Asymptotic Normality and Confidence Intervals for Prediction Risk of the Min-norm Least Squares Estimator

Zeng Li[1] **Chuanlong Xie**[2] **Qinwen Wang**[3]

Abstract

This paper quantifies the uncertainty of prediction risk for the min-norm least squares estimator in high-dimensional linear regression models. We establish the asymptotic normality of prediction risk when both the sample size and the number of features tend to infinity. Based on the newly established central limit theorems (CLTs), we derive the confidence intervals of the prediction risk under various scenarios. Our results demonstrate the sample-wise non-monotonicity of the prediction risk and confirm *"more data hurt"* phenomenon. Furthermore, the width of confidence intervals indicates that over-parameterization would enlarge the randomness of prediction performance.

1. Introduction

One major surprise of deep learning models is their accurate predictive performance while achieving zero training error (Zhang et al., 2016). This observation contradicts with the common wisdom of bias-variance trade-off when model complexity increases (Van der Vaart, 2000; Friedman et al., 2001). The theory of over-parameterization is a rapidly growing area, which makes attempts to explain the empirical success of large scale models in deep learning. Quite a few papers are trying to understand over-parameterized models from generalization perspective, e.g. Advani & Saxe (2017); Belkin et al. (2018; 2019a;b); Geiger et al. (2019); Spigler et al. (2019); Bartlett et al. (2020); Muthukumar et al. (2020).

As model capacity increases, the generalization error first decreases and then increases, and decreases again. Such a phenomenon has been summarized in the *"double descent"* curve, i.e. Fig 1B. in Belkin et al. (2019a). It subsumes the classical bias-variance trade-off, a U-shape curve, and further shows that the prediction error exhibits a second drop when the model capacity exceeds the interpolation threshold, which is the so-called over-parameterized settings. The second drop has been broadly quantified for certain parametric and non-parametric models, including linear model, nearest neighbours algorithm, random-feature model and one-hidden-layer neural network (Hastie et al., 2019; Xing et al., 2019; Mei & Montanari, 2019; Ba et al., 2019). However, the existing theoretical studies mostly focus on the first-order limit of prediction risk. The randomness caused by sampling remains unclear. On the other hand, the finite-sample results, e.g. Dereziński et al. (2019), only consider the expectation of the risk while ignoring the variance. In summary, the current existing results can capture the global trend of the prediction risk in terms of first-order limit and expectation. However, there is no information provided on the fluctuation or randomness around this limit. In the finite sample situations, how far away the empirical prediction risk is from its limit remains a mystery.

To fill this gap, this paper makes the first step toward the uncertainty quantification for the prediction risk, especially in the widely concerned over-parameterized settings. We consider a linear regression task with n data points, p features and develop CLTs for the prediction risk as $n, p \to +\infty$ and $p/n \to c \in (0, \infty)$. The main goal of this paper is to study the second-order asymptotic behaviors or CLTs of two different types of conditional prediction risk for the min-morn least squares estimator. One is $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ given both the training data $\mathbf{X}$ and regression coefficient $\boldsymbol{\beta}$ while the other is $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ given the training data $\mathbf{X}$ only. We summarize our main results as follows:

(1) The regression coefficient is assumed to be either random or nonrandom to cover more cases. Asymptotic normality and limiting distributions of prediction risk are proved and derived under various scenarios.

(2) Finite-sample distributions of the conditional prediction risk given both the training data and regression coefficient are derived and characterized in Theorem 4.2 and 4.5. Under certain assumptions, the *"more data hurt"* phenomenon can be confirmed by comparing the

[1] First author, Department of Statistics and Data Science, Southern University of Science and Technology, China [2] Huawei Noah's Ark Lab, Hong Kong [3] Corresponding author, School of Data Science, Fudan University, China. Correspondence to: Qinwen Wang <wqw@fudan.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

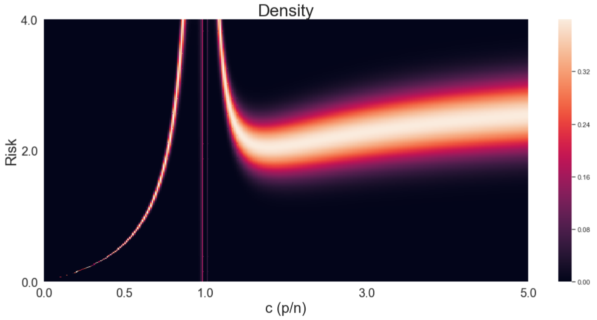

Figure 1. The heat map of asymptotic conditional density of the prediction risk $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ when c varies from 0 to 5. The density functions are plotted according to Theorem 4.2 for $c < 1$ and Theorem 4.5 for $c \geq 1$. We take $p = 100$, the signal strength $r^2 = 3$ and the noise level $\sigma^2 = 0.75$.

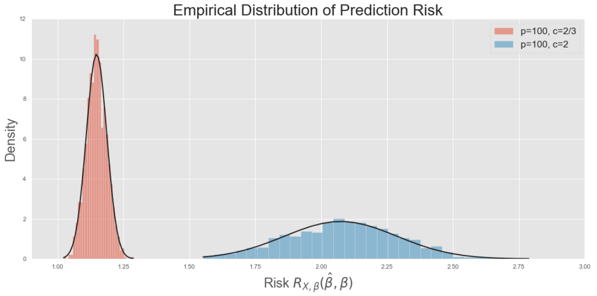

Figure 2. Histogram for empirical distribution of prediction risk $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, the black curves are corresponding to the theoretical limiting distributions we derived. We take $r^2 = 3$, $\sigma^2 = 0.75$, $p = 100$ and $n = 150, 50$, respectively.

confidence intervals built via the CLTs we established.

(3) Our results incorporate non-Gaussian observations. For Gaussian data, the limiting mean and variance in the CLTs have simpler forms, see Section 4.2 and 4.3 for more details.

