Sup-Norm Error under Proportional Asymptotics: Phase Transitions under Linear Regression

Lin Liu
Shanghai Jiao Tong University
linliu@sjtu.edu.cn

Debarghya Mukherjee
Boston University
mdeb@bu.edu

Rajarshi Mukherjee
Harvard University
ram521@mail.harvard.edu

Zixiao Jolene Wang
Harvard University
zixiaowang@fas.harvard.edu

Abstract

This paper characterizes the sharp asymptotics of the ℓ_∞-norm estimation error for estimating the coefficient vector in linear regression under proportional asymptotics. While ℓ_2-error is well-documented, we demonstrate that the ℓ_∞-error exhibits a distinct phase transition governed by the structure of the underlying signal. We explore this through the lens of a class of estimators that include ridge-regularized estimators, a method-of-moments estimator, and the ordinary least squares estimator for the under-parametrized regime. Our results provide a comparative study of these estimators and characterize regimes (as identified by the maximal signal component, number of large signals, total signal strength, and noise variance) where one estimator is preferable to the others. As a by-product, we obtain new double descent curves in ℓ_∞-error metric for the ridgeless regression, as well as ingredients for uniform confidence intervals for the signal coordinates. Our theoretical results are supported by extensive numerical simulations that confirm the predicted phase transitions and error limits.

1 Introduction

Statistical inference within the linear regression framework remains a foundational pillar of statistics and machine learning. While traditionally viewed as a textbook topic [10, 19], linear regression has also served as a primary template for understanding high-dimensional statistical phenomena over the past two decades [9, 21, 33, 31, 22]. In this paper, we study the canonical linear model

$$Y_i = X_i^\top \beta_0 + \varepsilon_i, \qquad i = 1, \dots, n,$$

where $X_i \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}^p$ is the unknown regression vector, and ε_i has mean zero and variance σ_ε^2. Classical fixed-dimensional conclusions have been substantially revisited in modern regimes where the dimension $p \gtrsim n$. This line of work has led to a precise understanding of estimation, testing, and uncertainty quantification, revealing phase transitions governed by the interplay among dimension, sample size, noise level, and signal structure. Within this high-dimensional paradigm, two sub-regimes have emerged as central. The first is the ultra-high-dimensional regime, where $p \gg n$ and one typically imposes sparsity on β_0, leading to statistically efficient procedures such as the Lasso and SLOPE [9, 33, 21]. The second is the proportional asymptotic regime [23, 16, 28], where $p/n \to \gamma \in (0, \infty)$. This regime permits rigorous analysis without requiring sparsity assumptions. Significant progress has been made in understanding regularized estimators in this setting, especially their mean-squared error and the associated optimal tuning of regularization parameters [15, 28]. However, unlike in fixed dimensions, the asymptotic behavior of estimation error can depend strongly on the norm under consideration. In particular, the behavior of the ℓ_q error changes substantially

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

between $q \in [2, \infty)$ and $q = \infty$ [35]. While this distinction has been studied for Lasso-type estimators under sparsity in ultra-high-dimensional settings [35], parallel results in the proportional asymptotic regime remain scarce. For ℓ_q errors with $q \in [2, \infty)$, tools such as Approximate Message Passing (AMP) provide sharp asymptotic descriptions for many regularized estimators. These methods, however, do not directly yield a precise analysis of the ℓ_∞ estimation error for convex-regularized linear regression under proportional asymptotics [28, 17]. The main motivation of this paper is to take a step toward addressing this gap.

At this point, it is useful to highlight a salient distinction between ℓ_∞ error and the more commonly studied ℓ_q errors, $q \in [2, \infty)$, in high-dimensional regression. When $p \geq n$, consistent estimation in ℓ_q norm for $q \in [2, \infty)$ is impossible without additional structural assumptions, such as sparsity, on the true signal. Accordingly, a large literature on sparse regularized regression seeks to identify regimes in which consistent estimation becomes possible and to establish sharp optimality properties once consistency is attained [9, 33, 21, 32, 1]. In contrast, in the proportional asymptotic regime, without sparsity, estimation in ℓ_q norm for $q \in [2, \infty)$ remains intrinsically inconsistent. Consequently, much of the existing work in this regime has focused on deriving precise asymptotic risk formulas for regularized estimators rather than on consistency itself. From a statistical optimality perspective, this distinction is important. For example, for ℓ_2-bounded signals in Gaussian linear regression, optimally tuned ridge regression attains the sharp minimax risk under squared ℓ_2 loss [13]; analogous minimax comparisons can also be formulated for general ℓ_q losses. However, because every estimator remains inconsistent in these metrics when $p/n \to \gamma \in (0, \infty)$, such results compare the risks of inconsistent estimators. The situation changes fundamentally for $q = +\infty$: consistent estimation in ℓ_∞ norm is possible in various regimes even without sparsity assumptions [35]. This makes the ℓ_∞ metric for estimation error especially attractive, since it allows one to compare consistent estimators whose errors decrease to 0 at a certain rate, even in the non-sparse proportional regime. Moreover, unlike the ℓ_2 error, the ℓ_∞ error is sensitive not only to the overall signal strength $\|\beta_0\|_2$, but also to the coordinatewise structure of β_0, in particular the size and distribution of its largest entries [35]. This paper aims to understand these subtleties through the lens of several natural estimators: the null estimator, a method-of-moments estimator, the OLS estimator (for $p < n$), and ridge-regularized estimators. Since a full, sharp minimax theory beyond rates is not yet available in this setting, our goal is to develop a precise estimator-level theory. We derive sharp asymptotic limits for the ℓ_∞ error of ridge regression and compare with the null estimator ($\widehat{\beta} \equiv 0$), the method-of-moments estimator, and the ordinary least squares estimator (for $p < n$). We then identify how the preferred estimator depends on the coordinatewise signal strength $\|\beta_0\|_\infty$, the global signal strength $\|\beta_0\|_2$, and the noise level. Our results reveal phase transitions that are invisible at the level of ℓ_2 risk. They also provide ℓ_∞-optimal tuning rules for ridge regression in the regimes considered, clarify when optimally tuned ridge outperforms the competing estimators, and uncover a double-descent phenomenon for uniform estimation error in the ridgeless limit. We summarize our contributions below.

1. We develop a theory for the limiting ℓ_∞ estimation error of several estimators of the regression parameter. These include the ridge-regularized estimator, a method-of-moments estimator, and the ordinary least squares estimator when $p < n$. Our analysis is organized around two regimes: the strong-signal regime, where $\|\beta_0\|_\infty$ dominates the noise level namely when $\|\beta_0\|_\infty \gg \sqrt{\log p/n}$, and the weak-signal regime, where $\|\beta_0\|_\infty$ is comparable to, or smaller than, the noise level (i.e., $\|\beta_0\|_\infty \asymp \sqrt{\log p/n}$ or even $\|\beta_0\|_\infty \ll \sqrt{\log p/n}$) . We further extend the analysis to a random-effects model, in which the coordinates of β_0 are generated from an underlying distribution (including both light- and heavy-tailed).
2. Our results provide a comparative study of these estimators and identify regimes in which one estimator is preferable to the others. These regimes are determined by the interplay among $(\|\beta_0\|_\infty, \|\beta_0\|_2, \gamma, \sigma_\epsilon^2)$. We further demonstrate a double-descent curve for the ridgeless regression case, corresponding to the limit $\lambda \downarrow 0$, thereby complementing the ℓ_2-error and prediction-risk results of [22]. Our results provide first-order ingredients for constructing uniform confidence intervals for the regression coordinates, with potential implications for simultaneous testing and false discovery control.
3. On the technical side, our proofs rely on a delicate leave-two-out argument to handle the maximum of correlated Gaussian random variables in the weak-signal regime. This technique may be of independent interest.
4. Finally, our theoretical findings are supported by extensive numerical experiments.

2 Problem Formulation and Preliminaries

We now formalize the model, the estimators, and the notation used in the main results. As mentioned in the Introduction, we observe independent samples $\{(X_i, Y_i)\}_{i=1}^n$ from the linear model $Y_i = X_i^\top \beta_0 + \epsilon_i$ for $1 \le i \le n$, where where $X_i \in \mathbb{R}^p$ is the covariate vector, $Y_i \in \mathbb{R}$ is a continuous response, $\beta_0 \in \mathbb{R}^p$ is the unknown regression coefficient vector, and $\varepsilon_i \in \mathbb{R}$ is a mean-zero random error with variance σ_ε^2.

Throughout the main text, $\beta_0 = \beta_{0,n}$ is allowed to depend on n, although we suppress this dependence in the notation. Our primary object of interest is the coordinatewise worst-case estimation error $\|\widehat{\beta} - \beta_0\|_\infty$. We work under the following isotropic Gaussian design assumption.

Assumption 1. The covariates $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ and independent of errors. The errors $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. Moreover, $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

A few remarks are in order regarding the isotropic Gaussian assumption. First, this is mostly used to yield a more compact proof for the weak signal case considered in Section 3.2. As we note there, one can indeed extend this using detailed Cramer-type moderate deviation bounds together with universality arguments (e.g., Lindeberg swapping). Since the proof is already long and delicate even in the Gaussian case, we do not consider them here. The more subtle point is the isotropic nature of the covariates. A close look at our proofs will reveal that this allows us to carefully reduce the problem of the sup-norm asymptotics into a relevant weakly correlated set of Gaussians. However, for arbitrary feature covariance, the problem of sup-norm asymptotics is equivalent to studying sharp asymptotics of the maxima of arbitrarily correlated Gaussians – a problem that is still being studied in modern probability through specific instances [8].

We use the notation $\mathbf{X} \in \mathbb{R}^{n \times p}$ to denote the matrix of all covariates stacked rowwise, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ to denote the noise vector and $\mathbf{y} \in \mathbb{R}^n$ to denote the response vector. We will use $\widehat{\Sigma} = \mathbf{X}^\top \mathbf{X}/n$ to denote the sample covariance matrix. The role of the coordinatewise signal strength $\|\beta_0\|_\infty$ will be more delicate and is specified separately in the strong- and weak-signal regimes.

Marchenko–Pastur Law and Stieltjes Transforms We use F_γ for the MP distribution with aspect ratio γ, and $\widetilde{F}_\gamma$ for its companion distribution. If $\mathbf{X} \in \mathbb{R}^{n \times p}$ has i.i.d. standardized entries, and $p/n \to \gamma$, then F_γ and $\widetilde{F}_\gamma$ are the limiting empirical spectral distributions of $\mathbf{X}^\top \mathbf{X}/n$ and $\mathbf{X}\mathbf{X}^\top/n$, respectively [5]. We defines

$$m_\gamma(z) = \int \frac{1}{x - z}\, dF_\gamma(x), \qquad \widetilde{m}_\gamma(z) = \int \frac{1}{x - z}\, d\widetilde{F}_\gamma(x),$$

as the Stieltjes transforms of F_γ and $\widetilde{F}_\gamma$. m_γ admits a precise expression

$$m_\gamma(z) = \frac{1 - \gamma - z - \sqrt{(z - \gamma - 1)^2 - 4\gamma}}{2\gamma z}.$$

Moreover,

$$\widetilde{m}_\gamma(z) = -\frac{1 - \gamma}{z} + \gamma m_\gamma(z), \qquad \widetilde{m}_\gamma(-\lambda) = \frac{1 - \gamma}{\lambda} + \gamma m_\gamma(-\lambda).$$

We use $m_\gamma'(z)$ to denote the derivative of $m_\gamma(z)$ with respect to z.

2.1 Estimators

We compare several natural estimators of β_0. The first is the ridge-regularized estimator

$$\widehat{\beta}_\lambda = \left(\widehat{\Sigma} + \lambda \mathbf{I}_p\right)^{-1} \tfrac{\mathbf{X}^\top \mathbf{y}}{n}, \qquad \lambda > 0.$$

We also consider the null estimator $\widehat{\beta}_{\text{null}} \equiv \mathbf{0}_p$. If $\|\beta_0\|_\infty \downarrow 0$ as $p \uparrow \infty$ then trivially $\widehat{\beta}_{\text{null}}$ is a consistent estimator. Our next estimator is a method-of-moments (MoM) estimator, which was extensively analyzed in [11] with respect ℓ_2 estimation and prediction error. Under the isotropic

normalization in Assumption 1, the moment equation, we have the moment condition $\mathbb{E}[X_i Y_i] = \beta_0$, which immediately suggests the method-of-moments estimator:

$$\widehat{\beta}_{\mathrm{MoM}} = \frac{\mathbf{X}^\top \mathbf{y}}{n}.$$

More generally, if X has a known covariance matrix $\Sigma = \mathrm{Var}(X)$, the corresponding method-of-moments estimator is $\widehat{\beta}_{\mathrm{MoM}} = \Sigma^{-1}(\mathbf{X}^\top \mathbf{y}/n)$. Thus, in the isotropic case $\Sigma = \mathbf{I}_p$, this reduces to the estimator above. Last but not least, when $\gamma < 1$, the ordinary least squares estimator is also well-defined with probability tending to one:

$$\widehat{\beta}_{\mathrm{OLS}} = \widehat{\Sigma}^{-1} \frac{\mathbf{X}^\top \mathbf{y}}{n}.$$

When $\gamma > 1$, the ridgeless limit of the ridge path corresponds to the minimum-norm interpolator

$$\widehat{\beta}_{\mathrm{ridgeless}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y},$$

again well-defined with probability tending to one under Assumption 1. The main results compare these estimators through the first-order asymptotics of their ℓ_∞ estimation errors. To understand the interplay between signal strength and noise, let us consider the decomposition of the error of the ridge estimator:

$$\widehat{\beta}_\lambda - \beta_0 = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \mathbf{y}}{n} - \beta_0 = -\lambda (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}.$$

The first term is the regularization bias, while the second term is the stochastic noise contribution. Under proportional asymptotics, the maximum of p approximately Gaussian coordinates with variance of order $1/n$ is naturally of order $\sqrt{2 \log p/n}$ [25]. Therefore, the behavior of $\|\widehat{\beta}_\lambda - \beta_0\|_\infty$ depends on the relative size of the coordinatewise signal strength $\|\beta_0\|_\infty$ and the noise scale r_n. This observation motivates the two regimes studied in Section 3. In the strong-signal regime, $\|\beta_0\|_\infty \gg \sqrt{\log p/n}$, the ridge error is dominated by the regularization bias, whose leading size is proportional to $\|\beta_0\|_\infty$. In the weak-signal regime, $\|\beta_0\|_\infty \lesssim \sqrt{\log p/n}$, the bias and stochastic terms are of comparable order and interact in the ℓ_∞ norm. This interaction is the main technical difficulty in the weak-signal analysis. However, the competing estimators have simpler decompositions. For the method-of-moments estimator,

$$\widehat{\beta}_{\mathrm{MoM}} - \beta_0 = (\widehat{\Sigma} - \mathbf{I}_p)\beta_0 + \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}.$$

Thus, its coordinate-wise error is driven by the combined fluctuation of the design and the noise. When $\gamma < 1$, the OLS estimator satisfies

$$\widehat{\beta}_{\mathrm{OLS}} - \beta_0 = \widehat{\Sigma}^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n},$$

and hence its leading ℓ_∞ behavior is purely noise-driven. These decompositions explain why the limiting constants below depend differently on $\|\beta_0\|_\infty$, $\|\beta_0\|_2$, σ_ϵ^2, and γ.

2.2 Comparison criterion

Our comparisons are based on first-order asymptotic constants for the normalized ℓ_∞ error. If two estimators converge on different scales, the estimator with the smaller scale is preferred. If they converge on the same scale, we compare the corresponding limiting constants. In particular, when we refer to an optimally tuned ridge estimator, we mean the choice of the fixed regularization parameter $\lambda > 0$ that minimizes the relevant limiting ℓ_∞ constant in the regime under consideration. The next section gives the main results. We first consider the strong-signal regime, where the ridge bias dominates the stochastic error. We then turn to the weak-signal regime, where the largest signal coordinates and the largest noise coordinates are on the same scale and must be analyzed jointly.

3 Theoretical Analysis

3.1 Strong Signal Regime

In this subsection, we study the scaled ℓ_∞ estimation error of the four estimators in the strong-signal regime, namely when $\|\beta_0\|_\infty \gg \sqrt{\log p/n}$. More precisely, we impose the following condition:

Assumption 2. The regression parameter β_0 satisfies $\|\beta_0\|_2^2 = B$ and there exists a deterministic sequence r_n such that

$$r_n / \sqrt{2 \log p/n} \to \infty, \qquad \|\beta_0\|_\infty / r_n \to c^* \qquad \text{as } n \to \infty.$$

The following theorem characterizes the limiting ℓ_∞ error of the four estimators under Assumption 2.

Theorem 3.1 (Strong signal limit)**.** *Under Assumption 1 and Assumption 2, for every fixed $\lambda > 0$,*

$$\frac{\|\widehat{\beta}_\lambda - \beta_0\|_\infty}{r_n} \xrightarrow{p} \lambda m_\gamma(-\lambda)c^*, \qquad \frac{\|\widehat{\beta}_{\mathrm{null}} - \beta_0\|_\infty}{r_n} \xrightarrow{p} c^*,$$
$$\frac{\|\widehat{\beta}_{\mathrm{MoM}} - \beta_0\|_\infty}{\sqrt{2\log p/n}} \xrightarrow{p} \sqrt{B + \sigma_\epsilon^2}, \qquad \frac{\|\widehat{\beta}_{\mathrm{OLS}} - \beta_0\|_\infty}{\sqrt{2\log p/n}} \xrightarrow{p} \frac{\sigma_\epsilon}{\sqrt{1-\gamma}} \ (\textit{for } \gamma \in (0,1)),$$

where $m_\gamma(\cdot)$ denotes the Stieltjes transform of the Marchenko-Pasteur law as defined in Section 2 Furthermore, if $r_n = n^{-1/2+\alpha_*}$ for some $\alpha_* \in (0, 1/2)$ and $\gamma > 1$, then the analysis extends to the ridgeless case $\lambda = 0$. In particular, for the minimum-norm interpolator $\widehat{\beta}_{\mathrm{ridgeless}}$,

$$\|\widehat{\beta}_{\mathrm{ridgeless}} - \beta_0\|_\infty / r_n \xrightarrow{p} c^*(1 - \gamma^{-1}).$$

Theorem 3.1 reveals a sharp distinction between ridge-type estimators and the simpler method-of-moments (MoM) estimator in the strong-signal regime. Since $r_n \gg \sqrt{2\log p/n}$, both the ridge estimator $\widehat{\beta}_\lambda$ for any fixed $\lambda > 0$ and the null estimator $\widehat{\beta}_{\mathrm{null}}$ have substantially larger ℓ_∞ error than $\widehat{\beta}_{\mathrm{MoM}}$. Thus, in this regime, fixed-regularization ridge estimators are strictly suboptimal relative to the moment estimator. When $\gamma \in (0, 1)$, the OLS estimator also achieves the faster $\sqrt{2\log p/n}$ scale. An important feature of these limits is that the leading constant for $\widehat{\beta}_{\mathrm{MoM}}$ depends on the signal only through $B = \|\beta_0\|_2^2$, and not through the stronger quantity $\|\beta_0\|_\infty$. By contrast, the limiting constant for $\widehat{\beta}_{\mathrm{OLS}}$ depends neither on $\|\beta_0\|_\infty$ nor on $\|\beta_0\|_2$. Therefore, for $\gamma \in (0, 1)$, there is a natural trade-off between $\widehat{\beta}_{\mathrm{MoM}}$ and $\widehat{\beta}_{\mathrm{OLS}}$: OLS has the smaller limiting constant precisely when

$$\sigma_\epsilon^2 \le (1-\gamma)(B + \sigma_\epsilon^2) \iff (1-\gamma)B \ge \gamma\sigma_\epsilon^2.$$

Otherwise, $\widehat{\beta}_{\mathrm{MoM}}$ has the smaller limiting constant. For $\gamma > 1$, the ridgeless estimator is the best member of the ridge path in this strong-signal scale. Indeed, among $\{\widehat{\beta}_\lambda : \lambda \ge 0\}$, the limiting constant $\lambda m_\gamma(-\lambda)$ is minimized at $\lambda = 0$ for $\gamma > 1$. Consequently, for $\gamma > 1$, the ranking is

$$\widehat{\beta}_{\mathrm{MoM}} \succ \widehat{\beta}_{\mathrm{ridgeless}} \succ \{\widehat{\beta}_\lambda : \lambda > 0\}.$$

For $\gamma < 1$, the ranking depends on the size of B. If $B > \gamma\sigma_\epsilon^2/(1-\gamma)$, then

$$\widehat{\beta}_{\mathrm{OLS}} \succ \widehat{\beta}_{\mathrm{MoM}} \succ \{\widehat{\beta}_\lambda : \lambda > 0\},$$

otherwise, the preference between $\widehat{\beta}_{\mathrm{OLS}}$ and $\widehat{\beta}_{\mathrm{MoM}}$ will be reversed.

Remark 1 (Beyond Gaussian designs). A careful inspection of the proof shows that the limit for the ridge family $\{\widehat{\beta}_\lambda : \lambda \ge 0\}$ is not intrinsically Gaussian. The same conclusion continues to hold when the entries X_{ij} are independent, mean-zero, variance-one sub-Gaussian random variables. We also expect the limits for $\widehat{\beta}_{\mathrm{MoM}}$ and $\widehat{\beta}_{\mathrm{OLS}}$ to remain valid beyond Gaussianity, through a suitable application of Cramér-type moderate deviation bounds. We leave this extension as an interesting direction for future work.

Extension of anisotropic designs. The preceding result was stated for isotropic covariates. We now present the corresponding extension to anisotropic Gaussian designs. Suppose $X_i = \Sigma^{1/2} Z_i$, where $Z_i \sim \mathcal{N}(0, \mathbf{I}_p)$ and Σ is positive definite with eigenvalues uniformly bounded away from zero and infinity. In this case, the ℓ_∞ error depends on the geometry of Σ and its alignment with β_0.

Corollary 3.2 (Anisotropic extension)**.** *Consider the anisotropic Gaussian design described above. Suppose there exists a sequence $r_n = n^{-1/2+\alpha}$, for some $\alpha > 0$, such that*

$$\|\{\mathbf{I}_p + \widetilde{m}_n(-\lambda)\Sigma\}^{-1}\beta_0\|_\infty \asymp r_n,$$

where $\widetilde{m}_n(\cdot)$ denotes the companion Stieltjes transform of the Marchenko–Pastur law, as defined in C.1. Then, for every fixed $\lambda > 0$,

$$r_n^{-1}\left|\|\widehat{\beta}_\lambda - \beta_0\|_\infty - \|\{\mathbf{I}_p + \widetilde{m}_n(-\lambda)\Sigma\}^{-1}\beta_0\|_\infty\right| \xrightarrow{p} 0.$$

Thus, under anisotropic designs, the leading ℓ_∞ error is governed by the effective signal $\{\mathbf{I}_p + \tilde{m}_n(-\lambda)\Sigma\}^{-1}\beta_0$, rather than by $\|\beta_0\|_\infty$ itself. When $\Sigma = \mathbf{I}_p$, this recovers the isotropic result through the usual relation between the Stieltjes transform and its companion. The dependence on Σ is especially transparent for the spiked covariance model $\Sigma = \mathbf{I}_p + \sum_{j=1}^k \delta_j v_j v_j^\top$, which includes the equicorrelation model as a special case; see, for example, [22]. A direct calculation yields

$$\{\mathbf{I}_p + \widetilde{m}_n(-\lambda)\Sigma\}^{-1}\beta_0 = \frac{\beta_0 - \sum_{j=1}^k \rho_j v_j \langle v_j, \beta_0\rangle}{1+\widetilde{m}_n(-\lambda)}, \qquad \rho_j = \frac{\delta_j \widetilde{m}_n(-\lambda)}{1+\widetilde{m}_n(-\lambda)+\delta_j \widetilde{m}_n(-\lambda)}.$$

Hence, the limit depends jointly on the spike magnitudes and directions, and their alignment with β_0.

Implications for uniform confidence intervals. . Theorem 3.1 suggests first-order simultaneous intervals for the coordinates of β_0 when $\|\beta_0\|_\infty$ is not vanishingly small. For $\gamma > 1$, the theorem suggests using $\widehat{\beta}_{\mathrm{MoM}}$, since it achieves the faster $\sqrt{2\log p/n}$ scale. To form a feasible interval, we need consistent estimators of $B = \|\beta_0\|_2^2$ and σ_ϵ^2. One possible choice is

$$\widehat{\sigma}_\epsilon^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^n \big(Y_i - X_i^\top \widehat{\beta}_{\mathrm{MoM}}\big)^2, \qquad \widehat{B} = \tfrac{1}{n(n-1)}\textstyle\sum_{i\neq j} Y_i Y_j X_i^\top X_j.$$

The consistency of $\widehat{\sigma}_\epsilon^2$ and $\widehat{B}$ has been established in [11]. Therefore, a uniform confidence interval based on $\widehat{\beta}_{\mathrm{MoM}}$ takes the form

$$\widehat{\mathrm{CI}}_j^{\mathrm{MoM}} = \left[\widehat{\beta}_{\mathrm{MoM},j} - \sqrt{2(\widehat{B}+\widehat{\sigma}_\epsilon^2)\log p/n},\ \widehat{\beta}_{\mathrm{MoM},j} + \sqrt{2(\widehat{B}+\widehat{\sigma}_\epsilon^2)\log p/n}\right], \quad j = 1,\dots,p.$$

For $\gamma < 1$, analogous uniform intervals can be constructed. If $\widehat{B} < \gamma\widehat{\sigma}_\epsilon^2/(1-\gamma)$, then $\widehat{\beta}_{\mathrm{MoM}}$ is preferable to $\widehat{\beta}_{\mathrm{OLS}}$, and the same MoM-based interval can be used. Otherwise, one may instead use the OLS estimator with the following interval simultaneously for all $j \in [p]$:

$$\widehat{\mathrm{CI}}_j^{\mathrm{OLS}} = \left[\widehat{\beta}_{\mathrm{OLS},j} - \sqrt{(2\widehat{\sigma}_\epsilon^2\log p)/(n(1-\gamma))},\ \ \widehat{\beta}_{\mathrm{OLS},j} + \sqrt{(2\widehat{\sigma}_\epsilon^2\log p)/(n(1-\gamma))}\right].$$

3.2 Weak Signal Regime

In this section, we extend our theoretical analysis to the more delicate weak-signal regime, where $\|\beta_0\|_\infty \asymp \sqrt{\log p/n}$ or even $\|\beta_0\|_\infty \ll \sqrt{\log p/n}$. This regime is substantially more challenging than the strong-signal regime, especially for the ridge estimator $\widehat{\beta}_\lambda$ and the method-of-moments estimator $\widehat{\beta}_{\mathrm{MoM}}$. The main difficulty is that the coordinates of the largest signal and the largest noise are now of the same order. Since the maximum of independent Gaussian variables is of order $\sqrt{\log p}$, the signal and noise components interact nontrivially in the ℓ_∞ norm. We first formalize the signal-strength condition, in analogy with Assumption 2.

