2020), there are some unnatural parameters λ_n, d_J that are hard to make connections to deep learning models, while the random features model is closely related to two-layer neural networks.

References

Abbaras, A., Aubin, B., Krzakala, F., and Zdeborová, L. Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning. In *Mathematical and Scientific Machine Learning*, pp. 27–54. PMLR, 2020.

Adlam, B. and Pennington, J. Understanding double descent requires a fine-grained bias-variance decomposition. *arXiv preprint arXiv:2011.03321*, 2020.

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

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/arora18b.html.

Bach, F. On the equivalence between quadrature rules and random features. *arXiv preprint arXiv:1502.06800*, pp. 135, 2015.

Bartlett, P. L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. doi: 10.1109/18.661502.

Bartlett, P. L. and Long, P. M. Failures of model-dependent generalization bounds for least-norm interpolation. *arXiv preprint arXiv:2010.08479*, 2020.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, March 2003. ISSN 1532-4435.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 6240–6249. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/

b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/content/early/2020/04/22/1907378117.

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 2300–2311. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper/2018/file/e22312179bf43e61576081a2f250f845-Paper.pdf.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 541–549, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018b. PMLR. URL http://proceedings.mlr.press/v80/belkin18a.html.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549. PMLR, 2018c.

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

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1611–1619. PMLR, 16–18 Apr 2019b. URL http://proceedings.mlr.press/v89/belkin19a.html.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Cantelli, F. Sulla determinazione empirica della legge di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 38(4):421–424, 1933.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., dtextquotesingle Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 10836–10846. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/cf9dc5e4e194fc21f397b4cac9cc3ae9-Paper.pdf.

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

Chihara, T. S. *An introduction to orthogonal polynomials*. Courier Corporation, 2011.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

Dhifallah, O. and Lu, Y. M. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data, 2017.

E, W., Ma, C., and Wu, L. Machine learning from a continuous viewpoint, i. *Science China Mathematics*, 63(11):2233–2266, 2020.

Efthimiou, C. and Frye, C. *Spherical harmonics in p dimensions*. World Scientific, 2014.

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

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33, 2020.

Glivenko, V. Sulla determinazione empirica della legge di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 38(4):92–99, 1933.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/golowich18a.html.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

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

Jacot, A., Şimşek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020b.

Liang, T., Rakhlin, A., and Zhai, X. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv:1908.10292*, 2019.

Liang, T., Rakhlin, A., et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48 (3):1329–1347, 2020.

Louart, C., Liao, Z., Couillet, R., et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Ma, C., Wojtowytsch, S., Wu, L., et al. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, art. arXiv:1908.05355, August 2019.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlche Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 11615–11626. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/05e97c207235d63ceb1db43c60db7bbb-Paper.pdf.

Nagarajan, V. and Kolter, Z. Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=Hygn2o0qKX.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Negrea, J., Dziugaite, G. K., and Roy, D. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pp. 7263–7272. PMLR, 2020.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015a. URL http://arxiv.org/abs/1412.6614.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, Paris, France, 03–06 Jul 2015b. PMLR. URL http://proceedings.mlr.press/v40/Neyshabur15.html.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *International Confer-*

ence on Learning Representations, 2019. URL https://openreview.net/forum?id=BygfghAcYX.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, pp. 1177–1184, 2007. URL http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pp. 1313–1320, 2009.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3215–3225, 2017.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 2019.

Szego, Gabor. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *0*, 2016. URL http://arxiv.org/abs/1611.03530. cite arxiv:1611.03530Comment: Published in ICLR 2017.

Zhou, L., Sutherland, D., and Srebro, N. On uniform convergence and low-norm interpolation learning. *arXiv preprint arXiv:2006.05942*, 2020.

Zhu, J., Gibson, B., and Rogers, T. T. Human rademacher complexity. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 22, pp. 2322–2330. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

A. Definitions of quantities in the main text

A.1. Full definitions of $\overline{\mathcal{U}}$, $\overline{\mathcal{T}}$, $\mathcal{A}_U$, and $\mathcal{A}_T$ in Proposition 2

We first define functions $m_1(\cdot), m_2(\cdot)$, which could be understood as the limiting partial Stieltjes transforms of $\boldsymbol{A}(\boldsymbol{q})$ (c.f. Definition 1).

Definition 3 (Limiting partial Stieltjes transforms)**.** *For $\xi \in \mathbb{C}_+$ and $\boldsymbol{q} \in \mathcal{Q}$ where*

$$\mathcal{Q} = \{(s_1, s_2, t_1, t_2, p) : |s_2 t_2| \le \mu_1^2 (1+p)^2/2\}, \tag{25}$$

define functions $\mathsf{F}_1(\,\cdot\,,\,\cdot\,;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star), \mathsf{F}_2(\,\cdot\,,\,\cdot\,;\xi;\boldsymbol{q},\psi_1,\psi_2,\mu_1,\mu_\star) : \mathbb{C}\times\mathbb{C} \to \mathbb{C}$ via:

$$\mathsf{F}_1(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) \equiv \psi_1 \Big(-\xi + s_1 - \mu_\star^2 m_2 + \frac{(1+t_2 m_2)s_2 - \mu_1^2(1+p)^2 m_2}{(1+s_2 m_1)(1+t_2 m_2) - \mu_1^2(1+p)^2 m_1 m_2} \Big)^{-1},$$
$$\mathsf{F}_2(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) \equiv \psi_2 \Big(-\xi + t_1 - \mu_\star^2 m_1 + \frac{(1+s_2 m_1)t_2 - \mu_1^2(1+p)^2 m_1}{(1+t_2 m_2)(1+s_2 m_1) - \mu_1^2(1+p)^2 m_1 m_2} \Big)^{-1}.$$

Let $m_1(\,\cdot\,;\boldsymbol{q};\boldsymbol{\psi})\ m_2(\,\cdot\,;\boldsymbol{q};\boldsymbol{\psi}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \ge C$ a sufficiently large constant, as the unique solution of the equations

$$\begin{aligned} m_1 &= \mathsf{F}_1(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star), \\ m_2 &= \mathsf{F}_2(m_1, m_2; \xi; \boldsymbol{q}, \psi_1, \psi_2, \mu_1, \mu_\star) \end{aligned} \tag{26}$$

subject to the condition $|m_1| \le \psi_1/\Im(\xi)$, $|m_2| \le \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$.

We next define the function $g(\cdot)$ that will be shown to be the limiting log determinant of $\boldsymbol{A}(\boldsymbol{q})$.

Definition 4 (Limiting log determinants)**.** *For $\boldsymbol{q} = (s_1, s_2, t_1, t_2, p)$ and $\boldsymbol{\psi} = (\psi_1, \psi_2)$, define*

$$\begin{aligned} \Xi(\xi, z_1, z_2; \boldsymbol{q}; \boldsymbol{\psi}) \equiv\ & \log[(s_2 z_1 + 1)(t_2 z_2 + 1) - \mu_1^2(1+p)^2 z_1 z_2] - \mu_\star^2 z_1 z_2 \\ & + s_1 z_1 + t_1 z_2 - \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \xi(z_1 + z_2) - \psi_1 - \psi_2. \end{aligned} \tag{27}$$

Let $m_1(\xi;\boldsymbol{q};\boldsymbol{\psi}), m_2(\xi;\boldsymbol{q};\boldsymbol{\psi})$ be defined as the analytic continuation of solution of Eq. (26) as defined in Definition 3. Define

$$g(\xi; \boldsymbol{q}; \boldsymbol{\psi}) = \Xi(\xi, m_1(\xi;\boldsymbol{q};\boldsymbol{\psi}), m_2(\xi;\boldsymbol{q};\boldsymbol{\psi}); \boldsymbol{q}; \boldsymbol{\psi}). \tag{28}$$

We next give the definitions of $\overline{\mathcal{U}}$, $\overline{\mathcal{T}}$, $\mathcal{A}_U$, and $\mathcal{A}_T$.

Definition 5 ($\overline{\mathcal{U}}$, $\overline{\mathcal{T}}$, $\mathcal{A}_U$, and $\mathcal{A}_T$ in Proposition 2)**.** *For any $\lambda \in \Lambda_U$, define*

$$\begin{aligned} \mathcal{A}_U(\lambda, \psi_1, \psi_2) &= -\lim_{u \to 0_+} \left[\psi_1 \left(F_1^2 \mu_1^2 \partial_{s_1 s_2} + F_1^2 \partial_{s_1 p} + F_1^2 \partial_{s_1 t_2} + \tau^2 \partial_{s_1 t_1} \right) g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}) \Big|_{\boldsymbol{q}=\boldsymbol{q}_U} \right], \\ \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) &= F_1^2 + \tau^2 - \lim_{u \to 0_+} \left[\left(F_1^2 \mu_1^2 \partial_{s_2} + F_1^2 \partial_p + F_1^2 \partial_{t_2} + \tau^2 \partial_{t_1} \right) g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}) \Big|_{\boldsymbol{q}=\boldsymbol{q}_U} \right], \\ \mathcal{A}_T(\lambda, \psi_1, \psi_2) &= -\lim_{u \to 0_+} \left[\psi_1 \left(F_1^2 \mu_1^2 \partial_{s_1 s_2} + F_1^2 \partial_{s_1 p} + F_1^2 \partial_{s_1 t_2} + \tau^2 \partial_{s_1 t_1} \right) g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}) \Big|_{\boldsymbol{q}=\boldsymbol{q}_T} \right], \\ \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) &= F_1^2 + \tau^2 - \lim_{u \to 0_+} \left[\left(F_1^2 \mu_1^2 \partial_{s_2} + F_1^2 \partial_p + F_1^2 \partial_{t_2} + \tau^2 \partial_{t_1} \right) g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}) \Big|_{\boldsymbol{q}=\boldsymbol{q}_T} \right], \end{aligned}$$

where $\boldsymbol{q}_U = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$, $\boldsymbol{q}_T = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$.

In the following, we give a simplified expression for $\overline{\mathcal{U}}$ and $\mathcal{A}_U$.

Remark 2 (Simplification of $\overline{\mathcal{U}}$ and $\mathcal{A}_U$)**.** *Define $\zeta, \overline{\lambda}$ as the rescaled version of μ_1^2 and λ*

$$\zeta = \frac{\mu_1^2}{\mu_\star^2}, \quad \overline{\lambda} = \frac{\lambda}{\mu_\star^2}.$$

Let $m_1(\,\cdot\,;\boldsymbol{\psi})\; m_2(\,\cdot\,;\boldsymbol{\psi}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \geq C$ a sufficiently large constant, as the unique solution of the equations

$$\begin{aligned} m_1 &= \psi_1 \left[-\xi + (1-\overline{\lambda}\psi_1) - m_2 + \frac{\zeta(1-m_2)}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1}, \\ m_2 &= -\psi_2 \left[\xi + \psi_2 - m_1 - \frac{\zeta m_1}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1}, \end{aligned} \tag{29}$$

subject to the condition $|m_1| \leq \psi_1/\Im(\xi)$, $|m_2| \leq \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$. Let

$$\begin{aligned} \overline{m}_1 &= \lim_{u\to\infty} m_1(\boldsymbol{i}u, \boldsymbol{\psi}), \\ \overline{m}_2 &= \lim_{u\to\infty} m_2(\boldsymbol{i}u, \boldsymbol{\psi}). \end{aligned}$$

Define

$$\begin{aligned} \chi_1 &= \overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta + 1, \\ \chi_2 &= \overline{m}_1 - \psi_2 + \frac{\overline{m}_1\zeta}{\chi_1}, \\ \chi_3 &= \overline{\lambda}\psi_1 + \overline{m}_2 - 1 + \frac{\zeta\left(\overline{m}_2 - 1\right)}{\chi_1}. \end{aligned}$$

Define two polynomials $\mathcal{E}_1, \mathcal{E}_2$ as

$$\begin{aligned} \mathcal{E}_1(\psi_1,\psi_2,\overline{\lambda},\zeta) &= \psi_1^2(\psi_2\chi_1^4 + \psi_2\chi_1^2\zeta), \\ \mathcal{E}_2(\psi_1,\psi_2,\overline{\lambda},\zeta) &= \psi_1^2(\chi_1^2\chi_2^2\overline{m}_2^2\zeta - 2\chi_1^2\chi_2^2\overline{m}_2\zeta + \chi_1^2\chi_2^2\zeta + \psi_2\chi_1^2 - \psi_2\overline{m}_1^2\overline{m}_2^2\zeta^3 + 2\psi_2\overline{m}_1^2\overline{m}_2\zeta^3 - \psi_2\overline{m}_1^2\zeta^3 + \psi_2\zeta), \\ \mathcal{E}_3(\psi_1,\psi_2,\overline{\lambda},\zeta) &= -\chi_1^4\chi_2^2\chi_3^2 + \psi_1\psi_2\chi_1^4 + \psi_1\chi_1^2\chi_2^2\overline{m}_2^2\zeta^2 - 2\psi_1\chi_1^2\chi_2^2\overline{m}_2\zeta^2 + \psi_1\chi_1^2\chi_2^2\zeta^2 \\ &\quad + \psi_2\chi_1^2\chi_3^2\overline{m}_1^2\zeta^2 + 2\psi_1\psi_2\chi_1^2\zeta - \psi_1\psi_2\overline{m}_1^2\overline{m}_2^2\zeta^4 + 2\psi_1\psi_2\overline{m}_1^2\overline{m}_2\zeta^4 - \psi_1\psi_2\overline{m}_1^2\zeta^4 + \psi_1\psi_2\zeta^2. \end{aligned}$$

Then

$$\begin{aligned} \overline{\mathcal{U}}(\overline{\lambda},\psi_1,\psi_2) &= -\frac{\left(\overline{m}_2 - 1\right)\left(\tau^2\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta) + F_1^2\right)}{\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta)}, \\ \mathcal{A}_U(\overline{\lambda},\psi_1,\psi_2) &= \frac{\tau^2\mathcal{E}_1(\psi_1,\psi_2,\overline{\lambda},\zeta) + F_1^2\mathcal{E}_1(\psi_1,\psi_2,\overline{\lambda},\zeta)}{\mathcal{E}_2(\psi_1,\psi_2,\overline{\lambda},\zeta)}. \end{aligned}$$

Remark 3 (Simplification of $\overline{\mathcal{T}}$ and $\mathcal{A}_T$)**.** *Define $\zeta, \overline{\lambda}$ as the rescaled version of μ_1^2 and λ*

$$\zeta = \frac{\mu_1^2}{\mu_\star^2}, \;\; \overline{\lambda} = \frac{\lambda}{\mu_\star^2}.$$

Let $m_1(\,\cdot\,;\boldsymbol{\psi})\; m_2(\,\cdot\,;\boldsymbol{\psi}) : \mathbb{C}_+ \to \mathbb{C}_+$ be defined, for $\Im(\xi) \geq C$ a sufficiently large constant, as the unique solution of the equations

$$\begin{aligned} m_1 &= \psi_1 \left[-\xi + (1-\overline{\lambda}\psi_1) - m_2 + \frac{\zeta(1-m_2)}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1}, \\ m_2 &= -\psi_2 \left[\xi + m_1 + \frac{\zeta m_1}{1+\zeta m_1 - \zeta m_1 m_2} \right]^{-1}, \end{aligned} \tag{30}$$

subject to the condition $|m_1| \leq \psi_1/\Im(\xi)$, $|m_2| \leq \psi_2/\Im(\xi)$. Extend this definition to $\Im(\xi) > 0$ by requiring m_1, m_2 to be analytic functions in $\mathbb{C}_+$. Let

$$\begin{aligned} \overline{m}_1 &= \lim_{u\to\infty} m_1(\boldsymbol{i}u, \boldsymbol{\psi}), \\ \overline{m}_2 &= \lim_{u\to\infty} m_2(\boldsymbol{i}u, \boldsymbol{\psi}). \end{aligned}$$

Define

$$\chi_4 = \overline{m}_1 + \frac{\overline{m}_1\zeta}{\chi_1(\overline{m}_1,\overline{m}_2,\zeta)},$$

and

$$\chi_1 = \overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta + 1,$$
$$\chi_3 = \overline{\lambda}\psi_1 + \overline{m}_2 - 1 + \frac{\zeta\left(\overline{m}_2 - 1\right)}{\chi_1},$$

where the definitions of χ_1, χ_3 are the same as in Remark 2. Define three polynomials $\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5$ as

$$\begin{aligned}
\mathcal{E}_4(\psi_1,\psi_2,\overline{\lambda},\zeta) =& \psi_1\Big(\psi_2\chi_1^4\chi_4^3 + \chi_1^4\chi_4^2\overline{m}_1^3\overline{m}_2^2\zeta^3 - 2\chi_1^4\chi_4^2\overline{m}_1^3\overline{m}_2\zeta^3 + \chi_1^4\chi_4^2\overline{m}_1^3\zeta^3 + 2\chi_1^3\chi_4^2\overline{m}_1^3\overline{m}_2^2\zeta^2 \\
&- 4\chi_1^3\chi_4^2\overline{m}_1^3\overline{m}_2\zeta^2 + 2\chi_1^3\chi_4^2\overline{m}_1^3\zeta^2 - \psi_2\chi_1^3\chi_4^2\overline{m}_1\zeta + \chi_1^2\chi_4^2\overline{m}_1^3\overline{m}_2^2\zeta - 2\chi_1^2\chi_4^2\overline{m}_1^3\overline{m}_2\zeta \\
&+ \chi_1^2\chi_4^2\overline{m}_1^3\zeta + \psi_2\chi_1^2\chi_4^2\overline{m}_1\zeta - \psi_2\chi_1^2\overline{m}_1^5\overline{m}_2^2\zeta^5 + 2\psi_2\chi_1^2\overline{m}_1^5\overline{m}_2\zeta^5 - \psi_2\chi_1^2\overline{m}_1^5\zeta^5 \\
&- 2\psi_2\chi_1\overline{m}_1^5\overline{m}_2^2\zeta^4 + 4\psi_2\chi_1\overline{m}_1^5\overline{m}_2\zeta^4 - 2\psi_2\chi_1\overline{m}_1^5\zeta^4 - \psi_2\overline{m}_1^5\overline{m}_2^2\zeta^3 \\
&+ 2\psi_2\overline{m}_1^5\overline{m}_2\zeta^3 - \psi_2\overline{m}_1^5\zeta^3\Big), \\
\mathcal{E}_5(\psi_1,\psi_2,\overline{\lambda},\zeta) =& \overline{m}_1\Big(\zeta + 1 + \overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta\Big)^2\Big(-\chi_1^4\chi_3^2\chi_4^2\overline{m}_1^2 \\
&+ \psi_1\psi_2\chi_1^4\chi_4^2 - 2\psi_1\psi_2\chi_1^3\chi_4\overline{m}_1\zeta + \psi_2\chi_1^2\chi_3^2\overline{m}_1^4\zeta^2 + \psi_1\chi_1^2\chi_4^2\overline{m}_1^2\overline{m}_2^2\zeta^2 \\
&- 2\psi_1\chi_1^2\chi_4^2\overline{m}_1^2\overline{m}_2\zeta^2 + \psi_1\chi_1^2\chi_4^2\overline{m}_1^2\zeta^2 + 2\psi_1\psi_2\chi_1^2\chi_4\overline{m}_1\zeta + \psi_1\psi_2\chi_1^2\overline{m}_1^2\zeta^2 \\
&- 2\psi_1\psi_2\chi_1\overline{m}_1^2\zeta^2 - \psi_1\psi_2\overline{m}_1^4\overline{m}_2^2\zeta^4 + 2\psi_1\psi_2\overline{m}_1^4\overline{m}_2\zeta^4 - \psi_1\psi_2\overline{m}_1^4\zeta^4 + \psi_1\psi_2\overline{m}_1^2\zeta^2\Big), \\
\mathcal{E}_6(\psi_1,\psi_2,\overline{\lambda},\zeta) =& \chi_1^2\chi_4^2\psi_1\psi_2\Big(\chi_4\chi_1^2 - \overline{m}_1\chi_1\zeta + \overline{m}_1\zeta\Big)\Big(\overline{m}_1\zeta - \overline{m}_1\overline{m}_2\zeta + 1\Big)^2.
\end{aligned}$$

Then

$$\overline{\mathcal{T}}(\overline{\lambda},\psi_1,\psi_2) = -\frac{\left(\overline{m}_2 - 1\right)\left(\tau^2\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta) + F_1^2\right)}{\chi_1(\psi_1,\psi_2,\overline{\lambda},\zeta)},$$
$$\mathcal{A}_T(\overline{\lambda},\psi_1,\psi_2) = -\psi_1\frac{F_1^2\mathcal{E}_4(\psi_1,\psi_2,\overline{\lambda},\zeta) + \tau^2\mathcal{E}_6(\psi_1,\psi_2,\overline{\lambda},\zeta)}{\mathcal{E}_5(\psi_1,\psi_2,\overline{\lambda},\zeta)}.$$

A.2. Definitions of $\mathcal{R}$ and $\mathcal{A}$

In this section, we present the expression of $\mathcal{R}$ and $\mathcal{A}$ from Mei & Montanari (2019) which are used in our results and plots.

Definition 6 (Formula for the prediction error of minimum norm interpolator)**.** *Define*

$$\zeta = \mu_1^2/\mu_\star^2, \quad \rho = F_1^2/\tau^2$$

Let the functions $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ be be uniquely defined by the following conditions: (i) ν_1, ν_2 are analytic on $\mathbb{C}_+$; (ii) For $\Im(\xi) > 0$, $\nu_1(\xi)$, $\nu_2(\xi)$ satisfy the following equations

$$\begin{aligned}
\nu_1 &= \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1 - \zeta^2\nu_1\nu_2}\Big)^{-1}, \\
\nu_2 &= \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1 - \zeta^2\nu_1\nu_2}\Big)^{-1};
\end{aligned} \tag{31}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ is the unique solution of these equations with $|\nu_1(\xi)| \le \psi_1/\Im(\xi)$, $|\nu_2(\xi)| \le \psi_2/\Im(\xi)$ for $\Im(\xi) > C$, with C a sufficiently large constant.

Let

$$\chi \equiv \lim_{u\to 0}\nu_1(\boldsymbol{i}u)\cdot\nu_2(\boldsymbol{i}u), \tag{32}$$

and

$$\begin{aligned}
E_0(\zeta,\psi_1,\psi_2) \equiv& -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
&+ (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
E_1(\zeta,\psi_1,\psi_2) \equiv& \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\,, \\
E_2(\zeta,\psi_1,\psi_2) \equiv& \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\,.
\end{aligned} \tag{33}$$

Then the expression for the asymptotic risk of minimum norm interpolator gives

$$\mathcal{R}(\psi_1,\psi_2) = F_1^2 \frac{E_1(\zeta,\psi_1,\psi_2)}{E_0(\zeta,\psi_1,\psi_2)} + \tau^2 \frac{E_2(\zeta,\psi_1,\psi_2)}{E_0(\zeta,\psi_1,\psi_2)} + \tau^2.$$

The expression for the norm of the minimum norm interpolator gives

$$\begin{aligned}
A_1 &= \frac{\rho}{1+\rho}\Big[-\chi^2(\chi\zeta^4 - \chi\zeta^2 + \psi_2\zeta^2 + \zeta^2 - \chi\psi_2\zeta^4 + 1)\Big] + \frac{1}{1+\rho}\Big[\chi^2(\chi\zeta^2-1)(\chi^2\zeta^4 - 2\chi\zeta^2 + \zeta^2 + 1)\Big],\\
A_0 &= -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2\\
&\quad + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2,\\
\mathcal{A}(\psi_1,\psi_2) &= \psi_1(F_1^2+\tau^2)A_1/(\mu_\star^2 A_0).
\end{aligned}$$

B. Experimental setup for simulations in Figure 2

In this section, we present additional details for Figure 2. We choose $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle$ for some $\|\boldsymbol{\beta}\|_2^2 = 1$, the ReLU activation function $\sigma(x) = \max\{x, 0\}$, and $\psi_1 = N/d = 2.5$ and $\psi_2 = n/d = 1.5$.