Figure 1 presents part of our CLT results. The x-axes represent model complexity, i.e. the value $c = p/n$. For each c, we draw the heat map of the asymptotic conditional density function of the corresponding prediction risk given in Theorems 4.2 and 4.5. In particular, one can find out that in the over-parameterized regime ($c > 1$), the density of the prediction risk does not concentrate on its first-order limit, but fluctuates within a wide band around it. It indicates that the randomness of the prediction risk is enlarged around its limit when c becomes large. Figure 2 draws the empirical densities of the prediction risk when $p = 100$ and $n = 150, 50$, respectively. Not only the empirical densities fit their theoretical counterparts perfectly, their different span further demonstrates the growing variance in over-parameterized cases, which confirms the findings in Figure 1 as well.

One application of our theoretical results is to explain the *"more data hurt"* phenomenon. *"More data hurt"* describes that training on more data may hurt the prediction performance of the learned model, especially for some deep learning tasks (Nakkiran et al., 2019). This concept is closely related to the second drop of the prediction risk in the *"double descent"* curve, where the over-parameterized regime ($p \geq n$) is concerned. In this regime, once the sample size increases (training on more data), the degree of overparameterization decreases and becomes closer to the interpolation boundary $p = n$ in Hastie et al. (2019). In this way, the first-order limit of the prediction risk increases according to the *"double descent"* curve and confirms the *"more data hurt"* phenomenon. However, the behaviour of the empirical prediction risk can not be fully represented by its first-order limit only. There is a non-negligible discrepancy between the finite sample prediction risk and its limit, especially when the sample size or dimension is small. We take Figure 3 as an example to illustrate this. For a fixed dimension $p = 100$, the first-order limits of prediction risk at sample size $n_1 = 65$, $n_2 = 75$ and $n_3 = 85$ are 2.09, 2.44 and 3.64 respectively (Hastie et al., 2019). *"More data hurt"* seems true because, from the first-order asymptotics, larger sample size corresponds to smaller model complexity and hence larger prediction risk. However, if we look at the confidence intervals for the prediction risk, which can be derived from our second-order asymptotic results (see Theorem 4.5), the 95% confidence interval of the prediction risk at sample size $n_1 = 65$ is $[1.62, 2.63]$, it overlaps with the 95% CI $[2.04, 2.97]$ at $n_2 = 75$, while keeps a distance from the 95% CI $[3.13, 4.57]$ at $n_3 = 85$. Hence the increment of risk from $n_1 = 65$ to $n_2 = 75$ is not statistically significant, but the increment from $n_1 = 65$ to $n_3 = 85$ is. Therefore, only knowing the first-order limit is not enough to illustrate the *"more data hurt"* phenomenon. Fine-grained second-order results are needed to fully characterize the discrepancy between the empirical prediction risk and its limit. More importantly, a confidence band for the prediction risk can be constructed to evaluate its finite sample performance and distinguish the statistical significance of *"more data hurt"* phenomenon based on our newly established CLT results.

The rest of this paper is organized as follows. Section 2 contains some related work. Section 3 introduces in detail our model settings and two different prediction risk. Section 4 presents the main results on CLTs for the two types of risk with some discussion. Section 5 conducts simulation experiments to verify our main results. All the technical proofs and lemmas are relegated to the Appendix in the supplementary file.

2. Related work

First-Order Limit As our results are based on the linear regression, we mainly focus on the literature of linear mod-

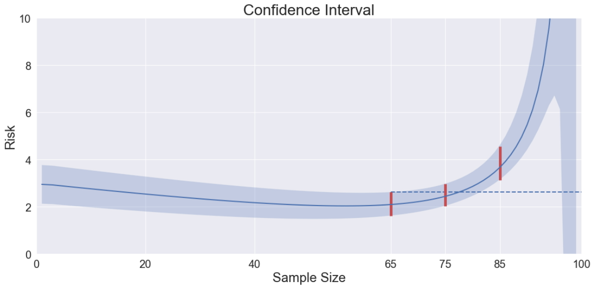

Figure 3. Sample-wise non-monotonicity and the 95%-confidence band (point-wise) of the prediction risk when sample size varies from 1 to 99. We take $r^2 = 3$, $\sigma^2 = 0.75$ and $p = 100$. The 95% CI of the prediction risk for $n_1 = 65$ overlaps with that for $n_2 = 75$, but is separated from the 95% CI for $n_3 = 85$.

els. Hastie et al. (2019) gives the first-order limit of the generalization error for linear regressions as $n, p \to +\infty$. Dereziński et al. (2019) provides an exact non-asymptotic expression for *“double descent”* of the min-norm least squares estimator. Wu & Xu (2020) extends the first-order limit of the prediction error of the generalized weighted ridge estimator to a more general setting with anisotropic features and signals. Montanari et al. (2019), Deng et al. (2019) and Kini & Thrampoulidis (2020) investigate the sharp asymptotic of binary classification tasks with the max-margin and maximum likelihood solution. Emami et al. (2020) and Gerbelot et al. (2020a) consider the *“double descent”* in generalized linear models. The *“double descent”* phenomenon is also observed on linear tasks with various problems and assumptions, e.g. LeJeune et al. (2020); Gerbelot et al. (2020b); Javanmard et al. (2020); Dar & Baraniuk (2020); Xu & Hsu (2019); Dar et al. (2020).

Second-Order Fluctuation There are very few second-order results in the literature. Shen & Bellec (2020) establishes the asymptotic normality for the derivatives of random-feature model, but not the exact limiting distribution of the risk. We are the first to develop results on second-order fluctuations of the prediction risk in linear regressions and provide its corresponding confidence intervals in this work.

More Data Hurt Loog et al. (2019) shows that various standard learners can lead to sample-wise non-monotonicity. Nakkiran et al. (2019) experimentally confirms the sample-wise non-monotonicity of the test accuracy on deep neural networks. This challenges the conventional understanding in large sample asymptotics: if an estimate is consistent, more data will make this estimate more stable and improves its finite-sample performance. Nakkiran (2019) considers adding one single data point to a linear regression task and analyzes its marginal effect to the test risk. Dereziński et al. (2019) gives an exact non-asymptotic risk of the min-norm least squares estimator and confirms the sample-wise non-monotonicity on mean square error. For adversarially robust models, Min et al. (2020) proves that more data may increase the gap between the generalization error of adversarially-trained models and standard models. Chen et al. (2020) shows that more training data causes the generalization error to increase in the strong adversary regime.