Assumption 3. Set $r_n = \sqrt{2\log p/n}$. Assume that there exists a constant $B \in (0,\infty)$ such that $\|\beta_0\|_2^2 = B$. Furthermore, assume that there exists a deterministic set $S_n \subset \{1,\dots,p\}$ with $s_n := |S_n|$ such that

$$\log s_n/\log p \longrightarrow \rho \in [0,1],$$

with the convention that $\rho = 0$ if s_n is bounded or $s_n = p^{o(1)}$. Moreover, for some $a \geq 0$,

$$\max_{j\in S_n}|(|\beta_{0,j}|/r_n) - a| \to 0, \qquad \max_{j\notin S_n}(|\beta_{0,j}|/r_n) \to 0.$$

Assumption 3 states that the largest coordinates of β_0 have magnitude approximately $a\sqrt{2\log p/n}$ and that there are roughly p^ρ such coordinates. The formulation also allows the case $a = 0$, which corresponds to $\|\beta_0\|_\infty \ll \sqrt{\log p/n}$, and also $\rho = 1$, in which case, s_n can be of the order at most $p/\log p$, which is needed for $\|\beta_0\|_2^2 = B \in (0,\infty)$. The following theorem characterizes the limiting ℓ_∞ error of the four estimators in this weak-signal regime.

Theorem 3.3 (Weak signal limit)**.** *Under Assumptions 1 and 3, we have, for every fixed $\lambda > 0$,*

$$\frac{\|\widehat{\beta}_\lambda - \beta_0\|_\infty}{r_n} \xrightarrow{p} \max\left\{\lambda m_\gamma(-\lambda)a + \tau_\lambda\sqrt{\rho},\ \tau_\lambda\right\}, \qquad \frac{\|\widehat{\beta}_{\mathrm{null}} - \beta_0\|_\infty}{r_n} \xrightarrow{p} a,$$

$$\frac{\|\widehat{\beta}_{\mathrm{MoM}} - \beta_0\|_\infty}{r_n} \xrightarrow{p} \sqrt{B+\sigma_\epsilon^2}, \qquad \frac{\|\widehat{\beta}_{\mathrm{OLS}} - \beta_0\|_\infty}{r_n} \xrightarrow{p} \frac{\sigma_\epsilon}{\sqrt{1-\gamma}}, \ \ (\textit{for } \gamma \in (0,1)).$$

Here

$$\tau_\lambda^2 = m_\gamma^2(-\lambda)\left\{\lambda^2\big(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda)\big)B + \sigma_\epsilon^2 \int \tfrac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_\gamma(x)\right\},$$

where $\widetilde{F}_\gamma$ denotes the companion Marchenko–Pastur distribution.

Compared to Theorem 3.1, all four estimators converge on the same scale $\sqrt{\log p/n}$ in the weak-signal regime. Therefore, no single estimator is uniformly preferred on the basis of rate alone. Indeed, as our simulations in Section 4 demonstrate, there are parameter regimes in which each estimator can outperform the others in terms of its limiting constant. Nevertheless, the theorem reveals an important structural distinction. The limits for $\widehat{\beta}_{\mathrm{OLS}}$ and $\widehat{\beta}_{\mathrm{MoM}}$ do not depend on $\|\beta_0\|_\infty$ directly, whereas the limits for $\widehat{\beta}_\lambda$ and $\widehat{\beta}_{\mathrm{null}}$ depend on the boundary signal strength through the parameter a. Thus, when a is large, one may expect $\widehat{\beta}_{\mathrm{OLS}}$ or $\widehat{\beta}_{\mathrm{MoM}}$ to outperform ridge-type estimators. The precise comparison, however, depends strongly on whether $\gamma > 1$ or $0 \le \gamma < 1$; the next theorem makes this comparison explicit. For $\gamma > 1$, if a is larger than a suitable threshold, then $\widehat{\beta}_{\mathrm{MoM}}$ strictly outperforms the optimally tuned ridge estimator. In contrast, for $0 < \gamma < 1$, the optimal ridge risk is always no larger than the OLS limiting risk. Consequently, when $B > \gamma\sigma_\epsilon^2/(1-\gamma)$, we obtain the following ranking:

$$\text{optimally tuned ridge} \succ \widehat{\beta}_{\mathrm{OLS}} \succ \widehat{\beta}_{\mathrm{MoM}}. \tag{3.1}$$

Theorem 3.4 (Comparison with optimally tuned ridge)**.** *Let*

$$R^*_{\mathrm{ridge}} := \inf_{\lambda>0} \max\left\{\tau_\lambda,\, a\lambda m_\gamma(-\lambda) + \sqrt{\rho}\,\tau_\lambda\right\}.$$

If $\gamma > 1$ and $a(\gamma-1) > \gamma\sqrt{B+\sigma_\epsilon^2}$, then $R^*_{\mathrm{ridge}} > \sqrt{B+\sigma_\epsilon^2}$. Thus, in this regime, the optimally tuned ridge estimator is strictly suboptimal compared to $\widehat{\beta}_{\mathrm{MoM}}$.

On the other hand, if $0 < \gamma < 1$, then $R^*_{\mathrm{ridge}} \le \sigma_\epsilon/\sqrt{1-\gamma}$, i.e., an optimally tuned ridge estimator is always no worse than $\widehat{\beta}_{\mathrm{OLS}}$. Moreover, the inequality is strict if either $0 \le \rho < 1$, or if $\rho = 1$ and $a < \sigma_\epsilon/(1-\gamma)^{3/2}$. Consequently, if $B > \gamma\sigma_\epsilon^2/(1-\gamma)$, then the ordering in (3.1) *holds.*

Remark 2 (Ridgeless limit and double descent in the weak-signal regime). The weak-signal limit also captures the double-descent behavior of nearly ridgeless regression. To see this, recall $R_{\mathrm{ridge}}(\lambda)$, the ridge limiting constant in Theorem 3.3. Although Theorem 3.3 is stated for fixed $\lambda > 0$, the behavior as $\lambda \downarrow 0$ describes the nearly ridgeless ridge estimator. On the under-parameterized side, $\gamma < 1$, this limit recovers the OLS constant:

$$\lim_{\lambda\downarrow 0} R_{\mathrm{ridge}}(\lambda) = \sigma_\epsilon/\sqrt{1-\gamma}\,.$$

which diverges as $\gamma \uparrow 1$. On the over-parameterized side, $\gamma > 1$, the limit $\lambda \downarrow 0$ corresponds to the minimum-norm interpolator. The corresponding limiting constant is finite for every fixed $\gamma > 1$, but diverges as $\gamma \downarrow 1$ and then decreases away from the interpolation threshold before approaching a signal-dominated floor. Thus, in the weak-signal regime, the ℓ_∞ error along the ridgeless (or nearly ridgeless, i.e. $\lambda \approx 0$), path exhibits a double-descent profile as a function of γ: it increases toward the interpolation threshold and undergoes a second descent in the over-parameterized regime.

Proof technique. The proof of Theorem 3.3 is substantially more involved than the proof of Theorem 3.1. In the weak-signal regime, the signal and noise are of comparable magnitude in the ℓ_∞ scale, and therefore they must be analyzed simultaneously. Our proof relies on a delicate surgery using *leave-two-out approach* which helps to separate the signal and the noise contributions. This reduction is then combined with a careful analysis of the maximum of a correlated Gaussian vector. Although the current proof uses the Gaussianity of both X and ϵ, we expect that the result can be extended beyond the Gaussian setting using Cramer-type moderate deviation bounds together with universality arguments (e.g., Lindeberg swapping). Since the proof is already long and delicate even in the Gaussian case, we leave such universality extensions for future work.

Uniform confidence interval construction. As in the strong-signal regime, Theorem 3.3 also suggests first-order simultaneous intervals for the coordinates of β_0. The key difference is that, in the weak-signal regime, the optimally tuned ridge estimator may be preferable to both $\widehat{\beta}_{\mathrm{MoM}}$ and $\widehat{\beta}_{\mathrm{OLS}}$; for example, this occurs in the setting described in Theorem 3.4. When the optimally tuned ridge estimator is preferred, one can construct simultaneous confidence intervals of the form

$$\widehat{\mathrm{CI}}_j^{\mathrm{ridge}} = \left[\widehat{\beta}_{\lambda^*,j} - R^*_{\mathrm{ridge}}\sqrt{2\log p/n},\ \widehat{\beta}_{\lambda^*,j} + R^*_{\mathrm{ridge}}\sqrt{2\log p/n}\right], \qquad j = 1,\dots,p,$$

where λ^* denotes the optimal tuning parameter. On the other hand, if $\widehat{\beta}_{\mathrm{MoM}}$ or $\widehat{\beta}_{\mathrm{OLS}}$ is preferred to the optimally tuned ridge estimator, then one can use the corresponding confidence intervals described in Section 3.1.

3.3 Extension to random effect model

We also extend our analysis to a random-effects setting, in which the coordinates of β_0 are generated independently from a distribution μ on $\mathbb{R}$ and then scaled by $1/\sqrt{n}$, so that $\|\beta_0\|_2^2$ remains of constant order. Such random-effects signal assumptions are common in proportional asymptotic regimes; see, for example, [15, 7, 34, 2, 14] and the references therein. The resulting size of the coordinatewise maximum $\|\beta_0\|_\infty$ depends on the tail behavior of μ. For instance, if $\mu = \mathcal{N}(0,1)$, then $\|\beta_0\|_\infty \sim \sqrt{\log p/n}$, whereas for a heavy-tailed distribution such as $\mu = t_\alpha$, $\|\beta_0\|_\infty \sim n^{1/\alpha - 1/2}$. In the interest of space, we defer the full discussion of this random-effects formulation to Appendix A.

4 Numerical Experiments

In this section, numerical experiments are performed to empirically verify our theoretical claims. All experiments use i.i.d. Gaussian design matrices $X_{ij} \sim \mathcal{N}(0,1)$ and noise $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 1$. We consider both under-parameterized ($\gamma = 0.5$) and over-parameterized ($\gamma = 2$) regimes.

4.1 Strong Signal Regime

Setup. We first validate Theorem 3.1 by instantiating the strong-signal regime, setting $r_n = n^\alpha \sqrt{2\log p/n}$, $\alpha \in \{0.12, 0.14, \ldots, 0.20\}$. $\beta_{0,j} = \begin{cases} r_n, & j = 1, \\ 1/\sqrt{n}, & j = 2, \ldots, p. \end{cases}$ Thus $\|\beta_0\|_\infty = r_n$, $\frac{r_n}{\sqrt{2\log p/n}} = n^\alpha$, and $B = \|\beta_0\|_2^2 \approx \gamma$ remains bounded. We fix $n = 1000$, $\gamma = \{0.5, 2\}$ ($p = \{500, 2000\}$), $\sigma_\epsilon = 1$, $\lambda = 1$ for the ridge estimator, and run $n_{\mathrm{MC}} = 500$ replications per α.

Results. Figure 1 displays the raw, unnormalized ℓ_∞ error versus α. We report the scale separation predicted by Theorem 3.1: Null and Ridge errors grow with α along their respective r_n-scale theory curves, while MoM and OLS remain flat near $\sqrt{2\log p/n}$. The empirical means (solid lines with points) track theoretical predictions (dashed lines). Table 1 reports the ratio of empirical mean to theoretical prediction for each estimator and each α; ratios are close to 1, confirming Theorem 3.1.

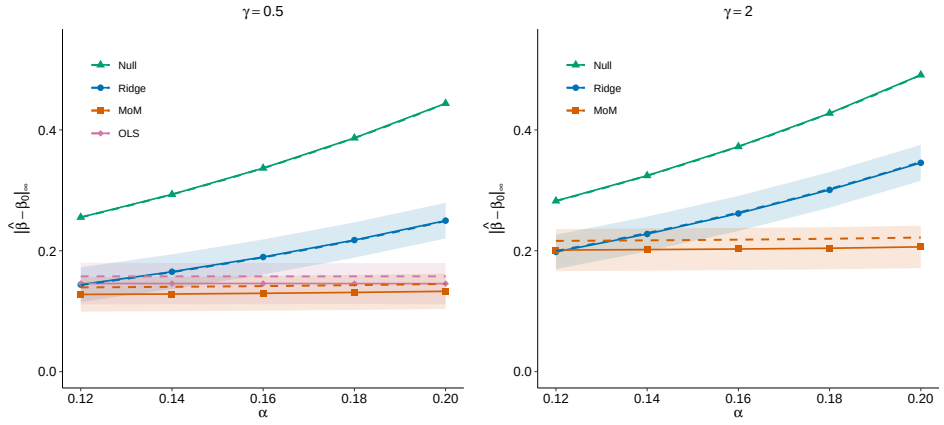


Figure 1: Verification of Theorem 3.1: raw $\|\hat{\beta} - \beta_0\|_\infty$ versus α, where $r_n = n^\alpha \sqrt{2\log p/n}$. Left: $\gamma = 0.5$, Right: $\gamma = 2$. Solid lines show Monte Carlo means; dashed lines show theoretical predictions (See ratio in Table 1; shaded bands show mean ± 2 empirical standard deviations.)

4.2 Weak-Signal Regime

We validate Theorem 3.3. In contrast to the strong-signal experiment, here all estimators operate at the scale $r_n = \sqrt{2\log p/n}$. We therefore plot the normalized error $\|\widehat{\beta} - \beta_0\|_\infty / r_n$ as a function of the ridge regularization parameter λ.

Setup. We fix $n = 1000$, $\sigma_\epsilon = 1$, and $\|\beta_0\|_2^2 = B = 1.2$. For a fixed boundary strength $a = 3.5$, we construct β_0 by setting $\beta_{0,1} = ar_n$, and distributing the remaining ℓ_2 mass uniformly across the other coordinates: $\beta_{0,j} = \sqrt{\frac{B-a^2r_n^2}{p-1}}$, $j = 2, \ldots, p$. This construction satisfies Assumption 3 with $S_n = \{1\}$, $s_n = 1$, and $\rho = 0$. Indeed, $\frac{|\beta_{0,1}|}{r_n} = a$, while $\max_{j \notin S_n} \frac{|\beta_{0,j}|}{r_n} = \sqrt{\frac{n(B-a^2r_n^2)}{2(p-1)\log p}} \to 0$ whenever $p/n \to \gamma \in (0, \infty)$. We consider $\gamma = 0.5$ and $\gamma = 2$, corresponding to $p = 500$ and $p = 2000$, respectively. We run $n_{\mathrm{MC}} = 500$ Monte Carlo replications for each λ. We evaluate the ridge estimator over a grid of 12 regularization parameters λ log-spaced from 10^{-3} to 10^2.

Results. Figure 2 plots $\frac{\|\widehat{\beta}-\beta_0\|_\infty}{r_n}$ against λ. The ridge curve varies with λ, while the null, method-of-moments, and OLS estimators appear horizontal. The empirical means (solid lines with points) track all theoretical predictions (dashed lines), confirming Theorem 3.3.

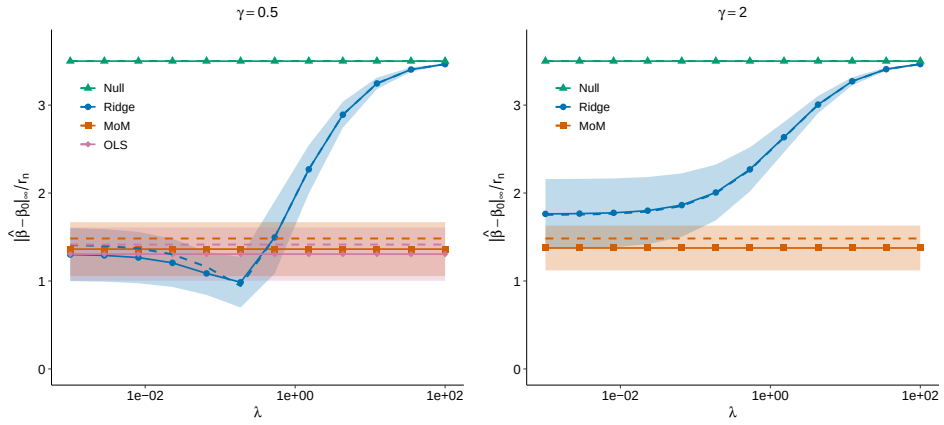


Figure 2: Verification of Theorem 3.3: normalized error $\|\hat{\beta} - \beta_0\|_\infty / r_n$ versus regularization parameter λ. Left: $\gamma = 0.5$, Right: $\gamma = 2$. Solid lines show Monte Carlo means; dashed lines show theoretical predictions (See ratio in Table 2; shaded bands show mean ± 2 empirical standard deviations.) $n = 1000$, $a = 3.5$, $B = 1.2$, $\sigma_\epsilon = 1$, $n_{\mathrm{MC}} = 500$.

4.3 Double Descent in the Weak-Signal Regime

We illustrate the double-descent implication of the weak-signal theory as discussed in Remark 2.

Setup. We use the same weak-signal construction as in Section 4.2. We fix $n = 1000$, $\sigma_\epsilon = 1$, $\|\beta_0\|_2^2 = B = 0.5$, and $a = 2$. For each value of γ, we set $p = \lfloor \gamma n \rfloor$ and define $r_n = \sqrt{2 \log p / n}$. The signal vector is given by $\beta_{0,1} = ar_n, \beta_{0,j} = \sqrt{\frac{B-a^2r_n^2}{p-1}}$, $j = 2, \ldots, p$. This satisfies Assumption 3 with $S_n = \{1\}$ and $\rho = 0$. We fix the ridge regularization parameter at $\lambda = 10^{-5}$ and vary the aspect ratio $\gamma = p/n$ over 20 values, with a finer grid near the interpolation threshold $\gamma = 1$. For each value of γ, we run 500 Monte Carlo replications.

5 Conclusion

In this paper, we present sharp asymptotic analyses of $\|\widehat{\beta}_\lambda - \beta_0\|_\infty$ for λ-ridge regularized least squares estimator $\widehat{\beta}_\lambda$ under proportional asymptotics under different regimes of signal structures for β_0 in a linear regression model through the lens of ridge regularized methods and competing unbiased methods based on moments and ordinary least squares (for under-parametrized regimes). Our explorations are motivated not only by complementing the extensive results available for ℓ_q-norm of error asymptotic in literature for $q \in [2, \infty)$ but also by trying to understand the impact of signal structure on the behavior of specific estimators under study. It turns out that the behavior of $\|\widehat{\beta}_\lambda - \beta_0\|_\infty$ depends subtly on the specific structure of β_0 and yields not only differential performance of estimators but also yields a new class of double descent curves in the ridgeless least squares estimator.

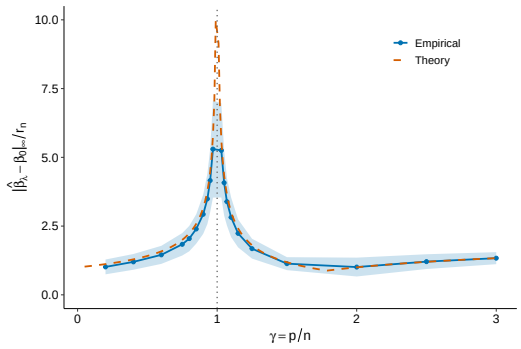


Figure 3: Double descent in the weak-signal regime: normalized error $\|\widehat{\beta}_\lambda - \beta_0\|_\infty / r_n$ versus aspect ratio $\gamma = p/n$ for nearly ridgeless regression ($\lambda = 10^{-5}$). Solid: Monte Carlo mean; dashed: theoretical prediction from Theorem 3.3. See the Ratio column in Table 3.

Finally, as potential limitations of the current paper, it remains to explore precise downstream inferential consequences such as variable selection and FDR control based on our uniform limit theorems. Moreover, we have only considered the well-specified linear model in this paper. The implications of these results and analyses in neural network models in proportional scaling regimes [27, 26] also constitute interesting future scopes.

Acknowledgments and Disclosure of Funding

The authors have no funding to disclose. The authors declare no competing interests.

References

[1] Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Special invited lecture: Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[2] Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu. When does preconditioning help or hurt generalization? In *International Conference on Learning Representations*, 2021.

[3] Zhi-Dong Bai and Jack W Silverstein. 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.

[4] Zhi-Dong Bai and Yong-Qua Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

[5] Zhidong Bai and Jack W Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer, 2010.

[6] Shankar Bhamidi, Partha S Dey, and Andrew B Nobel. Energy landscape for large average submatrix detection problems in Gaussian random matrices. *Probability Theory and Related Fields*, 168(3):919–983, 2017.

[7] Jérémie Bigot, Issa-Mbenard Dabo, and Camille Male. High-dimensional analysis of ridge regression for non-identically distributed data with a variance profile. *arXiv preprint arXiv:2403.20200*, 2024.

[8] Marek Biskup. Extrema of the two-dimensional discrete Gaussian free field. In *PIMS-CRM Summer School in Probability*, pages 163–407. Springer, 2017.

[9] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

[10] Samprit Chatterjee and Jeffrey S Simonoff. *Handbook of Regression Analysis*. John Wiley & Sons, 2013.

[11] Xingyu Chen, Lin Liu, and Rajarshi Mukherjee. Method-of-moments inference for GLMs and doubly robust functionals under proportional asymptotics. *arXiv preprint arXiv:2408.06103*, 2024.

[12] Clément Chouard. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv preprint arXiv:2211.13044*, 2022.

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

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

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

[16] Noureddine El Karoui. Random matrices and high-dimensional statistics: Beyond covariance matrices. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 2857–2876. World Scientific, 2018.

[17] Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15 (4):335–536, 2022.

[18] Boris Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453, 1943.

[19] Jürgen Groß. *Linear Regression*, volume 175. Springer Science & Business Media, 2003.

[20] Walid Hachem, Philippe Loubaton, Jamal Najim, and Pascal Vallet. On bilinear forms based on the resolvent of large random matrices. *Annales de l'IHP Probabilités et Statistiques*, 49(1): 36–63, 2013.

[21] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

[22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[23] Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253, 2009.

[24] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, 2017.

[25] Malcolm R Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.

[26] Song Mei and Andrea Montanari. 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.

[27] Theodor Misiakiewicz and Andrea Montanari. Six lectures on linearized neural networks. *arXiv preprint arXiv:2308.13431*, 2023.

[28] Andrea Montanari. Mean field asymptotics in high-dimensional statistics: From exact results to efficient algorithms. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 2973–2994. World Scientific, 2018.

[29] Sidney I Resnick. *Extreme Values, Regular Variation, and Point Processes*, volume 4. Springer Science & Business Media, 2008.

[30] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[31] Vadim Ivanovich Serdobolskii. *Multiparametric Statistics*. Elsevier, 2007.

[32] Weijie Su and Emmanuel Candes. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

[33] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[34] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10112–10123, 2020.

[35] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

A Random Effect Model

In the previous section, we analyzed the estimation error in the *strong-signal regime*, namely when $\|\beta_0\|_\infty \gg \sqrt{\log p/n}$. In this section, we turn to the complementary setting where $\|\beta_0\|_\infty = O(\sqrt{\log p/n})$. In this regime, the behavior of the estimation error changes markedly, as the noise effects become non-negligible, rendering the analysis substantially more delicate. A canonical example of such a signal arises when the coordinates of β_0 are drawn independently from a suitably scaled sub-Gaussian distribution so as to maintain a bounded signal-to-noise ratio. Our first result in this direction characterizes the resulting change in the rate of convergence relative to Theorem 3.3.

Theorem A.1 (Sub-Gaussian)**.** *Consider the linear regression problem (Assumption 1) with X_{ij} being iid centered subgaussian random variables with variance 1 and ϵ_i's are i.i.d. Gaussian random variable with mean 0 and variance σ_ϵ^2. Assume that $\beta_0 = \boldsymbol{\mu}/\sqrt{n}$, where $\boldsymbol{\mu}_j$'s are i.i.d. sub-Gaussian random variables with sub-Gaussian constant being σ_β^2. Then, for any $\lambda > 0$, the ridge estimator $\widehat{\beta}_\lambda$ satisfies:*

$$\sqrt{n/\log p}\,\|\widehat{\beta}_\lambda - \beta_0\|_\infty = O_p(1)\,.$$

The above theorem establishes that when the coordinates of β_0 are generated from the sub-Gaussian distribution, then the overall estimation error scales as $\sqrt{\log p/n}$. In fact, both the signal part and noise part $\widehat{\beta}_\lambda - \beta_0$ scales at this rate. Whereas the above result helps us understand the rate of convergence, isotropic Gaussian random effects allows derivation of sharp results and demonstrate a double descent phenomenon for the ridgeless least squares. We collect these results in our next theorem.

