Sketched Ridgeless Linear Regression: The Role of Downsampling

Xin Chen [*1] **Yicheng Zeng** [*2] **Siyue Yang** [3] **Qiang Sun** [3]

Abstract

Overparametrization often helps improve the generalization performance. This paper presents a dual view of overparametrization suggesting that downsampling may also help generalize. Focusing on the proportional regime $m \asymp n \asymp p$, where m represents the sketching size, n is the sample size, and p is the feature dimensionality, we investigate two out-of-sample prediction risks of the sketched ridgeless least square estimator. Our findings challenge conventional beliefs by showing that downsampling does not always harm generalization but can actually improve it in certain cases. We identify the optimal sketching size that minimizes out-of-sample prediction risks and demonstrate that the optimally sketched estimator exhibits stabler risk curves, eliminating the peaks of those for the full-sample estimator. To facilitate practical implementation, we propose an empirical procedure to determine the optimal sketching size. Finally, we extend our analysis to cover central limit theorems and misspecified models. Numerical studies strongly support our theory.

1. Introduction

According to international data corporation, worldwide data will grow to 175 zettabytes by 2025, with as much of the data residing in the cloud as in data centers. These massive datasets hold tremendous potential to revolutionize operations and analytics across various domains. However, their sheer size presents unprecedented computational challenges, as many traditional statistical methods and learning algorithms struggle to scale effectively.

In recent years, sketch-and-solve methods, also referred to as sketching algorithms, have emerged as a powerful solution for approximate computations over large datasets (Pilanci, 2016; Mahoney, 2011). Sketching algorithms first employ random sketching/projection or random sampling techniques to construct a small “sketch” of the full dataset, and then use this sketch as a surrogate to perform analyses of interest that would otherwise be computationally impractical on the full dataset.

This paper focuses on the linear regression problem. We assume that we have collected a set of independent and identically distributed (i.i.d.) data points following the model:

$$y_i = \beta^\top x_i + \varepsilon_i,\ i = 1, \cdots, n, \tag{1}$$

where $y_i \in \mathbb{R}$ represents the label of the i-th observation, $\beta \in \mathbb{R}^p$ is the unknown random regression coefficient vector, $x_i \in \mathbb{R}^p \sim x \sim P_x$ is the p-dimensional feature vector of the i-th observation, with P_x denoting a probability distribution on $\mathbb{R}^p$ having mean $\mathbb{E}(x) = 0$ and covariance $\operatorname{cov}(x) = \Sigma$. The i-th random noise term $\varepsilon_i \sim \varepsilon \sim P_\varepsilon$ is independent of x_i, with P_ε being a probability distribution on $\mathbb{R}$ having mean $\mathbb{E}(\varepsilon) = 0$ and variance $\operatorname{var}(\varepsilon) = \sigma^2$. In matrix form, the model can be expressed as:

$$Y = X\beta + \varepsilon,$$

where $X = (x_1, \cdots, x_n)^\top \in \mathbb{R}^{n\times p}$, $Y = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^\top \in \mathbb{R}^n$.

We consider the following ridgeless least square estimator

$$\widehat{\beta} := (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y = \lim_{\lambda\to 0^+}(X^{\mathrm{T}}X + n\lambda I_p)^{-1}X^{\mathrm{T}}Y,$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudoinverse and $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix. In the case where $\operatorname{rank}(X) = p$, the estimator $\widehat{\beta}$ reduces to the ordinary least square (OLS) estimator, which is the de-facto standard for linear regression due to its optimality properties. However, computing the OLS estimator, typically done via QR decomposition (Golub & Van Loan, 2013), has a computational complexity of $\mathcal{O}(np^2)$. This renders the computation of the full-sample OLS estimator infeasible when the sample size

[*] Equal contribution [1] Department of Operations Research and Financial Engineering, Princeton University, 98 Charlton St, Princeton, NJ 08544, USA. [2] Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong, 2001 Longxiang Boulevard, Shenzhen, Guangdong, China. [3] Department of Statistical Sciences, University of Toronto, 700 University Ave, Toronto, ON M5G 1X6, Canada. Correspondence to: Qiang Sun <qiang.sun@utoronto.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

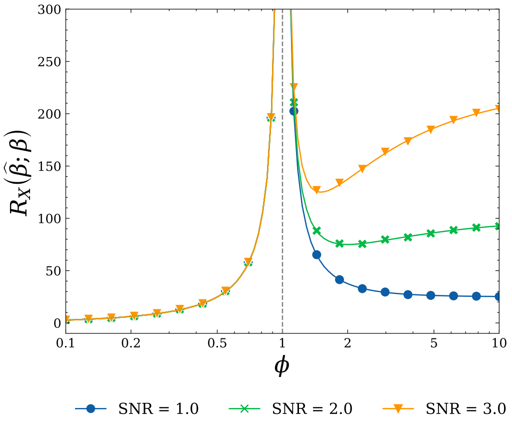

Figure 1. Asymptotic risk curves for the ridgeless least square estimator. The blue, green, and yellow lines are theoretical risk curves for $\mathrm{SNR} = \alpha/\sigma = 1, 2, 3$ with (α, σ) taking $(5, 5)$, $(10, 5)$ and $(15, 5)$, respectively. The blue dots, green crosses, and yellow triangles mark the finite-sample risks with $n = 400$, ϕ varying in $[0.1, 10]$ and $p = [n\phi]$. Each row of the feature matrix $X \in \mathbb{R}^{n\times p}$ is i.i.d. drawn from $\mathcal{N}(0, I_p)$.

n and dimensionality p reach the order of millions or even billions.

Sketching algorithms provide a solution to reduce the computational burden by reducing the data size, aka downsampling. This is achieved by multiplying the full dataset (X, Y) with a sketching matrix $S \in \mathbb{R}^{m\times n}$ to obtain the sketched dataset $(SX, SY) \in \mathbb{R}^{m\times p} \times \mathbb{R}^m$, where $m < n$ is the sketching size. Instead of computing the full-sample OLS estimator, we compute the sketched ridgeless least square estimator based on the sketched dataset:

$$\begin{aligned}\widehat{\beta}^S &= (X^\top S^\top SX)^+ X^\top S^\top SY \\ &= \lim_{\lambda\to 0^+} (X^\top S^\top SX + n\lambda I_p)^{-1} X^\top S^\top SY. \quad (2)\end{aligned}$$

The total computational complexity for computing the sketched data and sketched least square estimator is approximately $\mathcal{O}(np\log m + mp^2)$ when using fast orthogonal sketches (Pilanci, 2016). The prevailing belief is that sketching reduces the runtime complexity at the expense of statistical accuracy (Woodruff, 2014; Raskutti & Mahoney, 2016; Drineas & Mahoney, 2018; Dobriban & Liu, 2018). Indeed, as pointed out by (Dobriban & Liu, 2018), a larger number of samples leads to a higher accuracy. They showed that, in the case of orthogonal sketches, if one sketches to m samples such that $p < m < n$, the test error increases by a factor of $m(n - p)/n(m - p) > 1$, which equals 1.1 when $m = 10^6$, $n = 10^7$, and $p = 10^5$. (Raskutti & Mahoney, 2016) reported a similar phenomenon by considering the regime $n \gg p$ and various error criteria. However, these results only focus on the underparameterized regime ($p < m$) and do not reveal the statistical role of downsampling in a broader regime. It is therefore natural to ask the following questions:

> *What is the statistical role of downsampling? Does downsampling always hurt the statistical accuracy?*

This paper answers the questions above in the case of sketched ridgeless least square estimators (2), in both the underparameterized and overparameterized regimes, where downsampling is achieved through random sketching. Our intuition is that downsampling plays a similar role as that of increasing the model capacity. Because increasing the model capacity has been recently observed in modern machine learning to often help improve the generalization performance (He et al., 2016; Neyshabur et al., 2014; Novak et al., 2018; Belkin et al., 2018; Nakkiran et al., 2021), downsampling may also benefit generalization properties. This "dual view" can be seen clearly in the case of linear regression, where the out-of-sample prediction risk only depends on the model size and sample size via the quantity p/n (Hastie et al., 2022); see Figure 1. Thus increasing the model size p has the same effect to downsampling the sample size n.

Motivated by this dual view, we examine the out-of-sample prediction risks of the sketched ridgeless least square estimator in the proportional regime, where the sketching size m is comparable to the sample size n and the dimensionality p. We consider a broad class of sketching matrices that satisfy mild assumptions, as described in Assumption 4, which includes several existing sketching matrices as special cases. Our work makes the following key contributions.

1. First, we provide asymptotically exact formulas for the two out-of-sample prediction risks in the proportional regime. This allows us to reveal the statistical role of downsampling in terms of generalization performance. Perhaps surprisingly, we find that downsampling does not always harm the generalization performance and may even improve it in certain scenarios.

2. Second, we show that orthogonal sketching is optimal among all types of sketching matrices considered in the underparameterized case. In the overparameterized case however, all general sketching matrices are equivalent to each other.

3. Third, we identify the optimal sketching sizes that minimize the out-of-sample prediction risks. The optimally sketched ridgeless least square estimators exhibit universally better risk curves when varying the model size, indicating their improved stability compared with the full-sample estimator.

4. Fourth, we propose a practical procedure to empirically determine the optimal sketching size using an additional validation dataset, which can be relatively small in size.

5. Fifth, in addition to characterizing the first-order limits, we provide central limit theorems for the risks. Leveraging results from random matrix theory for covariance matrices (Zhang, 2007; El Karoui, 2009; Knowles & Yin, 2017; Zheng et al., 2015), we establish almost sure convergence results for the test risks. These results complement the work of (Dobriban & Liu, 2018), which focused on the asymptotic limits of expected risks. The expected risk results can be recovered from our findings using the dominated convergence theorem.

1.1. Related work

Generalization properties of overparameterized models The generalization properties of overparametrized models have received significant attention in recent years. It all began with the observation that overparameterized neural networks often exhibit benign generalization performance, even without the use of explicit regularization techniques (He et al., 2016; Neyshabur et al., 2014; Canziani et al., 2016; Novak et al., 2018; Zhang et al., 2021; Bartlett et al., 2020; Liang & Rakhlin, 2020). This observation challenges the conventional statistical wisdom that overfitting the training data leads to poor accuracy on new examples. To reconcile this discrepancy, (Belkin et al., 2018) introduced the unified "double descent" performance curve that reconciles the classical understanding with the modern machine learning practice. This double descent curve subsumes the textbook U-shape bias-variance-tradeoff curve (Hastie et al., 2009) by demonstrating how increasing model capacity beyond the interpolation threshold can actually lead to improved test errors. Subsequent research has aimed to characterize this double descent phenomenon in various simplified models, including linear models (Hastie et al., 2022; Richards et al., 2021), random feature models (Mei & Montanari, 2022), and partially optimized two-layer neural network (Ba et al., 2019), among others.

Implicit regularization and minimum ℓ_2-norm solutions Another line of research focuses on understanding the phenomenon of benign overfitting through implicit regularization mechanisms in overparameterized models (Neyshabur et al., 2014). For instance, (Gunasekar et al., 2018) and (Zhang et al., 2021) showed that gradient descent (GD) converges to the minimum ℓ_2-norm solutions in linear regression problems, which corresponds to the ridgeless least square estimators. (Hastie et al., 2022) characterized the exact out-of-sample prediction risk for the ridgeless least square estimator in the proportional regime. Minimum ℓ_2-norm solutions are also studied for other models, including kernel ridgeless regression (Liang & Rakhlin, 2020), classification (Chatterji & Long, 2021; Liang & Recht, 2021; Muthukumar et al., 2021), and the random feature model (Mei & Montanari, 2022).

Paper overview The rest of this paper proceeds as follows. Section 2 provides the necessary preliminaries for our analysis. In Section 3, we investigate the out-of-sample prediction risks under the assumption of isotropic features. Section 4 focuses on the case of correlated features. Section 5 presents the conclusions and discussions. Additionally, it includes simulation results of a simple and practical procedure for selecting the optimal sketching size using a validation dataset. Due to space limitations, the details of this practical procedure, results on central limit theorems and misspecified models, the details of some numerical experiments, computational cost comparisons, as well as all proofs, are provided in the appendix.

2. Preliminaries

In this section, we provide definitions for two types of random sketching matrices, introduce two out-of-sample prediction risks to measure the generalization performance, and present several standing assumptions that are crucial for our analysis.

2.1. Sketching matrix

A sketching matrix $S \in \mathbb{R}^{m\times}$ is used to construct the sketched dataset (SX, SY), allowing us to perform approximate computations on the sketched dataset for the interest of the full dataset. Recall m is the sketching size and we shall refer to m/n as the downsampling ratio. We consider two types of sketching matrices: orthogonal sketching matrices and i.i.d. sketching matrices, defined as follows.

Definition 1 (Orthogonal sketching matrix)**.** An orthogonal sketching matrix $S \in \mathbb{R}^{m\times n}$ is a partial orthogonal random matrix, i.e., S satisfies the condition $SS^{\mathrm{T}} = I_m$, where I_m denotes the identity matrix of size $m \times m$.

Definition 2 (i.i.d. sketching matrix)**.** An i.i.d. sketching matrix S is a random matrix whose entries are i.i.d. , each with mean zero, variance $1/n$, and a finite fourth moment.

For i.i.d. sketching, we consider i.i.d. Gaussian sketching matrices in all of our experiments, although our results hold for general i.i.d. sketching matrices. For orthogonal sketching, we construct an orthogonal sketching matrix based on the subsampled randomized Hadamard transforms (Ailon & Chazelle, 2006). Specifically, we use $S = BHDP$, where the rows of $B \in \mathbb{R}^{m\times n}$ are sampled without replacement from the standard basis of $\mathbb{R}^n$, $H \in \mathbb{R}^{n\times n}$ is a Hadamard matrix [1], $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix of i.i.d. Rademacher random variables, and $P \in \mathbb{R}^{n\times n}$ is a uniformly distributed permutation matrix. The time complexity of computing (SX, SY) is of order $\mathcal{O}(np\log m)$. Orthogonal sketching matrices can also be realized by, for example, subsampling and Haar distributed matrices (Mezzadri, 2006).

2.2. Out-of-sample prediction risk

Recall that $\widehat{\beta}^S$ is the sketched ridgeless regression estimator defined in Equation (2). Let us consider a test data point $x_{\text{new}} \sim P_x$, which is independent of the training data. Following (Hastie et al., 2022), we consider the following out-of-sample prediction risk as a measure of the generalization performance:

$$\begin{aligned} R_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) &= \mathbb{E}\left[\left(x_{\text{new}}^{\mathrm{T}}\widehat{\beta}^S - x_{\text{new}}^{\mathrm{T}}\beta\right)^2 \Big|\beta, S, X\right] \\ &= \mathbb{E}\left[\left\|\widehat{\beta}^S - \beta\right\|_\Sigma^2 \Big|\beta, S, X\right], \end{aligned}$$

where $\Sigma := \operatorname{cov}(x_i)$ is the covariance matrix of x_i, and $\|x\|_\Sigma^2 := x^{\mathrm{T}}\Sigma x$. The above conditional expectation is taken with respect to the randomness of $\{\varepsilon_i\}_{1\le i\le n}$ and x_{new}, while β, S, and X are fixed. We can decompose the out-of-sample prediction risk into bias and variance components:

$$\begin{aligned} &R_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) \\ =&B_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) + V_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right), \end{aligned} \tag{3}$$

where

$$B_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) = \left\|\mathbb{E}\left(\widehat{\beta}^S|\beta, S, X\right) - \beta\right\|_\Sigma^2, \tag{4}$$

$$V_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) = \operatorname{tr}\left[\operatorname{Cov}\left(\widehat{\beta}|\beta, S, X\right)\Sigma\right]. \tag{5}$$

We also consider a second out-of-sample prediction risk, defined as:

$$\begin{aligned} R_{(S,X)}\left(\widehat{\beta}^S;\beta\right) &= \mathbb{E}\left[\left(x_{\text{new}}^{\mathrm{T}}\widehat{\beta}^S - x_{\text{new}}^{\mathrm{T}}\beta\right)^2 \Big|S, X\right] \\ &= \mathbb{E}\left[\left\|\widehat{\beta}^S - \beta\right\|_\Sigma^2 \Big|S, X\right]. \end{aligned}$$

The second one also averages over the randomness of β. Similarly, we have the following bias-variance decomposition

$$R_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = B_{(S,X)}\left(\widehat{\beta}^S;\beta\right) + V_{(S,X)}\left(\widehat{\beta}^S;\beta\right),$$

where

$$B_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \mathbb{E}\left[\left\|\mathbb{E}\left(\widehat{\beta}^S|\beta, S, X\right) - \beta\right\|_\Sigma^2 \Big|S,X\right], \tag{6}$$

$$V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \mathbb{E}\left\{\operatorname{tr}\left[\operatorname{Cov}\left(\widehat{\beta}|\beta, S, X\right)\Sigma\right] \Big|S,X\right\}. \tag{7}$$

We shall also refer to the above out-of-sample prediction risks as test risks or simply risks, since they are the only risks considered in this paper. Throughout the paper, we study the above two out-of-sample prediction risks by examining their bias and variance terms respectively. Specifically, we study the behaviors of $R_{(\beta,S,X)}(\widehat{\beta}^S;\beta)$ and $R_{(S,X)}(\widehat{\beta}^S;\beta)$ in the proportional asymptotic limit where the sketching size m, sample size n, and dimensionality p all tend to infinity such that the aspect ratio converges as $\phi_n := p/n \to \phi$, and the downsampling ratio converges as $\psi_n := m/n \to \psi \in (0, 1)$. It is worth noting that $R_{(\beta,S,X)}(\widehat{\beta}^S;\beta)$ exhibits larger variability due to the additional randomness introduced by the random variable β, S, X when compared with $R_{(S,X)}(\widehat{\beta}^S;\beta)$.

2.3. Assumptions

This subsection collects standing assumptions.

Assumption 1 (Covariance and moment conditions)**.** For $i = 1, \cdots, n$, $x_i = \Sigma^{1/2}z_i$, where z_i has i.i.d. entries with mean zero, variance one and a finite moment of order $4 + \eta$ for some $\eta > 0$. The noise ε is independent of x, and follows a distribution P_ε on $\mathbb{R}$ with mean $\mathbb{E}(\varepsilon) = 0$ and variance $\operatorname{var}(\varepsilon) = \sigma^2$.

Assumption 2 (Correlated features)**.** The matrix Σ is a deterministic positive definite matrix, and there exist constants C_0, C_1 such that $0 < C_0 \le \lambda_{\min}(\Sigma) \le \lambda_{\min}(\Sigma) \le C_1$ for all n and p. The empirical spectral distribution (ESD) of Σ is defined as $F^\Sigma(x) = \frac{1}{p}\sum_{i=1}^p \mathbf{1}_{[\lambda_i(\Sigma),\infty)}(x)$. Assume that as $p \to \infty$, the ESD F^Σ converges weakly to a probability measure H.