Random Matrix Theory The primary tool for analyzing the second-order fluctuations of prediction risk comes from random matrix theory. In particular, Bai & Silverstein (2004) refines the CLTs for linear spectral statistics of large dimensional sample covariance matrix with general population and the population is not necessarily to be Gaussian. Similar CLTs are also developed for other random matrix ensembles, see Sinai & Soshnikov (1998); Bai & Yao (2005); Zheng (2012). Other than the CLTs for linear spectral statistics, Bai et al. (2007) and Pan & Zhou (2008) study the asymptotic fluctuation of eigenvectors of sample covariance matrices. Bai & Yao (2008) considers the fluctuation of quadratic forms. All these technical tools and results are adopted and fully utilized, especially those related to Stieltjes transform, which are closely connected to the prediction risk studied in this paper.

3. Preliminaries

3.1. Problem, data and estimator

Suppose that the training data $\{(\mathbf{x}_i, \mathrm{y}_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, 2, \dots, n\}$ is generated independently from the model (ground truth or teacher model):

$$\mathrm{y}_i = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i + \epsilon_i, \quad \text{and} \quad (\mathbf{x}_i, \epsilon_i) \sim (P_{\mathbf{x}}, P_\epsilon). \tag{1}$$

Here, $P_{\mathbf{x}}$ is a distribution on $\mathbb{R}^p$ such that $\mathbb{E}(\mathbf{x}_i) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}$, and P_ϵ is a distribution on $\mathbb{R}$ such that $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$. In particular, the coordinates of $\mathbf{x}_i$ are not necessarily independent, that is, $\boldsymbol{\Sigma}$ is not restricted to be diagonal. To proceed further, we denote

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^{\mathrm{T}}, \quad \mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_n)^{\mathrm{T}}.$$

The minimum ℓ_2 norm (min-norm) least squares estimator, of $\mathbf{y}$ on $\mathbf{X}$, is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{2}$$

where $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{+}$ denotes the Moore-Penrose pseudoinverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

3.2. Bias, variance and risk

Similar to Hastie et al. (2019), we define two different types of out-of-sample prediction risk. The first one is given by

$$\begin{aligned} R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \mathbb{E}\big[(\mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta})^2 \big| \mathbf{X}\big] \\ &= \mathbb{E}\big[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 \big| \mathbf{X}\big], \end{aligned} \tag{3}$$

where $\mathbf{x}_0 \sim P_{\mathbf{x}}$ is a test point and is independent of the training data, and the notation $\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2$ stands for $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}$. Here $\boldsymbol{\beta}$ is assumed to be a random vector independent of $\mathbf{x}_0$. In this definition, the expectation $\mathbb{E}$ stands for the conditional expectation for $\mathbf{x}_0$, $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ when $\mathbf{X}$ is given. According to the bias-variance decomposition, we have $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) := B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) + V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, where

$$\begin{aligned} B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \mathbb{E}\Big\{\|\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 \big| \mathbf{X}\Big\}, \\ V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \mathrm{Tr}\{\mathrm{Cov}(\hat{\boldsymbol{\beta}}|\mathbf{X})\boldsymbol{\Sigma}\}. \end{aligned}$$

Plugging the model (1) into the min-norm estimator (2), the bias and variance terms can be rewritten as

$$\begin{aligned} B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \mathbb{E}\{\boldsymbol{\beta}^{\mathrm{T}}\Pi\boldsymbol{\Sigma}\Pi\boldsymbol{\beta}\big|\mathbf{X}\}, \\ V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \frac{\sigma^2}{n}\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{+}\boldsymbol{\Sigma}), \end{aligned}$$

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

The second type of out-of-sample prediction risk is defined as

$$\begin{aligned} R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \mathbb{E}\big[(\mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta})^2\big|\mathbf{X}, \boldsymbol{\beta}\big] \\ &= \mathbb{E}\big[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2\big|\mathbf{X}, \boldsymbol{\beta}\big], \end{aligned} \quad (4)$$

where

$$\begin{aligned} B_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \boldsymbol{\beta}^{\mathrm{T}}\Pi\boldsymbol{\Sigma}\Pi\boldsymbol{\beta}, \\ V_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{\sigma^2}{n}\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{+}\boldsymbol{\Sigma}). \end{aligned}$$

In this definition, the parameter $\boldsymbol{\beta}$ is assumed to be given. The expectation $\mathbb{E}$ is the conditional expectation for $\mathbf{x}_0$ and $\hat{\boldsymbol{\beta}}$ when $\mathbf{X}$ and $\boldsymbol{\beta}$ are given. This is consistent with the commonly-used testing procedure, in which a trained model is evaluated by the average loss on those unseen testing data.

4. Main Results

Before stating our main results, we briefly highlight the challenges we faced in proving the *"more data hurt"* phenomenon. First, the finite-sample behaviors of the prediction risk is required. Hastie et al. (2019) gives the first-order limits of both $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ as $n, p \to +\infty$ and $p/n \to c \in (0, +\infty)$. However, to prove the *"more data hurt"* phenomenon, we should fix p and investigate the finite-sample risk with sample size n varies. This implies that only knowing the first-order limit is not enough, the convergence rate is also needed. To solve this problem, we have derived the CLTs for both $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, respectively, which characterize the second-order fluctuations of the risk. Then we can figure out the finite-sample behavior of the risk by computing the gap between the risk and its limit. The confidence intervals of the risk can be further obtained. Second, the parameter $\boldsymbol{\beta}$ also contributes randomness to the finite-sample risk, which further influences the convergence rate. To analyze the contribution of $\boldsymbol{\beta}$, we need to make use of the technical tools and asymptotic results for eigenvectors and quadratic forms developed in Bai et al. (2007) and Bai & Yao (2008). Another interesting finding is that, in the over-parameterized regime such that $p > n$, the two types of out-of-sample prediction risk $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ enjoy different convergence rates.

