Online Appendix of
Can Machines Learn Weak Signals?

Zhouyu Shen*
Booth School of Business
University of Chicago

Dacheng Xiu†
Booth School of Business
University of Chicago and NBER

January 21, 2025

Abstract

Appendix A presents additional results from Monte Carlo simulations. Appendix B discusses the selection of tuning parameters. Appendix C is devoted to the exposition of technical lemmas along with their corresponding proofs.

A Supplemental Simulation Results

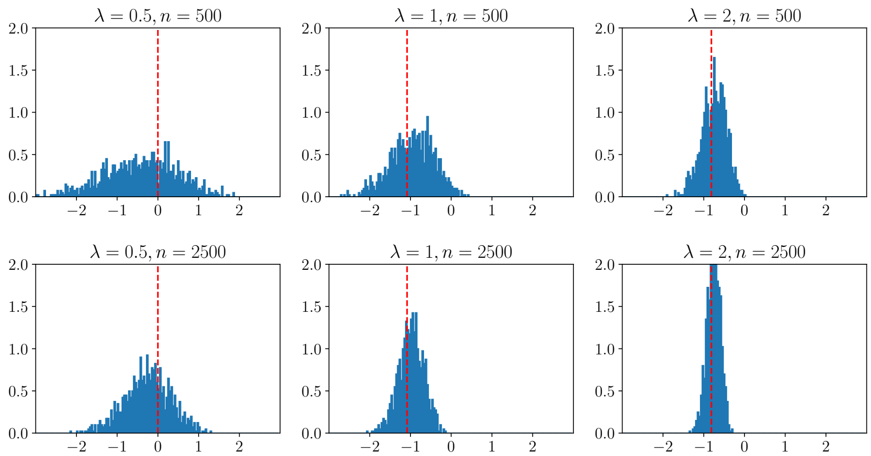
A.1 Additional Simulations with Fixed Tunings

In the simulation for the main paper, cross-validation is applied for Ridge and Lasso. In this section, we verify our theories with manually selected λ_n. We also experiment with two sample sizes, $n = 500$ and $n = 2,500$, while maintaining $p/n = 3/5$. We fix $q = 0.2$ and $R^2 = 5\%$. In the case of Ridge regression, we set λ as 0.5, 1 and 2, where $\lambda = 1$ corresponding to the optimal tuning. The histograms of relative prediction error are presented in Figure A1.

Several noteworthy observations can be made from these histograms. First, across all plots, the probability mass is concentrated around the red vertical line. As the sample size increases from 500 to 2,500 (and dimension increases from 300 to 1,500), the histograms become increasingly concentrated. This aligns with our theory, which predicts that the

*Address: 5807 S Woodlawn Avenue, Chicago, IL 60637 USA. Email: zshen10@chicagobooth.edu.
†Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: dacheng.xiu@chicagobooth.edu.

Figure A1: Simulation Results for Ridge with Fixed Tuning Parameters



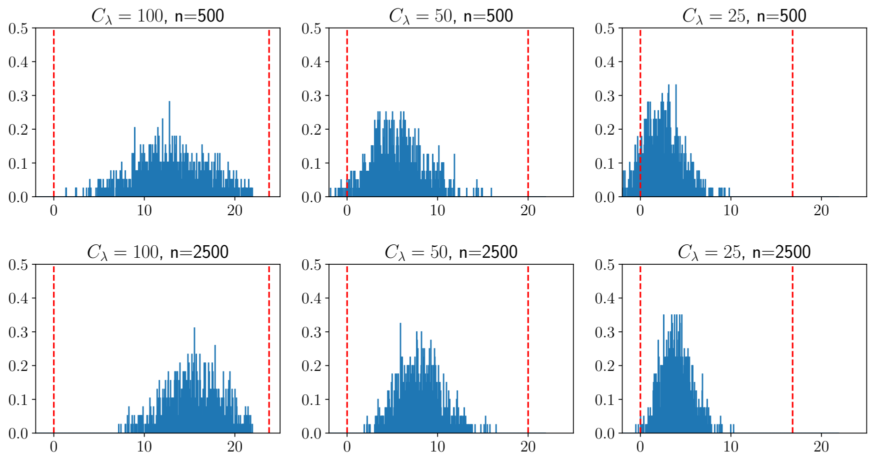
Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_r(\lambda_n))$ following equation (8) across 1,000 Monte Carlo samples. We consider two different sample sizes ($n = 500$ and $n = 2,500$) and examine three different values of λ, where $\lambda = 0.5, 1$, and 2. Notably, $\lambda = 1$ represents the optimal tuning parameter. The red dashed line indicates the values of α^*.

relative prediction error converges in probability to the limit α^* as the sample size grows. Second, the value of α^* corresponding to the optimal tuning parameter $\lambda = 1$ is the smallest. This is because the optimal Ridge estimator achieves the smallest prediction error. Moreover, almost all the probability mass corresponding to the optimal Ridge estimator is situated on the negative side of the x-axis, indicating that this estimator outperforms the zero estimator with high probability. Third, when $\lambda = 0.5$, it results in the worst performance, with a large portion of the probability mass on the positive side of zero. In contrast, for $\lambda = 2$, α^* gets closer to zero, and the variance of the relative prediction error decreases. This behavior is due to the increasing amount of penalization, which ultimately drives the estimator towards zero, and in turn, α^* towards zero as well.

In contrast to the results obtained for Ridge regression, our theoretical framework does not provide a precise error limit for Lasso. Instead, Theorem 4 offers high probability bounds on relative prediction errors. Figure A2 displays histograms of these errors for various tuning parameters and sample sizes, accompanied by two red vertical lines in each plot representing the lower and upper bounds, c_α and C_α.

These plots yield several interesting findings. First, as the sample size increases, we observe that the probability mass becomes more concentrated and largely falls within the intervals defined by the bounds. Second, regardless of the tuning parameter values, Lasso

Figure A2: Simulation Results for Lasso with Fixed Tuning Parameters



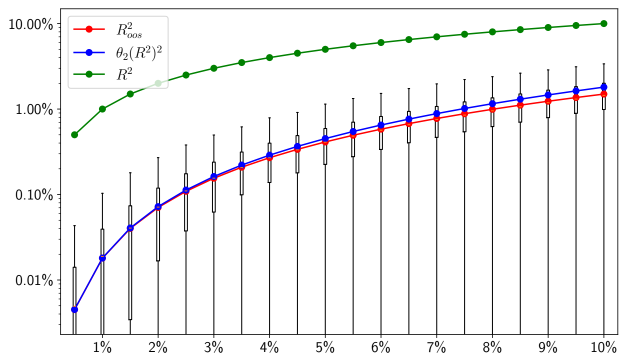
Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_l(\lambda_n))$ following equation (8) across 1,000 Monte Carlo samples. We consider two different sample sizes ($n = 500$ and $n = 2,500$) and examine three different values of C_λ. The two dashed lines in each figure indicate the values of c_α and C_α that are solutions to (10).

consistently underperforms the zero estimator in almost all samples when the sample size is large. Third, as the tuning parameter increases (indicated by a decrease in C_λ), both the lower and upper bounds approach zero. This behavior is a consequence of the increased regularization, which, in turn, steers the estimator closer to zero. In the end, Lasso becomes identical to the zero estimator.

A.2 Out-of-sample R^2

Continuing our investigation in the main text, we conduct an experiment to analyze R^2_{oos} based on the optimal Ridge. Proposition 1 describes the expected asymptotic behavior of R^2_{oos}. To empirically test this, we implement the optimal Ridge, setting $\lambda = 1$, on a training dataset comprising $n = 500$ observations. We then calculate R^2_{oos} based on predictions for a separate test dataset of size $n_{\text{oos}} = 10,000$. The comparative analysis between the population R^2, the empirically estimated R^2_{oos}, and the theoretically derived limit of R^2_{oos} is illustrated in Figure A3. For a clearer visual presentation, we apply a logarithmic transformation to the y-axis. We vary τ to compare against a range of population R^2 values from 0.5% to 10% on the x-axis. The red line represents the average R^2_{oos} over 1,000 Monte Carlo simulations. Additionally, we draw boxplots to describe the distributions of R^2_{oos} across these simulations. The theoretical limit, expressed as $p^{-1}n\theta_2(R^2)^2$, is traced by the blue line, and the green line illustrates the population R^2, which would align with a 45-degree line on a standard scale. Notably, in this weak signal setting, the population R^2 significantly surpasses the empirically achievable R^2_{oos}. Furthermore, the close alignment between the red and blue lines, particularly for scenarios with small R^2 values, substantiates our theoretical predictions.

Figure A3: Out-of-Sample R^2 for Optimal Ridge in Linear DGPs



Note: The figure presents boxplots showing the distributions of R^2_{oos} for optimal Ridge regression ($\lambda = 1$) over $1,000$ Monte Carlo repetitions, with $n = 500$, $p = 300$, $q = 0.2$, and $n_{oos} = 10,000$. We explore a range of population R^2 values, from 0.5% to 10% in increments of 0.5% by adjusting τ. The plot features red, blue, and green lines to represent the average R^2_{oos} over Monte Carlo samples, the theoretical limit as given by Proposition 1, and the population R^2. In this plot, we employ a logarithmic scale for the y-axis. Without the logarithmic transformation, the green line would align with a 45-degree line. Additionally, the lower boundaries of the boxplots surpass the axis limits in instances where the R^2_{oos} values are negative.

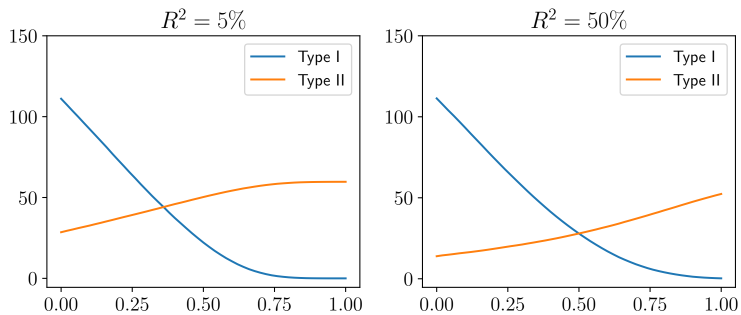
A.3 Why Lasso Fails?

A plausible explanation for the Lasso's suboptimal performance with weak signals is its difficulty in distinguishing between genuine and spurious signals. The failure to identify genuine weak signals has a minor impact on Lasso's performance relative to the zero estimator, which does not utilize any true signals. Hence, the primary challenge for the Lasso lies in its failure to adequately filter out irrelevant signals. This issue could be addressed with a sufficiently large tuning parameter. However, our theory indicates that only when the penalty is so substantial that the Lasso effectively becomes equivalent to the zero estimator does it apply an adequate penalty.

To empirically explore this issue, we quantify Type I and Type II errors in simulations of Lasso's selection relative to its tuning parameter λ_n. The findings are presented in Figure

A4. Considering our previous discussion, Type I errors represent a significant cost for Lasso. Indeed, a considerable portion of the variables selected by Lasso are incorrectly deemed genuine when λ_n is small. As λ_n increases, Type I errors decrease, enhancing Lasso's performance. Meanwhile, Type II errors persist and eventually converge to the number of non-zero betas in the DGP.

Figure A4: Lasso's Type I and Type II Errors



Note: The plots compare average Type I and Type II errors of Lasso using the linear DGP following equation (1) over 1,000 Monte Carlo samples. Two population R^2 are considered: $R^2 = 5\%$ (left panel) and $R^2 = 50\%$ (right panel). The horizontal axis of each plot represents the logarithm of λ_n, spanning a range from 0 to 1. The vertical axis measures the count of errors incurred while testing the null hypothesis $\mathbb{H}_{0,i} : \beta_i = 0$.

A.4 Robustness Check

In our subsequent series of experiments, we intentionally deviate from the assumptions originally established during the development of our theoretical framework. This deviation is aimed at evaluating the robustness and generalizability of our theoretical predictions beyond their premises and initial parameters. To facilitate this evaluation, we introduce specific modifications to the baseline configuration along three key dimensions: First, we adjust (R^2, q), exploring more extreme sparsity levels and reducing signal strength accordingly compared to the settings in the main text. Second, we increase the ratio p/n to 2 by increasing p while maintaining n, making it more challenging for both Ridge and Lasso to capture the underlying signals. Third, we modify the distribution of Z from standard Gaussian to a t-distribution characterized by four degrees of freedom, and with a mean of zero and a variance of one. In addition, we introduce heteroscedasticity into the error distribution, following the configuration outlined by Giannone et al. (2022). The error term's variance is defined by the function $\sigma^2 \exp(\alpha X_i^\top \delta / \sqrt{\sum_{i=1}^n (X_i^\top \delta)^2 / n})$ with $\alpha = 0.5$. Here, X_i represents the i-th row of X. σ serves as a scaling parameter to standardize the variance and match $\sigma_\varepsilon^2 = 1$. The vector δ is a $p \times 1$ vector with zero elements in the same positions as the zero elements of β_0, while non-zero elements are drawn from a standard Gaussian distribution.

Table A1 compare the summary statistics for various cases under consideration. In Case I, when q is small, the performance of the Lasso estimator improves relative to the baseline scenario (reproduced from Table 1 for ease of comparison). This improvement is evident at $R^2 = 5\%$ for all levels of q, as the Q1 values become negative, indicating that Lasso surpasses zero in predictive accuracy for a larger proportion of Monte Carlo repetitions. However, as R^2 is further reduced to 2%, Lasso once again becomes falls below the performance of zero. In contrast, Ridge's performance remains largely unaffected by changes in sparsity levels. As expected, its performance deteriorates in finite samples as the signal strength weakens (i.e., as R^2 decreases). Nonetheless, Ridge continues to outperform Lasso, although its relative advantage over the zero estimator diminishes. The theoretical support for these observations is discussed in Section 2.9. In Case II, we observe the increased ratio of p/n does not affect our conclusion. Case III demonstrates the robustness of our theoretical findings, as it aligns closely with the baseline scenario despite variations in distributional assumptions.

Table A1: Robustness Analysis of Ridge and Lasso in Alternative DGPs

			Lasso				Ridge			
	q	R^2 (%)	Q1	Q2	Q3	#Zero	Q1	Q2	Q3	#Zero
Case I	0.20	5%	-0.127	0.000	0.521	360	-0.992	-0.501	-0.129	97
	0.10	5%	-0.871	0.000	0.187	327	-0.981	-0.475	-0.077	113
	0.10	2%	0.000	0.000	3.435	493	-0.622	0.000	0.440	237
	0.05	5%	-2.688	-0.305	0.000	255	-1.037	-0.387	0.000	130
	0.05	2%	0.000	0.000	2.948	473	-0.642	0.000	0.426	238
	0.02	5%	-6.542	-2.050	0.000	215	-1.304	-0.230	0.000	149
	0.02	2%	0.000	0.000	1.695	432	-0.605	0.000	0.625	254
Case II	0.20	5%	0.000	0.000	3.228	470	-0.768	-0.416	0.000	183
Case III	0.20	5%	0.000	0.000	0.591	392	-0.848	-0.384	0.000	129

Note: The table illustrate the summary statistics of relative prediction error $\Delta(\hat{\beta}(\hat{\lambda}_n^{K-CV}))$ for Ridge and Lasso based on 1,000 Monte Carlo samples. We explore several distinct DGPs, each involving the alteration of a specific condition. In Case I, we try a series of different values of R^2 and q. In Case II, we adjust n/p to 0.5. In Case III, we introduce t-distributed covariates with heterogeneous variance of ε. The benchmark DGP adheres to the following specifications: $n = 500$, $p = 300$, $p/n = 3/5$, and complies with Assumptions 1 and 2. 10-fold cross-validation is used throughout these experiments.