Assumption 3 (Random β)**.** The coefficient vector $\beta \in \mathbb{R}^p$ is a random vector with i.i.d. entries satisfying $\mathbb{E}(\beta) = 0$, $\mathbb{E}\left((\sqrt{p}\beta_i)^2\right) = \alpha^2$, and $\sup_i \mathbb{E}\left((\sqrt{p}\beta_i)^{4+\eta}\right) < \infty$ for some $\eta > 0$. It is assumed to be independent of the data matrix X, the noise ε, and the sketching matrix S.

Assumption 4 (Sketching matrix)**.** Let $S \in \mathbb{R}^{m\times n}$ be a sketching matrix. Suppose the ESD of SS^{T} converges weakly to a probability measure B. Furthermore, there exist constants $\widetilde{C}_0, \widetilde{C}_1 > 0$ such that almost surely for all large n, it holds that $0 < \widetilde{C}_0 \le \lambda_{\min}\left(SS^{\mathrm{T}}\right) \le \lambda_{\max}\left(SS^{\mathrm{T}}\right) \le \widetilde{C}_1$.

Assumption 1 specifies the covariance matrix for features and moment conditions for both features and errors (Dobriban & Wager, 2018; Hastie et al., 2022; Li et al., 2021). While (Dobriban & Liu, 2018) requires only a finite fourth

[1] The definition of Hadamard matrices can be found, for example, in (Ailon & Chazelle, 2006).

moment for z_i, which is slightly weaker than our moment condition, this is because they studied the expected risk, which has less randomness compared to our risks. Assumption 2 considers correlated features. In this paper, we first focus on the random β case as stated in Assumption 3, where β follows an isotropic Gaussian distribution, allowing for clear presentation of optimal sketching size results. The assumption of random β is commonly adopted in the literature (Dobriban & Wager, 2018; Li et al., 2021). We also consider the deterministic β in Appendix B, where the interaction between β and Σ needs to be taken into account. Assumption 4 regarding the sketching matrix is relatively mild. For example, orthogonal sketching matrices naturally satisfy this assumption. According to (Bai & Silverstein, 1998), almost surely there are no eigenvalues outside the support of the limiting spectral distribution (LSD) of large-dimensional sample covariance matrices for sufficiently large sample size. Therefore, the i.i.d. sketching matrices in Definition 2 also satisfy this assumption.

3. A warm-up case: Isotropic features

As a warm-up, we first study the case of isotropic features, specifically when $\Sigma = I_p$, and postpone the investigation of the correlated case to Section 4. Before presenting the limiting behaviors, we establish the relationship between the two out-of-sample prediction risks through the following lemma, which is derived in the general context of correlated features.

Lemma 3.1. *Under Assumptions 1 and 3, the biases* (4) *and* (6)*, as well as the variances* (5) *and* (7)*, can be expressed as follows:*

$$B_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) = \beta^\intercal\left[(X^\intercal S^\intercal SX)^+ X^\intercal S^\intercal SX - I_p\right] \cdot \Sigma\left[(X^\intercal S^\intercal SX)^+ X^\intercal S^\intercal SX - I_p\right]\beta,$$

$$B_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \frac{\alpha^2}{p}\operatorname{tr}\left\{\left[I_p - (X^\intercal S^\intercal SX)^+ X^\intercal S^\intercal SX\right]\Sigma\right\},$$

$$V_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) = V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \operatorname{tr}\left[\sigma^2 (X^\intercal S^\intercal SX)^+ X^\intercal S^\intercal SS^\intercal SX (X^\intercal S^\intercal SX)^+ \Sigma\right].$$

Furthermore, suppose there exists C_1 such that $\lambda_{\max}(\Sigma) \le C_1$. Then as $n, p \to \infty$,

$$R_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) - R_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \overset{\text{a.s.}}{\to} 0. \tag{8}$$

The above lemma establishes the asymptotic equivalence of the two risks when β is random, with the variance terms being exactly equal and the bias terms converging asymptotically. Due to this asymptotic equivalence, our primary focus will be on analyzing the risk $R_{(S,X)}(\widehat{\beta}^S;\beta)$. However, it should be noted that the second-order inferential results do not align in general, and this discrepancy will be discussed in detail in Appendix B.

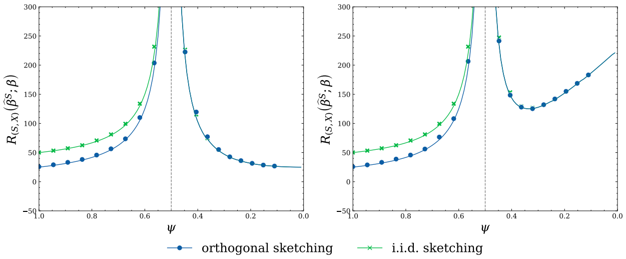

Figure 2. Asymptotic risk curves for sketched ridgeless least square estimators with orthogonal and i.i.d. sketching under isotropic features, as functions of ψ. The lines in the left panel and right panel are theoretical risk curves for $\mathrm{SNR} = \alpha/\sigma = 1$ with $(\alpha, \sigma) = (5, 5)$ and $\mathrm{SNR} = \alpha/\sigma = 3$ with $(\alpha, \sigma) = (15, 5)$, respectively. The blue lines are for orthogonal sketching, while the green lines are for i.i.d. sketching. The blue dots mark the finite-sample risks for orthogonal sketching, while the green crosses mark the finite-sample risks for i.i.d. sketching, with $n = 400$, $p = 200$, ψ varying in $(0, 1)$, and $m = [n\psi]$. Each row of $X \in \mathbb{R}^{n\times p}$ is i.i.d. drawn from $\mathcal{N}_p(0, I_p)$. The orthogonal sketching matrices are generated using subsampled randomized Hadamard transform, while the entries of the i.i.d. sketching matrices are drawn independently from $\mathcal{N}(0, 1/n)$.

3.1. Limiting risks

We first focus on the case of isotropic features, which enables us to obtain clean expressions for the limiting risks. We characterize the limiting risks with two types of sketching matrices: orthogonal and i.i.d. sketching matrices, which were introduced earlier. Recall that we consider $m, n, p \to \infty$ such that $p/n \to \phi$ and $m/n \to \psi \in (0, 1)$.

Theorem 3.2. *Under Assumptions 1, 3, 4, and $\Sigma = I_p$, the following results hold.*

(i) If S is an orthogonal sketching matrix, then

$$R_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \overset{\text{a.s.}}{\to} \begin{cases} \frac{\sigma^2\phi\psi^{-1}}{1-\phi\psi^{-1}}, & \phi\psi^{-1} < 1, \\ \alpha^2(1-\psi\phi^{-1}) + \frac{\sigma^2}{\phi\psi^{-1}-1}, & \phi\psi^{-1} > 1. \end{cases}$$

(ii) If S is an i.i.d. sketching matrix, then

$$R_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \overset{\text{a.s.}}{\to} \begin{cases} \frac{\sigma^2\phi}{1-\phi} + \frac{\sigma^2\phi\psi^{-1}}{1-\phi\psi^{-1}}, & \phi\psi^{-1} < 1, \\ \alpha^2(1-\psi\phi^{-1}) + \frac{\sigma^2}{\phi\psi^{-1}-1}, & \phi\psi^{-1} > 1. \end{cases}$$

Moreover, $R_{(\beta,S,X)}(\widehat{\beta}^S;\beta)$ with orthogonal sketching and i.i.d. sketching converge almost surely to the same limits, respectively.

In the above theorem, we have characterized the limiting risks of sketched ridgeless least square estimators with both orthogonal and i.i.d. sketching. The limiting risks are determined by theoretical risk curves in the underparameterized and overparameterized regimes after sketching, where the regimes are described by $\phi\psi^{-1} < 1$ and $\phi\psi^{-1} > 1$, respectively. We shall simply call these two regimes underparametrized and overparameterized regimes respectively.

Interestingly, orthogonal and i.i.d. sketching exhibit different behaviors in the underparameterized regime, while their limiting risks agree in the overparameterized regime. In the underparameterized regime, taking orthogonal sketching is strictly better than taking i.i.d. sketching in terms of out-of-sample prediction risks. This difference can be attributed to the distortion of the geometry of the least square regression estimator caused by the non-orthogonality in i.i.d. sketching, as pointed out by (Dobriban & Liu, 2018), but for a different risk. Their risk is the expected version of ours. By using the dominated convergence theorem, Theorem 3.2 can recover their results in the underparameterized case.

Moving to the overparameterized case however, both orthogonal and i.i.d. sketching yield identical limiting risks. Specifically, when $\phi\psi^{-1} > 1$, the bias term $B_{(S,X)}(\widehat{\beta}^S;\beta) \overset{\text{a.s.}}{\to} \alpha^2(1-\psi\phi^{-1})$ and the variance term $V_{(S,X)}(\widehat{\beta}^S;\beta) \overset{\text{a.s.}}{\to} \sigma^2(\phi\psi^{-1}-1)^{-1}$ hold for both types of sketching.

Let $\mathrm{SNR} = \alpha/\sigma$. Figure 2 plots the asymptotic risk curves as functions of ψ, for sketched ridgeless least square estimators with orthogonal and i.i.d. sketching when $\mathrm{SNR} = \alpha/\sigma = 1, 3$ with $(\alpha,\sigma) = (5,5)$ and $(\alpha,\sigma) = (15,5)$ respectively, along with finite-sample risks. As depicted in the figure, orthogonal sketching is strictly better than i.i.d. sketching in the underparameterized regime, while they are identical in the overparameterized regime.

Lastly, we compare the limiting risk $R_{(S,X)}(\widehat{\beta};\beta)$ of the orthogonally sketched estimator with that of the full-sample estimator, since orthogonal sketching is universally better than i.i.d. sketching. We can use a variant of (Hastie et al., 2022, Theorem 1) to obtain the limiting risk $R_X(\widehat{\beta};\beta)$ of the full-sample ridgeless least square estimator $\widehat{\beta}$ with isotropic features:

$$R_X\left(\widehat{\beta};\beta\right) \overset{\text{a.s.}}{\to} \begin{cases} \frac{\sigma^2\phi}{1-\phi}, & \phi < 1, \\ \alpha^2(1-\phi^{-1}) + \frac{\sigma^2}{\phi-1}, & \phi > 1. \end{cases}$$

Figure 1 displays the asymptotic risk curves and finite-sample risks of $\widehat{\beta}$. The limiting risk $R_X(\widehat{\beta};\beta)$ depends on the sample size n and dimensionality p only through the aspect ratio $\phi = \lim p/n$. Comparing this limiting risk with that of the orthogonally sketched estimator in Theorem 3.2, we observe that orthogonal sketching modifies the limiting risk by changing the *effective aspect ratio* from $\phi = \lim p/n$ for the original problem to $\phi\psi^{-1} = \lim p/m$ for the sketched problem. This is natural since sketching is a form of downsampling that affects the aspect ratio and, consequently, the limiting risk. Therefore, it is reasonable to ask the following question:

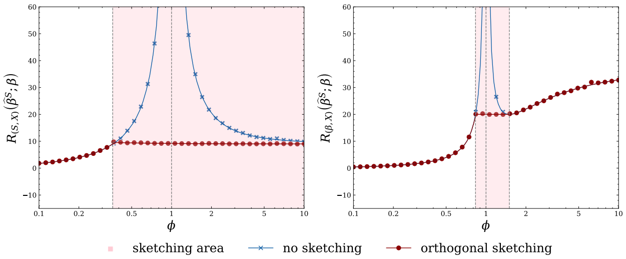

Figure 3. Asymptotic risk curves for the full-sample (no sketching) and orthogonally sketched ridgeless least square estimators under isotropic features, as functions of ϕ. For the sketched estimator, the optimal sketching size m^* is selected based on the SNR and ϕ, as described in Theorem 3.3. In the left panel and right panel, the lines represent the theoretical risk curves for $\mathrm{SNR} = \alpha/\sigma = 0.75$ with $(\alpha,\sigma) = (3,4)$ and $\mathrm{SNR} = \alpha/\sigma = 3$ with $(\alpha,\sigma) = (6,2)$, respectively. The blue crosses represent the finite-sample risks for the full-sample estimator, while the red dots indicate the finite-sample risks for the sketched estimator, with $n = 400$, ϕ varying in $[0.1, 10]$, and $p = [n\phi]$. The feature and orthogonal sketching matrices are generated in the same way as in Figure 2.

> *By carefully choosing the sketching size, can we potentially improve the out-of-sample prediction risks and, consequently, the generalization performance?*

This possibility arises due to the non-monotonicity of the asymptotic risk curves in Figure 1. In the following subsection, we investigate the optimal sketching size.

3.2. Optimal sketching size

In the previous subsection, we discussed the possibility of improving out-of-sample prediction risks and thus generalization performance by carefully choosing the sketching size. We now present the optimal sketching size m^* to minimize the limiting risks for both orthogonal and i.i.d. sketching.

Theorem 3.3 (Optimal sketching size for orthogonal and i.i.d. sketching)**.** *Assume Assumptions 1, 3, 4, and $\Sigma = I_p$. The optimal sketching size m^* for both orthogonal and i.i.d. sketching can be determined as follows.*

(a) If $\mathrm{SNR} > 1$ and $\phi \in (1-\frac{\sigma}{2\alpha}, \frac{\alpha}{\alpha-\sigma}]$, the optimal sketching size to minimize both limiting risks is $m^* = \frac{\alpha-\sigma}{\alpha}\phi \cdot n$.

(b) If $\mathrm{SNR} \leq 1$ and $\phi \in (\frac{\alpha^2}{\alpha^2+\sigma^2}, \infty)$, taking $\widetilde{\beta} = 0$ (corresponding to $m^* = 0$) yields the optimal solution.

(c) *No sketching is needed if either of the following two holds: (i)* $\mathrm{SNR} \leq 1$ *and* $\phi \in (0, \frac{\alpha^2}{\alpha^2+\sigma^2}]$*, or (ii)* $\mathrm{SNR} > 1$ *and* $\phi \in (0, 1 - \frac{\sigma}{2\alpha}] \bigcup (\frac{\alpha}{\alpha-\sigma}, \infty)$*.*

Theorem 3.3 reveals that both orthogonal and i.i.d. sketching can help improve out-of-sample prediction risks in certain cases. Specifically, when the signal-to-noise ratio is large with $\mathrm{SNR} > 1$ and the aspect ratio ϕ is within the range $(1-\sigma/(2\alpha), \alpha/(\alpha-\sigma)]$, a nontrivial sketching size of $m^* = (\alpha - \sigma)\phi n/\alpha$ leads to the optimal asymptotic risks. On the other hand, when the signal-to-noise ratio is low and the problem dimension is large, the null estimator $\widetilde{\beta} = 0$, which corresponds to $m^* = 0$, is the best among all sketched ridgeless least square estimators.

Figure 3 displays the asymptotic risk curves, as functions of ϕ, for the full-sample and optimally sketched ridgeless least square estimators using orthogonal sketching under isotropic features. As shown in the figure, optimal sketching can stabilize the asymptotic risk curves by eliminating the peaks, indicating that the optimally sketched estimator is a more stable estimator compared to the full-sample one. In Section A, we propose a practical procedure for selecting the optimally sketched estimator.

4. Correlated features

This section considers a general covariance matrix Σ. The results presented here apply to general sketching matrices captured by Assumption 4, including orthogonal and i.i.d. sketching as special cases. We will discuss the overparameterized and underparameterized cases separately.

4.1. Overparameterized regime

Recall that H is the limiting spectral distribution (LSD) of Σ, and $p, m, n \to \infty$ such that $\phi_n = p/n \to \phi$ and $\psi_n = m/n \to \psi \in (0, 1)$. In order to analyze the overparameterized case, we need the following lemma.

Lemma 4.1. *Assume Assumption 2. Suppose* $\phi\psi^{-1} > 1$*. Then the following equation (9) has a unique negative solution with respect to* c_0*,*

$$1 = \int \frac{x}{-c_0 + x\psi\phi^{-1}} \, dH(x). \tag{9}$$

The above lemma establishes the existence and uniqueness of a negative solution to the equation (9). Equations of this type, known as self-consistent equations (Bai & Silverstein, 2010), do not generally have closed-form solutions but the solutions can be computed numerically. Self-consistent equations are fundamental in calculating asymptotic risks. Lemma 4.1 provides the crucial result of the existence and uniqueness of a negative solution to (9) over $\mathbb{R}^-$, the negative real line. This result is essential for our asymptotic risk calculations and, to the best of our knowledge, is not available in the literature. We denote the unique negative solution to (9) as $c_0 = c_0(\phi, \psi, H)$, which will be used in our subsequent analysis. Our next result characterizes the limiting risks, as well as the limiting biases and variances, in the overparameterized regime.

Theorem 4.2. *Assume Assumptions 1- 4. Suppose* $\phi\psi^{-1} > 1$*. Then the following results hold:*

$$B_{(S,X)}(\widehat{\beta}^S; \beta),\ B_{(\beta,S,X)}(\widehat{\beta}^S; \beta) \overset{\text{a.s.}}{\to} -\alpha^2 c_0, \tag{10}$$

$$\begin{aligned} V_{(S,X)}(\widehat{\beta}^S; \beta) &= V_{(\beta,S,X)}(\widehat{\beta}^S; \beta) \\ &\overset{\text{a.s.}}{\to} \sigma^2 \frac{\int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2} \, dH(x)}{1 - \int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2} \, dH(x)}. \end{aligned} \tag{11}$$

Consequently, the limiting risks $R_{(S,X)}(\widehat{\beta}^S; \beta)$ *and* $R_{(\beta,S,X)}(\widehat{\beta}^S; \beta)$ *converge almost surely to the sum of the right-hand sides of (10) and (11).*

Different from the case with isotropic features, the asymptotic risk in the presence of correlated features does not admit closed-form solutions. However, it can be computed numerically. When $\Sigma = I_p$, the limiting spectral distribution H degenerates to the Dirac measure δ_1. In this case, we can show $c_0 = \psi\phi^{-1} - 1$, $B_{(S,X)}(\widehat{\beta}^S; \beta) \to \alpha^2(1 - \psi\phi^{-1})$, and $V_{(S,X)}(\widehat{\beta}^S; \beta) \to \frac{\sigma^2}{\phi\psi^{-1} - 1}$. These results recover Theorem 3.2 for the case of isotropic features. Furthermore, in the overparameterized regime, the limiting risks do not depend on a specific sketching matrix. This generalizes the same phenomenon observed in Theorem 3.2 for isotropic features.

4.2. Underparameterized regime

Recall from Assumption 4 that B is the limiting spectral distribution of $SS^\top$. We define $\widetilde{c}_0 = \widetilde{c}_0(\phi, \psi, B)$ as the unique negative solution to the self-consistent equation:

$$1 = \psi \int \frac{x}{-\widetilde{c}_0 + x\phi} \, dB(x). \tag{12}$$

Now we present the results for the limiting risks in the underparameterized regime, as well as the limiting biases and variances.