4.1. Assumptions and more notations

As follows are some notations used in this paper. The $p \times p$ identity matrix is denoted by $\boldsymbol{I}_p$. For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, we define its *empirical spectral distribution (ESD)* as

$$F^{\boldsymbol{A}}(x) = \frac{1}{p}\sum_{i=1}^{p}\mathbb{1}\{\lambda_i(\boldsymbol{A}) \leq x\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function and $\lambda_i(\boldsymbol{A})$, $i = 1, 2, \dots p$ are the eigenvalues of $\boldsymbol{A}$. The notation $\xrightarrow{d}$ stands for the convergence in distribution. $Z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution, $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote the largest and smallest eigenvalues of $\boldsymbol{A}$, respectively.

Here we list all the assumptions for $\mathbf{X}$, ε_i and $\boldsymbol{\beta}$ needed under different scenarios:

(A) $\mathbf{x}_j \sim P_{\mathbf{x}}$ is of the form $\mathbf{x}_j = \boldsymbol{\Sigma}^{1/2}\mathbf{z}_j$, where $\mathbf{z}_j$ is a p-dimensional random vector with i.i.d. entries, having zero mean, unit variance and finite 4-th moment $\mathbb{E}(\mathbf{z}_{ij}^4) = \nu_4$, $i = 1, \cdots, p$, $j = 1, \cdots, n$. $\varepsilon_i \sim P_\epsilon$ satisfies $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ and is independent of $\mathbf{X}$.

(B1) $\boldsymbol{\Sigma}$ is a deterministic positive definite matrix, such that $0 < C_0 \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C_1$ for all n, p and some constants C_0, C_1. As $p \to \infty$, we assume that the empirical spectral distribution $F^{\boldsymbol{\Sigma}}$ converges weakly to a probability measure H.

(B2) $\boldsymbol{\Sigma}$ is an identity matrix, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

(C1) $\boldsymbol{\beta}$ is a nonrandom constant vector, and $\|\boldsymbol{\beta}\|_2^2 = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = r^2$.

(C2) $\boldsymbol{\beta} \sim P_{\boldsymbol{\beta}}$ is independent of $\mathbf{X}$ and follows multivariate Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \frac{r^2}{p}\boldsymbol{I}_p)$.

Throughout this paper, we consider the limiting distributions and the convergence rates of the out-of-sample prediction risk when $n, p \to \infty$ such that $p/n = c_n \to c \in (0, \infty)$. If $c > 1$, the sample size n is smaller than the number of parameters p, we call this case *"over-parameterized"*. Otherwise when $c < 1$, we call it *"under-parameterized"*.

It's worth mentioning that in the under- parametrized case, Hastie et al. (2019) requires the finite 4-th moment as in our Assumption (A) as well. However, in the over-parametrized case, Hastie et al. (2019) considers more general framework allowing anisotropic features and hence requires higher moment conditions than ours. For isotropic features, our main technical tools are inherited from the CLTs for linear spectral statistics (Bai & Silverstein, 2004; Pan & Zhou, 2008), for eigenvectors (Bai et al., 2007) and for quadratic forms (Bai & Yao, 2008) of sample covariance matrices. The finite 4-th moment is sufficient for all these results.

4.2. Under-parametrized asymptotics

In this section, we focus on the risk of the min-norm estimator (2) in the under-parametrized regime. According to Theorem 1 of Hastie et al. (2019), both $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ converge to $\sigma^2 c/(1-c)$ almost surely. The following Theorem 4.1 and 4.2 show that both $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ converge to $\sigma^2 c/(1-c)$ at the rate $1/p$. Furthermore, the limiting distributions are derived by making use of the CLTs for linear spectral statistics of large-dimensional sample covariance matrices. To proceed further, we denote

$$R_{c_n} = \begin{cases} \frac{c_n}{1-c_n}\sigma^2, & \text{if } c_n < 1, \\ (1-\frac{1}{c_n})r^2 + \frac{1}{c_n - 1}\sigma^2, & \text{if } c_n > 1, \end{cases} \tag{5}$$

which is the finite counterpart for the first-order limit of $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ as $n, p \to \infty$ such that $p/n = c_n \to c$.

Theorem 4.1. *Suppose that the training data is generated from the model* (1)*, and the assumptions* (A) *and* (B1) *hold. Then the first type of out-of-sample prediction risk* $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ *of the min-norm estimator* (2) *satisfies that, as* $n, p \to \infty$ *such that* $p/n = c_n \to c < 1$,

$$p\Big(R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - R_{c_n}\Big) \xrightarrow{d} N(\mu_c, \sigma_c^2), \tag{6}$$

where

$$\begin{aligned} \mu_c &= \frac{c^2\sigma^2}{(c-1)^2} + \frac{\sigma^2 c^2(\nu_4 - 3)}{1-c} \\ \sigma_c^2 &= \frac{2c^3\sigma^4}{(c-1)^4} + \frac{c^3\sigma^4(\nu_4-3)}{(1-c)^2}. \end{aligned}$$

Conclusively,

$$P(L_{\alpha,c} \le R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le U_{\alpha,c}) \to 1 - \alpha, \tag{7}$$

where $1-\alpha$ *is the confidence level and*

$$\begin{aligned} L_{\alpha,c} &= \frac{c_n\sigma^2}{1-c_n} + \frac{1}{p}(\mu_c - Z_{\alpha/2}\sigma_c), \\ U_{\alpha,c} &= \frac{c_n\sigma^2}{1-c_n} + \frac{1}{p}(\mu_c + Z_{\alpha/2}\sigma_c). \end{aligned}$$

Under the assumptions of Theorem 4.1, we know that $B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = B_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 0$ and

$$V_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = V_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{\sigma^2}{n}\operatorname{Tr}(\hat{\boldsymbol{\Sigma}}^{+}\boldsymbol{\Sigma}).$$

Thus $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ equals to $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and the two risk share the same asymptotic limit.