B Choice of Tuning Parameters

In this section, we discuss the selection of tuning parameters for implementing machine learning methods. The selection process aims to balance performance and cost while ensuring fair method comparisons.

For Ridge and Lasso, each with one tuning parameter, we use the glmnet package, which optimizes it via ten-fold CV. The grid is adaptively selected for efficiency. Our implementation of RF involves three tuning parameters: the depth of each individual tree, the number of randomly selected features used in each tree split, and the proportion of sample data used for bootstrap.[1] [2] For GBRT, we tune tree depth, number of trees, and learning rate.[3] In the case of NNs, we adhere to a uniform architectural choice across our analyses, featuring a single hidden layer. The number of neurons in this hidden layer is approximately equal to the square root of the total number of neurons in the input layer, aligning the architecture with the complexity and dimensions of the dataset. By not tuning the NN architecture extensively, we streamline the model selection process while retaining adequate complexity for effective learning. For the remaining tuning parameters in trees and NNs, we select suitable ranges based on model performance from the cross-validation step. A critical element in selecting our grid is to ensure that the optimal tuning parameters are situated within the median range of the grid. The details regarding model configuration and tuning parameters for empirical studies are provided in Table B2.[4]

C Technical Lemmas and Their Proofs

For completeness, the following section introduces a collection of lemmas, including proofs for some. We start with the Convex Gaussian Min-max Theorem (CGMT), a pivotal theorem to our proof. For a detailed exposition of its proof, we direct readers to the work of Thrampoulidis et al. (2015). The CGMT pertains to the following optimization problems:

$$\Phi(G) := \min_{w\in\mathcal{S}_w}\max_{u\in\mathcal{S}_u} u^\top Gw + \psi(w,u), \text{ and } \phi(g,h) := \min_{w\in\mathcal{S}_w}\max_{u\in\mathcal{S}_u} \|w\|g^\top u - \|u\|h^\top w + \psi(w,u),$$

where $G \in \mathbb{R}^{m\times n}, g \in \mathbb{R}^m, h \in \mathbb{R}^n, \mathcal{S}_w \subset \mathbb{R}^n, \mathcal{S}_u \subset \mathbb{R}^m$, and $\psi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$.

[1] This procedure is known as subbagging, which helps address weak signals. In linear regressions, LeJeune et al. (2020) show that the asymptotic risk of subbagging least squares matches that of Ridge regression.

[2] In our simulations, tree depth varies from 5 to 20, selected features from 10 to 300, and bootstrap sample proportion from 0.1 to 0.2. The RF ensemble size is fixed at 5,000 trees, as 10,000 offers no significant improvement.

[3] The learning rate varies from 0.001 to 0.5, tree depth from 1 to 6, and the maximum number of trees is 100, though training usually stops earlier, reflecting GBRT's preference for shallower and fewer trees.

[4] We follow the same approach for the empirical analysis, except for Finance 2. Due to its scale and computational constraints, we use two-fold CV and a narrower grid: $\log(\lambda)$ ranges from 6 to 7 for Ridge and from -3.5 to -2.5 for Lasso. We ensure that the optimal tuning parameters fall within the central range of these specified grids.

Table B2: Model Configuration for Machine Learning Methods

	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
Finance 1	depth=1∼20 #trees=500 #features=1∼15 %samples=0.05∼1	depth=1∼5 #trees=1∼10 lr ∈ {0.01,0.02, 0.05,0.1,0.2,0.5,1}	architecture∼{16,4,1} batch size=16 (lr,epochs)={(0.1,5), (0.01,50), (0.0025,200)} log(λ) ∈ $[-2, 1]$	architecture∼{16,4,1} batch size=16 (lr,epochs)={(0.4,1), (0.08,5), (0.02,20)} log(λ) ∈ $[-2, 1]$
Finance 2	depth=2∼12 #trees=500 #features ∈ {1, 2,3,5} %samples=0.5∼1	depth=1∼6 #trees=10∼400 lr ∈ {0.0001,0.001, 0.01,0.02,0.05}	architecture∼{920,32,1} batch size=10000 (lr,epochs)={(0.5,2), (0.1,10), (0.067,15)} log(λ) ∈ $[-4, 0]$	architecture∼{920,32,1} batch size=10000 (lr,epochs)={(0.5,2), (0.2,5), (0.067,15), (0.05,20),(0.04,25)} log(λ) ∈ $[-5, -3]$
Macro 1	depth=5∼50 #trees=500 #features=1∼60 %samples=0.5∼1	depth=1∼5 #trees=1∼600 lr ∈ {0.005,0.01, 0.02,0.05,0.1, 0.2,0.5}	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.008,10), (0.004,20), (0.002,40), (0.0008,100), (0.0005,160)} log(λ) ∈ $[-2, 2]$	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.05,2), (0.02,5),(0.01,10), (0.005,20), (0.002,50)} log(λ) ∈ $[-10.5, 1.5]$
Macro 1b	depth=5∼50 #trees=500 #features=5∼100 %samples=0.5∼1	depth=1∼5 #trees=1∼200 lr ∈ {0.01,0.02, 0.05,0.1,0.2,0.5}	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.024,75), (0.012,150), (0.006,300) (0.003,600)} log(λ) ∈ $[-1, -0.5]$	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.004,25), (0.002,50),(0.001,100), (0.0005,200)} log(λ) ∈ $[2, 3]$
Macro 2	depth=1∼30 #trees=500 #features=1∼60 %samples=0.5∼1	depth=1∼10 #trees=1∼500 lr ∈ {0.01,0.02, 0.05,0.1,0.2,0.5}	architecture∼{61,8,1} batch size=16 (lr,epochs)={(0.02,50), (0.005,200), (0.00125,800)} log(λ) ∈ $[-3, -3]$	architecture∼{61,8,1} batch size=16 (lr,epochs)={(0.05,20), (0.02,50), (0.002,500)} log(λ) ∈ $[-7, 4]$
Micro 1	depth=1∼20 #trees=500 #features=1∼20 %samples=0.005∼0.5	depth=1∼5 #trees=1∼20 lr ∈ {10^{-15},10^{-14}, ...,0.05,0.1,0.2,0.5}	architecture∼{297,16,1} batch size=16 (lr,epochs)={(0.1,1), (0.01,10), (0.001,100), (0.0001,1000), (0.00005,2000)} log(λ) ∈ $[-11, -7]$	architecture∼{297,16,1} batch size=16 (lr,epochs)={(0.4,1), (0.2,2), (0.08,5), (0.04,10), (0.02,20)} log(λ) ∈ $[-8, 0]$
Micro 2	depth=5∼30 #trees=500 #features=1∼30 %samples=0.5∼1.0	depth=1∼6 #trees=1∼30 lr ∈ {0.05,0.1, 0.15,...,1}	architecture∼{217,16,1} batch size=16 (lr,epochs)={(0.1,1),(0.02,5), (0.01,10),(0.005,20)} log(λ) ∈ $[-10, -6]$	architecture∼{217,16,1} batch size=16 (lr,epochs)={(0.1,1),(0.01,10), (0.001,100), (0.0001,1000)} log(λ) ∈ $[-12, -9]$
Micro 2b	depth=1∼50 #trees=500 #features=1∼3 %samples=0.5∼1	depth=1∼10 #trees=1∼50 lr ∈ {10^{-10},10^{-9}, ...,0.1,0.2,0.5,1}	architecture∼{215,16,1} batch size=16 (lr,epochs)={(0.04,5), (0.02,10), (0.01,20)} log(λ) ∈ $[0, 2]$	architecture∼{215,16,1} batch size=16 (lr,epochs)={(0.01,50), (0.005,100), (0.0025,200)} log(λ) ∈ $[0, 2]$

Note: The table reports the range of tuning parameters for RF, GBRT, and NNs, as well as the architecture of NNs applied across six datasets. For RF, we fix the number of trees at #trees= 500, and tune three other parameters: the depth of the tree (depth), the number of features (#features), and the ratio of bootstrapped samples (%samples) within a predefined grid. In the case of GBRT, we tune depth and #trees, and the learning rate (lr). For NNs, we adopt a fixed model architecture, denoted by the number of neurons in each layer indicated in brackets. Additionally, we fix the batch size for SGD and focus on jointly tuning the learning rate (lr) and the number of epochs (epochs), as well as the ℓ_1- or ℓ_2-penalty parameter (log(λ)).

Lemma 1 (CGMT)**.** *Suppose that $\mathcal{S}_w$ and $\mathcal{S}_u$ are compact sets, ψ is continuous on $\mathcal{S}_w \times \mathcal{S}_u$, and the entries of G, g, and h are i.i.d. Gaussian. Then we have* $\mathrm{P}(\Phi(G) < c) \leq 2\mathrm{P}(\phi(g,h) \leq c)$, $\forall c \in \mathbb{R}$. *Moreover, if $\mathcal{S}_w$ and $\mathcal{S}_u$ are convex sets, and ψ is convex-concave on $\mathcal{S}_{\mathrm{w}} \times \mathcal{S}_u$, then* $\mathrm{P}(\Phi(G) > c) \leq 2\mathrm{P}(\phi(g,h) \geq c)$, $\forall c \in \mathbb{R}$.

Lemma 2 (Lemma B.26 from Bai and Silverstein (2009))**.** *Let $x = (x_1, \cdots, x_n)^\top$ be a random vector of i.i.d. entries. Assume that $\mathbb{E}x_i = 0$, $\mathbb{E}x_i^2 = 1$, and $\mathbb{E}x_i^4 \leq v_4$. Then, for any $A \in \mathbb{R}^{n\times n}$, it holds that $x^\top Ax - \mathrm{Tr}(A) = O_{\mathrm{P}}\big(\sqrt{v_4\,\mathrm{Tr}(AA^\top)}\big)$.*

Lemma 3. *Let $x = (x_1, \cdots, x_n)^\top$ and $y = (y_1, \cdots, y_m)^\top$ be two independent random vectors with i.i.d. entries. Assume that each element has a mean of zero and a variance of one. Then, for any $A \in \mathbb{R}^{n\times m}$, it holds that $x^\top Ay = O_{\mathrm{P}}\big(\sqrt{\mathrm{Tr}(AA^\top)}\big)$.*

Proof. The conclusion follows from the fact that $\mathbb{E}(x^\top Ay)^2 = \mathrm{Tr}(AA^\top)$. □

The lemma below pertains to the Neumann series. See Meyer (2000) for a detailed proof.

Lemma 4. *If A is a square matrix with $\|A\| < 1$, then $\mathbb{I} - A$ is nonsingular and $(\mathbb{I} - A)^{-1} = \sum_{k=0}^{\infty} A^k$. As a consequence, $\|(\mathbb{I} - A)^{-1} - \sum_{k=0}^{\ell} A^k\| \leq \sum_{k=\ell+1}^{\infty} \|A\|^k = \|A\|^{\ell+1}/(1 - \|A\|)$.*

Lemma 5. *Assume $x = (x_1, \cdots, x_n)^\top$ and $y = (y_1, \cdots, y_p)^\top$ are two independent random vectors with i.i.d. sub-exponential random variables with their sub-exponential norm bounded by K. Then for any $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times p}$, there exists a constant $c > 0$ such that*

$$\mathrm{P}\left(|x^\top Ax - \mathbb{E}x^\top Ax| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|A\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|A\|^{1/2}}\right\}\right), \tag{C1}$$

$$\mathrm{P}\left(|x^\top By| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|B\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|B\|^{1/2}}\right\}\right). \tag{C2}$$

Proof. Inequality (C1) is given by Proposition 1.1 presented in Götze et al. (2021) for the case of symmetric A. To extend it for the asymmetric case, we use the identity that $x^\top Ax = x^\top(A + A^\top)x/2$, which allows us to apply (C1) to $(A + A^\top)/2$. Using triangle inequalities, we have $\big\|(A + A^\top)/2\big\|_{\mathrm{F}}^2 \leq \|A\|_{\mathrm{F}}^2$ and $\big\|(A + A^\top)/2\big\|^{1/2} \leq \|A\|^{1/2}$. Thus, (C1) holds for asymmetric A. For (C2), let $z = (x^\top, y^\top)^\top$ and $C = \begin{pmatrix} 0_{n\times n} & B \\ 0_{p\times n} & 0_{p\times p} \end{pmatrix}$. By (C1), we obtain

$$\begin{aligned}\mathrm{P}\left(|x^\top By| \geq t\right) = \mathrm{P}\left(|z^\top Cz| \geq t\right) &\leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|C\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|C\|^{1/2}}\right\}\right) \\ &= 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|B\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|B\|^{1/2}}\right\}\right).\end{aligned}$$

□

The next lemma is established in Bai and Silverstein (2009) and Chen and Pan (2012).

Lemma 6. *Suppose Z is an $n \times p$ matrix with i.i.d. Gaussian entries. Then for any positive constant $\epsilon > 0$, it holds that $n^{-1}Z^\top Z \le (1+\epsilon)(1+\sqrt{c_n})^2$, w.p.a.1, for $c_n = p/n \in [0, \infty]$.*

Lemma 7 (Convexity)**.** *Let $O \subseteq \mathbb{R}^d$ be open and convex and D be a dense subset of O. For $\theta \in O$, both $M_n(\theta)$ and $M(\theta)$ are convex in θ. If $M_n(\theta) \xrightarrow{\mathrm{P}} M(\theta)$, for any $\theta \in D$, then $\sup_{\theta \in K} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0$, for any compact subset $K \subset O$.*

This lemma has been shown by Lemma 7.75 of Liese and Miescke (2008). Next, we present a min-convergence theorem for functions defined on an open set $(0, \infty)$, as shown by Lemma 10 of Thrampoulidis et al. (2018).

Lemma 8. *Consider a sequence of proper, convex stochastic functions $M_n : \mathbb{R}^+ \to \mathbb{R}$, and a deterministic function $M : \mathbb{R}^+ \to \mathbb{R}$, satisfying (a) $M_n(x) \xrightarrow{\mathrm{P}} M(x)$, $\forall x > 0$; (b) there exists $z > 0$ such that $M(x) > \inf_{y>0} M(y)$, $\forall x \ge z$. Then we have $\inf_{x>0} M_n(x) \xrightarrow{\mathrm{P}} \inf_{x>0} M(x)$.*

Relatedly, we introduce a lemma for functions on a diverging sequence of closed sets.