Theorem A.2 (Gaussian)**.** *Consider the same setup as above, but assume that $\beta_0 = \mathbf{z}/\sqrt{n}$, where $\mathbf{z} \sim \mathcal{N}(0, \sigma_\beta^2 \mathbf{I}_p)$. Then for any $\lambda > 0$, the ridge estimator $\widehat{\beta}_\lambda$ satisfies:*

$$\sqrt{\tfrac{n}{2\log p}}\|\widehat{\beta}_\lambda - \beta_0\|_\infty \xrightarrow{p} \Psi_\gamma(\lambda; \sigma_\beta^2; \sigma_\epsilon^2)\,,$$

where $\Psi_\gamma(\lambda; \sigma_\beta^2; \sigma_\epsilon^2)$ is defined as:

$$\Psi_\gamma(\lambda; \sigma_\beta^2; \sigma_\epsilon^2) = \sigma_\beta^2 \lambda^2 m_\gamma'(-\lambda) + \sigma_\epsilon^2 \left(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda)\right).$$

Furthermore, for $\gamma > 1$, the min-norm interpolator $\widehat{\beta}$ admits the following limit:

$$\sqrt{\tfrac{n}{2\log p}}\|\widehat{\beta}_\lambda - \beta_0\|_\infty \xrightarrow{p} \Psi_\gamma(0; \sigma_\beta^2; \sigma_\epsilon^2)\,,$$

where $\Psi_\gamma(0; \sigma_\beta^2; \sigma_\epsilon^2) = \lim_{\lambda \downarrow 0} \Psi_\gamma(\lambda; \sigma_\beta^2; \sigma_\epsilon^2)$ is defined as:

$$\Psi_\gamma(0; \sigma_\beta^2; \sigma_\epsilon^2) = \sigma_\beta^2 \left(1 - \tfrac{1}{\gamma}\right) + \sigma_\epsilon^2 \tfrac{1}{\gamma(\gamma - 1)}\,.$$

On the other hand, for $\gamma < 1$, the OLS estimator $\widehat{\beta}$ satisfies:

$$\sqrt{\tfrac{n}{2\log p}}\|\widehat{\beta}_\lambda - \beta_0\|_\infty \xrightarrow{p} \Psi_\gamma(0;\sigma_\beta^2;\sigma_\epsilon^2)\,,$$

where $\Psi_\gamma(0;\sigma_\beta^2;\sigma_\epsilon^2) = \sigma_\epsilon^2(1-\gamma)^{-1}$.

Remark 3 (Double descent of $\Psi_\gamma(0)$). Theorem A.2 also reveals an interesting *double-descent* phenomenon in the ℓ_∞ estimation error. When $\gamma < 1$, the ordinary least squares estimator exists and its scaled limiting estimation error with respect to ℓ_∞ norm is driven purely by noise as $\Psi_\gamma(0;\sigma_\beta^2;\sigma_\epsilon^2)$ only depends on σ_ϵ^2 (noted in the previous theorem). This limiting risk which approaches to σ_ϵ^2 as $\gamma \downarrow 0$ and diverges as $\gamma \uparrow 1$ at rate $(1-\gamma)^{-1}$. On the other hand, for $\gamma > 1$, the min-norm interpolator exists and its limiting risk, relies on both signal and noise, i.e., both on $(\sigma_\beta^2, \sigma_\epsilon^2)$. The limiting risk of the min-norm interpolator decreases as γ increases; in particular, $\gamma \uparrow \infty$, the risk approaches σ_β^2. Notably, this limiting value is *strictly smaller* than the corresponding limit as $\gamma \to 0$ as soon as $\sigma_\beta^2 < \sigma_\epsilon^2$. This contrasts sharply with the classical ℓ_2 risk, for which the minimum risk in the overparameterized regime is never below the minimum risk in the underparametrized regimes (which is 0). The result highlights that, for estimation with respect to ℓ_∞ norm, extreme overparameterization can be beneficial, leading to a second descent in the ℓ_∞ error beyond the interpolation threshold provided that signal variance is smaller than noise variance.

So far, our analysis has focused on light-tailed random effects, such as Gaussian or sub-Gaussian signals. In many modern applications, however, the coordinates of β_0 may exhibit *heavy-tailed* behavior, arising for instance from Student-t, Pareto, or more generally power-law–distributed effects. Such distributions naturally arise in high-dimensional scientific data, finance, genomics, and networked systems. In these settings, classical concentration arguments break down, and the ℓ_∞ error is no longer governed by Gaussian-type fluctuations. Instead, extreme-value effects dominate, suggesting that a proper asymptotic description should be formulated through the lens of *extreme-value theory*. In particular, a broad and canonical class of heavy-tailed distributions, including Student-t and Pareto laws, falls within the domain of attraction of the Fréchet distribution. This observation motivates the following result, which characterizes the limiting behavior of the coordinatewise estimation error when the signal exhibits heavy tails.

Theorem A.3 (Heavy-tailed)**.** *Let $\Phi_\alpha(x)$ denotes the Fréchet distribution with parameter α for some $\alpha > 2$. Suppose $\beta_0 = \sigma_\beta \boldsymbol{\mu}/\sqrt{n}$ where $\mu_i \overset{\text{i.i.d.}}{\sim} F$ such that F belongs to the domain of attraction of Φ_α. Then, for any $\lambda > 0$, the ridge estimator $\widehat{\beta}_\lambda$ satisfies:*

$$\mathbb{P}\left(\tfrac{n^{\frac{1}{2}-\frac{1}{\alpha}}}{\ell(n)}\,\|\widehat{\beta}_\lambda - \beta_0\|_\infty \le t\right) \to e^{-\gamma\left(\frac{t}{\lambda m_\gamma(-\lambda)\sigma_\beta}\right)^{-\alpha}}\,,$$

for some slowly varying sequence $\{\ell(n)\}_{n\in\mathbb{N}}$.

A brief discussion of slowly and regularly varying functions can be found in Appendix H. This theorem complements the results for fixed effects and sub-Gaussian random effects, yielding yet another rate of convergence and a different asymptotic limit for the ℓ_∞-norm of the estimation error. As expected, the signal dominates the effect of the noise here, so that the limiting distribution only depends on σ_β^2.

Theorem A.3 applies to a wide class of heavy-tailed random effects and admits transparent interpretations in several canonical examples. First, if $\mu_i \sim t_\alpha$ follows a Student-t distribution with $\alpha > 2$ degrees of freedom, then F belongs to the Fréchet domain of attraction with tail index α, and the associated slowly varying function satisfies $\ell(n) \to c_\alpha$ for some finite constant $c_\alpha > 0$. Second, if μ_i follows a Pareto distribution with tail index $\alpha > 2$, that is, $\mathbb{P}(|\mu_i| > x) \asymp x^{-\alpha}$, then $\ell(n)$ converges to a constant as well, and the same Fréchet limit arises. These examples illustrate that, across a broad class of heavy-tailed signals, the asymptotic behavior of $\|\widehat{\beta}_\lambda - \beta_0\|_\infty$ is universal and dictated by extreme-value effects rather than Gaussian-type fluctuations. We illustrate this through Figure 5 in Section 4 where we also report the precise expression of c_α.

B Proof of Theorem 3.1

B.1 Proof of Ridge estimator

Proof. Recall the definition of $\widehat{\beta}_\lambda$:

$$\widehat{\beta}_\lambda = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \mathbf{y}}{n} = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} \beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \,.$$

This immediately implies:

$$\begin{aligned}
\widehat{\beta}_{\lambda,j} - \beta_{0,j} &= -\lambda e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \beta_0 + e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \\
&= \underbrace{-\lambda \left(e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \beta_0 - m(-\lambda) \beta_{0,j} \right)}_{:= W_{n,j}} - \lambda m(-\lambda) \beta_{0,j} + \underbrace{e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}}_{:= V_{n,j}} \\
&:= -\lambda m(-\lambda) \beta_{0,j} + W_{n,j} + V_{n,j}
\end{aligned}$$

To bound $W_{n,j}$ we use Proposition 2.4 of [12], which yields for $1 \le j \le p$, almost-surely we have:

$$|W_{n,j}| \le C_\lambda \|\beta_0\|_2 \sqrt{\frac{\log p}{n}} \implies \|W_n\|_\infty \le C_\lambda \|\beta_0\|_2 \sqrt{\frac{\log p}{n}} \,.$$

The last inequality follows from the fact that C depends on λ, but does not depend on j. Now for $V_{n,j}$, we use the tail bound on Gaussian random variables. Note that:

$$V_n \mid \mathbf{X} = e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \mid \mathbf{X} \sim \frac{1}{\sqrt{n}} \mathcal{N}\left(0, (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\right)$$

The covariance matrix of $V_n \mid \mathbf{X}$ can be rewritten as:

$$(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - \lambda (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} \,.$$

Now, from anisotropic local law (Theorem 3.17 of [24]), we have that for any small $\alpha_0, \alpha_1 > 0$ and large $D > 0$, there exists $N_0 \equiv N_0(\alpha_0, \alpha_1, D)$ such that the following event:

$$\Omega_n = \left\{ \mathbf{X} : \left| e_i^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \beta_0 - m_\gamma(-\lambda) \delta_{ij} \right| \le \frac{1}{n^{\frac{1}{2} - \alpha_0} \lambda}, \quad \forall\, 1 \le i, j \le p \right\}$$

holds with probability $\ge 1 - p^2 n^{-D} = 1 - \gamma^2 n^{D-2}$ for all $n \ge N_0$ and $\lambda \ge n^{-\frac{2}{3} + \alpha_1}$. Here we have implicitly used a union bound over $1 \le i, j \le p$. The above conclusion of can be extended to $-\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2}$ by observing that:

$$-\lambda u^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} v = \lambda \partial_\lambda u^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} v$$

Using the same technique employed in [22] (e.g., see Lemma 5), we can conclude that the following event $\Omega_{n,1}$ holds with probability $\ge 1 - p^2 n^{-D} = 1 - \gamma^2 n^{D-2}$:

$$\Omega_{n,1} = \left\{ \mathbf{X} : \lambda |e_i^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} e_j - m'(-\lambda) \delta_{ij}| \le \frac{1}{n^{\frac{1}{2} - \alpha_0}} + \frac{C\lambda}{n}, \quad \forall\, i, j \right\}$$

for some constant $C > 0$. Therefore, on the event $\Omega_{n,2} = \Omega_n \cap \Omega_{n,1}$, we have:

$$|e_i^\top (-\lambda (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2}) e_j + \lambda m'(-\lambda) \delta_{ij}| \le \frac{1}{n^{\frac{1}{2} - \alpha_0} \lambda} + \frac{1}{n^{\frac{1}{2} - \alpha_0}} + \frac{C\lambda}{n} := a(n, \lambda) \quad \forall\, i, j \,.$$

For notational simplicity, let us further define $\Psi(\lambda) = m(-\lambda) - \lambda m'(-\lambda)$ and $\widehat{\sigma}_j^2 = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}_{jj} - \lambda (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2}_{jj}$. On $\Omega_{n,2}$, we have $\widehat{\sigma}_j^2 \le \Psi(\lambda) + a(n, \lambda)$ for all $1 \le j \le p$. We have:

$$\mathbb{P}\left(\|V_n\|_\infty \ge \sqrt{\frac{2(1+\delta)\Psi(\lambda) \log p}{n}} \right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}}, \Omega_n\right) + \mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\log p}{n}}, \Omega_n^c\right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}} \mid \mathbf{X}\right)\mathbb{1}_{\mathbf{X}\in\Omega_n}\right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p \mathbb{P}\left(|V_{n,j}| \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}} \mid \mathbf{X}\right)\mathbb{1}_{\mathbf{X}\in\Omega_n}\right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p \mathbb{P}\left(|Z| \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{\widehat{\sigma}_j^2}} \mid \mathbf{X}\right)\mathbb{1}_{\mathbf{X}\in\Omega_n}\right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p e^{-\frac{(1+\delta)\Psi(\lambda)\log p}{\widehat{\sigma}_j^2}}\mathbb{1}_{\mathbf{X}\in\Omega_n}\right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p e^{-\frac{(1+\delta)\Psi(\lambda)\log p}{\Psi(\lambda)+n^{-\frac{1}{2}+\alpha}\lambda^{-\frac{1}{4}}}}\mathbb{1}_{\mathbf{X}\in\Omega_n}\right] + \mathbb{P}(\Omega_n^c) \\
&\leq e^{\log p - \frac{(1+\delta)\Psi(\lambda)\log p}{\Psi(\lambda)+n^{-\frac{1}{2}+\alpha}\lambda^{-\frac{1}{4}}}} \longrightarrow 0\,.
\end{aligned}
$$

As a consequence, we can conclude $\|V_n\|_\infty = O_p(\sqrt{\log p/n})$. Therefore, using the fact that $r_n = o(\sqrt{n/\log p})$

$$
\frac{1}{r_n}\left|\|\widehat{\beta}_\lambda - \beta_0\|_\infty - \lambda m(-\lambda)\|\beta_0\|_\infty\right| \leq \frac{1}{r_n}(\|W_n\|_\infty + \|V_n\|_\infty) = o_p(1)\,.
$$

Finally, using the fact that $\|\beta_0\|_\infty/r_n \to c_*$, we conclude that:

$$
\frac{\|\widehat{\beta}_\lambda - \beta_0\|_\infty}{r_n} \xrightarrow{p} \lambda m(-\lambda)c_*.
$$

□

B.2 Proof for null estimator

As $\widehat{\beta}_{\text{null}} \equiv 0$, we have from Assumption 2,

$$
\frac{\|\widehat{\beta}_{\text{null}} - \beta_0\|_\infty}{r_n} = \frac{\|\beta_0\|_\infty}{r_n} \to c_*\,.
$$

B.3 Proof of MoM estimator

Proof. For notational simplicity, we use $\widehat{\beta}_{\text{MoM}} = \widehat{\beta}$ in this subsection as there will be no confusion. We work under the same data generating process and assumptions as before, but now we consider the estimator:

$$
\widehat{\beta} = \frac{1}{n}\sum_{i=1}^n X_i Y_i = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \beta_0 + \frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i = \widehat{\Sigma}\beta_0 + \frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i\,.
$$

Therefore

$$
\widehat{\beta} - \beta_0 = (\widehat{\Sigma} - \mathbf{I}_p)\beta_0 + \frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i.
$$

Our goal is to prove

$$
\frac{\|\widehat{\beta} - \beta_0\|_\infty}{\sqrt{2\log p/n}} \xrightarrow{p} \sqrt{B + \sigma_\varepsilon^2}.
$$

Set $a = \beta_0/\sqrt{B}$ and $b = \sqrt{B}$ for notational simplicity, assuming $b > 0$. Now, for each i, decompose X_i as

$$q_i = a^\top X_i, \qquad X_i = q_i a + \xi_i,$$

where

$$q_i \sim N(0,1), \qquad \xi_i \sim N(0, \mathbf{I}_p - aa^\top), \qquad q_i \perp\!\!\!\perp \xi_i.$$

Since

$$Y_i = X_i^\top \beta_0 + \varepsilon_i = bq_i + \varepsilon_i,$$

we have

$$X_i Y_i - \beta_0 = (q_i a + \xi_i)(bq_i + \varepsilon_i) - ba.$$

Expanding the right-hand side gives

$$X_i Y_i - \beta_0 = a\{b(q_i^2 - 1) + q_i\varepsilon_i\} + (bq_i + \varepsilon_i)\xi_i.$$

Therefore

$$\widehat{\beta} - \beta_0 = a\left\{b\frac{1}{n}\sum_{i=1}^n (q_i^2 - 1) + \frac{1}{n}\sum_{i=1}^n q_i\varepsilon_i\right\} + \frac{1}{n}\sum_{i=1}^n (bq_i + \varepsilon_i)\xi_i.$$

The first term is a rank-one, common-direction term. Since

$$b\frac{1}{n}\sum_{i=1}^n (q_i^2 - 1) + \frac{1}{n}\sum_{i=1}^n q_i\varepsilon_i = O_p(n^{-1/2}),$$

and $\|a\|_\infty \le b$, we have

$$\frac{\left\|a\left\{bn^{-1}\sum_{i=1}^n (q_i^2 - 1) + n^{-1}\sum_{i=1}^n q_i\varepsilon_i\right\}\right\|_\infty}{\sqrt{2\log p/n}} = o_p(1).$$

Therefore,

$$\frac{\|\widehat{\beta} - \beta_0\|_\infty}{\sqrt{2\log p/n}} = \frac{\left\|\frac{1}{n}\sum_{i=1}^n (bq_i + \varepsilon_i)\xi_i\right\|_\infty}{\sqrt{2\log p/n}} + o_p(1) := \frac{\|R_n\|_\infty}{\sqrt{2\log p/n}} + o_p(1)\,.$$

It remains to analyze

$$R_n := \frac{1}{n}\sum_{i=1}^n (bq_i + \varepsilon_i)\xi_i.$$

Conditional on $\{q_i, \varepsilon_i\}_{i=1}^n$, this term is Gaussian with mean zero and covariance

$$\mathsf{Cov}\left(R_n \mid \{q_i, \varepsilon_i\}_{i=1}^n\right) = \frac{S_n}{n}(\mathbf{I}_p - aa^\top), \quad S_n := \frac{1}{n}\sum_{i=1}^n (bq_i + \varepsilon_i)^2.$$

Equivalently, conditionally on $\{q_i, \varepsilon_i\}_{i=1}^n$, we may write

$$R_n \overset{d}{=} \sqrt{\frac{S_n}{n}}\,(\mathbf{I}_p - aa^\top)Z, \qquad Z \sim N(0, \mathbf{I}_p).$$

By the law of large numbers,

$$S_n \xrightarrow{p} B + \sigma_\varepsilon^2, \quad \text{and} \quad S_n = (B + \sigma_\epsilon^2) + O_p(n^{-1/2})\,.$$

Hence, we can write:

$$\sqrt{\tfrac{S_n}{n}}\,(\mathbf{I}_p - aa^\top)Z = \sqrt{\tfrac{B+\sigma_\epsilon^2}{n}}\,(\mathbf{I}_p - aa^\top)Z + \left(\sqrt{\tfrac{S_n}{n}} - \sqrt{\tfrac{B+\sigma_\epsilon^2}{n}}\right)(\mathbf{I}_p - aa^\top)Z \tag{B.1}$$

Let us first consider the first term:

$$\frac{\|(\mathbf{I}_p - aa^\top)Z\|_\infty}{\sqrt{2\log p}} = \frac{\|Z\|_\infty}{\sqrt{2\log p}} + \frac{\|(\mathbf{I}_p - aa^\top)Z\|_\infty - \|Z\|_\infty}{\sqrt{2\log p}}$$

Now for the second term on the RHS of the above equation:

$$\frac{|\|(\mathbf{I}_p - aa^\top)Z\|_\infty - \|Z\|_\infty|}{\sqrt{2\log p}} \le \frac{\|a\|_\infty |a^\top Z|}{\sqrt{2\log p}} = O_p(1/\sqrt{\log p}) = o_p(1)\,.$$

Therefore,

$$\frac{\|(\mathbf{I}_p - aa^\top)Z\|_\infty}{\sqrt{2\log p}} \xrightarrow{p} 1\,.$$

Now, going back to the second term on the RHS of Equation (B.1), we have:

$$\begin{aligned}
&\frac{1}{\sqrt{2\log p/n}} \left\| \left(\sqrt{\frac{S_n}{n}} - \sqrt{\frac{B+\sigma_\epsilon^2}{n}} \right) (\mathbf{I}_p - aa^\top)Z \right\|_\infty \\
&= \frac{1}{\sqrt{2\log p}} \left\| (\sqrt{S_n} - \sqrt{B+\sigma_\epsilon^2})(\mathbf{I}_p - aa^\top)Z \right\|_\infty \\
&\leq \frac{\|(\mathbf{I}_p - aa^\top)Z\|_\infty}{\sqrt{2\log p}} \frac{|S_n - (B+\sigma_\epsilon^2)|}{\sqrt{S_n} + \sqrt{B+\sigma_\epsilon^2}} \\
&\leq \frac{\|(\mathbf{I}_p - aa^\top)Z\|_\infty}{\sqrt{2\log p}} \frac{|S_n - (B+\sigma_\epsilon^2)|}{\sqrt{B+\sigma_\epsilon^2}} \\
&= O_p(n^{-1/2}) = o_p(1)\,.
\end{aligned}$$

Therefore,

$$\frac{\|\widehat{\beta} - \beta_0\|_\infty}{\sqrt{2\log p/n}} \xrightarrow{p} \sqrt{B+\sigma_\varepsilon^2}.$$

□

B.4 Proof of OLS estimator for $\gamma < 1$

Proof. For notational simplicity, we use $\widehat{\beta}_{\mathrm{OLS}} = \widehat{\beta}$ in this subsection as there will be no confusion. Since $\mathbf{y} = \mathbf{X}\beta_0 + \boldsymbol{\epsilon}$, we have

$$\widehat{\beta} - \beta_0 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon} \implies \widehat{\beta} - \beta_0 \mid \mathbf{X} \sim N\left(0, \sigma_\epsilon^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

We now use the exact marginal distribution of the OLS error under Gaussian design. Since X_i's are i.i.d. $\mathcal{N}(0, \mathbf{I}_p)$, we have

$$\mathbf{X}^\top\mathbf{X} \sim W_p(n, \mathbf{I}_p),$$

a p-dimensional Wishart random matrix with n degrees of freedom. Conditional on $\mathbf{X}^\top\mathbf{X} = W$,

$$\widehat{\beta} - \beta_0 \mid W \sim N(0, \sigma_\varepsilon^2 W^{-1}).$$

Next, we show that, integrating out W,we obtain the standard multivariate t-representation of $\widehat{\beta} - \beta_0$, i.e.,

$$\widehat{\beta} - \beta_0 \overset{d}{=} \frac{\sigma_\varepsilon}{\sqrt{Q_\nu}} Z, \qquad Z \sim N(0, \mathbf{I}_p), \qquad Q_\nu \sim \chi^2_\nu,$$

where $\nu = n - p + 1$, and Z and Q_ν are independent. To see why, recall that the density of $W = \mathbf{X}^\top\mathbf{X}$ is proportional to

$$f_{W_p(n,\mathbf{I}_p)}(W) \propto |W|^{(n-p-1)/2} \mathsf{exp}\left\{ -\frac{1}{2}\mathsf{tr}(W) \right\}.$$

Now, given W, the density of $u := \widehat{\beta} - \beta_0$ is proportional to

$$f_{\widehat{\beta}-\beta_0 | W=w}(u) \propto |W|^{1/2} \mathsf{exp}\left\{ -\frac{1}{2\sigma_\varepsilon^2} u^\top W u \right\}.$$

Therefore, the marginal density $\widehat{\beta} - \beta_0$ is proportional to

$$f_{\widehat{\beta}-\beta_0}(u) \propto \int_{W>0} |W|^{(n-p)/2} \mathsf{exp}\left\{ -\frac{1}{2}\mathsf{tr}\left[W\left(\mathbf{I}_p + \frac{uu^\top}{\sigma_\varepsilon^2} \right) \right] \right\} dW.$$

Using the Wishart integral,

$$\int_{W>0} |W|^{(m-p-1)/2} e^{-\frac{1}{2}\mathsf{tr}(AW)} dW \propto |A|^{-m/2},$$

with $m = n + 1$, we get

$$f_{\widehat{\beta}-\beta_0}(u) \propto \left| \mathbf{I}_p + \frac{uu^\top}{\sigma_\varepsilon^2} \right|^{-(n+1)/2}.$$

By the matrix determinant lemma,

$$\left| \mathbf{I}_p + \frac{uu^\top}{\sigma_\varepsilon^2} \right| = 1 + \frac{\|u\|_2^2}{\sigma_\varepsilon^2}.$$

Hence

$$f_{\widehat{\beta}-\beta_0}(u) \propto \left(1 + \frac{\|u\|_2^2}{\sigma_\varepsilon^2} \right)^{-(n+1)/2}.$$

This is the density of a multivariate t-random vector with degrees of freedom $\nu = n - p + 1$, and scale matrix $(\sigma_\epsilon^2/\nu)\mathbf{I}_p$. Equivalently, marginally,

$$\widehat{\beta} - \beta_0 \overset{d}{=} \frac{\sigma_\varepsilon}{\sqrt{Q_\nu}} Z, \qquad Z \sim N(0, \mathbf{I}_p), \qquad Q_\nu \sim \chi^2_\nu,$$

with $Z \perp\!\!\!\perp Q_\nu$. Therefore,

$$\|\widehat{\beta} - \beta_0\|_\infty \overset{d}{=} \frac{\sigma_\varepsilon}{\sqrt{Q_\nu}} \|Z\|_\infty.$$

Thus

$$\frac{\|\widehat{\beta} - \beta_0\|_\infty}{\sqrt{2 \log p / n}} \overset{d}{=} \sigma_\varepsilon \left(\frac{\|Z\|_\infty}{\sqrt{2 \log p}} \right) \left(\sqrt{\frac{n}{Q_\nu}} \right).$$

We now analyze the two factors separately. First, since $Z_1, \ldots, Z_p \overset{\text{iid}}{\sim} N(0, 1)$, the standard Gaussian maximum satisfies

$$\frac{\|Z\|_\infty}{\sqrt{2 \log p}} \xrightarrow{\text{p}} 1.$$

Second, since $Q_\nu \sim \chi^2_\nu$, we have $Q_\nu / \nu \to 1$ in probability. Moreover, $\nu = n - p + 1$, and because $p/n \to \gamma \in (0, 1)$,

$$\frac{\nu}{n} = 1 - \frac{p}{n} + \frac{1}{n} \to 1 - \gamma.$$

Hence

$$\frac{Q_\nu}{n} = \frac{Q_\nu}{\nu} \frac{\nu}{n} \xrightarrow{p} 1 - \gamma,$$

and therefore

$$\sqrt{\frac{n}{Q_\nu}} \xrightarrow{p} \frac{1}{\sqrt{1 - \gamma}}.$$

Combining the two limits and using Slutsky's theorem gives

$$\frac{\|\widehat{\beta} - \beta_0\|_\infty}{\sqrt{2 \log p / n}} \xrightarrow{p} \sigma_\epsilon \cdot 1 \cdot \frac{1}{\sqrt{1 - \gamma}} = \frac{\sigma_\varepsilon}{\sqrt{1 - \gamma}}.$$

This completes the proof.