Theorem 4.2. *Under the assumptions of Theorem 4.1, the second type of out-of-sample prediction risk* $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ *of the min-norm estimator* (2) *satisfies that, as* $n, p \to \infty$ *such that* $p/n = c_n \to c < 1$,

$$p\big(R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - R_{c_n}\big) \xrightarrow{d} N(\mu_c, \sigma_c^2),$$

and

$$P(L_{\alpha,c} \le R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le U_{\alpha,c}) \to 1 - \alpha,$$

where μ_c, σ_c^2, $L_{\alpha,c}$ *and* $U_{\alpha,c}$ *are the same as those in Theorem 4.1.*

4.3. Over-parameterized asymptotics

In this section, we consider the min-norm estimator (2) in the over-parameterized case $c > 1$. The bias term, either $B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ or $B_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, is generally nonzero. According to Lemma 2 in Hastie et al. (2019), both $B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $B_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ converge to $r^2(1-1/c)$ as $n, p \to +\infty$ and $p/n \to c > 1$. This implies that the bias term can influence the asymptotic behavior of the prediction risk, including the convergence rate. Hence to derive the CLT of the out-of-sample prediction risk, we need to consider both the bias and variance terms in (3) and (4).

In the following, we investigate the asymptotic properties of the two prediction risk $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ under various combinations of the assumptions (A), (B2) for $\mathbf{X}$ and scenarios (C1), (C2) for both random and nonrandom $\boldsymbol{\beta}$. We start with the case when $\boldsymbol{\beta}$ is a constant vector.

Theorem 4.3. *Suppose that the training data is generated from the model (1), and the assumptions* (A), (B2) *and* (C1) *hold. Then the first type of out-of-sample prediction risk* $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ *of the min-norm estimator* (2) *satisfies that, as* $n, p \to \infty$ *such that* $p/n = c_n \to c > 1$,

$$\sqrt{p}\Big\{R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - R_{c_n}\Big\} \xrightarrow{d} N(\mu_{c,1}, \sigma_{c,1}^2), \tag{8}$$

where $\mu_{c,1} = 0$ and $\sigma^2_{c,1} = \frac{2(c-1)}{c^2} r^4$. A more practical version is to replace $\mu_{c,1}$ and $\sigma^2_{c,1}$ with

$$
\begin{aligned}
\tilde{\mu}_{c,1} &= \frac{1}{\sqrt{p}}\Big\{\frac{c\sigma^2}{(1-c)^2} + \frac{\sigma^2(\nu_4-3)}{c-1}\Big\} \\
\tilde{\sigma}^2_{c,1} &= \frac{2(c-1)}{c^2} r^4 + \frac{1}{p}\Big\{\frac{2c^3\sigma^4}{(1-c)^4} + \frac{c\sigma^4(\nu_4-3)}{(c-1)^2}\Big\}.
\end{aligned}
$$

Conclusively,

$$
P(L_{\alpha,c} \le R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le U_{\alpha,c}) \to 1-\alpha, \tag{9}
$$

where $1-\alpha$ is the confidence level and

$$
\begin{aligned}
L_{\alpha,c} &= (1-\frac{1}{c_n})r^2 + \frac{\sigma^2}{c_n-1} + \frac{1}{\sqrt{p}}(\tilde{\mu}_{c,1} - Z_{\alpha/2}\tilde{\sigma}_{c,1}), \\
U_{\alpha,c} &= (1-\frac{1}{c_n})r^2 + \frac{\sigma^2}{c_n-1} + \frac{1}{\sqrt{p}}(\tilde{\mu}_{c,1} + Z_{\alpha/2}\tilde{\sigma}_{c,1}).
\end{aligned}
$$

Remark 4.1. *Under the assumption (C1),*

$$
B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = B_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}), \quad R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}).
$$

Thus Theorem 4.3 still holds if we replace $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ with $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$.

Remark 4.2. *Under Assumption (B2), the eigenvector of $\hat{\boldsymbol{\Sigma}}$ is asymptotically Haar distributed. Therefore, the bias term $B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ is only related to the length of $\boldsymbol{\beta}$. However, in the anisotropic settings with general $\boldsymbol{\Sigma}$, the eigenvector of the $\hat{\boldsymbol{\Sigma}}$ is no longer asymptotically Haar distributed. The limiting behavior of $B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ heavily relies on the interaction between $\boldsymbol{\beta}$ and the eigenvectors of $\hat{\boldsymbol{\Sigma}}$. Therefore, we conjecture that there is no universal convergence rate for the bias term $B_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ that can cover arbitrary non-random $\boldsymbol{\beta}$ and anisotropic $\boldsymbol{\Sigma}$ in the over-parameterized case, not to mention the prediction risk $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. A small simulation experiment is conducted in Appendix to confirm our conjecture on this point.*

Next we consider the case when $\boldsymbol{\beta}$ is a random vector that follows assumption (C2).

Theorem 4.4. *Suppose that the training data is generated from the model* (1)*, and the assumptions* (A), (B2) *and* (C2) *hold. Then as $n, p \to \infty$ such that $p/n = c_n \to c > 1$, the first type of out-of-sample prediction risk $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ of the min-norm estimator* (2) *satisfies,*

$$
p\Big\{R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - R_{c_n}\Big\} \xrightarrow{d} N(\mu_{c,2}, \sigma^2_{c,2}),
$$

where

$$
\begin{aligned}
\mu_{c,2} &= \frac{c\sigma^2}{(1-c)^2} + \frac{\sigma^2(\nu_4-3)}{c-1}, \\
\sigma^2_{c,2} &= \frac{2c^3\sigma^4}{(1-c)^4} + \frac{c\sigma^4(\nu_4-3)}{(c-1)^2}.
\end{aligned}
$$

Hence we have

$$
P(L_{\alpha,c} \le R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le U_{\alpha,c}) \to 1-\alpha,
$$

where

$$
\begin{aligned}
L_{\alpha,c} &= \frac{\sigma^2}{c_n-1} + (1-\frac{1}{c_n})r^2 + \frac{1}{p}(\mu_{c,2} - Z_{\alpha/2}\sigma_{c,2}), \\
U_{\alpha,c} &= \frac{\sigma^2}{c_n-1} + (1-\frac{1}{c_n})r^2 + \frac{1}{p}(\mu_{c,2} + Z_{\alpha/2}\sigma_{c,2}).
\end{aligned}
$$

As for $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, we have the following theorem.