For the theoretical curves (in solid lines), we choose $\lambda \in [0.426, 2]$, so that $\mathcal{A}_U(\lambda) \in [0, 15]$, and plot the parametric curve $(\mathcal{A}_U(\lambda), \overline{\mathcal{U}}(\lambda) + \lambda\mathcal{A}_U(\lambda))$ for the uniform convergence. For the uniform convergence over interpolators, we choose $\lambda \in [0.21, 2]$ so that $\mathcal{A}_T(\lambda) \in [6.4, 15]$, and plot $(\mathcal{A}_T(\lambda), \overline{\mathcal{T}}(\lambda) + \lambda\mathcal{A}_T(\lambda))$. The definitions of these theoretical predictions are given in Definition 5, Remark 2 and Remark 3

For the empirical simulations (in dots), first recall that in Proposition 2, we defined

$$\begin{aligned}
\boldsymbol{a}_U(\lambda) &= \arg\max_{\boldsymbol{a}} \Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda\|\boldsymbol{a}\|_2^2\Big],\\
\boldsymbol{a}_T(\lambda) &= \arg\max \boldsymbol{a} \inf_{\boldsymbol{\mu}} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle\Big].
\end{aligned}$$

After picking a value of λ, we sample 20 independent problem instances, with the number of features $N = 500$, number of samples $n = 300$, covariate dimension $d = 200$. We compute the corresponding $(\psi_1\|\boldsymbol{a}_U\|_2^2, R(\boldsymbol{a}_U) - \hat{R}_n(\boldsymbol{a}_U))$ and $(\psi_1\|\boldsymbol{a}_T\|_2^2, R(\boldsymbol{a}_T))$ for each instance. Then, we plot the empirical mean and $1/\sqrt{20}$ times the empirical standard deviation (around the mean) of each coordinate.

C. Proof of Proposition 1

The proof of Proposition 1 contains two parts: standard uniform convergence U and uniform convergence over interpolators T. The proof for the two cases are essentially the same, both based on the fact that strong duality holds for quadratic program with single quadratic constraint (c.f. Boyd & Vandenberghe (2004), Appendix A.1).

C.1. Standard uniform convergence U

Recall that the uniform convergence bound U is defined as in Eq. (4)

$$U(A,N,n,d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a})\Big).$$

Since the maximization problem in (4) is a quadratic program with a single quadratic constraint, the strong duality holds. So we have

$$\sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A^2} R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) = \inf_{\lambda\ge 0}\sup_{\boldsymbol{a}} \Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda(\|\boldsymbol{a}\|_2^2 - \psi_1^{-1}A)\Big].$$

Finally, by the definition of $\overline{U}$ as in Eq. (20), we get

$$U(A,N,n,d) = \inf_{\lambda\ge 0}\Big[\overline{U}(\lambda,N,n,d) + \lambda A\Big].$$

C.2. Uniform convergence over interpolators T

Without loss of generality, we consider the regime when $N > n$.

Recall that the uniform convergence over interpolators T is defined as in Eq. (5)

$$T(A, N, n, d) = \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}).$$

When the set $\{\boldsymbol{a} \in \mathbb{R}^N : (N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a}) = 0\}$ is empty, we have

$$T(A, N, n, d) = \inf_{\lambda \ge 0} \Big[\overline{T}(\lambda, N, n, d) + \lambda A\Big] = -\infty.$$

In the following, we assume that the set $\{\boldsymbol{a} \in \mathbb{R}^N : (N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a}) = 0\}$ is non-empty, i.e., there exists $\boldsymbol{a} \in \mathbb{R}^N$ such that $\widehat{R}_n(\boldsymbol{a}) = 0$ and $(N/d)\|\boldsymbol{a}\|_2^2 \le A$.

Let m be the dimension of the null space of $\boldsymbol{Z} \in \mathbb{R}^{n\times N}$, i.e. $m = \dim(\{\boldsymbol{u} : \boldsymbol{Z}\boldsymbol{u} = \boldsymbol{0}\})$. Note that $\boldsymbol{Z} \in \mathbb{R}^{N\times n}$ and $N > n$, we must have $N - n \le m \le N$. We let $\boldsymbol{R} \in \mathbb{R}^{N\times m}$ be a matrix whose column space gives the null space of matrix $\boldsymbol{Z}$. Let $\boldsymbol{a}_0$ be the minimum norm interpolating solution (whose existence is given by the assumption that $\{\boldsymbol{a} \in \mathbb{R}^N : \widehat{R}_n(\boldsymbol{a}) = 0\}$ is non-empty)

$$\boldsymbol{a}_0 = \lim_{\lambda \to 0_+} \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \Big[\widehat{R}_n(\boldsymbol{a}) + \lambda\|\boldsymbol{a}\|_2^2\Big] = \arg\min_{\boldsymbol{a}\in\mathbb{R}^N : \widehat{R}_n(\boldsymbol{a})=0} \|\boldsymbol{a}\|_2^2.$$

Then we have

$$\{\boldsymbol{a} \in \mathbb{R}^N : \widehat{R}_n(\boldsymbol{a}) = 0\} = \{\boldsymbol{a} \in \mathbb{R}^N : \boldsymbol{y} = \sqrt{d}\boldsymbol{Z}\boldsymbol{a}\} = \{\boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0 : \boldsymbol{u} \in \mathbb{R}^m\}.$$

Then T can be rewritten as a maximization problem in terms of $\boldsymbol{u}$:

$$\begin{aligned}\sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}) &= \sup_{\boldsymbol{u}\in\mathbb{R}^m : \|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2 \le \psi_1^{-1}A} \Big[\langle \boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0, \boldsymbol{U}(\boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0)\rangle - 2\langle \boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0, \boldsymbol{v}\rangle + \mathbb{E}(y^2)\Big] \\ &= R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}\in\mathbb{R}^m : \|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2 \le \psi_1^{-1}A} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle\Big].\end{aligned}$$

Note that the optimization problem only has non-feasible region when $A > (N/d)\|\boldsymbol{a}_0\|_2^2$. By strong duality of quadratic programs with a single quadratic constraint, we have

$$\begin{aligned}&\sup_{\boldsymbol{u}\in\mathbb{R}^m : \|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2 \le \psi_1^{-1}A} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle\Big] \\ &= \inf_{\lambda\ge 0} \sup_{\boldsymbol{u}\in\mathbb{R}^m} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle - \lambda(\psi_1\|\boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0\|_2^2 - A)\Big].\end{aligned}$$

The maximization over $\boldsymbol{u}$ can be restated as the maximization over $\boldsymbol{a}$:

$$R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}\in\mathbb{R}^m} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0 - \boldsymbol{v}\rangle - \lambda\psi_1\|\boldsymbol{R}\boldsymbol{u} + \boldsymbol{a}_0\|_2^2\Big] = \sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big].$$

Moreover, since $\sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0}[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2]$ is a quadratic programming with linear constraints, we have

$$\sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big] = \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle\Big].$$

Combining all the equality above and the definition of $\overline{T}$ as in Eq. (21), we have

$$
\begin{aligned}
T(A,N,n,d) &= \sup_{(N/d)\|\boldsymbol{a}\|_2^2\le A,\widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}) \\
&= R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}\in\mathbb{R}^m:\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\le\psi_1^{-1}A} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0-\boldsymbol{v}\rangle\Big] \\
&= R(\boldsymbol{a}_0) + \inf_{\lambda\ge0}\sup_{\boldsymbol{u}} \Big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0-\boldsymbol{v}\rangle - \lambda(\psi_1\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2 - A)\Big] \\
&= \inf_{\lambda\ge0}\Big\{\lambda A + R(\boldsymbol{a}_0) + \sup_{\boldsymbol{u}}\big[\langle \boldsymbol{u}, \boldsymbol{R}^\mathsf{T}\boldsymbol{U}\boldsymbol{R}\boldsymbol{u}\rangle + 2\langle \boldsymbol{R}\boldsymbol{u}, \boldsymbol{U}\boldsymbol{a}_0-\boldsymbol{v}\rangle - \lambda\psi_1\|\boldsymbol{R}\boldsymbol{u}+\boldsymbol{a}_0\|_2^2\big]\Big\} \\
&= \inf_{\lambda\ge0}\Big\{\lambda A + \sup_{\boldsymbol{a}:\widehat{R}_n(\boldsymbol{a})=0}\Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2\Big]\Big\} \\
&= \inf_{\lambda\ge0}\Big\{\lambda A + \sup_{\boldsymbol{a}}\inf_{\boldsymbol{\mu}}\Big[R(\boldsymbol{a}) - \lambda\psi_1\|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle\Big]\Big\} \\
&= \inf_{\lambda\ge0}\Big[\overline{T}(\lambda,N,n,d) + \lambda A\Big].
\end{aligned}
$$

This concludes the proof.

D. Proof of Proposition 2

Note that the definitions of $\overline{U}$ and $\overline{T}$ as in Eq. (20) and (21) depend on $\boldsymbol{\beta} = \boldsymbol{\beta}^{(d)}$, where $\boldsymbol{\beta}^{(d)}$ gives the coefficients of the target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\beta}^{(d)}\rangle$. Suppose we explicitly write their dependence on $\boldsymbol{\beta} = \boldsymbol{\beta}^{(d)}$, i.e., $\overline{U}(\lambda,N,n,d) = \overline{U}(\boldsymbol{\beta},\lambda,N,n,d)$ and $\overline{T}(\lambda,N,n,d) = \overline{T}(\boldsymbol{\beta},\lambda,N,n,d)$, then we can see that for any fixed $\boldsymbol{\beta}_\star$ and $\tilde{\boldsymbol{\beta}}$ with $\|\tilde{\boldsymbol{\beta}}\|_2 = \|\boldsymbol{\beta}_\star\|_2$, we have $\overline{U}(\boldsymbol{\beta}_\star,\lambda,N,n,d) \overset{d}{=} \overline{U}(\tilde{\boldsymbol{\beta}},\lambda,N,n,d)$ and $\overline{T}(\boldsymbol{\beta}_\star,\lambda,N,n,d) \overset{d}{=} \overline{T}(\tilde{\boldsymbol{\beta}},\lambda,N,n,d)$ where the randomness comes from $\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}$. This is by the fact that the distribution of $\boldsymbol{x}_i$'s and $\boldsymbol{\theta}_a$'s are rotationally invariant. As a consequence, for any fixed deterministic $\boldsymbol{\beta}_\star$, if we take $\boldsymbol{\beta}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\|\boldsymbol{\beta}_\star\|_2))$, we have

$$
\begin{aligned}
\overline{U}(\boldsymbol{\beta}_\star,\lambda,N,n,d) &\overset{d}{=} \overline{U}(\boldsymbol{\beta},\lambda,N,n,d), \\
\overline{T}(\boldsymbol{\beta}_\star,\lambda,N,n,d) &\overset{d}{=} \overline{T}(\boldsymbol{\beta},\lambda,N,n,d).
\end{aligned}
$$

where the randomness comes from $\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon},\boldsymbol{\beta}$.

Consequently, as long as we are able to show the equation

$$
\overline{U}(\boldsymbol{\beta},\lambda,N,n,d) = \overline{\mathcal{U}}(\lambda,\psi_1,\psi_2) + o_{d,\mathbb{P}}(1)
$$

for random $\boldsymbol{\beta}\sim\mathrm{Unif}(\mathbb{S}^{n-1}(F_1))$, this equation will also hold for any deterministic $\boldsymbol{\beta}_\star$ with $\|\boldsymbol{\beta}_\star\|_2^2 = F_1^2$. Vice versa for $\overline{T}$, $\|\overline{\boldsymbol{a}}_U\|_2^2$ and $\|\overline{\boldsymbol{a}}_T\|_2^2$.

As a result, in the following, we work with the assumption that $\boldsymbol{\beta} = \boldsymbol{\beta}^{(d)} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$. That is, in proving Proposition 2, we replace Assumption 1 by Assumption 6 below. By the argument above, as long as Proposition 2 holds under Assumption 6, it also holds under the original assumption, i.e., Assumption 1.

Assumption 6 (Linear Target Function)**.** *We assume that $f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ with $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}^{(d)}, \boldsymbol{x}\rangle$, where $\boldsymbol{\beta}^{(d)} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$.*

D.1. Expansions

Denote $\boldsymbol{v} = (v_i)_{i\in[N]} \in \mathbb{R}^N$ and $\boldsymbol{U} = (U_{ij})_{i,j\in[N]} \in \mathbb{R}^{N\times N}$ where their elements are defined via

$$
\begin{aligned}
v_i &\equiv \mathbb{E}_{\varepsilon,\boldsymbol{x}}[y\sigma(\langle \boldsymbol{x},\boldsymbol{\theta}_i\rangle/\sqrt{d})], \\
U_{ij} &\equiv \mathbb{E}_{\boldsymbol{x}}[\sigma(\langle \boldsymbol{x},\boldsymbol{\theta}_i\rangle/\sqrt{d})\sigma(\langle \boldsymbol{x},\boldsymbol{\theta}_j\rangle/\sqrt{d})].
\end{aligned}
$$

Here, $y = \langle \boldsymbol{x},\boldsymbol{\beta}\rangle + \varepsilon$, where $\boldsymbol{\beta}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$, $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\varepsilon\sim\mathcal{N}(0,\tau^2)$, and $(\boldsymbol{\theta}_j)_{j\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ are mutually independent. The expectations are taken with respect to the test sample $\boldsymbol{x}\sim\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\varepsilon\sim\mathcal{N}(0,\tau^2)$ (especially, the expectations are conditional on $\boldsymbol{\beta}$ and $(\boldsymbol{\theta}_i)_{i\in[N]}$).

Moreover, we denote $\boldsymbol{y} = (y_1, \dots, y_n)^\mathsf{T} \in \mathbb{R}^n$ where $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \boldsymbol{\varepsilon}_i$. Recall that $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $(\varepsilon_i)_{i\in[n]} \sim_{iid} \mathcal{N}(0, \tau^2)$ are mutually independent and independent from $\boldsymbol{\beta} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. We further denote $\boldsymbol{Z} = (Z_{ij})_{i\in[n], j\in[N]}$ where its elements are defined via

$$Z_{ij} = \sigma(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle / \sqrt{d}) / \sqrt{d}.$$

The population risk (1) can be reformulated as

$$R(\boldsymbol{a}) = \langle \boldsymbol{a}, \boldsymbol{U}\boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \boldsymbol{v} \rangle + \mathbb{E}[y^2],$$

where $\boldsymbol{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$. The empirical risk (2) can be reformulated as

$$\widehat{R}_n(\boldsymbol{a}) = \psi_2^{-1} \langle \boldsymbol{a}, \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{a} \rangle - 2\psi_2^{-1} \frac{\langle \boldsymbol{Z}^\mathsf{T}\boldsymbol{y}, \boldsymbol{a} \rangle}{\sqrt{d}} + \frac{1}{n}\|\boldsymbol{y}\|_2^2.$$

By the Appendix A in Mei & Montanari (2019) (we include in the Appendix F for completeness), we can expand $\sigma(x)$ in terms of Gegenbauer polynommials

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d} \cdot x),$$

where $Q_k^{(d)}$ is the k'th Gegenbauer polynomial in d dimensions, $B(d,k)$ is the dimension of the space of polynomials on $\mathbb{S}^{d-1}(\sqrt{d})$ with degree exactly k. Finally, $\lambda_{d,k}(\sigma)$ is the k'th Gegenbauer coefficient. More details of this expansion can be found in Appendix F.

By the properties of Gegenbauer polynomials (c.f. Appendix F.2), we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\boldsymbol{x} Q_k(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle)] &= \boldsymbol{0}, & \forall k \neq 1, \\ \mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\boldsymbol{x} Q_1(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle)] &= \boldsymbol{\theta}_i / d, & k = 1. \end{aligned}$$

As a result, we have

$$v_i = \mathbb{E}_{\varepsilon, \boldsymbol{x}}[y\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle / \sqrt{d})] = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) \mathbb{E}_{\boldsymbol{x}}[\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle Q_k^{(d)}(\sqrt{d} \cdot x)] = \lambda_{d,1}(\sigma) \langle \boldsymbol{\theta}_i, \boldsymbol{\beta} \rangle. \tag{34}$$

D.2. Removing the perturbations

By Lemma 6 and 7 as in Appendix D.6, we have the following decomposition

$$\boldsymbol{U} = \mu_1^2 \boldsymbol{Q} + \mu_\star^2 \mathbf{I}_N + \boldsymbol{\Delta}, \tag{35}$$

with $\boldsymbol{Q} = \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}/d$, $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$, and μ_1^2 and $\mu_\star^2$ are given in Assumption 2.

In the following, we would like to show that $\boldsymbol{\Delta}$ has vanishing effects in the asymptotics of $\overline{U}$, $\overline{T}$, $\|\overline{\boldsymbol{a}}_U\|_2^2$ and $\|\overline{\boldsymbol{a}}_T\|_2^2$.

For this purpose, we denote

$$\begin{aligned} \boldsymbol{U}_c &= \mu_1^2 \boldsymbol{Q} + \mu_\star^2 \mathbf{I}_N, \\ R_c(\boldsymbol{a}) &= \langle \boldsymbol{a}, \boldsymbol{U}_c \boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \boldsymbol{v} \rangle + \mathbb{E}[y^2], \\ \widehat{R}_{c,n}(\boldsymbol{a}) &= \langle \boldsymbol{a}, \psi_2^{-1} \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{a} \rangle - 2\langle \boldsymbol{a}, \psi_2^{-1} \boldsymbol{Z}^\mathsf{T}\boldsymbol{y}/\sqrt{d} \rangle + \mathbb{E}[y^2], \\ \overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(R_c(\boldsymbol{a}) - \widehat{R}_{c,n}(\boldsymbol{a}) - \psi_1 \lambda \|\boldsymbol{a}\|_2^2 \Big), \\ \overline{T}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \Big[R_c(\boldsymbol{a}) - \lambda\psi_1 \|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d} \rangle \Big]. \end{aligned} \tag{36}$$

For a fixed $\lambda \in \Lambda_U$, note we have

$$\begin{aligned}\overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, (\boldsymbol{U}_c - \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} - \psi_1 \lambda \mathbf{I}_N) \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \boldsymbol{v} - \psi_2^{-1} \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}} \rangle \Big) \\ &= \sup_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, \overline{\boldsymbol{M}} \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \overline{\boldsymbol{v}} \rangle \Big)\end{aligned} \tag{37}$$

where $\overline{\boldsymbol{M}} = \boldsymbol{U}_c - \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{Z} - \psi_1 \lambda \mathbf{I}_N$ and $\overline{\boldsymbol{v}} = \boldsymbol{v} - \psi_2^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{y} / \sqrt{d}$. When $\boldsymbol{X}, \boldsymbol{\Theta}$ are such that the good event in Assumption 4 happens (which says that $\overline{\boldsymbol{M}} \preceq -\varepsilon \mathbf{I}_N$ for some $\varepsilon > 0$), the inner maximization can be uniquely achieved at

$$\overline{\boldsymbol{a}}_{U,c}(\lambda) = \arg\max_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, \overline{\boldsymbol{M}} \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \overline{\boldsymbol{v}} \rangle \Big) = \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}}. \tag{38}$$

and when the good event $\{ \|\boldsymbol{\Delta}\|_{\mathrm{op}} \le \varepsilon/2 \}$ also happens, the maximizer in the definition of $\overline{U}(\lambda, N, n, d)$ (c.f. Eq. (20)) can be uniquely achieved at

$$\overline{\boldsymbol{a}}_U(\lambda) = \arg\max_{\boldsymbol{a}} \Big(\langle \boldsymbol{a}, (\overline{\boldsymbol{M}} + \boldsymbol{\Delta}) \boldsymbol{a} \rangle - 2 \langle \boldsymbol{a}, \overline{\boldsymbol{v}} \rangle \Big) = (\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \overline{\boldsymbol{v}}.$$

Note we have

$$\overline{\boldsymbol{a}}_U(\lambda) - \overline{\boldsymbol{a}}_{U,c}(\lambda) = (\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \overline{\boldsymbol{v}} - \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}} = (\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta} \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}},$$

so by the fact that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, we have

$$\|\overline{\boldsymbol{a}}_U(\lambda) - \overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2 \le \|(\overline{\boldsymbol{M}} + \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta}\|_{\mathrm{op}} \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2 = o_{d,\mathbb{P}}(1) \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2.$$

This gives $\|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2 = (1 + o_{d,\mathbb{P}}(1)) \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$.

Moreover, by the fact that $\|\boldsymbol{\Delta}\|_{\mathrm{op}} = o_{d,\mathbb{P}}(1)$, we have

$$\begin{aligned}\overline{U}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1 \lambda \|\boldsymbol{a}\|_2^2 - \langle \boldsymbol{a}, \boldsymbol{\Delta} \boldsymbol{a} \rangle \Big) + \mathbb{E}[y^2] - \|\boldsymbol{y}\|_2^2 / n \\ &= \overline{U}(\lambda, N, n, d) + o_{d,\mathbb{P}}(1) (\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2 + 1).\end{aligned}$$

As a consequence, as long as we can prove the asymptotics of $\overline{U}_c$ and $\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$, it also gives the asymptotics of $\overline{U}$ and $\|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2$. Vice versa for $\overline{T}$ and $\|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2$.

D.3. The asymptotics of $\overline{U}_c$ and $\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$

In the following, we derive the asymptotics of $\overline{U}_c(\lambda, N, n, d)$ and $\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$. When we refer to $\overline{\boldsymbol{a}}_{U,c}(\lambda)$, it is always well defined with high probability, since it can be well defined under the condition that the good event in Assumption 4 happens. Note that this good event only depend on $\boldsymbol{X}, \boldsymbol{\Theta}$ and is independent of $\boldsymbol{\beta}, \boldsymbol{\varepsilon}$.