□

C Proof of Corollary 3.2

Proof. Here, we assume $X_i = Z_i \Sigma^{1/2}$ where Z_{ij} are iid from some distribution with mean 0, variance 1 and sub-Gaussian tails. As before, define an event Ω_n as:

$$\Omega_n = \left\{ \mathbf{X} : \left| \lambda e_i^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \beta_0 - e_i^\top \left(\mathbf{I}_p + \tilde{m}_n(-\lambda) \Sigma \right)^{-1} \beta_0 \right| \leq \frac{1}{n^{\frac{1}{2} - \alpha_0} \lambda} \quad \forall\, 1 \leq i \leq p \right\}$$

for all $\lambda \geq n^{-\frac{2}{3} + \alpha_1}$. Here $\tilde{m}_n(z)$ is the Stieltjes transform of the companion of the Marchenko-Pasteur law, defined as the solution of the following fixed-point equation:

$$\frac{1}{\tilde{m}_n(z)} = -z + \gamma \frac{1}{p} \sum_{j=1}^{p} \frac{s_i(\Sigma)}{1 + s_i(\Sigma) \tilde{m}_n(z)} = -z + \gamma \int \frac{x}{1 + x \tilde{m}_n(z)} \, dF_{\Sigma,n}(x), \tag{C.1}$$

where $\{s_i(\Sigma)\}_{i\in[p]}$ are eigenvalues of Σ. By anisotropic local law (Theorem 3.17 of [24], see also proof of Theorem 5 of [22]), given any large $D > 0$ and small $(\alpha_0, \alpha_1) > 0$, there exists $N_0 \equiv N_0(D, \alpha_0, \alpha_1)$, such that $\mathbb{P}(\Omega_n^c) \leq n^{-(D-1)}$ for $n \geq N_0$. Let

$$\widehat{\Sigma} = \mathbf{U} \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U}^\top$$

be the eigendecompositon of $\widehat{\Sigma}$. Then we have:

$$(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} = \mathbf{U} \begin{pmatrix} \mathbf{S}(\mathbf{S} + \lambda \mathbf{I}_p)^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U}^\top$$

Furthermore, for $\lambda > 0$:

$$e_i^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} e_i \leq \|(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\|_{\text{op}} \leq \max_j \frac{s_j}{(s_j + \lambda)^2}$$

Therefore,

$$\|(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\|_\infty \leq \|(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\|_{\text{op}} \leq \max_j \frac{s_j}{(s_j + \lambda)^2} \leq \frac{1}{s_{\min} + \lambda}$$

where $s_{\min}$ is the minimum non-zero singular value of $\widehat{\Sigma}$. Now, if we define $\Omega_{n,1}$ to be event as:

$$\Omega_{n,1} = \left\{ \mathbf{X} : s_{\min}(\widehat{\Sigma}) \geq \frac{1}{2} c (1 - \sqrt{\gamma})^2 \right\},$$

where c is the lower bound on the eigenvalues of Σ, then, by [3] we have $\mathbb{P}(\Omega_{n,1}^c) \to 0$ as $n \uparrow \infty$. As a consequence, on $\Omega_{n,1}$

$$\|(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\|_\infty \leq \frac{1}{\frac{c}{2}(1 - \sqrt{\gamma})^2 + \lambda}$$

Using this fact, we first show that on $\Omega_n \cap \Omega_{n,1}$:

$$\left\| (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \right\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}} \right).$$

To prove this, first, from the normality of ϵ, we have:

$$(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \mid \mathbf{X} \sim \frac{1}{\sqrt{n}} \mathcal{N}\left(0, (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\right)$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left\| (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \right\|_\infty \geq t\sqrt{\log p} \right) \\
&\leq \mathbb{P}\left(\left\| (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \right\|_\infty \geq t\sqrt{\log p}, \Omega_{n,1} \right) + \mathbb{P}(\Omega_{n,1}^c) \\
&\leq \sum_{j=1}^p \mathbb{P}\left(\left| e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \right| \geq t\sqrt{\log p}, \Omega_{n,1} \right) + \mathbb{P}(\Omega_{n,1}^c) \\
&= \sum_{j=1}^p \mathbb{E}\left[\mathbb{P}\left(\left| e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}} \right| \geq t\sqrt{\log p} \mid \mathbf{X} \right) \mathbb{1}_{\mathbf{X} \in \Omega_{n,1}} \right] + \mathbb{P}(\Omega_{n,1}^c) \\
&\leq \sum_{j=1}^p \mathbb{E}\left[2\mathsf{exp}\left(-\frac{t \log p}{2 e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} e_j} \right) \mathbb{1}_{\mathbf{X} \in \Omega_{n,1}} \right] + \mathbb{P}(\Omega_{n,1}^c) \\
&\leq \sum_{j=1}^p \mathbb{E}\left[2\mathsf{exp}\left(-\frac{1}{2}\left(\frac{c}{2}(1 - \sqrt{\gamma})^2 + \lambda \right) t \log p \right) \mathbb{1}_{\mathbf{X} \in \Omega_{n,1}} \right] + \mathbb{P}(\Omega_{n,1}^c) \\
&\leq 2\mathsf{exp}\left(\log p - \frac{1}{2}\left(\frac{c}{2}(1 - \sqrt{\gamma})^2 + \lambda \right) t \log p \right) + \mathbb{P}(\Omega_{n,1}^c) \to 0 \text{ as } t \uparrow \infty .
\end{aligned}$$

This completes the proof. Now, coming back the decomposition of $\widehat{\beta}_\lambda$:

$$\widehat{\beta}_\lambda - \beta_0 = -\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \\ - \left(\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - (\mathbf{I}_p + \tilde{m}_n(-\lambda)\Sigma)^{-1}\right)\beta_0 - (\mathbf{I}_p + \tilde{m}_n(-\lambda)\Sigma)^{-1}\beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}.$$

Hence,

$$\frac{1}{r_n}\left|\|\widehat{\beta}_\lambda - \beta_0\|_\infty - \|(\mathbf{I}_p + \tilde{m}_n(-\lambda)\Sigma)^{-1}\beta_0\|_\infty\right| \\ \leq \frac{1}{r_n}\left\|\left(\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - (\mathbf{I}_p + \tilde{m}_n(-\lambda)\Sigma)^{-1}\right)\beta_0\right\|_\infty + \frac{1}{r_n}\left\|(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}\right\|_\infty$$

That the second term is $o_p(1)$ follows from the fact that $r_n \gg \sqrt{\log p/n}$. Now, on Ω_n,

$$\frac{1}{r_n}\left\|\left(\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - (\mathbf{I}_p + \tilde{m}_n(-\lambda)\Sigma)^{-1}\right)\beta_0\right\|_\infty \leq \frac{1}{r_n n^{\frac{1}{2}-\alpha_0}\lambda} \to 0$$

for any $\alpha_0 < \alpha$. This completes the proof. □

D Proof of Theorem 3.3

D.1 Proof of Ridge estimator

Let $\widehat{\beta}_\lambda$ be the ridge estimator with penalty parameter λ, i.e.

$$\widehat{\beta}_\lambda = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \mathbf{y}}{\lambda} = (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\widehat{\Sigma}\beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}$$

which immediately implies:

$$\widehat{\beta}_\lambda - \beta_0 = -\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \\ = -\lambda m(-\lambda)\beta_0 - \lambda\left\{(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - m(-\lambda)\right\}\beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}$$

Therefore, for any $1 \leq j \leq p$:

$$\widehat{\beta}_{\lambda,j} - \beta_{0,j} = -\lambda m(-\lambda)\beta_{0,j} - \lambda e_j^\top\left\{(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - m(-\lambda)\right\}\beta_0 + e_j^\top(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \quad \text{(D.1)}$$

We will start with the leave-one-out operation. Fix some $1 \leq j \leq p$ Assume for the time being, the columns are rotation, so that we can write $\mathbf{X}^{(j)} = (\mathbf{x}_j, \mathbf{X}_{-j})$, i.e., we write j^{th} column in the first column. Then, we have:

$$\widehat{\Sigma}^{(j)} + \lambda \mathbf{I}_p = \frac{(\mathbf{X}^{(j)})^\top \mathbf{X}^{(j)}}{n} + \lambda \mathbf{I}_p = \begin{bmatrix} \frac{\|\mathbf{x}_j\|_2^2}{n} + \lambda & \frac{\mathbf{x}_j^\top \mathbf{X}_{-j}}{n} \\ \frac{\mathbf{X}_{-j}^\top \mathbf{x}_j}{n} & \frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1} \end{bmatrix} := \begin{bmatrix} a & \mathbf{b}^\top \\ \mathbf{b} & \mathbf{C} \end{bmatrix}.$$

As a consequence, by Schur complement,

$$e_j^\top(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\beta_0 = e_1^\top(\widehat{\Sigma}^{(j)} + \lambda \mathbf{I}_p)^{-1}\begin{pmatrix} \beta_{0,j} \\ \beta_{0,-j} \end{pmatrix} \\ = \begin{bmatrix} (a - \mathbf{b}^\top \mathbf{C}^{-1}\mathbf{b})^{-1} & -(a - \mathbf{b}^\top \mathbf{C}^{-1}\mathbf{b})^{-1}\mathbf{b}^\top \mathbf{C}^{-1} \end{bmatrix}\begin{pmatrix} \beta_{0,j} \\ \beta_{0,-j} \end{pmatrix}$$

Now, from the definition of $(a, \mathbf{b}, \mathbf{C})$, we have:

$$a - \mathbf{b}^\top \mathbf{C}^{-1}\mathbf{b} = \left[\frac{\|\mathbf{x}_j\|_2^2}{n} + \lambda - \frac{\mathbf{x}_j^\top \mathbf{X}_{-j}}{n}\left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1}\frac{\mathbf{X}_{-j}^\top \mathbf{x}_j}{n}\right]$$

Let us define:

$$A_j = \mathbf{I}_n - \frac{1}{n}\mathbf{X}_{-j}\left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1}\mathbf{X}_{-j}^\top = \lambda\left(\frac{\mathbf{X}_{-j}\mathbf{X}_{-j}^\top}{n} + \lambda \mathbf{I}_n\right)^{-1}. \tag{D.2}$$

Then we have:

$$a - \mathbf{b}^\top \mathbf{C}^{-1}\mathbf{b} = \lambda + \frac{\mathbf{x}_j^\top A_j \mathbf{x}_j}{n}.$$

Now further defining $q_j = 1/(a - \mathbf{b}^\top \mathbf{C}^{-1}\mathbf{b})$, we have:

$$e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\beta_0 = q_j \beta_{0,j} - q_j \mathbf{b}^\top \mathbf{C}^{-1}\beta_{0,-j}, \tag{D.3}$$

$$e_j^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} = q_j \frac{\mathbf{x}_j^\top \boldsymbol{\epsilon}}{n} - q_j \mathbf{b}^\top \mathbf{C}^{-1}\frac{\mathbf{X}_{-j}^\top \boldsymbol{\epsilon}}{n}. \tag{D.4}$$

Therefore, from Equation (D.1), we obtain:

$$\begin{aligned}\widehat{\beta}_{\lambda,j} - \beta_{0,j} &= -\lambda m(-\lambda)\beta_{0,j} - \lambda(q_j - m(-\lambda))\beta_{0,j} + q_j \frac{\mathbf{x}_j^\top \boldsymbol{\epsilon}}{n} \\ &\quad + \lambda q_j \mathbf{b}^\top \mathbf{C}^{-1}\beta_{0,-j} - q_j \mathbf{b}^\top \mathbf{C}^{-1}\frac{\mathbf{X}_{-j}^\top \boldsymbol{\epsilon}}{n} \\ &:= -\lambda m(-\lambda)\beta_{0,j} - \lambda(q_j - m(-\lambda))\beta_{0,j} + q_j x_j^\top \mathbf{c}_j\,, \end{aligned} \tag{D.5}$$

where

$$\begin{aligned}\mathbf{c}_j &= \frac{1}{n}\left[\boldsymbol{\epsilon} + \lambda \mathbf{X}_{-j}\left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1}\beta_{0,-j} - \mathbf{X}_{-j}\left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1}\frac{\mathbf{X}_{-j}^\top \boldsymbol{\epsilon}}{n}\right] \\ &= \lambda \frac{\mathbf{X}_{-j}}{n}\left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1}\beta_{0,-j} + \frac{A_j \boldsymbol{\epsilon}}{n}.\end{aligned}$$

Now, observe that $\mathbf{c}_j$ depends on $(\mathbf{X}_{-j}, \boldsymbol{\epsilon})$ which is independent of $\mathbf{x}_j$. Therefore,

$$\mathbf{x}_j^\top c_j \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \|\mathbf{c}_j\|_2^2) \implies \frac{\mathbf{x}_j^\top \mathbf{c}_j}{\|\mathbf{c}_j\|} \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1).$$

For the rest of the calculation, define, for notational simplicity, $r_n = \sqrt{2\log p/n}$. The following lemma will be helpful to establish the final result:

Lemma 1 (Uniform leave-one-out resolvent estimates)**.** *Under Assumption 1 -3, we have:*

$$\max_{1\le j\le p} |q_j - m| = O_P\left(\sqrt{\frac{\log p}{n}}\right).$$

Lemma 2. *Under Assumption 1 -3 we have with probability $\ge 1 - 7p^{-K+1}$:*

$$\max_{1\le j\le p}\left|n\|\mathbf{c}_j\|_2^2 - c_{\lambda,n}\right| \le \frac{C_1}{p^{\frac{1}{2}-\epsilon_0}\lambda} + \sqrt{C_2\frac{\log p}{n}} + \frac{C_3\lambda}{p^{1-\epsilon_0}} := \delta_{n,\mathbf{c}}\,,$$

$$c_{\lambda,n} = \lambda^2(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda))\|\beta_{0,-j}\|_2^2 + \sigma_\epsilon^2 \int \frac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_\gamma(x)$$

where $\widetilde{F}_\gamma$ is the distribution corresponding to the companion of the Marchenko-Pastur distribution, and $K > 1$, a large number, can be chosen as large as possible, at the cost of increasing the other constants (C_1, C_2, C_3). In particular, defining c_λ as:

$$c_\lambda = \lambda^2(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda))B + \sigma_\epsilon^2 \int \frac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_\gamma(x)$$

where have $|c_{\lambda,n} - c_\lambda| = O((\log p)/n)$.

Lemma 3. *Under Assumption 1 -3, we have:*

$$\frac{\max_{1\le j\le p} \mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{\frac{2c_\lambda \log p}{n}}} \xrightarrow{p} 1\,.$$

Corollary D.1. *Under Assumption 1 - 3, and for any subset $T_n \subseteq [p]$ with $\log|T_n|/\log p \to \rho \in [0,1]$ we have:*

$$\frac{\max_{j\in T_n} s_j \mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{\frac{2c_\lambda \log p}{n}}} \xrightarrow{p} \sqrt{\rho}\,, \qquad \frac{\max_{j\in T_n} |\mathbf{x}_j^\top \mathbf{c}_j|}{\sqrt{\frac{2c_\lambda \log p}{n}}} \xrightarrow{p} \sqrt{\rho}\,.$$

where $s_j \in \{-1,1\}$ are some fixed assignment of signs independent of $(\mathbf{X},\epsilon)$.

Let us first complete the proof with these lemmas, and then we will complete the proof of these lemmas. Recall that from Equation (D.5):

$$\widehat{\beta}_{\lambda,j} - \beta_j = -\lambda m(-\lambda)\beta_{0,j} - \lambda(q_j - m(-\lambda))\beta_{0,j} + (q_j - m)\mathbf{x}_j^\top \mathbf{c}_j + m\,\mathbf{x}_j^\top \mathbf{c}_j\,.$$

Our first claim is that:

$$\max_j \frac{|(q_j - m(-\lambda))\beta_{0,j}|}{r_n} = o_p(1), \qquad \max_j \frac{\left|(q_j - m)\mathbf{x}_j^\top \mathbf{c}_j\right|}{r_n} = o_p(1), \quad \text{in probability.}$$

The first conclusion follows from Lemma 1 and the fact that $\|\beta_0\|_\infty = O(\sqrt{\log p/n})$ and the second conclusion follows from Lemma 1 and Lemma 3. Therefore,

$$\frac{\|\widehat{\beta}_\lambda - \beta_0\|_\infty}{r_n} = \max_{1\le j\le p} \frac{|-\lambda m(-\lambda)\beta_{0,j} + m\,\mathbf{x}_j^\top \mathbf{c}_j|}{r_n} + o_p(1)\,.$$

Also, recall that definition of S_n in Assumption 3; on S_n $|\beta_{0,j}| \sim r_n$ and on S_n^c, $|\beta_{0,j}| = o(r_n)$. Therefore

$$\max_{j\notin S_n} |\lambda m \beta_j| = o(r_n).$$

Since $|S_n^c| = p^{1+o(1)}$, Lemma 3, in particular Corollary D.1, gives

$$\max_{j\notin S_n} \frac{|m\mathbf{x}_j^\top \mathbf{c}_j|}{r_n} \xrightarrow{p} \tau_\lambda := m_\gamma(-\lambda)\sqrt{c_\lambda}\,.$$

Thus

$$\max_{j\notin S_n} \frac{|\widehat{\beta}_{\lambda,j} - \beta_j|}{r_n} \xrightarrow{p} \tau_\lambda\,.$$

Now, on S_n, $|\beta_j| = (a + o(1))r_n$ uniformly. Hence

$$|\lambda m \beta_j| = \lambda m a\, r_n + o(r_n)$$

uniformly over $j \in S_n$. For the lower bound, define deterministic signs

$$s_j = \mathsf{sign}(-\lambda m \beta_j), \qquad j \in S_n,$$

with an arbitrary choice if $\beta_j = 0$. Then

$$\frac{|-\lambda m\beta_j + m\mathbf{x}_j^\top \mathbf{c}_j|}{r_n} \ge \lambda m \frac{|\beta_j|}{r_n} + \frac{s_j\, m\mathbf{x}_j^\top \mathbf{c}_j}{r_n} = \lambda m a + \frac{s_j\, m\mathbf{x}_j^\top \mathbf{c}_j}{r_n} + \lambda m\left(\frac{|\beta_j|}{r_n} - a\right).$$

As the last term is $o(1)$, we have, taking the maximum over S_n and applying Corollary D.1, with $|S_n| = p^{\rho+o(1)}$:

$$\limsup_{p\uparrow\infty} \max_{j\in S_n} \frac{|\widehat{\beta}_{\lambda,j} - \beta_j|}{r_n} \ge (\lambda m a + \tau_\lambda\sqrt{\rho})\,.$$

For the upper bound,

$$|-\lambda m\beta_j + m\mathbf{x}_j^\top \mathbf{c}_j| \le \lambda m|\beta_j| + |m\mathbf{x}_j^\top \mathbf{c}_j|,$$

and consequently:

$$\max_{j\in S_n}\frac{|\widehat{\beta}_{\lambda,j}-\beta_j|}{r_n} \le \lambda m \max_{j\in S_n}\frac{|\beta_{0,j}|}{r_n} + \max_{j\in S_n}\frac{|m\mathbf{x}_j^\top \mathbf{c}_j|}{r_n}$$
$$= \lambda m a + \max_{j\in S_n}\frac{|m\mathbf{x}_j^\top \mathbf{c}_j|}{r_n} + \lambda m \max_{j\in S_n}\left(\frac{|\beta_{0,j}|}{r_n} - a\right)$$

Again, the last term is $o(1)$ by Assumption 3, and therefore, an application of Corollary D.1 yields:

$$\limsup_{p\uparrow\infty}\max_{j\in S_n}\frac{|\widehat{\beta}_{\lambda,j}-\beta_j|}{r_n} \le \lambda m a + \tau_\lambda\sqrt{\rho}\,.$$

Hence

$$\max_{j\in S_n}\frac{|\widehat{\beta}_{\lambda,j}-\beta_j|}{r_n} \xrightarrow{p} (\lambda m a + \tau_\lambda\sqrt{\rho})\,.$$

Combining the estimates on S_n and S_n^c, we conclude:

$$\frac{\|\widehat{\beta}_\lambda-\beta\|_\infty}{r_n} \xrightarrow{p} \max\{\lambda m a + \tau_\lambda\sqrt{\rho}, \tau_\lambda\}\,.$$

D.2 Proof of other estimates

The proof for $\widehat{\beta}_{\mathrm{null}}$, $\widehat{\beta}_{\mathrm{MoM}}$ and $\widehat{\beta}_{\mathrm{OLS}}$ is same as Theorem 3.1 and hence skipped.

E Proof of Theorem 3.4

Proof. For notational simplicity, we use

$$m_\lambda := m_\gamma(-\lambda), \qquad \widetilde{m}_\lambda := \widetilde{m}_\gamma(-\lambda),$$

where $\widetilde{m}_\gamma$ denotes the Stieltjes transform of the companion Marchenko–Pastur law. Recall the companion identity

$$\widetilde{m}_\lambda = \frac{1-\gamma}{\lambda} + \gamma m_\lambda.$$

The MP fixed-point equation may be written as

$$m_\lambda = \frac{1}{\lambda + \lambda\widetilde{m}_\lambda} \iff \frac{1}{m_\lambda} = \lambda(1+\widetilde{m}_\lambda).$$

We first simplify τ_λ^2. Differentiating

$$\frac{1}{m_\lambda} = \lambda(1+\widetilde{m}_\lambda)$$

with respect to λ, and noting that

$$\frac{d}{d\lambda}m_\gamma(-\lambda) = -m_\gamma'(-\lambda), \qquad \frac{d}{d\lambda}\widetilde{m}_\gamma(-\lambda) = -\int\frac{1}{(x+\lambda)^2}\,d\widetilde{F}_{\gamma,\mathrm{MP}}(x),$$

we obtain

$$\frac{m_\gamma'(-\lambda)}{m_\lambda^2} = 1 + \widetilde{m}_\lambda - \lambda\int\frac{1}{(x+\lambda)^2}\,d\widetilde{F}_{\gamma,\mathrm{MP}}(x).$$

Since $1+\widetilde{m}_\lambda = (\lambda m_\lambda)^{-1}$. This gives

$$\lambda\int\frac{1}{(x+\lambda)^2}\,d\widetilde{F}_{\gamma,\mathrm{MP}}(x) = \frac{1}{\lambda m_\lambda} - \frac{m_\gamma'(-\lambda)}{m_\lambda^2}.$$

Multiplying by λm_λ^2, we get

$$\lambda^2 m_\lambda^2\int\frac{1}{(x+\lambda)^2}\,d\widetilde{F}_{\gamma,\mathrm{MP}}(x) = m_\lambda - \lambda m_\gamma'(-\lambda).$$

Therefore,

$$m_\lambda^2 \int \frac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_{\gamma,\mathrm{MP}}(x) = m_\lambda - \lambda m_\gamma'(-\lambda).$$

Define $h_\lambda := m_\lambda - \lambda m_\gamma'(-\lambda)$. Then the above identity gives

$$m_\lambda^2 \int \frac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_{\gamma,\mathrm{MP}}(x) = h_\lambda.$$

Hence the definition of τ_λ^2 simplifies to

$$\tau_\lambda^2 = m_\lambda^2 \lambda^2 h_\lambda B + \sigma_\varepsilon^2 h_\lambda = h_\lambda \left\{\sigma_\varepsilon^2 + \lambda^2 m_\lambda^2 B\right\} = h_\lambda \left\{\sigma_\varepsilon^2 + \lambda^2 m_\gamma^2(-\lambda) B\right\}.$$

Now set $t := \lambda m_\lambda$. Using

$$\frac{1}{m_\lambda} = \lambda + \lambda \widetilde{m}_\lambda = \lambda + 1 - \gamma + \gamma \lambda m_\lambda,$$

we get

$$\frac{\lambda}{t} = \lambda + 1 - \gamma + \gamma t \implies \lambda\left(\frac{1}{t} - 1\right) = 1 - \gamma + \gamma t \implies \lambda = \frac{t(1-\gamma+\gamma t)}{1-t}.$$

This reparametrizes $\lambda > 0$ by t. When $\gamma > 1$, $(\gamma-1)/\gamma < t < 1$, whereas when $0 < \gamma < 1$, $t \in (0,1)$. Differentiating λ with respect to t, we obtain

$$\frac{d\lambda}{dt} = \frac{1 - \gamma + 2\gamma t - \gamma t^2}{(1-t)^2}.$$

Since

$$h_\lambda = m_\lambda - \lambda m_\gamma'(-\lambda) = \frac{d}{d\lambda}\{\lambda m_\gamma(-\lambda)\} = \frac{dt}{d\lambda},$$

we have

$$h_\lambda = \frac{(1-t)^2}{1-\gamma+2\gamma t - \gamma t^2} \implies \tau_\lambda^2 = \frac{(1-t)^2}{1-\gamma+2\gamma t-\gamma t^2}\left\{\sigma_\varepsilon^2 + B t^2\right\}.$$

Define

$$\tau(t) := (1-t)\sqrt{\frac{\sigma_\varepsilon^2 + B t^2}{1-\gamma+2\gamma t - \gamma t^2}}.$$

Then

$$R^*_{\mathrm{ridge}} = \inf_{t \in I_\gamma} L(t), \quad L(t) := \max\{\tau(t),\, at + \sqrt{\rho}\,\tau(t)\},$$

with

$$I_\gamma = \begin{cases} (0,1), & 0 < \gamma < 1, \\ \left(1 - \frac{1}{\gamma}, 1\right), & \gamma > 1. \end{cases}$$

First suppose $\gamma > 1$. Let $t_- := 1 - \gamma^{-1}$. Then $t \in (t_-, 1)$, and hence for every admissible t,

$$L(t) = \max\{\tau(t),\, at + \sqrt{\rho}\,\tau(t)\} \geq at \geq at_-.$$

Taking the infimum over $t \in (t_-, 1)$, we get $R^*_{\mathrm{ridge}} \geq a(1-\gamma^{-1})$. If $a(1-\gamma^{-1}) > \sqrt{B + \sigma_\epsilon^2}$, then

$$R^*_{\mathrm{ridge}} > \sqrt{B + \sigma_\varepsilon^2} = R_{\mathrm{MoM}}.$$

Thus, an optimally tuned ridge is strictly worse than MoM under the stated condition. Next suppose $0 < \gamma < 1$. Then $t \in (0,1)$, and $t \downarrow 0$ corresponds to $\lambda \downarrow 0$. Since

$$\tau(t)^2 = \frac{(1-t)^2(\sigma_\varepsilon^2 + B t^2)}{1-\gamma+2\gamma t - \gamma t^2},$$

we have

$$\lim_{t \downarrow 0} \tau(t)^2 = \frac{\sigma_\varepsilon^2}{1-\gamma} \implies \lim_{t \downarrow 0} \tau(t) = \frac{\sigma_\varepsilon}{\sqrt{1-\gamma}}.$$