Lemma 9. *Consider a sequence of closed intervals $\{[x_n, y_n]\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} x_n = -\infty$ and $\lim_{n\to\infty} y_n = +\infty$. Additionally, let there be a sequence of proper random and convex functions $M_n : [x_n, y_n] \to \mathbb{R}$, and a convex, continuous, and deterministic function $M : \mathbb{R} \to \mathbb{R}$ that satisfy: (a) $M_n(x) \xrightarrow{\mathrm{P}} M(x)$ for every $x \in \mathbb{R}$; (b) there exists $z > 0$ such that $M(x) > \inf_{y\in\mathbb{R}} M(y)$ holds for all $|x| \ge z$. Then it holds that $\inf_{x\in[x_n,y_n]} M_n(x) \xrightarrow{\mathrm{P}} \inf_{x\in\mathbb{R}} M(x)$.*

Proof. For n sufficiently large, $z \in [x_n, y_n]$. Assume $x^* \in [-z, z]$ minimizes $M(x)$. Condition (b) in fact implies that $x^\star \in (-z, z)$ and that $M(x^\star) = \inf_{x\in\mathbb{R}} M(x)$. Consider the event $\inf_{|x|>z, x\in[x_n,y_n]} M_n(x) < M_n(x^*)$. Under this event, there exists $|z_n| > z$ and $z_n \in [x_n, y_n]$ such that $M_n(z_n) < M_n(x^*)$. The geometry implies that there exists $\theta_n \in (0, 1)$, such that either $z_n\theta_n + x^\star(1-\theta_n) = z$ or $z_n\theta_n + x^\star(1-\theta_n) = -z$ holds. Using convexity, we have $\min(M_n(z), M_n(-z)) \le \theta_n M_n(z_n) + (1-\theta_n)M_n(x^*) < M_n(x^*)$. By taking limits on both sides, we have $\min(M(z), M(-z)) \le M(x^*)$, which contradicts condition (b). Therefore, w.p.a.1, we have $\inf_{|x|>z, x\in[x_n,y_n]} M_n(x) \ge M_n(x^*)$. Furthermore, by Lemma 7, for all arbitrarily small $\epsilon > 0$, w.p.a.1, $\sup_{|x|\le z} |M_n(x) - M(x)| < \epsilon$. In addition, by definition, there exists a sequence of z_n, such that $|z_n| \le z$ and $\inf_{|x|\le z} M_n(x) \ge M_n(z_n) - \epsilon$. Combining these two inequalities with the fact that $M(x^\star)$ minimizes M on $\mathbb{R}$ leads to $\inf_{|x|\le z} M_n(x) \ge M_n(z_n) - \epsilon \ge M(z_n) - 2\epsilon \ge M(x^*) - 2\epsilon$, w.p.a.1. On the other hand, $\inf_{|x|\le z} M_n(x) \le M_n(x^*) \xrightarrow{\mathrm{P}} M(x^*)$. Since ϵ

is arbitrary, we have $\inf_{|x|\le z} M_n(x) \xrightarrow{\text{P}} M(x^*)$. Along with $\inf_{|x|>z, x\in[x_n,y_n]} M_n(x) \ge M_n(x^*)$ and $M_n(x^\star) \xrightarrow{\text{P}} M(x^*)$ by (a), we have

$$\inf_{x\in[x_n,y_n]} M_n(x) = \min\left(\inf_{|x|\le z} M_n(x), \inf_{|x|>z, x\in[x_n,y_n]} M_n(x) \right) \xrightarrow{\text{P}} M(x^*). \qquad \square$$

Lemma 10. *Suppose X is a standard Gaussian random variable, then for $x > 0$,*

$$\frac{\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{x^2}{2}\right) \left(2x^{-3} - 12x^{-5} - 15x^{-7}\right) \le \mathbb{E}(|X| - x)_+^2 \le \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{x^2}{2}\right) \left(2x^{-3} + 3x^{-5}\right).$$

Proof. With integration by parts, we find

$$\mathbb{E}(|X| - x)_+^2 = \sqrt{\frac{2}{\pi}} \int_x^\infty (t - x)^2 \exp\left(-\frac{t^2}{2}\right) dt = \sqrt{\frac{2}{\pi}} \left(-x \exp\left(-\frac{x^2}{2}\right) + (x^2 + 1) f_G(x) \right),$$

where $f_G(x) = \int_x^\infty \exp(-t^2/2)dt$. Lemma 10 then follows from the tail inequality:

$$\exp\left(-\frac{x^2}{2}\right) \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}\right) \le f_G(x) \le \exp\left(-\frac{x^2}{2}\right) \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}\right). \qquad \square$$

Lemma 11. *Given that X is a standard Gaussian random variable, the following inequalities hold when $x > 0$ and x is sufficiently large:*

$$\mathbb{E}|X|(|X| - x)_+^2 \le 2x\mathbb{E}(|X| - x)_+^2 \quad \text{and} \quad \mathbb{E}X^2(|X| - x)_+^2 \le 2x^2\mathbb{E}(|X| - x)_+^2.$$

Proof. The proof is analogous to that of Lemma 10 and is therefore omitted. $\square$

Definition 1. *A centered random variable X belongs to the sub-exponential class $SE(\nu^2, \alpha)$ with $\nu > 0$ and $\alpha > 0$, if $\mathbb{E}e^{\lambda X} \le e^{\frac{\lambda^2\nu^2}{2}}$, for all λ such that $|\lambda| < \alpha^{-1}$.*

Lemma 12. *Let $\{x_k\}_{k=1}^\infty$ be a sequence of diverging positive numbers. Then as $p \to \infty$, we have w.p.a.1, $\|\Sigma_2 b_0\|_\infty < x_p q^{-1/2} \log(p)$ and $\|\Sigma_2^{1/2} h\|_\infty < x_p \sqrt{\log(p)}$, where Σ_2 and b_0 are defined in Assumptions 1 and 3, respectively, and $h \in \mathbb{R}^p$ is a standard Gaussian vector.*

Proof. We only present the proof for the first inequality, noting that the proof for the second inequality follows similarly. By definition, there exist $b_{1i} \sim B(1, q)$ and a sub-exponential random variable b_{2i} such that $b_{0,i} = q^{-1/2} b_{1i} b_{2i}$. Note that $b_{1i} b_{2i}$ is still sub-exponential. Without loss of generality, assume $q^{1/2} b_{0,i} = b_{1i} b_{2i} \in \text{SE}(1, 1)$.

Write the (i, j)-th element of Σ_2 as $\Sigma_{2,ij}$. By the properties of sub-exponential variables, we have $\left(\Sigma_2 q^{1/2} b_0\right)_i \in \text{SE}\left(\sum_{j=1}^p \Sigma_{2,ij}^2, \max_j |\Sigma_{2,ij}|\right)$. Given that $\sum_{j=1}^p \Sigma_{2,ij}^2 = (\Sigma_2^2)_{i,i} \le$

$\lambda_1(\Sigma_2^2) = C_2^2$ and $\max_j |\Sigma_{2,ij}| \le C_2$, we conclude that $\left(\Sigma_2 q^{1/2} b_0\right)_i \in \mathrm{SE}(C_2^2, C_2)$. The tail bound of sub-exponential variables yields $\mathrm{P}\left(|(\Sigma_2 q^{1/2} b_0)_i| > x_p \log(p)\right) \le 2\exp\left(-\frac{x_p \log(p)}{2C_2}\right)$. The conclusion follows by union bound inequality and $2p\exp\left(-\frac{x_p\log(p)}{2C_2}\right) \to 0$. □

Lemma 13. *For any given M_1, it holds that $\lim_{n\to\infty} S_{1n} = 0$ for S_{1n} defined in Eq. (21).*

Proof. Write $\tilde{x}_k = (\Sigma_\varepsilon^{-1/2} X_{\cdot,i})_k$ and $\tilde{y}_k = \beta_{0,i}\tilde{x}_k + z_k$ for $k = 1, \ldots, n$. By definition, we have

$$\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i) = \frac{\int b\mathbf{1}_{q^{1/2}b\in[M_1,M_2]} \exp\left(-\sum_{k=1}^n \left(\tilde{y}_k - p^{-1/2}\tau^{1/2}\tilde{x}_k b\right)^2/2\right) dF(b)}{\int \exp\left(-\sum_{k=1}^n \left(\tilde{y}_k - p^{-1/2}\tau^{1/2}\tilde{x}_k b\right)^2/2\right) dF(b)}$$

$$= \frac{\int b\mathbf{1}_{q^{1/2}b\in[M_1,M_2]} \exp\left(-p^{-1}\tau b^2 \sum_{k=1}^n \tilde{x}_k^2/2 + bp^{-1/2}\tau^{1/2}\sum_{k=1}^n \tilde{y}_k\tilde{x}_k\right) dF(b)}{\int \exp\left(-p^{-1}\tau b^2 \sum_{k=1}^n \tilde{x}_k^2/2 + bp^{-1/2}\tau^{1/2}\sum_{k=1}^n \tilde{y}_k\tilde{x}_k\right) dF(b)} := \frac{Q_{1n}}{Q_{2n}},$$

where F is the distribution function of $b_{0,i}$. By the facts that $\int b\mathbf{1}_{q^{1/2}b\in[M_1,M_2]}dF(b) = 0$ and $dF(b) = (1-q)\delta_0 + qdF_{b_2}(q^{1/2}b)$, we have

$$\begin{aligned}
|Q_{1n}| =& \left|\int qb\mathbf{1}_{q^{1/2}b\in[M_1,M_2]}\left[\exp\left(-p^{-1}\tau b^2\sum_{k=1}^n \tilde{x}_k^2/2 + bp^{-1/2}\tau^{1/2}\sum_{k=1}^n \tilde{y}_k\tilde{x}_k\right) - 1\right] dF_{b_2}(q^{1/2}b)\right| \\
\le& q^{1/2}\tilde{M}\int \left|\exp\left(-p^{-1}\tau q^{-1}\tilde{b}^2\sum_{k=1}^n \tilde{x}_k^2/2 + \tilde{b}q^{-1/2}p^{-1/2}\tau^{1/2}\sum_{k=1}^n \tilde{y}_k\tilde{x}_k\right) - 1\right| dF_{b_2}(\tilde{b}),
\end{aligned}$$

where $\tilde{M} := \max(|M_1|, |M_2|)$. Define the event

$$A_n := \left\{\left|p^{-1/2}\tau^{1/2}\sum_{k=1}^n \tilde{y}_k\tilde{x}_k\right| \le \tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p) \text{ and } p^{-1}\tau\sum_{k=1}^n \tilde{x}_k^2 \le \tilde{C}p^{-1}n\tau\right\}, \quad \text{(C3)}$$

where $\tilde{C} := 5C_1C_2c_\varepsilon^{-1}$. Under this event, we observe that

$$\begin{aligned}
&\left|\exp\left(-p^{-1}\tau q^{-1}\tilde{b}^2\sum_{k=1}^n \tilde{x}_k^2/2 + \tilde{b}q^{-1/2}p^{-1/2}\tau^{1/2}\sum_{k=1}^n \tilde{y}_k\tilde{x}_k\right) - 1\right| \\
\le& \exp\left(\tilde{C}|\tilde{b}|p^{-1/2}\tau^{1/2}n^{1/2}q^{-1/2}\log^2(p)\right) - \exp\left(-\tilde{C}\tilde{b}^2p^{-1}n\tau q^{-1} - \tilde{C}|\tilde{b}|p^{-1/2}\tau^{1/2}n^{1/2}q^{-1/2}\log^2(p)\right).
\end{aligned}$$

Since $p^{-1/2}\tau^{1/2}n^{1/2}q^{-1/2}\log^2(p) \to 0$ and $p^{-1}n\tau q^{-1} \to 0$ by Assumption 4, and given that F_{b_2} follows a sub-exponential distribution, the integral of both terms on the right-hand-side converges to zero as $n \to \infty$. Therefore, for any $\epsilon > 0$, there exists n_0 such that for all $n > n_0$, we have $|Q_{1n}| \leq \epsilon$ under the event A_n. Similarly, it can be proven that there exists n_1 such that for all $n > n_1$, we obtain $Q_{2n} \geq 1/2$ under the event A_n.

Next we analyze the event A_n. We start with the second inequality in A_n. Note that

$$\begin{aligned}\sum_{k=1}^{n}\tilde{x}_k^2 =& X_{\cdot,i}^\top\Sigma_\varepsilon^{-1}X_{\cdot,i} \leq c_\varepsilon^{-1}X_{\cdot,i}^\top X_{\cdot,i} = c_\varepsilon^{-1}e_i^\top\Sigma_2^{1/2}Z^\top\Sigma_1 Z\Sigma_2^{1/2}e_i \\ \leq& c_\varepsilon^{-1}C_1\|Z\Sigma_2^{1/2}e_i\|^2 \overset{d}{=} c_\varepsilon^{-1}C_1\|\Sigma_2^{1/2}e_i\|^2\chi^2(n) \leq c_\varepsilon^{-1}C_1C_2\chi^2(n),\end{aligned}$$

where e_i is the i-th standard basis vector. By Lemma 5, with probability at least $1 - 2\exp(-cp)$ for some fixed constant $c > 0$, $\chi^2(n) \leq 5n$, which implies the second inequality.

For the first inequality in A_n, using the second inequality, we observe that,

$$\begin{aligned}p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n}\tilde{x}_k\tilde{y}_k\Big| &= p^{-1/2}\tau^{1/2}\Big|\beta_{0,i}\sum_{k=1}^{n}\tilde{x}_k^2 + \sum_{k=1}^{n}\tilde{x}_k z_k\Big| \leq \tilde{C}np^{-1/2}\tau^{1/2}|\beta_{0,i}| + p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n}\tilde{x}_k z_k\Big| \\ &= \tilde{C}p^{-1}\tau n|q^{-1/2}b_{1i}b_{2i}| + p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n}\tilde{x}_k z_k\Big| \leq \tilde{C}p^{-1}q^{-1/2}\tau n|b_{2i}| + p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n}\tilde{x}_k z_k\Big|.\end{aligned}$$

Using the property of a sub-exponential random variable, for some constant $c > 0$, with probability at least $1 - 2\exp(-c\log^2(p))$, we have $|b_2| \leq \log^2(p)/2$, which implies $\tilde{C}p^{-1}q^{-1/2}\tau n|b_{2i}| \leq \tilde{C}p^{-1}q^{-1/2}\tau n\log^2(p)/2 = o(\tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p)/2)$ by Assumption 4. In addition, by Lemma 5, with probability at least $1 - 2\exp(-c\log^2(p))$, we have $\Big|\sum_{k=1}^{n}\tilde{x}_k z_k\Big| \leq \tilde{C}n^{1/2}\log^2(p)/2$, which implies $p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n}\tilde{x}_k z_k\Big| \leq \tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p)/2$.

In sum, using the facts that $\max(\exp(-cp), \exp(-c\log^2(p))) = o(p^{-1})$, we conclude that with probability at least $1 - p^{-1}$ as $p \to \infty$, A_n holds. Hence we have

$$\begin{aligned}\lim_{n\to\infty} S_{1n} &= \lim_{n\to\infty}\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i)\big)^2\mathbf{1}_{A_n} + \lim_{n\to\infty}\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i)\big)^2\mathbf{1}_{A_n^c} \\ &\leq 4\epsilon^2 + \lim_{n\to\infty} q^{-1}\tilde{M}^2\mathrm{P}(A_n^c) \leq 4\epsilon^2 + \lim_{n\to\infty} p^{-1}q^{-1}\tilde{M}^2 = 4\epsilon^2.\end{aligned}$$

The conclusion then follows from the arbitrariness of ϵ. □

Lemma 14. *The objective function in Eq. (27) is convex with respect to α and jointly concave with respect to (δ, γ). Additionally, as long as Eq. (29) holds, we have Eq. (26).*

Proof. By Lemma 15, it suffices to prove Eq. (26) holds for $\hat{w}^B$, which equals

$$\arg\min_{w \in S_w^n} \frac{c_n}{n} \left\| \tau^{1/2} \Sigma_1^{1/2} Z w - \tau^{-1} \varepsilon \right\|^2 + c_n^2 \lambda \left\| \Sigma_2^{-1/2} w + \tau^{-3/2} \beta_0 \right\|^2 - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi, \tag{C4}$$

and $S_w^n = \{w | c_n \tau^{-1} \sigma_x \sigma_\beta - K_\alpha \le c_n \|w\| \le c_n \tau^{-1} \sigma_x \sigma_\beta + K_\alpha\}$ for some sufficiently large K_α. With slight abuse of notation, we refer to the optimal solution $\hat{w}$ instead of using $\hat{w}^B$.