By Eq. (37) and (38), simple calculation shows that

$$\begin{aligned}\overline{U}_c(\lambda, N, n, d) &\equiv -\langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1} \overline{\boldsymbol{v}} \rangle = -\Psi_1 - \Psi_2 - \Psi_3, \\ \|\overline{\boldsymbol{a}}_{U,c}\|_2^2 &\equiv \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-2} \overline{\boldsymbol{v}} \rangle = \Phi_1 + \Phi_2 + \Phi_3,\end{aligned}$$

where

$$\begin{aligned}\Psi_1 &= \langle \boldsymbol{v}, \overline{\boldsymbol{M}}^{-1} \boldsymbol{v} \rangle, & \Phi_1 &= \langle \boldsymbol{v}, \overline{\boldsymbol{M}}^{-2} \boldsymbol{v} \rangle, \\ \Psi_2 &= -2\psi_2^{-1} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-1} \boldsymbol{v} \rangle, & \Phi_2 &= -2\psi_2^{-1} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-2} \boldsymbol{v} \rangle, \\ \Psi_3 &= \psi_2^{-2} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-1} \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}} \rangle, & \Phi_3 &= \psi_2^{-2} \langle \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}}, \overline{\boldsymbol{M}}^{-2} \frac{\boldsymbol{Z}^\mathsf{T} \boldsymbol{y}}{\sqrt{d}} \rangle.\end{aligned}$$

The following lemma gives the expectation of Ψ_i's and Φ_i's with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Lemma 1 (Expectation of Ψ_i's and Φ_i's)**.** *Denote* $\boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$. *We have*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &= \mu_1^2 F_1^2 \cdot \frac{1}{d}\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &= -\frac{2F_1^2}{\psi_2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}_1^\mathsf{T}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= \frac{F_1^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\Big) + \frac{\tau^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\Big),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &= \mu_1^2 F_1^2 \cdot \frac{1}{d}\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Q}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &= -\frac{2F_1^2}{\psi_2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}_1^\mathsf{T}\Big) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= \frac{F_1^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\Big) + \frac{\tau^2}{\psi_2^2} \cdot \frac{1}{d}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\Big).
\end{aligned}$$

Here the definitions of $\boldsymbol{Q}$, $\boldsymbol{H}$, *and* $\boldsymbol{Z}_1$ *are given by Eq. (19).*

Furthermore, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &= \mu_1^2 F_1^2 \cdot \partial_{s_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &= F_1^2 \cdot \partial_p G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= F_1^2 \cdot (\partial_{t_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) - 1) + \tau^2 \cdot (\partial_{t_1} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) - 1),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &= -\mu_1^2 F_1^2 \cdot \partial_{s_1}\partial_{s_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &= -F_1^2 \cdot \partial_{s_1}\partial_p G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= -F_1^2 \cdot \partial_{s_1}\partial_{t_2} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} G_d(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi})).
\end{aligned}$$

The definition of G_d *is as in Definition 1, and* $\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q})$ *for* $k \in \{1, 2\}$ *stands for the* k*'th derivatives (as a vector or a matrix) of* $G_d(\boldsymbol{i}u; \boldsymbol{q})$ *with respect to* $\boldsymbol{q}$ *in the* $u \to 0+$ *limit (with its elements given by partial derivatives)*

$$\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q}) = \lim_{u \to 0+} \nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q}).$$

We next state the asymptotic characterization of the log-determinant which was proven in (Mei & Montanari, 2019).

Proposition 3 (Proposition 8.4 in (Mei & Montanari, 2019))**.** *Define*

$$\begin{aligned}
\Xi(\xi, z_1, z_2; \boldsymbol{q}; \boldsymbol{\psi}) \equiv\ & \log[(s_2 z_1 + 1)(t_2 z_2 + 1) - \mu_1^2 (1 + p)^2 z_1 z_2] - \mu_\star^2 z_1 z_2\\
& + s_1 z_1 + t_1 z_2 - \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \xi(z_1 + z_2) - \psi_1 - \psi_2.
\end{aligned} \tag{39}$$

For $\xi \in \mathbb{C}_+$ *and* $\boldsymbol{q} \in \mathcal{Q}$ *(c.f. Eq. (25)), let* $m_1(\xi; \boldsymbol{q}; \boldsymbol{\psi}), m_2(\xi; \boldsymbol{q}; \boldsymbol{\psi})$ *be defined as the analytic continuation of solution of Eq. (26) as defined in Definition 3. Define*

$$g(\xi; \boldsymbol{q}; \boldsymbol{\psi}) = \Xi(\xi, m_1(\xi; \boldsymbol{q}; \boldsymbol{\psi}), m_2(\xi; \boldsymbol{q}; \boldsymbol{\psi}); \boldsymbol{q}; \boldsymbol{\psi}). \tag{40}$$

Consider proportional asymptotics $N/d \to \psi_1$, $N/d \to \psi_2$, *as per Assumption 3. Then for any fixed* $\xi \in \mathbb{C}_+$ *and* $\boldsymbol{q} \in \mathcal{Q}$, *we have*

$$\lim_{d \to \infty} \mathbb{E}[|G_d(\xi; \boldsymbol{q}) - g(\xi; \boldsymbol{q}; \boldsymbol{\psi})|] = 0. \tag{41}$$

Moreover, for any fixed $u \in \mathbb{R}_+$ *and* $\boldsymbol{q} \in \mathcal{Q}$, *we have*

$$\lim_{d \to \infty} \mathbb{E}[\|\partial_{\boldsymbol{q}} G_d(\boldsymbol{i}u; \boldsymbol{q}) - \partial_{\boldsymbol{q}} g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})\|_2] = 0, \tag{42}$$

$$\lim_{d \to \infty} \mathbb{E}[\|\nabla_{\boldsymbol{q}}^2 G_d(\boldsymbol{i}u; \boldsymbol{q}) - \nabla_{\boldsymbol{q}}^2 g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})\|_{\mathrm{op}}] = 0. \tag{43}$$

Remark 4. *Note that Proposition 8.4 in (Mei & Montanari, 2019) stated that the Eq. (42) and (43) holds at* $\boldsymbol{q} = \boldsymbol{0}$. *However, by a simple modification of their proof, one can show that these equations also holds at any* $\boldsymbol{q} \in \mathcal{Q}$.

Combining Assumption 5 with Proposition 3, we have

Proposition 4. *Let Assumption 5 holds. For any $\lambda \in \Lambda_U$, denote $\boldsymbol{q}_U = \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$, then we have, for $k = 1, 2$,*

$$\|\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q}_U) - \lim_{u \to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}_U; \boldsymbol{\psi})\| = o_{d,\mathbb{P}}(1).$$

As a consequence of Proposition 4, we can calculate the asymptotics of Ψ_i's and Φ_i's. Combined with the concentration result in Lemma 2 latter in the section, the proposition below completes the proof of the part of Proposition 2 regarding the standard uniform convergence U. Its correctness follows directly from Lemma 1 and Proposition 4.

Proposition 5. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_U$, denote $\boldsymbol{q}_U(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$, then we have*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] \xrightarrow{\mathbb{P}} \mu_1^2 F_1^2 \cdot \partial_{s_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] \xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_p g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] \xrightarrow{\mathbb{P}} F_1^2 \cdot \Big(\partial_{t_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - 1\Big) + \tau^2 \Big(\partial_{t_1} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - 1\Big),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] \xrightarrow{\mathbb{P}} -\mu_1^2 F_1^2 \cdot \partial_{s1}\partial_{s2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] \xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s1}\partial_p g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
&\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] \xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s_1}\partial_{t_2} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} g(0_+; \boldsymbol{q}_U(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),
\end{aligned}$$

where $\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi})$ for $k \in \{1, 2\}$ stands for the k'th derivatives (as a vector or a matrix) of $g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi}) = \lim_{u \to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}).$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{U}_c(\lambda, N, n, d)] \xrightarrow{\mathbb{P}} \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2), \quad \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2] \xrightarrow{\mathbb{P}} \mathcal{A}_U(\lambda, \psi_1, \psi_2),$$

where the definitions of $\overline{\mathcal{U}}$ and $\mathcal{A}_U$ are given in Definition 5. Here $\xrightarrow{\mathbb{P}}$ stands for convergence in probability as $N/d \to \psi_1$ and $n/d \to \psi_2$ (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$).

Lemma 2. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_U$, we have*

$$\begin{aligned}
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= o_{d,\mathbb{P}}(1),\\
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= o_{d,\mathbb{P}}(1),
\end{aligned}$$

so that

$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{U}_c(\lambda, N, n, d)], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2] = o_{d,\mathbb{P}}(1).$$

Here, $o_{d,\mathbb{P}}(1)$ stands for converges to 0 in probability (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$) as $N/d \to \psi_1$ and $n/d \to \psi_2$ and $d \to \infty$.

Now, combining Lemma 2 and Proposition 5, we have

$$\overline{U}_c(\lambda, N, n, d) \xrightarrow{\mathbb{P}} \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2), \quad \psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2 \xrightarrow{\mathbb{P}} \mathcal{A}_U(\lambda, \psi_1, \psi_2),$$

Finally, combining with the arguments in Appendix D.2 proves the asymptotics of $\overline{U}$ and $\psi_1 \|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2$.

D.4. The asymptotics of $\overline{T}_c$ and $\psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2$

In the following, we derive the asymptotics of $\overline{T}_c(\lambda, N, n, d)$ and $\psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2$. This follows the same steps as the proof of the asymptotics of $\overline{U}_c$ and $\psi_1 \|\overline{\boldsymbol{a}}_{U,c}(\lambda)\|_2^2$. We will give an overview of its proof. The detailed proof is the same as that of $\overline{U}_c$, and we will not include them for brevity.

For a fixed $\lambda \in \Lambda_T$, recalling that the definition of $\overline{T}_c$ as in Eq. (36), we have

$$\begin{aligned}\overline{T}_c(\lambda, N, n, d) &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \left[R_c(\boldsymbol{a}) - \lambda\psi_1 \|\boldsymbol{a}\|_2^2 + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\,\rangle \right] \\ &= \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \left(\langle \boldsymbol{a}, (\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a}\rangle - 2\langle \boldsymbol{a}, \boldsymbol{v}\rangle + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a}\rangle - 2\langle \boldsymbol{\mu}, \boldsymbol{y}/\sqrt{d}\rangle \right) + \mathbb{E}[y^2] \\ &= \sup_{\sqrt{d}\boldsymbol{Z}\boldsymbol{a}=y} \langle \boldsymbol{a}, (\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a}\rangle - 2\langle \boldsymbol{a}, \boldsymbol{v}\rangle + \mathbb{E}[y^2]\end{aligned} \tag{44}$$

Whenever the good event in Assumption 4 happens, $(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)$ is negative definite in null$(\boldsymbol{Z})$. The optimum of the above variational equation exists. By KKT condition, the optimal $\boldsymbol{a}$ and dual variable $\boldsymbol{\mu}$ satisfies

- Stationary condition: $(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a} + \boldsymbol{Z}^\mathsf{T}\mu = \boldsymbol{v}$.
- Primal Feasible: $\boldsymbol{Z}\boldsymbol{a} = \boldsymbol{y}/\sqrt{d}$.

The two conditions can be written compactly as

$$\begin{bmatrix} \boldsymbol{U}_c - \psi_1\lambda \mathbf{I}_N & \boldsymbol{Z}^\mathsf{T} \\ \boldsymbol{Z} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{y}/\sqrt{d} \end{bmatrix}. \tag{45}$$

We define

$$\overline{\boldsymbol{M}} \equiv \begin{bmatrix} \boldsymbol{U}_c - \psi_1\lambda \mathbf{I}_N & \boldsymbol{Z}^\mathsf{T} \\ \boldsymbol{Z} & \mathbf{0} \end{bmatrix}, \qquad \overline{\boldsymbol{v}} \equiv \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{y}/\sqrt{d} \end{bmatrix}. \tag{46}$$

Under Assumption 4, $\overline{\boldsymbol{M}}$ is invertible. To see this, suppose there exists vector $[\boldsymbol{a}_1^\mathsf{T}, \boldsymbol{\mu}_1^\mathsf{T}]^\mathsf{T} \neq \mathbf{0} \in \mathbb{R}^{N+n}$ such that $\overline{\boldsymbol{M}}[\boldsymbol{a}_1^\mathsf{T}, \boldsymbol{\mu}_1^\mathsf{T}]^\mathsf{T} = \mathbf{0}$, then

$$\begin{aligned}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\boldsymbol{a}_1 + \boldsymbol{Z}^\mathsf{T}\boldsymbol{\mu}_1 &= 0, \\ \boldsymbol{Z}\boldsymbol{a}_1 &= 0.\end{aligned}$$

As in Assumption 4, let $\mathsf{P}_{\text{null}} = \mathbf{I}_N - \boldsymbol{Z}^\dagger \boldsymbol{Z}$. We write $\boldsymbol{a}_1 = \mathsf{P}_{\text{null}}\boldsymbol{v}_1$ for some $\boldsymbol{v}_1 \neq \mathbf{0} \in \mathbb{R}^N$. Then,

$$\begin{aligned}&(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}\boldsymbol{v}_1 + \boldsymbol{Z}^\mathsf{T}\boldsymbol{\mu}_1 = 0, \\ \Rightarrow\ &\mathsf{P}_{\text{null}}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}\boldsymbol{v}_1 + \mathsf{P}_{\text{null}}\boldsymbol{Z}^\mathsf{T}\boldsymbol{\mu}_1 = 0, \\ \Rightarrow\ &\mathsf{P}_{\text{null}}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}\boldsymbol{v}_1 = 0,\end{aligned}$$

where the last relation come from the fact that $\boldsymbol{Z}\mathsf{P}_{\text{null}} = \mathbf{0}$. However by Assumption 4, $\mathsf{P}_{\text{null}}(\boldsymbol{U}_c - \lambda\psi_1 \boldsymbol{I}_N)\mathsf{P}_{\text{null}}$ is negative definite, which leads to a contradiction.

In the following, we assume the event in Assumption 4 happens so that $\overline{\boldsymbol{M}}$ is invertible. In this case, the maximizer in Eq. (44) can be well defined as

$$\overline{\boldsymbol{a}}_{T,c}(\lambda) = [\mathbf{I}_N, \mathbf{0}_{N\times n}]\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}.$$

Moreover, we can write $\overline{T}_c$ as

$$\overline{T}_c(\lambda, N, n, d) = \mathbb{E}[y^2] - \overline{\boldsymbol{v}}^\mathsf{T}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}.$$

We further define

$$\overline{\boldsymbol{v}}_1 = [\boldsymbol{v}^\mathsf{T}, \mathbf{0}_{n\times 1}^\mathsf{T}]^\mathsf{T}, \quad \overline{\boldsymbol{v}}_2 = [\mathbf{0}_{N\times 1}^\mathsf{T}, \boldsymbol{y}^\mathsf{T}/\sqrt{d}]^\mathsf{T}, \quad \boldsymbol{E} \equiv \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N\times n} \\ \mathbf{0}_{n\times N} & \mathbf{0}_{n\times n} \end{bmatrix}.$$

Simple calculation shows that

$$\begin{aligned}\overline{T}_c(\lambda, N, n, d) &\equiv \mathbb{E}[y^2] - \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}\rangle = F_1^2 + \tau^2 - \Psi_1 - \Psi_2 - \Psi_3, \\ \|\overline{\boldsymbol{a}}_{U,c}\|_2^2 &\equiv \langle \overline{\boldsymbol{v}}, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}\rangle = \Phi_1 + \Phi_2 + \Phi_3,\end{aligned}$$

where

$$\begin{aligned}\Psi_1 &= \langle \overline{\boldsymbol{v}}_1, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, & \Phi_1 &= \langle \overline{\boldsymbol{v}}_1, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, \\ \Psi_2 &= 2\langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, & \Phi_2 &= 2\langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_1\rangle, \\ \Psi_3 &= \langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_2\rangle, & \Phi_3 &= \langle \overline{\boldsymbol{v}}_2, \overline{\boldsymbol{M}}^{-1}\boldsymbol{E}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{v}}_2\rangle.\end{aligned}$$

The following lemma gives the expectation of Ψ_i's and Φ_i's with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Lemma 3 (Expectation of Ψ_i's and Φ_i's)**.** *Denote* $\boldsymbol{q}_T(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$. *We have*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &= \mu_1^2 F_1^2 \cdot \partial_{s_2} G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &= F_1^2 \cdot \partial_p G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= F_1^2 \cdot \partial_{t_2} G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})) + \tau^2 \cdot \partial_{t_1} G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &= -\mu_1^2 F_1^2 \cdot \partial_{s_1}\partial_{s_2} G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &= -F_1^2 \cdot \partial_{s_1}\partial_p G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})) \times (1 + o_d(1)),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= -F_1^2 \cdot \partial_{s_1}\partial_{t_2} G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} G_d(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi})).
\end{aligned}$$

The definition of G_d is as in Definition 1, and $\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q})$ for $k \in \{1, 2\}$ stands for the k'th derivatives (as a vector or a matrix) of $G_d(\boldsymbol{i}u; \boldsymbol{q})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q}) = \lim_{u \to 0+} \nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q}).$$

The proof of Lemma 3 follows from direct calculation and is identical to the proof of Lemma 1. Combining Assumption 5 with Proposition 3, we have

Proposition 6. *Let Assumption 5 holds. For any $\lambda \in \Lambda_T$, denote $\boldsymbol{q}_T = \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$, then we have, for $k = 1, 2$,*

$$\|\nabla_{\boldsymbol{q}}^k G_d(0_+; \boldsymbol{q}_T) - \lim_{u \to 0+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}_T; \boldsymbol{\psi})\| = o_{d,\mathbb{P}}(1).$$

As a consequence of Proposition 6, we can calculate the asymptotics of Ψ_i's and Φ_i's.

Proposition 7. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_T$, denote $\boldsymbol{q}_T(\lambda, \boldsymbol{\psi}) = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)$, then we have*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1] &\xrightarrow{\mathbb{P}} \mu_1^2 F_1^2 \cdot \partial_{s_2} g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2] &\xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_p g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &\xrightarrow{\mathbb{P}} F_1^2 \cdot \partial_{t_2} g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) + \tau^2 \cdot \partial_{t_1} g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1] &\xrightarrow{\mathbb{P}} -\mu_1^2 F_1^2 \cdot \partial_{s1}\partial_{s2} g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2] &\xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s1}\partial_p g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),\\
\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &\xrightarrow{\mathbb{P}} -F_1^2 \cdot \partial_{s_1}\partial_{t_2} g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}) - \tau^2 \cdot \partial_{s_1}\partial_{t_1} g(0_+; \boldsymbol{q}_T(\lambda, \boldsymbol{\psi}); \boldsymbol{\psi}),
\end{aligned}$$

where $\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi})$ for $k \in \{1, 2\}$ stands for the k'th derivatives (as a vector or a matrix) of $g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi})$ with respect to $\boldsymbol{q}$ in the $u \to 0+$ limit (with its elements given by partial derivatives)

$$\nabla_{\boldsymbol{q}}^k g(0_+; \boldsymbol{q}; \boldsymbol{\psi}) = \lim_{u \to 0_+} \nabla_{\boldsymbol{q}}^k g(\boldsymbol{i}u; \boldsymbol{q}; \boldsymbol{\psi}).$$

As a consequence, we have

$$\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{T}_c(\lambda, N, n, d)] \xrightarrow{\mathbb{P}} \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2), \quad \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2] \xrightarrow{\mathbb{P}} \mathcal{A}_T(\lambda, \psi_1, \psi_2),$$

where the definitions of $\overline{\mathcal{T}}$ and $\mathcal{A}_T$ are given in Definition 5. Here $\xrightarrow{\mathbb{P}}$ stands for convergence in probability as $N/d \to \psi_1$ and $n/d \to \psi_2$ (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$).

The Proposition above suggests that Ψ_i and Φ_i concentrates with respect to the randomness in $\boldsymbol{X}$ and $\boldsymbol{\Theta}$. To complete the concentration proof, we need to show that Ψ_i and Φ_i concentrates with respect to the randomness in $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Lemma 4. *Follow the assumptions of Proposition 2. For any $\lambda \in \Lambda_T$, we have*

$$\begin{aligned}
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Psi_3] &= o_{d,\mathbb{P}}(1),\\
\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_1], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_2], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\Phi_3] &= o_{d,\mathbb{P}}(1),
\end{aligned}$$

so that

$$\mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\overline{T}_c(\lambda, N, n, d)], \mathit{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}[\|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2] = o_{d,\mathbb{P}}(1).$$

Here, $o_{d,\mathbb{P}}(1)$ stands for converges to 0 in probability (with respect to the randomness of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$) as $N/d \to \psi_1$ and $n/d \to \psi_2$ and $d \to \infty$.

Now, combining Proposition 7 and 4, we have

$$\overline{T}_c(\lambda, N, n, d) \xrightarrow{\mathbb{P}} \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2), \quad \psi_1 \|\overline{\boldsymbol{a}}_{T,c}(\lambda)\|_2^2 \xrightarrow{\mathbb{P}} \mathcal{A}_T(\lambda, \psi_1, \psi_2).$$

The results above combined with the arguments in Appendix D.2 completes the proof for the asymptotics of $\overline{T}$ and $\psi_1 \|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2$.

D.5. Proof of Lemma 1 and Lemma 2

Proof of Lemma 1. Note that by Assumption 4, the matrix $\overline{\boldsymbol{M}} = \boldsymbol{U}_c - \psi_2^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} - \psi_1\lambda\mathbf{I}_N$ is negative definite (so that it is invertible) with high probability. Moreover, whenever $\overline{\boldsymbol{M}}$ is negative definite, the matrix $\boldsymbol{A}(\boldsymbol{q}_U)$ for $\boldsymbol{q}_U = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$ is also invertible. In the following, we condition on this good event happens.

From the expansion for $\boldsymbol{v}_i$ in (34), we have

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Psi_1 = \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Big[\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\Big)\Big] = \frac{1}{d}\lambda_{d,1}(\sigma)^2 F_1^2 \cdot \Big[\mathrm{Tr}\Big(\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}\Big)\Big] = \frac{1}{d}\mu_1^2 F_1^2 \mathrm{Tr}\left(\overline{\boldsymbol{M}}^{-1}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}}{d}\right) \times (1 + o_d(1)),$$

where we used the relation $\lambda_{d,1} = \mu_1/\sqrt{d} \times (1 + o_d(1))$ as in Eq. (66). Similarly, the second term is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Psi_2 =& -\frac{2}{\psi_2\sqrt{d}}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Big[\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{v}\boldsymbol{y}^\mathsf{T}\Big)\Big] \\
=& -\frac{2}{\psi_2 d\sqrt{d}}\lambda_{d,1}(\sigma)F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T}\Big) \\
=& -\frac{2}{\psi_2 d^2}\mu_1 F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T}\Big) \times (1 + o_d(1)).
\end{aligned}$$

To compute Ψ_3, note we have

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}[\boldsymbol{y}\boldsymbol{y}^\mathsf{T}] = F_1^2 \cdot (\boldsymbol{X}\boldsymbol{X}^\mathsf{T})/d + \tau^2\mathbf{I}_n.$$

This gives the expansion for Ψ_3

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Psi_3 &= \psi_2^{-2}d^{-1}\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}\boldsymbol{y}^\mathsf{T}\Big) \\
&= \psi_2^{-2}d^{-2}F_1^2\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big) + \psi_2^{-2}d^{-1}\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}\Big)\tau^2.
\end{aligned}$$

Through the same algebraic manipulation above, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Phi_1 &= \frac{1}{d}\mu_1^2 F_1^2 \mathrm{Tr}\left(\overline{\boldsymbol{M}}^{-2}\frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}}{d}\right) \times (1 + o_d(1)), \\
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Phi_2 &= -\frac{2}{\psi_2 d^2}\mu_1 F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T}\Big) \times (1 + o_d(1)), \\
\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\varepsilon}}\Phi_3 &= \psi_2^{-2}d^{-2}F_1^2 \cdot \mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\Big) + \psi_2^{-2}d^{-1}\tau^2\mathrm{Tr}\Big(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\Big).
\end{aligned}$$

Next, we express the trace of matrices products as the derivative of the function $G_d(\xi, \boldsymbol{q})$ (c.f. Definition 1). The derivatives of G_d are (which can we well-defined at $\boldsymbol{q} = \boldsymbol{q}_U = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$ with high probability by Assumption 4)

$$\partial_{q_i}G_d(0, \boldsymbol{q}) = \frac{1}{d}\mathrm{Tr}(\boldsymbol{A}(\boldsymbol{q})^{-1}\partial_i\boldsymbol{A}(\boldsymbol{q})), \quad \partial_{q_i}\partial_{q_j}G_d(0, \boldsymbol{q}) = -\frac{1}{d}\mathrm{Tr}(\boldsymbol{A}(\boldsymbol{q})^{-1}\partial_{q_i}\boldsymbol{A}(\boldsymbol{q})\boldsymbol{A}(\boldsymbol{q})^{-1}\partial_{q_j}\boldsymbol{A}(\boldsymbol{q})). \tag{47}$$

As an example, we consider evaluating $\partial_{s_2} G_d(0, \boldsymbol{q})$ at $\boldsymbol{q} = \boldsymbol{q}_U \equiv (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)$. Using the formula for block matrix inversion, we have

$$\boldsymbol{A}(\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)^{-1} = \begin{bmatrix} (\mu_\star^2 - \lambda\psi_1)\mathbf{I}_N + \mu_1^2 \boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T} \\ \boldsymbol{Z} & \psi_2 \mathbf{I}_n \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{U}_c - \psi_2^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} - \psi_1\lambda\mathbf{I}_N)^{-1} & \cdots \\ \cdots & \cdots \end{bmatrix}.$$

Then we have

$$\partial_{s_2} G_d(0, \boldsymbol{q}_U) = \frac{1}{d}\mathrm{Tr}\left(\begin{bmatrix} \overline{\boldsymbol{M}}^{-1} & \cdots \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} \boldsymbol{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) = \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q})/d.$$

Applying similar argument to compute other derivatives, we get

1. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T})/d^2 = \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q})/d = \partial_{s_2} G_d(0, \boldsymbol{q}_U)$.
2. $\mu_1 \cdot \mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{X}^\mathsf{T})/d^2 = \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}_1^\mathsf{T}\boldsymbol{Z})/d = -\psi_2 \partial_p G_d(0, \boldsymbol{q}_U)/2$.
3. $\mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}\boldsymbol{X}^\mathsf{T})/d^2 = \mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H})/d = \psi_2^2 \partial_{t_2} G_d(0, \boldsymbol{q}_U) - \psi_2^2$.
4. $\mathrm{Tr}(\boldsymbol{Z}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Z}^\mathsf{T})/d = \psi_2^2 \partial_{t_1} G_d(0, \boldsymbol{q}_U) - \psi_2^2$.
5. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Q})/d = -\partial_{s_1}\partial_{s_2} G_d(0, \boldsymbol{q}_U)$.
6. $(2/d\psi_2) \cdot \mathrm{Tr}(\boldsymbol{Z}_1^\mathsf{T}\boldsymbol{Z}\overline{\boldsymbol{M}}^{-2}) = \partial_{s_1}\partial_p G_d(0, \boldsymbol{q}_U)$.
7. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z})/(d\psi_2^2) = -\partial_{s_1}\partial_{t_2} G_d(0, \boldsymbol{q}_U)$.
8. $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})/(d\psi_2^2) = -\partial_{s_1}\partial_{t_1} G_d(0, \boldsymbol{q}_U)$.