Theorem 4.5. *Suppose that the training data is generated from the model* (1)*, and the assumptions* (A), (B2) *and* (C2) *hold. Then, as $n, p \to \infty$ such that $p/n = c_n \to c > 1$, the second type of out-of-sample prediction risk $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ of the min-norm estimator* (2) *satisfies,*

$$
\sqrt{p}\Big\{R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - R_{c_n}\Big\} \xrightarrow{d} N(\mu_{c,3}, \sigma^2_{c,3}), \tag{10}
$$

where $\mu_{c,3} = 0$ and $\sigma^2_{c,3} = 2(1-\frac{1}{c})r^4$. A more practical version is to replace $\mu_{c,3}$ and $\sigma^2_{c,3}$ with

$$
\begin{aligned}
\tilde{\mu}_{c,3} &= \frac{1}{\sqrt{p}}\left\{\frac{c\sigma^2}{(1-c)^2} + \frac{\sigma^2(\nu_4-3)}{c-1}\right\}, \\
\tilde{\sigma}^2_{c,3} &= 2(1-\frac{1}{c})r^4 + \frac{1}{p}\left\{\frac{2c^3\sigma^4}{(1-c)^4} + \frac{c\sigma^4(\nu_4-3)}{(c-1)^2}\right\}.
\end{aligned}
$$

The corresponding $(1-\alpha)$-confidence interval is given by

$$
P(L_{\alpha,c} \le R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le U_{\alpha,c}) \to 1-\alpha, \tag{11}
$$

with

$$
\begin{aligned}
L_{\alpha,c} &= \frac{\sigma^2}{c_n-1} + (1-\frac{1}{c_n})r^2 + \frac{1}{\sqrt{p}}(\tilde{\mu}_{c,3} - Z_{\alpha/2}\tilde{\sigma}_{c,3}), \\
U_{\alpha,c} &= \frac{\sigma^2}{c_n-1} + (1-\frac{1}{c_n})r^2 + \frac{1}{\sqrt{p}}(\tilde{\mu}_{c,3} + Z_{\alpha/2}\tilde{\sigma}_{c,3}).
\end{aligned}
$$

Remark 4.3. *Note that besides the leading constants in $(\mu_{c,3}, \sigma_{c,3})$, the version $(\tilde{\mu}_{c,3}, \tilde{\sigma}_{c,3})$ also contains smaller order terms, including terms of order $O(1/\sqrt{p})$ in $\tilde{\mu}_{c,3}$ and terms of order $O(1/p)$ in $\tilde{\sigma}_{c,3}$. These smaller order terms will vanish when p and n grow very large, but for finite sample situations, these smaller order terms will provide a finer approximation for the finite sample distribution of $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. As shown in the following experiments, these terms have indeed made non-negligible contributions to fitting the empirical distribution of $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, which sheds new lights for practitioners.*

Remark 4.4. *If we compare the results in Theorem 4.3 and 4.5, we will find out that $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ with constant β and $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ with random β share the same first-order limit and second-order error rate $O(p^{-1/2})$. This is quite* *intuitive because both risk treat β as a constant. Their differences are reflected in their limiting variances. Nevertheless, it's very interesting to observe from Theorem 4.4 that, $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ with random β under the over-parameterized case has a smaller second-order error rate $O(p^{-1})$. It enjoys the same rate as the under-parametrized case in Theorem 4.1. A possible explanation would be that averaging over the randomness in β can partially offset the curse of dimensionality so that $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ achieves the same error rate for all p, n combinations.*

4.4. Discussion

In this section, we first make a short conclusion of what we have done theoretically in this paper and further discuss some possible directions of extension.

We have systematically investigated the second-order fluctuations of two types of prediction risk, $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, for the high-dimensional least squares estimator $\hat{\boldsymbol{\beta}}$. Theorem 4.1 and 4.4 are for $R_{\mathbf{X}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ while Theorem 4.2 and 4.5 are for $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. Both fixed and random regression coefficients $\boldsymbol{\beta}$ are discussed following the settings in Hastie et al. (2019). Asymptotic results are categorized into the under-parameterized case ($p < n$) and the over-parameterized case ($p > n$).

The first-order limits of the prediction risk in high-dimensional linear models have already been well studied in recent years, including general extensions to anisotropic features and signals in Wu & Xu (2020). The *"double descent"* risk curve is depicted as a function of the limiting ratio $\lim p/n$. However, there is still a non-negligible discrepancy between the finite-sample prediction risk and its first-order limit on the *"double descent"* curve. How large is this discrepancy? How fast does the risk converge to its limit? Our CLTs provide answers to such questions and give a fine-grained characterization of the second-order fluctuations of the prediction risk. Not only explicit forms of the leading constants in the limiting means and variances are shown in our main theorems, smaller order terms are also derived to improve the empirical performance for practitioners.

It is also important to recognize the limitations of our results. First, the present paper only concerns linear regression task since the linear regression task is simple but important as well. For example, some recent works linearize neural networks at the initialization and employ Neural Tangent Kernels (Jacot et al., 2018) to approximate the training procedure of a strongly over-parameterized neural network by solving a linear regression task, e.g. Du et al. (2018); Arora et al. (2019); Lee et al. (2019). Though the setting considered in this paper is simple and limited, the problem has not been fully understood so far in the literature. Therefore, we are among the first to take the task and develop the second-order fluctuation results for the prediction risk. Second, we assume general covariance $\boldsymbol{\Sigma}$ and non-Gaussianity for the under-parameterized case, which fits the most updated and realistic settings in the literature, however, we only investigate the isotropic settings while still allow for non-Gaussianity under over-parameterization. We haven't extended it to the more general anisotropic settings yet. The reasons are two-fold. On the one hand, according to Wu & Xu (2020), the first-order limits depend on the Stieltjes transforms of the unknown spectral distribution of $\boldsymbol{\Sigma}$. Since $\boldsymbol{\Sigma}$ is unknown, we cannot obtain any explicit characterization of the first-order limits, not to mention the second-order fluctuations. The CLTs would only be written as certain complicated implicit functions of $\boldsymbol{\Sigma}$ and would be too abstract to evaluate practically. More restrictions would be imposed on $\boldsymbol{\Sigma}$ to guarantee the second-order convergence. On the other hand, from the technical perspective, the techniques required for anisotropic over-parameterized cases are very different from the isotropic cases due to difference in the bias-variance decomposition. The tools in random matrix theory have not been fully developed yet for anisotropic cases. Since we have considered various scenarios in this paper, including random and nonrandom signals $\boldsymbol{\beta}$ for both conditional and unconditional risk, it will take great efforts and continuous work to extend all of them to the most general settings, which would lead to many subsequent works in the field of machine learning and random matrix theory literature.