Therefore,

$$\lim_{t\downarrow 0}\{at+\sqrt{\rho}\,\tau(t)\}=\sqrt{\rho}\,\frac{\sigma_\varepsilon}{\sqrt{1-\gamma}}\le\frac{\sigma_\varepsilon}{\sqrt{1-\gamma}}\implies\lim_{t\downarrow 0}L(t)=\frac{\sigma_\varepsilon}{\sqrt{1-\gamma}},$$

and consequently $R^*_{\text{ridge}}\le R_{\text{OLS}}$. This proves that an optimally tuned ridge is always no worse than OLS when $0<\gamma<1$. Now, for the strict improvement, a direct differentiation of $\tau(t)$ gives

$$\frac{\tau'(t)}{\tau(t)}=-\frac{1}{1-t}+\frac{Bt}{\sigma_\varepsilon^2+Bt^2}-\frac{\gamma(1-t)}{1-\gamma+2\gamma t-\gamma t^2}.$$

Hence

$$\tau'(0)=-\frac{\sigma_\varepsilon}{(1-\gamma)^{3/2}}<0.$$

If $0\le\rho<1$, then $\sqrt{\rho}<1$. Since $\tau(t)<R_{\text{OLS}}$, for all sufficiently small $t>0$, and

$$at+\sqrt{\rho}\,\tau(t)\to\sqrt{\rho}\,R_{\text{OLS}}<R_{\text{OLS}},$$

we have $L(t)<R_{\text{OLS}}$ for all sufficiently small $t>0$. Therefore $R^*_{\text{ridge}}<R_{\text{OLS}}$. If $\rho=1$, then $L(t)=at+\tau(t)$. Since

$$\frac{d}{dt}\{at+\tau(t)\}\Big|_{t=0}=a-\frac{\sigma_\varepsilon}{(1-\gamma)^{3/2}},$$

we have $L(t)<R_{\text{OLS}}$ for all sufficiently small $t>0$ whenever $a<\sigma_\epsilon/(1-\gamma)^{3/2}$. Thus in this case also $R^*_{\text{ridge}}<R_{\text{OLS}}$. The proof is complete. □

F Proof of Theorem A.1

Proof. Let us start from the definition of $\widehat{\beta}_\lambda$:

$$\widehat{\beta}_\lambda-\beta_0=-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\beta_0+(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top\epsilon}{n}=-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}+(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top\epsilon}{n}$$

We have already established in the proof of Theorem 3.1 that:

$$\left\|(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top\epsilon}{n}\right\|_\infty=O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

Therefore, all we need to prove is that:

$$\left\|-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right\|_\infty=O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

Define the event Ω_n as:

$$\Omega_n=\left\{\mathbf{X}:|(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-2}_{ij}-m'(-\lambda)\delta_{ij}|\le\frac{1}{n^{\frac{1}{2}-\alpha_0}\lambda}+\frac{C}{n}\right\}$$

for some constant $C>0$ independent of (n,λ). As argued in the proof of Theorem 3.3, $\mathbb{P}(\Omega_{n,1})\ge 1-n^{D-2}$. As a consequence, on Ω_n, we have:

$$\left|\lambda^2(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-2}_{ii}-\lambda^2m'(-\lambda)\right|\le\frac{\lambda}{n^{\frac{1}{2}-\alpha_0}}+\frac{C\lambda^2}{n}:=a(n,\lambda).$$

Now,

$$\begin{aligned}
&\mathbb{P}\left(\left\|-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right\|_\infty \geq t\right)\\
&= \mathbb{P}\left(\left\|-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right\|_\infty \geq t, \Omega_n\right) + \mathbb{P}\left(\left\|-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right\|_\infty \geq t, \Omega_n^c\right)\\
&\leq \mathbb{P}\left(\left\|-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right\|_\infty \geq t, \Omega_n\right) + \mathbb{P}\left(\Omega_n^c\right)\\
&\leq \sum_{j=1}^{p}\mathbb{P}\left(\left|-\lambda e_j^\top(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\boldsymbol{\mu}\right| \geq t\sqrt{n}\right) + \mathbb{P}\left(\Omega_n^c\right)\\
&\leq \sum_{j=1}^{p} 2\mathsf{exp}\left(-\frac{nt^2}{2\lambda^2 e_j^\top(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-2}e_j}\right) + \mathbb{P}\left(\Omega_n^c\right)\\
&\leq \sum_{j=1}^{p} 2\mathsf{exp}\left(-\frac{nt^2}{2\lambda^2 m'(-\lambda)+2a(n,\lambda)}\right) + \mathbb{P}\left(\Omega_n^c\right)\\
&\leq 2\mathsf{exp}\left(\log p - \frac{nt^2}{2\lambda^2 m'(-\lambda)+2a(n,\lambda)}\right) + \mathbb{P}\left(\Omega_n^c\right)
\end{aligned}$$

Hence, it is immediate that:

$$\left\|-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

which completes the proof. □

G Proof of Theorem A.2

Proof. Recall that we assume $\beta = \sigma_\beta \mathbf{z}/\sqrt{n}$ where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_p)$ independent of (X, ϵ). Again, from the definition of $\widehat{\beta}_\lambda$ we have:

$$\begin{aligned}
\widehat{\beta}_\lambda - \beta_0 &= -\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\beta_0 + (\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top\boldsymbol{\epsilon}}{n}\\
&= \frac{1}{\sqrt{n}}\left\{-\lambda(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\sigma_\beta\mathbf{z} + (\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top\boldsymbol{\epsilon}}{\sqrt{n}}\right\}
\end{aligned}$$

From the Gaussianity and independence of $(\mathbf{z}, \boldsymbol{\epsilon})$, we have:

$$\widehat{\beta}_\lambda - \beta_0 \mid \mathbf{X} \sim \frac{1}{\sqrt{n}}\mathcal{N}\left(0, \sigma_\beta^2\lambda^2(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-2} + \sigma_\epsilon^2(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\right)$$

Now we know by Marchenko-Pasteur law, we know for any fixed $\lambda > 0$:

$$\begin{aligned}
\frac{1}{p}\mathsf{tr}\left(\Psi_n(\lambda)\right) &:= \frac{1}{p}\mathsf{tr}\left(\sigma_\beta^2\lambda^2(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-2} + \sigma_\epsilon^2(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\right)\\
&\longrightarrow \sigma_\beta^2\lambda^2 m_\gamma'(-\lambda) + \sigma_\epsilon^2\left(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda)\right)\\
&:= \Psi(\lambda).
\end{aligned} \tag{G.1}$$

Our goal is to show that:

$$\sqrt{\frac{n}{2\Psi(\lambda)\log p}}\|\widehat{\beta}_\lambda - \beta_0\|_\infty \to 1 \quad \text{in probability.}$$

First, we bound the difference between $\Psi_n(\lambda)$ and Ψ_λ using local laws. Towards that end, we use Theorem 3.17 of [24] to have:

$$\left|e_i^\top(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}e_j - m(-\lambda)\delta_{ij}\right| \leq \frac{1}{n^{1/2-\alpha_0}\lambda}$$

for all $\lambda \geq n^{-\frac{2}{3}+\epsilon_0}$ with probability $\geq 1 - n^{-D}$ for all $n \geq N_0 \equiv N_0(\alpha_0, \epsilon_0, D)$. Define the event $\Omega_{n,1}$ as:

$$\Omega_n = \left\{ \mathbf{X} : \left| e_i^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} e_j - m(-\lambda)\delta_{ij} \right| \leq \frac{1}{n^{\frac{1}{2}-\alpha_0}\lambda} \forall\, i, j \right\}$$

Therefore, $\mathbb{P}(\Omega_n) \geq 1 - p^2 n^D = 1 - \gamma^2 n^{D-2}$, which is asymptotically negligible as soon as $D > 2$. Furthermore, using the same techniques used in the proof of Theorem 5 of [22] along with the fact that:

$$(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} = -\partial_\eta (\widehat{\Sigma} + (\lambda + \eta)\mathbf{I}_p)^{-1} \mid_{\eta=0}$$

we conclude that on Ω_n:

$$\left| e_i^\top (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} e_j - m'(-\lambda)\delta_{ij} \right| \leq \frac{1}{n^{\frac{1}{2}-2\alpha_0}\lambda} + \frac{c_1}{n} \quad \forall\, i, j\,.$$

As a conesquence, we have on Ω_n, for all (i, j):

$$|\Psi_n(\lambda)_{ij} - \Psi(\lambda)\delta_{ij}| \leq \frac{\sigma_\beta^2 \lambda}{n^{\frac{1}{2}-2\alpha_0}} + \frac{c_1 \sigma_\beta^2 \lambda^2}{n} + \frac{\sigma_\epsilon^2}{n^{\frac{1}{2}-\alpha_0}\lambda} + \frac{\sigma_\epsilon^2}{n^{\frac{1}{2}-2\alpha_0}} + \frac{c_1 \sigma_\epsilon^2 \lambda}{n} := a(n, \lambda)\,.$$

We prove these in two steps, i.e., we show that:

1. We first show that $\mathbb{P}(\|\widehat{\beta}_\lambda - \beta_0\|_\infty \geq \sqrt{2(1+\delta)\Psi(\lambda)\log p/n}) \to 0$ for any $\delta > 0$.
2. We then show that $\mathbb{P}(\|\widehat{\beta}_\lambda - \beta_0\|_\infty \geq \sqrt{2(1-\delta)\Psi(\lambda)\log p/n}) \to 1$ for any $\delta > 0$.

For notational simplicity, define $V_n = \widehat{\beta}_\lambda - \beta_0$ and $\widehat{\sigma}_j^2 = \Psi_{n,jj}$. The proof of the first part is relatively simple:

$$\begin{aligned}
&\mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}} \right) \\
&= \mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}}, \Omega_n \right) + \mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\log p}{n}}, \Omega_n^c \right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}} \mid \mathbf{X} \right) \mathbb{1}_{\mathbf{X}\in\Omega_n} \right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p \mathbb{P}\left(|V_{n,j}| \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{n}} \mid \mathbf{X} \right) \mathbb{1}_{\mathbf{X}\in\Omega_n} \right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p \mathbb{P}\left(|Z| \geq \sqrt{\frac{2(1+\delta)\Psi(\lambda)\log p}{\widehat{\sigma}_j^2}} \mid \mathbf{X} \right) \mathbb{1}_{\mathbf{X}\in\Omega_n} \right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p e^{-\frac{(1+\delta)\Psi(\lambda)\log p}{\widehat{\sigma}_j^2}} \mathbb{1}_{\mathbf{X}\in\Omega_n} \right] + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{E}\left[\sum_{j=1}^p e^{-\frac{(1+\delta)\Psi(\lambda)\log p}{\Psi(\lambda)+a(n,\lambda)}} \mathbb{1}_{\mathbf{X}\in\Omega_n} \right] + \mathbb{P}(\Omega_n^c) \\
&\leq e^{\log p - \frac{(1+\delta)\Psi(\lambda)\log p}{\Psi(\lambda)+a(n,\lambda)}} + \mathbb{P}(\Omega_n^c) \longrightarrow 0\,.
\end{aligned}$$

Now, proving the other part requires more delicate analysis; in particular, we will use the second moment method. Define Y_j as:

$$Y_j = \mathbb{1}_{|V_{n,j}| \geq \sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{n}}}$$

and $Y = \sum_{j=1}^p Y_j$. Then

$$\mathbb{P}\left(\|V_n\|_\infty \geq \sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{n}} \right) = \mathbb{P}(Y > 0)\,.$$

Now, using the second moment method, we have:

$$\mathbb{P}(Y > 0) \geq \frac{(\mathbb{E}[Y])^2}{\mathbb{E}[Y^2]} .$$

As a consequence, if we can show that the RHS of the above equation goes to 1 as $n \uparrow \infty$, we are done. As Y is sum of indicators, we have:

$$\mathbb{E}[Y^2] = \mathbb{E}[Y] + \sum_{i \neq j} \mathbb{E}[Y_i Y_j] .$$

Our first goal is to bound $\mathbb{E}[Y_i Y_j]$ for $i \neq j$. By definition of (Y_i, Y_j), we have:

$$\begin{aligned}
\mathbb{E}[Y_i Y_j] &= \mathbb{P}\left(|V_{n,i}| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{n}}, |V_{n,j}| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{n}} \right) \\
&= \mathbb{P}\left(|Z_i| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\widehat{\sigma}_i^2}}, |Z_j| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\widehat{\sigma}_j^2}} \right) \\
&\leq \mathbb{P}\left(|Z_i| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\widehat{\sigma}_i^2}}, |Z_j| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\widehat{\sigma}_j^2}}, \Omega_n \right) + \mathbb{P}(\Omega_n^c) \\
&\leq \mathbb{P}\left(|Z_i| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, |Z_j| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, \Omega_n \right) + \mathbb{P}(\Omega_n^c) \\
&= \mathbb{E}\left[\mathbb{P}\left(|Z_i| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, |Z_j| \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \mid \mathbf{X} \right) \mathbb{1}_{\mathbf{X} \in \Omega_n} \right] + \mathbb{P}(\Omega_n^c) \\
&= 4\mathbb{E}\left[\mathbb{P}\left(Z_i \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, Z_j \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \mid \mathbf{X} \right) \mathbb{1}_{\mathbf{X} \in \Omega_n} \right] + \mathbb{P}(\Omega_n^c)
\end{aligned}$$

Note that, here (Z_i, Z_j) follows Bivariate Gaussian distribution with mean 0, variance 1 and correlation $\widehat{\rho}_{ij}$, where

$$\widehat{\sigma}_i \widehat{\sigma}_j \widehat{\rho}_{ij} = \Psi_{n,ij} .$$

To bound the joint tail probability of (Z_i, Z_j), we use Lemma 3.5 of [6], which says that:

$$\begin{aligned}
&\mathbb{P}\left(Z_i \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, Z_j \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \mid \mathbf{X} \right) \\
&= \mathbb{P}\left(Z_i \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \mid Z_j \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, \mathbf{X} \right) \mathbb{P}\left(Z_j \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \mid \mathbf{X} \right) \\
&\leq (1 + \widehat{\rho}_{ij}) \bar{\Phi}\left(\sqrt{\frac{2(1-\widehat{\rho}_{ij}^2)\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \right) \bar{\Phi}\left(\sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \right)
\end{aligned}$$

Now, on the event Ω_n, we have $\widehat{\sigma}_i \widehat{\sigma}_j |\rho_{ij}| \leq a(n,\lambda)$. As a consequence, we have, on Ω_n:

$$|\rho_{ij}| \leq \frac{1}{\widehat{\sigma}_i \widehat{\sigma}_j} a(n,\lambda) \leq \frac{a(n,\lambda)}{\Psi(\lambda) - a(n,\lambda)} \leq \frac{2a(n,\lambda)}{\Psi(\lambda)}$$

where the last inequality holds for all large n provided $\Psi(\lambda)$ is bounded away from 0 and $a(n,\lambda)$ goes to 0 as $n \uparrow \infty$. Therefore, we have:

$$\begin{aligned}
&\mathbb{P}\left(Z_i \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}}, Z_j \geq \sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \mid \mathbf{X} \right) \\
&\leq \left(1 + \frac{2a(n,\lambda)}{\Psi(\lambda)} \right) \bar{\Phi}\left(\sqrt{\frac{2\left(1 - \frac{4a^2(n,\lambda)}{\Psi^2(\lambda)}\right)\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \right) \bar{\Phi}\left(\sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda) + a(n,\lambda)}} \right)
\end{aligned}$$

As an immediate consequence, we have:

$$\sum_{i\neq j}\mathbb{E}[Y_iY_j] \leq 4p(p-1)\times\left(1+\frac{2a(n,\lambda)}{\Psi(\lambda)}\right)\bar{\Phi}\left(\sqrt{\frac{2\left(1-\frac{4a^2(n,\lambda)}{\Psi^2(\lambda)}\right)\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda)+a(n,\lambda)}}\right)\bar{\Phi}\left(\sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda)+a(n,\lambda)}}\right)\mathbb{P}(\Omega_n)$$
$$+\,p(p-1)\mathbb{P}(\Omega_n^c)\,.$$

To complete the proof, we need a lower bound on $\mathbb{E}[Y]$ as well. By definition of Y, we have:

$$\begin{aligned}
\mathbb{E}[Y]=\sum_{j=1}^{p}\mathbb{E}[Y_j] &= \sum_j \mathbb{P}\left(|V_{n,j}|\geq\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{n}}\right)\\
&\geq \sum_j \mathbb{E}\left[\mathbb{P}\left(|V_{n,j}|\geq\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{n}}\mid\mathbf{X}\right)\mathbb{1}_{\mathbf{X}\in\Omega_n}\right]\\
&= \sum_j \mathbb{E}\left[\mathbb{P}\left(|Z_j|\geq\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\widehat{\sigma}_j^2}}\mid\mathbf{X}\right)\mathbb{1}_{\mathbf{X}\in\Omega_n}\right]\\
&\geq \sum_j \mathbb{E}\left[\mathbb{P}\left(|Z_j|\geq\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}\mid\mathbf{X}\right)\mathbb{1}_{\mathbf{X}\in\Omega_n}\right]\\
&= \mathbb{P}(\Omega_n)\times 2p\times\bar{\Phi}\left(\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}\right)
\end{aligned}$$

First we claim that for any $\delta>0$, $\mathbb{E}[Y]\to\infty$ as $(n,p)\uparrow\infty$. Towards that end, we use the lower bound on Mill's ratio, i.e. $\bar{\Phi}(x)\geq\phi(x)(x/(1+x^2))$. Therefore, we have:

$$\begin{aligned}
p\times\bar{\Phi}\left(\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}\right) &\geq \frac{p\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}}{1+\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}\frac{1}{\sqrt{2\pi}}\mathsf{exp}\left(-\frac{(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}\right)\\
&\gtrsim \frac{p^{\delta}}{\sqrt{\log p}}\to\infty\,.
\end{aligned}$$

Now, coming to the final part of the proof:

$$\mathbb{P}(Y>0)\geq\frac{(\mathbb{E}[Y])^2}{\mathbb{E}[Y^2]}=\frac{(\mathbb{E}[Y])^2}{\mathbb{E}[Y]+\sum_{i,j}\mathbb{E}[Y_iY_j]}=\frac{1}{\frac{1}{\mathbb{E}[Y]}+\frac{\sum_{i,j}\mathbb{E}[Y_iY_j]}{(\mathbb{E}[Y])^2}}\,.$$

Now, as $\mathbb{E}[Y]\to\infty$, $1/\mathbb{E}[Y]\to 0$. On the other hand:

$$\begin{aligned}
\frac{\sum_{i,j}\mathbb{E}[Y_iY_j]}{(\mathbb{E}[Y])^2} \leq\;& \frac{4p(p-1)\times\left(1+\frac{2a(n,\lambda)}{\Psi(\lambda)}\right)\bar{\Phi}\left(\sqrt{\frac{2\left(1-\frac{4a^2(n,\lambda)}{\Psi^2(\lambda)}\right)\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda)+a(n,\lambda)}}\right)\bar{\Phi}\left(\sqrt{\frac{2\Psi(\lambda)(1-\delta)\log p}{\Psi(\lambda)+a(n,\lambda)}}\right)\mathbb{P}(\Omega_n)}{\mathbb{P}^2(\Omega_n)\times 4p^2\times\bar{\Phi}^2\left(\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}\right)}\\
&+\frac{p(p-1)\mathbb{P}(\Omega_n^c)}{\mathbb{P}^2(\Omega_n)\times 4p^2\times\bar{\Phi}^2\left(\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{\Psi(\lambda)-a(n,\lambda)}}\right)}
\end{aligned}$$

Now, it is immediate that the first term goes to 1 and the second term goes to 0 as $(n,p)\uparrow\infty$. Therefore, we have:

$$\liminf_{n,p\uparrow\infty}\mathbb{P}(Y>0)\geq 1\,,$$

which implies:

$$\mathbb{P}\left(\|V_n\|_\infty\geq\sqrt{\frac{2(1-\delta)\Psi(\lambda)\log p}{n}}\right)\to 1\,.$$

This completes the proof.

G.1 Min-norm interpolator

In this section, we extend our calculation for the min-norm interpolator when $\gamma > 1$. Recall that the min-norm interpolator is:

$$\widehat{\beta} = \widehat{\Sigma}^\dagger \frac{\mathbf{X}^\top \mathbf{y}}{n} = \widehat{\Sigma}^\dagger \widehat{\Sigma} \beta_0 + \widehat{\Sigma}^\dagger \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} = \sigma_\beta \widehat{\Sigma}^\dagger \widehat{\Sigma} \frac{\mathbf{z}}{\sqrt{n}} + \widehat{\Sigma}^\dagger \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}$$

Therefore,

$$\sqrt{n}(\widehat{\beta} - \beta_0) = (\widehat{\Sigma}^\dagger \widehat{\Sigma} - \mathbf{I}_p)\sigma_\beta \mathbf{z} + \widehat{\Sigma}^\dagger \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}}$$

Therefore,

$$\sqrt{n}(\widehat{\beta} - \beta_0) \mid \mathbf{X} \sim \mathcal{N}\left(0, \sigma_\beta^2 (\widehat{\Sigma}^\dagger \widehat{\Sigma} - \mathbf{I}_p)^2 + \sigma_\epsilon^2 \widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger\right)$$

For notational simplicity, let us define:

$$\tilde{\Psi}_n = \sigma_\beta^2 (\widehat{\Sigma}^\dagger \widehat{\Sigma} - \mathbf{I}_p)^2 + \sigma_\epsilon^2 \widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger \,.$$

Now, for $\lambda = \lambda_n > 0$, define $\Psi_n(\lambda)$ as:

$$\Psi_n(\lambda) = \sigma_\beta^2 \lambda^2 (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} + \sigma_\epsilon^2 (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \,.$$

First, we claim:

$$\|\Psi_n(\lambda) - \tilde{\Psi}_n\|_\infty \le \sigma_\beta^2 \left(\frac{\lambda}{s_{\min} + \lambda}\right)^2 + \frac{2\lambda\sigma_\epsilon^2}{s_{\min}^2} \le \frac{\sigma_\beta^2 \lambda^2 + 2\lambda\sigma_\epsilon^2}{s_{\min}^2} \,. \tag{G.2}$$

where $s_{\min}$ is the minimum non-zero eigenvalue of $\widehat{\Sigma}$. To prove this, first observe that, by the definition of Moore-Penrose inverse:

$$(\widehat{\Sigma}^\dagger \widehat{\Sigma} - \mathbf{I}_p)^2 = \widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger \widehat{\Sigma} - \widehat{\Sigma}^\dagger \widehat{\Sigma} - \widehat{\Sigma}^\dagger \widehat{\Sigma} + \mathbf{I}_p = \widehat{\Sigma}^\dagger \widehat{\Sigma} - 2\widehat{\Sigma}^\dagger \widehat{\Sigma} + \mathbf{I}_p = \mathbf{I}_p - \widehat{\Sigma}^\dagger \widehat{\Sigma}$$

where we have used the fact $\widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger = \widehat{\Sigma}^\dagger$. Define the eigendecomposition of $\widehat{\Sigma}$ as:

$$\widehat{\Sigma} = \mathbf{U} \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U} \implies \widehat{\Sigma}^\dagger = \mathbf{U} \begin{pmatrix} \mathbf{S}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U} \,.$$

This immediately implies:

$$\mathbf{I}_p - \widehat{\Sigma}^\dagger \widehat{\Sigma} = \mathbf{I}_p - \mathbf{U} \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U} = \mathbf{U} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{pmatrix} \mathbf{U} \,.$$

On the other hand:

$$\lambda^2 (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2} = \mathbf{U} \begin{pmatrix} \lambda^2 (\mathbf{S} + \lambda \mathbf{I}_p)^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{pmatrix} \mathbf{U}$$

Therefore,

$$\|(\mathbf{I}_p - \widehat{\Sigma}^\dagger \widehat{\Sigma}) - \lambda^2 (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2}\|_{\mathrm{op}} \le \max_j \left(\frac{\lambda}{s_j + \lambda}\right)^2 \le \left(\frac{\lambda}{s_{\min} + \lambda}\right)^2$$

Finally using $\|A\|_\infty \le \|A\|_{\mathrm{op}}$, we have:

$$\|(\mathbf{I}_p - \widehat{\Sigma}^\dagger \widehat{\Sigma}) - \lambda^2 (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-2}\|_\infty \le \left(\frac{\lambda}{s_{\min} + \lambda}\right)^2 .$$

Similarly, we have:

$$\widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger = \widehat{\Sigma}^\dagger = \mathbf{U} \begin{pmatrix} \mathbf{S}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U}, \quad (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} = \mathbf{U} \begin{pmatrix} \mathbf{S}(\mathbf{S} + \lambda \mathbf{I}_p)^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{U}$$

As a consequence,

$$\|\widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger - (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\|_{\mathrm{op}} \le \max_j \left| \frac{1}{s_j} - \frac{s_j}{(s_j + \lambda)^2} \right|$$

$$
\begin{aligned}
&= \max_j \frac{2s_j\lambda + \lambda^2}{s_j(s_j+\lambda)^2} \\
&\le \max_j \frac{2\lambda(s_j+\lambda)}{s_j(s_j+\lambda)^2} \\
&\le \max_j \frac{2\lambda}{s_j(s_j+\lambda)} \le \frac{2\lambda}{s_{\min}^2}\,.
\end{aligned}
$$

Again, by $\|A\|_\infty \le \|A\|_{\mathrm{op}}$, we conclude:

$$
\|\widehat{\Sigma}^\dagger\widehat{\Sigma}\widehat{\Sigma}^\dagger - (\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}\|_\infty \le \frac{2\lambda}{s_{\min}^2}\,.
$$

Combining these two bounds concludes the proof of Equation (G.2). Now, from the proof of the previous part we already have an event Ω_n with probability $\ge 1-\gamma^2 n^{D-2}$ such that:

$$
|\Psi_n(\lambda)_{ij} - \Psi(\lambda)\delta_{ij}| \le \frac{\sigma_\beta^2\lambda}{n^{\frac{1}{2}-2\alpha_0}} + \frac{c_1\sigma_\beta^2\lambda^2}{n} + \frac{\sigma_\epsilon^2}{n^{\frac{1}{2}-\alpha_0}\lambda} + \frac{\sigma_\epsilon^2}{n^{\frac{1}{2}-2\alpha_0}} + \frac{c_1\sigma_\epsilon^2\lambda}{n} := a(n,\lambda)\,.
$$

where $\Psi(\lambda)$ is as defined in Equation (G.1). Furthermore, from the definition $\Psi(\lambda)$, we have:

$$
\Psi(0) = \lim_{\lambda\downarrow 0}\Psi(\lambda) = \lim_{\lambda\downarrow 0}\left\{\sigma_\beta^2\lambda^2 m_\gamma'(-\lambda) + \sigma_\epsilon^2\left(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda)\right)\right\} = \sigma_\beta^2\left(1-\frac{1}{\gamma}\right) + \sigma_\epsilon^2\frac{1}{\gamma(\gamma-1)}\,. \tag{G.3}
$$

We next prove that:

$$
|\Psi(\lambda) - \Psi(0)| \le \frac{\sigma_\beta^2\lambda^2 + (2\sigma_\epsilon^2/\gamma)\lambda}{(1-\sqrt{\gamma})^4}\,. \tag{G.4}
$$

Towards that end, recall that for $\gamma > 1$, the MP-law has an atom of mass $1-\gamma^{-1}$ at 0 and a continuous part supported on $[a,b]$, where $a = (1-\sqrt{\gamma})^2 > 0$ and $b = (1+\sqrt{\gamma})^2$. By the definition of $m_\gamma(z)$, we have:

$$
m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda) = \frac{1}{\gamma}\int_a^b \frac{x}{(x+\lambda)^2}\,dF_\gamma(x) \xrightarrow{\lambda\downarrow 0} \frac{1}{\gamma}\int_a^b \frac{1}{x}\,dF_\gamma(x) = \frac{1}{\gamma(\gamma-1)}\,.
$$

Therefore,

$$
\begin{aligned}
\left|m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda) - \frac{1}{\gamma(\gamma-1)}\right| &= \frac{1}{\gamma}\left|\int_a^b\left(\frac{x}{(x+\lambda)^2} - \frac{1}{x}\right)dF_\gamma(x)\right| \\
&= \frac{1}{\gamma}\int_a^b \frac{\lambda(2x+\lambda)}{x(x+\lambda)^2}\,dF_\gamma(x) \\
&\le \sup_{x\in[a,b]}\frac{\lambda(2x+\lambda)}{\gamma x(x+\lambda)^2} \\
&\le \sup_{x\in[a,b]}\frac{2\lambda}{\gamma x(x+\lambda)} = \frac{2\lambda}{\gamma a^2} = \frac{2\lambda}{\gamma(1-\sqrt{\gamma})^4}\,.
\end{aligned}
$$

For the other part in $\Psi(\lambda)$, we have:

$$
\lambda^2 m_\gamma'(-\lambda) = \left(1-\frac{1}{\gamma}\right) + \frac{1}{\gamma}\int_a^b \frac{\lambda^2}{(x+\lambda)^2}\,dF_\gamma(x) \xrightarrow{\lambda\downarrow 0} \left(1-\frac{1}{\gamma}\right)\,.
$$

Furthermore,

$$
\begin{aligned}
\left|\lambda^2 m_\gamma'(-\lambda) - \left(1-\frac{1}{\gamma}\right)\right| &= \lambda^2\int_a^b \frac{1}{(x+\lambda)^2}\,dF_{\gamma,\mathrm{cont}}(x) \\
&\le \lambda^2 \sup_{x\in[a,b]}\frac{1}{(x+\lambda)^2} = \left(\frac{\lambda}{a+\lambda}\right)^2 \le \frac{\lambda^2}{a^2} = \frac{\lambda^2}{(1-\sqrt{\gamma})^4}.
\end{aligned}
$$

Combining these two bounds completes the proof of Equation (G.4).