Combining these equations concludes the proof. □

Proof of Lemma 2. We prove this lemma by assuming that $\boldsymbol{\beta}$ follows a different distribution: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, (\|F_1\|_2^2/d)\mathbf{I}_d)$. The case when $\boldsymbol{\beta} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(F_1))$ can be treated similarly.

By directly calculating the variance, we can show that, there exists scalers $(c_{ik}^{(d)})_{k\in[K_i]}$ with $c_{ik}^{(d)} = \Theta_d(1)$, and matrices $(\boldsymbol{A}_{ik}, \boldsymbol{B}_{ik})_{k\in[K_i]} \subseteq \{\mathbf{I}_N, \boldsymbol{Q}, \boldsymbol{Z}^\mathsf{T}\boldsymbol{H}\boldsymbol{Z}, \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\}$, such that the variance of Ψ_i's can be expressed in form

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Psi_i) = \frac{1}{d}\sum_{k=1}^{K_i} c_{ik}^{(d)} \mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{A}_{ik}\overline{\boldsymbol{M}}^{-1}\boldsymbol{B}_{ik})/d.$$

For example, by Lemma 8, we have

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\beta}\sim\mathcal{N}(\mathbf{0},(F_1^2/d)\mathbf{I}_d)}(\Psi_1) &= \lambda_{d,1}(\sigma)^4 \mathrm{Var}_{\boldsymbol{\beta}\sim\mathcal{N}(\mathbf{0},(F_1^2/d)\mathbf{I}_d))}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Theta}^\mathsf{T}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\boldsymbol{\beta}) = 2\lambda_{d,1}(\sigma)^4 F_1^4 \|\boldsymbol{\Theta}^\mathsf{T}\overline{\boldsymbol{M}}^{-1}\boldsymbol{\Theta}\|_F^2/d^2 \\ &= c_1^{(d)}\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q}\overline{\boldsymbol{M}}^{-1}\boldsymbol{Q})/d^2,\end{aligned}$$

where $c_1^{(d)} = 2d^2\lambda_{d,1}(\sigma)^4 F_1^4 = O_d(1)$. The variance of Ψ_2 and Ψ_3 can be calculated similarly.

Note that each $\mathrm{Tr}(\overline{\boldsymbol{M}}^{-1}\boldsymbol{A}_{ik}\overline{\boldsymbol{M}}^{-1}\boldsymbol{B}_{ik})/d$ can be expressed as an entry of $\nabla_{\boldsymbol{q}}^2 G_d(0; \boldsymbol{q})$ (c.f. Eq. (47)), and by Proposition 4, they are of order $O_{d,\mathbb{P}}(1)$. This gives

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Psi_i) = o_{d,\mathbb{P}}(1).$$

Similarly, for the same set of scalers $(c_{ik}^{(d)})_{k\in[K_i]}$ and matrices $(\boldsymbol{A}_{ik}, \boldsymbol{B}_{ik})_{k\in[K_i]}$, we have

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Phi_i) = \frac{1}{d}\sum_{k=1}^{K_i} c_{ik}\mathrm{Tr}(\overline{\boldsymbol{M}}^{-2}\boldsymbol{A}_{ik}\overline{\boldsymbol{M}}^{-2}\boldsymbol{B}_{ik})/d.$$

Note that for two semidefinite matrices $\boldsymbol{A}, \boldsymbol{B}$, we have $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) \le \|\boldsymbol{A}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{B})$. Moreover, note we have $\|\overline{\boldsymbol{M}}\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1)$ (by Assumption 4). This gives

$$\mathrm{Var}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}(\Phi_i) = o_{d,\mathbb{P}}(1).$$

This concludes the proof. □

D.6. Auxiliary Lemmas

The following lemma (Lemma 5) is a reformulation of Proposition 3 in (Ghorbani et al., 2019). We present it in a stronger form, but it can be easily derived from the proof of Proposition 3 in (Ghorbani et al., 2019). This lemma was first proved in (El Karoui, 2010) in the Gaussian case. (Notice that the second estimate —on $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})$— follows by applying the first one whereby $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{W} = [\boldsymbol{\Theta}^{\mathsf{T}}|\boldsymbol{X}^{\mathsf{T}}]^{\mathsf{T}}$

Lemma 5. *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^{\mathsf{T}} \in \mathbb{R}^{n\times d}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Assume $1/c \le n/d, N/d \le c$ for some constant $c \in (0,\infty)$. Then*

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}) - \mathbf{I}_N\|_{\mathrm{op}}^2\Big] = o_d(1)\,, \tag{48}$$

$$\mathbb{E}\Big[\sup_{k\ge 2} \|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}}^2\Big] = o_d(1). \tag{49}$$

Notice that the second estimate —on $Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})$— follows by applying the first one —Eq. (48)— whereby $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{W} = [\boldsymbol{\Theta}^{\mathsf{T}}|\boldsymbol{X}^{\mathsf{T}}]^{\mathsf{T}}$, and we use $\|Q_k(\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}})\|_{\mathrm{op}} \le \|Q_k(\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}) - \mathbf{I}_{N+n}\|_{\mathrm{op}}$.

The following lemma (Lemma 6) can be easily derived from Lemma 5. Again, this lemma was first proved in (El Karoui, 2010) in the Gaussian case.

Lemma 6. *Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)^{\mathsf{T}} \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_a)_{a\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Let activation function σ satisfies Assumption 2. Assume $1/c \le N/d \le c$ for some constant $c \in (0,\infty)$. Denote*

$$\boldsymbol{U} = \Big(\mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(\langle \boldsymbol{\theta}_a, \boldsymbol{x}\rangle/\sqrt{d})\sigma(\langle \boldsymbol{\theta}_b, \boldsymbol{x}\rangle/\sqrt{d})]\Big)_{a,b\in[N]} \in \mathbb{R}^{N\times N}.$$

Then we can rewrite the matrix $\boldsymbol{U}$ to be

$$\boldsymbol{U} = \lambda_{d,0}(\sigma)^2 \mathbf{1}_N \mathbf{1}_N^{\mathsf{T}} + \mu_1^2 \boldsymbol{Q} + \mu_\star^2(\mathbf{I}_N + \boldsymbol{\Delta}),$$

with $\boldsymbol{Q} = \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}/d$ and $\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathrm{op}}^2] = o_d(1)$.

In the following, we show that, under sufficient regularity condition of σ, we have $\lambda_{d,0}(\sigma) = O(1/d)$.

Lemma 7. *Let $\sigma \in C^2(\mathbb{R})$ with $|\sigma'(x)|, |\sigma''(x)| < c_0 e^{c_1|x|}$ for some $c_0, c_1 \in \mathbb{R}$. Assume that $\mathbb{E}_{G\sim\mathcal{N}(0,1)}[\sigma(G)] = 0$. Then we have*

$$\lambda_{d,0}(\sigma) \equiv \mathbb{E}_{\boldsymbol{x}\sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))}[\sigma(x_1)] = O(1/d).$$

Proof of Lemma 7. Let $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\gamma \sim \chi(d)/\sqrt{d}$ independently. Then we have $\gamma\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, so that by the assumption, we have $\mathbb{E}[\sigma(\gamma x_1)] = 0$.

As a consequence, by the second order Taylor expansion, and by the independence of γ and $\boldsymbol{x}$, we have (for $\xi(x_1) \in [\gamma, 1]$)

$$\begin{aligned}|\lambda_{d,0}(\sigma)| = |\mathbb{E}[\sigma(x_1)]| \le |\mathbb{E}[\sigma(x_1)] - \mathbb{E}[\sigma(\gamma x_1)]| &\le \Big|\mathbb{E}[\sigma'(x_1)x_1]\mathbb{E}[\gamma - 1]\Big| + \Big|(1/2)\mathbb{E}[\sigma''(\xi(x_1)x_1)(\gamma-1)^2]\Big| \\ &\le \Big|\mathbb{E}[\sigma'(x_1)x_1]\Big| \cdot \Big|\mathbb{E}[\gamma-1]\Big| + (1/2)\mathbb{E}\Big[\sup_{u\in[\gamma,1]} \sigma''(ux_1)^2\Big]^{1/2} \mathbb{E}[(\gamma-1)^4]^{1/2}.\end{aligned}$$

By the assumption that $|\sigma'(x)|, |\sigma''(x)| < c_0 e^{c_1|x|}$ for some $c_0, c_1 \in \mathbb{R}$, there exists constant K that only depends on c_0 and c_1 such that

$$\sup_d \Big|\mathbb{E}[\sigma'(x_1)x_1]\Big| \le K, \qquad \sup_d \Big|(1/2)\mathbb{E}\Big[\sup_{u\in[\gamma,1]} \sigma''(ux_1)^2\Big]^{1/2}\Big| \le K.$$

Moreover, by property of the χ distribution, we have

$$|\mathbb{E}[\gamma - 1]| = O(d^{-1}), \qquad \mathbb{E}[(\gamma-1)^4]^{1/2} = O(d^{-1}).$$

This concludes the proof. □

The following lemma is a simple variance calculation and can be found as Lemma C.5 in (Mei & Montanari, 2019). We restate here for completeness.

Lemma 8. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times N}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$. Let $\boldsymbol{g} = (g_1, \dots, g_n)^\mathsf{T}$ with $g_i \sim_{iid} \mathbb{P}_g$, $\mathbb{E}_g[g] = 0$, and $\mathbb{E}_g[g^2] = 1$. Let $\boldsymbol{h} = (h_1, \dots, h_N)^\mathsf{T}$ with $h_i \sim_{iid} \mathbb{P}_h$, $\mathbb{E}_h[h] = 0$, and $\mathbb{E}_h[h^2] = 1$. Further we assume that $\boldsymbol{h}$ is independent of $\boldsymbol{g}$. Then we have*

$$\mathit{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{A}\boldsymbol{h}) = \|\boldsymbol{A}\|_F^2,$$
$$\mathit{Var}(\boldsymbol{g}^\mathsf{T}\boldsymbol{B}\boldsymbol{g}) = \sum_{i=1}^n B_{ii}^2(\mathbb{E}[g^4] - 3) + \|\boldsymbol{B}\|_F^2 + \mathrm{Tr}(\boldsymbol{B}^2).$$

E. Proof of Theorem 1

Here we give the whole proof for U. The proof for T is the same.

For fixed $A^2 \in \Gamma_U \equiv \{\mathcal{A}_U(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_U\}$, we denote

$$\lambda_\star(A^2) = \inf_\lambda \Big\{\lambda : \mathcal{A}_U(\lambda, \psi_1, \psi_2) = A^2\Big\}.$$

By the definition of Γ_U, the set $\{\lambda : \mathcal{A}_U(\lambda, \psi_1, \psi_2) = A^2\}$ is non-empty and lower bounded, so that $\lambda_\star(A^2)$ can be well-defined. Moreover, we have $\lambda_\star(A^2) \in \Lambda_U$. It is also easy to see that we have

$$\lambda_\star(A^2) \in \arg\min_{\lambda\geq 0} \Big[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A^2\Big]. \tag{50}$$

E.1. Upper bound

Note we have

$$\begin{aligned}
U(A, N, n, d) &= \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \leq A^2} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a})\Big) \\
&\leq \inf_\lambda \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \leq A^2} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda(\|\boldsymbol{a}\|_2^2 - \psi_1^{-1}A^2)\Big) \\
&\leq \inf_\lambda \Big[\overline{U}(\lambda, N, n, d) + \lambda A^2\Big] \\
&\leq \overline{U}(\lambda_\star(A^2), N, n, d) + \lambda_\star(A^2)A^2.
\end{aligned}$$

Note that $\lambda_\star(A^2) \in \Lambda_U$, so by Lemma 5, in the limit of Assumption 3, we have

$$U(A, N, n, d) \leq \overline{\mathcal{U}}(\lambda_\star(A^2), \psi_1, \psi_2) + \lambda_\star(A^2)A^2 + o_{d,\mathbb{P}}(1) = \mathcal{U}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1),$$

where the last equality is by Eq. (50). This proves the upper bound.

E.2. Lower bound

For any $A^2 > 0$, we define a random variable $\hat{\lambda}(A^2)$ (which depend on $\boldsymbol{X}$, $\boldsymbol{\Theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$) by

$$\hat{\lambda}(A^2) = \inf\Big\{\lambda : \lambda \in \arg\min_{\lambda\geq 0} \Big[\overline{U}(\lambda, N, n, d) + \lambda A^2\Big]\Big\}.$$

By Proposition 1, the set is should always be non-empty, so that $\hat{\lambda}(A^2)$ can always be well-defined.

Moreover, since $\lambda_\star(A^2) \in \Lambda_U$, by Assumption 4, as we have shown in the proof in Proposition 2, we can uniquely define $\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))$ with high probability, where

$$\overline{\boldsymbol{a}}_U(\lambda_\star(A^2)) = \arg\max_{\boldsymbol{a}} \Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1\lambda_\star(A^2)\|\boldsymbol{a}\|_2^2\Big].$$

As a consequence, for a small $\varepsilon > 0$, the following event $\mathcal{E}_{\varepsilon,d}$ can be well-defined with high probability

$$\begin{aligned}
\mathcal{E}_{\varepsilon,d} &= \Big\{\psi_1\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 \geq A^2 - \varepsilon\Big\} \cap \Big\{\hat{\lambda}(A^2 + \varepsilon) \leq \lambda_\star(A^2)\Big\} \\
&= \Big\{A^2 - \varepsilon \leq \psi_1\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 \leq A^2 + \varepsilon\Big\}.
\end{aligned}$$

Now, by Proposition 2, in the limit of Assumption 3, we have

$$\lim_{d\to\infty} \mathbb{P}_{\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\beta},\boldsymbol{\varepsilon}}(\mathcal{E}_{\varepsilon,d}) = 1, \tag{51}$$

and we have

$$\mathcal{U}(\lambda_\star(A^2), \psi_1, \psi_2) = \overline{\mathcal{U}}(\lambda_\star(A^2), \psi_1, \psi_2) + o_{d,\mathbb{P}}(1). \tag{52}$$

By the strong duality as in Proposition 1, for any $A^2 \in \Gamma_U$, we have

$$U(A, N, n, d) = \overline{U}(\hat{\lambda}(A^2), N, n, d) + \hat{\lambda}(A^2)A^2.$$

Consequently, for small $\varepsilon > 0$, when the event $\mathcal{E}_{\varepsilon,d}$ happens, we have

$$\begin{aligned}
&U((A^2+\varepsilon)^{1/2}, N, n, d) \\
&= \sup_{\boldsymbol{a}} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1 \hat{\lambda}(A^2+\varepsilon) \cdot \big(\|\boldsymbol{a}\|_2^2 - \psi_1^{-1}(A^2+\varepsilon) \big) \Big) \\
&\geq R(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \widehat{R}_n(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \psi_1 \hat{\lambda}(A^2+\varepsilon) \cdot \big(\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 - \psi_1^{-1}(A^2+\varepsilon) \big) \\
&\geq R(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \widehat{R}_n(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \psi_1 \hat{\lambda}(A^2+\varepsilon) \cdot \big(\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 - \psi_1^{-1}(A^2-\varepsilon) \big) \\
&\geq R(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \widehat{R}_n(\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))) - \psi_1 \lambda_\star(A^2) \cdot \big(\|\overline{\boldsymbol{a}}_U(\lambda_\star(A^2))\|_2^2 - \psi_1^{-1}(A^2-\varepsilon) \big) \\
&= \overline{U}(\lambda_\star(A^2), N, n, d) + \lambda_\star(A^2) \cdot (A^2 - \varepsilon).
\end{aligned}$$

As a consequence, by Eq. (51) and (52), we have

$$U((A^2+\varepsilon)^{1/2}, N, n, d) \geq \overline{\mathcal{U}}(\lambda_\star(A^2), \psi_1, \psi_2) + \lambda_\star(A^2) \cdot (A^2 - \varepsilon) - o_{d,\mathbb{P}}(1) = \mathcal{U}(A, \psi_1, \psi_2) - \varepsilon\lambda_\star(A^2) - o_{d,\mathbb{P}}(1).$$

where the last equality is by the definition of $\mathcal{U}$ as in Definition 2, and by the fact that $\lambda_\star(A^2) \in \arg\min_{\lambda\geq 0}[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A^2]$. Taking ε sufficiently small proves the lower bound. This concludes the proof of Theorem 1.

F. Technical background

In this section we introduce additional technical background useful for the proofs. In particular, we will use decompositions in (hyper-)spherical harmonics on the $\mathbb{S}^{d-1}(\sqrt{d})$ and in Hermite polynomials on the real line. We refer the readers to (Efthimiou & Frye, 2014; Szego, Gabor, 1939; Chihara, 2011; Ghorbani et al., 2019; Mei & Montanari, 2019) for further information on these topics.

F.1. Functional spaces over the sphere

For $d \geq 1$, we let $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ denote the sphere with radius r in $\mathbb{R}^d$. We will mostly work with the sphere of radius $\sqrt{d}$, $\mathbb{S}^{d-1}(\sqrt{d})$ and will denote by γ_d the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in the following are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d)$, with scalar product and norm denoted as $\langle\,\cdot\,,\,\cdot\,\rangle_{L^2}$ and $\|\,\cdot\,\|_{L^2}$:

$$\langle f, g\rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x})\, g(\boldsymbol{x})\, \gamma_d(\mathrm{d}\boldsymbol{x})\,. \tag{53}$$

For $\ell \in \mathbb{Z}_{\geq 0}$, let $\tilde{V}_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree ℓ on $\mathbb{R}^d$ (i.e. homogeneous polynomials $q(\boldsymbol{x})$ satisfying $\Delta q(\boldsymbol{x}) = 0$), and denote by $V_{d,\ell}$ the linear space of functions obtained by restricting the polynomials in $\tilde{V}_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the following orthogonal decomposition

$$L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}\,. \tag{54}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(d,\ell) = \frac{2\ell+d-2}{\ell}\binom{\ell+d-3}{\ell-1}\,. \tag{55}$$

For each $\ell \in \mathbb{Z}_{\geq 0}$, the spherical harmonics $\{Y_{\ell j}^{(d)}\}_{1\leq j\leq B(d,\ell)}$ form an orthonormal basis of $V_{d,\ell}$:

$$\langle Y_{ki}^{(d)}, Y_{sj}^{(d)}\rangle_{L^2} = \delta_{ij}\delta_{ks}.$$

Note that our convention is different from the more standard one, that defines the spherical harmonics as functions on $\mathbb{S}^{d-1}(1)$. It is immediate to pass from one convention to the other by a simple scaling. We will drop the superscript d and write $Y_{\ell,j} = Y_{\ell,j}^{(d)}$ whenever clear from the context.

We denote by P_k the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \gamma_d)$. This can be written in terms of spherical harmonics as

$$\mathsf{P}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(d,k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{56}$$

Then for a function $f \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$, we have

$$f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \mathsf{P}_k f(\boldsymbol{x}) = \sum_{k=0}^{\infty}\sum_{l=1}^{B(d,k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}).$$

F.2. Gegenbauer polynomials

The ℓ-th Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell\geq 0}$ forms an orthogonal basis on $L^2([-d, d], \tilde{\tau}_d)$ (where $\tilde{\tau}_d$ is the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{i.i.d.} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), satisfying the normalization condition:

$$\langle Q_k^{(d)}, Q_j^{(d)}\rangle_{L^2(\tilde{\tau}_d)} = \frac{1}{B(d,k)}\,\delta_{jk}\,. \tag{57}$$

In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$. As above, we will omit the superscript d when clear from the context (write it as Q_ℓ for notation simplicity).

Gegenbauer polynomials are directly related to spherical harmonics as follows. Fix $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and consider the subspace of V_ℓ formed by all functions that are invariant under rotations in $\mathbb{R}^d$ that keep $\boldsymbol{v}$ unchanged. It is not hard to see that this subspace has dimension one, and coincides with the span of the function $Q_\ell^{(d)}(\langle \boldsymbol{v}, \cdot\rangle)$.