Note that for any vector x, $\|x\|^2 = \max_u \sqrt{n} u^\top x - n\|u\|^2/4$, where its argmax is $2x/\sqrt{n}$, and similarly $\|x\|^2 = \max_v v^\top x - \|v\|^2/4$. Applying these equalities to $\|\tau^{1/2}\Sigma_1^{1/2} Z w - \tau^{-1}\varepsilon\|^2$ and $\|\Sigma_2^{-1/2} w + \tau^{-3/2}\beta_0\|^2$, setting $\tilde{u} = \Sigma_1^{1/2} u$, and $\tilde{v} = \Sigma_2^{-1/2} v$, we can rewrite (C4) as

$$\begin{aligned} \min_{w \in S_w^n} \max_{\tilde{u}, \tilde{v}} & \frac{c_n \tau^{1/2}}{\sqrt{n}} \tilde{u}^\top Z w - \frac{c_n \tau^{-1}}{\sqrt{n}} \tilde{u}^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} \tilde{u}\|^2}{4} + c_n^2 \lambda \tilde{v}^\top w + c_n^2 \lambda \tau^{-3/2} \tilde{v}^\top \Sigma_2^{1/2} \beta_0 \\ & - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} \tilde{v}\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi. \end{aligned} \tag{C5}$$

To simplify notation and without ambiguity, we continue using u and v in place of $\tilde{u}$ and $\tilde{v}$.

For a given w, the argmax of Eq. (C5), denoted by $\hat{u}$, is equal to $\frac{2}{\sqrt{n}}(\tau^{1/2}\Sigma_1 Z w - \tau^{-1}\Sigma_1^{1/2}\varepsilon)$. Given the definition of S_w^n and Assumptions 1 and 2, we have $\|w\| \le \tau^{-1}\sigma_x\sigma_\beta + c_n^{-1} K_\alpha$, $\|\Sigma_1\| \le C_1$, $\|\Sigma_\varepsilon\| \le C_\varepsilon$. Furthermore, w.p.a. 1, $\|z\| \le \sqrt{2n}$ by the law of large numbers, which implies $\|\varepsilon\| \le \sqrt{2C_\varepsilon n}$. Together with Lemma 6 and that $\tau c_n \to 0$ by Assumption 4, we have the following upper bound for $\|\hat{u}\|$ as n is large enough: $\|\hat{u}\| \le \frac{2\tau^{1/2}}{\sqrt{n}}\|\Sigma_1 Z w\| + \frac{2}{\sqrt{n}}\|\tau^{-1}\Sigma_1^{1/2}\varepsilon\| \le 4\tau^{-1}\sqrt{C_1 C_\varepsilon}$. Let $S_u^n = \{u | \|u\| \le 4\tau^{-1}\sqrt{C_1 C_\varepsilon}\}$. Based on the above result, w.a.p.1, the following optimization problem is equivalent to (C5):

$$\begin{aligned} \min_{w \in S_w^n} \max_{\substack{u \in S_u^n \\ v}} & \frac{c_n \tau^{1/2}}{\sqrt{n}} u^\top Z w - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4} + c_n^2 \lambda v^\top w + c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 \\ & - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi. \end{aligned} \tag{C6}$$

Next, we need introduce an auxiliary problem for the purpose of applying CGMT:

$$\phi(g,h) = \max_{0\le\delta\le 4\tau_v^{-1}\sqrt{C_1C_\varepsilon}} \min_{w\in S_w^n} \max_{\|u\|=\delta} \mathcal{R}_n(w,v,u), \quad \text{where}$$

$$\begin{aligned}\mathcal{R}_n(w,v,u) =& \frac{c_n\tau^{1/2}}{\sqrt{n}}\|w\|g^\top u - \frac{c_n\tau^{1/2}}{\sqrt{n}}\delta h^\top w - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} \\ &+ c_n^2\lambda v^\top w + c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi,\end{aligned} \tag{C7}$$

and $g\in\mathbb{R}^n$ and $h\in\mathbb{R}^p$ are standard Gaussian vectors, independent of the other random variables. Similarly, let $\tilde{\mathcal{S}}_n := \{w\big||c_n\|w\| - c_n\tau^{-1}\sigma_x\sigma_\beta - \alpha_2^*| < \epsilon\}$, define $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$ as the optimal value of an analogous optimization problem to (C7), with w restricted to $S_w^n \cap \tilde{\mathcal{S}}_n^c$.

Lemma 16 characterizes the limiting behavior of the optimal solution to (C6), $\hat{w}$, and in turn, proves the desired (26), under conditions pertaining to the optimization problem (C7). Therefore, we only need show that conditions outlined in Lemma 16 hold as long as (29) holds. That is, under (29), we need to prove the existence of the constants $\bar{\phi} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c}$ such that for all $\eta > 0$, w.p.a.1 in the limit of $n\to\infty$, $\phi(g,h) < \bar{\phi} + \eta$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h) > \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \eta$.

Let $\bar{u} = u/\delta$, maximizing part of $\mathcal{R}_n(w,v,u)$ over u simplifies to the following problem:

$$\begin{aligned}&\max_{\|u\|=\delta} \frac{c_n\tau^{1/2}}{\sqrt{n}}\|w\|g^\top u - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} \\ =& \max_{\|\bar{u}\|=1} \frac{c_n\delta}{\sqrt{n}}(\tau^{1/2}\|w\|g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top\bar{u} - \frac{c_n\delta^2}{4}\bar{u}^\top\Sigma_1^{-1}\bar{u}.\end{aligned}$$

The latter is a quadratic programming problem, which has been studied in Gander et al. (1989). The optimal value associated with this problem is given by:

$$-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) \tag{C8}$$

where $\alpha := \|w\|$ and $\mu_n(\alpha,\delta)$ is the solution to

$$\frac{1}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-2}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{\delta^2}{4} = 0, \tag{C9}$$

under the condition $\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I}$ is positive semidefinite. With this, Eq. (C7) equals:

$$\begin{aligned}\max_{0\le\delta\le 4\tau_v^{-1}\sqrt{C_1C_\varepsilon}} \min_{w\in S_w^n} &-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1} \\ &\times(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n\tau^{1/2}}{\sqrt{n}}\delta h^\top w + c_n^2\lambda v^\top w\end{aligned}$$

$$+ c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi.$$

Solving the inside minimization problem with respect to w/α while fixing α leads to

$$\max_{\substack{0 \le \delta \le 4\tau^{-1}\sqrt{C_1 C_\varepsilon} \\ v}} \min_{|c_n\alpha - c_n \tau^{-1}\sigma_x\sigma_\beta| \le K_\alpha} - \frac{c_n \delta^2}{4} \mu_n(\alpha, \delta) + \frac{c_n}{n} (\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top (\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}$$
$$\times (\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - c_n \|c_n \lambda v - n^{-1/2}\tau^{1/2}\delta h\| \alpha \tag{C10}$$
$$+ c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi.$$

By Lemma 17, the objective function of the above optimization is convex in α and jointly concave in (δ, v). As a result, we can switch the order of min and max by Corollary 3.3 in Sion (1958). Also, note that for any vector x, $\|x\| = \min_{\gamma > 0} \frac{1}{2\gamma}\|x\|^2 + \frac{\gamma}{2}$. Applying this equation to $\|c_n \lambda v - n^{-1/2}\tau^{1/2}\delta h\|\alpha$, Eq. (C10) becomes

$$\min_{c_n|\alpha - \tau^{-1}\sigma_x\sigma_\beta| \le K_\alpha} \max_{\substack{\gamma > 0 \\ 0 \le \delta \le 4\tau^{-1}\sqrt{C_1 C_\varepsilon}}} \max_v - \frac{c_n \delta^2}{4} \mu_n(\alpha, \delta) + \frac{c_n}{n} (\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top$$
$$\times (\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1} (\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n \gamma}{2} - \frac{c_n \alpha^2}{2\gamma} \|c_n \lambda v - n^{-1/2}\tau^{1/2}\delta h\|^2$$
$$+ c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi.$$

Note that the objective function above is jointly concave in (δ, γ, v). To see why this is true, it is sufficient to prove that $-\frac{\alpha^2}{2\gamma}\|c_n \lambda v - n^{-1/2}\tau^{1/2}\delta h\|^2$ is jointly concave in (δ, γ, v), which follows by Lemma 13 in Thrampoulidis et al. (2018). Consequently, after solving the first maximization problem over v, the resulting function remains jointly concave in (δ, γ). Maximizing over v is again a standard quadratic programming problem, which leads to Eq. (27). Thus, we conclude that (27) is convex with respect to α and jointly concave with respect to (δ, γ).

For any compact set A, define $\phi_A(g, h) := \min_{\alpha_2 \in A} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n(\alpha_2, \delta_3, \gamma_1)$. Based on the above argument and condition (i) in (29), we can deduce

$$\phi(g, h) = \phi_{[-K_\alpha, K_\alpha]}(g, h) \xrightarrow{\text{P}} \min_{\alpha_2 \in [-K_\alpha, K_\alpha]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in \mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1),$$

$$\phi_{\tilde{\mathcal{S}}_n^c}(g, h) = \min\{\phi_{[-K_\alpha, \alpha_2^* - \epsilon]}(g, h), \phi_{[\alpha_2^* + \epsilon, K_\alpha]}(g, h)\} \xrightarrow{\text{P}} \min_{\alpha_2 \in [-K_\alpha, \alpha_2^* - \epsilon] \cup [\alpha_2^* + \epsilon, K_\alpha]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in \mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1).$$

Together with condition (ii), the conditions in Lemma 16 are satisfied by letting $\bar{\phi} = \min_{\alpha_2 \in [-K_\alpha, K_\alpha]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in \mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1)$ and $\bar{\phi}_{\tilde{\mathcal{S}}_n^c} = \min_{\alpha_2 \in [-K_\alpha, \alpha_2^* - \epsilon] \cup [\alpha_2^* + \epsilon, K_\alpha]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in \mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1)$. □

We now introduce two lemmas whose proofs follow the same reasoning as those of Lemma 5 and Lemma 7 in Thrampoulidis et al. (2018), and are therefore omitted here.

Lemma 15. *Under the conditions of Theorem 2, define* $S_w^n := \{w | c_n \tau^{-1} \sigma_x \sigma_\beta - K_\alpha \le c_n \|w\| \le c_n \tau^{-1} \sigma_x \sigma_\beta + K_\alpha\}$ *for some* K_α *such that* $|\alpha_2^*| < K_\alpha$. *If the solution* $\hat{w}^B$ *to*

$$\arg\min_{w \in S_w^n} \frac{c_n}{n} \left\| \tau^{1/2} \Sigma_1^{1/2} Z w - \tau^{-1} \varepsilon \right\|^2 + c_n^2 \lambda \left\| \Sigma_2^{-1/2} w + \tau^{-3/2} \beta_0 \right\|^2 - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi$$

satisfies $c_n \|\hat{w}^B\| - c_n \tau^{-1} \sigma_x \sigma_\beta \to \alpha_2^*$, *then the same holds true for* $\hat{w}$ *of Eq. (25).*

Lemma 16. *Let* $\hat{w}$ *denote an optimal solution of Eq. (C6). Regarding* $\phi(g, h)$ *and* $\phi_{\tilde{\mathcal{S}}_n^c}(g, h)$, *as introduced and discussed in relation to Eq. (C7), suppose there are constants* $\bar{\phi}$ *and* $\bar{\phi}_{\tilde{\mathcal{S}}_n^c}$ *with* $\bar{\phi} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c}$, *such that for all* $\eta > 0$, *the following hold w.a.p.1 as* $n \to \infty$: *(a)* $\phi(g, h) < \bar{\phi} + \eta$, *(b)* $\phi_{\tilde{\mathcal{S}}_n^c}(g, h) > \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \eta$. *Under these conditions, we have* $\hat{w} \in \tilde{\mathcal{S}}_n$ *w.p.a.1.*

Lemma 17. *The objective function of Eq. (C10) is convex in* α *and jointly concave in* (δ, v).

Proof. First, we prove the objective function is convex in $\alpha = \|w\|$. Revisiting the term $f(\alpha, u) := \frac{c_n \tau^{1/2}}{\sqrt{n}} \alpha g^\top u - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4}$ in Eq. (C7), we observe that it is convex in α. After maximizing over the direction of u, the term remains convex in α since $\max_{\|u\|=\delta} f(\theta\alpha_1 + (1-\theta)\alpha_2, u) \le \max_{\|u\|=\delta} \{\theta f(\alpha_1, u) + (1-\theta) f(\alpha_2, u)\} \le \theta \max_{\|u\|=\delta} f(\alpha_1, u) + (1-\theta) \max_{\|u\|=\delta} f(\alpha_2, u)$, for $\theta \in (0, 1)$. Note that from Eq. (C8), $\max_{\|u\|=\delta} f(\alpha, u) = -\frac{c_n \delta^2}{4} \mu_n(\alpha, \delta) + \frac{c_n}{n} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top (\Sigma_1^{-1} - \mu_n(\alpha, \delta) \mathbb{I})^{-1} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)$, which yield the first two terms in Eq. (C10). Meanwhile, the term $-\|c_n \lambda v - n^{-1/2} \tau^{1/2} \delta h\| \alpha$ is also convex in α. Consequently, we deduce that the objective function of Eq. (C10) is convex in α.