Theorem 4.3. *Assume Assumptions 1-4. Suppose* $\phi\psi^{-1} < 1$*. Then*

$$B_{(S,X)}(\widehat{\beta}^S; \beta),\ B_{(\beta,S,X)}(\widehat{\beta}^S; \beta) \overset{\text{a.s.}}{\to} 0, \tag{13}$$

$$\begin{aligned} V_{(S,X)}(\widehat{\beta}^S; \beta) &= V_{(\beta,S,X)}(\widehat{\beta}^S; \beta) \\ &\overset{\text{a.s.}}{\to} \sigma^2 \frac{\psi \int \frac{x^2\phi}{(\widetilde{c}_0 - x\phi)^2} \, dB(x)}{1 - \psi \int \frac{x^2\phi}{(\widetilde{c}_0 - x\phi)^2} \, dB(x)}. \end{aligned} \tag{14}$$

Consequently, both $R_{(S,X)}(\widehat{\beta}^S; \beta)$ *and* $R_{(\beta,S,X)}(\widehat{\beta}^S; \beta)$ *converge almost surely to the right hand side of (14).*

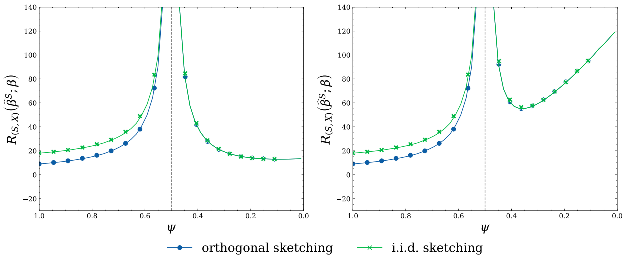

Figure 4. Asymptotic risk curves for sketched ridgeless least square estimators with orthogonal and i.i.d. sketching under correlated features, as functions of ψ. The lines in the left panel and the right panel are theoretical risk curves for SNR $= \alpha/\sigma = 1$ with $(\alpha, \sigma) = (3, 3)$ and SNR $= \alpha/\sigma = 3$ with $(\alpha, \sigma) = (9, 3)$, respectively. The blue dots mark the finite-sample risks for orthogonal sketching, while the green crosses mark the risks for i.i.d. sketching, with $n = 400$, $p = 200$, ψ varying in $(0, 1)$, and $m = [n\psi]$. Each row of $X \in \mathbb{R}^{n\times p}$ is i.i.d. drawn from $\mathcal{N}_p(0, \Sigma)$ and Σ has empirical spectral distribution $F^\Sigma(x) = \frac{1}{p}\sum_{i=1}^p \mathbf{1}_{[\lambda_i(\Sigma),\infty)}(x)$ with $\lambda_i = 2$ for $i = 1, \dots, [p/2]$, and $\lambda_i = 1$ for $i = [p/2] + 1, \dots, p$. The orthogonal and i.i.d. sketching matrices are generated in the same way as in Figure 2.

In the underparameterized case, the biases vanish, and the variances depend on the sketching matrix S and are independent of the covariance matrix Σ. The following corollary presents the limiting variances for orthogonal and i.i.d. sketching.

Corollary 4.4. *Assume the same assumptions as in Theorem 4.3. The following hold.*

(i) If S is an orthogonal sketching matrix, then

$$V_{(S,X)}(\widehat{\beta}^S;\beta) = V_{(\beta,S,X)}(\widehat{\beta}^S;\beta) \overset{\text{a.s.}}{\to} \sigma^2 \frac{\phi\psi^{-1}}{1-\phi\psi^{-1}}. \tag{15}$$

(ii) If S is an i.i.d. sketching matrix, then

$$V_{(S,X)}(\widehat{\beta}^S;\beta) = V_{(\beta,S,X)}(\widehat{\beta}^S;\beta) \overset{\text{a.s.}}{\to} \sigma^2\left(\frac{\phi}{1-\phi} + \frac{\phi\psi^{-1}}{1-\phi\psi^{-1}}\right). \tag{16}$$

The corollary above once again confirms that taking i.i.d. sketching yields a larger limiting variance compared to taking orthogonal sketching, extending the corresponding results for isotropic features. This naturally raises the question:

> *Is orthogonal sketching matrix optimal among all sketching matrices?*

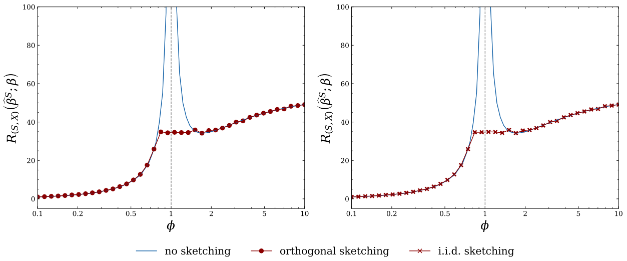

Figure 5. Asymptotic risk curves for the full-sample (no sketching) and sketched ridgeless least square estimators with orthogonal or i.i.d. sketching under correlated features, as functions of ϕ. For the sketched estimator, the optimal sketching size m^* is selected based on theoretical risk curves, as described in Appendix C.2. The blue lines are the theoretical risk curves for the full-sample estimator with SNR $= \alpha/\sigma = 2$, where $(\alpha, \sigma) = (6, 3)$. The red dots and crosses mark the finite-sample risks of the orthogonally and i.i.d. sketched estimators, respectively, with $n = 400$, ϕ varying in $[0.1, 10]$, and $p = [n\phi]$. The feature matrix, orthogonal sketching matrices, and i.i.d. sketching matrices are generated in the same way as in Figure 4.

We provide a positive answer to this question by utilizing the variance formula (14). Specifically, the following result demonstrates that the Dirac measure, which corresponds to orthogonal sketching, minimizes the variance formula (14) and therefore minimizes the limiting risks.

Corollary 4.5 (Optimal sketching matrix)**.** *Taking $B = \delta_a$ with some $a > 0$, which corresponds to orthogonal sketching, minimizes the limiting variance* (14)*, and therefore minimizes the limiting risks, among all choices of B supported on the positive real line $\mathbb{R}_{>0}$.*

Figure 4 displays the asymptotic risk curves of sketched ridgeless least square estimators with orthogonal or i.i.d. sketching, under correlated features, as functions of ψ. The figure highlights that, when considering a general feature covariance matrix Σ, employing orthogonal sketching outperforms i.i.d. sketching in the underparameterized regime. However, both approaches yield identical limiting risks in the overparameterized regime. Furthermore, Figure 5 compares the full-sample and sketched least square estimators. It demonstrates that optimal orthogonal and i.i.d. sketching techniques can enhance the stability of the risk curve by eliminating the peaks observed in the risk curves for the full-sample estimator.

5. Conclusions and Discussions

This paper introduces a dual view of overparametrization suggesting that downsampling may also help improve generalization performance. Motiviated by this insight, we investigates the statistical roles of downsampling through random

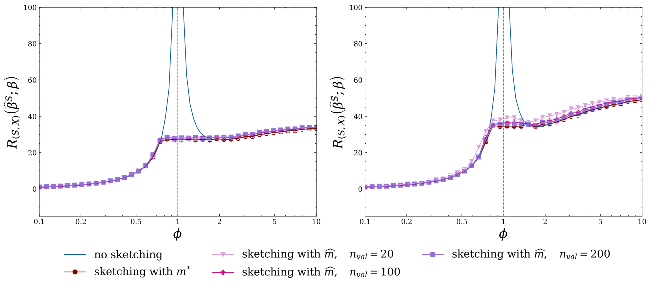

Figure 6. Asymptotic risk curves for the full-sample (no sketching) and sketched ridgeless least square estimators with orthogonal sketching under isotropic and correlated features, respectively, as functions of ϕ. The blue lines in the left panel and the right panel are theoretical risk curves for the full-sample estimator under isotropic features and correlated features, respectively. For both figures, we set $\mathrm{SNR} = \alpha/\sigma = 2$ with $(\alpha, \sigma) = (6, 3)$. The red dots mark finite-sample risks of the sketched estimator with the theoretically optimal sketching size m^*, while the plum triangles, pink diamonds, and purple squares mark finite-sample risks of sketched estimators with the empirically optimal sketching size $\widehat{m}$ determined using the validation datasets of sizes $n_{\text{val}} = 20$, $n_{\text{val}} = 100$, $n_{\text{val}} = 200$, where $n = 400$, ϕ varies in $[0.1, 10]$, and $p = [n\phi]$. The feature matrix and orthogonal sketching matrices are generated in the same way as in Figure 4.

sketching in linear regression estimators, uncovering several intriguing phenomena. First, contrary to conventional beliefs, our findings demonstrate that downsampling does not always harm the generalization performance. In fact, it can be beneficial in certain cases, challenging the prevailing notion. Second, we establish that orthogonal sketching is optimal among all types of sketching considered in the underparameterized regime. In the overparameterized regime however, all general sketching matrices are equivalent. Third, we provide central limit theorems for the risks and discuss the implications of our results for misspecified models, which are presented in the appendix due to space constraints. Lastly, we identify the optimal sketching sizes that minimize the out-of-sample prediction risks under isotropic features. The optimally sketched ridgeless least square estimators exhibit universally better risk curves, indicating their improved stability compared with the full-sample estimator.

Motivated by these findings, we propose a practical procedure to select the optimal sketching size for constructing the optimally sketched estimator, leveraging an additional validation dataset. We provide more details on this procedure in Appendix A. Briefly, we show in Figure 6 that the sketching size selected using the validation dataset leads to risk curves that closely resemble those of the sketched estimator with the optimal sketching size m^*, even with a small validation dataset size of $n_{\text{val}} = 20$. This demonstrates the effectiveness of our proposed procedure in producing sketched estimators with stabler risk curves.

We point out that the benefit of optimal sketching arises from the non-monotonic nature of the risk function with respect to the aspect ratio. Interestingly, recent studies (Hastie et al., 2022) have observed that this non-monotonicity disappears when optimally-tuned ridge regularization is applied. The motivation behind investigating minimum norm estimators, including ridgeless linear regression estimators, stems from the surprising behavior of deep neural networks. Despite lacking explicit regularizers like weight decay or data augmentation, deep neural networks often exhibit a minimal gap between training and test performance (Zhang et al., 2021). The ridgeless least square estimator closely mimics the practice in neural networks, making it an intriguing subject for analysis in the context of linear regression.

Furthermore, comparing with downsampling, the optimally-tuned ridge regression is usually more computationally intensive, as there is no computational reduction from downsampling. Downsampling can provide a potential tool for mitigating the risk with less computational cost. Additionally, we demonstrate in the appendix that, surprisingly, in certain cases, the sketched ridgeless estimator can have a smaller asymptotic variance compared to the full-sample estimator. This is unclear for ridge regression.

As future research directions, it would be interesting to compare the statistical behaviors of ridge and downsampled estimators, as their comparative properties remain unclear. From a broader perspective, viewing downsampling as a form of regularization raises the question of which regularization approach is optimal among all possibilities. Additionally, we hypothesize that Assumption 4 on the sketching matrix can be further relaxed to accommodate cases such as subsampling with replacement, where the limiting spectral distribution of $SS^\top$ contains zero as a mass. Refining the proposed practical procedure with provable guarantees and establishing central limit theorems for more general cases, such as i.i.d. sketching and correlated features, are also promising directions for future exploration.

Acknowledgements

Yicheng Zeng is partially supported by the Shenzhen Outstanding Scientific and Technological Innovation Talents PhD Startup Project (Grant RCBS20221008093336086) and by the Internal Project Fund from Shenzhen Research Institute of Big Data (Grant J00220230012). Siyue Yang and Qiang Sun are partially supported by Natural Sciences and Engineering Research Council of Canada (Grant RGPIN-2018-06484) and a Data Sciences Institute Catalyst Grant.

References

Ailon, N. and Chazelle, B. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 557–563, 2006.

Ba, J., Erdogdu, M., Suzuki, T., Wu, D., and Zhang, T. Generalization of two-layer neural networks: An asymptotic viewpoint. In *Proceedings of the seventh International Conference on Learning Representations*, 2019.

Bai, Z. and Silverstein, J. W. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.

Bai, Z. and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, New York, 2010.

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.

Bartlett, P. L., Long, P. M., 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., Ma, S., and Mandal, S. Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

Canziani, A., Paszke, A., and Culurciello, E. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

Chatterji, N. S. and Long, P. M. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(1): 5721–5750, 2021.

Couillet, R. and Hachem, W. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

Couillet, R. and Liao, Z. *Random Matrix Methods for Machine Learning*. Cambridge University Press, Cambridge, 2022.

Dobriban, E. and Liu, S. Asymptotics for sketching in least squares regression. *arXiv preprint arXiv:1810.06089*, 2018.

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

Drineas, P. and Mahoney, M. W. Lectures on randomized numerical linear algebra. *The Mathematics of Data*, 25 (1), 2018.

El Karoui, N. Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362–2405, 2009.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2013.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the thirty-fifth International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2009.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the twenty-ninth IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Knowles, A. and Yin, J. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1): 257–352, 2017.

Li, Z., Xie, C., and Wang, Q. Asymptotic normality and confidence intervals for prediction risk of the min-norm least squares estimator. In *Proceedings of the thirty-eighth International Conference on Machine Learning*, pp. 6533–6542. PMLR, 2021.

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

Liang, T. and Recht, B. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Mahoney, M. W. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2): 123–224, 2011.

Marcenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Mezzadri, F. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.

Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., and Sahai, A. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(1):10104–10172, 2021.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: on the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Paul, D. and Silverstein, J. W. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

Pilanci, M. *Fast Randomized Algorithms for Convex Optimization and Statistical Estimation*. PhD thesis, University of California, Berkeley, 2016.

Raskutti, G. and Mahoney, M. W. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge (less) regression under general source condition. In *Proceedings of the twenty-fourth International Conference on Artificial Intelligence and Statistics*, pp. 3889–3897. PMLR, 2021.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Yao, J., Zheng, S., and Bai, Z. *Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, Cambridge, 2015.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, L. *Spectral Analysis of Large Dimensional Random Matrices*. PhD thesis, National University of Singapore, Singapore, 2007.

Zheng, S., Bai, Z., and Yao, J. Substitution principle for clt of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *The Annals of Statistics*, 43(2):546–591, 2015.

Appendix

We present a simple yet practical procedure to determine the optimal sketching size in Appendix A. In Appendix B, we extend our results in several directions. The details of our numerical studies are included in Appendix C. We compare the computational cost between the sketched and full-sample estimators in Appendix D. We provide the proofs for results under isotropic features in Appendix E and the proofs for results under correlated features in Appendix F. The proof of Theorem B.1 is provided in Appendix G, while the proofs for the results on central limit theorems are presented in Appendix H.

Throughout the appendix, we use $\|\cdot\|_2$ for the spectral norm of a matrix and use $\|\cdot\|$ for the ℓ_2 norm of a vector.

A. A practical procedure

Determining the optimal sketching size based on the theoretical risk curves requires the knowledge of SNR, which is often unknown in practice. Therefore, estimating the optimal sketching size m^* would require new and potentially complicated methodologies for estimating the SNR, which is beyond the scope of this work.

In this section, we present a simple yet practical procedure to pick the best possible sketching size when we have access to an additional validation dataset. This is not very restrictive, especially in applications with large and streaming data where a validation dataset can be easily obtained. Alternatively, we can manually split the dataset into two parts: a training dataset and a validation dataset. The training dataset is used to obtain the sketched estimators, while the validation dataset is used to select the best sketching size. Finally, the test risk $R_{(S,X)}(\widehat{\beta}^S;\beta)$ can be evaluated on the testing dataset using the tuned sketched least square estimator.

To evaluate the performance of this procedure, we conducted numerical studies with 500 replications. For each replication, we generated $\beta\sim\mathcal{N}_p(0,\frac{\alpha^2}{p}I_p)$ and created a training dataset (X,Y) with $n=400$ training samples, a validation dataset $\{(x_{\text{val},i},y_{\text{val},i}):1\le i\le n_{\text{val}}\}$ with $n_{\text{val}}=\{20,100,200\}$ validation samples, and a testing dataset $\{(x_{\text{new},i},y_{\text{new},i}):1\le i\le n_{\text{new}}\}$ with $n_{\text{new}}=100$ testing samples. The feature matrix $X\in\mathbb{R}^{n\times p}$, orthogonal sketching, and i.i.d. sketching matrices were generated in the same way as in Figure 4 and were fixed across all replications.

Next, we provide details on how the sketching size was selected in each replication and how the empirical out-of-sample prediction risks were calculated.

Selection of the optimal sketching size. The empirically optimal sketching size $\widehat{m}$ was selected if it minimized the empirical risk across a set of values for m evaluated on the validation dataset. Specifically, given fixed p and n, we varied ψ by taking a grid of $\psi\in(0,1)$ with $|\psi_i-\psi_{i+1}|=\delta$ for $\delta=0.05$. This led to a set of potential values for $\widehat{m}$, i.e., $m_i=[\psi_i n]$. For each m_i, we fitted a sketched ridgeless least square estimator $\widehat{\beta}^{S_{m_i}}$ using the training dataset and calculated the empirical risks on the validation dataset:

$$\widehat{R}^{\text{val}}_{(S_{m_i},X)}\left(\widehat{\beta}^{S_{m_i}};\beta\right)=\frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}}\left(x^{\mathrm{T}}_{\text{val},i}\,\widehat{\beta}^{S_{m_i}}-x^{\mathrm{T}}_{\text{val},i}\,\beta\right)^2. \tag{17}$$

The empirical optimal sketching size $\widehat{m}$ was picked as the one that minimized the empirical risks across all m_i.

We briefly discuss the computational cost of using a validation set. For a given m, suppose the computational complexity of orthogonal sketching is $C(np\log m+mp^2)$ where C is a constant. When m varies with $|m_i-m_{i+1}|=[\delta n]$, the total computational complexity would be $\sum_i C(np\log m_i+m_ip^2)\sim C(\frac{1}{\delta}np\log n+\frac{1}{2\delta}np^2)$ where $a_n\sim b_n$ means $\lim_n a_n/b_n=1$. We compare it with ridge regression which also requires a validation set (or CV) to tune the parameter and should have a computational complexity of $C(\frac{1}{\delta}np^2)$. Although they both have the same order, we can still see sketching reduces almost half of the computational cost and the improvement would be significant especially when p is large.

Evaluation of the out-of-sample prediction performance. In the k-th replication, we first generate the coefficient vector $\beta(k)$ if the empirically best sketching size was $\widehat{m}(k)=n$, we fitted a ridgeless least square estimator $\widehat{\beta}(k)$ on the training set; if $\widehat{m}(k)<n$, we fitted a sketched ridgeless least square estimator with the selected $\widehat{m}(k)$. Denote this final estimator by $\widehat{\beta}(k)^{S_{\widehat{m}(k)}}$. The empirical risk of this final estimator was then evaluated on the testing dataset:

$$\widehat{R}_{(S,X)}\left(\widehat{\beta}^S;\beta\right)=\frac{1}{500}\sum_{k=1}^{500}\left\{\frac{1}{n_{\text{new}}}\sum_{r=1}^{n_{\text{new}}}\left(x^{\mathrm{T}}_{\text{new},r}\widehat{\beta}(k)^{S_{\widehat{m}(k)}}-x^{\mathrm{T}}_{\text{new},r}\beta(k)\right)^2\right\}. \tag{18}$$

Figure 6 plots the asymptotic risk curves for the full-sample and sketched least square estimators with orthogonal sketching, correlated features, and the the theoretically and empirically optimal sketching sizes. The performance of the orthogonal sketched estimator with $\widehat{m}$ is comparable to that of sketched estimators with m^* when $n_{\text{val}} = \{20, 100, 200\}$. As the size of the validation dataset increases, the finite-sample risk curve of the orthogonally sketched estimator with $\widehat{m}$ becomes stabler and closer to that of the orthogonally sketched estimator with m^*. Moreover, a particularly small validation dataset with $n_{\text{val}} = 20$ already suffices for producing an estimator with a stable and monotone risk curve.