We will use the following properties of Gegenbauer polynomials

1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}),\gamma_d)} = \frac{1}{B(d,k)}\delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{58}$$

2. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \frac{1}{B(d,k)} \sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{59}$$

Note in particular that property 2 implies that –up to a constant– $Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ is a representation of the projector onto the subspace of degree-k spherical harmonics

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(d,k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)\, f(\boldsymbol{y})\, \gamma_d(\mathrm{d}\boldsymbol{y})\,. \tag{60}$$

For a function $\sigma \in L^2([-\sqrt{d}, \sqrt{d}], \tau_d)$ (where τ_d is the distribution of $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / \sqrt{d}$ when $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), denoting its spherical harmonics coefficients $\lambda_{d,k}(\sigma)$ to be

$$\lambda_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]} \sigma(x) Q_k^{(d)}(\sqrt{d}x) \tau_d(x), \tag{61}$$

then we have the following equation holds in $L^2([-\sqrt{d}, \sqrt{d}], \tau_d)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k) Q_k^{(d)}(\sqrt{d}x). \tag{62}$$

F.3. Hermite polynomials

The Hermite polynomials $\{\mathrm{He}_k\}_{k \geq 0}$ form an orthogonal basis of $L^2(\mathbb{R}, \mu_G)$, where $\mu_G(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G \sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\, \mathrm{He}_k(G)\big\} = k!\, \delta_{jk}\,. \tag{63}$$

As a consequence, for any function $\sigma \in L^2(\mathbb{R}, \mu_G)$, we have the decomposition

$$\sigma(x) = \sum_{k=1}^{\infty} \frac{\mu_k(\sigma)}{k!}\, \mathrm{He}_k(x)\,, \qquad \mu_k(\sigma) \equiv \mathbb{E}\big\{\sigma(G)\, \mathrm{He}_k(G)\big\}\,. \tag{64}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram-Schmidt orthogonalization of the monomials $\{x^k\}_{k \geq 0}$ with respect to the measure τ_d, while Hermite polynomial are obtained by Gram-Schmidt orthogonalization with respect to μ_G. Since $\tau_d \Rightarrow \mu_G$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d \to \infty} \mathrm{Coeff}\{Q_k^{(d)}(\sqrt{d}x)\, B(d,k)^{1/2}\} = \mathrm{Coeff}\left\{\frac{1}{(k!)^{1/2}}\, \mathrm{He}_k(x)\right\}\,. \tag{65}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d \to \infty} \lambda_{d,k}(\sigma)(B(d,k)k!)^{1/2}, \tag{66}$$

where $\mu_k(\sigma)$ and $\lambda_{d,k}(\sigma)$ are given in Eq. (64) and (61).

Rethinking Bias-Variance Trade-off for Generalization of Neural Networks

1. Introduction

Bias-variance trade-off is a fundamental principle for understanding the generalization of predictive learning models (Hastie et al., 2001). The *bias* is an error term that stems from a mismatch between the model class and the underlying data distribution, and is typically monotonically non-increasing as a function of the complexity of the model. The *variance* measures sensitivity to fluctuations in the training set and is often attributed to a large number of model parameters. Classical wisdom predicts that model variance increases and bias decreases *monotonically* with model complexity (Geman et al., 1992). Under this perspective, we should seek a model that has neither too little nor too much capacity and achieves the best trade-off between bias and variance.

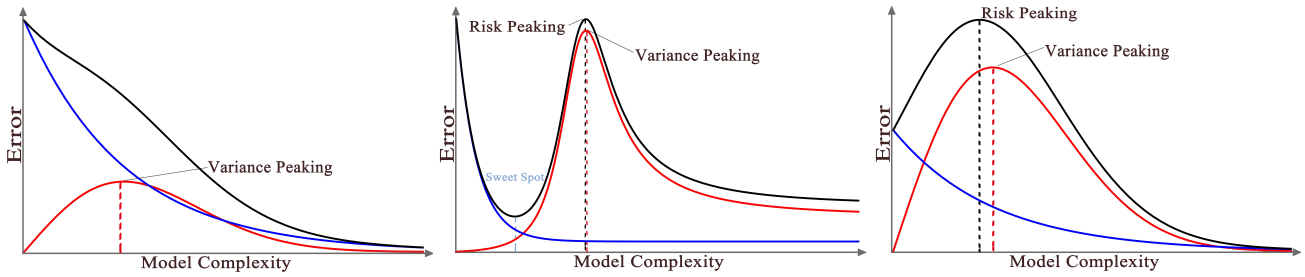
In contrast, modern practice for neural networks repeatedly demonstrates the benefit of increasing the number of neurons (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Zhang et al., 2017), even up to the point of saturating available memory. The inconsistency between classical theory and modern practices suggests that some arguments in the classical theory can not be applied to modern neural networks.

Geman et al. (1992) first studied the bias and variance of the neural networks and give experimental evidence that the variance is indeed increasing as the width of the neural network increases. Since Geman et al. (1992), Neal et al. (2019) first experimentally measured the variance of modern neural network architectures and shown that the variance can actually be decreasing as the width increases to a highly overparameterized regime. Recently, Belkin et al. (2019a; 2018; 2019b) directly studied the risk of modern machine learning models and proposed a *double descent* risk curve, which has also been analytically characterized for certain regression and classification models (Mei & Montanari, 2019; Hastie et al., 2019; Spigler et al., 2019; Deng et al., 2019; Advani & Saxe, 2017; Bartlett et al., 2020; Chatterji & Long, 2020). However, there exists two mysteries around the double descent risk curve. First, the double descent phenomenon can not be robustly observed (Nakkiran et al., 2019; Ba et al., 2020). In particular, to observe it in modern neural network architectures, we sometimes have to artificially inject label noise (Nakkiran et al., 2019). Second, there lacks an explanation for *why* the double descent risk curve should occur. In this work, we offer an simple explanation for these two mysteries by proposing an unexpected *unimodal* variance curve.

Specifically, we measure the bias and variance of modern deep neural networks trained on commonly used computer vision datasets. Our main finding is that while the bias is monotonically decreasing with network width as in the classical theory, the variance curve is *unimodal* or *bell-shaped*: it first increases and then decreases (see Figure 2). Therefore, the unimodal variance is consistent with the finding of Neal et al. (2019), who observed that the variance eventually decreases in the overparameterized regime. In particular, the unimodal variance curve can also be observed in Neal et al. (2019, Figure 1, 2, 3). However, Neal et al. (2019) did not point out the characteristic shape of the variance or connect it to double descent. More importantly, we demonstrate that the unimodal variance phenomenon can be robustly observed for varying network architecture and dataset. Moreover, by using a generalized bias-variance decomposition for Bregman divergences (Pfau, 2013), we verify that it occurs for both squared loss and cross-entropy loss.

This unimodal variance phenomenon initially appears to contradict recent theoretical work suggesting that both bias and variance are non-monotonic and exhibit a peak in some regimes (Mei & Montanari, 2019; Hastie et al., 2019) . The difference is that this previous work considered the *fixed-design* bias and variance, while we measure the *random-design* bias and variance (we describe the differences in detail in §2.1). Prior to our work, Nakkiran (2019) also considered the variance of linear regression in the random-design setting, and Rosset & Tibshirani (2017) discussed additional ways to decompose risk into the bias and the variance term.

A key finding of our work is that the complex behavior of the risk curve arises due to the simple but non-classical variance unimodality phenomenon. Indeed, since the expected risk (test loss) is the sum of bias and variance, monotonic bias and unimodal variance can lead to three characteristic behaviors, illustrated in Figure 1, depending on the relative size of the bias and variance. If the bias completely dominates, we obtain monotonically decreasing risk curve (see Figure 1(a)). Meanwhile, if the variance dominates, we obtain a bell-shaped risk curve that first increases then de-



(a) Case 1 (b) Case 2 (c) Case 3

Figure 1. Typical cases of expected risk curve (in black) in neural networks. Blue: squared bias curve. Red: variance curve.

creases (see Figure 1(c)). The most complex behavior is if bias and variance dominate in different regimes, leading to the double-descent risk curve in Figure 1(b). All three behaviors are well-aligned with the empirical observation in deep learning that larger models typically perform better. The most common behavior in our experiments is the first case (monotonically decreasing risk curve) as bias is typically larger than variance. We can observe the double-descent risk curve when label noise is added to the training set (see §3.3), and can observe the unimodal risk curve when we use the generalized bias-variance decomposition for cross-entropy loss (see §3.2).

Further Implications. The investigations described above characterize bias and variance as a function of network width, but we can explore the dependence on other quantities as well, such as model depth (§4.2). Indeed, we find that deeper models tend to have lower bias but higher variance. Since bias is larger at current model sizes, this confirms the prevailing wisdom that we should generally use deeper models when possible. On the other hand, it suggests that this process may have a limit—eventually very deep models may have low bias but high variance such that increasing the depth further harms performance.

We also investigate the commonly observed drop in accuracy for models evaluated on out-of-distribution data, and attribute it primarily to increased bias. Combined with the previous observation, this suggests that increasing model depth may help combat the drop in out-of-distribution accuracy, which is supported by experimental findings in Hendrycks & Dietterich (2019).

Theoretical Analysis of A Two-Layer Neural Network. Finally, we conduct a theoretical study of a two-layer linear network with a random Gaussian first layer. While this model is much simpler than those used in practice, we nevertheless observe the same characteristic behaviors for the bias and variance. In particular, by working in the asymptotic setting where the input data dimension, amount of training data, and network width go to infinity with fixed ratios, we show that the bias is monotonically decreasing while the variance curve is unimodal. Our analysis also characterizes the location of the variance peak as the point where the number of hidden neurons is approximately half of the dimension of the input data.

2. Preliminaries

In this section we present the bias-variance decomposition for squared loss. We also present a generalized bias-variance decomposition for cross-entropy loss in §2.2. The task is to learn a function $f : \mathbb{R}^d \to \mathbb{R}^c$, based on i.i.d. training samples $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ drawn from a joint distribution P on $\mathbb{R}^d \times \mathbb{R}^c$, such that the mean squared error $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\|\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T})\|_2^2\right]$ is minimal, where $(\boldsymbol{x}, \boldsymbol{y}) \sim P$. Here we denote the learned function by $f(\boldsymbol{x}; \mathcal{T})$ to make the dependence on the training samples clear.

Note that the learned predictor $f(\boldsymbol{x}; \mathcal{T})$ is a random quantity depending on $\mathcal{T}$. We can assess its performance in two different ways. The first way, random-design, takes the expectation over $\mathcal{T}$ such that we consider the expected error $\mathbb{E}_{\mathcal{T}}\left[\|\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T})\|_2^2\right]$. The second way, fixed-design, holds the training covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ fixed and only takes expectation over $\{\boldsymbol{y}_i\}_{i=1}^n$, i.e., $\mathbb{E}_{\mathcal{T}}\left[\|\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T})\|_2^2 \mid \{\boldsymbol{x}_i\}_{i=1}^n\right]$. The choice of random/fixed-design leads to different bias-variance decompositions. Throughout the paper, we focus on random-design, as opposed to fixed-design studied in Mei & Montanari (2019); Hastie et al. (2019); Ba et al. (2020).

2.1. Bias Variance Decomposition

Random Design. In the random-design setting, decomposing the quantity $\mathbb{E}_{\mathcal{T}}\left[\|\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T})\|_2^2\right]$ gives the usual bias-variance trade-off from machine learning, e.g. Geman et al. (1992); Hastie et al. (2001).

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\mathbb{E}_{\mathcal{T}}\left[\|\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T})\|_2^2\right] = \underbrace{\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\|\boldsymbol{y} - \bar{f}(\boldsymbol{x})\|_2^2\right]}_{\textbf{Bias}^2} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\mathcal{T}}\left[\|f(\boldsymbol{x}, \mathcal{T}) - \bar{f}(\boldsymbol{x})\|_2^2\right]}_{\textbf{Variance}},$$

where $\bar{f}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{T}} f(\boldsymbol{x}, \mathcal{T})$. Here $\mathbb{E}_{\mathcal{T}}\left[\|(\boldsymbol{y} - f(\boldsymbol{x}, \mathcal{T})\|_2^2\right]$ measures the average prediction error over different realiza-

tions of the training sample. In addition to take the expectation $\mathbb{E}_{\mathcal{T}}$, we also average over $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}$, as discussed in Bishop (2006, §3.2). For future reference, we define

$$\textbf{Bias}^2 = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\|\boldsymbol{y} - \bar{f}(\boldsymbol{x})\|_2^2\right], \tag{1}$$

$$\textbf{Variance} = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{\mathcal{T}}\left[\|f(\boldsymbol{x},\mathcal{T}) - \bar{f}(\boldsymbol{x})\|_2^2\right]. \tag{2}$$

In §2.2, we present our estimator for bias and variance in equation (1) and (2).

Fixed Design. In fixed-design setting, the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$ are held be fixed, and the only randomness in the training set $\mathcal{T}$ comes from $\boldsymbol{y}_i \sim P(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}_i)$. As presented in Mei & Montanari (2019); Hastie et al. (2019); Ba et al. (2020), a more natural way to present the fixed-design assumption is to hold $\{\boldsymbol{x}_i\}_{i=1}^n$ to be fixed and let $\boldsymbol{y}_i = f_0(\boldsymbol{x}) + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$, where $f_0(\boldsymbol{x})$ is a ground-truth function and $\boldsymbol{\epsilon}_i$ are random noises. Under this assumption, the randomness in $\mathcal{T}$ all comes from the random noise $\boldsymbol{\epsilon}_i$. To make this clear, we write $\mathcal{T}$ as $\mathcal{T}_{\boldsymbol{\epsilon}_i}$. Then, we obtain the *fixed-design* bias-variance decomposition

$$\mathbb{E}_{\boldsymbol{\epsilon}_i}\left[\|(\boldsymbol{y} - f(\boldsymbol{x},\mathcal{T}_{\boldsymbol{\epsilon}_i})\|_2^2\right] = \underbrace{\left[\|(\boldsymbol{y} - \bar{f}(\boldsymbol{x})\|_2^2\right]}_{\textbf{Bias}^2} + \underbrace{\mathbb{E}_{\boldsymbol{\epsilon}_i}\left[\|(f(\boldsymbol{x},\mathcal{T}_{\boldsymbol{\epsilon}_i}) - \bar{f}(\boldsymbol{x})\|_2^2\right]}_{\textbf{Variance}},$$

where $\bar{f}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\epsilon}_i} f(\boldsymbol{x},\mathcal{T}_{\boldsymbol{\epsilon}_i})$. In most practical settings, the expectation $\mathbb{E}_{\boldsymbol{\epsilon}_i} f(\boldsymbol{x},\mathcal{T}_{\boldsymbol{\epsilon}_i})$ *cannot be estimated* from training samples $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, because we do not have access to independent copies of $f(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$. In comparison to the random-design setting, the fixed-design setting tends to have larger bias and smaller variance, since more "randomness" is introduced into the variance term.

2.2. Estimating Bias and Variance

In this section, we present the estimator we use to estimate the bias and variance as defined in equation (1) and (2). The high level idea is to approximate the expectation $\mathbb{E}_{\mathcal{T}}$ by computing the sample average using multiple training sets $\mathcal{T}_1, \ldots, \mathcal{T}_N$. When evaluating the expectation $\mathbb{E}_{\mathcal{T}}$, there is a trade-off between having larger training sets (n) within each training set and having larger number of splits (N), since $n \times N =$ total number of training samples.

Mean Squared Error (MSE). To estimate bias and variance in equation (1) and (2), we introduce an *unbiased* estimator for variance, and obtain bias by subtracting the variance from the risk. Let $\mathcal{T} = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_N$ be a random disjoint split of training samples. In our experiment, we mainly take $N = 2$ (for CIFAR10 each $\mathcal{T}_i$ has 25k samples). To estimate the variance, we use the unbiased estimator

$$\widehat{\mathsf{var}}(\boldsymbol{x},\mathcal{T}) = \frac{1}{N-1}\sum_{j=1}^{N}\left\|f(\boldsymbol{x},\mathcal{T}_j) - \sum_{j=1}^{N}\frac{1}{N}f(\boldsymbol{x},\mathcal{T}_j)\right\|_2^2,$$

Algorithm 1 Estimating Generalized Variance

Input: Test point $\boldsymbol{x}$, Training set $\mathcal{T}$.
for $i = 1$ **to** k **do**
  Split the $\mathcal{T}$ into $\mathcal{T}_1^{(i)}, \ldots, \mathcal{T}_N^{(i)}$.
  for $j = 1$ **to** N **do**
    Train the model using $\mathcal{T}_j^{(i)}$;
    Evaluate the model at $\boldsymbol{x}$; call the result $\pi_j^{(i)}$;
  end for
end for
Compute $\widehat{\pi} = \exp\left\{\frac{1}{N\cdot k}\sum_{ij}\log\left(\pi_j^{(i)}\right)\right\}$
(*using element-wise log and exp; $\widehat{\pi}$ estimates $\bar{\pi}$*).
Normalize $\widehat{\pi}$ to get a probability distribution.
Compute the variance $\frac{1}{N\cdot k}\sum_{ij} D_{\mathrm{KL}}\left(\widehat{\pi}\|\pi_j^{(i)}\right)$.

where var depends on the test point $\boldsymbol{x}$ and on the random training set $\mathcal{T}$. While var is unbiased, its variance can be reduced by using multiple random splits to obtain estimators $\widehat{\mathsf{var}}_1, \ldots, \widehat{\mathsf{var}}_k$ and taking their average. This reduces the variance of the variance estimator since:

$$\mathrm{Var}_{\mathcal{T}}\left(\frac{1}{k}\sum_{i=1}^{k}\widehat{\mathsf{var}}_i\right) = \frac{\sum_{ij}\mathrm{Cov}_{\mathcal{T}}(\widehat{\mathsf{var}}_i, \widehat{\mathsf{var}}_j)}{k^2} \le \mathrm{Var}_{\mathcal{T}}(\widehat{\mathsf{var}}_1),$$

where the $\{\widehat{\mathsf{var}}_i\}_{i=1}^k$ are identically distributed but not independent, and we used the Cauchy-Schwarz inequality.

Cross-Entropy Loss (CE). In addition to the classical bias-variance decomposition for MSE loss, we also consider a generalized bias-variance decomposition for cross-entropy loss. Let $\pi(\boldsymbol{x},\mathcal{T}) \in \mathbb{R}^c$ be the output of the neural network (a probability distribution over the class labels). $\pi(\boldsymbol{x},\mathcal{T})$ is a random variable since the training set $\mathcal{T}$ is random. Let $\pi_0(\boldsymbol{x}) \in \mathbb{R}^c$ be the one-hot encoding of the ground-truth label. Then, omitting the dependence of π and π_0 on $\boldsymbol{x}$ and $\mathcal{T}$, the cross entropy loss

$$H(\pi_0, \pi) = \sum_{l=1}^{c}\pi_0[l]\log(\pi[l])$$

can be decomposed as

$$\mathbb{E}_{\mathcal{T}}\left[H(\pi_0,\pi)\right] = \underbrace{D_{\mathrm{KL}}(\pi_0\|\bar{\pi})}_{\textbf{Bias}^2} + \underbrace{\mathbb{E}_{\mathcal{T}}\left[D_{\mathrm{KL}}(\bar{\pi}\|\pi)\right]}_{\textbf{Variance}}, \tag{3}$$

where $\pi[l]$ is the l-th element of π, and $\bar{\pi}$ is the average of log-probability after normalization, i.e.,

$$\bar{\pi}[l] \propto \exp\{\mathbb{E}_{\mathcal{T}}\log(\pi[l])\} \quad \text{for} \quad l = 1, \ldots, c.$$

This decomposition is a special case of the general decomposition for Bregman divergence discussed in Pfau (2013).

We apply Algorithm 1 to estimate the generalized variance in (3). Here we could not obtain an unbiased estimator, but

the estimate is better if we take more random splits (larger k). In practice, we choose k to be large enough so that the estimated variance stabilizes when we further increase k (see §3.4). Similar to the case of squared loss, we estimate the bias by subtracting the variance from the risk.

3. Measuring Bias and Variance for Neural Networks

In this section, we study the bias and variance (equations (1) and (2)) of deep neural networks. While the bias is monotonically decreasing as folk wisdom would predict, the variance is unimodal (first increases to a peak and then decreases). We conduct extensive experiments to verify that this phenomenon appears robustly across architectures, datasets, optimizer, and loss function. Our code can be found at `https://github.com/yaodongyu/Rethink-BiasVariance-Tradeoff`.

3.1. Mainline Experimental Setup

We first describe our mainline experimental setup. In the next subsection, we vary each design choice to check robustness of the phenomenon. More extensive experimental results are given in the appendix.

For the mainline experiment, we trained a ResNet34 (He et al., 2016) on the CIFAR10 dataset (Krizhevsky et al., 2009). We trained using stochastic gradient descent (SGD) with momentum 0.9. The initial learning rate is 0.1. We applied stage-wise training (decay learning rate by a factor of 10 every 200 epochs), and used weight decay 5×10^{-4}. To change the model complexity of the neural network, we scale the number of filters (i.e., width) of the convolutional layers. More specifically, with width $= w$, the number of filters are $[w, 2w, 4w, 8w]$. We vary w from 2 to 64 (the width w of a regular ResNet34 designed for CIFAR10 in He et al. (2016) is 16).

Relative to the standard experimental setup (He et al., 2016), there are two main differences. First, since bias-variance is usually defined for the squared loss (see (1) and (2)), our loss function is the squared error (squared ℓ_2 distance between the softmax probabilities and the one-hot class vector) rather than the log-loss. In the next section we also consider models trained with the log-loss and estimate the bias and variance by using a generalized bias-variance decomposition, as described in §2.2. Second, to measure the variance (and hence bias), we need two models trained on independent subsets of the data as discussed in §2.2. Therefore, the training dataset is split in half and each model is trained on only $n = 25,000 = 50,000/2$ data points. We estimate the variance by averaging over $N = 3$ such random splits (i.e., we train $6 = 3 \times 2$ copies of each model).

In Figure 2, we can see that the variance as a function of the width is unimodal and the bias is monotonically decreasing. Since the scale of the variance is small relative to the bias, the overall behavior of the risk is monotonically decreasing.

3.2. Varying Architectures, Loss Functions, Datasets

Architectures. We observe the same monotonically descreasing bias and unimodal variance phenomenon for ResNext29 (Xie et al., 2017). To scale the "width" of the ResNext29, we first set the number of channels to 1 and increase the *cardinality*, defined in (Xie et al., 2017), from 2 to 4, and then fix the cardinality at 4 and increase channel size from 1 to 32. Results are shown in Figure 3(a), where the width on the x-axis is defined as the cardinality times the filter size.

Loss Function. In addition to the bias-variance decomposition for MSE loss, we also considered a similar decomposition for cross-entropy loss as described in §2.2. We train with cross-entropy loss and use $n = 10,000$ training samples (5 splits), repeating $N = 4$ times with independent random splits. As shown in Figure 3(b), the behavior of the generalized bias and variance for cross entropy is consistent with our earlier observations: the bias is monotonically decreasing and the variance is unimodal. The risk first increases and then decreases, corresponding to the unimodal risk pattern in Figure 1(c).

Datasets. In addition to CIFAR10, we study bias and variance on MNIST (LeCun, 1998) and Fashion-MNIST (Xiao et al., 2017). For these two datasets, we use a fully connected neural network with one hidden layer with ReLU activation function. The "width" of the network is the number of hidden nodes. We use 10,000 training samples ($N = 5$). As seen in Figure 3(c) and 10 (in Appendix B), for both MNIST and Fashion-MNIST, the variance is again unimodal and the bias is monotonically decreasing.

In addition to the above experiments, we also conduct experiments on the CIFAR100 dataset, the VGG network architecture (Simonyan & Zisserman, 2015), various training sample sizes, and different weight decay regularization and present the results in Appendix B. We observe the same monotonically descreasing bias and unimodal variance phenomenon in *all* of these experiments.