Next, we demonstrate that this function is jointly concave in (δ, v). It is easy to verify that $-\|c_n \lambda v - n^{-1/2} \tau^{1/2} \delta h\| \alpha$ is jointly concave in (δ, v), since $\alpha \ge 0$. Moreover, $\lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 - \lambda \|\Sigma_2^{1/2} v\|^2 / 4$ is concave in v. Therefore, it suffices to prove

$$-\frac{\delta^2}{4} \mu_n(\alpha, \delta) + \frac{1}{n} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top (\Sigma_1^{-1} - \mu_n(\alpha, \delta) \mathbb{I})^{-1} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon) \quad \text{(C11)}$$

is concave in δ. Let the eigenvalues and normalized eigenvectors of Σ_1 be $\{(\lambda_i, v_i)\}_{i=1}^n$, and let $w_i = (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top v_i$, for $i = 1, 2, \ldots, n$. Then (C11) equals $-\frac{\delta^2}{4} \mu_n(\alpha, \delta) +$

$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1/\lambda_i-\mu_n(\alpha,\delta)}w_i^2$. The first order derivative of this equation with respect to δ is

$$-\frac{\delta}{2}\mu_n(\alpha,\delta)-\frac{\delta^2}{4}\partial_\delta\mu_n(\alpha,\delta)+\frac{\partial_\delta\mu_n(\alpha,\delta)}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n(\alpha,\delta))^2}w_i^2=-\frac{\delta}{2}\mu_n(\alpha,\delta), \quad \text{(C12)}$$

where the last equation follows from the definition of the function $\mu_n(\alpha,\delta)$:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n(\alpha,\delta))^2}w_i^2=\frac{\delta^2}{4}. \quad \text{(C13)}$$

Further, the second-order derivative with respect to δ can be calculated as: $-\frac{1}{2}\mu_n(\alpha,\delta)-\frac{\delta}{2}\partial_\delta\mu_n(\alpha,\delta)$. By the chain rule of differentiation, $\partial_\delta\mu_n(\alpha,\delta)$ is the reciprocal of $\partial\mu_n\delta$. The latter can be calculated directly using the definition of μ_n via Eq. (C13): $\partial_{\mu_n}\delta=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^2}w_i^2\right)^{-1/2}\cdot\frac{2}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^3}w_i^2$. With this, we can write the second-order derivative as follows:

$$-\frac{1}{2}\mu_n(\alpha,\delta)-\frac{\delta}{2}\partial_\delta\mu_n(\alpha,\delta)=-\frac{1}{2}\mu_n-\frac{1}{2}\cdot\frac{\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^2}w_i^2}{\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^3}w_i^2}=-\frac{1}{2}\cdot\frac{\sum_{i=1}^{n}\frac{1/\lambda_i}{(1/\lambda_i-\mu_n)^3}w_i^2}{\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^3}w_i^2}.$$

Since $\Sigma_1^{-1}-\mu_n\mathbb{I}$ is positive semidefinite, the right-hand-side is no larger than zero, which concludes the proof. □

Lemma 18. *For $\tilde{Q}_n=\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ in Eq. (28), Eq. (29) holds.*

Proof. The notation below is defined in the proof of Theorem 2. Let $\delta_2=\delta_2^*+c_n^{-1/2}\delta_3$. First, we demonstrate that $c_n\tau^{-1}\mu_n(\alpha,\delta)-c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\xrightarrow{\mathrm{P}}0$. Let $f(x):=\frac{1}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-x\mathbb{I})^{-2}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)$. Recall that $\mu_n(\alpha,\delta)$ is the solution to $f(x)=\delta^2/4$. Note that $f(x)$ exhibits a monotonic increase in x when $x\le 1/C_1$. Therefore, it suffices to show that, given any arbitrarily small $\epsilon>0$, w.p.a.1, the following inequalities hold: $c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)-c_n\delta^2\tau/4>c_+>0$ and $c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)-\tau c_n^{-1}\epsilon)-c_n\delta^2\tau/4<c_-<0$, for some constants c_+ and c_-.

By Lemmas 2 and 3, we can deduce the following equations:

$$\begin{aligned}&\frac{c_n\tau^{-1}}{n}\varepsilon^\top\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\varepsilon\\&\quad-\frac{c_n\tau^{-1}}{n}\mathrm{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]=O_{\mathrm{P}}(c_n\tau^{-1}n^{-1/2}),\\&\frac{c_n\tau^2}{n}\alpha^2g^\top\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}g\end{aligned}$$

$$-\frac{c_n\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=O_{\mathrm{P}}(c_nn^{-1/2}),$$

$$\frac{c_n\tau^{1/2}\alpha}{n}\varepsilon\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}g=O_{\mathrm{P}}(c_n\tau^{-1/2}n^{-1/2}).$$

Therefore, using the definition of $f(\cdot)$ we can deduce that:

$$\begin{aligned}&c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)\\&\quad-\frac{c_n\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\&\quad-\frac{c_n\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=O_{\mathrm{P}}(c_n\tau^{-1}n^{-1/2})=o_{\mathrm{P}}(1).\end{aligned}\tag{C14}$$

Note that for sufficiently small x such that $x\|\Sigma_1\|<1$,

$$\begin{aligned}&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-x\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}-\Sigma_\varepsilon^{1/2}\Sigma_1^{1/2}\left(\mathbb{I}+2x\Sigma_1\right)\Sigma_1^{1/2}\Sigma_\varepsilon^{1/2}\right]\\=&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{1/2}\left(\mathbb{I}-x\Sigma_1\right)^{-2}\Sigma_1^{1/2}\Sigma_\varepsilon^{1/2}-\Sigma_\varepsilon^{1/2}\Sigma_1^{1/2}\left(\mathbb{I}+2x\Sigma_1\right)\Sigma_1^{1/2}\Sigma_\varepsilon^{1/2}\right]\\\leq&\tau^{-1}C_1C_\varepsilon\|\left(\mathbb{I}-x\Sigma_1\right)^{-2}-\left(\mathbb{I}+2x\Sigma_1\right)\|\lesssim\tau^{-1}x^2,\end{aligned}$$

where we apply Lemma 4 in the last inequality. As a consequence, we have:

$$\begin{aligned}&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\=&\frac{1}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\tau^{-1}\Sigma_1^2+2\left(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon\right)\Sigma_1^3\right)\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]+O(\tau)\\=&\tau^{-1}\sigma_\varepsilon^2\theta_1+2(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+O(\tau)+o(c_n^{-1}),\end{aligned}\tag{C15}$$

where the last equation follows by Assumption 5. By the same argument, it follows that:

$$\frac{\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=\sigma_x^2\sigma_\beta^2\theta_4+O(\tau)+o(c_n^{-1}).$$

Applying the above estimates to the left-hand-side of (C14), we can deduce that:

$$\begin{aligned}&c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)-\frac{c_n\delta^2\tau}{4}\\=&2c_n(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+c_n\sigma_x^2\sigma_\beta^2\theta_4-\frac{c_n\delta_1^*\delta_2}{2}+o_{\mathrm{P}}(1).\end{aligned}$$

By the definition of $\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)$, the right-hand side of the above equation is positive w.p.a.1. The proof of the other inequality is similar. Hence, we have proved

$$c_n\tau^{-1}\mu_n(\alpha,\delta)-c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\xrightarrow{\mathrm{P}}0. \tag{C16}$$

Next, to analyze $\tilde{Q}_n$, we first investigate the limiting behavior of:

$$-\frac{\delta^2}{4}\mu_n(\alpha,\delta)+\frac{1}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right).$$

By (C16), we have $\|\mu_n(\alpha,\delta)\Sigma_1\|=O_{\mathrm{P}}(\tau)$. Applying Lemma 4 again, we deduce:

$$\begin{aligned}&\|\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-2}-\Sigma_1^2-2\mu_n(\alpha,\delta)\Sigma_1^3-3\mu_n^2(\alpha,\delta)\Sigma_1^4\|\lesssim_{\mathrm{P}}\tau^3\\&\|\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}-\Sigma_1-\mu_n(\alpha,\delta)\Sigma_1^2-\mu_n^2(\alpha,\delta)\Sigma_1^3\|\lesssim_{\mathrm{P}}\tau^3.\end{aligned}$$

Furthermore, by the fact that $\|\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\|=O_{\mathrm{P}}(n\tau^{-1})$ and Eq. (C9), we have

$$\frac{\delta^2}{4}\mu_n(\alpha,\delta)=\frac{1}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\mu_n(\alpha,\delta)\Sigma_1^2+2\mu_n^2(\alpha,\delta)\Sigma_1^3)(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)+O_{\mathrm{P}}(\tau).$$

With a similar approach, we have

$$\begin{aligned}&\frac{1}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)\\&=\frac{1}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1+\mu_n(\alpha,\delta)\Sigma_1^2+\mu_n^2(\alpha,\delta)\Sigma_1^3\right)\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)+O_{\mathrm{P}}(\tau).\end{aligned}$$

As a consequence, based on Lemmas 2 and 3, as well as the definition of α_2 and the fact that $c_n\tau^{-1}\mu_n(\alpha,\delta)-c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\xrightarrow{\mathrm{P}}0$, we have:

$$\begin{aligned}&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)\\&=\frac{c_n}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1-\mu_n^2(\alpha,\delta)\Sigma_1^3\right)\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)+O_{\mathrm{P}}(c_n\tau)\\&=c_n\tau^{-1}\sigma_x^2\sigma_\beta^2-c_n\sigma_\varepsilon^2\theta_3\mu^2(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+2\sigma_x\sigma_\beta\alpha_2+\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2+o_{\mathrm{P}}(1).\end{aligned} \tag{C17}$$

Finally, we examine the remainder term that contributes to $\tilde{Q}_n$:

$$\frac{c_n^2\lambda^2}{4}\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0+\frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right)^\top\left(\frac{\lambda}{4}\Sigma_2+\frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0+\frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right)-\frac{c_n\tau\alpha^2\delta^2}{2\gamma n}\|h\|^2.$$

Using Lemmas 2-3, $p^{1/2}\tau^{-1}n^{-1}q^{-1/2} = o(1)$ by Assumption 4, and the assumptions on Σ_2, this term converges in probability to:

$$\frac{c_n^2\lambda^2\tau^{-2}\sigma_\beta^2}{4p}\operatorname{Tr}\left[\Sigma_2^{1/2}\left(\frac{\lambda}{4}\Sigma_2+\frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}\Sigma_2^{1/2}\right]+c_n\tau\operatorname{Tr}\left[\frac{c_n\lambda^2\alpha^4\delta^2}{4n\gamma_h^2}\left(\frac{\lambda}{4}\Sigma_2+\frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}-\frac{\alpha^2\delta^2}{2\gamma n}\mathbb{I}\right]$$
$$=\frac{c_n\gamma n_1}{2}\tau^{-1}-\frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}-\tau^{-1}\frac{(\delta_1^*)^2\sigma_x^2c_n}{4\lambda}-\frac{c_n\sigma_x^2\delta_1^*\delta_2}{2\lambda}+\frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2+o_n(1),$$

where we apply Lemma 4 and the same argument in proving Eq. (C15). Combining this estimate with (C17) we conclude that

$$\tilde{Q}_n=-\frac{\delta_3^2\theta_1}{4\theta_3}+2\sigma_x\sigma_\beta\alpha_2-\frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}+\frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2+o_{\mathrm{P}}(1).$$

We now proceed to establish Claims (i) to (ii) in Eq. (29). Fix $\alpha_2 \in A$ and $\gamma_1 > 0$, and observe that $\lim_{\delta_3\to\pm\infty}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)\to-\infty$. By the concave version of Lemma 9, we conclude that $\max_{\delta_3\in K_{\delta_3}}\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)\xrightarrow{\mathrm{P}}\max_{\delta_3\in\mathbb{R}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. Since $\tilde{Q}_n$ is jointly concave in (δ_3,γ_1), after maximizing with respect to δ_3, the function should remain concave in γ_1. Moreover, consider the following equation: $\max_{\delta_3\in\mathbb{R}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)=2\sigma_x\sigma_\beta\alpha_2-\frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}+\frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2$. As a result, $\lim_{\gamma_1\to\infty}\max_{\delta_3\in\mathbb{R}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)\to-\infty$. By Lemma 8, we conclude that $\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}}\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)\xrightarrow{\mathrm{P}}\max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. Since $\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ is convex in α_2, it should retain its convexity in α_2 after being maximized with respect to δ_2 and γ_1. Since the above equation holds for any $\alpha_2\in A$, by Lemma 7, we conclude that Claim (i) holds.

The first-order condition implies a unique solution: $\alpha_2^*:=\arg\min_{\alpha_2}\max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)$, which is given by $\theta_2\sigma_x^3\left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2\sigma_\beta}-\frac{\sigma_\beta}{\lambda}\right)$. Thus, Claim (ii) holds true, concluding the proof. □

Lemma 19. *Under the conditions of Theorem 3, there exists a constant $\tilde{c}>0$ that depends solely on fixed constants, such that w.p.a.1, inequality (32) holds. In addition, as $n\to\infty$, for any given fixed $\lambda>0$, Eq. (33) holds.*

Proof. Under the condition that $\lambda_n \geq \epsilon$, using (30) and (31), we have, w.p.a.1,

$$\|\hat{\beta}^i_{\lambda_n}\|^2=\left\|\frac{1}{n}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_n\mathbb{I}\right)^{-1}X_{(-i)}y_{(-i)}\right\|^2\leq\frac{\|X_{(-i)}y_{(-i)}\|^2}{c_n^2\epsilon^2n^2}\leq\frac{2C_2\sigma_\varepsilon^2(1+\sqrt{c_n})^2}{c_n^2\epsilon^2},$$

$$\|\hat{\beta}^i_{\lambda_1}-\hat{\beta}^i_{\lambda_2}\|^2=c_n^2(\lambda_1-\lambda_2)^2\left\|\frac{1}{n}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_2\right)^{-1}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_1\right)^{-1}X_{(-i)}y_{(-i)}\right\|^2$$

$$\leq \frac{(\lambda_1-\lambda_2)^2}{n^2 c_n^2 \lambda_1^2\lambda_2^2}\|X_{(-i)}y_{(-i)}\|^2 \leq \frac{2C_2\sigma_\varepsilon^2(1+\sqrt{c_n})^2(\lambda_1-\lambda_2)^2}{c_n^2\epsilon^4}. \tag{C18}$$

With the inequalities above and triangle inequalities, we obtain, w.a.p.1,

$$|\hat{R}^{K-CV}(\lambda_1) - \hat{R}^{K-CV}(\lambda_2)| = \frac{1}{n}\left|\sum_{i=1}^{K}\left(\|y_{(i)} - X_{(i)}\hat{\beta}^i_{\lambda_1}\|^2 - \|y_{(i)} - X_{(i)}\hat{\beta}^i_{\lambda_2}\|^2\right)\right|$$

$$\leq \frac{2}{n}\sum_{i=1}^{K}\|X_{(i)}\|\|\hat{\beta}^i_{\lambda_1} - \hat{\beta}^i_{\lambda_2}\|\left(\|y_{(i)}\| + \frac{2C_2^{1/2}\sigma_\varepsilon(1+\sqrt{c_n})}{c_n\epsilon}\|X_{(i)}\|\right) \leq \tilde{C}|\lambda_1-\lambda_2|, \tag{C19}$$

where $\tilde{C}$ is some fixed constant.