B. Extensions

B.1. Deterministic β case

Previously, we assume that the coefficient vector β is independent of the data matrix X and follows a multivariate normal distribution $\mathcal{N}_p\left(0, \frac{\alpha^2}{p} I_p\right)$. This section considers deterministic β as specified in the following assumption.

Assumption 5 (Deterministic β)**.** The coefficient vector β is deterministic.

Denote the eigenvalue decomposition of Σ by $\Sigma = \sum_{i=1}^p \lambda_i u_i u_i^{\mathrm{T}}$ where, under Assumption 2, $C_1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq C_0 > 0$. We define the eigenvector empirical spectral distribution (VESD) to be

$$G_n(x) = \frac{1}{\|\beta\|^2} \sum_{i=1}^p \langle \beta, u_i \rangle^2 \mathbf{1}_{[\lambda_i, \infty)}(x), \tag{19}$$

where the indicator function $\mathbf{1}_{[\lambda_i,\infty)}(x)$ takes value 1 if and only if $x \in [\lambda_i, \infty)$. (19) characterizes the relation of Σ and β. Theorem B.1 presents the asymptotic risk when β is deterministic. According to the Lemma 3.1, the variance term is exactly the same as in the previous two subsections. Besides, the bias vanishes in the under-parameterized regime. Thus, the only nontrivial case is the bias for the over-parameterized case. Let

$$c_1 := \frac{\int \frac{x^2 \psi \phi^{-1}}{(c_0 - x\psi\phi^{-1})^2} \, dH(x)}{1 - \int \frac{x^2 \psi \phi^{-1}}{(c_0 - x\psi\phi^{-1})^2} \, dH(x)}, \tag{20}$$

where c_0 is defined in (9). The c_1 can be treated as a rescaled limiting variance of the sketched estimator in the overparameterized regime; see (11).

Theorem B.1. *Assume Assumptions 1, 2, 4, and 5. Then the followings hold.*

(i) If $p/m \to \phi\psi^{-1} < 1$,

$$B_{(\beta,S,X)}(\widehat{\beta}^S; \beta) \overset{\text{a.s.}}{\to} 0. \tag{21}$$

(ii) If $p/m \to \phi\psi^{-1} > 1$ and assume the VESD G_n defined in (19) converges weakly to a probability measure G, then

$$B_{(\beta,S,X)}(\widehat{\beta}^S; \beta) / \|\beta\|^2 \overset{\text{a.s.}}{\to} (1 + c_1) \int \frac{c_0^2 x}{(c_0 - x\psi\phi^{-1})^2} \, dG(x). \tag{22}$$

For the variance term, $V_{(\beta,S,X)}(\widehat{\beta}^S; \beta)$ converges to the same limit as (11) and (14) respectively for the over-parameterized and under-parameterized cases.

(Hastie et al., 2022) obtained a similar result in the case of the full-sample ridgeless least square estimator. Because we are dealing with sketched estimators where additional random sketching matrices are involved, our proofs are more challenging. Specifically, we utilize results for separable covariance matrices. If we further assume $\|\beta\|^2 \to \alpha^2$ and $\Sigma = I_p$, then Theorem B.1 shall recover the same limiting risks in Theorem 3.2.

B.2. Central limit theorem

This subsection establishes central limit theorems for both out-of-sample prediction risks $R_{(\beta,S,X)}(\widehat{\beta}^S; \beta)$ and $R_{(S,X)}(\widehat{\beta}^S; \beta)$. (Li et al., 2021) studies the central limit theorems for risks of full-sample ridgeless least square estimator. Compared with their work, our results show the risks of sketched estimators may have smaller asymptotic variances. We start with the following assumptions.

Assumption 6 (Random β)**.** The coefficient vector β follows a multivariate normal distribution $\mathcal{N}_p\left(0, \frac{\alpha^2}{p} I_p\right)$, and is independent of the data matrix X, the noise ε, and the sketching matrix S.

Assumption 7. Suppose $\{X_{ij}\}$ share the fourth moment $\nu_4 := \mathbb{E}|X_{ij}|^4 < \infty$. Furthermore, they satisfy the following Lindeberg condition

$$\frac{1}{np} \sum_{1 \le i \le n,\ 1 \le j \le p} \mathbb{E}\left(|X_{ij}|^4 \mathbf{1}_{[\eta\sqrt{n}, \infty)}(|X_{ij}|)\right) \to 0, \quad \text{for any fixed } \eta > 0,$$

where the indicator function $\mathbf{1}_{[\eta\sqrt{n}, \infty)}(|X_{ij}|)$ takes value 1 if and only if $|X_{ij}| \ge \eta\sqrt{n}$.

CLTs for $R_{(S,X)}(\widehat{\beta}^S; \beta)$. The following theorems give CLTs for $R_{(S,X)}(\widehat{\beta}^S; \beta)$ in the underparameterized and overparameterized regimes. Recall that $m, n, p \to \infty$ such that $\phi_n = p/n \to \phi$ and $\psi_n = m/n \to \psi \in (0, 1)$.

Theorem B.2. *Assume Assumptions 1, 2, 4, 6, and 7. Suppose $\phi\psi^{-1} < 1$ and S is an orthogonal sketching matrix. Then it holds that*

$$p\left(R_{(S,X)}(\widehat{\beta}^S; \beta) - \frac{\sigma^2 \phi_n \psi_n^{-1}}{1 - \phi_n \psi_n^{-1}}\right) \xrightarrow{D} \mathcal{N}(\mu_1, \sigma_1^2),$$

where

$$\mu_1 = \frac{\sigma^2 \phi^2 \psi^{-2}}{(\phi\psi^{-1} - 1)^2} + \frac{\sigma^2 \phi^2 \psi^{-2}(\nu_4 - 3)}{1 - \phi\psi^{-1}}, \quad \sigma_1^2 = \frac{2\sigma^4 \phi^3 \psi^{-3}}{(\phi\psi^{-1} - 1)^4} + \frac{\sigma^4 \phi^3 \psi^{-3}(\nu_4 - 3)}{(1 - \phi\psi^{-1})^2}.$$

Theorem B.3. *Assume Assumptions 1, 6, 7 and $\Sigma = I_p$. Suppose $\phi\psi^{-1} > 1$ and S is any sketching matrix that satisfies Assumption 4. Then it holds that*

$$p\left(R_{(S,X)}(\widehat{\beta}^S; \beta) - \alpha^2(1 - \psi_n \phi_n^{-1}) - \frac{\sigma^2}{\phi_n \psi_n^{-1} - 1}\right) \xrightarrow{D} \mathcal{N}(\mu_2, \sigma_2^2),$$

where

$$\mu_2 = \frac{\sigma^2 \phi \psi^{-1}}{(\phi\psi^{-1} - 1)^2} + \frac{\sigma^2(\nu_4 - 3)}{\phi\psi^{-1} - 1}, \quad \sigma_2^2 = \frac{2\sigma^4 \phi^3 \psi^{-3}}{(\phi\psi^{-1} - 1)^4} + \frac{\sigma^4 \phi \psi^{-1}(\nu_4 - 3)}{(\phi\psi^{-1} - 1)^2}.$$

The CLT of $R_{(S,X)}(\widehat{\beta}^S; \beta)$ after an orthogonal sketching $(X, Y) \mapsto (SX, SY)$ coincides with that by (Li et al., 2021) after replacing p/n by p/m. According to Theorems B.2, B.3 and 3.3, we provide the asymptotic variance of $R_{(S,X)}(\widehat{\beta}^S; \beta)$ for the orthogonal sketched estimator with the optimal sketching size m^* given by Theorem 3.3.

Corollary B.4. *Denote the asymptotic variance of the risk $R_{(S,X)}(\widehat{\beta}^S; \beta)$ for the orthogonal sketched estimator with the optimal sketching size m^* by σ_S^2. The followings hold.*

(a) If $\mathrm{SNR} > 1$ and $\phi \in (1 - \frac{\sigma}{2\alpha}, \frac{\alpha}{\alpha - \sigma}]$, then $\sigma_S^2 = 2\alpha^3(\alpha - \sigma) + \sigma^2(\nu_4 - 3)\alpha(\alpha - \sigma)$.

(b) If $\mathrm{SNR} \le 1$ and $\phi \in (\frac{\alpha^2}{\alpha^2 + \sigma^2}, \infty)$, then $\sigma_S^2 = O(\frac{m^*}{n}) \to 0$.

(c) If either of the following two holds: (i) $\mathrm{SNR} \le 1$ and $\phi \in (0, \frac{\alpha^2}{\alpha^2 + \sigma^2}]$, or (ii) $\mathrm{SNR} > 1$ and $\phi \in (0, 1 - \frac{\sigma}{2\alpha}] \bigcup (\frac{\alpha}{\alpha - \sigma}, \infty)$, then

$$\sigma_S^2 = \begin{cases} \dfrac{2\sigma^4 \phi^3}{(\phi - 1)^4} + \dfrac{\sigma^4 \phi^3 (\nu_4 - 3)}{(1 - \phi)^2}, & \text{if } \phi < 1, \\ \dfrac{2\sigma^4 \phi^5}{(\phi - 1)^4} + \dfrac{\sigma^4 \phi^3 (\nu_4 - 3)}{(\phi - 1)^2}, & \text{if } \phi > 1. \end{cases}$$

Comparing the asymptotic variance σ_S^2 of $R_{(S,X)}(\widehat{\beta}^S; \beta)$ with optimal sketching and that of $R_{S,X}(\widehat{\beta}; \beta)$ without sketching, we have following observations. First, the non-trivial and optimal sketching in case (a) may result in a smaller asymptotic variance σ_S^2 than that for the full-sample estimator. Take standard Gaussian features with $\phi > 1$, for which the forth (central) moment ν_4 is 3, as an example. Then it can be verified that $\sigma_S^2 \le 2\sigma^4 \phi^3 (\phi - 1)^{-4} =: \sigma_0^2$ for $\phi \in (1, \alpha(\alpha - \sigma)^{-1})$ and $\sigma_S^2 \le \sigma^4 (2\phi - 1)(1 - \phi)^{-4}/8 < \sigma_0^2$ for $\phi \in (1 - \sigma\alpha^{-1}/2, 1]$ when $\mathrm{SNR} > 1$. Second, the trivial sketching in case (b) has a zero limiting variance because in this case the null estimator $\widetilde{\beta} = 0$ is optimal.

CLTs for $R_{(\beta,S,X)}(\widehat{\beta}^S;\beta)$. In the underparameterized regime, for sufficiently large n, $B_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) \sim B_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \sim 0$, and $V_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right) \sim V_{(S,X)}\left(\widehat{\beta}^S;\beta\right)$. Thus, $B_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right)$ has exactly the same CLT as $B_{(S,X)}\left(\widehat{\beta}^S;\beta\right)$ in Theorem B.2. We now present the corresponding CLT in the overparameterized case.

Theorem B.5. *Assume Assumptions 1, 6, 7 and $\Sigma = I_p$. Suppose $\phi\psi^{-1} > 1$ and S is any sketching matrix S that satisfies Assumption 4. Then it holds that*

$$\sqrt{p}\left(R_{(\beta,S,X)}(\widehat{\beta}^S;\beta) - \alpha^2(1-\psi_n\phi_n^{-1}) - \frac{\sigma^2}{\phi_n\psi_n^{-1}-1}\right) \xrightarrow{D} \mathcal{N}(\mu_3,\sigma_3^2),$$

where

$$\mu_3 = 0, \quad \sigma_3^2 = 2(1-\phi^{-1}\psi)\alpha^4.$$

More precise versions of μ_3 and σ_3^2 are

$$\widetilde{\mu}_3 = \frac{1}{\sqrt{p}}\left(\frac{\sigma^2\phi\psi^{-1}}{(\phi\psi^{-1}-1)^2} + \frac{\sigma^2(\nu_4-3)}{\phi\psi^{-1}-1}\right),$$
$$\widetilde{\sigma}_3^2 = 2(1-\phi^{-1}\psi)\alpha^4 + \frac{1}{p}\left(\frac{2\sigma^4\phi^3\psi^{-3}}{(\phi\psi^{-1}-1)^4} + \frac{\sigma^4\phi\psi^{-1}(\nu_4-3)}{(\phi\psi^{-1}-1)^2}\right).$$

B.3. Misspecified model

This subsection briefly discusses the misspecified model. When the misspecification error, aka model bias, is included, the risk will decrease at first and then increase for the full-sample ridgeless least square estimator in the underparameterized case. This aligns with the classic statistical idea of "underfitting" and "overfitting". This subsection studies the effect of sketching on the selection of the optimal sketching size.

We consider a misspecified in which we observe only a subset of the features. A similar model is also discussed in the section 5.1 of (Hastie et al., 2022). Suppose the true model is

$$y_i = \beta^{\mathrm{T}}x_i + \theta^{\mathrm{T}}w_i + \varepsilon_i,\ i = 1,\cdots,n \tag{23}$$

where $x_i \in \mathbb{R}^p, w_i \in \mathbb{R}^q$ and the noise ε_i is independent of (x_i, w_i). Further assume (x_i, w_i) are jointly Gaussian with mean zero and covariance matrix

$$\Sigma = \begin{bmatrix}\Sigma_{xx}, \Sigma_{xw}\\ \Sigma_{xw}^{\mathrm{T}}, \Sigma_{ww}\end{bmatrix}.$$

We can only observe the data matrix $X = (x_1,\cdots,x_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$. Still, we use the sketched data $\widetilde{Y} := SY \in \mathbb{R}^m$, $\widetilde{X} := SX \in \mathbb{R}^{m\times p}$ and its corresponding minimum-norm least square estimator $\widehat{\beta}^S$ defined in (2). Let $(x_{\text{new}}, w_{\text{new}})$ be a test point. The out-of-sample prediction risk is defined as

$$R_{(S,X)}\left(\widehat{\beta}^S;\beta,\theta\right) = \mathbb{E}\left[\left(x_{\text{new}}^{\mathrm{T}}\widehat{\beta}^S - x_{\text{new}}^{\mathrm{T}}\beta - w_{\text{new}}^{\mathrm{T}}\theta\right)^2 \middle| S, X\right].$$

Here we let β and θ are nonrandom parameters and the expectation is taken over $x_{\text{new}}, w_{\text{new}}, \epsilon$ and also $W = (w_1,\cdots,w_n)^{\mathrm{T}} \in \mathbb{R}^{n\times q}$. Similar to lemma 2 in (Hastie et al., 2022), we can decompose the risk into two terms,

$$R_{(S,X)}\left(\widehat{\beta}^S;\beta,\theta\right) = \underbrace{\mathbb{E}\left[\left(x_{\text{new}}^{\mathrm{T}}\widehat{\beta}^S - \mathbb{E}\left(y_{\text{new}}|x_{\text{new}}\right)\right)^2 \middle| S, X\right]}_{R^*_{(S,X)}\left(\widehat{\beta}^S;\beta,\theta\right)} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}\left(y_{\text{new}}|x_{\text{new}}\right) - \mathbb{E}\left(y_{\text{new}}|x_{\text{new}}, w_{\text{new}}\right)\right)^2\right]}_{M(\beta,\theta)},$$

where $M(\beta,\theta)$ can be seen as the misspecification bias. Notice that conditioning on x_i, model (23) is equivalent to $y_i = \widetilde{\beta}^{\mathrm{T}}x_i + \widetilde{\epsilon}_i$ where $\widetilde{\beta} = \beta + \Sigma_{xx}^{-1}\Sigma_{xw}\theta$ and $\widetilde{\epsilon}_i \sim N(0,\widetilde{\sigma}^2)$ is independent of x_i, $\widetilde{\sigma}^2 = \sigma^2 + \theta^{\mathrm{T}}\Sigma_{w|x}\theta$. Here

$\Sigma_{w|x}$ is the covariance matrix of w_i given x_i, i.e., $\Sigma_{w|x} = \Sigma_{ww} - \Sigma_{xw}^{\mathrm{T}}\Sigma_{xx}^{-1}\Sigma_{xw}$. Moreover, simple calculation shows $M(\beta,\theta) = \theta^{\mathrm{T}}\Sigma_{w|x}\theta$. We refer readers to Remark 2 in (Hastie et al., 2022) for more details.

We conclude that even for this misspecified model, since $M(\beta,\theta)$ is independent of the sketching matrix S, and $R^*_{(S,X)}\left(\widehat{\beta}^S;\beta,\theta\right)$ can still be approximated using Theorem B.1, random sketching cannot improve the limiting risks by sketching the estimator to the under-parameterized regime. We expect in more complicated models, for example, the random feature model in (Mei & Montanari, 2022), sketching to the underparameterized regime might help reduce the limiting risks. We leave this problem to the future.

C. Details on numerical studies

C.1. Numerical studies for isotropic features

This section provides additional details on the numerical studies for isotropic features to replicate Figures 2 and 3.

C.1.1. Figure 2

For Figure 2, numerical simulations were run 500 replications. For each replication, we generated $\beta \sim \mathcal{N}_p\left(0, \frac{\alpha^2}{p}I_p\right)$ and a training dataset (X,Y) with $n = 400$ training samples, and a testing dataset $\{(x_{\text{new},i}, y_{\text{new},i}) : 1 \le i \le n_{\text{new}}\}$ with $n_{\text{new}} = 100$ testing samples. The feature, orthogonal sketching, and i.i.d. sketching matrices were generated first and then fixed across all replications. The orthogonal sketching matrix was generated using subsampled randomized Hadamard transform, which relies on the fast Fourier transform. This approach is commonly regarded as a rapid and reliable method for implementing sketching algorithms (Dobriban & Liu, 2018). The feature matrix and i.i.d. sketching matrices were generated using Python library `NumPy`. Other details are given in the caption of Figure 2. Our implementation is available at https://github.com/statsle/SRLR_python.