3.3. Connection to Double-Descent Risk

When the relative scale of bias and variance changes, the risk displays one of the three patterns, *monotonically decreasing*, *double descent*, and *unimodal*, as presented in Figure 1(a), 1(b) and 1(c). In particular, the recent stream of observations on double descent risk (Belkin et al., 2019a) can be explained by unimodal variance and monotonically decreasing bias. In our experiments, including the experiments in previous sections, we typically observe monotonically

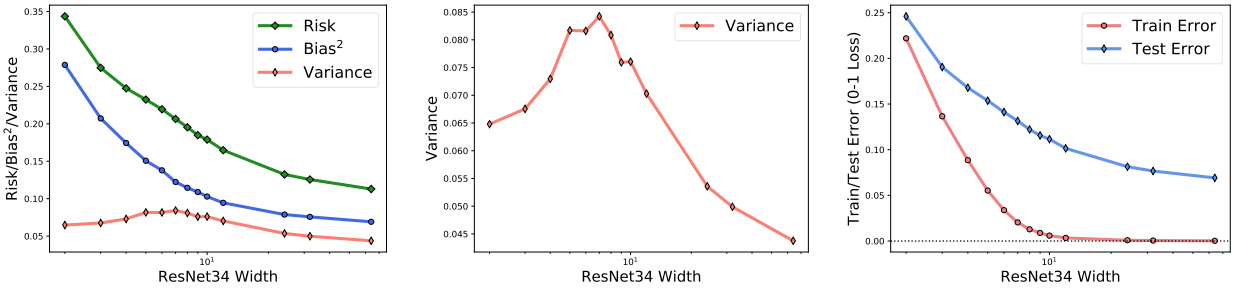
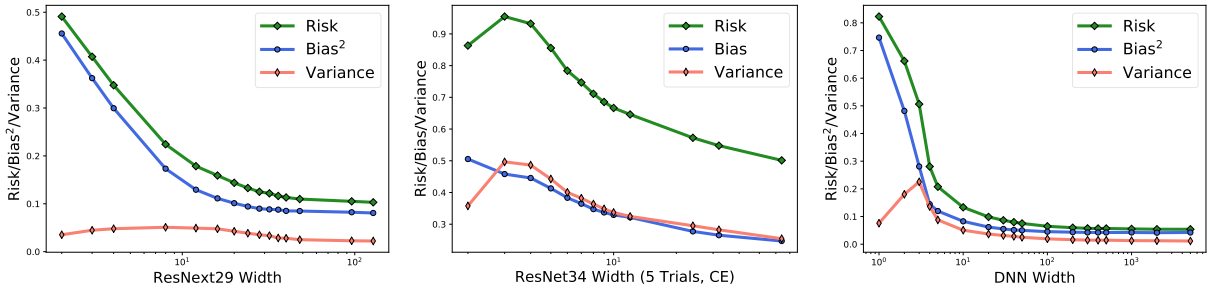


Figure 2. Mainline experiment on ResNet34, CIFAR10 dataset (25,000 training samples). (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.



(a) ResNext29, MSE loss, CIFAR10 (b) ResNet34, CE loss, CIFAR10 (c) DNN, MSE loss, MNIST

Figure 3. Risk, bias, and variance with respect to different network architectures, training loss functions, and datasets. (a). ResNext29 trained by MSE loss on CIFAR10 dataset (25,000 training samples). (b). ResNet34 trained by CE loss (estimated by generalized bias-variance decomposition using Bregman divergence) on CIFAR10 dataset (10,000 training samples). (c). Fully connected network with one hidden layer and ReLU activation trained by MSE loss on MNIST dataset (10,000 training samples).

decreasing risk; but with more label noise, the variance will increase and we observe the double descent risk curve.

Label Noise. Similar to the setup in Nakkiran (2019), for each split, we sample training data from the whole training dataset, and replace the label of each training example with a uniform random class with independent probability p. Label noise increases the variance of the model and hence leads to double-descent risk as seen in Figure 4. If the variance is small, the risk does not have the double-descent shape because the variance peak is not large enough to overwhelm the bias, as observed in Figures 2, 3(a), 3(c) and 10.

3.4. Discussion of Possible Sources of Error

In this section, we briefly describe the possible sources of error in our estimator defined in §2.2.

Mean Squared Error. As argued in §2.2, the variance estimator is unbiased. To understand the variance of the estimator, we first split the data into two parts. For each part, we compute the bias and variance for varying network width by using our estimator. Averaging across different model width, the relative difference between the two parts is 0.6% for bias and 3% for variance, so our results for MSE are minimally sensitive to finite-sample effects. The complete experiments can be found in the appendix (see Figure 17).

Cross Entropy Loss. For cross entropy loss, we are currently unable to obtain an unbiased estimator. We can assess the quality of our estimator using the following scheme. We partition the dataset into five parts $\mathcal{T}_1, \ldots, \mathcal{T}_5$, i.e., set $N = 5$ in Algorithm 1. Then, we sequentially plot the estimate of bias and variance using $k = 1, 2, 3, 4$ as described in Algorithm 1. We observe that using larger k gives better estimates. In Figure 18 of Appendix B.9, we observe that as k increases, the bias curve systematically decreases and the variance curve increases. Therefore our estimator over-estimates the bias and under-estimates the variance, but the overall behaviors of the curves remain consistent.

4. What Affects the Bias and Variance?

In this section, through the Bias-Variance decomposition analyzed in §3, we investigate the role of depth for neural networks and the robustness of neural networks on out-of-distribution examples.

4.1. Bias-Variance Tradeoff for Out-of-Distribution (OOD) Example

For many real-world computer vision applications, inputs can be corrupted by random noise, blur, weather, etc. These common occurring corruptions are shown to signifi-

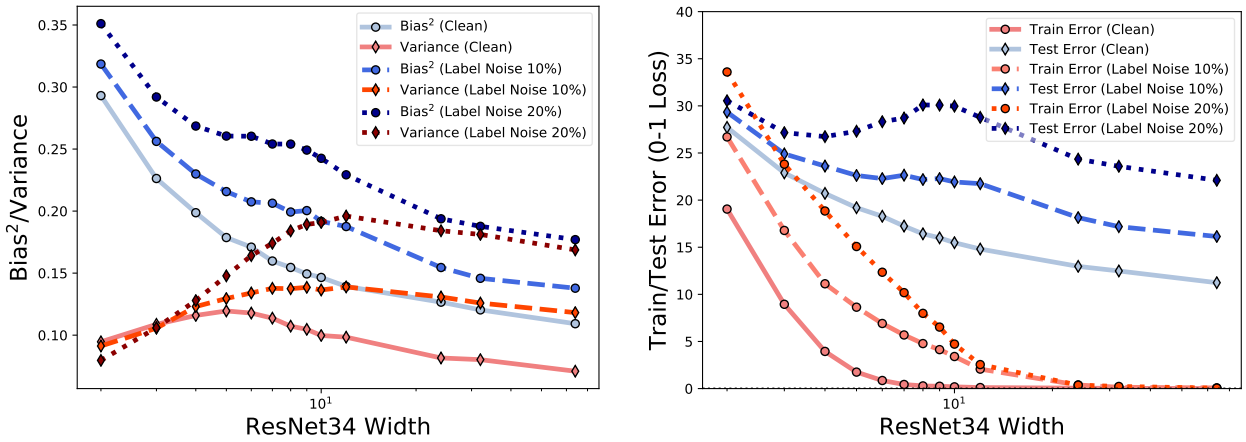
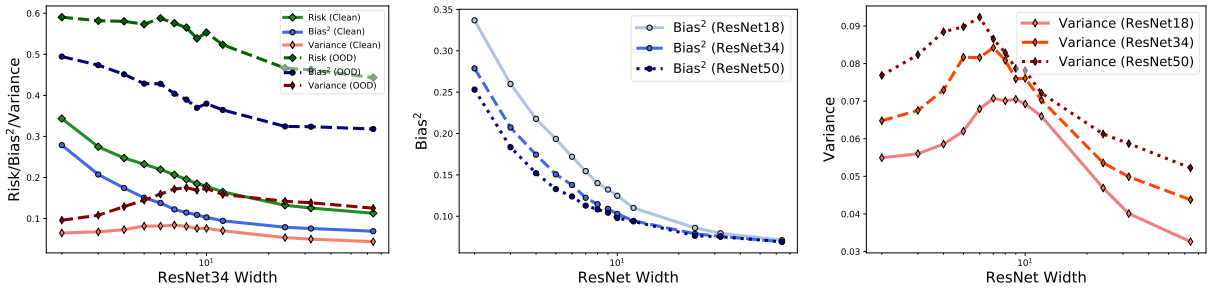


Figure 4. Increasing label noise leads to double-descent. (**Left**) Bias and variance under different label noise percentage. (**Right**) Training error and test error under different label noise percentage.



(a) OOD Example (b) Bias of model with different depth (c) Variance of model with different depth

Figure 5. (a). Risk, bias, and variance for ResNet34 on out-of-distribution examples (CIFAR10-C dataset). (b)-(c). Bias and variance for ResNet with different depth trained by MSE loss on CIFAR10 (25,000 training samples).

cantly decrease model performance (Azulay & Weiss, 2019; Hendrycks & Dietterich, 2019). To better understand the "generalization gap" between in-distribution test examples and out-of-distribution test examples, we empirically evaluate the bias and variance on the CIFAR10-C dataset developed by Hendrycks & Dietterich (2019), which is a common corruption benchmark and includes 15 types of corruption.

By applying the models trained in the mainline experiment, we are able to evaluate the bias and variance on CIFAR10-C test dataset according to the definitions in (1) and (2). As we can see from Figure 5(a), both the bias and variance increase relative to the original CIFAR10 test set. Consistent with the phenomenon observed in the mainline experiment, the bias dominates the overall risk. The results indicate that the "generalization gap" mainly comes from increased bias, with relatively less contribution from variance as well.

4.2. Effect of Model Depth on Bias and Variance

In addition to the ResNet34 considered in the mainline experiment, we also evaluate the bias and variance for ResNet18 and ResNet50. Same as the mainline experiment setup, we estimate the bias and variance for ResNet using 25,000 training samples ($N = 2$) and three independent random splits ($k = 3$). The standard building block of ResNet50 architecture in He et al. (2016) is bottleneck block, which is different from the basic block used in ResNet18 and ResNet34. To ensure that depth is the only changing variable across three architectures, we apply the basic block for ResNet50. Same training epochs and learning rate decays are applied to three models.

From Figure 5(b) and 5(c), we observe that *the bias decreases as the depth increases, while the variance increases as the depth increases*. For each model, the bias is monotonically decreasing and the variance is unimodal. The differences in variance are small (around 0.01) compared with the changes in bias. Overall, the risk typically decreases as the depth increases. Our experimental results suggest that the improved generalization for deeper models, with the same network architecture, are mainly attributed to lower bias.

For completeness, we also include the bias and variance versus depth when basic blocks in ResNet are replaced by bottleneck blocks (see Figure 20 in the appendix). We observe similar qualitative trend of bias and variance.

Note that at high width, the bias of ResNet50 is slightly higher than the bias of ResNet18 and ResNet34. We attribute this inconsistency to difficulties when training ResNet50 without bottleneck blocks at high width. Lastly, we also include the bias and variance versus depth for out-of-distribution test samples, in which case we also observed decreased bias and increased variance as depth increases, as shown in Figure 19 of Appendix B.10.

5. Theoretical Insights from a Two-layer Linear Model

While the preceding experiments show that the bias and variance robustly exhibit monotonic-unimodal behavior in the *random-design* setting, existing theoretical analyses hold instead for the *fixed-design* setting, where the behavior of the bias and variance are more complex, with both the bias and variance exhibiting a peak and the risk exhibiting double descent pattern (Mei & Montanari (2019, Figure 6)). In general, while the risk should be the same (in expectation) for the random and fixed design setting, the fixed-design setting has lower bias and higher variance.

Motivated by the more natural behavior in the random-design setting, we work to extend the existing fixed-design theory to the random-design case. Our starting point is Mei & Montanari (2019), who consider two-layer non-linear networks with random hidden layer weights. However, the randomness in the design complicates the analysis, so we make two points of departure to help simplify: first, we consider two-layer *linear* rather than non-linear networks, and second, we consider a different scaling limit ($n/d \to \infty$ rather than n/d going to some constant). In this setting, we rigorously show that the variance is indeed unimodal and the bias is monotonically decreasing (Figure 6). Our precise assumptions are given below.

5.1. Model Assumptions

We consider the task of learning a function $y = f(\boldsymbol{x})$ that maps each input vector $\boldsymbol{x} \in \mathbb{R}^d$ to an output (label) value $y \in \mathbb{R}$. The input-output pair $(\boldsymbol{x}, y)$ is assumed to be drawn from a distribution where $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d/d)$ and

$$y = f_0(\boldsymbol{x}) := \boldsymbol{x}^\top \boldsymbol{\theta}, \tag{4}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a weight vector. Given a training set $\mathcal{T} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with training samples drawn independently from the data distribution, we learn a two-layer linear neural network parametrized by $\boldsymbol{W} \in \mathbb{R}^{p\times d}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ as

$$f(\boldsymbol{x}) = (\boldsymbol{W}\boldsymbol{x})^\top \boldsymbol{\beta},$$

where p is the number of hidden units in the network. In above, we take $\boldsymbol{W}$ as a parameter independent of the training data $\mathcal{T}$ whose entries are drawn from i.i.d. Gaussian distribution $\mathcal{N}(0, 1/d)$. Given $\boldsymbol{W}$, the parameter $\boldsymbol{\beta}$ is estimated by solving the following *ridge regression*[1] problem

$$\boldsymbol{\beta}_\lambda(\mathcal{T}, \boldsymbol{W}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|(\boldsymbol{W}\boldsymbol{X})^\top \boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{5}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ denotes a matrix that contains training data vectors as its columns, $\boldsymbol{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ denotes a vector containing training labels as its entries, and $\lambda \in \mathbb{R}^+$ is the regularization parameter. By noting that the solution to (5) is given by

$$\boldsymbol{\beta}_\lambda(\mathcal{T}, \boldsymbol{W}) = (\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{W}^\top + \lambda\boldsymbol{I})^{-1}\boldsymbol{W}\boldsymbol{X}\boldsymbol{y},$$

our estimator $f : \mathbb{R}^d \to \mathbb{R}$ is given as

$$f_\lambda(\boldsymbol{x}; \mathcal{T}, \boldsymbol{W}) = \boldsymbol{x}^\top \boldsymbol{W}^\top \boldsymbol{\beta}_\lambda(\mathcal{T}, \boldsymbol{W}). \tag{6}$$

5.2. Bias-Variance Analysis

We may now calculate the bias and variance of the model described above via the following formulations:

$$\textbf{Bias}_\lambda(\boldsymbol{\theta})^2 = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{T},\boldsymbol{W}} f_\lambda(\boldsymbol{x}; \mathcal{T}, \boldsymbol{W}) - f_0(\boldsymbol{x})\right]^2,$$
$$\textbf{Variance}_\lambda(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\mathrm{Var}_{\mathcal{T},\boldsymbol{W}}\left[f_\lambda(\boldsymbol{x}; \mathcal{T}, \boldsymbol{W})\right],$$

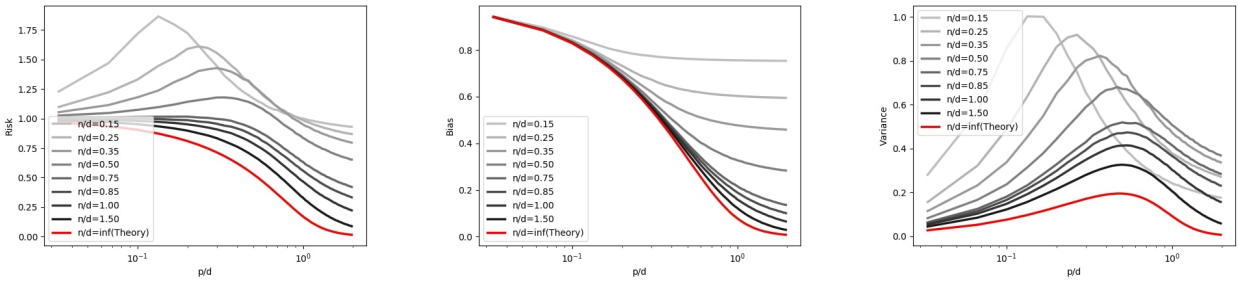
where $f_0(\boldsymbol{x})$ and $f_\lambda(\boldsymbol{x}; \mathcal{T}, \boldsymbol{W})$ are defined in (4) and (6), respectively. Note that the bias and variance are functions of the model parameter $\boldsymbol{\theta}$. To simplify the analysis, we introduce a prior $\boldsymbol{\theta} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and calculate the expected bias and expected variance as

$$\textbf{Bias}_\lambda^2 := \mathbb{E}_{\boldsymbol{\theta}}\textbf{Bias}_\lambda(\boldsymbol{\theta})^2, \tag{7}$$
$$\textbf{Variance}_\lambda := \mathbb{E}_{\boldsymbol{\theta}}\textbf{Variance}_\lambda(\boldsymbol{\theta}). \tag{8}$$

The precise formulas for the expected bias and the expected variance are parametrized by the dimension of the input feature d, the number of training points n, the number of hidden units p and also λ.

Previous literatures (Mei & Montanari, 2019) suggests that both the risk and the variance achieves a peak at the interpolation threshold ($n = p$). In the regime when n is very large, the risk no longer exhibits a peak, but the unimodal pattern of variance still holds. In the rest of the section, we consider the regime where the n is large (monotonically decreasing risk), and derive the precise expression for the bias and variance of the model. From our expression, we obtain the location where the variance achieves the peak. For this purpose, we consider the following asymptotic regime of n, p and d:

[1] ℓ_2 regularization on weight parameters is arguably the most widely used technique in training neural network, known for improving generalization (Krogh & Hertz, 1992). Other regularization such as ℓ_1 can also be used and leads to qualitatively similar behaviors.



(a) Risk v.s. γ for different n (b) Bias v.s. γ for different n (c) Variance v.s. γ for different n

Figure 6. Risk, bias, and variance for a two-layer linear neural network.

Assumption 1. *Let $\{(d, n(d), p(d))\}_{d=1}^{\infty}$ be a given sequence of triples. We assume that there exists a $\gamma > 0$ such that*

$$\lim_{d\to\infty} \frac{p(d)}{d} = \gamma, \quad \textit{and} \quad \lim_{d\to\infty} \frac{n(d)}{d} = \infty.$$

For simplicity, we will write $n := n(d)$ and $p := p(d)$.

With the assumption above, we have the expression of the expected bias, variance and risk as a function of γ and λ.

Theorem 1. *Given $\{(d, n(d), p(d))\}_{d=1}^{\infty}$ that satisfies Assumption 1, let $\lambda = \frac{n}{d}\lambda_0$ for some fixed $\lambda_0 > 0$. The asymptotic expression of expected bias and variance are given by*

$$\lim_{d\to\infty} \textbf{\textit{Bias}}_\lambda^2 = \frac{1}{4}\Phi_3(\lambda_0, \gamma)^2, \tag{9}$$

$$\lim_{d\to\infty} \textbf{\textit{Variance}}_\lambda = \begin{cases} \frac{\Phi_1(\lambda_0, \frac{1}{\gamma})}{2\Phi_2(\lambda_0, \frac{1}{\gamma})} - \frac{(1-\gamma)(1-2\gamma)}{2\gamma} - \frac{1}{4}\Phi_3(\lambda_0, \gamma)^2, & \gamma \le 1, \\ \frac{\Phi_1(\lambda_0, \gamma)}{2\Phi_2(\lambda_0, \gamma)} - \frac{\gamma-1}{2} - \frac{1}{4}\Phi_3(\lambda_0, \gamma)^2, & \gamma > 1, \end{cases}$$

where

$$\Phi_1(\lambda_0, \gamma) = \lambda_0(\gamma+1) + (\gamma-1)^2,$$
$$\Phi_2(\lambda_0, \gamma) = \sqrt{(\lambda_0+1)^2 + 2(\lambda_0-1)\gamma + \gamma^2},$$
$$\Phi_3(\lambda_0, \gamma) = \Phi_2(\lambda_0, \gamma) - \lambda_0 - \gamma + 1.$$

The proof is given in Appendix C.

The risk can be obtained through $\textbf{Bias}_\lambda^2 + \textbf{Variance}_\lambda$. The expression in Theorem 1 is plotted as the red curves in Figure 6. In addition to the case when $n/d \to \infty$, we also plot the shape of bias, variance and risk when $n/d \to \{0.15, 0.25, 0.35, \ldots, 1.00, 1.50\}$. We find that the risk of the model grows from unimodal to monotonically decreasing as the number of samples increased (see Figure 6(a)). Moreover, the bias of the model is monotonically decreasing (see Figure 6(b)) and the variance is unimodal (see Figure 6(c)).

Corollary 1 (Monotonicity of Bias)**.** *The derivative of the limiting expected Bias in (9) can be calculated as*

$$-\frac{\left(\sqrt{2(\gamma+1)\lambda_0 + (\gamma-1)^2 + \lambda_0^2} - \gamma - \lambda_0 + 1\right)^2}{2\sqrt{\gamma^2 + 2\gamma(\lambda_0 - 1) + (\lambda_0+1)^2}}. \tag{10}$$

When $\lambda_0 \ge 0$, the expression in (10) is strictly non-positive, therefore the limiting expected bias is monotonically non-increasing as a function of γ, as classical theories predicts.

To gain further insight into the above formulas, we also consider the case when the ridge regularization amount λ_0 is small. In particular, we consider the first order effect of λ_0 on the bias and variance term, and compute the value of γ where the variance attains the peak.

Corollary 2 (Unimodality of Variance – small λ_0 limit)**.** *Under the assumptions of Theorem 1, the first order effect of λ_0 on variance is given by*

$$\lim_{d\to\infty} \mathbb{E}\textbf{\textit{Variance}}_\lambda = \begin{cases} O\left(\lambda_0^2\right), & \gamma > 1, \\ -(\gamma-1)\gamma - 2\gamma\lambda_0 + O\left(\lambda_0^2\right), & o.w. \end{cases}$$

and the risk is given by

$$\lim_{d\to\infty} \mathbb{E}\textbf{\textit{Risk}}_\lambda = \begin{cases} 1 - \gamma + O\left(\lambda_0^2\right), & \gamma \le 1, \\ O\left(\lambda_0^2\right), & \gamma > 1. \end{cases}$$

Moreover, up to first order, the peak in the variance is

$$\textbf{\textit{Peak}} = \frac{1}{2} - \lambda_0 + O\left(\lambda_0^2\right).$$

Theorem 2 suggests that when λ_0 is sufficiently small, the variance of the model is maximized when $p = d/2$, and the effect of λ_0 is to shift the peak slightly to $d/2 - \lambda_0 d$.

From a technical perspective, to compute the variance in the random-design setting, we need to compute the element-wise expectation of certain random matrix. For this purpose, we apply the combinatorics of counting non-cross partitions to characterize the asymptotic expectation of products of Wishart matrices.

6. Conclusion and Discussion

In this paper we re-examine the classical theory of bias and variance trade-off as the width of a neural network increases. Through extensive experimentation, our main finding is that, while the bias is monotonically decreasing as classical theory would predict, the variance is unimodal. This combination leads to three typical risk curve patterns, all observed in practice. Theoretical analysis of a two-layer linear network corroborates these experimental observations.

The seemingly varied and baffling behaviors of modern neural networks are thus in fact consistent, and explainable through classical bias-variance analysis. The main unexplained mystery is the unimodality of the variance. We conjecture that as the model complexity approaches and then goes beyond the data dimension, it is regularization in model estimation (the ridge penalty in our theoretical example) that helps bring down the variance. Under this account, the decrease in variance for large dimension comes from better conditioning of the empirical covariance, making it better-aligned with the regularizer.

In the future, it would be interesting to see if phenomena characterized by the simple two-layer model can be rigorously generalized to deeper networks with nonlinear activation, probably revealing other interplays between model complexity and regularization (explicit or implicit). Such a study could also help explain another phenomenon we (and others) have observed: bias decreases with more layers as variance increases. We believe that the (classic) bias-variance analysis remains a powerful and insightful framework for understanding the behaviors of deep networks; properly used, it can guide practitioners to design more generalizable and robust networks in the future.