Let us fix a constant $\tilde{c}$ such that the inequality $\frac{c_2^2\sigma_\varepsilon^2}{2K\tilde{c}^2} - 4C_2^2\frac{(1+\sqrt{c_n})^2}{c_n\tilde{c}} > 100$ remains true as $n, p \to \infty$. This is possible because $(1+\sqrt{c_n})^2/c_n$ is bounded as $n, p \to \infty$. In addition, let $S := \{\lambda_j = \epsilon + p^{-9}(j-1) : 1 \leq j \leq 1 + [p^9(\tilde{c}\tau^{-1} - \epsilon)]\}$. Given $\tau^{-1} = o(p)$, the cardinality of the set satisfies $|S| \leq p^{10}$. By definition, for any $\lambda \in [\epsilon, \tilde{c}\tau^{-1}]$, there exists a $\lambda_{j^*} \in S$ such that $|\lambda - \lambda_{j^*}| \leq p^{-9}$. By Eq. (C19), we have $|\hat{R}^{K-CV}(\lambda) - \hat{R}^{K-CV}(\lambda_{j^*})| \leq \tilde{C}|\lambda - \lambda_{j^*}| \leq \tilde{C}p^{-9}$. Therefore, if we show that

$$\inf_{\lambda_j \in S}\left\{\hat{R}^{K-CV}(\lambda_j) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} > np^{-1}\tau^2 \tag{C20}$$

holds w.p.a.1, we have

$$\inf_{\lambda\in[\epsilon,\tilde{c}\tau^{-1}]}\left\{\hat{R}^{K-CV}(\lambda_j) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} > np^{-1}\tau^2 - \tilde{C}p^{-9} > \frac{np^{-1}\tau^2}{2}, \tag{C21}$$

which implies Eq. (32). By the definition of $\hat{R}^{K-CV}$, it is easy to verify that we need prove:

$$\inf_{\lambda_j\in S}\left\{n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - 2n^{-1}K\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0) - \|\Sigma_2^{1/2}\beta_0\|^2\right\} > np^{-1}\tau^2$$

holds w.p.a.1 for all $i = 1, \ldots, K$. By the independence of $Z_{(i)}$ and $\hat{\beta}^i_{\lambda_j}$, the first term on the left-hand-side is distributed as: $n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 \overset{d}{=} n^{-1}K\chi^2(K^{-1}n)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2$, where $\chi^2(K^{-1}n)$ denotes a Chi-squared random variable with $K^{-1}n$ degrees of freedom. Consequently, we can deduce that:

$$\mathrm{P}\left(\left|n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2\right| \geq \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2\right)$$

$$= \mathrm{P}\left(\left|n^{-1}K\chi^2\left(K^{-1}n\right) - 1\right| \geq \frac{\log(p)}{\sqrt{n}}\right) \leq 2\exp(-\tilde{c}_1 \log^2(p)),$$

where the last step uses Lemma 5, and $\tilde{c}_1$ is a fixed positive constant. Analogously, we have:

$$\mathrm{P}\left(\left|n^{-1}K\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\right| \geq \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|\right) \leq 2\exp(-\tilde{c}_2 \log^2(p)),$$

with $\tilde{c}_2$ being another fixed positive constant. For simplicity, we consolidate the constants $\tilde{c}_1$ and $\tilde{c}_2$ into a unified constant denoted as $\tilde{c}_1$. By the union bound inequality, we have that with probability exceeding $1 - 4p^{10}\exp(-\tilde{c}_1 \log^2(p))$, the following relation holds:

$$\begin{aligned}\inf_{\lambda_j \in S}&\left\{n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - 2n^{-1}K\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0) - \|\Sigma_2^{1/2}\beta_0\|^2\right\} \\ &\geq \inf_{\lambda_j \in S}\left\{\left(1 - \frac{\log(p)}{\sqrt{n}}\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\| - \|\Sigma_2^{1/2}\beta_0\|^2\right\}.\end{aligned}$$

Assume for now that $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 \geq 50np^{-1}\tau^2$ holds. In this scenario, $\left(1 - \frac{\log(p)}{\sqrt{n}}\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|$ is monotonically increasing in $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|$ since $\log(p)/\sqrt{n} = o(\tau)$, hence it achieves its minimum when $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 = 50np^{-1}\tau^2$. As a result, it can be shown that $\left(1 - \frac{\log(p)}{\sqrt{n}}\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\| - \|\Sigma_2^{1/2}\beta_0\|^2 \geq np^{-1}\tau^2$. Therefore, we only need to prove $\inf_{\lambda_j \in S}\{\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\} \geq 50np^{-1}\tau^2$ holds w.p.a.1.

We now establish a uniform lower bound for $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2$, which can be written as: $\|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2 - 2\beta_0^\top\Sigma_2\hat{\beta}^i_{\lambda_j}$. By direct calculation, we have for each i,

$$\begin{aligned}&\|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2 \geq c_2\|\hat{\beta}^i_{\lambda_j}\|^2 = c_2\left\|\frac{1}{n}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}X_{(-i)}y_{(-i)}\right\|^2 \\ &\geq \frac{c_2}{n^2}\left\|\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right\|^{-2}\left\|X_{(-i)}y_{(-i)}\right\|^2 \geq \frac{c_2}{n^2}\left(C_2(1+\sqrt{c_n})^2 + c_n\lambda_j\right)^{-2}\|X_{(-i)}^\top y_{(-i)}\|^2.\end{aligned}$$

Further, by Lemmas 2 and 3, we have

$$\|X_{(-i)}^\top y_{(-i)}\|^2 = \sigma_\varepsilon^2\operatorname{Tr}(X_{(-i)}X_{(-i)}^\top) + p^{-1}\tau\sigma_\beta^2\operatorname{Tr}(X_{(-i)}^\top X_{(-i)}X_{(-i)}^\top X_{(-i)}) + o_{\mathrm{P}}(n^{-1/2}).$$

By the fact that $\lambda_{\min}(A)\operatorname{Tr}(B) \leq \operatorname{Tr}(AB) \leq \lambda_{\max}(A)\operatorname{Tr}(B)$ when A, B are positive semidefinite, we have $c_2\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top) \leq \operatorname{Tr}(X_{(-i)}X_{(-i)}^\top) \leq C_2\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top)$, which, along with

$(np)^{-1}\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top) \xrightarrow{\mathrm{P}} (K-1)/K$ and Eq. (30), imply that

$$p^{-1}\tau \operatorname{Tr}(X_{(-i)}^\top X_{(-i)} X_{(-i)}^\top X_{(-i)}) \le p^{-1}\tau \|X_{(-i)}^\top X_{(-i)}\| \operatorname{Tr}(X_{(-i)}^\top X_{(-i)}) \lesssim_{\mathrm{P}} \tau p n = o_{\mathrm{P}}(np).$$

Therefore, w.p.a.1, we obtain

$$\frac{c_2\sigma_\varepsilon^2 pn}{2K} \le \|X_{(-i)}^\top y_{(-i)}\|^2 \le 2C_2\sigma_\varepsilon^2 pn. \tag{C22}$$

Consequently, uniformly over $\lambda_j \in S$, we deduce:

$$\|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2 \ge \frac{c_2^2\sigma_\varepsilon^2 p}{2nK}\left(C_2(1+\sqrt{c_n})^2 + c_n\lambda_j\right)^{-2}. \tag{C23}$$

On the other hand, we have

$$\begin{aligned} |\beta_0^\top \Sigma_2 \hat{\beta}^i_{\lambda_j}| \le{}& \frac{1}{n}\left|\beta_0^\top \Sigma_2 \left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j \mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\beta_0\right| \\ &+ \frac{1}{n}\left|\varepsilon_{(-i)}^\top X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2\beta_0\right| =: L_1 + L_2. \end{aligned} \tag{C24}$$

To bound L_1, note that $\operatorname{Tr}(AB) \le \|AB\| \operatorname{rank}(AB) \le \|A\|\,\|B\| \operatorname{rank}(B)$, which implies

$$\begin{aligned} &\frac{1}{n}\operatorname{Tr}\left(\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\right) \\ &\le \|\Sigma_2\| \left\|\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}\frac{1}{n}X_{(-i)}^\top X_{(-i)}\right\| \operatorname{rank}(X_{(-i)}^\top X_{(-i)}) \le \frac{nC_2^2(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j}, \end{aligned}$$

where the last inequality uses $\lambda_1((A+\mathbb{I})^{-1}A) = (\lambda_1(A)+1)^{-1}\lambda_1(A)$ and $n^{-1}\|X_{(-i)}^\top X_{(-i)}\| \le C_2(1+\sqrt{c_n})^2$. In addition, by Lemma 5 and the fact that the sub-exponential norm of $b_{0,i}$ is of order $O(q^{-1/2})$, we have, with probability exceeding $1 - 2p^{10}\exp(-\tilde{c}_1\log^2(p))$,

$$\sup_{\lambda_j\in S}\left|L_1 - \frac{p^{-1}\tau}{n}\operatorname{Tr}\left(\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\right)\right| \le q^{-1}p^{-1}\tau n^{1/2}\log(p).$$

Combining the above two inequalities, we have

$$L_1 \le np^{-1}\tau C_2^2 \frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j} + q^{-1}p^{-1}\tau n^{1/2}\log(p), \quad \forall \lambda_j \in S.$$

To bound L_2 in (C24), by definition, we have $L_2 = |n^{-1}p^{-1/2}\tau^{1/2}q^{-1/2}z_{(-i)}^\top X_{(-i)}(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I})^{-1}\Sigma_2(\sqrt{q}b_0)|$. Using the facts that $\lambda_{\min}(n^{-1}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}) \geq c_n\lambda_j \geq c_n\epsilon$, $\|A\|_{\mathrm{F}} \leq \sqrt{\operatorname{rank}(A)}\,\|A\|$, and Eq. (30), we have

$$\big\|X_{(-i)}\big(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\big)^{-1}\Sigma_2\big\| \leq C_2c_n^{-1}\epsilon^{-1}\|X_{(-i)}\| \lesssim np^{-1/2},$$
$$\big\|X_{(-i)}\big(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\big)^{-1}\Sigma_2\big\|_{\mathrm{F}}^2 \lesssim \operatorname{rank}(X_{(-i)})n^2p^{-1} \lesssim n^3p^{-1}.$$

Therefore, by Lemma 5 and the fact that $\sqrt{q}b_{0,i}$ has bounded sub-exponential norm, it holds that, for some constant $\tilde{c}_1$, $\mathrm{P}\left(L_2 > q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p)\right) \leq 2\exp(-\tilde{c}_1\log^2(p))$. As a consequence, with probability at least $1-2p^{10}\exp(-\tilde{c}_1\log^2(p))$, we have $\sup_{\lambda_j\in S} L_2 \leq q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p)$. Therefore, taking bounds for L_1 and L_2 together, we have, w.p.a.1,

$$\begin{aligned}|\beta_0^\top\Sigma_2\hat{\beta}^i_{\lambda_j}| &\leq np^{-1}\tau C_2^2\frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j} + p^{-1}q^{-1}\tau n^{1/2}\log(p) + q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p)\\ &\leq 2np^{-1}\tau C_2^2\frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j},\end{aligned} \tag{C25}$$

for each $\lambda_j \in S$. In the last inequality, we use $p^{-1}q^{-1}\tau n^{1/2}\log(p), q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p) = o(np^{-1}\tau^2)$ by the assumptions of Theorem 3. With (C23) and (C25), we have

$$\begin{aligned}&\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 = \|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2 - 2\beta_0^\top\Sigma_2\hat{\beta}^i_{\lambda_j}\\ \geq&\frac{c_2^2\sigma_\varepsilon^2 p}{2nK}\left(C_2(1+\sqrt{c_n})^2 + c_n\lambda_j\right)^{-2} - 4np^{-1}\tau C_2^2\frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j}.\end{aligned}$$

This inequality holds w.p.a. 1 as $n, p \to \infty$ uniformly for all $\lambda_j \in S$. Given our initial choice for $\tilde{c}$, it is easy to check that the right-hand side exceeds $50np^{-1}\tau^2$, which implies Eq. (32).

To prove Eq. (33), note that

$$\frac{1}{n}\|y_{(i)} - X_{(i)}\hat{\beta}^i_{\tau^{-1}\lambda}\|^2 - \frac{1}{n}\|\varepsilon_{(i)}\|^2 = \frac{1}{n}\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)\|^2 + \frac{2}{n}\varepsilon_{(i)}^\top Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0).$$

By the facts $Z_{(i)} \perp \hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0$ and $n^{-1}\chi^2(K^{-1}n) = K^{-1} + O_p(n^{-1/2})$, we have

$$\begin{aligned}\frac{1}{n}\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)\|^2 &\overset{d}{=}\frac{1}{n}\chi^2(K^{-1}n)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)\|^2\\ &=\frac{1}{K}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)\|^2 + O_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)\|^2\right).\end{aligned}$$

Additionally, by Theorem 2, we deduce:

$$\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2-\|\Sigma_2^{1/2}\beta_0\|^2=\frac{2(K-1)}{K}np^{-1}\tau^2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2}{2\lambda^2}-\frac{\sigma_\beta^2}{\lambda}\right)+o_{\mathrm{P}}(\tau^2np^{-1}).$$

Hence, using the fact that $\|\Sigma_2^{1/2}\beta_0\|^2\asymp_{\mathrm{P}}\tau$, we derive the following equation:

$$\frac{1}{n}\sum_{i=1}^K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2-\|\Sigma_2^{1/2}\beta_0\|^2=\frac{2(K-1)}{K}np^{-1}\tau^2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2}{2\lambda^2}-\frac{\sigma_\beta^2}{\lambda}\right)+o_{\mathrm{P}}(\tau^2np^{-1}).$$

Thus, to prove Eq. (33), it remains to show that: $\frac{2}{n}\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)=o_{\mathrm{P}}(\tau^2np^{-1})$. Given that $\varepsilon_{(i)}\perp Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)$ and $n^{-3/2}\tau^{-3/2}p\to 0$ by Assumption 4, we have

$$\frac{2}{n}\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\overset{d}{=}\frac{2}{n}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|\varepsilon_{(i)}^\top x=O_{\mathrm{P}}(n^{-1/2}\tau^{1/2})=o_{\mathrm{P}}(\tau^2np^{-1}),$$

where x is a standard Gaussian vector independent of $\varepsilon_{(i)}$. This concludes the proof. □

Lemma 20. *There exists a constant $\tilde{C}_1$ such that, w.p.a.1, uniformly for $\mu_1,\mu_2\in[0,\tilde{c}^{-1}]$,*

$$pn^{-1}\tau^{-2}|\widetilde{R}^{K-CV}(\mu_1)-\widetilde{R}^{K-CV}(\mu_2)|\le\tilde{C}_1|\mu_1-\mu_2|+o_{\mathrm{P}}\left(pn^{-1}\tau^{-2}\right).$$

Proof. By the Woodbury identity, we deduce that

$$\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\tau^{-1}\mu^{-1}\mathbb{I}\right)^{-1}-c_n^{-1}\tau\mu\mathbb{I}=-\frac{c_n^{-2}\tau^2\mu^2}{n}X_{(-i)}^\top\left(\mathbb{I}+\frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X_{(-i)}^\top\right)^{-1}X_{(-i)}.$$

Hence, we arrive at:

$$\begin{aligned}
&\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}} c_n\tau^{-3}\log^{-1}(p)\left\|\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\tau^{-1}\mu^{-1}\mathbb{I}\right)^{-1}-c_n^{-1}\tau\mu\mathbb{I}+\frac{c_n^{-2}\tau^2\mu^2}{n}X_{(-i)}^\top X_{(-i)}\right\|\\
&=\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}} c_n^{-1}\mu^2\tau^{-1}\log^{-1}(p)\left\|\frac{1}{n}X_{(-i)}^\top\Big[\left(\mathbb{I}+\frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X_{(-i)}^\top\right)^{-1}-\mathbb{I}\Big]X_{(-i)}\right\|\\
&\le\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}} \mu^3c_n^{-2}\log^{-1}(p)\left\|\frac{1}{n}X_{(-i)}^\top X_{(-i)}\right\|^2\xrightarrow{\mathrm{P}}0.
\end{aligned}\tag{C26}$$

The last inequality is a consequence of Eq. (30) and the fact that

$$\left\|\left(\mathbb{I}+\frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X_{(-i)}^\top\right)^{-1}-\mathbb{I}\right\| \leq \left\|\left(\mathbb{I}+\frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X_{(-i)}^\top\right)^{-1}\right\|\cdot c_n^{-1}\tau\mu\left\|\frac{1}{n}X_{(-i)}^\top X_{(-i)}\right\|$$
$$\leq c_n^{-1}\tau\mu\left\|\frac{1}{n}X_{(-i)}^\top X_{(-i)}\right\|.$$