The finite-sample risks, aka the dots and crosses in Figure 2, were calculated as functions of ψ. Specifically, given fixed n, we varied ψ by taking a grid of $\psi \in (0,1)$ with $|\psi_i - \psi_{i+1}| = \delta$ for $\delta = 0.05$. This led to a grid of values for m, i.e., $m_i = [\psi_i n]$. For each replication k, we randomly generated a coefficient vector $\beta(k)$. For each replication k, we first randomly generated a coefficient vector $\beta(k)$. Within replication k and for each m_i, we fitted a sketched ridgeless least square estimator $\widehat{\beta}(k)^{S_{m_i}}$ using the training dataset and calculated the empirical risks on the testing dataset:

$$\widehat{R}_{(S_{m_i},X)}\left(\widehat{\beta}^{S_{m_i}};\beta\right) = \frac{1}{500}\sum_{k=1}^{500}\left\{\frac{1}{n_{\text{new}}}\sum_{r=1}^{n_{\text{new}}}\left(x_{\text{new}_r}^{\mathrm{T}}\widehat{\beta}(k)^{S_{m_i}} - x_{\text{new}_r}^{\mathrm{T}}\beta(k)\right)^2\right\}. \tag{24}$$

C.1.2. Figure 3

The finite-sample risks, aka the dots in Figure 3, were calculated as functions of ϕ. Numerical simulation procedure and data generation followed Section C.1.1. The optimal sketching size m^* was selected based on Theorem 3.3. If $m^* = n$, we fitted a ridgeless least square estimator $\widehat{\beta}$ on the training set; if $m^* < n$, we fitted a sketched estimator $\widehat{\beta}^S$ with m^*. The empirical risks $\widehat{R}_{(S,X)}\left(\widehat{\beta}^S;\beta\right)$ were evaluated on the testing dataset in a similar way as in Equation (24). To indicate how SNR and ϕ influence the selection of m^*, the left panel of Figure 3 presents risks for SNR < 1, and the right panel presents risks for SNR > 1.

C.2. Numerical studies for correlated features

This section provides additional details on numerical studies for correlated features to replicate Figures 4 and 5.

C.2.1. Figure 4

The numerical simulation procedure generally followed Section C.1.1. Instead of isotropic features, we generated correlated features. Other details are given in the caption of Figure 4.

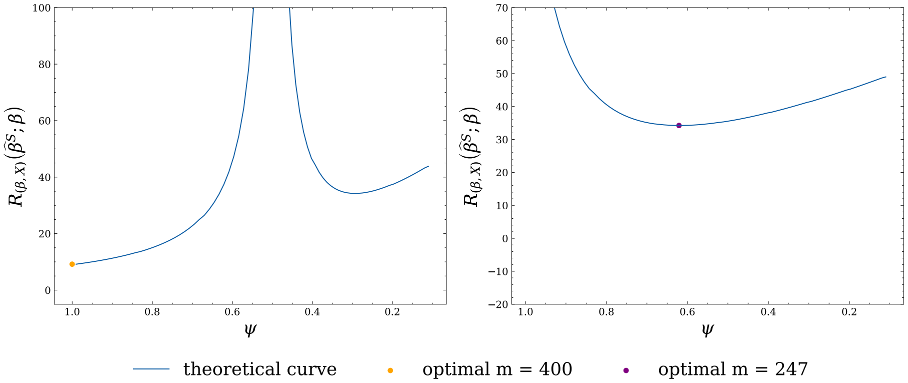

Figure 7. Asymptotic risk curves for sketched ridgeless least square estimators with correlated features, orthogonal sketching, as functions of ψ. The lines are theoretical risk curves for $n = 400$ with $p = 200$ and $p = 424$ respectively, where $\text{SNR} = \alpha/\sigma = 2$ with $(\alpha, \sigma) = (6, 3)$, ψ varying in $(0, 1)$, and $m = [n\psi]$. The dot marks the minimum of a theoretical risk curve within $\psi \in (0, 1)$.

C.2.2. Figure 5

The simulation procedure followed Section C.1.2 and the data generation followed Section C.2.1. In the case of correlated features, the theoretically optimal sketching size m^* does not have a closed-form representation, and m^* can be picked by minimizing the theoretical risk function across a set of values for m. Specifically, given fixed p and n, we varied ψ by taking a grid of $\psi \in (0, 1)$ with $|\psi_i - \psi_{i+1}| = \delta$ for $\delta = 0.05$. This led to a set of potential values for m^*, i.e., $m_i = [\psi_i n]$. For each m_i, we calculated the negative solutions c_0 in (9) and (12) numerically using the functon `fsolve` in the Python library `SciPy`. These values of c_0 were then used to generate the theoretical risk curves in the overparameterized and underparameterized regime as described in Theorem 4.2 and Theorem 4.3, respectively. The optimal sketching size m^* was selected as the one that minimized the theoretical risks across all m_i. With the optimal sketching size m^*, the empirical risks were calculated the same way as in Section C.1.2.

We further illustrate how the sketching size was selected using two examples shown in Figure 7. For $p = 200$ in the left panel, the risk attained minimum at $\psi \approx 1$ and $m^* = 400$, so no sketching was needed. For $p = 424$ in the right panel, the risk attained minimum at $\psi \approx 0.6175$ and we set $m^* = 0.6175 \times 400 = 247$.

D. Computational cost

We analyze the computational cost when the optimal sketching size m^* is given *a priori*. The time for the full sketching and orthogonal sketching (realized by the subsampled randomized Hadamard transform) is

$$t_{\text{full}} = C_1 np^2, \; t_{\text{orthogonal}} = C_2 pn \log n + C_3 m^* p^2.$$

where C_1, C_2, C_3 are some constants. It is clear that the optimal orthogonal sketching can reduce computational costs when the condition $\frac{C_3}{C_1}\frac{m^*}{n} + \frac{C_2}{C_1}\frac{\log n}{p} < 1$ is satisfied. This condition is typically met in the overparameterized regime, where p is large compared to n.

We conducted timing experiments on a MacMini with an Apple M1 processor and 16GB of memory to measure the computational time required for the full-sample (no sketching) and sketched ridgeless least square estimators with orthogonal sketching (implemented through the subsampled randomized Hadamard transform) under isotropic features. These experiments were designed to investigate the impact of the sketching size m, sample size n, and feature dimension p on the computational time. To mitigate variations resulting from runtime disparities, we computed the average time from ten separate runs.

Figure 8 compares the run time in seconds for different values of p in both the underparameterized and overparameterized regimes, with a fixed sample size of $n = 10,000$. The figure demonstrates a significant computational benefit of sketching in the overparameterized regime. In this regime, as the feature dimension p further deviates from the sample size n, sketching becomes increasingly time-efficient. Notably, when $p = 11,500$, sketching saves time for almost every ψ.

In Figure 9, we fix the aspect ratio $\phi = p/n$ and the SNR to be $\text{SNR} = \alpha/\sigma = 3$ with $(\alpha, \sigma) = (6, 2)$, which implies a

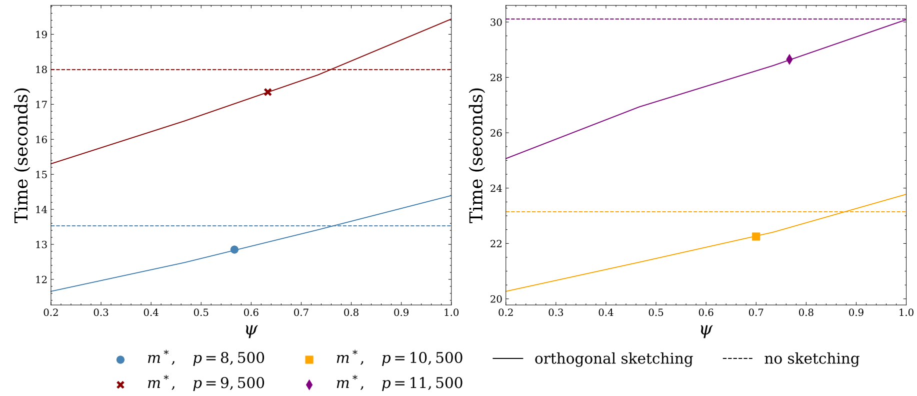

Figure 8. Computational costs associated with the full-sample (no sketching) and sketched ridgeless least square estimators with orthogonal sketching under isotropic features, as functions of ψ, for a fixed sample size of $n = 10,000$ and varying p. The left panel shows the time required for the full-sample and orthogonally sketched estimators in the underparameterized regime, represented by the dotted and solid lines, respectively. The right panel depicts the time for estimators in the overparameterized regime. The dots and crosses mark the computational time required for the sketched estimators with the optimal sketching size m^*. We set $\mathrm{SNR} = \alpha/\sigma = 3$ with $(\alpha, \sigma) = (6, 2)$.

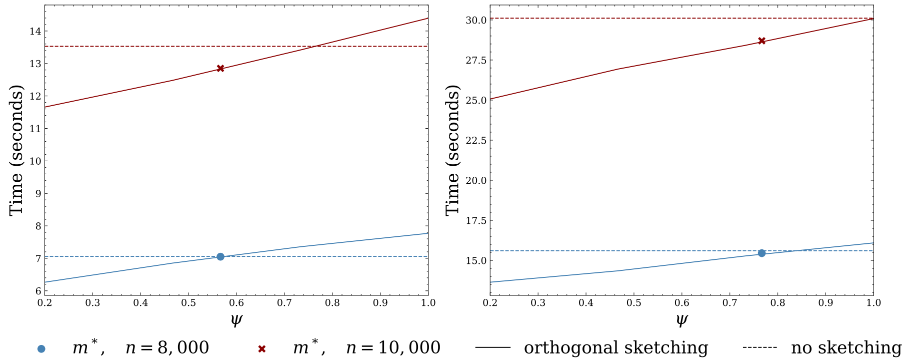

Figure 9. Computational costs associated with the full-sample (no sketching) and sketched ridgeless least square estimators with orthogonal sketching under isotropic features, as functions of ψ, for fixed $\phi = p/n$. The left panel shows the time required for the full-sample and orthogonally sketched estimators in the underparameterized regime with $\phi = 0.85$, represented by the dotted and solid lines, respectively. The right panel depicts the time for estimators in the overparameterized regime with $\phi = 1.15$. The dots and crosses mark the computational time required for the sketched estimators with the optimal sketching size m^*. We set $\mathrm{SNR} = \alpha/\sigma = 3$ with $(\alpha, \sigma) = (6, 2)$.

fixed optimal sketching ratio $\psi^* := m^*/n$, while varying the sample size n. In this scenario, as the sample size n increases, the optimal orthogonal sketching becomes even more time-efficient. This observation encourages the use of sketching when dealing with large sample sizes, which aligns with our intuition.

Figure 10 illustrates a scenario with a fixed aspect ratio ϕ and different SNR, which results in smaller values for the optimal sketching size m^*. In such cases, employing the optimal orthogonal sketching significantly reduces computational time.

In closing, we expect even larger computational improvements when using sketching in more complex models, such as neural networks, while simultaneously mitigating the out-of-time prediction risk. We leave these experiments for future research.

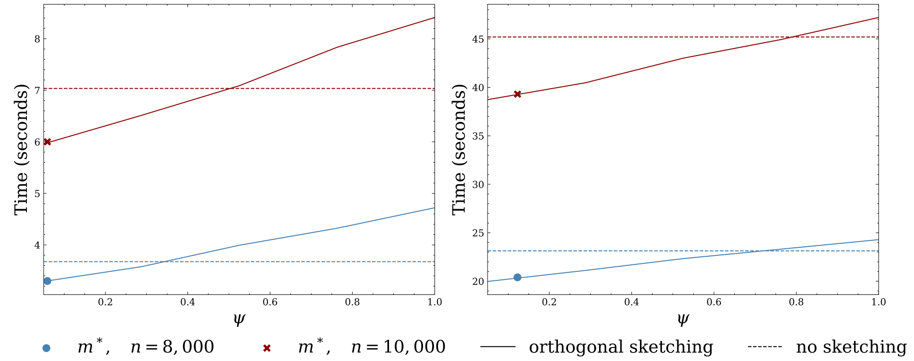

Figure 10. Computational costs associated with the full-sample (no sketching) and sketched ridgeless least square estimators with orthogonal sketching under isotropic features , as functions of ψ, for fixed $\phi = p/n$. The left panel shows the time required for the full-sample and orthogonally sketched estimators in the underparameterized regime with $\phi = 0.65$, represented by the dotted and solid lines, respectively. The right panel depicts the time for estimators in the overparameterized regime with $\phi = 1.35$. The dots and crosses mark the computational time required for the sketched estimators with optimal sketching size m^*. We set $\mathrm{SNR} = \alpha/\sigma = 1.1$, where $(\alpha, \sigma) = (22, 20)$.

E. Proofs for isotropic features

E.1. Proof of Lemma 3.1

Because

$$\mathbb{E}\left(\widehat{\beta}^S|\beta, S, X\right) = (X^\top S^\top SX)^+ X^\top S^\top SX\beta,$$
$$\mathrm{Cov}\left(\widehat{\beta}^S|\beta, S, X\right) = \sigma^2 (X^\top S^\top SX)^+ X^\top S^\top SS^\top SX(X^\top S^\top SX)^+,$$

we can derive the expressions of $B_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right)$, $V_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right)$ and $V_{(S,X)}\left(\widehat{\beta}^S;\beta\right)$ from their respective definitions. $B_{(S,X)}\left(\widehat{\beta}^S;\beta\right)$ expression follows from the formula of the expectation of quadratic form and the fact that $(X^\top S^\top SX)^+ X^\top S^\top SX$ is idempotent.

Now since the eigenvalues of $I_p - (X^\top S^\top SX)^+ X^\top S^\top SX$ are either 0 or 1, the eigenvalues of $[(X^\top S^\top SX)^+ X^\top S^\top SX - I_p]\,\Sigma\,[(X^\top S^\top SX)^+ X^\top S^\top SX - I_p]$ are uniformly bounded over $[0, C_1]$. Then (8) can be obtained by applying (Bai & Silverstein, 2010, Lemma B.26).

E.2. Proof of Theorem 3.2

The proof for the underparameterized case directly follows from the Corollary 4.4 since the bias vanishes. As for the overparameterized case, according to Theorem 4.2, when $\Sigma = I_p$, a simple calculation shows $c_0 = \psi\phi^{-1} - 1$, and $B_{(S,X)}(\widehat{\beta}^S;\beta) \to \alpha^2(1 - \psi\phi^{-1})$, $V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \to \sigma^2(\phi\psi^{-1} - 1)^{-1}$, which then leads to the desired results. We collect the proofs for Corollary 4.4 and Theorem 4.2 in Sections F.2.2 and F.1.2 respectively.

E.3. Proof of Theorem 3.3

We prove this theorem for orthogonal and i.i.d. sketching separately. We start with orthogonal sketching.

Orthogonal sketching We start with orthogonal sketching first. Let

$$f(x) = \alpha^2(1 - x^{-1}) + \frac{\sigma^2}{x - 1},\ x > 1.$$

According to Theorem 3.2, for orthogonal sketching, both limiting risks in the overparameterized regime are $f(\phi\psi^{-1})$.

For the case of $\alpha \leq \sigma$, i.e., SNR ≤ 1,

$$f'(x) = \alpha^2 x^{-2} - \frac{\sigma^2}{(x-1)^2} = \frac{(\alpha^2 - \sigma^2)x^2 - 2\alpha^2 x + \alpha^2}{x^2(x-1)^2} < 0, \ \forall x > 1,$$

and $\lim_{x\to\infty} f(x) = \alpha^2$. Thus, if $\phi \geq 1$, $f(\phi\psi^{-1})$ decreases as ψ decreases. If $\phi < 1$, the limiting risks without sketching ($\psi = 1$) are $\frac{\sigma^2\phi}{1-\phi}$, which exceeds α^2 if and only if $\phi > \frac{\alpha^2}{\alpha^2+\sigma^2}$. So the optimal sketching size is $m^* \ll n$ if $\phi > \frac{\alpha^2}{\alpha^2+\sigma^2}$, and $m^* = n$ otherwise.

For the case of $\alpha > \sigma$, i.e., SNR > 1, $f(x)$ decreases when $x \in (1, \frac{\alpha}{\alpha-\sigma})$, and increases when $x \in [\frac{\alpha}{\alpha-\sigma}, \infty)$, and $f(\frac{\alpha}{\alpha-\sigma}) = \sigma(2\alpha - \sigma)$. Thus, if $\frac{\alpha}{\alpha-\sigma} \geq \phi > 1$, orthogonal sketching can help reduce the limiting risks to $\min_{x>1} f(x)$. If $\phi < 1$, the same improvement holds if and only if $\frac{\sigma^2\phi}{1-\phi} > \sigma(2\alpha-\sigma)$, or equivalently $\phi > 1 - \frac{\sigma}{2\alpha}$. Thus, the optimal sketching size is $m^* = \phi\frac{\alpha-\sigma}{\alpha}\cdot n = \frac{\alpha-\sigma}{\alpha}\cdot p$ if $1 - \frac{\sigma}{2\alpha} < \phi \leq \frac{\alpha}{\alpha-\sigma}$; and $m^* = n$ otherwise.

i.i.d. sketching Because in the underparameterized case, sketching always increases the risk which follows the classic statistical intuition: a larger sample size is better. In other words, only sketching to the overparameterized case can help reduce the risk. Because i.i.d. sketching shares the same limiting risks with orthogonal sketching in the overparameterized regime, it has the same optimal sketching size.

F. Proofs for correlated features

F.1. Proofs for the over-parameterized case

F.1.1. Proof of Lemma 4.1

Let $f(c) = 1 - \int \frac{x}{-c+x\psi\phi^{-1}}\, dH(x)$. Because $f(0) = 1 - \phi\psi^{-1} < 0$, $f(-\infty) = 1$, and f is smooth, f has at least one negative root. Suppose c_1 and c_2 are two negative roots with $c_1 < c_2$. Then we have

$$0 = f(c_1) - f(c_2) = \int \frac{x(c_2 - c_1)}{(-c_2 + x\psi\phi^{-1})(-c_1 + x\psi\phi^{-1})}\, dH(x) > 0,$$

where the last inequality follows from the fact that the numerator and denominator are both larger than 0. This is a contradiction and thus f has a unique negative root.

F.1.2. Proof of Theorem 4.2

Bias part To prove the bias part (10), we first need some lemmas. Lemmas F.1 and F.2 show that the minimal nonzero eigenvalues of conrresponding matrices are lower bounded, which will be used to guarantee to exchange limits. Lemma F.1 also proves, in the overparameterized case, $\frac{1}{p}SZZ^{\mathrm{T}}S^{\mathrm{T}}$ is invertible almost surely for all large n.