Acknowledgements. We would like to thank Emmanuel Candés for first bringing the double-descent phenomenon to our attention, Song Mei for helpful discussion regarding random v.s. fixed design regression, and Nikhil Srivastava for pointing out to relevant references in random matrix theory. We would also like to thank Preetum Nakkiran, Mihaela Curmei, and Chloe Hsu for valuable feedback during preparation of this manuscript. The authors acknowledge support from Tsinghua-Berkeley Shenzhen Institute Research Fund and BAIR.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *ArXiv*, abs/1710.03667, 2017.

Azulay, A. and Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019.

Ba, J., Erdogdu, M., Suzuki, T., Wu, D., and Zhang, T. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gBsgBYwH.

Bai, Z. and Silverstein, J. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 01 2010. doi: 10.1007/978-1-4419-0661-8.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/content/early/2020/04/22/1907378117.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549, 2018.

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

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619, 2019b.

Bishop, A. N., Del Moral, P., and Niclas, A. *An Introduction to Wishart Matrix Moments*. now, 2018. URL https://ieeexplore.ieee.org/document/8572806.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Chatterji, N. S. and Long, P. M. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime, 2020.

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

Dietterich, T. G. and Kong, E. B. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Technical report, Department of Computer Science, Oregon State University, 1995.

Geman, S. A limit theorem for the norm of random matrices. *Ann. Probab.*, 8(2):252–261, 04 1980. doi: 10.1214/aop/1176994775. URL https://doi.org/10.1214/aop/1176994775.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Ghaoui, L. E. Inversion error, condition number, and approximate inverses of uncertain matrices. *Linear Algebra and its Applications*, 343-344:171 – 193, 2002. ISSN 0024-3795. doi: https://doi.org/10.1016/S0024-3795(01)00273-7. URL http://www.sciencedirect.com/science/article/pii/S0024379501002737. Special Issue on Structured and Infinite Systems of Linear equations.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints*, art. arXiv:1903.08560, Mar 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Krizhevsky, A. et al. Learning multiple layers of features from tiny images, 2009.

Krogh, A. and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, art. arXiv:1908.05355, Aug 2019.

Nakkiran, P. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. A modern take on the bias-variance tradeoff in neural networks, 2019. URL https://openreview.net/forum?id=HkgmzhC5F7.

Pfau, D. A generalized bias-variance decomposition for bregman divergences, 2013.

Rosset, S. and Tibshirani, R. J. From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation. *arXiv e-prints*, art. arXiv:1704.08160, April 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 2019.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

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

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

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

A. Summary of Experiments

We summarize the experiments in Table 1, each row corresponds to one experiment, some include several independent splits, in this paper. Every experiment is related to one or multiple figures, which is specified in the last column "Figure".

Dataset	Architecture	Loss	Optimizer	Train Size	#Splits(k)	Label Noise	Figure	Comment
CIFAR10	ResNet34	MSE	SGD(wd=5e-4)	25000	3	✗	2, 5	Mainline
CIFAR10	ResNext29	MSE	SGD(wd=5e-4)	25000	3	✗	3(a), 7	Architecture
	VGG11	MSE	SGD(wd=5e-4)	10000	1	✗	8	
CIFAR10	ResNet34	CE	SGD(wd=5e-4)	10000	4	✗	3(b), 9	Loss
MNIST	DNN	MSE	SGD(wd=5e-4)	10000	1	✗	3(c)	Dataset
Fashion-MNIST	DNN	MSE	SGD(wd=5e-4)	10000	1	✗	10	
CIFAR100	ResNet34	CE	SGD(wd=5e-4)	10000	1	✗	11	
CIFAR10	ResNet34	MSE	SGD(wd=5e-4)	10000	1	10%/20%	4	Label noise
CIFAR10	ResNet18	MSE	SGD(wd=5e-4)	25000	3	✗	5	Depth
	ResNet50	MSE	SGD(wd=5e-4)	25000	3	✗	5	
CIFAR10	ResNet34	MSE	SGD(wd=5e-4)	10000	1	✗	12	Train size
	ResNet34	MSE	SGD(wd=5e-4)	2500	1	✗	13	
CIFAR10	ResNet34	MSE	SGD(wd=1e-4)	10000	1	✗	14	Weight decay
CIFAR10	ResNet26-B	MSE	SGD(wd=5e-4)	25000	3	✗	20	Depth (with bottleneck block)
	ResNet38-B	MSE	SGD(wd=5e-4)	25000	3	✗	20	
	ResNet50-B	MSE	SGD(wd=5e-4)	25000	3	✗	20	
CIFAR10	VGG9	MSE	SGD(wd=5e-4)	25000	3	✗	21	Depth
	VGG11	MSE	SGD(wd=5e-4)	25000	3	✗	21	

Table 1. Summary of Experiments.

B. Additional Experiments

In this section, we provide additional experimental results, some of them are metioned in §3 and §4.

Network Architecture: The implementation of the deep neural networks used in this work is mainly adapted from https://github.com/kuangliu/pytorch-cifar.

Training Details: For CIFAR10 dataset and CIFAR100 dataset, when training sample size is 25,000, we use 500 epochs for training and decay by a factor of 10 the learning rate every 200 epoch. When training sample size is 10,000/5,000, we use 1000 epochs for training and decay by a factor of 10 the learning rate every 400 epoch. For MNIST dataset and FMNIST dataset, we use 200 epochs for training and decay by a factor of 10 the learning rate every 100 epoch. For all the experiments in this paper, we sampled data without replacement to train the models as described in §2.2.

B.1. Architecture

We provide additional results on ResNext29 presented in §3.2. The results are shown in Figure 7. We also study the behavior of risk, bias, and variance of VGG network (Simonyan & Zisserman, 2015) on CIFAR10 dataset. Here we use VGG11 and the number of filters are $[k, 2k, 4k, 4k, 8k, 8k, 8k, 8k]$, where k is the width in Figure 8. The number of training samples of each split is 10,000. We use the same optimization setup as the mainline experiment (ResNet34 in Figure2).

B.2. Loss

We provide additional results on cross-entropy loss presented in §3.2, the results are shown in Figure 9.

B.3. Dataset

We provide the results on Fashion-MNIST dataset in Figure 10, which is mentioned in §3.2. We study the behavior of risk, bias, and variance of ResNet34 on CIFAR100 dataset. Because the number of class is large, we use cross-entropy

during training, and apply the classical Bias-Vairance decomposition for MSE in (1) and (2) to estimate the risk, bias, and variance. As shown in Figure 11, we observe the bell-shaped variance curve and the monotonically decreasing bias curve on CIFAR100 dataset.

B.4. Training Size

Appart from the 2 splits case in Figure 2, we also consider 5 splits (10,000 training samples) and 20 splits case (2,500 training samples). We present the 5 splits case (10,000 training samples) in Figure 12, which corresponds to the label 0% case in Figure 4. We present the 20 splits (2,500 training samples) in Figure 13. With less number of training samples, both the bias and the variance will increase.

B.5. Weight Decay

We study another different weight decay parameter, (wd=`1e-4`) for ResNet34 on CIFAR10 dataset (10,000 training samples). The risk, bias, variance, and train/test error curves are shown in Figure 14. Compared with Figure 12, we observe that larger weight decay can decrease the variance.

B.6. Label Noise

We provide the risk curve for ResNet34 under different label noise percentage as described in §3.3, and the results are shown in Figure 15.

B.7. 0-1 Loss Bias-Variance Decomposition

We evaluated the bias and variance for 0-1 loss (defined in Dietterich & Kong (1995)) on the CIFAR10 dataset with 10,000 training samples using ResNet34. The results are shown in Figure 16. We can consistently observe that the bias is monotonically decreasing and the variance is unimodal.

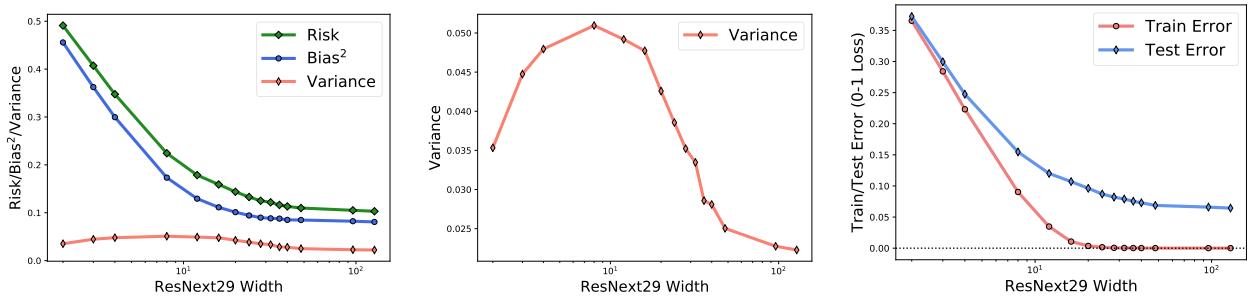


Figure 7. Risk, bias, variance, train/test error for ResNext29 trained by MSE loss on CIFAR10 dataset (25,000 training samples). (**Left**) Risk, bias, and variance for ResNext29. (**Middle**) Variance for ResNext29. (**Right**) Train error and test error for ResNext29.

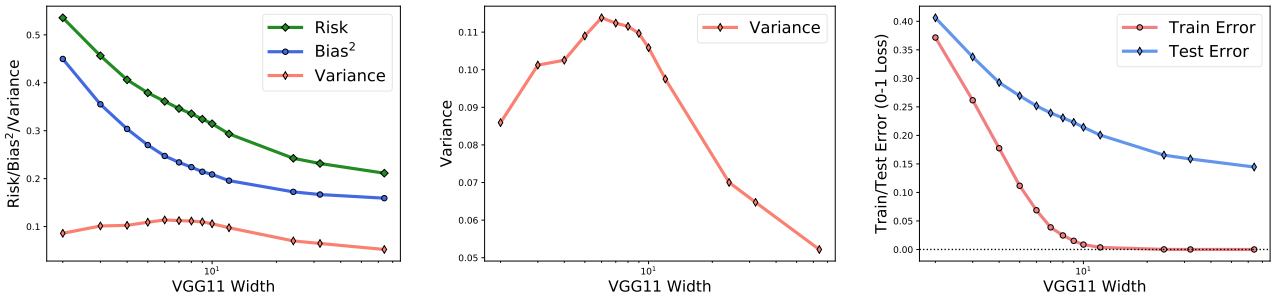


Figure 8. Risk, bias, variance, train/test error for VGG11 trained by MSE loss on CIFAR10 dataset (10,000 training samples). (**Left**) Risk, bias, and variance for VGG11. (**Middle**) Variance for VGG11. (**Right**) Train error and test error for VGG11.

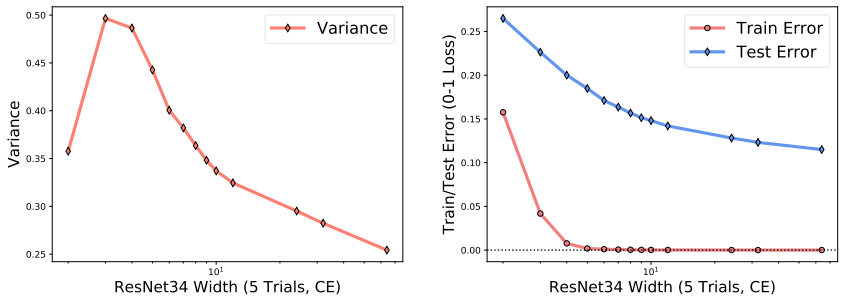


Figure 9. Variance and train/test error for ResNet34 trained by cross-entropy loss (estimated by generalized bias-variance decomposition using Bregman divergence) on CIFAR10 dataset (10,000 training samples). (**Left**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.

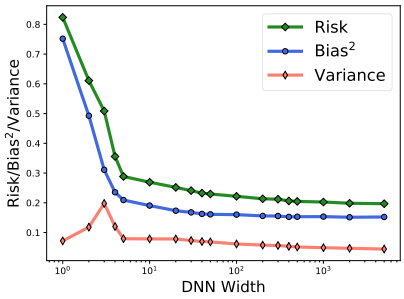


Figure 10. Fully connected network with one-hidden-layer and ReLU activation trained by MSE loss on Fashion-MNIST dataset (10,000 training samples).

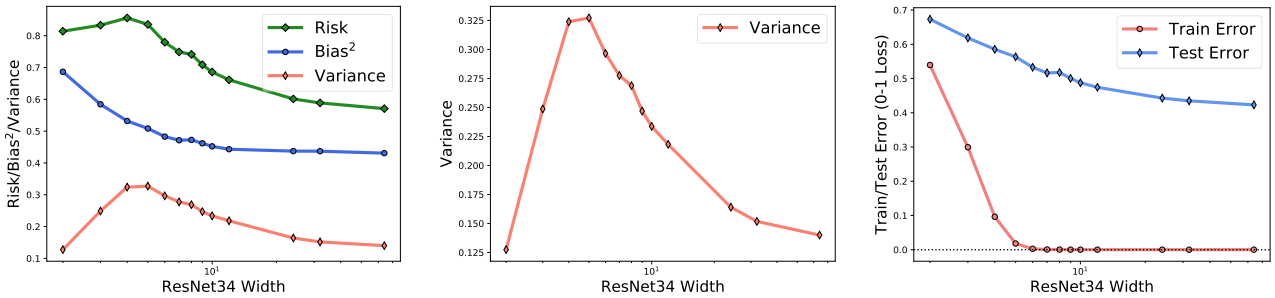


Figure 11. Risk, bias, variance, and train/test error for ResNet34 trained by cross-entropy loss (estimated by MSE bias-variance decomposition) on CIFAR100 (10,000 training samples). (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.

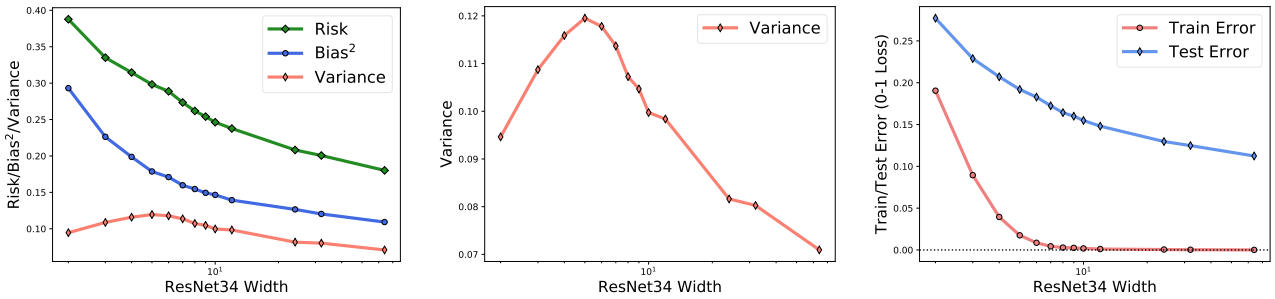


Figure 12. Risk, bias, variance, train/test error for ResNet34 trained by MSE loss on CIFAR10 dataset (10,000 training samples). (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.

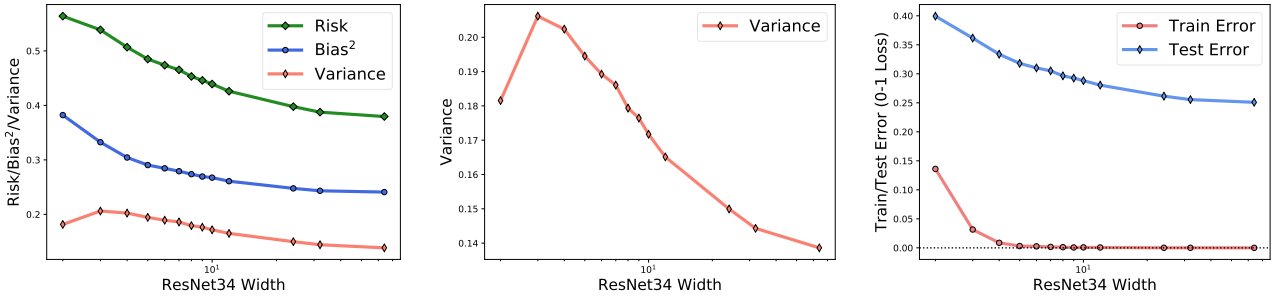


Figure 13. Risk, bias, variance, train/test error for ResNet34 trained by MSE loss on CIFAR10 dataset (2,500 training samples). (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.

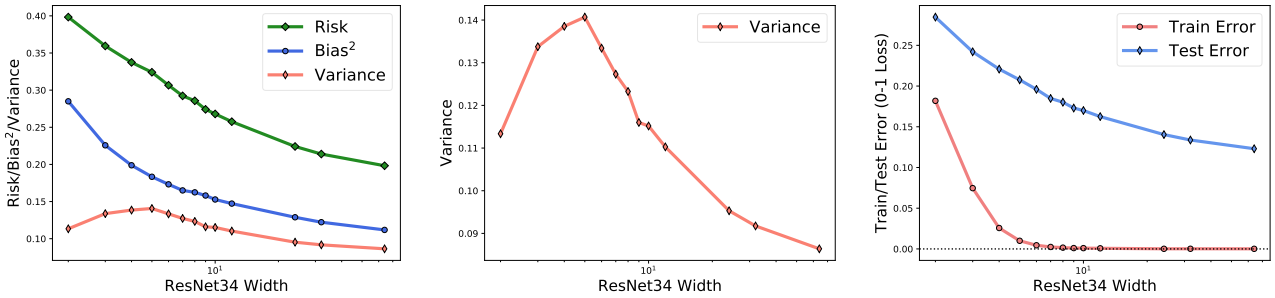


Figure 14. Risk, bias, variance, train/test error for ResNet34 trained by MSE loss on CIFAR10 dataset (10,000 training samples), the weight decay parameter of SGD is 1e-4. (**Left**) Risk, bias, and variance for ResNet34. (**Middle**) Variance for ResNet34. (**Right**) Train error and test error for ResNet34.

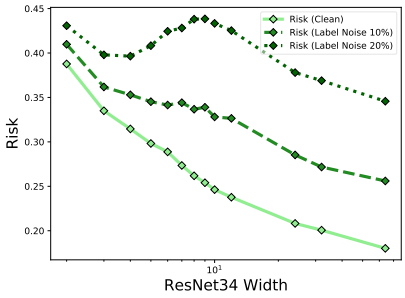


Figure 15. Risk under different label noise percentage. Increasing label noise leads to double descent risk curve.

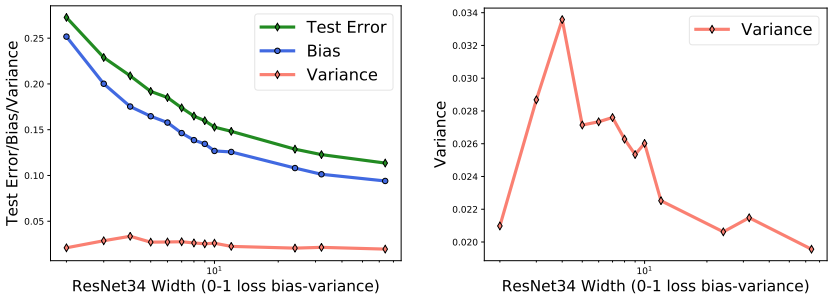


Figure 16. Bias-variance (0-1 loss), and test error for ResNet34 trained by MSE loss on CIFAR10 dataset (10,000 training samples). (**Left**) Bias and variance (0-1 loss), and test error for ResNet34. (**Right**) Variance (0-1 loss) for ResNet34.

B.8. Sources of Error for Mean Squared Error (MSE)

As argued in §2.2 the estimator for variance is unbiased estimator. To understand the variance of the estimator, we first split the data into two parts, A and B. For each part, we take multiple random splits (k) and estimate the variance by taking the average of those estimators, and vary the number of random splits k. The results are shown in Figure 17. We can see that the variation between to parts of data is small. Quantitatively, veraging across different model width, the relative difference between two parts of data is 0.65% for bias and 3.15% for variance.

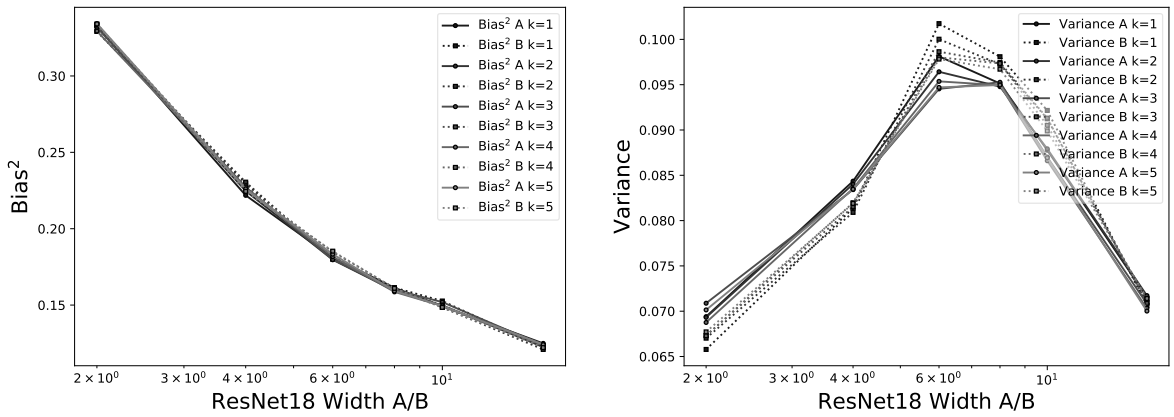


Figure 17. Bias and variance for two portions of data with k from 1 to 5. (**Left**) Bias for ResNet18. (**Right**) Variance for ResNet18.

B.9. Sources of Error for Cross Entropy Loss (CE)

For cross entropy loss, we are currently unable to obtain an unbiased estimator. We can access the quality of our estimator using the following scheme. We partition the dataset into five parts $\mathcal{T}_1, \ldots, \mathcal{T}_5$, i.e., set $N = 5$ in Algorithm 1. Then, we sequentially plot the estimate of bias and variance using $k = 1, 2, 3, 4$ as described in Algorithm 1. Using larger k gives better estimate. As shown in Figure 18, when k is small, our estimator over-estimate the bias and under-estimate the variance, but the overall behavior of the curves are consistent.

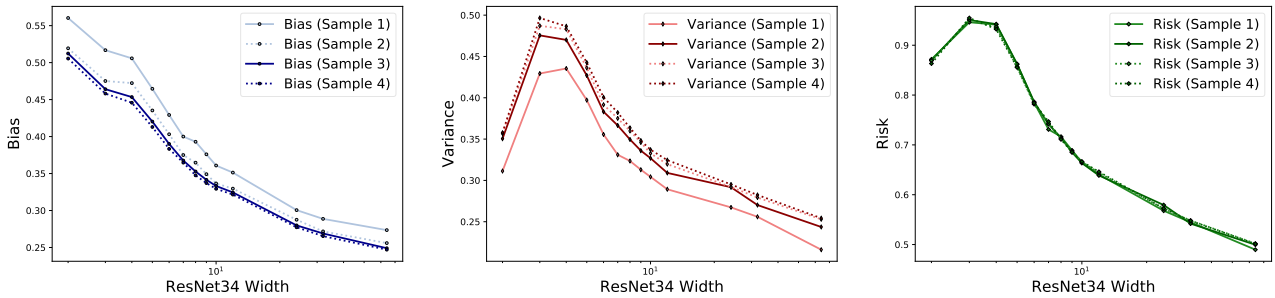


Figure 18. Estimate of bias, variance, and risk using varying number of sample (k in Algorithm 1). (**Left**) Bias (CE) for ResNet34. (**Middle**) Variance (CE) for ResNet34. (**Right**) Risk (CE) for ResNet34.