On the other hand, by direct calculation we have that $\widetilde{R}^{K-CV}(\mu_1)-\widetilde{R}^{K-CV}(\mu_2)$ equals:

$$\sum_{i=1}^K\left(\frac{1}{n}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\|^2-\frac{1}{n}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}}\|^2\right)-\frac{2}{n}y_{(i)}^\top X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})$$
$$=:\sum_{i=1}^K W_{1i}(\mu_1,\mu_2)-W_{2i}(\mu_1,\mu_2).$$

We next investigate $W_{1i}(\mu_1,\mu_2)$ and $W_{2i}(\mu_1,\mu_2)$ separately. For $W_{1i}(\mu_1,\mu_2)$, we have

$$W_{1i}(\mu_1,\mu_2)=\frac{1}{n}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})^\top X_{(i)}^\top X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}+\frac{1}{n}\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}}X_{(i)}^\top X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})$$
$$\leq\frac{1}{n}\left\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\right\|\cdot\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\|+\frac{1}{n}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}}\|\cdot\left\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\right\|.$$

Define $\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}=\frac{1}{n}\left[c_n^{-1}\tau\mu_1\mathbb{I}-\frac{c_n^{-2}\tau^2\mu_1^2}{n}X_{(-i)}^\top X_{(-i)}\right]X_{(-i)}^\top y_{(-i)}$. Observe that

$$\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-X_{(i)}\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}\right\|$$
$$\leq\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{n^{3/2}}\left\|X_{(i)}\right\|\left\|\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\tau^{-1}\mu_1^{-1}\right)^{-1}-c_n^{-1}\tau\mu_1\mathbb{I}+\frac{c_n^{-2}\tau^2\mu_1^2}{n}X_{(-i)}^\top X_{(-i)}\right\|$$
$$\times\left\|X_{(-i)}^\top y_{(-i)}\right\|=O_{\mathrm{P}}(\tau^3\log(p))=o_{\mathrm{P}}(c_n^{-1/2}\tau), \tag{C27}$$

where we use Eq. (C26), Eq. (30), and Eq. (C22). Additionally, it is easy to verify that

$$\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|\frac{1}{n}X_{(i)}\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}\right\|=\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|\frac{1}{n}X_{(i)}\left[c_n^{-1}\tau\mu_1\mathbb{I}-\frac{c_n^{-2}\tau^2\mu_1^2}{n}X_{(-i)}^\top X_{(-i)}\right]X_{(-i)}^\top y_{(-i)}\right\|$$
$$\leq\frac{c_n^{-1}\tau\tilde{c}^{-1}}{n\sqrt{n}}\left\|X_{(i)}X_{(-i)}^\top y_{(-i)}\right\|+\frac{c_n^{-2}\tau^2\tilde{c}^{-2}}{n^2\sqrt{n}}\left\|X_{(i)}X_{(-i)}^\top X_{(-i)}X_{(-i)}^\top y_{(-i)}\right\|.$$

For the first term, by Eq. (C22), it equals

$$\frac{c_n^{-1}\tau\tilde{c}^{-1}}{n\sqrt{n}}\left\|Z_{(i)}\Sigma_2^{1/2}X_{(-i)}^\top y_{(-i)}\right\|\overset{d}{=}\frac{c_n^{-1}\tau\tilde{c}^{-1}}{n\sqrt{n}}\sqrt{\chi^2(n/K)}\left\|\Sigma_2^{1/2}X_{(-i)}^\top y_{(-i)}\right\|\leq\frac{\tilde{C}_1}{2}c_n^{-1/2}\tau,$$

w.p.a.1 for some constant $\tilde{C}_1$ that only depends on fixed constants. The second term can be bounded in the same way. Therefore, we have

$$\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\| \le \tilde{C}_1 c_n^{-1/2}\tau + o_{\mathrm{P}}(c_n^{-1/2}\tau). \tag{C28}$$

Analogously, we can prove that $\frac{1}{\sqrt{n}}\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\| \le \tilde{C}_1|\mu_1-\mu_2|c_n^{-1/2}\tau+o_{\mathrm{P}}(c_n^{-1/2}\tau)$ holds uniformly for $\mu_1,\mu_2\in[0,\tilde{c}^{-1}]$, where $\tilde{C}_1$ is a fixed constant that may vary from line to line. In light of this, we deduce that: $\sup_{1\le i\le K}W_{1i}(\mu_1,\mu_2)\le\tilde{C}_1^2c_n^{-1}\tau^2|\mu_1-\mu_2|+o_{\mathrm{P}}(c_n^{-1}\tau^2)$ holds w.p.a.1 uniformly for $\mu_1,\mu_2\in[0,\tilde{c}^{-1}]$.

To bound $W_{2i}(\mu_1,\mu_2)$, we first define $\tilde{W}_{2i}(\mu_1,\mu_2)=\frac{2}{n}y_{(i)}^\top X_{(i)}(\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\tilde{\beta}^i_{\tau^{-1}\mu_2^{-1}})$. By Eq. (C27), it holds that

$$\begin{aligned}\sup_{\mu_1,\mu_2\in[0,\tilde{c}^{-1}]}|\tilde{W}_{2i}(\mu_1,\mu_2)-W_{2i}(\mu_1,\mu_2)| &\le \frac{2}{n}\|y_{(i)}^\top\|\|X_{(i)}(\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}})\|\\ &+\frac{2}{n}\|y_{(i)}^\top\|\|X_{(i)}(\tilde{\beta}^i_{\tau^{-1}\mu_1^{-2}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\| = O_{\mathrm{P}}(\tau^3\log(p))=o_{\mathrm{P}}(c_n^{-1}\tau^2).\end{aligned}$$

Moreover, employing a similar argument to that used in proving Eq. (C28), we have

$$\begin{aligned}&\sup_{\mu_1,\mu_2\in[0,\tilde{c}^{-1}]}|\tilde{W}_{2i}(\mu_1,\mu_2)|\\ &=\sup_{\mu_1,\mu_2\in[0,\tilde{c}^{-1}]}\left|\frac{2}{n}y_{(i)}^\top X_{(i)}\frac{1}{n}\left[c_n^{-1}\tau(\mu_1-\mu_2)\mathbb{I}-\frac{c_n^{-2}\tau^2(\mu_1^2-\mu_2^2)}{n}X_{(-i)}^\top X_{(-i)}\right]X_{(-i)}^\top y_{(-i)}\right|\\ &\lesssim|\mu_1-\mu_2|\frac{c_n^{-1}\tau}{n^2}\left|y_{(i)}^\top X_{(i)}X_{(-i)}^\top y_{(-i)}\right|+|\mu_1-\mu_2|\frac{c_n^{-2}\tau^2}{n^3}\left|y_{(i)}^\top X_{(i)}X_{(-i)}^\top X_{(-i)}X_{(-i)}^\top y_{(-i)}\right|.\end{aligned}$$

For the first term, by Lemmas 2 and 3, it is easy to verify that

$$\begin{aligned}\frac{c_n^{-1}\tau}{n^2}\left|y_{(i)}^\top X_{(i)}X_{(-i)}^\top y_{(-i)}\right| &\le \frac{c_n^{-1}\tau}{n^2}\left|\varepsilon_{(i)}^\top X_{(i)}X_{(-i)}^\top\varepsilon_{(-i)}\right|+\frac{c_n^{-1}\tau}{n^2}\left|\varepsilon_{(i)}^\top X_{(i)}X_{(-i)}^\top X_{(-i)}\beta_0\right|\\ &+\frac{c_n^{-1}\tau}{n^2}\left|\beta_0^\top X_{(i)}^\top X_{(i)}X_{(-i)}^\top\varepsilon_{(-i)}\right|+\frac{c_n^{-1}\tau}{n^2}\left|\beta_0^\top X_{(i)}^\top X_{(i)}X_{(-i)}^\top X_{(-i)}\beta_0\right|\le\tilde{C}_1c_n^{-1}\tau^2,\end{aligned}$$

for some constant $\tilde{C}_1$ w.p.a.1. The second term can be shown analogously. As a result, we have $\sup_{1\le i\le K}W_{2i}(\mu_1,\mu_2)\le\tilde{C}_1c_n^{-1}\tau^2|\mu_1-\mu_2|+o_{\mathrm{P}}(c_n^{-1}\tau^2)$ w.p.a.1, uniformly for $\mu_1,\mu_2\in[0,\tilde{c}^{-1}]$. Combining the bounds for $W_{1i}(\mu_1,\mu_2)$ and $W_{2i}(\mu_1,\mu_2)$ concludes the proof. □

Lemma 21. *A_1 to A_3 defined in (34) converge to zero w.p.a. 1.*

Proof. For A_1, by using the same argument as Eq. (C18) and noting that $\lambda_n^{opt}\asymp\tau^{-1}$,

$\hat{\lambda}_n^{K-CV} \asymp_{\mathrm{P}} \tau^{-1}$ and $\tau(\lambda_n^{opt} - \hat{\lambda}_n^{K-CV}) = o_{\mathrm{P}}(1)$, we have

$$|A_1| \lesssim c_n\tau^{-2}\|\hat{\beta}_{cv} - \hat{\beta}_{opt}\|\|\hat{\beta}_{cv}\| \lesssim_{\mathrm{P}} c_n\tau^{-2} \cdot c_n^{-1/2}|\lambda_n^{opt} - \hat{\lambda}_n^{K-CV}|\tau^2 \cdot c_n^{-1/2}\tau = o_{\mathrm{P}}(1).$$

Similarly, $A_2 = o_{\mathrm{P}}(1)$. To prove $A_3 = o_{\mathrm{P}}(1)$, define $\tilde{\beta}(\lambda_n) = \frac{1}{n}\left[c_n^{-1}\lambda_n^{-1}\mathbb{I} - \frac{c_n^{-2}\lambda_n^{-2}}{n}X^\top X\right]X^\top y$ and write $\tilde{\beta}(\hat{\lambda}_n^{K-CV})$ as $\tilde{\beta}_{cv}$ and $\tilde{\beta}(\lambda_n^{opt})$ as $\tilde{\beta}_{opt}$ for simplicity. By using a similar result as Eq. (C26) as well as the fact $n^{-1}\|X^\top y\| \lesssim n^{-1}\|X\|\|y\| \lesssim_{\mathrm{P}} c_n^{1/2}$ according to Lemma 6, we have

$$\begin{aligned}
&c_n\tau^{-2}|(\hat{\beta}_{opt} - \tilde{\beta}_{opt})^\top\Sigma_2\beta_0| \lesssim c_n\tau^{-2}\|\hat{\beta}_{opt} - \tilde{\beta}_{opt}\|\|\beta_0\| \\
\lesssim& c_n\tau^{-2}\left\|\left(\frac{1}{n}X^\top X + c_n\lambda_n^{opt}\mathbb{I}\right)^{-1} - c_n^{-1}(\lambda_n^{opt})^{-1}\mathbb{I} + \frac{c_n^{-2}(\lambda_n^{opt})^{-2}}{n}X^\top X\right\|\left\|\frac{1}{n}X^\top y\right\| \cdot \|\beta_0\| \\
=& o_{\mathrm{P}}(c_n\tau^{-2} \cdot c_n^{-1}\tau^3\log(p) \cdot c_n^{1/2} \cdot \tau^{1/2}) = o_{\mathrm{P}}(c_n^{1/2}\tau^{3/2}\log(p)) = o_{\mathrm{P}}(1),
\end{aligned}$$

where the last equation holds by Assumption 4. Similarly, $c_n\tau^{-2}|(\hat{\beta}_{cv} - \tilde{\beta}_{cv})^\top\Sigma_2\beta_0| = o_{\mathrm{P}}(1)$. Therefore, $A_3 = o_{\mathrm{P}}(1)$ follows by $c_n\tau^{-2}(\tilde{\beta}_{opt} - \tilde{\beta}_{cv})^\top\Sigma_2\beta_0 = o_{\mathrm{P}}(1)$. Note that

$$\begin{aligned}
c_n\tau^{-2}(\tilde{\beta}_{opt} - \tilde{\beta}_{cv})^\top\Sigma_2\beta_0 &= n^{-1}\tau^{-2}\big((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1}\big)\beta_0^\top\Sigma_2 X^\top y \\
&- n^{-2}c_n^{-1}\tau^{-2}\big((\hat{\lambda}_n^{K-CV})^{-2} - (\lambda_n^{opt})^{-2}\big)\beta_0^\top\Sigma_2 X^\top X X^\top y =: B_1 + B_2.
\end{aligned}$$

By Lemma 2 and Lemma 3, we have

$$\begin{aligned}
B_1 =& n^{-1}\tau^{-2}\big((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1}\big)\beta_0^\top\Sigma_2 X^\top X\beta_0 + n^{-1}\tau^{-2}\beta_0^\top\Sigma_2((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1})X^\top\varepsilon \\
\asymp_{\mathrm{P}}& n^{-1}\tau^{-1}p^{-1}\big((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1}\big)\operatorname{Tr}(\Sigma_2 X^T X) = o_{\mathrm{P}}(\tau^{-1}\big((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1})\big) = o_{\mathrm{P}}(1).
\end{aligned}$$

The same argument proves $B_2 = o_{\mathrm{P}}(1)$, leading to $A_3 = o_{\mathrm{P}}(1)$, which concludes the proof. $\square$

Lemma 22. *The objecive function in (36) is convex with respect to α and jointly concave with respect to (δ, γ). Additionally, as long as Eq. (38) holds, we have Eq. (35).*

Proof. Define $S_w^n = \{w \big| c_n\tau^{-1}\sigma_x\sigma_\beta + c_\alpha/4\sigma_\beta \le c_n\|w\| \le c_n\tau^{-1}\sigma_x\sigma_\beta + C_\alpha/\sigma_\beta\}$. Analogous to the result proved by Lemma 15, if the solution $\hat{w}^B$ to the following problem

$$\min_{w\in S_w^n} \frac{c_n}{n}\|\tau^{1/2}\Sigma_1^{1/2}Z\tilde{w} - \tau^{-1}\varepsilon\|^2 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\|_1 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi \quad \text{(C29)}$$

satisfies $c_n\|\hat{w}^B\| - c_n\tau^{-1}\sigma_x\sigma_\beta \in [c_\alpha/2\sigma_\beta + \epsilon, C_\alpha/2\sigma_\beta - \epsilon]$ w.a.p.1, then the same holds true for $\hat{w}$, which leads to the desired result, (35). In light of this, without ambiguity we now directly focus on (C29), and refer to $\hat{w}^B$ as $\hat{w}$ for ease of notation.