Lemma F.1. *Let $Z \in \mathbb{R}^{n\times p}$ be a matrix with i.i.d. entries Z_{ij} such that $\mathbb{E}[Z_{ij}] = 0$, $\mathbb{E}[Z_{ij}^2] = 1$, and $\mathbb{E}[Z_{ij}^4] < \infty$. Assume Assumption 4 and suppose $m, n, p \to \infty$ such that $p/n \to \phi$, $m/n \to \psi \in (0, 1)$. Then, there exists some constant $\tau > 0$ such that almost surely for all large n, it holds that $\lambda_{\min}^{+}\left(\frac{1}{p}SZZ^{\mathrm{T}}S^{\mathrm{T}}\right) = \lambda_{\min}^{+}\left(\frac{1}{p}Z^{\mathrm{T}}S^{\mathrm{T}}SZ\right) \geq \tau$, where $\lambda_{\min}^{+}$ denotes the smallest nonzero eigenvalue. Furthermore, if (i) $\phi\psi^{-1} > 1$, then almost surely for all large n, $\frac{1}{p}SZZ^{\mathrm{T}}S^{\mathrm{T}}$ is invertable; (2) $\phi\psi^{-1} < 1$, then almost surely for all large n, $\frac{1}{p}Z^{\mathrm{T}}S^{\mathrm{T}}SZ$ is invertible.*

Proof of Lemma F.1. Denote the limiting spectral measure of $\frac{1}{p}Z^{\mathrm{T}}S^{\mathrm{T}}SZ$ by μ. By (Yao et al., 2015, Proposition 2.17), the support of μ is completely determined by $\Psi(\alpha)$, known as the functional inverse of the function $a(x) := -1/s(x)$, where $s(x)$ is the Stieltjes transform of μ. Specifically, if we let Γ be the support of μ, then $\Gamma^c \cap (0, \infty) = \{\Psi(a) : \Psi'(a) > 0\}$. Under Assumption 4, we can assume that the ESD of $S^{\mathrm{T}}S$ converges to a nonrandom measure $\underline{B}$, which is the companion of B. Then, $\Psi(a) = a + \phi^{-1}a \int \frac{t}{a-t}\, d\underline{B}(t)$, and hence $\Psi'(a) = 1 + \phi^{-1}\int \frac{t}{a-t}\, d\underline{B}(t) - \phi^{-1}a\int \frac{t}{(a-t)^2}\, d\underline{B}(t)$, which is smooth.

(i) If $\phi^{-1}\psi < 1$, then $\lim_{a\to 0^+} \Psi'(a) = 1 - \phi^{-1}(1 - \underline{B}(\{0\})) = 1 - \phi^{-1}\psi > 0$. Thus, there exists some small enough $\epsilon > 0$ such that Ψ is increasing on $(0, \epsilon)$. Besides, under Assumption 4, the support of $\underline{B}$ is a subset of $\{0\} \cup [\widetilde{C}_0, \widetilde{C}_1]$. Thus, when ϵ is small enough, Ψ is well defined on $(0, \epsilon)$. Since $\lim_{a\to 0^+} \Psi(a) = 0$ and Ψ' is smooth, we know that there exists some $\tau > 0$ such that $\{\Psi(a) : \Psi'(a) > 0\} \supseteq (0, \tau)$.

(ii) If $\phi^{-1}\psi > 1$, then $\lim_{a\to 0^-}\Psi'(a) < 0$, and hence we can find some small enough ϵ such that Ψ is decreasing on $(-\epsilon, 0)$. Since $\lim_{a\to-\infty}\Psi(a) = -\infty$, and by the smoothness of Ψ, we know $\{\Psi(a) : \Psi'(a) > 0\} \supseteq (0, \Psi(-\epsilon))$. Overall, by combining these two situations and using (Bai & Silverstein, 1998, Theorem 1.1), we can show that there exists some $\tau > 0$ such that almost surely for all large n, $\lambda_{\min}^{+}(\frac{1}{p}Z^{\mathrm{T}}S^{\mathrm{T}}SZ) \geq \tau$.

To prove the invertibility of $\frac{1}{p}SZZ^{\mathrm{T}}S^{\mathrm{T}}$ when $\phi\psi^{-1} > 1$, we first denote the limiting spectral measure of $\frac{1}{p}SZZ^{\mathrm{T}}S^{\mathrm{T}}$ by $\underline{\mu}$. With (Couillet & Hachem, 2014, Proposition 2.2), it holds that $\underline{\mu}\left(\{0\}\right) = 1 - \min\{1 - B\left(\{0\}\right), \frac{n}{m}\min\{\frac{p}{n}, 1\}\} = 0$ since $B\left(\{0\}\right) = 0$ under Assumption 4. We thus obtain the invertibility of $\frac{1}{p}SZZ^{\mathrm{T}}S^{\mathrm{T}}$ almost surely for all large n by using again (Bai & Silverstein, 1998, Theorem 1.1). When $\phi\psi^{-1} < 1$, the invertibility of $\frac{1}{p}Z^{\mathrm{T}}S^{\mathrm{T}}SZ$ follows from a similar argument. □

Lemma F.2. *Let $a, b > 0$ be two positive constants. Let $A \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix such that $\lambda_{\min}^{+}(A) \geq a$ and $\Sigma \in \mathbb{R}^{n\times n}$ a positive definite matrix such that $\lambda_{\min}(\Sigma) \geq b$. Then, $\lambda_{\min}^{+}\left(\Sigma^{1/2}A\Sigma^{1/2}\right) \geq ab$.*

Proof of Lemma F.2. The result follows from

$$\begin{aligned}\lambda_{\min}^{+}\left(\Sigma^{1/2}A\Sigma^{1/2}\right) &\geq \min_{x\in\mathbb{R}^n: x^{\mathrm{T}}\Sigma^{1/2}A\Sigma^{1/2}x\neq 0} \frac{x^{\mathrm{T}}\Sigma^{1/2}A\Sigma^{1/2}x}{\|x\|^2} \\ &\geq \min_{x\in\mathbb{R}^n: x^{\mathrm{T}}\Sigma^{1/2}A\Sigma^{1/2}x\neq 0} \frac{x^{\mathrm{T}}\Sigma^{1/2}A\Sigma^{1/2}x}{\left\|\Sigma^{1/2}x\right\|^2} \cdot \min_{x\in\mathbb{R}^n: x\neq 0} \frac{\left\|\Sigma^{1/2}x\right\|^2}{\|x\|^2} \\ &= ab.\end{aligned}$$

□

Lemma F.3. *Assume Assumption 2 and suppose $\phi, \psi > 0$. Then for any $z < 0$, the following equation (25) has a unique negative solution $c(z) = c(z, \phi, \psi, H)$,*

$$c(z) = \int \frac{(z + c(z))x}{-z - c(z) + x\psi\phi^{-1}}\, dH(x). \tag{25}$$

Furthermore, $\lim_{z\to 0^-} c(z) = c_0$ where c_0 is defined by (9).

Proof of Lemma F.3. Given $z < 0$, let $f(c(z)) = c(z) - \int \frac{(z+c(z))x}{-z-c(z)+x\psi\phi^{-1}}\, dH(x)$. We have $f(-\infty) = -\infty$ and $f(0) = -\int \frac{zx}{-z+x\psi\phi^{-1}}\, dH(x) > 0$ since $z < 0$ and $x, \phi, \psi > 0$. By the smoothness of f, we know f has at least one negative solution. Suppose $c_1(z)$ and $c_2(z)$ are two negative solutions with $c_1(z) > c_2(z)$. Then, we have

$$\begin{aligned}0 &= \int \frac{x(\frac{z}{c_1(z)} + 1)}{-z - c_1(z) + x\psi\phi^{-1}}\, dH(x) - \int \frac{x(\frac{z}{c_2(z)} + 1)}{-z - c_2(z) + x\psi\phi^{-1}}\, dH(x) \\ &= \int \Bigg\{ \frac{x(c_1(z) - c_2(z))}{(-z - c_1(z) + x\psi\phi^{-1})(-z - c_2(z) + x\psi\phi^{-1})} + \frac{z^2 x(\frac{c_1(z)-c_2(z)}{c_1(z)c_2(z)})}{(-z - c_1(z) + x\psi\phi^{-1})(-z - c_2(z) + x\psi\phi^{-1})} \\ &\quad + \frac{zx^2\psi\phi^{-1}(\frac{c_2(z)-c_1(z)}{c_1(z)c_2(z)})}{(-z - c_1(z) + x\psi\phi^{-1})(-z - c_2(z) + x\psi\phi^{-1})} + \frac{zx(\frac{(c_1(z)-c_2(z))(c_1(z)+c_2(z))}{c_1(z)c_2(z)})}{(-z - c_1(z) + x\psi\phi^{-1})(-z - c_2(z) + x\psi\phi^{-1})} \Bigg\}\, dH(x).\end{aligned}$$

Since $z, c_1(z), c_2(z) < 0$, it is easy to find that each term above is larger than 0. This contradiction shows for given $z < 0$, (25) has a unique negative solution, denoted by $c(z)$.

Next, we show $\lim_{z\to 0^-} c(z) = c_0$. Given $z < 0$, let

$$g(a, z) = 1 - \int \frac{x}{-z - a + x\psi\phi^{-1}}\, dH(x) - \int \frac{z/a}{-z - a + x\psi\phi^{-1}}\, dH(x).$$

Since $c(z)$ is the solution of (25), $g(c(z), z) = 0$. For any small $\epsilon > 0$, we can find a sufficiently small $\delta_1 > 0$ such that, when $-\delta_1 < z < 0$, we have $0 < -z - c_0 - \epsilon + x\psi\phi^{-1} < -c_0 + x\psi\phi^{-1}$. The second inequality is satisfied by taking

$\delta_1 < \epsilon$. Because x lies in the support of the measure H, $x > C_0 > 0$. Moreover, since $c_0 < 0$, the first inequality holds when ϵ and δ_1 are sufficiently small. Because c_0 is the root of g when $z = 0$, in this case, we have $g(c_0 + \epsilon, z) < 0$ when $z \in (-\delta_1, 0)$. Furthermore, we can find a sufficiently small $\delta_2 > 0$ such that $g(c_0 - \epsilon, z) > 0$ when $z \in (-\delta_2, 0)$. This is because $g(c_0, 0) = 0$, $1 - \int \frac{x}{-c_0+\epsilon+x\psi\phi^{-1}}\, dH(x) > 0$, and $\lim_{z\to 0^-} \int \frac{z/a}{-z-a+x\psi\phi^{-1}}\, dH(x) = 0$.

To conclude, taking $\delta = \min\{\delta_1, \delta_2\}$, we have, when $z \in (-\delta, 0)$, $g(c_0 + \epsilon, z) < 0$ and $g(c_0 - \epsilon, z) > 0$. By the smoothness of $g(a, z)$ with respect to a, we know $c(z) \in (c_0 - \epsilon, c_0 + \epsilon)$. Using the definition of a limit completes the proof.

□

Now we prove the bias part (10).

Proof of the bias part (10). For all large n, we have almost surely

$$\begin{aligned}
& \left(X^\intercal S^\intercal S X\right)^{+} X^\intercal S^\intercal S X \\
= \; & (SX)^{+} SX \\
= \; & \lim_{\delta\to 0^+} X^\intercal S^\intercal \left(S X X^\intercal S^\intercal + \delta I_m\right)^{-1} SX \\
= \; & X^\intercal S^\intercal \left(S X X^\intercal S^\intercal\right)^{-1} SX, \qquad (26)
\end{aligned}$$

where the first equality uses $A^{+} = \left(A^\intercal A\right)^{+} A^\intercal$ for any matrix A, the second inequality uses $A^{+} = \lim_{\delta\to 0^+} A^\intercal \left(AA^\intercal + \delta I\right)^{-1}$, and the third equality follows from Lemma F.1 and Assumption 2. Specifically, when $SZZ^\intercal S^\intercal$ is invertible and Σ satisfies Assumption 2, then $SXX^\intercal S^\intercal = SZ\Sigma Z^\intercal S^\intercal$ is also invertible. Let the singular value decomposition (SVD) of S be $S = UDV$ where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{m\times n}$ are both orthogonal matrices, $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix. By Assumption 4, we know almost surely for all large n, D is invertible. Then the RHS (right hand side) of (26) can be writen as

$$X^\intercal S^\intercal \left(SXX^\intercal S^\intercal\right)^{-1} SX = X^\intercal V^\intercal \left(VXX^\intercal V^\intercal\right)^{-1} VX = \left(X^\intercal V^\intercal V X\right)^{+} X^\intercal V^\intercal V X. \qquad (27)$$

Thus, by (26), (27) and Lemma 3.1, we have

$$B_{(S,X)}(\widehat{\beta}^S; \beta) = \frac{\alpha^2}{p} \mathrm{tr}\left\{\left[I_p - \left(\frac{1}{p} X^\intercal V^\intercal V X\right)^{+} \frac{1}{p} X^\intercal V^\intercal V X\right]\Sigma\right\}. \qquad (28)$$

For any $z < 0$,

$$\begin{aligned}
& \left|\frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}X^\intercal V^\intercal V X\right)^{+} \frac{1}{p} X^\intercal V^\intercal V X \Sigma\right] - \frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p} X^\intercal V^\intercal V X - z I_p\right)^{-1} \frac{1}{p} X^\intercal V^\intercal V X\Sigma\right]\right| \\
\leq \; & \frac{|z|\, \|\Sigma\|_2}{\lambda^{+}_{\min}\left(\frac{1}{p} X^\intercal V^\intercal V X\right) - z} \\
= \; & \frac{|z|\, \|\Sigma\|_2}{\lambda^{+}_{\min}\left(\frac{1}{p} \Sigma^{1/2} Z^\intercal V^\intercal V Z \Sigma^{1/2}\right) - z} \leq \frac{|z|\, C_1}{C_0 \tau - z},
\end{aligned}$$

where the last inequality follows from Lemmas F.1, F.2 and Assumption 2. Thus, taking limites on both sides of (28) gives

$$\begin{aligned}
\lim_{n\to\infty} B_{(S,X)}(\widehat{\beta}^S; \beta) &= \alpha^2 \lim_{n\to\infty} \lim_{z\to 0^-} \frac{1}{p}\mathrm{tr}\left\{\left[I_p - \left(\frac{1}{p} X^\intercal V^\intercal V X - z I_p\right)^{-1} \frac{1}{p} X^\intercal V^\intercal V X\right]\Sigma\right\} \\
&= \alpha^2 \lim_{n\to\infty} \lim_{z\to 0^-} \frac{1}{p}\mathrm{tr}\left\{\left[I_p - \left(\frac{1}{p} X^\intercal V^\intercal V X - z I_p\right)^{-1} \left(\frac{1}{p} X^\intercal V^\intercal V X - z I_p + z I_p\right)\right]\Sigma\right\} \\
&= -\alpha^2 \lim_{n\to\infty} \lim_{z\to 0^-} z \frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p} X^\intercal V^\intercal V X - z I_p\right)^{-1} \Sigma\right]. \qquad (29)
\end{aligned}$$

Now we can follow a similar argument to the proof of Theorem 1 in (Hastie et al., 2022) to show the validity of exchanging the limits $n \to \infty$ and $z \to 0^-$. Define $f_n(z) = -\frac{z}{p}\text{tr}\left[\left(\frac{1}{p}X^\top V^\top VX - zI_p\right)^{-1}\Sigma\right]$. Since $|f_n(z)| \le |z|\left\|(\frac{1}{p}X^\top V^\top VX - zI_p)^{-1}\right\|_2 \|\Sigma\|_2 \le C_1$, we know $f_n(z)$ is uniformly bounded. Besides,

$$
\begin{aligned}
|f_n'(z)| &\le \frac{1}{p}\left|\text{tr}\left[\left(\frac{1}{p}X^\top V^\top VX - zI_p\right)^{-1}\Sigma\right] + z\,\text{tr}\left[\left(\frac{1}{p}X^\top V^\top VX - zI_p\right)^{-2}\Sigma\right]\right| \\
&\le \frac{\lambda_{\min}^+\left(\frac{1}{p}X^\top V^\top VX\right)\|\Sigma\|_2}{\left[\lambda_{\min}^+\left(\frac{1}{p}X^\top V^\top VX\right) - z\right]^2} \\
&\le \frac{\|\Sigma\|_2}{\lambda_{\min}^+\left(\frac{1}{p}X^\top V^\top VX\right)} \le \frac{C_1}{C_0\tau}
\end{aligned} \tag{30}
$$

where the last inequality holds almost surely for all large n. As its derivatives are bounded, the sequence $\{f_n\}_{n=1}^\infty$ is equicontinuous, and hence, by the Arzela-Ascoli theorem, f_n converges uniformly. Thus, we can use Moore-Osgood theorem to conclude the validity of exchanging the limits $n \to \infty$ and $\lambda \to 0^-$. Define

$$
m_{1n}(z) = \frac{1}{p}\text{tr}\left[\left(\frac{1}{p}\Sigma^{1/2}Z^\top V^\top VZ\Sigma^{1/2} - zI_p\right)^{-1}\Sigma\right], \quad m_{2n}(z) = \frac{1}{p}\text{tr}\left[\left(\frac{1}{p}VZ\Sigma Z^\top V^\top - zI_m\right)^{-1}\right]. \tag{31}
$$

According to (Zhang, 2007), almost surely, as $n \to \infty$, $m_{1n}(z) \to m_1(z)$, and $m_{2n}(z) \to m_2(z)$ where for given $z < 0$, $(m_1(z), m_2(z)) \in \mathbb{R}^+ \times \mathbb{R}^+$ is the unique solution of the self-consistent equations

$$
\begin{aligned}
m_1(z) &= \int \frac{x}{-z\left[1 + xm_2(z)\right]}\,dH(x), \\
m_2(z) &= \psi\phi^{-1}\frac{1}{-z\left[1 + m_1(z)\right]}.
\end{aligned} \tag{32}
$$

Substituting m_2 into m_1 in (32) and mutiplying both sides by z, we obtain

$$
zm_1(z) = \int \frac{(z + zm_1(z))\,x}{-z - zm_1(z) + x\psi\phi^{-1}}\,dH(x).
$$

By lemma F.3, we know $\lim_{z\to 0^-} zm_1(z) = c_0$. Exchanging the limits in (29), we have almost surely

$$
\begin{aligned}
\lim_{n\to\infty} B_{(S,X)}(\widehat{\beta}^S;\beta) &= -\alpha^2 \lim_{z\to 0^-}\lim_{n\to\infty} z\frac{1}{p}\text{tr}\left[\left(\frac{1}{p}X^\top V^\top VX - zI_p\right)^{-1}\Sigma\right] \\
&= -\alpha^2 \lim_{z\to 0^-} zm_1(z) = -\alpha^2 c_0.
\end{aligned}
$$

Lemma 3.1 assures that $B_{(\beta,S,X)}(\widehat{\beta}^S;\beta)$ converges almost surely to the same limit. □

Variance part To prove the variance part (11), we need the following theorem, often known as the Vitali's theorem (Bai & Silverstein, 2010, Lemma 2.14). This theorem ensures the convergence of the derivatives of converging analytic functions.