5. Experiments

In this section, we carry out simulation experiments to examine the CLTs and the corresponding confidence intervals in Theorem 4.2 and Theorem 4.5. We generate data points from the linear model (1) and directly compute the prediction risk via the bias-variance decomposition in (4). To make sure the assumption (A) holds, the generative distribution $P_{\mathbf{x}}$ is taken to be the standard normal distribution, the centered gamma with shape 4.0 and scale 0.5, and the normalized Student-t distribution with 6.0 degree of freedom. The noise distribution P_ϵ is taken to be $N(0, 1)$. In the following, we present the gap between the finite-sample distribution of the prediction risk and the corresponding limiting distribution to check the CLTs and use the coverage to measure the effectiveness of the confidence intervals. More simulation results are relegated to the Appendix due to space limitations.

Example 1. This example examines the results in Theorem 4.2. We define a standardized statistic:

$$T_n = \frac{p}{\sigma_c}\Big(R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) - R_{c_n}\Big) - \frac{\mu_c}{\sigma_c}.$$

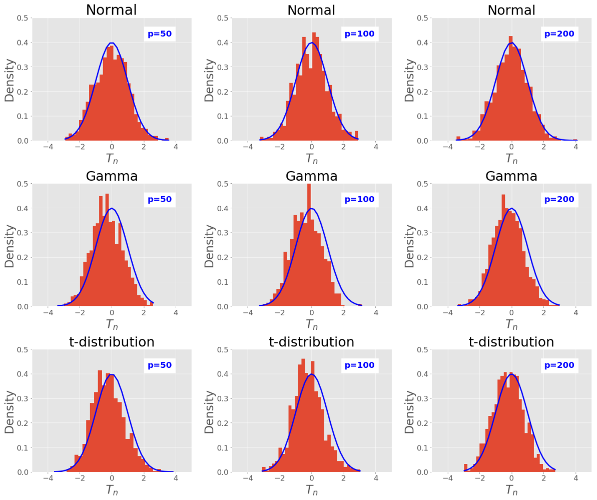

Figure 4. The histogram of T_n. The solid line is the density of the standard normal distribution.

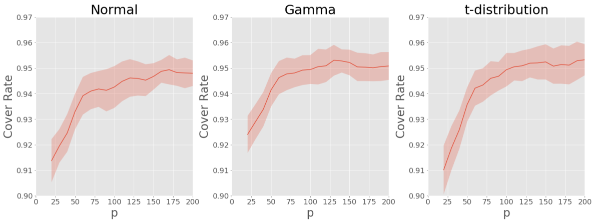

Figure 5. The cover rate of the confidence interval (7) as p creases. The confidence level is 95%.

According to Theorem 4.2, T_n weakly converges to the standard normal distribution as $n, p \to \infty$. In this example, $c = 1/2$ and $p = 50, 100, 200$. The finite-sample distribution of T_n is estimated by the histogram of T_n under 1000 repetitions. The results are presented in Figure 4. It can be seen that the finite-sample distribution of T_n is very consistent with the density function of the standard normal distribution, especially when n, p become larger. When $\alpha = 0.05$, the coverage of the 95%-confidence interval is reported in Figure 5. According to the mean and confidence band of the cover rate, we can find that the empirical coverage converges to 95% as $n, p \to \infty$. All these experiments verify the correctness of our theoretical results.

Example 2. This example verifies the results in Theorem 4.5. Here we consider two standardized statistics:

$$T_{n,0} = \frac{\sqrt{p}}{\sigma_{c,3}}\{R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) - R_{c_n}\} - \frac{\mu_{c,3}}{\sigma_{c,3}},$$

$$T_{n,1} = \frac{\sqrt{p}}{\tilde{\sigma}_{c,3}}\{R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) - R_{c_n}\} - \frac{\tilde{\mu}_{c,3}}{\tilde{\sigma}_{c,3}}.$$

According to the CLT (10) and its practical version, both $T_{n,0}$ and $T_{n,1}$ weakly converge to the standard normal distribution as $n, p \to +\infty$. Compared to $T_{n,0}$, $T_{n,1}$ provides a better approximation for the finite sample distribution of $R_{\mathbf{X},\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})$ because it contains smaller order terms in the asymptotic mean and variance. We take $c = 2$ and $p = 100, 200, 400$. Similarly the finite-sample distributions of $T_{n,0}$ and $T_{n,1}$ are presented by the histogram of $T_{n,0}$ and $T_{n,1}$ with 1000 repetitions. The results are presented in Figure 6 and Figure 7. It can also be seen that the finite sample distributions of $T_{n,0}$ and $T_{n,1}$ both match the standard normal distribution quite well, especially $T_{n,1}$ with more precise characterization. When $\alpha = 0.05$, the empirical coverage of the 95%-confidence interval (11) are reported in Figure 8.

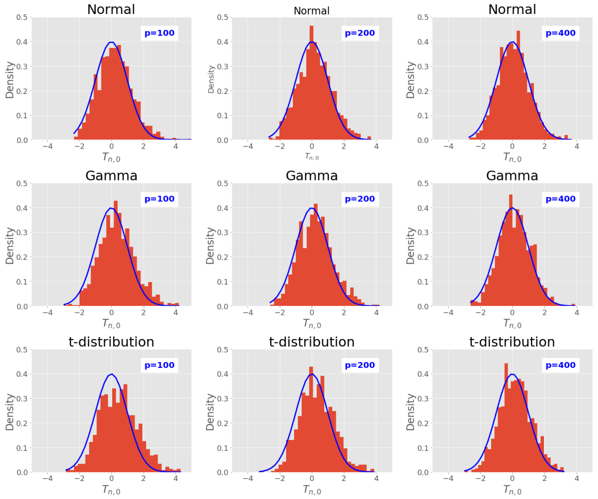

Figure 6. The histogram of $T_{n,0}$. The solid line is the density of the standard normal distribution.