For the rest of the analysis, we choose $\lambda_n = n^{-\frac{1}{2}+\alpha_1}$ for some $\alpha_1 < \alpha_0$. Furthermore, define $\Omega_{n,1}$ to be event such that $s_{\min} \geq (1-\sqrt{\gamma})^2/2 = a/2$. By [4], $\mathbb{P}(\Omega_{n,1})$ goes to 1 and $n \uparrow \infty$. Therefore, on $\Omega_n \cap \Omega_{n,1}$, we have:

$$\|\tilde{\Psi}_n - \Psi(0)\mathbf{I}\|_\infty \leq \frac{C_1}{n^{\frac{1}{2}-\alpha_0}} + \frac{C_2}{n^{\alpha_0-\alpha_1}} := r_n \,.$$

The rest of the proof is same as the previous part; we show that the probability that $\sqrt{n}\|\widehat{\beta} - \beta_0\|_\infty$ is both $\geq \sqrt{2(1+\delta)\Psi(0)\log p}$ and $\leq \sqrt{2(1-\delta)\Psi(0)\log p}$ goes to 0 for any $\delta > 0$ and in the proof we use the fact that

$$|\widehat{\sigma}_j^2 - \Psi(0)| = |\tilde{\Psi}_{n,jj} - \Psi(0)| \leq r_n,$$

and

$$|\widehat{\rho}_{ij}| = \frac{\tilde{\Psi}_{n,ij}}{\widehat{\sigma}_i, \widehat{\sigma}_j} \leq \frac{r_n}{\Psi(0) - r_n} \leq \frac{2r_n}{\Psi(0)} \,.$$

Therefore, replacing $a(n,\lambda)$ with r_n in the proof of the previous part, we complete the claim.

G.2 Limit of OLS when $\gamma < 1$

When $p/n = \gamma < 1$, a natural choice for the estimator of β_0 is the OLS estimator:

$$\widehat{\beta}_0 = \widehat{\Sigma}^{-1}\frac{\mathbf{X}^\top \mathbf{y}}{n} = \beta_0 + \widehat{\Sigma}^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}$$

Therefore,

$$\widehat{\beta} - \beta_0 \mid \mathbf{X} \sim \mathcal{N}(0, \sigma_\epsilon^2 \Sigma^{-1}).$$

Now, by another application of Theorem 3.17 of [24], we have the following event Ω_n, which occurs with probability $\geq 1 - p^2 n^{-D} = 1 - \gamma^2 n^{-(D-2)}$:

$$\Omega_n = \left\{ \mathbf{X} : \left| e_i^\top \widehat{\Sigma}^{-1} e_j - m_\gamma(0)\delta_{ij} \right| \leq \frac{C}{n^{\frac{1}{2}-\alpha_0}}, \quad \forall\, i,j \right\}$$

for any small $\alpha_0 > 0$, and $n \geq N_0 \equiv N_0(\alpha_0, D)$. Note that here we can take $\lambda = 0$ because $\gamma < 1$ and consequently $\widehat{\Sigma}$ is invertible with high probability and its limiting spectral distribution is supported on $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$, which is bounded away from 0. Furthermore, from the standard properties of MP distribution, we know:

$$m_\gamma(0) = \int_{(1-\sqrt{\gamma})^2}^{(1+\sqrt{\gamma})^2} \frac{1}{x}\, dF_\gamma(x) = \frac{1}{1-\gamma} \,.$$

Therefore, on Ω_n, we have:

$$\left| e_i^\top \widehat{\Sigma}^{-1} e_j - \frac{1}{1-\gamma}\delta_{ij} \right| \leq \frac{C}{n^{\frac{1}{2}-\alpha_0}}, \quad \forall\, i,j \,.$$

The rest of the proof is exactly the same as the previous one, by setting $\Psi_\gamma(0) = \sigma_\epsilon^2(1-\gamma)^{-1}$. This completes the proof. □

H Proof of Theorem A.3

Proof. Before going into the proof, let us first quickly recall the definition of a regularly varying function and a slowly varying function from Section 0.4.1 of [29]. A function U is said to be regularly varying at ∞ with index α, denoted at $U \in \mathrm{RV}_\alpha$, if for $x > 0$:

$$\lim_{t\to\infty} \frac{U(tx)}{U(x)} = x^\alpha \,.$$

Furthermore, U is said to be slowly varying if $U \in \mathrm{RV}_0$. Now, by Gnedenko's theorem [18] given any distribution F, it belongs to the domain of attraction of Fréchet distribution $\Phi_\alpha(x) = \mathsf{exp}(-x^{-\alpha})$, iff $1 - F(x) \in \mathrm{RV}_{-\alpha}$ (Proposition 1.11 of [29]). As a consequence, we know $1 - F(x)$ must satisfy:

$$\lim_{x\to\infty} \frac{1-F(x)}{x^{-\alpha}L(x)} = 1 \tag{H.1}$$

for some slowly varying function $L(x)$. Furthermore, by Proposition 1.11 of [29], we also have that if $X_1, \ldots, X_p \overset{\text{i.i.d.}}{\sim} F$, then

$$\mathbb{P}\left(\frac{\max_{1\le i\le p}|X_i|}{a_p} \le t\right) \to e^{-t^{-\alpha}}. \tag{H.2}$$

where $a_p = (1/(1-F))^{\leftarrow}(p)$. Now using Equation (H.1), we have:

$$1 - F(a_p) \sim \frac{1}{p} \implies a_p^{-\alpha} L(a_p) \sim \frac{1}{p} \implies a_p \sim p^{\frac{1}{\alpha}} \ell(p)$$

for some slowly varying sequence $\ell(p)$. As a consequence, we can conclude from Equation (H.2):

$$\mathbb{P}\left(\frac{\max_{1\le i\le p}|X_i|}{p^{\frac{1}{\alpha}}\ell(p)} \le t\right) \to e^{-t^{-\alpha}}. \tag{H.3}$$

Now coming back to the difference $\widehat{\beta}_\lambda - \beta_0$:

$$\begin{aligned}
\widehat{\beta}_\lambda - \beta_0 &= -\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\beta_0 + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n} \\
&= \frac{1}{\sqrt{n}}\left\{-\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\sigma_\beta \boldsymbol{\mu} + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}}\right\} \\
&= \frac{1}{\sqrt{n}}\left\{-\lambda(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\sigma_\beta \boldsymbol{\mu} + \lambda m(-\lambda)\boldsymbol{\mu} - \lambda m(-\lambda)\boldsymbol{\mu} + (\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}}\right\}
\end{aligned}$$

where $\mu_i \overset{\text{i.i.d.}}{\sim} t_\alpha$ and independent of (X_i, ϵ_i). Therefore,

$$\frac{n^{\frac{1}{2}-\frac{1}{\alpha}}}{\ell(n)}(\widehat{\beta}_\lambda - \beta_0) = \frac{-\lambda\left((\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - m(-\lambda)\mathbf{I}\right)\sigma_\beta \boldsymbol{\mu}}{n^{\frac{1}{\alpha}}\ell(n)} + \frac{-\lambda m(-\lambda)\sigma_\beta \boldsymbol{\mu}}{n^{\frac{1}{\alpha}}\ell(n)} + \frac{(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}}}{n^{\frac{1}{\alpha}}\ell(n)} := T_1 + T_2 + T_3\,.$$

Now, for the result of the maxima of i.i.d. t_α distributed random variable, as stated in Equation (H.3), we have:

$$\mathbb{P}(\|T_2\|_\infty \le t) \longrightarrow e^{-\gamma\left(\frac{t}{\lambda m(-\lambda)\sigma_\beta}\right)^{-\alpha}}$$

where we use the fact that $\ell(p)/\ell(n) \to 1$ as $p/n \to \gamma \in (0,1)$, as $\ell(n)$ is a slowly varying sequence. Furthermore, it follows from the proof of Theorem 3.1 that:

$$\left\|(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}\frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{\sqrt{n}}\right\|_\infty = O_p(\sqrt{\log p})$$

which implies $\|T_3\|_\infty = o_p(1)$. Finally, to analyze T_1, we further divide it into two parts:

$$\begin{aligned}
T_1 &= \frac{-\lambda\left((\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - m(-\lambda)\mathbf{I}\right)\sigma_\beta \boldsymbol{\mu}}{n^{\frac{1}{\alpha}}} \\
&= \frac{-\lambda\left((\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - \mathbb{E}[(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}]\right)\sigma_\beta \boldsymbol{\mu}}{n^{\frac{1}{\alpha}}} + \frac{-\lambda\left(\mathbb{E}[(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}] - m(-\lambda)\mathbf{I}\right)\sigma_\beta \boldsymbol{\mu}}{n^{\frac{1}{\alpha}}} \\
&= T_{11} + T_{12}
\end{aligned}$$

Following Proposition 3.9 of [20], we have:

$$|T_{12,j}| = \frac{\lambda\sigma_\beta}{n^{\frac{1}{\alpha}}}\left|e_j^\top\left(\mathbb{E}[(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}] - m(-\lambda)\mathbf{I}\right)\boldsymbol{\mu}\right| \le \frac{\lambda\sigma_\beta}{n^{\frac{1}{\alpha}}}\frac{c\|\boldsymbol{\mu}\|_2}{n\lambda^k} = \frac{C\sigma_\beta}{n^{\frac{1}{\alpha}+\frac{1}{2}}\lambda^{k-1}}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{n}}$$

Consequently,

$$\|T_{12}\|_\infty \le \frac{C\sigma_\beta}{n^{\frac{1}{\alpha}+\frac{1}{2}}\lambda^{k-1}}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{n}} = O_p(n^{-\frac{1}{\alpha}-\frac{1}{2}}) = o_p(1)\,.$$

where we have used the fact that $\|\boldsymbol{\mu}\|_2/\sqrt{n} = O_p(1)$. Now, for T_{11}, we use the fact that

$$\begin{aligned}
|T_{11,j}| &= \frac{\lambda\sigma_\beta}{n^{\frac{1}{\alpha}}}\sqrt{n}\left|e_j^\top\left\{(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - \mathbb{E}[(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}]\right\}\frac{\boldsymbol{\mu}}{\sqrt{n}}\right| \\
&= \frac{\lambda\sigma_\beta}{n^{\frac{1}{\alpha}}}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{n}}\sqrt{n}\left|e_j^\top\left\{(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1} - \mathbb{E}[(\widehat{\Sigma} + \lambda \mathbf{I}_p)^{-1}]\right\}\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right|
\end{aligned}$$

We now use the sub-exponential concentration of the centered bilinear form $T_{11,j}$, which essentially follows from Proposition 3.7 of [20]. Let $K = C(\alpha/(\alpha-2))$ for some $C > \gamma$, where, recall that $\alpha/(\alpha-2)$ is the variance of t_α distribution.

$$
\begin{aligned}
&\mathbb{P}(\|T_{11}\|_\infty \geq t) \\
&\leq \mathbb{P}\left(\|T_1\|_\infty \geq t, \sqrt{\gamma}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{p}} \leq K\right) + \mathbb{P}\left(\sqrt{\gamma}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{p}} > K\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(|T_{11,j}| \geq t, \sqrt{\gamma}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{p}} \leq K\right) + \mathbb{P}\left(\frac{1}{p}\sum_j \mu_j^2 > \frac{C}{\gamma}\frac{\alpha}{\alpha-2}\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{n}}\sqrt{n}\left|e_j^\top\left\{(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1} - \mathbb{E}[(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}]\right\}\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right| \geq \frac{tn^{\frac{1}{\alpha}}}{\lambda\sigma_\beta}, \sqrt{\gamma}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{p}} \leq K\right) \\
&\qquad\qquad + \mathbb{P}\left(\frac{1}{p}\sum_j \mu_j^2 > \frac{C}{\gamma}\frac{\alpha}{\alpha-2}\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(K\sqrt{n}\left|e_j^\top\left\{(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1} - \mathbb{E}[(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}]\right\}\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right| \geq \frac{tn^{\frac{1}{\alpha}}}{\lambda\sigma_\beta}, \sqrt{\gamma}\frac{\|\boldsymbol{\mu}\|_2}{\sqrt{p}} \leq K\right) + \mathbb{P}\left(\frac{1}{p}\sum_j \mu_j^2 > \frac{C}{\gamma}\frac{\alpha}{\alpha-2}\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left(\sqrt{n}\left|e_j^\top\left\{(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1} - \mathbb{E}[(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}]\right\}\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right| \geq \frac{tn^{\frac{1}{\alpha}}}{K\lambda\sigma_\beta}\right) + \mathbb{P}\left(\frac{1}{p}\sum_j \mu_j^2 > \frac{C}{\gamma}\frac{\alpha}{\alpha-2}\right) \\
&\leq C\mathsf{exp}\left(\log p - c\frac{tn^{\frac{1}{\alpha}}}{K\lambda\sigma_\beta}\right) + \mathbb{P}\left(\frac{1}{p}\sum_j \mu_j^2 > \frac{C}{\gamma}\frac{\alpha}{\alpha-2}\right) \\
&\longrightarrow 0, \quad \text{as } n \uparrow \infty\,.
\end{aligned}
$$

The second summand goes to 0 by SLLN. Therefore, $\|T_{11}\|_\infty = o_p(1)$. This completes the proof.

□

I Proof of Auxiliary Lemmas of Theorem 3.3

I.1 Proof of Lemma 1

Proof. Fix $\lambda > 0$ and write, for simplicity,

$$m = m_\gamma(-\lambda), \qquad \widetilde{m} = \widetilde{m}_\gamma(-\lambda),$$

where

$$m_\gamma(-\lambda) = \int \frac{1}{x+\lambda}\, dF_\gamma(x), \qquad \widetilde{m}_\gamma(-\lambda) = \int \frac{1}{x+\lambda}\, d\widetilde{F}_\gamma(x).$$

Here F_γ denotes the Marchenko–Pastur law associated with $p^{-1}\operatorname{tr}(\widehat{\Sigma}+\lambda\mathbf{I}_p)^{-1}$, while $\widetilde{F}_\gamma$ denotes its companion law associated with $n^{-1}\operatorname{tr}(XX^\top/n+\lambda I_n)^{-1}$. We shall use the standard identity

$$m_\gamma(-\lambda) = \frac{1}{\lambda + \lambda\widetilde{m}_\gamma(-\lambda)}. \tag{I.1}$$

Recall the definition q_j:

$$q_j = \left(\lambda + \frac{x_j^\top A_j x_j}{n}\right)^{-1}, \quad A_j = \lambda\left(\frac{X_{-j}X_{-j}^\top}{n} + \lambda I_n\right)^{-1}.$$

Since the eigenvalues of A_j belong to $[0,1]$, we have

$$\|A_j\|_{o_P} \leq 1, \qquad \|A_j\|_F^2 \leq n.$$

We first prove a uniform concentration bound for the quadratic forms. Since x_j is independent of X_{-j}, conditionally on X_{-j} we may apply the Hanson-Wright inequality [30]. For every $u > 0$,

$$\mathbb{P}\left(\left|\frac{x_j^\top A_j x_j}{n} - \frac{1}{n}\operatorname{tr}(A_j)\right| > u \;\middle|\; X_{-j}\right) \le 2\exp\left\{-cn\min(u^2, u)\right\},$$

where $c > 0$ is a universal constant. Taking

$$u = C\sqrt{\frac{\log p}{n}},$$

and using $\log p/n \to 0$, we obtain, for all sufficiently large n,

$$\mathbb{P}\left(\left|\frac{x_j^\top A_j x_j}{n} - \frac{1}{n}\operatorname{tr}(A_j)\right| > C\sqrt{\frac{\log p}{n}}\right) \le 2p^{-cC^2}.$$

Therefore, by a union bound over $j = 1, \dots, p$,

$$\mathbb{P}\left(\max_{1\le j\le p}\left|\frac{x_j^\top A_j x_j}{n} - \frac{1}{n}\operatorname{tr}(A_j)\right| > C\sqrt{\frac{\log p}{n}}\right) \le 2p^{1-cC^2}.$$

Choosing C sufficiently large gives

$$\max_{1\le j\le p}\left|\frac{x_j^\top A_j x_j}{n} - \frac{1}{n}\operatorname{tr}(A_j)\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right). \tag{I.2}$$

It remains to control the normalized traces uniformly in j. Define

$$A = \lambda\left(\frac{XX^\top}{n} + \lambda I_n\right)^{-1}.$$

Since

$$\frac{XX^\top}{n} = \frac{X_{-j}X_{-j}^\top}{n} + \frac{x_j x_j^\top}{n},$$

the two matrices $XX^\top/n$ and $X_{-j}X_{-j}^\top/n$ differ by a rank-one matrix. The function

$$f(x) = \frac{\lambda}{x+\lambda}$$

is bounded between 0 and 1 and has total variation at most 1 on $[0, \infty)$. Hence, by the rank inequality for empirical spectral distributions (follows from Theorem A.43 of [5]),

$$\left|\frac{1}{n}\operatorname{tr}(A_j) - \frac{1}{n}\operatorname{tr}(A)\right| \le \frac{1}{n}.$$

Consequently,

$$\max_{1\le j\le p}\left|\frac{1}{n}\operatorname{tr}(A_j) - \frac{1}{n}\operatorname{tr}(A)\right| \le \frac{1}{n}. \tag{I.3}$$

By the standard deterministic-equivalent bound for Wishart resolvents at a fixed negative spectral parameter,

$$\frac{1}{n}\operatorname{tr}(A) = \lambda\widetilde{m}_\gamma(-\lambda) + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Combining this with (I.3), we obtain

$$\max_{1\le j\le p}\left|\frac{1}{n}\operatorname{tr}(A_j) - \lambda\widetilde{m}_\gamma(-\lambda)\right| = O_{\mathbb{P}}\left(\frac{1}{n}\right). \tag{I.4}$$

In particular, since $p/n = \gamma$ and hence $\log p \asymp \log n$,

$$\frac{1}{n} = O\left(\sqrt{\frac{\log p}{n}}\right).$$

Combining (I.2) and (I.4) gives

$$\max_{1\le j\le p}\left|\frac{x_j^\top A_j x_j}{n}-\lambda\widetilde{m}_\gamma(-\lambda)\right|=O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right). \tag{I.5}$$

Now define

$$h(u)=\frac{1}{\lambda+u},\qquad u\ge 0.$$

Since $\lambda>0$,

$$|h'(u)|=\frac{1}{(\lambda+u)^2}\le\frac{1}{\lambda^2}.$$

Thus h is Lipschitz on $[0,\infty)$ with Lipschitz constant λ^{-2}. Hence, by (I.5),

$$\max_{1\le j\le p}\left|q_j-\frac{1}{\lambda+\lambda\widetilde{m}_\gamma(-\lambda)}\right|=O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right).$$

Using the identity (I.1), we conclude that

$$\boxed{\max_{1\le j\le p}|q_j-m_\gamma(-\lambda)|=O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right).}$$

This proves the lemma. □

I.2 Proof of Lemma 2

Proof. Let us unpack $\mathbf{c}_j$, we can write it as:

$$\mathbf{c}_j=\lambda\frac{\mathbf{X}_{-j}}{n}\left(\frac{\mathbf{X}_{-j}^\top\mathbf{X}_{-j}}{n}+\lambda\mathbf{I}_{p-1}\right)^{-1}\beta_{0,-j}+\frac{A_j\boldsymbol{\epsilon}}{n}:=\mathbf{s}_j+\boldsymbol{\eta}_j$$

As a consequence, $\|\mathbf{c}_j\|_2^2=\|\mathbf{s}_j\|_2^2+\|\boldsymbol{\eta}_j\|_2^2+2\langle\mathbf{s}_j,\boldsymbol{\eta}_j\rangle$. Let us first consider $\|\mathbf{s}_j\|_2^2$:

$$\begin{aligned}
n\|\mathbf{s}_j\|_2^2&=\lambda^2\beta_{0,-j}^\top\left(\frac{\mathbf{X}_{-j}^\top\mathbf{X}_{-j}}{n}+\lambda\mathbf{I}_{p-1}\right)^{-1}\frac{\mathbf{X}_{-j}^\top\mathbf{X}_{-j}}{n}\left(\frac{\mathbf{X}_{-j}^\top\mathbf{X}_{-j}}{n}+\lambda\mathbf{I}_{p-1}\right)^{-1}\beta_{0,-j}\\
&=-\lambda^2\partial_\omega\beta_{0,-j}^\top(\widehat{\Sigma}_{-j}+\lambda\mathbf{I}_{p-1}+\omega\widehat{\Sigma}_{-j})^{-1}\beta_{0,-j}|_{\omega=0}\\
&=-\lambda^2\partial_\omega\left\{\frac{1}{1+\omega}\beta_{0,-j}^\top(\widehat{\Sigma}_{-j}+(\lambda/(1+\omega))\mathbf{I}_{p-1})^{-1}\beta_{0,-j}\right\}|_{\omega=0}\\
&=-\lambda^2\partial_\omega\left\{\beta_{0,-j}^\top(\widehat{\Sigma}_{-j}+(\lambda/(1+\omega))\mathbf{I}_{p-1})^{-1}\beta_{0,-j}\right\}|_{\omega=0}+\lambda^2\beta_{0,-j}^\top(\widehat{\Sigma}_{-j}+\lambda\mathbf{I}_{p-1})^{-1}\beta_{0,-j}
\end{aligned}$$

Now, by anisotropic local law, we have for any large K and any small $\epsilon_0>0$ and for all $n\ge N(\epsilon_0,K)$:

$$\mathbb{P}\left(\left|\beta_{0,-j}^\top(\widehat{\Sigma}_{-j}+\lambda\mathbf{I}_{p-1})^{-1}\beta_{0,-j}-m(-\lambda)\|\beta_{0,-j}\|_2^2\right|>\frac{C}{p^{\frac12-\epsilon_0}\lambda}\right)\le p^{-K}$$

Similarly, using the same technique as in the proof of Theorem 5 of [22], we have:

$$\mathbb{P}\left(\left|\partial_\omega\left\{\beta_{0,-j}^\top(\widehat{\Sigma}_{-j}+(\lambda/(1+\omega))\mathbf{I})^{-1}\beta_{0,-j}\right\}|_{\omega=0}-\lambda m'(-\lambda)\|\beta_{0,-j}\|_2^2\right|>\frac{C}{p^{\frac12-\epsilon_0}\lambda}\right)\le p^{-K}.$$