Lemma F.4 (Vitali's convergence theorem)**.** *Let $f_1, f_2, \cdots$ be analytic on the domain D, satisfying $|f_n(z)| \le M$ for every n and $z \in D$. Suppose that there is an analytic function f on D such that $f_n(z) \to f(z)$ for all $z \in D$. Then it also holds that $f_n'(z) \to f'(z)$ for all $z \in D$.*

Proof of the variance part (11). Let the singular value decomposition (SVD) of S be $S = UDV$ where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{m\times n}$ are both orthogonal matrices, $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix. According to Lemma 3.1,

$$
V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \sigma^2\text{tr}\left[(X^\top S^\top SX)^+ X^\top S^\top SS^\top SX\,(X^\top S^\top SX)^+\,\Sigma\right]
$$

$$
\begin{aligned}
&= \sigma^2 \mathrm{tr}\left[(SX)^+ SS^\mathrm{T} (X^\mathrm{T}S^\mathrm{T})^+ \Sigma\right] \\
&= \sigma^2 \mathrm{tr}\left[\lim_{\delta\to 0+} X^\mathrm{T}S^\mathrm{T}\left(SXX^\mathrm{T}S^\mathrm{T} + \delta I_m\right)^{-1} SS^\mathrm{T}\left(SXX^\mathrm{T}S^\mathrm{T}+\delta I_m\right)^{-1} SX\Sigma\right] \\
&= \sigma^2 \mathrm{tr}\left[X^\mathrm{T}S^\mathrm{T}\left(SXX^\mathrm{T}S^\mathrm{T}\right)^{-1} SS^\mathrm{T}\left(SXX^\mathrm{T}S^\mathrm{T}\right)^{-1} SX\Sigma\right] \\
&= \sigma^2 \mathrm{tr}\left[X^\mathrm{T}V^\mathrm{T}\left(VXX^\mathrm{T}V^\mathrm{T}\right)^{-2} VX\Sigma\right] \\
&= \sigma^2 \mathrm{tr}\left[(VX)^+ (X^\mathrm{T}V^\mathrm{T})^+ \Sigma\right] \\
&= \sigma^2 \mathrm{tr}\left[(X^\mathrm{T}V^\mathrm{T}VX)^+ \Sigma\right] \\
&= \frac{\sigma^2}{p} \mathrm{tr}\left[\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right)^+ \Sigma\right],
\end{aligned} \tag{33}
$$

where similar to the proof of the bias part in Theorem 3.2, we use the identity $A^+ = (A^\mathrm{T}A)^+ A^\mathrm{T} = \lim_{\delta\to 0^+} A^\mathrm{T}\left(AA^\mathrm{T}+\delta I\right)^{-1}$ for any matrix A, and the fact that almost surely for all large n, $SXX^\mathrm{T}S^\mathrm{T}$ is invertible. Define

$$
g_n(z) = \frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right)\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-2}\Sigma\right].
$$

Since for any $z \le 0, x > 0$, we have

$$
\left|\frac{x}{(x-z)^2} - \frac{1}{x}\right| \le \frac{2\left|z\right|}{x^2}.
$$

Thus, by Lemma F.2 and Assumption 2, for any $z < 0$,

$$
\left|\frac{V_{(S,X)}\left(\widehat{\beta}^S;\beta\right)}{\sigma^2} - g_n(z)\right| \le \frac{2\left|z\right|}{\left[\lambda_{\min}^+\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right)\right]^2}\left\|\Sigma\right\|_2 \le \frac{2\left|z\right|C_1}{\left(C_0\tau\right)^2}. \tag{34}
$$

By (34), we can continue (33),

$$
\begin{aligned}
&\lim_{n\to\infty} V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \\
=\;& \sigma^2 \lim_{n\to\infty}\lim_{z\to 0^-}\frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right)\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-2}\Sigma\right] \\
=\;& \sigma^2 \lim_{n\to\infty}\lim_{z\to 0^-}\frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p + zI_p\right)\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-2}\Sigma\right] \\
=\;& \sigma^2 \lim_{n\to\infty}\lim_{z\to 0^-}\frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-1}\Sigma\right] + \frac{1}{p}\mathrm{tr}\left[z\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-2}\Sigma\right],
\end{aligned} \tag{35}
$$

We now verify the validity of exchanging the limits $n\to\infty$ and $z\to 0^-$. As in the proof of the bias part of Theorem 3.2, in order to use Arzela-Ascoli theorem and Moore-Osgood theorem, it suffices to show $g_n(z)$ and $g_n'(z)$ are both uniformly bounded. We know it holds almost surely for all large n that for any $z < 0$,

$$
\left|g_n(z)\right| \le \frac{\left\|\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right\|_2\left\|\Sigma\right\|_2}{\left[\lambda_{\min}^+\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right) - z\right]^2} \le \frac{\left\|\frac{1}{p}ZZ^\mathrm{T}\right\|_2\left\|V^\mathrm{T}V\right\|_2\left\|\Sigma\right\|_2^2}{\left[\lambda_{\min}^+\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX\right) - z\right]^2} \le \frac{\left(1+\sqrt{\phi^{-1}}\right)^2 C_1^2}{(C_0\tau)^2}.
$$

Moreover,

$$
\left|g_n'(z)\right| = \left|\frac{2}{p}\mathrm{tr}\left[\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-2}\Sigma\right] + \frac{2}{p}\mathrm{tr}\left[z\left(\frac{1}{p}X^\mathrm{T}V^\mathrm{T}VX - zI_p\right)^{-3}\Sigma\right]\right|
$$

$$\leq \frac{2\left\|\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right\|_2 \|\Sigma\|_2}{\left[\lambda_{\min}^{+}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right)-z\right]^3} \leq \frac{2\left(1+\sqrt{\phi^{-1}}\right)^2 C_1^2}{(C_0\tau)^3}.$$

Thus, $g_n(z)$ and $g_n'(z)$ are both uniformly bounded and hence, we can exchange the limits. Recall the definition of $m_{1n}(z)$ in (31), we know $g_n(z) = m_{1n}(z) + zm_{1n}'(z) = (zm_{1n}(z))'$. We will use Lemma F.4 to show $g_n(z) \to m_1(z) + zm_1'(z)$ almost surely as $n \to \infty$. Since $zm_{1n}(z)$ and $zm_1(z)$ are analytic on $(-\infty, 0)$ such that $zm_{1n}(z) \to zm_1(z)$; see (Zhang, 2007). In addition, as in the proof of the bias part of Theorem 3.2, almost surely for all large n, it holds that $|zm_{1n}(z)| \leq C_1$. Thus the conditions of Lemma F.4 are satisfied. By exchanging the limits in (35) and using Lemma F.4, we obtain

$$\lim_{n\to\infty} V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \sigma^2 \lim_{z\to 0^-}\lim_{n\to\infty} g_n(z) = \sigma^2 \lim_{z\to 0^-} m_1(z) + zm_1'(z). \tag{36}$$

Recall the self-consistent equations in (32). A direct calculation yields

$$m_1(z) + zm_1'(z) = \int \frac{(1+m_1(z)+zm_1'(z))\, x^2\psi\phi^{-1}}{(z+zm_1(z)-x\psi\phi^{-1})^2}\, dH(x).$$

Taking $z \to 0^-$ in the above equality and using $\lim_{z\to 0^-} zm_1(z) = c_0$ in Lemma F.3, we derive

$$\lim_{z\to 0^-} m_1(z) + zm_1'(z) = \frac{\int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2}\, dH(x)}{1 - \int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2}\, dH(x)}.$$

Combining the above limit with (36), we have almost surely

$$\lim_{n\to\infty} V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = \sigma^2 \frac{\int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2}\, dH(x)}{1 - \int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2}\, dH(x)}.$$

Lemma 3.1 assures that $V_{(\beta,S,X)}\left(\widehat{\beta}^S;\beta\right)$ converges almost surely to the same limit. □

F.2. Proofs for the under-parameterized case

F.2.1. Proof of Theorem 4.3

According to Lemma F.1 and Assumption 2, we know almost surely for all large n, $\frac{1}{p}X^{\mathrm{T}}S^{\mathrm{T}}SX$ is invertible. Thus by Lemma 3.1, it holds that almost surely for all large n, $B_{(S,X)}(\widehat{\beta}^S;\beta) = B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) = 0$ and hence (13) holds. To show the limiting variance (14), we follow from a similar proof to that for the bias part in Theorem 3.2. To be concise, we only sketch the proof here. Similar to (33) and (35), we have

$$\begin{aligned}
& V_{(S,X)}\left(\widehat{\beta}^S;\beta\right) \\
= \;& \sigma^2 \mathrm{tr}\left[\left(X^{\mathrm{T}}S^{\mathrm{T}}SX\right)^{-1} X^{\mathrm{T}}S^{\mathrm{T}}SS^{\mathrm{T}}SX \left(X^{\mathrm{T}}S^{\mathrm{T}}SX\right)^{-1}\Sigma\right] \\
= \;& \sigma^2 \mathrm{tr}\left[\left(Z^{\mathrm{T}}S^{\mathrm{T}}SZ\right)^{-1} Z^{\mathrm{T}}S^{\mathrm{T}}SS^{\mathrm{T}}SZ \left(Z^{\mathrm{T}}S^{\mathrm{T}}SZ\right)^{-1}\right] \\
= \;& \sigma^2 \mathrm{tr}\left[\left(Z^{\mathrm{T}}S^{\mathrm{T}}\right)^{+}(SZ)^{+} SS^{\mathrm{T}}\right] \\
= \;& \sigma^2 \mathrm{tr}\left[\left(SZZ^{\mathrm{T}}S^{\mathrm{T}}\right)^{+} SS^{\mathrm{T}}\right] \\
= \;& \frac{\sigma^2}{n}\mathrm{tr}\left[\left(\frac{1}{n}SZZ^{\mathrm{T}}S^{\mathrm{T}}\right)^{+} SS^{\mathrm{T}}\right] \\
= \;& \sigma^2 \lim_{z\to 0^-}\frac{1}{n}\mathrm{tr}\left[\left(\frac{1}{n}SZZ^{\mathrm{T}}S^{\mathrm{T}}\right)\left(\frac{1}{n}SZZ^{\mathrm{T}}S^{\mathrm{T}} - zI_m\right)^{-2} SS^{\mathrm{T}}\right] \\
= \;& \sigma^2 \lim_{z\to 0^-}\frac{1}{n}\mathrm{tr}\left[\left(\frac{1}{n}SZZ^{\mathrm{T}}S^{\mathrm{T}} - zI_m\right)^{-1} SS^{\mathrm{T}}\right] + \frac{1}{n}\mathrm{tr}\left[z\left(\frac{1}{n}SZZ^{\mathrm{T}}S^{\mathrm{T}} - zI_m\right)^{-2} SS^{\mathrm{T}}\right].
\end{aligned} \tag{37}$$

Define

$$\widetilde{m}_{1n}(z) = \frac{1}{n}\mathrm{tr}\left[\left(\frac{1}{n}SZZ^{\mathrm{T}}S^{\mathrm{T}} - zI_m\right)^{-1} SS^{\mathrm{T}}\right], \; \widetilde{m}_{2n}(z) = \frac{1}{n}\mathrm{tr}\left[\left(\frac{1}{n}Z^{\mathrm{T}}S^{\mathrm{T}}SZ - zI_p\right)^{-1}\right]. \tag{38}$$

Then $\widetilde{m}_{1n}(z) \to \widetilde{m}_1(z)$ and $\widetilde{m}_{2n}(z) \to \widetilde{m}_2(z)$ almost surely as $n \to \infty$, where $(\widetilde{m}_1(z), \widetilde{m}_2(z)) \in \mathbb{R}^+ \times \mathbb{R}^+$ is the unique solution of the self-consistent equations (Zhang, 2007)

$$\begin{aligned}
\widetilde{m}_1(z) &= \psi \int \frac{x}{-z\left[1 + x\widetilde{m}_2(z)\right]}\, dB(x), \\
\widetilde{m}_2(z) &= \phi \frac{1}{-z\left[1 + \widetilde{m}_1(z)\right]},
\end{aligned} \tag{39}$$

for any $z < 0$. Substituting $\widetilde{m}_2$ into $\widetilde{m}_1$ in (39) and multiplying both sides by z, we obtain

$$z\widetilde{m}_1(z) = \int \psi \frac{\left(z + z\widetilde{m}_1(z)\right)x}{-z - z\widetilde{m}_1(z) + x\phi}\, dH(x).$$

Following the similar proofs to Lemma F.3 and the bias part in Theorem 3.2, we can obtain $\lim_{z\to 0^-} z\widetilde{m}_1(z) = \widetilde{c}_0$ where $\widetilde{c}_0$ is defined in (12). Following the same argument for verifying interchange of the limits and Lemma F.4, we have

$$\begin{aligned}
&\lim_{n\to\infty} V_{(S,X)}(\widehat{\beta}^S;\beta) = \lim_{n\to\infty} V_{(\beta,S,X)}(\widehat{\beta}^S;\beta) \\
= \;& \sigma^2 \lim_{z\to 0^-} \widetilde{m}_1(z) + z\widetilde{m}_1'(z) = \sigma^2 \frac{\psi \int \frac{x^2\phi}{(\widetilde{c}_0 - x\phi)^2}\, dB(x)}{1 - \psi \int \frac{x^2\phi}{(\widetilde{c}_0 - x\phi)^2}\, dB(x)}.
\end{aligned}$$

F.2.2. PROOF OF COROLLARY 4.4

When S is an orthogonal sketching matrix, i.e., $SS^{\mathrm{T}} = I_m$, we have $B(x) = \delta_{\{1\}}(x)$ where δ is the Dirac function. A simple calculation shows $\widetilde{c}_0 = \phi - \psi$, and hence

$$V_{(S,X)}(\widehat{\beta}^S;\beta) = V_{(\beta,S,X)}(\widehat{\beta}^S;\beta) \to \sigma^2 \frac{\psi \int \frac{x^2\phi}{(\widetilde{c}_0 - x\phi)^2}\, dB(x)}{1 - \psi \int \frac{x^2\phi}{(\widetilde{c}_0 - x\phi)^2}\, dB(x)} = \sigma^2 \frac{\phi\psi^{-1}}{1 - \phi\psi^{-1}}.$$

When S is an i.i.d. sketching matrix, we know that almost surely, the ESD of SS^{T} converges to the M-P law with parameter ψ, whose CDF (cumulative distribution function) is denoted by F_ψ, i.e., $B = F_\psi$. The self-consistent equation (12) reduces to

$$1 = \frac{\psi}{\phi} + \frac{\psi\widetilde{c}_0}{\phi^2} s_\psi\left(\frac{\widetilde{c}_0}{\phi}\right)$$

where s_ψ is the Stieltjes transform of M-P law with parameter ψ. According to the seminal work (Marcenko & Pastur, 1967), we know for any $z < 0$,

$$s_\psi(z) = \frac{1 - \psi - z - \sqrt{(z - 1 - \psi)^2 - 4\psi}}{c\psi z}.$$

A direct calculation shows $\widetilde{c}_0 = -\psi - \phi^2 + \phi + \psi\phi$. Furthermore,

$$\begin{aligned}
\psi \int \frac{x^2\phi}{\left(\widetilde{c}_0 - x\phi\right)^2}\, dB(x) &= \psi\phi^{-1} \int \frac{x^2}{\left(x - \widetilde{c}_0\phi^{-1}\right)^2}\, dF_\psi(x) \\
&= \psi\phi^{-1}\left[1 + 2\widetilde{c}_0\phi^{-1} s_\psi\left(\widetilde{c}_0\phi^{-1}\right) + \left(\widetilde{c}_0\phi^{-1}\right)^2 s_\psi'\left(\widetilde{c}_0\phi^{-1}\right)\right] \\
&= \frac{\phi - 2\phi^2\psi^{-1} + \phi\psi^{-1}}{1 - \phi^2\psi^{-1}}.
\end{aligned}$$

Plugging the above equality into (14), we get

$$V_{(S,X)}(\widehat{\beta}^S;\beta) = V_{(\beta,S,X)}(\widehat{\beta}^S;\beta) \to \sigma^2 \left(\frac{\phi}{1 - \phi} + \frac{\phi\psi^{-1}}{1 - \phi\psi^{-1}}\right).$$

F.2.3. Proof of Corollary 4.5

According to (12), we have

$$1 = \psi\phi^{-1} + \psi\widetilde{c}_0\phi^{-2}\int \frac{1}{x - \widetilde{c}_0\phi^{-1}}\,dB(x) = \psi\phi^{-1}\left(1 + ts_B(t)\right),$$

where $t = \widetilde{c}_0\phi^{-1}$ and $s_B(z) = \int \frac{1}{x-z}\,dB(x)$ is the Stieltjes transform of the measure B. Thus, we have $ts_B(t) = \psi^{-1}\phi - 1$. In order to minimize (14), it suffices to minimize the numerator $\psi\phi^{-1}\int \frac{x^2}{(t-x)^2}\,dB(x)$, which after simplification is $\psi\phi^{-1}\left[1 + 2ts_B(t) + t^2 s_B'(t)\right]$. Therefore it suffices to minimize $t^2 s_B'(t)$. By the Cauchy-Schwartz inequality, we have

$$t^2 s_B'(t) = \int \frac{t^2}{(x-t)^2}\,dB(x) \geq \left(\int \frac{t}{x-t}\,dB(x)\right)^2 = \left(\psi^{-1}\phi - 1\right)^2,$$

and the minimum is achieved at $B = \delta_{\{a\}}(a > 0)$.