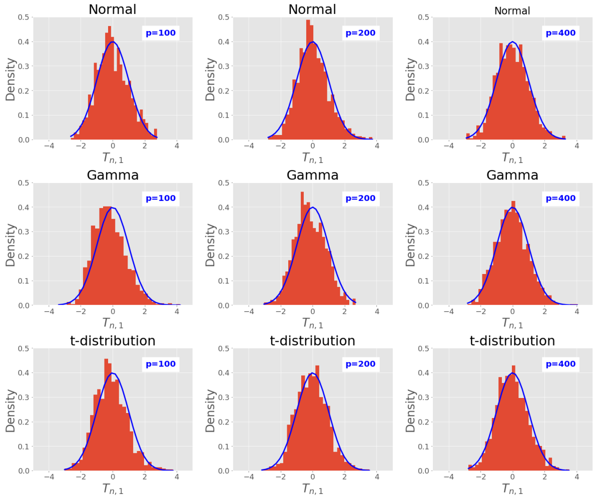

Figure 7. The histogram of $T_{n,1}$. The solid line is the density of the standard normal distribution.

Acknowledgements

Zeng Li's research is partially supported by NSFC (No. 12031005). Qinwen Wang's research is partially supported

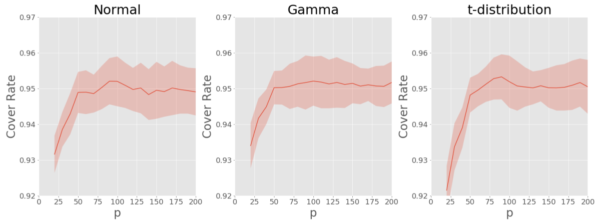

Figure 8. The cover rate of the confidence interval (11) as p creases. The confidence level is 95%.

by Shanghai Sailing Program (No. 18YF1401500) and NSFC (No. 11801085).

References

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

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 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*, 2019.

Bai, Z. and Silverstein, J. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability*, 32:553–605, 2004.

Bai, Z. and Yao, J. On the convergence of the spectral empirical process of wigner matrices. *Bernoulli*, 11(6): 1059–1092, 2005.

Bai, Z. and Yao, J. Central limit theorems for eigenvalues in a spiked population model. *Annales de l'IHP Probabilités et statistiques*, 44(3):447–474, 2008.

Bai, Z., Miao, B., and Pan, G. On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532–1572, 2007.

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

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pp. 2300–2311, 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., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv:1903.07571*, 2019b.

Chen, L., Min, Y., Zhang, M., and Karbasi, A. More data can expand the generalization gap between adversarially robust and standard models. *arXiv:2002.04725*, 2020.

Dar, Y. and Baraniuk, R. Double double descent: On generalization errors in transfer learning between linear regression tasks. *arXiv:2006.07002*, 2020.

Dar, Y., Mayer, P., Luzi, L., and Baraniuk, R. Subspace fitting meets regression: The effects of supervision and orthonormality constraints on double descent of generalization errors. *arXiv:2002.10614*, 2020.

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

Dereziński, M., Liang, F., and Mahoney, M. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv:1912.04533*, 2019.

Du, S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv:1810.02054*, 2018.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. Generalization error of generalized linear models in high dimensions. *arXiv:2005.00180*, 2020.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Gerbelot, C., Abbara, A., and Krzakala, F. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *arXiv:2006.06581*, 2020a.

Gerbelot, C., Abbara, A., and Krzakala, F. Asymptotic errors for convex penalized linear regression beyond gaussian matrices. *arXiv:2002.04372*, 2020b.

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

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. *arXiv:2002.10477*, 2020.

Kini, G. and Thrampoulidis, C. Analytic study of double descent in binary classification: The impact of loss. *arXiv:2001.11572*, 2020.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pp. 8572–8583, 2019.

LeJeune, D., Javadi, H., and Baraniuk, R. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 3525–3535, 2020.

Loog, M., Viering, T., and Mey, A. Minimizers of the empirical risk and risk monotonicity. In *Advances in Neural Information Processing Systems*, pp. 7478–7487, 2019.

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

Min, Y., Chen, L., and Karbasi, A. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv:2002.11080*, 2020.

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:1911.01544*, 2019.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Nakkiran, P. More data can hurt for linear regression: Sample-wise double descent. *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. In *International Conference on Learning Representations*, 2019.

Pan, G. and Zhou, W. Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *The Annals of Applied Probability*, 18(3):1232–1270, 2008.

Shen, Y. and Bellec, P. Asymptotic normality and confidence intervals for derivatives of 2-layers neural network in the random features model. In *Neural Information Processing Systems*, 2020.

Sinai, Y. and Soshnikov, A. Central limit theorem for traces of large random symmetric matrices with independent matrix elements. *Boletim da Sociedade Brasileira de Matemática-Bulletin/Brazilian Mathematical Society*, 29 (1):1–24, 1998.

Spigler, S., Geiger, M., dAscoli, 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*, 52 (47):474001, 2019.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wu, D. and Xu, J. On the optimal weighted l2 regularization in overparametrized linear regression. *arXiv:2006.05800*, 2020.

Xing, Y., Song, Q., and Cheng, G. Benefit of interpolation in nearest neighbor algorithms. *arXiv:1909.11720*, 2019.

Xu, J. and Hsu, D. On the number of variables to use in principal component regression. In *Advances in Neural Information Processing Systems*, pp. 5094–5103, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.

Zheng, S. Central limit theorems for linear spectral statistics of large dimensional f-matrices. *Annales de l'IHP Probabilités et statistiques*, 48(2):444–476, 2012.