Therefore, there exists an event $\Omega_{n,1}$ with probability $\ge 1-2p^{-K}$, such that on $\Omega_{n,1}$:

$$\left|n\|\mathbf{s}_j\|_2^2-\lambda^2\left(m(-\lambda)-\lambda m'(-\lambda)\right)\|\beta_{0,-j}\|_2^2\right|\le\frac{2C}{p^{\frac12-\epsilon_0}\lambda}.$$

Now, let us consider $n\|\boldsymbol{\eta}_j\|_2^2=(\boldsymbol{\epsilon}^\top A_j^2\boldsymbol{\epsilon})/n$; given $\epsilon\perp\!\!\!\perp\mathbf{X}$, we have by Hanson-Wright inequality:

$$\mathbb{P}\left(\left|\frac{\boldsymbol{\epsilon}^\top A_j^2\boldsymbol{\epsilon}}{n}-\frac{\sigma_\epsilon^2\mathsf{tr}(A_j^2)}{n}\right|>t\mid\mathbf{X}\right)\le 2\mathsf{exp}\left(-\min\left\{\frac{n^2t^2}{\sigma_\epsilon^4\|A_j\|_F^2},\frac{nt}{\sigma_\epsilon^2\|A\|_{\mathrm{op}}}\right\}\right)$$

Now, let us recall the definition of A_j from Equation (D.2). It is immediate that all the eigenvalues of A_j lie in $[0,1]$. Therefore, we have $\|A\|_{\text{op}} \leq 1$ and $\|A\|_F^2 \leq n$ almost surely for all $\mathbf{X}$. Therefore, we have:

$$\mathbb{P}\left(\left|\frac{\boldsymbol{\epsilon}^\top A_j^2 \boldsymbol{\epsilon}}{n} - \frac{\sigma_\epsilon^2 \mathsf{tr}(A_j^2)}{n}\right| > t \mid \mathbf{X}\right) \leq 2\mathsf{exp}\left(-\min\left\{\frac{nt^2}{\sigma_\epsilon^4}, \frac{nt}{\sigma_\epsilon^2}\right\}\right)$$

As the right hand side does not depend on $\mathbf{X}$, we can conclude this unconditionally as well. Therefore, we have choose $t = \sqrt{C\log p/n}$ for some large enough constant C such that:

$$\mathbb{P}\left(\left|\frac{\boldsymbol{\epsilon}^\top A_j^2 \boldsymbol{\epsilon}}{n} - \frac{\sigma_\epsilon^2 \mathsf{tr}(A_j^2)}{n}\right| > \sqrt{\frac{C\log p}{n}} \mid \mathbf{X}\right) \leq 2p^{-K}$$

Therefore, there exists an event $\Omega_{n,2}$ with probability $\geq 1 - 2p^{-K}$ such that on $\Omega_{n,2}$:

$$\left|n\|\boldsymbol{\eta}\|_2^2 - \frac{\sigma_\epsilon^2 \mathsf{tr}(A_j^2)}{n}\right| \leq \sqrt{\frac{C\log p}{n}}$$

Next, we need the concentration of $\mathsf{tr}(A_j^2)/n$ to its deterministic equivalence. From Theorem 3.6 of [24], we have:

$$\mathbb{P}\left(\left|\frac{1}{n}\mathsf{tr}(A_j^2) - \int \frac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_\gamma(x)\right| \geq \frac{C\lambda}{p^{1-\epsilon_0}}\right) \leq p^{-K}\,,$$

where $\widetilde{F}_\gamma$ is the companion of Marchenko-Pasteur distribution, i.e., the LSD of the eigenvalues of $\mathbf{X}_{-j}\mathbf{X}_{-j}^\top/n$. Therefore, there exists an event $\Omega_{n,3}$ with probability $\geq 1 - p^{-K}$ such that, on the event:

$$\left|\frac{1}{n}\mathsf{tr}(A_j^2) - \int \frac{\lambda^2}{(x+\lambda)^2}\, d\widetilde{F}_\gamma(x)\right| \leq \frac{C\lambda}{p^{1-\epsilon_0}}\,.$$

Finally, let us deal with the cross term: $2\mathbf{s}_j^\top \boldsymbol{\eta}_j$, which by definition:

$$2n\mathbf{s}_j^\top \boldsymbol{\eta}_j = 2\lambda\beta_{0,-j}^\top \left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1} \frac{\mathbf{X}_{-j}^\top A_j \boldsymbol{\epsilon}}{n}$$

Therefore, conditional on $\mathbf{X}_{-j}$, $2n\mathbf{s}_j^\top \boldsymbol{\eta}_j$ follows $\mathcal{N}(0, \sigma_j^2)$ where

$$\sigma_j^2 = 4\sigma_\epsilon^2\lambda^2\beta_{0,-j}^\top \left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1} \frac{\mathbf{X}_{-j}^\top A_j^2 \mathbf{X}_{-j}}{n^2} \left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1} \beta_{0,-j}\,.$$

Now, we have already argued that all the eigenvalues of A_j (hence A_j^2) lie within $[0,1]$. Therefore,

$$\sigma_j^2 \leq \frac{4\sigma_\epsilon^2\lambda^2}{n}\beta_{0,-j}^\top \left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1} \frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} \left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1} \beta_{0,-j}$$

Now using the fact that $x/(x+\lambda)^2 \leq 1/(4\lambda)$ for all $x \in \mathbb{R}^+$, we have:

$$\lambda_i\left(\left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1} \frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} \left(\frac{\mathbf{X}_{-j}^\top \mathbf{X}_{-j}}{n} + \lambda \mathbf{I}_{p-1}\right)^{-1}\right) \leq \frac{1}{4\lambda},\ \ \forall\, 1 \leq i \leq p-1\,.$$

Therefore,

$$\sigma_j^2 \leq \frac{\sigma_\epsilon^2\lambda\|\beta_{0,-j}\|_2^2}{n}\,.$$

Hence, we have:

$$\mathbb{P}\left(|2\mathbf{s}_j^\top \boldsymbol{\eta}_j| \geq t\right) \leq 2e^{-\frac{nt^2}{2\sigma_\epsilon^2\lambda\|\beta_{0,-j}\|_2^2}}$$

Therefore, choosing $t = \sqrt{C\log p/n}$ for large enough C, we have:

$$\mathbb{P}\left(|2\mathbf{s}_j^\top \boldsymbol{\eta}_j| \geq \sqrt{\frac{C\log p}{n}}\right) \leq 2p^{-K}\,.$$

As a consequence, there is an event $\Omega_{n,4}$ with probability $\geq 1 - 2p^{-K}$ such that on $\Omega_{n,4}$:

$$|2\mathbf{s}_j^\top \boldsymbol{\eta}_j| \leq \sqrt{\frac{C \log p}{n}}\,.$$

Combining our bounds so far, we have on an event $\cap_{j=1}^4 \Omega_{n,j}$, with probability $\geq 1 - 7p^{-K}$:

$$\left|n\|\mathbf{c}_j\|_2^2 - c_\lambda\right| \leq \frac{C_1}{p^{\frac{1}{2}-\epsilon_0}\lambda} + \sqrt{C_2\frac{\log p}{n}} + \frac{C_3\lambda}{p^{1-\epsilon_0}}\,. \tag{I.6}$$

where

$$c_{\lambda,n} = \lambda^2\left(m(-\lambda) - \lambda m'(-\lambda)\right)\|\beta_{0,-j}\|_2^2 + \sigma_\epsilon^2\int \frac{\lambda^2}{(x+\lambda)^2}\,d\widetilde{F}_\gamma(x) \tag{I.7}$$

Finally, taking the maximum over j, a simple union bound yields that with probability $\geq 1 - 7p^{-K+1}$:

$$\max_{1\leq j\leq p}\left|n\|\mathbf{c}_j\|_2^2 - c_{\lambda,n}\right| \leq \frac{C_1}{p^{\frac{1}{2}-\epsilon_0}\lambda} + \sqrt{C_2\frac{\log p}{n}} + \frac{C_3\lambda}{p^{1-\epsilon_0}} = \delta_{n,\mathbf{c}}\,. \tag{I.8}$$

□

I.3 Proof of Lemma 3

Proof. We will prove it in two steps: upper bound and lower bound.

Step I: In this step, we will show that for any fixed $\eta > 0$:

$$\mathbb{P}\left(\frac{\max_{1\leq j\leq p}\mathbf{x}_j^\top\mathbf{c}_j}{\sqrt{\frac{2c_\lambda\log p}{n}}} > \sqrt{1+\eta}\right) \longrightarrow 0\,. \tag{I.9}$$

This essentially follows from a union bound:

$$\begin{aligned}
\mathbb{P}\left(\frac{\max_{1\leq j\leq p}\mathbf{x}_j^\top\mathbf{c}_j}{\sqrt{\frac{2c_\lambda\log p}{n}}} > \sqrt{1+\eta}\right) &\leq \sum_j \mathbb{P}\left(\frac{\mathbf{x}_j^\top\mathbf{c}_j}{\|\mathbf{c}_j\|} > \sqrt{\frac{2c_\lambda(1+\eta)\log p}{n\|\mathbf{c}_j\|_2^2}}\right)\\
&\leq \sum_j \mathbb{P}\left(\frac{\mathbf{x}_j^\top\mathbf{c}_j}{\|\mathbf{c}_j\|} > \sqrt{\frac{2c_\lambda(1+\eta)\log p}{n\|\mathbf{c}_j\|_2^2}}, \mathcal{E}_j\right) + \sum_j \mathbb{P}(\mathcal{E}_j^c)\\
&\leq \sum_j \mathbb{P}\left(\frac{\mathbf{x}_j^\top\mathbf{c}_j}{\|\mathbf{c}_j\|} > \sqrt{\frac{2(1+\eta)\log p}{1+\rho_n}}, \mathcal{E}_j\right) \sum_j \mathbb{P}(\mathcal{E}_j^c)\\
&\leq p\bar{\Phi}\left(\sqrt{\frac{2(1+\eta)\log p}{1+\rho_n}}\right) + cp^{-(K-1)}\\
&\leq p\mathsf{exp}\left(-\frac{(1+\eta)\log p}{1+\rho_n}\right) + cp^{-(K-1)}\\
&= p^{1-\frac{1+\eta}{1+\rho_n}} + cp^{-(K-1)} = p^{-\frac{\eta-\rho_n}{1+\rho_n}} + cp^{-(K-1)} \longrightarrow 0\,.
\end{aligned}$$

Step II: In this step, we will establish the lower bound, i.e., we show that for any fixed $\eta > 0$:

$$\mathbb{P}\left(\frac{\max_{1\leq j\leq p}\mathbf{x}_j^\top\mathbf{c}_j}{\sqrt{\frac{2c_\lambda\log p}{n}}} > \sqrt{1-\eta}\right) \longrightarrow 1\,. \tag{I.10}$$

To prove this, we will use the Paley-Zygmund inequality and the Mill's ratio. Define random variables $\{N_j\}_{j=1}^p$ and N_+ as:

$$N_j = \mathbb{1}\left(\mathbf{x}_j^\top\mathbf{c}_j > \sqrt{\frac{2(1-\eta)c_\lambda\log p}{n}}\right), \qquad N_+ = \sum_{j=1}^p N_j\,.$$

Therefore,

$$\mathbb{P}\left(\frac{\max_{1\le j\le p}\mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{\frac{2c_\lambda \log p}{n}}} > \sqrt{1-\eta}\right) = \mathbb{P}(N_+ > 0) \ge \frac{(\mathbb{E}[N_+])^2}{\mathbb{E}[N_+^2]} = \frac{(\mathbb{E}[N_+])^2}{\mathbb{E}[N_+] + \sum_{j\neq k}\mathbb{E}[N_j N_k]}$$

where the inequality follows from Paley-Zigmund inequality. We now complete Step II. First, we will show that $\mathbb{E}[N_+] \to \infty$ as $p \uparrow \infty$, and

$$\sum_{j\neq k}\mathbb{E}[N_j N_k] \le (1+o(1))(\mathbb{E}N_+)^2.$$

These two statements will immediately imply $\mathbb{P}(N_+ > 0) \to 1$. Towards that end, define

$$z_n := \sqrt{2(1-\eta)\log p}, \qquad q_n := \bar{\Phi}(z_n),$$

where $\bar{\Phi}(t) = \mathbb{P}(Z > t)$, $Z \sim N(0,1)$. We shall use the following elementary form of Mills' ratio: for all $z > 0$,

$$\frac{\phi(z)}{z + z^{-1}} \le \bar{\Phi}(z) \le \frac{\phi(z)}{z}, \tag{1.11}$$

where $\phi(z) = (2\pi)^{-1/2}\mathsf{exp}(-z^2/2)$. Since $z_n \ge 1$ for all large n, this gives

$$\frac{1}{2\sqrt{2\pi}z_n}p^{-(1-\eta)} \le q_n \le \frac{1}{\sqrt{2\pi}z_n}p^{-(1-\eta)}. \tag{1.12}$$

In particular,

$$pq_n \ge \frac{1}{2\sqrt{2\pi}z_n}p^{\eta} \longrightarrow \infty. \tag{1.13}$$

Let us define an event $\mathcal{E}_j$ as:

$$\mathcal{E}_j = \left|n\|\mathbf{c}_j\|_2^2 - c_\lambda\right| \le \rho_n c_\lambda, \quad \rho_n = \frac{\delta_{\mathbf{c},n} + \frac{C\log p}{n}}{c_\lambda}.$$

Now, from Lemma 2, we have $\mathcal{E}_j$ happens with probability $\ge 1 - cp^{-K}$ for large K. Now, going back to $\mathbb{E}[N_+]$:

$$\begin{aligned}
\mathbb{E}[N_+] = \sum_{j=1}^p \mathbb{P}\left(x_j^\top \mathbf{c}_j > z_n\sqrt{\frac{c_\lambda}{n}}\right) &= \sum_{j=1}^p \mathbb{P}\left(Z > z_n\sqrt{\frac{c_\lambda}{n\|\mathbf{c}_j\|_2^2}}\right) \\
&= \sum_j \mathbb{E}\left[\mathbb{P}\left(Z > z_n\sqrt{\frac{c_\lambda}{n\|\mathbf{c}_j\|_2^2}} \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon}\right)\right] \\
&\ge \sum_j \mathbb{E}\left[\mathbb{P}\left(Z > z_n\sqrt{\frac{c_\lambda}{n\|\mathbf{c}_j\|_2^2}} \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon}\right)\mathbb{1}_{\mathcal{E}_j}\right] \\
&\ge \sum_j \mathbb{E}\left[\mathbb{P}\left(Z > z_n\sqrt{\frac{1}{1-\rho_n}} \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon}\right)\mathbb{1}_{\mathcal{E}_j}\right] \\
&= \bar{\Phi}\left(z_n\sqrt{\frac{1}{1-\rho_n}}\right)\sum_j \mathbb{P}(\mathcal{E}_j) \\
&:= \bar{\Phi}\left(a_n\right)\sum_j \mathbb{P}(\mathcal{E}_j) \\
&\ge p(1-cp^{-K})\frac{\phi(a_n)}{a_n + a_n^{-1}} = p(1-cp^{-K})\frac{e^{-\frac{a_n^2}{2}}}{\sqrt{2\pi}(a_n + a_n^{-1})}.
\end{aligned}$$

In particular, since

$$\frac{a_n^2}{2} = \frac{(1-\eta)\log p}{1-\rho_n},$$

we have

$$\mathbb{E}[N_+] \geq \frac{1 - cp^{-K}}{\sqrt{2\pi}(a_n + a_n^{-1})} p^{1-(1-\eta)/(1-\rho_n)}.$$

Because $\rho_n \to 0$, for all sufficiently large n,

$$1 - \frac{1-\eta}{1-\rho_n} = \frac{\eta - \rho_n}{1-\rho_n} \geq \frac{\eta}{2}.$$

Also $a_n \asymp \sqrt{\log p}$. Hence

$$\mathbb{E}[N_+] \to \infty. \tag{1.14}$$

On the other hand, we have the following upper bound on $\mathbb{E}[N_+]$:

$$\begin{aligned}
\mathbb{E}[N_+] &\leq \sum_j \mathbb{E}\left[\mathbb{P}\left(Z > z_n \sqrt{\frac{c_\lambda}{n\|\mathbf{c}_j\|_2^2}} \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon}\right) \mathbb{1}_{\mathcal{E}_j}\right] + \sum_j \mathbb{P}(\mathcal{E}_j^c) \\
&\leq \bar{\Phi}\left(z_n \sqrt{\frac{1}{1+\rho_n}}\right) \sum_j \mathbb{P}(\mathcal{E}_j) + \sum_j \mathbb{P}(\mathcal{E}_j^c) \\
&\leq p \,\mathsf{exp}\left(-\frac{(1-\eta)\log p}{1+\rho_n}\right) + cp^{-(K-1)}.
\end{aligned}$$

Hence, we obtain the following two sided bound for $\mathbb{E}[N_+]$:

$$p(1-o(1)) \frac{\mathsf{exp}\left(-\frac{(1-\eta)\log p}{1-\rho_n}\right)}{2\sqrt{2\pi}\sqrt{\frac{2(1-\eta)\log p}{1-\rho_n}}} \leq \mathbb{E}[N_+] \leq p \,\mathsf{exp}\left(-\frac{(1-\eta)\log p}{1+\rho_n}\right) + cp^{-(K-1)}. \tag{I.11}$$

It remains to control the off-diagonal term. Now, observe that eahc off-diagonal term is of the form

$$\mathbb{P}(N_j = 1, N_k = 1) = \mathbb{P}\left(\mathbf{x}_j^\top \mathbf{c}_j > z_n \sqrt{\frac{c_\lambda}{n}},\ \mathbf{x}_k^\top \mathbf{c}_k > z_n \sqrt{\frac{c_\lambda}{n}}\right)$$

These two terms are correlated to $\mathbf{c}_j$ and $\mathbf{c}_k$ who shares $(\mathbf{x}_j, \mathbf{x}_k)$. Therefore, to decorrelate them, we need a leave-two-out approach. Fix $j \neq k$, and let

$$\mathbf{W}_{jk} := \mathbf{X}_{-(j,k)}, \qquad b_{jk} := \beta_{0,-(j,k)}.$$

Define

$$A_{jk} := \lambda \left(\frac{\mathbf{W}_{jk}\mathbf{W}_{jk}^\top}{n} + \lambda I_n\right)^{-1},$$

and

$$\mathbf{c}_{jk} := \lambda \frac{\mathbf{W}_{jk}}{n} \left(\frac{\mathbf{W}_{jk}^\top \mathbf{W}_{jk}}{n} + \lambda \mathbf{I}_{p-2}\right)^{-1} b_{jk} + \frac{A_{jk}\boldsymbol{\epsilon}}{n} = \frac{A_{jk}}{n}\left(\mathbf{W}_{jk} b_{jk} + \epsilon\right).$$

Then $\mathbf{c}_{jk}$ is measurable with respect to $\mathcal{F}_{jk} := \sigma(\mathbf{W}_{jk}, \boldsymbol{\epsilon})$, and $\mathbf{x}_j, \mathbf{x}_k$ are independent of $\mathcal{F}_{jk}$. Moreover, conditionally on $\mathcal{F}_{jk}$, $\mathbf{x}_j^\top \mathbf{c}_{jk}$ and $\mathbf{x}_k^\top c_{jk}$ are independent centered Gaussian random variables with common conditional variance $\|\mathbf{c}_{jk}\|_2^2$. We use the following quantitative leave-two-out estimate. By the same calculation as in Lemma 2, along with a union bound, we have, with probability at least $1 - cp^{-(K-2)}$, uniformly over $j \neq k$,

$$\left|n\|\mathbf{c}_{jk}\|_2^2 - c_\lambda\right| \leq \rho_n c_\lambda, .$$

Let us denote the above event $\mathcal{E}_{jk}^{\mathrm{var}}$. For the rest of the proof, we need one additional lemma:

Lemma 4. *Under Assumption 1-3, there exists deterministic $\rho_n \downarrow 0$ and some fixed constant $C > 0$ satisfying $\rho_n \log p \to 0$,*

$$\max_{j \neq k} \left\{|\mathbf{x}_j^\top (\mathbf{c}_j - \mathbf{c}_{jk})| + |\mathbf{x}_k^\top (\mathbf{c}_k - \mathbf{c}_{jk})|\right\} \leq \rho_n \sqrt{\frac{C \log p}{n}}$$

with probability $\geq 1 - cp^{-K}$ for large K.

The proof is deferred to the end of the section. We denote the above event by $\mathcal{E}_{jk}^{\text{pair}}$. We have the following upper bound:

$$\mathbb{P}(N_j = N_k = 1) \leq \mathbb{P}(N_j = N_k = 1, \mathcal{E}_{jk}^{\text{var}}, \mathcal{E}_{jk}^{\text{pair}}) + \mathbb{P}\big((\mathcal{E}_{jk}^{\text{var}})^c\big) + \mathbb{P}\big((\mathcal{E}_{jk}^{\text{pair}})^c\big).$$

On $\mathcal{E}_{jk}^{\text{pair}}$,

$$\{N_j = N_k = 1\} \subseteq \left\{x_j^\top c_{jk} > r_{n,\lambda}(\sqrt{1-\eta} - \rho_n),\, x_k^\top c_{jk} > r_{n,\lambda}(\sqrt{1-\eta} - \rho_n)\right\}.$$

Hence

$$\mathbb{P}(N_j = N_k = 1) \leq \mathbb{P}\Big(x_j^\top c_{jk} > r_{n,\lambda}(\sqrt{1-\eta} - \rho_n),\, x_k^\top c_{jk} > r_{n,\lambda}(\sqrt{1-\eta} - \rho_n),\, \mathcal{E}_{jk}^{\text{var}}\Big) + \mathbb{P}\big((\mathcal{E}_{jk}^{\text{var}})^c\big) + \mathbb{P}\big((\mathcal{E}_{jk}^{\text{pair}})^c\big).$$

where

$$r_{n,\lambda} = \sqrt{\frac{2c_\lambda \log p}{n}}.$$

Hence, we have:

$$\begin{aligned}
&\mathbb{P}\left(N_j = 1, N_k = 1\right) \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\mathbf{x}_j^\top \mathbf{c}_{jk} \geq r_{n,\lambda}(\sqrt{1-\eta} - \rho_n),\, \mathbf{x}_k^\top \mathbf{c}_{jk} \geq r_{n,\lambda}(\sqrt{1-\eta} - \rho_n) \mid \mathcal{F}_{jk}\right) \mathbb{1}_{\mathcal{E}_{jk}^{\text{var}}}\right] + cp^{-K} \\
&= \mathbb{E}\left[\mathbb{P}\left(Z_j \geq \frac{r_{n,\lambda}}{\|\mathbf{c}_{jk}\|}(\sqrt{1-\eta} - \rho_n),\, Z_k \geq \frac{r_{n,\lambda}}{\|\mathbf{c}_{jk}\|}(\sqrt{1-\eta} - \rho_n) \mid \mathcal{F}_{jk}\right) \mathbb{1}_{\mathcal{E}_{jk}}\right] + cp^{-K} \\
&\leq \mathbb{E}\left[\mathbb{P}\left(Z_j \geq \sqrt{\frac{2\log p}{1+\rho_n}}(\sqrt{1-\eta} - \rho_n),\, Z_k \geq \sqrt{\frac{2\log p}{1+\rho_n}}(\sqrt{1-\eta} - \rho_n) \mid \mathcal{F}_{jk}\right) \mathbb{1}_{\mathcal{E}_{jk}}\right] + cp^{-K} \\
&= \bar{\Phi}^2\left(\sqrt{\frac{2\log p}{1+\rho_n}}(\sqrt{1-\eta} - \rho_n)\right) \mathbb{P}(\mathcal{E}_{jk}) + cp^{-K} \\
&:= \bar{\Phi}^2\left(b_n\right) \mathbb{P}(\mathcal{E}_{jk}) + cp^{-K} \qquad \left[b_n = \sqrt{\frac{2\log p}{1+\rho_n}}(\sqrt{1-\eta} - \rho_n)\right] \\
&\leq \frac{\phi^2(b_n)}{b_n^2} \mathbb{P}(\mathcal{E}_{jk}) + cp^{-K} \\
&\leq \mathsf{exp}\left(-\frac{2\log p}{1+\rho_n}(\sqrt{1-\eta} - \rho_n)^2\right) \mathbb{P}(\mathcal{E}_{jk}) + cp^{-K} \\
&= \frac{1}{2\pi b_n^2} \mathsf{exp}(-b_n^2) + cp^{-K}
\end{aligned}$$

Therefore, summing over all pairs $(j \neq k)$, we have:

$$\sum_{j \neq k} \mathbb{E}[N_j N_k] \leq \frac{p(p-1)}{2\pi b_n^2} \mathsf{exp}(-b_n^2) + cp^{-(K-2)}. \tag{I.12}$$

Combining these two bounds, we have:

$$\frac{\sum_{j \neq k} \mathbb{E}[N_j N_k]}{(\mathbb{E}[N_+])^2} \leq \frac{p(p-1)}{p^2} \frac{(a_n + a_n^{-1})^2}{b_n^2} \mathsf{exp}(a_n^2 - b_n^2) \frac{1}{(1 - cp^{-K})^2} + R_n, \tag{I.13}$$

where

$$R_n := \frac{cp^{2-K}}{p^2(1-cp^{-K})^2\{2\pi(a_n + a_n^{-1})^2\}^{-1} \mathsf{exp}(-a_n^2)}.$$

We first show $R_n \to 0$. Since

$$a_n^2 = \frac{2(1-\eta)\log p}{1-\rho_n},$$

we have

$$\mathsf{exp}(a_n^2) = p^{2(1-\eta)/(1-\rho_n)}.$$

Also

$$(a_n + a_n^{-1})^2 \lesssim \log p.$$

Hence

$$R_n \le C(\log p)p^{-K+2(1-\eta)/(1-\rho_n)}.$$

Choosing K large enough, for instance $K > 3$, gives

$$R_n \to 0. \tag{I.14}$$

It remains to control the first term in (1.17). By definition,

$$a_n^2 = \frac{2(1-\eta)\log p}{1-\rho_n},$$

whereas

$$b_n^2 = \frac{2\log p}{1+\rho_n}(\sqrt{1-\eta} - \rho_n)^2.$$

Therefore,

$$\frac{a_n^2}{b_n^2} = \frac{(1-\eta)(1+\rho_n)}{(1-\rho_n)(\sqrt{1-\eta}-\rho_n)^2} \to 1. \tag{I.15}$$