G. Proof of Theorem B.1

The proof to the variance part is the same as those for Theorem 4.2 and Theorem 4.3. As for the bias part, when $p/m \to \phi\psi^{-1} < 1$, same as (13), it is easy to show almost surely for all large n, $B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) = 0$. Hence, we only need to prove the bias part (22) for $p/m \to \phi\psi^{-1} > 1$. Without loss of generality, we assume $\|\beta\| = 1$ throughout the proof. Let the SVD of S be $S = UDV$ where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{m\times n}$ are both orthogonal matrices, and $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix. According to (26) and (27), we have

$$B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) = \left\|\Sigma^{1/2}\left[(X^{\mathrm{T}}V^{\mathrm{T}}VX)^{+}X^{\mathrm{T}}V^{\mathrm{T}}VX - I_p\right]\beta\right\|^2.$$

Let

$$h_n(z) = \left\|\Sigma^{1/2}\left[\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-1}\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - I_p\right]\beta\right\|^2.$$

Then, for any $z < 0$,

$$\begin{aligned}
&\left|B_{(\beta,S,X)}(\widehat{\beta}^S;\beta)^{1/2} - h_n(z)^{1/2}\right|\\
\leq\ &\left\|\Sigma^{1/2}\left[\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right)^{+}\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - \left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-1}\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right]\beta\right\|\\
\leq\ &\left\|\Sigma^{1/2}\right\|_2\|\beta\|\frac{|z|}{\lambda_{\min}^{+}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right) - z}\\
\leq\ &\frac{\sqrt{C_1}\,|z|}{\lambda_{\min}^{+}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right) - z} \leq \frac{\sqrt{C_1}\,|z|}{C_0\tau},
\end{aligned}\tag{40}$$

where the last inequality uses Lemma F.1 and F.2. By the fact that $\left|B_{(\beta,S,X)}(\widehat{\beta}^S;\beta)\right| \leq C_1$ and (40), we conclude

$$B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) = \lim_{z\to 0^-} h_n(z) = \lim_{z\to 0^-} z^2\beta^{\mathrm{T}}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-1}\Sigma\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-1}\beta.$$

Next, we follow the same idea as in the proof to the bias part in Theorem 3.2 to verify the interchange of the limits $n \to \infty$ and $z \to 0^-$. Since for any $z < 0$, $|h_n(z)| \leq C_1$ and

$$\begin{aligned}
&|h_n'(z)|\\
\leq& 2\|\Sigma\|_2\|\beta\|^2\left\|z\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-1}\right\|_2\left\|\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-1} + z\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p\right)^{-2}\right\|_2
\end{aligned}$$

$$\leq 2\frac{\lambda_{\min}^{+}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right)\|\Sigma\|_2}{\left[\lambda_{\min}^{+}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX\right)-z\right]^2} \leq \frac{2C_1}{C_0\tau},$$

where the second and the last inequalities follow (30) and hold almost surely for all large n, we can exchange limits by Arzela-Ascoli theorem and Moore-Osgood theorem, that is,

$$\lim_{n\to\infty} B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) = \lim_{n\to\infty}\lim_{z\to 0^-} h_n(z) = \lim_{z\to 0^-}\lim_{n\to\infty} h_n(z). \tag{41}$$

Next, we aim to find $\lim_{z\to 0^-}\lim_{n\to\infty} h_n(z)$. Let $\mathbb{D} = \{(z,w)\in\mathbb{R}^2 : z<0, w>-\frac{1}{2C_1}\}$,

$$\mathcal{H}_n(z,w) = z\beta^{\mathrm{T}}\left(\frac{1}{p}X^{\mathrm{T}}V^{\mathrm{T}}VX - zI_p - zw\Sigma\right)^{-1}\beta$$

which is defined on $\mathbb{D}$, and

$$\Sigma_w = \Sigma\left(I_p + w\Sigma\right)^{-1}, \quad \beta_w = \left(I_p + w\Sigma\right)^{-1/2}\beta.$$

Then, $\mathcal{H}_n(z,w)$ is analytic on $\mathbb{D}$ such that

$$\mathcal{H}_n(z,w) = z\beta_w^{\mathrm{T}}\left(\frac{1}{p}\Sigma_w^{1/2}Z^{\mathrm{T}}V^{\mathrm{T}}VZ\Sigma_w^{1/2} - zI_p\right)^{-1}\beta_w$$
$$h_n(z) = \frac{\partial\mathcal{H}_n}{\partial w}(z,0).$$

Further write

$$m_{1n}(z,w) = \frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}\Sigma_w^{1/2}Z^{\mathrm{T}}V^{\mathrm{T}}VZ\Sigma_w^{1/2} - zI_p\right)^{-1}\Sigma_w\right],\ m_{2n}(z,w) = \frac{1}{p}\mathrm{tr}\left[\left(\frac{1}{p}VZ\Sigma_w Z^{\mathrm{T}}V^{\mathrm{T}} - zI_m\right)^{-1}\right].$$

According to (Paul & Silverstein, 2009) or (Couillet & Liao, 2022, Theorem 2.7), $-\frac{1}{z}\left(I_p + m_{2n}(z,w)\Sigma_w\right)^{-1}$ is the deterministic equivalent of $\left(\frac{1}{p}\Sigma_w^{1/2}Z^{\mathrm{T}}V^{\mathrm{T}}VZ\Sigma_w^{1/2} - zI_p\right)^{-1}$. Thus, it holds that, for any given $(z,w)\in\mathbb{D}$, as $n\to\infty$,

$$\mathcal{H}_n(z,w) - \widetilde{\mathcal{H}}_n(z,w) \to 0 \quad \text{almost surely},$$

where $\widetilde{\mathcal{H}}_n(z,w)$ is defined as

$$\widetilde{\mathcal{H}}_n(z,w) = -\beta_w^{\mathrm{T}}\left(I_p + m_{2n}(z,w)\Sigma_w\right)^{-1}\beta_w.$$

Furthermore, it is easy to show that $\mathcal{H}_n(z,w)$ and $\widetilde{\mathcal{H}}_n(z,w)$ are both uniformly bounded on $\mathbb{D}$. Thus, using the Vitali's theorem colleted as Lemma F.4, we have for $z<0$

$$\begin{aligned}\lim_{n\to\infty} h_n(z) &= \lim_{n\to\infty}\frac{\partial\widetilde{\mathcal{H}}_n}{\partial w}(z,0)\\ &= \lim_{n\to\infty}\left(1+\frac{\partial m_{2n}}{\partial w}(z,0)\right)\beta^{\mathrm{T}}\left[1+m_{2n}(z,0)\Sigma\right]^{-2}\Sigma\beta\\ &= \left(1+\frac{\partial m_2}{\partial w}(z,0)\right)\int\frac{x}{\left[1+m_2(z,0)x\right]^2}\,dG(x)\end{aligned} \tag{42}$$

almost surely, where, according to (Zhang, 2007), $m_{1n}(z,w)\to m_1(z,w)$ and $m_{1n}(z,w)\to m_1(z,w)$ almost surely as $n\to\infty$. Moreover, for any given $(z,w)\in\mathbb{D}$, $(m_1(z,w), m_2(z,w))\in\mathbb{R}^+\times\mathbb{R}^+$ is the unique solution to the self-consistent equations

$$m_1(z,w) = \int\frac{x}{-z\left[1+wx+xm_2(z,w)\right]}\,dH(x),$$

$$m_2(z,w) = \psi\phi^{-1} \frac{1}{-z\left[1 + m_1(z,w)\right]}. \tag{43}$$

Substituting m_2 into m_1 in (43) and using $m_1(z,0) = m_1(z)$ as defined in (32), we have after some calculations

$$\lim_{z\to 0^-} z\frac{\partial m_1}{\partial w}(z,0) = \frac{\int \frac{c_0^2 x^2}{(c_0 - x\psi\phi^{-1})^2}\, dH(x)}{1 - \int \frac{x^2\psi\phi^{-1}}{(c_0 - x\psi\phi^{-1})^2}\, dH(x)} = \phi\psi^{-1}c_1c_0^2.$$

where c_0 is defined in (9) and $\lim_{z\to 0^-} zm_1(z) = c_0$. Using (41) and continue (42), we have almost surely

$$\begin{aligned}
\lim_{n\to\infty} B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) &= \lim_{z\to 0^-}\lim_{n\to\infty} h_n(z) \\
&= \lim_{z\to 0^-}\left(1 + \frac{\partial m_2}{\partial w}(z,0)\right)\int \frac{x}{\left[1 + m_2(z,0)x\right]^2}\, dG(x) \\
&= \lim_{z\to 0^-}\left(1 + \psi\phi^{-1}\frac{z\frac{\partial m_1}{\partial w}(z,0)}{\left[z + zm_1(z)\right]^2}\right)\int \frac{x}{\left[1 - \psi\phi^{-1}x\frac{1}{z + zm_1(z)}\right]^2}\, dG(x) \\
&= (1 + c_1)\int \frac{c_0^2 x}{\left(c_0 - x\psi\phi^{-1}\right)^2}\, dG(x).
\end{aligned}$$

H. Proofs for central limit theorems

H.1. Proof of Theorem B.2

For the underparameterized case with $\phi\psi^{-1} < 1$, by Lemma 3.1, it holds that

$$B_{(S,X)}(\widehat{\beta}^S,\beta) = 0, \quad V_{(S,X)}(\widehat{\beta}^S,\beta) = \sigma^2 \mathrm{tr}\left\{(X^{\mathrm{T}}S^{\mathrm{T}}SX)^{+}\right\}.$$

Assume that $p < m < n$. Then $X^{\mathrm{T}}S^{\mathrm{T}}SX$ is of rank p and then invertible. So

$$\begin{aligned}
R_{(S,X)}\left(\widehat{\beta}^S;\beta\right) = V_{(S,X)}(\widehat{\beta}^S,\beta) &= \sigma^2\mathrm{tr}\left\{(X^{\mathrm{T}}S^{\mathrm{T}}SX)^{-1}\right\} = \frac{\sigma^2}{p}\mathrm{tr}\left\{(X^{\mathrm{T}}S^{\mathrm{T}}SX/p)^{-1}\right\} \\
&= \sigma^2\int \frac{1}{t} dF^{X^{\mathrm{T}}S^{\mathrm{T}}SX/p}(t) =: \sigma^2 s_{1n}(0),
\end{aligned}$$

where $s_{1n}(\cdot)$ denotes the Stieltjes transformation of the ESD $F^{X^{\mathrm{T}}S^{\mathrm{T}}SX/n}$ of $X^{\mathrm{T}}S^{\mathrm{T}}SX/n \in \mathbb{R}^{p\times p}$.

Let H_n denote the ESD of $S^{\mathrm{T}}S \in \mathbb{R}^{n\times n}$ and H its LSD. Define

$$B_n := \frac{1}{p}(S^{\mathrm{T}}S)^{1/2}XX^{\mathrm{T}}(S^{\mathrm{T}}S)^{1/2} \in \mathbb{R}^{n\times n}.$$

Under Assumptions 7, the matrix B_n has the LSD $F^{\phi^{-1},H}$ which is the Marcehnko-Pastur law. Further, we define

$$G_n(t) : = n\left\{F^{B_n}(t) - F^{\phi_n^{-1},H_n}(t)\right\}$$

where we use $F^{\phi_n^{-1},H_n}$ instead of $F^{\phi^{-1},H}$ to avoid discussing the convergence of (ϕ_n^{-1},H_n) to (ϕ^{-1},H). For orthogonal sketching, $H_n = (1-\psi_n)\delta_0 + \psi_n\delta_1$ and $H = (1-\psi)\delta_0 + \psi\delta_1$. Notice that

$$G_n(t) = p\left\{F^{X^{\mathrm{T}}S^{\mathrm{T}}SX/p}(t) - \underline{F}^{\phi_n^{-1},H_n}(t)\right\}$$

with $\underline{F}^{\phi_n^{-1},H_n} := (1-\phi_n^{-1})\delta_0 + \phi_n^{-1}F^{\phi_n^{-1},H_n}$. By Theorem 3.2, we have

$$R_{(S,X)}(\widehat{\beta}^S,\beta) \xrightarrow{a.s.} \frac{\sigma^2\phi\psi^{-1}}{1-\phi\psi^{-1}}.$$

Further, we can rewrite

$$p\left(R_{(S,X)}(\widehat{\beta}^S,\beta)-\frac{\sigma^2\phi_n\psi_n^{-1}}{1-\phi_n\psi_n^{-1}}\right)=\sigma^2\int\frac{1}{t}dG_n(t), \tag{44}$$

where we replaced (ϕ,ψ) by (ϕ_n,ψ_n) when centering.

We prove the CLT for (44) in the following two steps.

Step 1. Given the sketching matrix S, by (Zheng et al., 2015, Theorem 2.1), the RHS of (44) converges to a Gaussian distribution with mean μ_1 and variance σ_1^2 specified as

$$\begin{aligned}\mu_1=&-\frac{\sigma^2}{2\pi i}\oint_{\mathcal{C}}\frac{1}{z}\frac{\phi^{-1}\int\underline{s}_\phi(z)^3t^2(1+t\underline{s}_\phi(z))^{-3}dH(t)}{\left\{1-\phi^{-1}\int\underline{s}_\phi^2t^2(1+t\underline{s}_\phi(z))^{-2}dH(t)\right\}^2}dz\\&-\frac{\sigma^2(\nu_4-3)}{2\pi i}\oint_{\mathcal{C}}\frac{1}{z}\frac{\phi^{-1}\int\underline{s}_\phi(z)^3t^2(1+t\underline{s}_\phi(z))^{-3}dH(t)}{1-\phi^{-1}\int\underline{s}_\phi^2t^2(1+t\underline{s}_\phi(z))^{-2}dH(t)}dz\end{aligned}$$

and

$$\begin{aligned}\sigma_1^2=&-\frac{2\sigma^4}{4\pi^2}\oint_{\mathcal{C}_1}\oint_{\mathcal{C}_2}\frac{1}{z_1z_2}\frac{1}{(\underline{s}_\phi(z_1)-\underline{s}_\phi(z_2))^2}d\underline{s}_\phi(z_1)d\underline{s}_\phi(z_2)\\&-\frac{\phi^{-1}(\nu_4-3)}{4\pi^2}\oint_{\mathcal{C}_1}\oint_{\mathcal{C}_2}\frac{1}{z_1z_2}\left\{\int\frac{t}{(\underline{s}_\phi(z_1)+1)^2}\frac{t}{(\underline{s}_\phi(z_2)+1)^2}dH(t)\right\}d\underline{s}_\phi(z_1)d\underline{s}_\phi(z_2),\end{aligned}$$

where $\underline{s}_\phi(\cdot)$ denotes the Stieltjes transformation of $\underline{F}^{\phi^{-1},H}$, and $\mathcal{C}$, $\mathcal{C}_1$ and $\mathcal{C}_2$ are contours containing the support of the LSD of $\underline{s}_\phi(z)$.

Following the calculation of μ_c and σ_c^2 in the proof of (Li et al., 2021, Theorem 4.1), we get

$$\mu_1=\frac{\sigma^2\phi^2\psi^{-2}}{(\phi\psi^{-1}-1)^2}+\frac{\sigma^2\phi^2\psi^{-2}(\nu_4-3)}{1-\phi\psi^{-1}},\quad\sigma_1^2=\frac{2\sigma^4\phi^3\psi^{-3}}{(\phi\psi^{-1}-1)^4}+\frac{\phi^3\psi^{-3}\sigma^4(\nu_4-3)}{(1-\phi\psi^{-1})^2}.$$

Step 2. Note that the mean μ_1 and variance σ_1^2 are nonrandom. It means that the limiting distribution of the RHS of (44) is independent of conditioning SS^{T}. So it asymptotically follows the Gaussian distribution $\mathcal{N}(\mu_1,\sigma_1^2)$.

H.2. Proof of Theorem B.3

For the overparameterized regime with $\phi\psi^{-1}>1$, when $\Sigma=I_p$, we have

$$\begin{aligned}B_{(S,X)}(\widehat{\beta}^S,\beta)&=\frac{\alpha^2}{p}\mathrm{tr}\left\{I_p-(X^{\mathrm{T}}S^{\mathrm{T}}SX)^+X^{\mathrm{T}}S^{\mathrm{T}}SX\right\}=\alpha^2-\frac{\alpha^2}{p}\mathrm{tr}\left\{(X^{\mathrm{T}}S^{\mathrm{T}}SX)^+X^{\mathrm{T}}S^{\mathrm{T}}SX\right\}\\&=\alpha^2-\frac{\alpha^2}{p}\mathrm{tr}\left\{(SXX^{\mathrm{T}}S^{\mathrm{T}})^+SXX^{\mathrm{T}}S^{\mathrm{T}}\right\}=\alpha^2\left(1-\phi^{-1}\psi_n\right)\end{aligned}$$

and

$$\begin{aligned}V_{(S,X)}(\widehat{\beta}^S,\beta)&=\sigma^2\mathrm{tr}\left\{(X^{\mathrm{T}}S^{\mathrm{T}}SX)^+\right\}=\sigma^2\frac{1}{n}\mathrm{tr}\left\{(SXX^{\mathrm{T}}S^{\mathrm{T}}/n)^{-1}\right\}\\&=\sigma^2\int\frac{1}{t}dF^{SXX^{\mathrm{T}}S^{\mathrm{T}}/n}(t)=:\sigma^2\underline{s}_{1n}(0),\end{aligned}$$

where $\underline{s}_{1n}(z)$ denotes the Stieltjes transformation of the ESD $F^{SXX^{\mathrm{T}}S^{\mathrm{T}}/n}$ of $SXX^{\mathrm{T}}S^{\mathrm{T}}/n$ and it satisfies

$$\frac{p}{m}s_{1n}(z)=-\frac{1}{z}(\frac{p}{m}-1)+\underline{s}_{1n}(z).$$

Following the proof of (Li et al., 2021, Theorem 4.3), we get

$$\mu_2 = \frac{\sigma^2 \phi \psi^{-1}}{(\phi\psi^{-1}-1)^2} + \frac{\sigma^2(\nu_4 - 3)}{\phi\psi^{-1}-1}, \quad \sigma_2^2 = \frac{2\sigma^4\phi^3\psi^{-3}}{(\phi\psi^{-1}-1)^4} + \frac{\sigma^4\phi\psi^{-1}(\nu_4-3)}{(\phi\psi^{-1}-1)^2}.$$

H.3. Proof of Theorem B.5

Leveraging (Bai & Yao, 2008, Theorem 7.2) or following a similar proof to that of (Li et al., 2021, Theorem 4.5), we obtain

$$\begin{aligned}&\sqrt{p}\left\{B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) - \alpha^2(1-\phi_n^{-1}\psi_n)\right\}\\ =&\sqrt{p}\left\{\beta^{\mathrm{T}}\left[I_p - \left(X^{\mathrm{T}}V^{\mathrm{T}}VX\right)^{+}X^{\mathrm{T}}V^{\mathrm{T}}VX\right]\beta - \alpha^2(1-\phi_n^{-1}\psi_n)\right\} \xrightarrow{D} \mathcal{N}(0, d^2 = d_1^2 + d_2^2),\end{aligned}$$

where

$$d_1^2 = wp^2\left(\mathbb{E}(\beta_1^4) - \gamma^2\right), \; d_2^2 = 2p^2(\theta - w)\gamma^2$$

and

$$\gamma = \mathbb{E}(\beta_1^2) = \frac{\alpha^2}{p}, \; \theta = \lim_{p\to\infty}\frac{1}{p}\mathrm{tr}\left\{I_p - (X^{\mathrm{T}}S^{\mathrm{T}}SX)^{+}X^{\mathrm{T}}S^{\mathrm{T}}SX\right\} = 1 - \phi^{-1}\psi.$$

Here w is the limit of the average of squared diagonal elements of $\left[I_p - \left(X^{\mathrm{T}}V^{\mathrm{T}}VX\right)^{+}X^{\mathrm{T}}V^{\mathrm{T}}VX\right]$, which will be canceled out in d^2 under the assumption that β is multivariate normal. After some simple calculation, we have $d^2 = 2(1-\phi^{-1}\psi)\alpha^4$. Thus,

$$\sqrt{p}\left\{B_{(\beta,S,X)}(\widehat{\beta}^S;\beta) - \alpha^2(1-\phi_n^{-1}\psi_n)\right\} \xrightarrow{D} \mathcal{N}(0, 2(1-\phi^{-1}\psi)\alpha^4). \tag{45}$$

Moreover, we have proved in the Theorem B.3 that

$$p\left(V_{(\beta,S,X)}(\widehat{\beta}^S;\beta) - \frac{\sigma^2}{\phi_n\psi_n^{-1}-1}\right) \xrightarrow{D} \mathcal{N}(\mu_2, \sigma_2^2) \tag{46}$$

where

$$\mu_2 = \frac{\sigma^2 \phi \psi^{-1}}{(\phi\psi^{-1}-1)^2} + \frac{\sigma^2(\nu_4 - 3)}{\phi\psi^{-1}-1}, \quad \sigma_2^2 = \frac{2\sigma^4\phi^3\psi^{-3}}{(\phi\psi^{-1}-1)^4} + \frac{\sigma^4\phi\psi^{-1}(\nu_4-3)}{(\phi\psi^{-1}-1)^2}.$$

Combining (45) and (46) completes the proof.