B.10. Effect of Depth on Bias and Variance for Out-Of-Distribution Data

We study the role of depth on out-of-distribution test data. In Figure 19, we observe that increasing the depth can decrease the bias and increase the variance. Also, deeper ResNet can generalize better on CIFAR10-C dataset as shown in Figure 19.

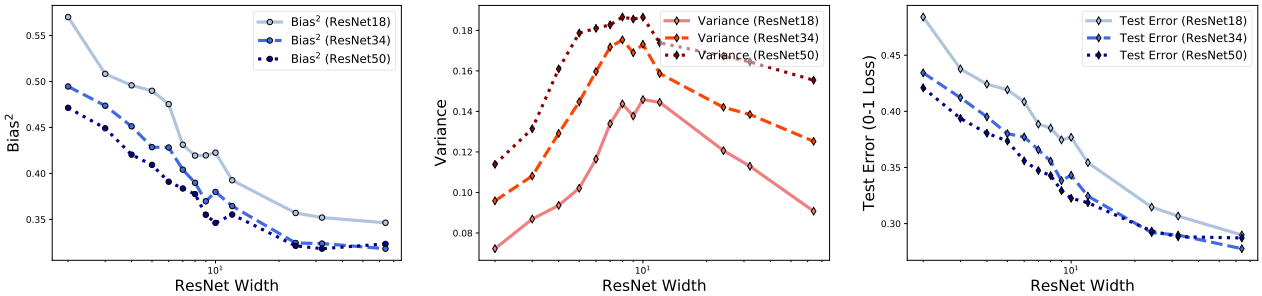


Figure 19. Bias, variance, and test error for ResNet with different depth (ResNet18, ResNet34 and ResNet50 trained by MSE loss on 25,000 CIFAR10 training samples) evaluated on out-of-distribution examples (CIFAR10-C dataset). (**Left**) Bias for ResNet18, ResNet34 and ResNet50. (**Middle**) Variance for ResNet18, ResNet34 and ResNet50. (**Right**) Test error for ResNet18, ResNet34 and ResNet50.

B.11. Effect of Depth on ResNet using Bottleneck Blocks

In order to study the role of depth for ResNet on bias and variance, we apply basic residual block for ResNet50. To better investigate the depth of ResNet, we use Bottleneck block for ResNet26, ResNet38, and ResNet50. More specifically, the number of 3-layer bottleneck blocks for ResNet26, ResNet38, and ResNet50 are $[2, 2, 2, 2]$, $[3, 3, 3, 3]$, and $[3, 4, 6, 3]$. As shown in Figure 20, we observe that deeper ResNet with Bottleneck blocks has lower bias and higher variance.

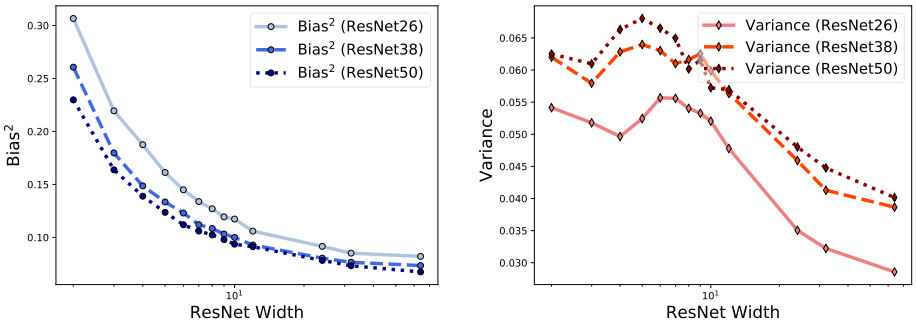


Figure 20. Bias and variance for ResNet (bottleneck block) with different depth. (**Left**) Bias for ResNet26, ResNet38 and ResNet50. (**Right**) Variance for ResNet26, ResNet38 and ResNet50.

B.12. Effect of Depth on VGG

We study the role of depth for VGG network on bias and variance. As shown in Figure 21, we observe that deeper VGG has lower bias and higher variance.

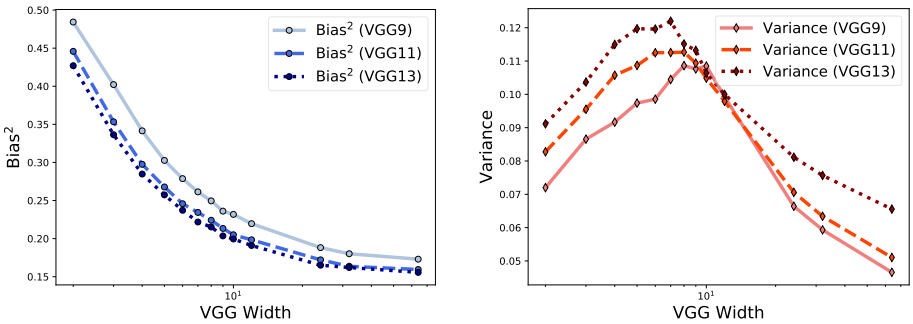


Figure 21. Bias and variance for VGG with different depth. (**Left**) Bias for VGG9, VGG11 and VGG13. (**Right**) Variance for VGG9, VGG11 and VGG13.

B.13. Additional Synthetic Experiment

In Figure 22, we plot the result of performing regression on synthetic data using a two-layer linear fully connected linear network with varying width. The data are generated as $y = \boldsymbol{\beta}^\top \boldsymbol{x}$, $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d/d)$, where $\|\boldsymbol{\beta}\|_2 = 1$ is randomly generated and fixed weight vector. The first layer of the network is drawn from i.i.d. zero-mean Gaussian distribution with variance $1/\sqrt{d}$, and the second layer is trained using gradient descent with weight decay 0.1. The horizontal axis is the number of parameters of the hidden layer normalized by the dimension of the data (i.e., p/d). The dots indicate actual experimental results, while the lines indicate theoretically predicted results. We can observe that they align well and the peak occurs at the predicted value.

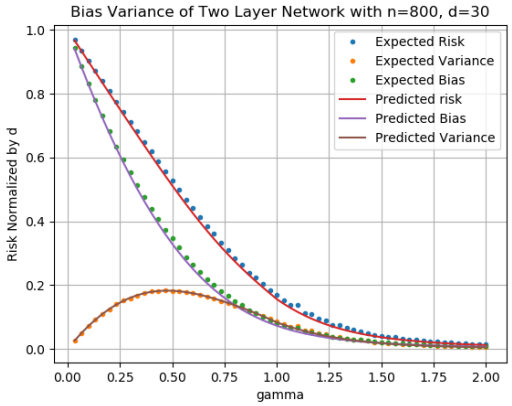


Figure 22. Bias, Variance, and Risk for two layer linear network with parameters $n = 800$ and $d = 30$.

C. Proof of Theorems in §5

Throughout this section, we use $\|\cdot\|$ and $\|\cdot\|_2$ to denote the Frobenius norm and spectral norm of a matrix, respectively. Recall that for any given $\boldsymbol{\theta}$, the training set $\mathcal{T} = (\boldsymbol{X}, \boldsymbol{y})$ satisfies the relation $\boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{\theta}$. By plugging this relation into (6), we get

$$f_\lambda(\boldsymbol{x}; \mathcal{T}, \boldsymbol{W}) = \boldsymbol{x}^\top \boldsymbol{M}_\lambda(\mathcal{T}, \boldsymbol{W})\boldsymbol{\theta}, \tag{11}$$

where we define

$$\boldsymbol{M}_\lambda(\mathcal{T}, \boldsymbol{W}) := \boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{W}^\top + \lambda \boldsymbol{I})^{-1}\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^\top. \tag{12}$$

To avoid cluttered notations, we omit the dependency of $\boldsymbol{M}$ on $\lambda, \mathcal{T}$ and $\boldsymbol{W}$.

By using (11), the expected bias and expected variance in (7) and (8) can be written as functions on the statistics of $\boldsymbol{M}$. This is stated in the following proposition. To proceed, we introduce the change of variable

$$\eta := \gamma^{-1} = \frac{d}{p}$$

in order to be consistent with conventions in random matrix theory.

Proposition 1 (Expected Bias/Variance)**.** *The expected bias and expected variance are given by*

$$\mathbb{E}\textbf{\textit{Bias}}_\lambda^2 = \frac{1}{d}\|\mathbb{E}\boldsymbol{M} - \boldsymbol{I}\|^2, \text{ and}$$
$$\mathbb{E}\textbf{\textit{Variance}}_\lambda = \frac{1}{d}\mathbb{E}\|\boldsymbol{M} - \mathbb{E}\boldsymbol{M}\|^2,$$

where $\boldsymbol{M}$ is defined in (12)*.*

Proof. By plugging (11) into (7), and using the prior that $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d/d)$ and $\boldsymbol{\theta} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, we get

$$\begin{aligned}
\mathbb{E}\textbf{Bias}_\lambda^2 &= \mathbb{E}\{\mathbb{E}(\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{x}^\top\boldsymbol{\theta}\}^2 \\
&= \mathbb{E}\left[\boldsymbol{x}^\top(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})\boldsymbol{\theta}\right]^2 \\
&= \mathbb{E}\boldsymbol{x}^\top(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})\boldsymbol{\theta}\boldsymbol{\theta}^\top(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})\boldsymbol{x} \\
&= \mathbb{E}\mathrm{tr}\Big[\boldsymbol{x}^\top(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})\boldsymbol{\theta}\boldsymbol{\theta}^\top(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})^\top\boldsymbol{x}\Big] \\
&= \mathrm{tr}\Big[(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)(\mathbb{E}\boldsymbol{M} - \boldsymbol{I})^\top\mathbb{E}(\boldsymbol{\theta}\boldsymbol{\theta}^\top)\Big] \\
&= \frac{1}{d}\|\mathbb{E}\boldsymbol{M} - \boldsymbol{I}\|^2.
\end{aligned}$$

Similarly, by plugging (11) into (8) we get

$$\begin{aligned}
&\mathbb{E}\textbf{Variance}_\lambda \\
&= \mathbb{E}\Big\{\mathbb{E}\big[(\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{x}^\top\boldsymbol{M}\theta|\boldsymbol{x}, \boldsymbol{\theta}))^2|\boldsymbol{x}, \boldsymbol{\theta}\big]\Big\} \\
&= \mathbb{E}\Big\{\mathbb{E}\big[(\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{\theta} - \boldsymbol{x}^\top(\mathbb{E}\boldsymbol{M})\boldsymbol{\theta})^2|\boldsymbol{x}, \boldsymbol{\theta}\big]\Big\} \\
&= \mathbb{E}(\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{\theta} - \boldsymbol{x}^\top(\mathbb{E}\boldsymbol{M})\boldsymbol{\theta})^2 \\
&= \mathbb{E}\Big[\boldsymbol{x}^\top(\boldsymbol{M} - \mathbb{E}\boldsymbol{M})\boldsymbol{\theta}\Big]^2 \\
&= \frac{1}{d}\mathbb{E}\|\boldsymbol{M} - \mathbb{E}\boldsymbol{M}\|^2.
\end{aligned}$$

□

The risk is given by

$$\mathbb{E}\textbf{Bias}_\lambda^2 + \mathbb{E}\textbf{Variance}_\lambda = \frac{1}{d}\mathbb{E}\|\boldsymbol{M} - \boldsymbol{I}\|^2 = \frac{1}{d}\mathbb{E}\mathrm{tr}(\boldsymbol{M}^\top\boldsymbol{M}) - \frac{2}{d}\mathbb{E}\mathrm{tr}(\boldsymbol{M}) + 1.$$

First, we show that in the asymptotic setting defined in Assumption 1, the expected Bias and expected Variance can be calculated as functions on the statistics of the following matrix:

$$\widetilde{\boldsymbol{M}}_{\lambda_0}(\boldsymbol{W}) = \boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}\boldsymbol{W}. \tag{13}$$

In the following, we omit the dependency of $\widetilde{\boldsymbol{M}}$ on λ_0 and $\boldsymbol{W}$.

Proposition 2 (Gap between $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$)**.** *Under Assumption 1 with $\lambda - \frac{n}{d}\lambda_0$, we have*

$$\frac{1}{d}\|\mathbb{E}\boldsymbol{M} - \boldsymbol{I}\|^2 = \frac{1}{d}\|\mathbb{E}\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2, \text{ and}$$
$$\frac{1}{d}\mathbb{E}\|\boldsymbol{M} - \boldsymbol{I}\|^2 = \frac{1}{d}\mathbb{E}\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2.$$

Proof. It suffices to show that $\|\boldsymbol{M} - \widetilde{\boldsymbol{M}}\|_2 = 0$ almost surely. From (12) and (13), we have

$$\boldsymbol{M} - \widetilde{\boldsymbol{M}} = \boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W} + \boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{\Delta} + \boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}\boldsymbol{W}\boldsymbol{\Delta},$$

where $\boldsymbol{\Delta} := (d/n)\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{I}$ and $\boldsymbol{\Omega} := (\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I} + \boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}^\top)^{-1} - (\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}$.

By using triangle inequality and the sub-multiplicative property of spectral norm, we have

$$\|\boldsymbol{M} - \widetilde{\boldsymbol{M}}\|_2 \le \|\boldsymbol{W}\|_2^2 \cdot \|\boldsymbol{\Omega}\|_2 + \|\boldsymbol{W}\|_2^2 \cdot \|\boldsymbol{\Omega}\|_2 \cdot \|\boldsymbol{\Delta}\|_2 + \|\widetilde{\boldsymbol{M}}\|_2 \cdot \|\boldsymbol{\Delta}\|_2. \tag{14}$$

Furthermore, by a classical result on the perturbation of matrix inverse (see e.g., Ghaoui (2002, equation (1.1))), we have

$$\|\boldsymbol{\Omega}\|_2 \le \|(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}\|_2^2\|\boldsymbol{W}\|_2^2\|\boldsymbol{\Delta}\|_2 + O(\|\boldsymbol{\Delta}\|_2^2).$$

Combining this bound with (14) gives

$$\|\boldsymbol{M} - \widetilde{\boldsymbol{M}}\|_2 \le \|\boldsymbol{W}\|_2^4 \cdot \|(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}\|_2^2 \cdot \|\boldsymbol{\Delta}\|_2 + \|\widetilde{\boldsymbol{M}}\|_2 \cdot \|\boldsymbol{\Delta}\|_2 + O(\|\boldsymbol{\Delta}\|_2^2).$$

It remains to show that $\|\boldsymbol{\Delta}\|_2 = 0$ and that $\|\boldsymbol{W}\|_2$, $\|(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}\|_2^2$, and $\|\widetilde{\boldsymbol{M}}\|_2$ are bounded from above almost surely. By Wainwright (2019, Example 6.2), $\forall\delta > 0$ and $n > d$,

$$\mathbb{P}\Big(\|\boldsymbol{\Delta}\|_2 \le 2\epsilon + \epsilon^2\Big) \ge 1 - e^{-n\delta^2/2}, \text{ where } \epsilon = \delta + \sqrt{\frac{d}{n}}.$$

By letting $\delta = \sqrt{d/n}$ and taking the asymptotic limit as in Assumption 1, we have

$$\|\boldsymbol{\Delta}\|_2 \overset{\text{a.s.}}{=} 0.$$

From Geman (1980), the largest eigenvalue of $\boldsymbol{W}\boldsymbol{W}^\top$ is almost surely $(1 + \sqrt{\eta})^2 < \infty$. Therefore, we have

$$\|\boldsymbol{W}\|_2 \overset{\text{a.s.}}{=} 1 + \sqrt{\eta} < \infty.$$

Finally, note that

$$\|(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0\boldsymbol{I})^{-1}\|_2 \le \frac{1}{\lambda_0 + \sigma_{\min}(\boldsymbol{W})^2} \le \frac{1}{\lambda_0} < \infty,$$
$$\|\widetilde{\boldsymbol{M}}\|_2 = \frac{\sigma_{\max}(\boldsymbol{W})^2}{\sigma_{\max}(\boldsymbol{W})^2 + \lambda_0} \le 1.$$

We therefore conclude that $\|\boldsymbol{M} - \widetilde{\boldsymbol{M}}\|_2 = 0$ almost surely, as desired. □

Proposition 3 (Asymptotic Risk)**.** *Given the expression for Bias and Variance in Proposition 1, under the asymptotic assumptions from Assumption 1,*

$$\frac{1}{d}\mathbb{E}\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2 = \begin{cases} (1 - \frac{1}{\eta}) + f_{\lambda_0^{-1}}(\frac{1}{\eta}), & \text{if } d > p, \\ f_{\lambda_0^{-1}}(\eta), & \text{if } d \le p, \end{cases}$$

where $\eta = d/p$, and for any $\eta, \alpha \in \mathbb{R}$,

$$f_\alpha(\eta) = \frac{\alpha + \eta(1 + \eta - 2\alpha + \eta\alpha)}{2\eta\sqrt{\eta^2 + 2\eta\alpha(1+\eta) + \alpha^2(1-\eta)^2}} - \frac{1-\eta}{2\eta}.$$

Proof. Recall that $\widetilde{\boldsymbol{M}} = \boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{W}^\top + \lambda_0 \boldsymbol{I})^{-1}\boldsymbol{W}$, by Sherman-Morrision,

$$\widetilde{\boldsymbol{M}} = \boldsymbol{I} - (\boldsymbol{I} + \lambda_0^{-1}\boldsymbol{W}^\top\boldsymbol{W})^{-1},$$

where $(d/p)\boldsymbol{W}^\top\boldsymbol{W} \in \mathbb{R}^{d\times d}$. Let $\lambda_i \geq 0, i = 1, \ldots, d$ be the eigenvalues of $(d/p)\boldsymbol{W}^\top\boldsymbol{W}$. For notational simplicity, let $\alpha = \lambda_0^{-1}$. Then

$$\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2 = \|[\boldsymbol{I} + (\alpha/\eta)(d/p)\boldsymbol{W}^\top\boldsymbol{W}]^{-1}\|^2 = \sum_{i=1}^{d} \frac{1}{(1+\frac{\alpha}{\eta}\lambda_i)^2}.$$

Let $\boldsymbol{A} = (d/p)\boldsymbol{W}^\top\boldsymbol{W}$, and $\mu_{\boldsymbol{A}}$ be the spectral measure of $\boldsymbol{A}$. Then

$$\frac{1}{d}\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2 = \int_{\mathbb{R}^+} \frac{1}{(1+\frac{\alpha}{\eta}x)^2} d\mu_{\boldsymbol{A}}(dx).$$

According to Marchenko-Pastur Law (Bai & Silverstein, 2010), in the limit when $d \to \infty$ when $\eta \leq 1$,

$$\frac{1}{d}\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|_F^2 \overset{\text{a.s.}}{=} \frac{1}{2\pi}\int_{\eta_-}^{\eta_+} \frac{\sqrt{(\eta_+ - x)(x-\eta_-)}}{\eta x(1+\frac{\alpha}{\eta}x)^2} dx,$$

where $\eta_+ = (1+\sqrt{\eta})^2$, and $\eta_- = (1-\sqrt{\eta})^2$. For convenience, define

$$f_\alpha(\eta) = \frac{1}{2\pi}\int_{\eta_-}^{\eta_+} \frac{\sqrt{(\eta_+ - x)(x-\eta_-)}}{\eta x(1+\frac{\alpha}{\eta}x)^2} dx = \frac{\alpha + \eta(1+\eta-2\alpha+\eta\alpha)}{2\eta\sqrt{\eta^2 + 2\eta\alpha(1+\eta) + \alpha^2(1-\eta)^2}} - \frac{1-\eta}{2\eta}.$$

When $\eta > 1$,

$$\frac{1}{d}\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|_F^2 = \left(1 - \frac{1}{\eta}\right) + f_\alpha\left(\frac{1}{\eta}\right).$$

Then, in the asymptotic regime,

$$\frac{1}{d}\|\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|_F^2 \overset{\text{a.s.}}{=} \begin{cases} (1-\frac{1}{\eta}) + f_\alpha(\frac{1}{\eta}), & \text{if } d > p, \\ f_\alpha(\eta), & \text{if } d < p. \end{cases}$$

□

Proposition 4 (Asymptotic Bias)**.** *Given the expression for Bias in Proposition 1, under the asymptotic assumptions in Assumption 1, the Bias for the model is given by*

$$\frac{1}{d}\|\mathbb{E}\boldsymbol{M} - \boldsymbol{I}\|^2 = \left[1 - \frac{\lambda_0\eta + (1+\eta) - \sqrt{\lambda_0^2\eta^2 + 2\lambda_0\eta(1+\eta) + (1-\eta)^2}}{2\eta}\right]^2.$$

Proof. Recall that

$$\boldsymbol{M} = \boldsymbol{W}^\top(\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{W}^\top + \lambda\boldsymbol{I})^{-1}\boldsymbol{W}\boldsymbol{X}\boldsymbol{X}^\top.$$

Recall that $\widetilde{\boldsymbol{M}} = \boldsymbol{I} - (\boldsymbol{I} + \lambda_0^{-1}\boldsymbol{W}^\top\boldsymbol{W})^{-1}$. Thus

$$\frac{1}{d}\|\mathbb{E}\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2 = \frac{1}{d}\|\mathbb{E}(\boldsymbol{I} + \lambda_0^{-1}\boldsymbol{W}^\top\boldsymbol{W})^{-1}\|^2.$$

By Neumann series,

$$\mathbb{E}(\boldsymbol{I} + \lambda_0^{-1}\boldsymbol{W}^\top\boldsymbol{W})^{-1} = \sum_{m\geq 0}\mathbb{E}(-\lambda_0^{-1}\boldsymbol{W}^\top\boldsymbol{W})^m = \boldsymbol{I} + \sum_{m\geq 1}(-1)^m(\lambda_0\eta)^{-m}\mathbb{E}\boldsymbol{A}^m,$$

where $\eta = d/p, \boldsymbol{A} = (d/p)\boldsymbol{W}^\top\boldsymbol{W}$. According to Corollary 3.3 in Bishop et al. (2018) (recall we are considering the asymptotic regime of $d, p \to \infty$),

$$\mathbb{E}\boldsymbol{A}^m = \sum_{k=1}^{m}\eta^{m-k}N_{m,k}\cdot\boldsymbol{I},$$

where

$$N_{m,k} = \frac{1}{k}\binom{m-1}{k-1}\binom{m}{k-1}$$

is the Narayana number. Therefore,

$$\frac{1}{d}\|\mathbb{E}\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|^2 = \left(1 + \eta^{-1}\sum_{m=1}^{\infty}\sum_{k=1}^{k}(-\lambda_0^{-1})^m(\eta^{-1})^{k-1}N_{m,k}\right)^2.$$

Observe that the double sum in the previous equation is just the generating series for the Narayana number,

$$\sum_{m=1}^{\infty}\sum_{k=1}^{k}(-\lambda_0^{-1})^m(\eta^{-1})^{k-1}N_{m,k} = -\frac{\lambda_0\eta + (1+\eta) - \sqrt{\lambda_0^2\eta^2 + 2\lambda_0\eta(1+\eta) + (1-\eta)^2}}{2\eta}.$$

This completes the proof. □

Finally, the statement of Theorem 1 follows directly from the above propositions.