Since $a_n \to \infty$,

$$\frac{(a_n + a_n^{-1})^2}{b_n^2} = \frac{a_n^2}{b_n^2}(1 + a_n^{-2})^2 \to 1. \tag{I.16}$$

Next,

$$a_n^2 - b_n^2 = 2\log p\left[\frac{1-\eta}{1-\rho_n} - \frac{(\sqrt{1-\eta}-\rho_n)^2}{1+\rho_n}\right] \implies |a_n^2 - b_n^2| \le C_\eta \rho_n \log p,$$

By assumption, $\rho_n \log p \to 0$. Thus

$$\mathsf{exp}(a_n^2 - b_n^2) \to 1. \tag{I.17}$$

Combining Equation (I.13)-(I.17), we obtain

$$\limsup_{n\to\infty} \frac{\sum_{j\neq k}\mathbb{E}[N_j N_k]}{(\mathbb{E}[N_+])^2} \le 1. \tag{I.18}$$

Finally, by Paley–Zygmund,

$$\mathbb{P}(N_+ > 0) \ge \frac{(\mathbb{E}[N_+])^2}{\mathbb{E}[N_+] + \sum_{j\neq k}\mathbb{E}[N_j N_k]} = \frac{1}{(\mathbb{E}[N_+])^{-1} + \frac{\sum_{j\neq k}\mathbb{E}[N_j N_k]}{(\mathbb{E}[N_+])^2}}.$$

We have already established $(\mathbb{E}[N_+])^{-1} \to 0$, and by Equation (I.18),

$$\limsup_{n\to\infty} \frac{\sum_{j\neq k}\mathbb{E}[N_j N_k]}{(\mathbb{E}[N_+])^2} \le 1.$$

Therefore,

$$\liminf_{n\to\infty} \mathbb{P}(N_+ > 0) \ge 1 \implies \mathbb{P}(N_+ > 0) \to 1\,.$$

Consequently,

$$\mathbb{P}\left(\max_{1\le j\le p} x_j^\top c_j > \sqrt{\frac{2(1-\eta)c_\lambda \log p}{n}}\right) \to 1.$$

Equivalently,

$$\mathbb{P}\left(\frac{\max_{1\le j\le p} x_j^\top c_j}{\sqrt{2c_\lambda \log p/n}} > \sqrt{1-\eta}\right) \to 1.$$

This proves Step II and completes the proof. □

I.4 Proof of Corollary D.1

Proof. Let

$$m_n := |T_n|, \qquad \frac{\log m_n}{\log p} \to \rho \in [0,1].$$

We first prove the signed one-sided statement

$$\frac{\max_{j\in T_n} s_j \mathbf{x}_j^\top c_j}{\sqrt{2c_\lambda \log p/n}} \xrightarrow{p} \sqrt{\rho}.$$

The proof is identical to the proof of Lemma 3, with two changes. First, all union bounds and first/second moment calculations are taken over $j \in T_n$ instead of $j \in [p]$. Thus the effective logarithmic size of the maximum is $\log m_n$, not $\log p$. Consequently Lemma 3 gives

$$\frac{\max_{j\in T_n} s_j x_j^\top c_j}{\sqrt{2c_\lambda \log m_n/n}} \xrightarrow{p} 1,$$

provided $m_n \to \infty$. Second, the deterministic signs do not affect the argument. Indeed, conditionally on $\sigma(\mathbf{X}_{-j}, \boldsymbol{\epsilon})$,

$$s_j \mathbf{x}_j^\top \mathbf{c}_j \mid \mathbf{X}_{-j}, \boldsymbol{\epsilon} \sim N(0, \|\mathbf{c}_j\|_2^2),$$

because $s_j \in \{-1, 1\}$ is deterministic and $s_j \mathbf{x}_j \overset{d}{=} \mathbf{x}_j$. Similarly, in the leave-two-out step, conditionally on $\mathcal{F}_{jk} := \sigma(\mathbf{X}_{-(j,k)}, \boldsymbol{\epsilon})$, the two random variables $s_j \mathbf{x}_j^\top \mathbf{c}_{jk}$ and $s_k \mathbf{x}_k^\top \mathbf{c}_{jk}$ are independent centered Gaussians with common variance $\|\mathbf{c}_{jk}\|_2^2$. The leave-two-out replacement bounds are unchanged because

$$|s_j \mathbf{x}_j^\top (\mathbf{c}_j - \mathbf{c}_{jk})| = |\mathbf{x}_j^\top (\mathbf{c}_j - \mathbf{c}_{jk})|.$$

Therefore, the entire proof of Lemma 3 applies verbatim after replacing $[p]$ by T_n and $\log p$ by $\log m_n$. Hence, if $m_n \to \infty$,

$$\frac{\max_{j\in T_n} s_j \mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{2c_\lambda \log p/n}} = \frac{\max_{j\in T_n} s_j \mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{2c_\lambda \log m_n/n}} \sqrt{\frac{\log m_n}{\log p}} \xrightarrow{p} \sqrt{\rho}.$$

If m_n is bounded, or more generally $m_n = p^{o(1)}$, then $\rho = 0$. In this case, for every fixed $\eta > 0$, the same upper-tail union bound used in Lemma 3 gives

$$\mathbb{P}\left(\max_{j\in T_n} s_j x_j^\top c_j > \eta \sqrt{\frac{2c_\lambda \log p}{n}} \right) \le m_n p^{-\eta^2 + o(1)} \to 0.$$

Therefore, again,

$$\frac{\max_{j\in T_n} s_j \mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{2c_\lambda \log p/n}} \xrightarrow{p} 0 = \sqrt{\rho}.$$

This proves the signed statement for all $\rho \in [0,1]$.

It remains to prove the absolute-value statement. The lower bound follows immediately from the one-sided result with $s_j \equiv 1$:

$$\max_{j\in T_n} |\mathbf{x}_j^\top \mathbf{c}_j| \ge \max_{j\in T_n} \mathbf{x}_j^\top \mathbf{c}_j.$$

Therefore

$$\liminf_{n\to\infty} \frac{\max_{j\in T_n} |\mathbf{x}_j^\top \mathbf{c}_j|}{\sqrt{2c_\lambda \log p/n}} \ge \sqrt{\rho} \qquad \text{in probability.}$$

For the upper bound, use the same union-bound part of Lemma 3, but with the two-sided Gaussian tail. Equivalently, for every fixed $\eta > 0$,

$$\mathbb{P}\left(\max_{j\in T_n} |\mathbf{x}_j^\top \mathbf{c}_j| > (1+\eta) \sqrt{\frac{2c_\lambda \log m_n}{n}} \right) \to 0$$

when $m_n \to \infty$. If $m_n = p^{o(1)}$, the same bound with $\log p$ in the threshold gives convergence to zero. Hence

$$\frac{\max_{j \in T_n} |\mathbf{x}_j^\top \mathbf{c}_j|}{\sqrt{2c_\lambda \log m_n/n}} \le 1 + o_p(1),$$

and therefore

$$\limsup_{n\to\infty} \frac{\max_{j \in T_n} |\mathbf{x}_j^\top \mathbf{c}_j|}{\sqrt{2c_\lambda \log p/n}} \le \sqrt{\rho} \qquad \text{in probability.}$$

Combining the lower and upper bounds yields

$$\frac{\max_{j \in T_n} |\mathbf{x}_j^\top \mathbf{c}_j|}{\sqrt{2c_\lambda \log p/n}} \xrightarrow{p} \sqrt{\rho}.$$

Thus, we have proved

$$\frac{\max_{j \in T_n} s_j \mathbf{x}_j^\top \mathbf{c}_j}{\sqrt{2c_\lambda \log p/n}} \xrightarrow{p} \sqrt{\rho}, \qquad \frac{\max_{j \in T_n} |\mathbf{x}_j^\top \mathbf{c}_j|}{\sqrt{2c_\lambda \log p/n}} \xrightarrow{p} \sqrt{\rho}.$$

□

I.5 Proof of Lemma 4

Proof. Fix distinct indices $j \neq k$. Notice that $\mathbf{c}_{jk}$ is measurable with respect to $\mathcal{F}_{jk} = \sigma(\mathbf{W}_{jk}, \boldsymbol{\epsilon})$, and $\mathbf{x}_j, \mathbf{x}_k$ are independent of $\mathcal{F}_{jk}$. We start with the following exact rank-one update formula. Since

$$\mathbf{X}_{-j}\mathbf{X}_{-j}^\top = \mathbf{W}_{jk}\mathbf{W}_{jk}^\top + \mathbf{x}_k\mathbf{x}_k^\top,$$

we have

$$A_j = \lambda \left(\frac{\mathbf{W}_{jk}\mathbf{W}_{jk}^\top}{n} + \frac{\mathbf{x}_k\mathbf{x}_k^\top}{n} + \lambda I_n \right)^{-1}.$$

By the Sherman-Morrison formula,

$$A_j = A_{jk} - q_{k|jk} \frac{A_{jk}\mathbf{x}_k\mathbf{x}_k^\top A_{jk}}{n}, \qquad q_{k|jk} := \left(\lambda + \frac{\mathbf{x}_k^\top A_{jk}\mathbf{x}_k}{n} \right)^{-1}.$$

Since A_{jk} is positive semidefinite, $q_{k|jk} \le \lambda^{-1}$. Defining $r_{jk} := \mathbf{W}_{jk}b_{jk} + \boldsymbol{\epsilon}$, we have $\mathbf{c}_{jk} = (A_{jk}r_{jk})/n$. Also, from the relation between $\mathbf{c}_{jk}$ and $\mathbf{c}_j$, we have:

$$\mathbf{c}_j = \frac{A_j(r_{jk} + \mathbf{x}_k\beta_{0,k})}{n}.$$

Using the Sherman–Morrison expansion,

$$\mathbf{c}_j = \frac{A_{jk}r_{jk}}{n} + \frac{\beta_{0,k}A_{jk}\mathbf{x}_k}{n} - q_{k|jk}\frac{A_{jk}\mathbf{x}_k\mathbf{x}_k^\top A_{jk}r_{jk}}{n^2} - q_{k|jk}\frac{\beta_{0,k}A_{jk}\mathbf{x}_k\mathbf{x}_k^\top A_{jk}\mathbf{x}_k}{n^2}.$$

Since

$$\frac{\mathbf{x}_k^\top A_{jk}r_{jk}}{n} = \mathbf{x}_k^\top \mathbf{c}_{jk}, \qquad \frac{\mathbf{x}_k^\top A_{jk}\mathbf{x}_k}{n} := u_{k|jk},$$

we obtain

$$c_j - c_{jk} = \frac{A_{jk}x_k}{n}\left[\beta_{0,k} - q_{k|jk}x_k^\top c_{jk} - q_{k|jk}\beta_{0,k}u_{k|jk}\right].$$

Because

$$q_{k|jk}(\lambda + u_{k|jk}) = 1,$$

we have

$$1 - q_{k|jk}u_{k|jk} = \lambda q_{k|jk}.$$

Therefore,

$$c_j - c_{jk} = q_{k|jk}\frac{A_{jk}x_k}{n}\left\{\lambda\beta_{0,k} - x_k^\top c_{jk}\right\}. \tag{I.19}$$

Similarly,

$$c_k - c_{jk} = q_{j|jk} \frac{A_{jk} x_j}{n} \left\{ \lambda \beta_{0,j} - x_j^\top c_{jk} \right\}, \tag{I.20}$$

where

$$q_{j|jk} := \left(\lambda + \frac{x_j^\top A_{jk} x_j}{n} \right)^{-1} \leq \lambda^{-1}.$$

Taking inner products in (I.19) and (I.20), we get

$$|x_j^\top (c_j - c_{jk})| \leq \lambda^{-1} \left| \frac{x_j^\top A_{jk} x_k}{n} \right| \left| \lambda \beta_{0,k} - x_k^\top c_{jk} \right|, \tag{I.21}$$

$$|x_k^\top (c_k - c_{jk})| \leq \lambda^{-1} \left| \frac{x_k^\top A_{jk} x_j}{n} \right| \left| \lambda \beta_{0,j} - x_j^\top c_{jk} \right|. \tag{I.22}$$

Thus it remains to prove uniform bounds for the two factors. First, by Assumption 3,

$$\|\beta_0\|_\infty \leq C_\beta \sqrt{\frac{\log p}{n}}$$

for all sufficiently large n. Next, since $0 \preceq A_{jk} \preceq I_n$, we have

$$\|A_{jk}\|_{o_P} \leq 1.$$

Moreover,

$$r_{jk} := W_{jk} b_{jk} + \varepsilon \sim N\left(0, \{\|b_{jk}\|_2^2 + \sigma_\varepsilon^2\} I_n\right),$$

and

$$\sup_{j \neq k} \|b_{jk}\|_2^2 \leq \|\beta_0\|_2^2 = O(1).$$

Therefore, by a standard Gaussian norm concentration bound and a union bound over all pairs (j, k), for every fixed $K > 0$ there exists $C_0 > 0$ such that, with probability at least $1 - cp^{-K}$,

$$\max_{j \neq k} \|c_{jk}\|_2 = \max_{j \neq k} \left\| \frac{A_{jk} r_{jk}}{n} \right\|_2 \leq \frac{C_0}{\sqrt{n}}. \tag{I.23}$$

Indeed, $\|A_{jk} r_{jk}\|_2 \leq \|r_{jk}\|_2$, and $\max_{j \neq k} \|r_{jk}\|_2 = O_{\mathbb{P}}(\sqrt{n})$ with polynomially high probability. We now bound

$$\max_{j \neq k} |x_k^\top c_{jk}|.$$

Conditionally on $\mathcal{F}_{jk}$, x_k is independent of c_{jk} and

$$x_k^\top c_{jk} \mid \mathcal{F}_{jk} \sim N(0, \|c_{jk}\|_2^2).$$

Using (I.23), for any $L > 0$,

$$\mathbb{P}\left(|x_k^\top c_{jk}| > L \sqrt{\frac{\log p}{n}}, \, \|c_{jk}\|_2 \leq \frac{C_0}{\sqrt{n}} \,\middle|\, \mathcal{F}_{jk} \right) \leq 2\exp\left\{ -\frac{L^2 \log p}{2C_0^2} \right\}.$$

Taking a union bound over $j \neq k$ and choosing L sufficiently large gives, with probability at least $1 - cp^{-K}$,

$$\max_{j \neq k} |x_k^\top c_{jk}| \leq L \sqrt{\frac{\log p}{n}}. \tag{I.24}$$

Combining (I.24) with $\|\beta_0\|_\infty \leq C_\beta \sqrt{\log p / n}$, we obtain

$$\max_{j \neq k} \left| \lambda \beta_{0,k} - x_k^\top c_{jk} \right| \leq C_1 \sqrt{\frac{\log p}{n}} \tag{I.25}$$

with probability at least $1 - cp^{-K}$. The same bound holds with j and k interchanged. It remains to control the bilinear term

$$\frac{x_j^\top A_{jk} x_k}{n}.$$

Conditionally on $\mathcal{F}_{jk}$ and x_k,

$$x_j^\top A_{jk} x_k \sim N(0, \|A_{jk}x_k\|_2^2).$$

Since $\|A_{jk}\|_{o_P} \le 1$,

$$\|A_{jk}x_k\|_2 \le \|x_k\|_2.$$

By Gaussian norm concentration and a union bound,

$$\max_{1\le k\le p} \|x_k\|_2 \le C_2\sqrt{n}$$

with probability at least $1 - cp^{-K}$. Therefore, for any $M > 0$,

$$\mathbb{P}\left(\left|\frac{x_j^\top A_{jk} x_k}{n}\right| > M\sqrt{\frac{\log p}{n}}\right) \le 2\exp\left\{-\frac{M^2 \log p}{2C_2^2}\right\} + cp^{-K-2}.$$

Taking a union bound over all $j \ne k$ and choosing M sufficiently large gives

$$\max_{j\ne k}\left|\frac{x_j^\top A_{jk} x_k}{n}\right| \le C_3\sqrt{\frac{\log p}{n}} \tag{I.26}$$

with probability at least $1 - cp^{-K}$. Combining (I.21), (I.22), (I.25), and (I.26), we obtain, with probability at least $1 - cp^{-K}$,

$$\max_{j\ne k} |x_j^\top (c_j - c_{jk})| \le \lambda^{-1} C_1 C_3 \frac{\log p}{n},$$

and similarly,

$$\max_{j\ne k} |x_k^\top (c_k - c_{jk})| \le \lambda^{-1} C_1 C_3 \frac{\log p}{n}.$$

Therefore,

$$\max_{j\ne k} \left\{|x_j^\top (c_j - c_{jk})| + |x_k^\top (c_k - c_{jk})|\right\} \le C_4 \frac{\log p}{n}.$$

Now set

$$\rho_n := \sqrt{\frac{\log p}{n}}.$$

Then

$$C_4 \frac{\log p}{n} = \rho_n\, C_4 \sqrt{\frac{\log p}{n}} = \rho_n \sqrt{\frac{C_4^2 \log p}{n}}.$$

Thus, the desired conclusion holds with $C = C_4^2$. Finally, because $p/n = \gamma \in (0, \infty)$,

$$\rho_n \log p = \sqrt{\frac{\log p}{n}} \log p = \frac{(\log p)^{3/2}}{\sqrt{n}} \to 0.$$

This proves the lemma. □

J Simulation Tables

The code for simulation can be found on Github: https://github.com/zixiaowang17/sup-norm-error

J.1 Double Descent Phenomenon for Random Effects

In the Gaussian random effects regime, next we examine the behavior of the ℓ_∞ risk in the ridgeless limit ($\lambda \to 0$) across varying aspect ratios $\gamma = p/n$. We generate Gaussian signals with $\sigma_\beta \in \{0.25, 0.5, 1, 2\}$ and compute the min-norm interpolator (using $\lambda = 10^{-6}$) for γ ranging from 0.05 to 10.

Figure 4 displays the scaled error $\sqrt{n/(2\log p)}\|\hat{\beta} - \beta_0\|_\infty$ as a function of γ. The characteristic double descent curve is clearly visible: the risk increases as $\gamma \to 1^-$, exhibits a peak near the interpolation threshold $\gamma = 1$, and then decreases in the over-parameterized regime before eventually increasing again for large γ. The empirical results (crosses) closely match the theoretical predictions (solid curves) across all signal strengths, demonstrating that our ℓ_∞ risk analysis captures the double descent phenomenon.

Table 1: Empirical mean / theory ratio ($\lambda = 1$, $s_0 = 1$, $n_{\mathrm{MC}} = 500$) for Theorem 3.1

γ	α	n^α	Ridge	Null	MoM	OLS
0.5	0.12	2.29	1.004	1.000	0.916	0.925
	0.14	2.63	1.004	1.000	0.915	0.925
	0.16	3.02	1.003	1.000	0.916	0.925
	0.18	3.47	1.003	1.000	0.916	0.924
	0.20	3.98	1.003	1.000	0.916	0.924
2.0	0.12	2.29	0.992	1.000	0.930	—
	0.14	2.63	0.993	1.000	0.930	—
	0.16	3.02	0.994	1.000	0.929	—
	0.18	3.47	0.995	1.000	0.929	—
	0.20	3.98	0.995	1.000	0.930	—

Table 2: Empirical mean / theory ratio ($a = 3.5$, $B = 1.2$, $n = 1000$, $n_{\mathrm{MC}} = 500$) for Theorem 3.3.

γ	λ	Ridge	Null	MoM	OLS
0.5	0.001	0.923	1.000	0.919	0.923
	0.023	0.926	1.000	0.919	0.923
	0.187	1.046	1.000	0.919	0.923
	4.329	1.000	1.000	0.919	0.923
	100.000	1.000	1.000	0.919	0.923
2.0	0.001	1.006	1.000	0.927	—
	0.023	1.006	1.000	0.927	—
	0.187	1.005	1.000	0.927	—
	4.329	1.001	1.000	0.927	—
	100.000	1.000	1.000	0.927	—

J.2 Heavy-Tailed Signal

Finally, we verify Theorem A.3 for heavy-tailed signals. We generate $\beta_0 = \sigma_\beta \boldsymbol{\mu}/\sqrt{n}$ where $\mu_j \sim t_\alpha$ i.i.d. with $\alpha = 4$ degrees of freedom and $\sigma_\beta = 2$. We fix $\lambda = 2$ and compute 2000 Monte Carlo replications to ensure stable estimation of the tail behavior. According to Theorem A.3, the scaled error $Z = n^{1/2-1/\alpha}\|\hat{\beta}_\lambda - \beta_0\|_\infty$ should follow a Fréchet distribution with CDF

$$F(t) = \exp\left(-\gamma\left(\frac{t}{c_\alpha^{1/\alpha}\lambda m(-\lambda)\sigma_\beta}\right)^{-\alpha}\right),$$

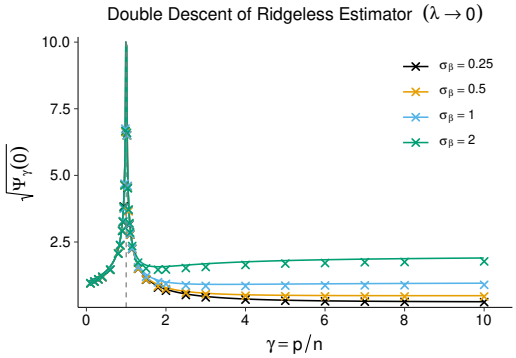
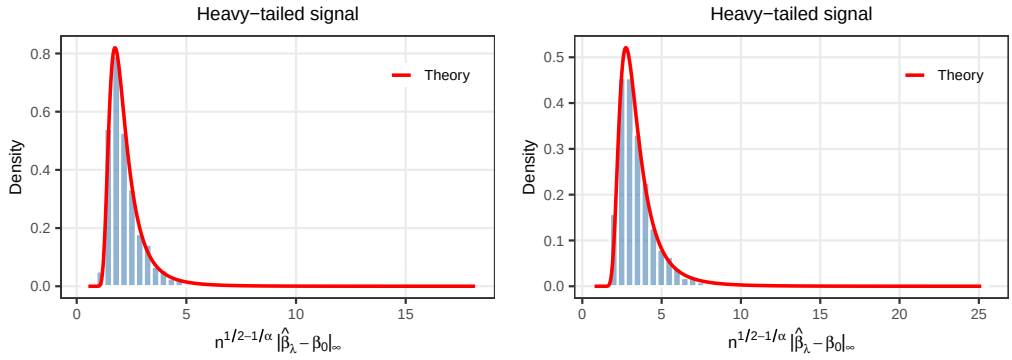


Figure 4: Double descent in ℓ_∞ risk for ridgeless regression from Theorem A.2. The scaled error $\sqrt{n/(2\log p)}\|\hat{\beta} - \beta_0\|_\infty$ is plotted against the aspect ratio $\gamma = p/n$. Solid curves show theoretical predictions; crosses show empirical means. Colors correspond to signal strengths $\sigma_\beta \in \{0.25, 0.5, 1, 2\}$. The vertical dashed line marks the interpolation threshold $\gamma = 1$.

Table 3: Empirical mean / theory ratio for nearly ridgeless regression ($\lambda = 10^{-5}$, $a = 2$, $B = 0.5$, $n_{\mathrm{MC}} = 500$).

γ	Empirical	Theory	Ratio
0.20	1.015	1.118	0.908
0.40	1.199	1.291	0.929
0.60	1.454	1.581	0.920
0.75	1.838	2.000	0.919
0.80	2.045	2.236	0.914
0.85	2.395	2.582	0.927
0.90	2.929	3.162	0.926
0.93	3.490	3.780	0.923
0.95	4.157	4.472	0.929
0.97	5.300	5.774	0.918
1.03	5.252	5.627	0.933
1.05	4.075	4.349	0.937
1.07	3.389	3.650	0.929
1.10	2.820	3.018	0.934
1.15	2.236	2.417	0.925
1.25	1.679	1.806	0.930
1.50	1.134	1.186	0.956
2.00	1.010	1.000	1.010
2.50	1.208	1.200	1.006
3.00	1.334	1.333	1.000



(a) Under-parameterized ($\gamma = 0.5$, $p = 500$)

(b) Over-parameterized ($\gamma = 2$, $p = 2000$)

Figure 5: Verification of Theorem A.3 for heavy-tailed signals ($\beta_0 \sim \sigma_\beta t_4/\sqrt{n}$). Histograms show the empirical distribution of $Z = n^{1/2-1/\alpha}\|\hat{\beta}_\lambda - \beta_0\|_\infty$ over 2000 replications. Red curves show the theoretical Fréchet density. Parameters: $\alpha = 4$, $\lambda = 2$, $\sigma_\beta = 2$.
where $c_\alpha = \frac{2\,\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\,\Gamma\left(\frac{\alpha}{2}\right)} \cdot \alpha^{\frac{\alpha-1}{2}}$. is the tail constant satisfying $\mathbb{P}(|\mu| > x) \sim c_\alpha x^{-\alpha}$.

Figure 5 compares the empirical distribution (histogram) of the simulated Z values against the theoretical Fréchet density (red curve). The excellent agreement in both regimes validates the heavy-tailed limiting distribution predicted by our theory.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper characterizes the sharp asymptotics of the L infty-norm estimation error for estimating the coefficient vector in linear regression under proportional asymptotics.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: See conclusion section paragraph 2.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer:[Yes],

   Justification: All assumptions are collected in Assumption 1 and Assumption 2 and and each theorem statement explicitly references them. Full proofs of all the theorems and associated results are collected in Appendices **A**-**I**.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies the full simulation protocol, and parameter configurations.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer:[No]

   Justification: We will be happy to open code when camera ready.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We report in detials in simulation section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[Yes]

   Justification: Figures are reported with bands being empirical mean ±2 empirical standard deviations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run using 4 CPU cores on a local machine.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes],

   Justification: We reviewed the code and ethics and acknowledge that we conduct following proper code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification:This is a theoretical work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:[N/A]

    Justification: Not applicable. This is a theory work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Not applicable, this is a theory work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer:[N/A]

Justification: Not applicable, this is thoery work.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Not applicable, this is theory work.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLMs were not used in the methodological development, theoretical analysis, or experimental design for this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.