Note that for any vector x, it holds that $\|x\|^2 = \max_u \sqrt{n} u^\top x - n\|u\|^2/4$, and $\|x\|_1 = \max_{\|v\|_\infty \le 1} v^\top x$. By applying these equations to $\|\tau^{1/2}\Sigma_1^{1/2} Z\tilde{w} - \tau^{-1}\varepsilon\|^2$ and $\|\Sigma_2^{-1/2} w + \tau^{-3/2}\beta_0\|_1$, and letting $\tilde{u} := \Sigma_1^{1/2} u$, the problem (C29) can be reformulated as:

$$\min_{w\in S_w^n} \max_{\substack{\tilde{u} \\ \|v\|_\infty \le 1}} \frac{c_n \tau^{1/2}}{\sqrt{n}} \tilde{u}^\top Z w - \frac{c_n \tau^{-1}}{\sqrt{n}} \tilde{u}^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2}\tilde{u}\|^2}{4} + \frac{c_n \tau^{-2}\lambda_n}{\sqrt{n}} v^\top \beta_0 + \frac{c_n \tau^{-1/2}\lambda_n}{\sqrt{n}} v^\top \Sigma_2^{-1/2} w - \frac{c_n \tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi. \tag{C30}$$

For convenience, we shall continue to employ u in place of $\tilde{u}$ throughout the remainder of the proof. Let $S_u^n = \{u \big| \|u\| \le 4\tau^{-1}\sqrt{C_1 C_\varepsilon}\}$. Similar to the proof of Lemma 14, w.a.p.1,

$$\min_{w\in S_w^n} \max_{\substack{u\in S_u^n \\ \|v\|_\infty \le 1}} \frac{c_n \tau^{1/2}}{\sqrt{n}} u^\top Z w - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4} + \frac{c_n \tau^{-2}\lambda_n}{\sqrt{n}} v^\top \beta_0 + \frac{c_n \tau^{-1/2}\lambda_n}{\sqrt{n}} v^\top \Sigma_2^{-1/2} w - \frac{c_n \tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi \tag{C31}$$

is equivalent to Eq. (C30). Next, we construct an auxiliary optimization problem:

$$\phi(g,h) = \max_{\substack{0\le\delta\le 4\tau^{-1}\sqrt{C_1 C_\varepsilon} \\ \|v\|_\infty \le 1}} \min_{w\in S_w^n} \max_{\|u\|=\delta} \mathcal{R}_n(w,v,u), \quad \text{where}$$

$$\mathcal{R}_n(w,v,u) = \frac{c_n \tau^{1/2}}{\sqrt{n}} \|w\| g^\top u - \frac{c_n \tau^{1/2}}{\sqrt{n}} \|u\| h^\top w - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4} + \frac{c_n \tau^{-2}\lambda_n}{\sqrt{n}} v^\top \beta_0 + \frac{c_n \tau^{-1/2}\lambda_n}{\sqrt{n}} v^\top \Sigma_2^{-1/2} w - \frac{c_n \tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \tag{C32}$$

and both $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are standard Gaussian vectors, independent of all other random variables. Moreover, let $\tilde{\mathcal{S}}_n := \{w \big| c_\alpha/2\sigma_\beta + \epsilon < c_n\|w\| - c_n\tau^{-1}\sigma_x\sigma_\beta < C_\alpha/2\sigma_\beta - \epsilon\}$, define $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$ as the optimal value of the optimization problem (C32), with $w \in S_w^n \cap \tilde{\mathcal{S}}_n^c$.

Lemma 23 characterizes the limiting behavior of the optimal solution to (C31), $\hat{w}$, and in turn, proves the desired (35), under conditions pertaining to the optimization problem (C32). Therefore, we only need show that conditions outlined in Lemma 23 hold as long as (38) holds. That is, under (38), we need to prove the existence of the constants $\bar{\phi} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c}$ such that for all $\eta > 0$, w.a.p.1, $\phi(g,h) < \bar{\phi} + \eta$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h) > \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \eta$.

Following the same argument as in the proof of Lemma 14, after maximizing over the direction of u and minimizing over the direction of w, Eq. (C32) becomes equivalent to:

$$
\begin{aligned}
\max_{\substack{0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}\\ \|v\|_\infty\le 1}} \min_{\alpha\in K_\alpha} & -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}\\
&\times(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)-c_n\left\|n^{-1/2}\tau^{1/2}\delta h-n^{-1/2}\tau^{-1/2}\lambda_n\Sigma_2^{-1/2}v\right\|\alpha\\
&+\frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}}v^\top\beta_0-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2-C_n^\phi,
\end{aligned}
$$

where $K_\alpha=\{\alpha|c_n\alpha-c_n\tau^{-1}\sigma_x\sigma_\beta\in[c_\alpha/4\sigma_\beta,C_\alpha/\sigma_\beta]\}$. By Lemma 17, the objective function of the above optimization problem is convex in α and jointly concave in (δ,v). Consequently, we can interchange the order of min and max by applying Corollary 3.3 in Sion (1958). Applying $\|x\|=\min_{\gamma>0}\frac{1}{2\gamma}\|x\|^2+\frac{\gamma}{2}$ to $\left\|n^{-1/2}\tau^{1/2}\delta h^\top-n^{-1/2}\tau^{-1/2}\lambda_n v^\top\Sigma_2^{-1/2}\right\|\alpha$, and By completing the square for terms associated with v, we can rewrite this problem as (36). As a consequence, we conclude that (36) is convex with respect to α and jointly concave with respect to (δ,γ).

Finally, based on the above argument, we conclude that under (38), for all $\eta>0$, w.a.p.1, $\phi(g,h)<\bar\phi+\eta$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)>\bar\phi_{\tilde{\mathcal{S}}_n^c}-\eta$ by choosing $\bar\phi=-\frac{C_\lambda}{8C_2}$ and $\bar\phi_{\tilde{\mathcal{S}}_n^c}=-\frac{C_\lambda}{100C_2}$, thereby verifying conditions outlined in Lemma 23. □

Next, we introduce a lemma that resembles Lemma 16.

Lemma 23. *Let $\hat w$ denote an optimal solution of Eq. (C31). Regarding $\phi(g,h)$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$, as introduced and discussed in relation to Eq. (C32), suppose there are constants $\bar\phi$ and $\bar\phi_{\tilde{\mathcal{S}}_n^c}$ with $\bar\phi<\bar\phi_{\tilde{\mathcal{S}}_n^c}$, such that for all $\eta>0$, the following hold w.a.p.1 as $n\to\infty$: (a) $\phi(g,h)<\bar\phi+\eta$, (b) $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)>\bar\phi_{\tilde{\mathcal{S}}_n^c}-\eta$. Under these conditions, we have $\hat w\in\tilde{\mathcal{S}}_n$ w.p.a.1.*

Lemma 24. *There exists some sufficiently small $\epsilon>0$, such that for any $\eta>0$, w.p.a.1, the inequalities in (38) hold.*

Proof. By Eq. (C17) in Lemma 18, we have the following result:

$$
\begin{aligned}
&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)\\
&\quad=c_n\tau^{-1}\sigma_x^2\sigma_\beta^2-c_n\sigma_\varepsilon^2\theta_3\mu^2(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+2\sigma_x\sigma_\beta\alpha_2+\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2+o_{\mathrm{P}}(1).
\end{aligned}
$$

Additionally, by Lemmas 2-3 and $p^{1/2}\tau^{-1}n^{-1}q^{-1/2}=o(1)$ by Assumption 4, we deduce that $-\frac{c_n\gamma}{2}+\frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0+\frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0\xrightarrow{\mathrm{P}}-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}$. In the sequel, we examine the asymptotic behavior of the remaining term in $\tilde Q_n(\alpha_2,\delta_3,\gamma_1)$:

$$\min_{\|v\|_\infty \le 1} \left\{ \frac{c_n \alpha^2}{2\gamma} \left\| n^{-1/2}\tau^{-1/2}\lambda_n \Sigma_2^{-1/2} v - n^{-1/2}\tau^{1/2}\delta h - \frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2^{1/2}\beta_0 \right\|^2 \right\}. \tag{C33}$$

By using $\|\Sigma_2^{-1}\| \le c_2^{-1}$, we see (C33) is upper bounded by

$$\begin{aligned}
&\frac{c_n\alpha^2}{2\gamma c_2} \min_{\|v\|_\infty \le 1} \left\{ \left\| n^{-1/2}\tau^{-1/2}\lambda_n v - n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h - \frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0 \right\|^2 \right\} \\
=&\frac{c_n\alpha^2}{2\gamma c_2} \left\| \left(\left| n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h + \frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0 \right| - n^{-1/2}\tau^{-1/2}\lambda_n \right)_+ \right\|^2.
\end{aligned}$$

Similarly, with $\lambda_{\min}\Sigma_2^{-1} \ge C_2^{-1}$, (C33) is lower bounded by $\frac{c_n\alpha^2}{2\gamma C_2}\big\|\big(|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h + \frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0| - n^{-1/2}\tau^{-1/2}\lambda_n\big)_+\big\|^2$. Together with Lemma 25, we deduce that, w.p.a.1, (C33) lies in $\left[\frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1 C_2}C_\lambda, \frac{\sigma_x^2\sigma_\beta^2}{\gamma_1 c_2}C_\lambda\right]$.

Recall that $\tilde{Q}_n(\alpha_2, \delta_3, \gamma_1)$ is defined in (37). We introduce $\tilde{Q}_n^{\text{upper}}(\alpha_2, \delta_3, \gamma_1)$, defined as:

$$\begin{aligned}
&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) \\
&-\frac{c_n\gamma}{2} - \frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1 C_2}C_\lambda + \frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0 + \frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.
\end{aligned}$$

Similarly, we define $\tilde{Q}_n^{\text{lower}}$ with the term $-\frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1 C_2}C_\lambda$ in $\tilde{Q}_n^{\text{upper}}$ being replaced by $-\frac{\sigma_x^2\sigma_\beta^2}{\gamma_1 c_2}C_\lambda$. Consequently, $\tilde{Q}_n^{\text{lower}} \le \tilde{Q}_n \le \tilde{Q}_n^{\text{upper}}$. Note also that $\tilde{Q}_n^{\text{lower}}(\alpha_2, \delta_2, \gamma_1)$ and $\tilde{Q}_n^{\text{upper}}(\alpha_2, \delta_3, \gamma_1)$ maintain their convexity in α_2 and joint concavity in (δ_3, γ_1). By employing a similar line of reasoning as presented in Lemma 18, alongside the definitions of c_α and C_α, it becomes evident that there exists a sufficiently small $\epsilon > 0$ such that

$$\min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n^{\text{upper}} \xrightarrow{\text{P}} \min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} -\frac{\delta_3^2\theta_1}{4\theta_3} + 2\sigma_x\sigma_\beta\alpha_2 - \frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta} - \frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1 C_2}C_\lambda = -\frac{C_\lambda}{8C_2},$$

and $$\min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{c_\alpha}{2\sigma_\beta}+\epsilon] \cup [\frac{C_\alpha}{2\sigma_\beta}-\epsilon, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n^{\text{lower}} \xrightarrow{\text{P}} -\frac{C_\lambda}{100C_2}.$$

These results immediately yield the desired inequalities. □

Lemma 25. *For any $\alpha_2, \delta_3 \in \mathbb{R}$ and $\gamma_1 > 0$, w.p.a.1, we have*

$$\frac{C_\lambda}{2} \le c_n\tau^{-1} \left\| \left(\left| n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h + \frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0 \right| - n^{-1/2}\tau^{-1/2}\lambda_n \right)_+ \right\|^2 \le 2C_\lambda. \tag{C34}$$

Proof. We first establish the following:

$$c_n n^{-1}\tau^{-2}\left\{\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n\right)_+\right\|^2-\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n\right)_+\right\|^2\right\}\xrightarrow{\mathrm{P}}0. \tag{C35}$$

Let $\tilde{h}:=\Sigma_2^{1/2}h\sim\mathcal{N}(0,\Sigma_2)$. Let us denote the (i,j)-th element of Σ_2 as $\Sigma_{2,ij}$, thus we have $\tilde{h}_j|\tilde{h}_i\overset{d}{=}\Sigma_{2,ij}\Sigma_{2,ii}^{-1}\tilde{h}_i+\sqrt{\Sigma_{2,jj}-\Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1$, where g_1 is a standard Gaussian random variable independent of $\tilde{h}_i$. Consequently, $\mathrm{Cov}\left((|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2,(|\delta_1^*\tilde{h}_j|-\lambda_n)_+^2\right)$ equals

$$\begin{aligned}&\mathbb{E}\left\{(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\mathbb{E}\left[(|\delta_1^*\tilde{h}_j|-\lambda_n)_+^2-\mathbb{E}(|\delta_1^*\tilde{h}_j|-\lambda_n)_+^2|\tilde{h}_i\right]\right\}\\=&\mathbb{E}\left\{(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\mathbb{E}\left[\eta(\tilde{h}_i)^2-\eta(g_2)^2\Big|\tilde{h}_i\right]\right\},\end{aligned}$$

where $\eta(x):=\left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}x+\sqrt{\Sigma_{2,jj}-\Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right|-\lambda_n\right)_+$ and $g_2\sim\mathcal{N}(0,\Sigma_{2,ii})$ is independent of both g_1 and $\tilde{h}_i$. In addition, note that $\left|\eta(\tilde{h}_i)^2-\eta(g_2)^2\right|\le|\delta_1^*\Sigma_{2,ij}\Sigma_{2,ii}^{-1}(\tilde{h}_i-g_2)|\cdot\left|\eta(\tilde{h}_i)+\eta(g_2)\right|$. Applying the Cauchy-Schwarz inequality to the above inequality yields

$$\begin{aligned}&\mathbb{E}\Big[\eta(\tilde{h}_i)^2-\eta(g_2)^2\Big|\tilde{h}_i\Big]\lesssim\left(\mathbb{E}\left(|\Sigma_{2,ij}\Sigma_{2,ii}^{-1}(\tilde{h}_i-g_2)|^2\Big|\tilde{h}_i\right)\right)^{1/2}\left(\mathbb{E}\left(\eta(g_2)^2+\eta(\tilde{h}_i)^2\Big|\tilde{h}_i\right)\right)^{1/2}\\&\lesssim|\Sigma_{2,ij}\Sigma_{2,ii}^{-1}|\sqrt{\Sigma_{2,ii}+\tilde{h}_i^2}\sqrt{\Sigma_{2,ij}^2\Sigma_{2,ii}^{-2}\tilde{h}_i^2+\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2}\\&\lesssim|\Sigma_{2,ij}|(1+|\tilde{h}_i|)\left(|\Sigma_{2,ij}||\tilde{h}_i|+\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2}\right),\end{aligned}$$

where the last step is due to $c_2\le\Sigma_{2,ii}\le C_2$. Therefore, by Lemma 11, we have

$$\begin{aligned}&\mathrm{Cov}\left((|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2,(|\delta_1^*\tilde{h}_j|-\lambda_n)_+^2\right)\\\lesssim&|\Sigma_{2,ij}|(\lambda_n+1)\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2}\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2+\Sigma_{2,ij}^2(\lambda_n^2+\lambda_n)\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2.\end{aligned}$$

Further, by Lemma 10 and Eq. (9), we have $\lambda_n=o(\log(p))$. The above inequality leads to:

$$\begin{aligned}&\mathrm{Var}\left(c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n\right)_+\right\|^2\right)=\sum_{i,j=1}^p c_n^2 n^{-2}\tau^{-4}\,\mathrm{Cov}\left((|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2,(|\delta_1^*\tilde{h}_j|-\lambda_n)_+^2\right)\\&\lesssim c_n^2 n^{-2}\tau^{-4}\log(p)\sum_{i=1}^p\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\left(\sum_{j=1}^p|\Sigma_{2,ij}|\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2}\right)\\&+c_n^2 n^{-2}\tau^{-4}(\log(p))^2\sum_{i=1}^p\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\sum_{j=1}^p\Sigma_{2,ij}^2\end{aligned}$$

$$\leq c_n^2 n^{-2}\tau^{-4}\Bigg\{\log(p)C_2\sum_{i=1}^p \mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\left(\sum_{j=1}^p \mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2\right)^{1/2}$$
$$+(\log(p))^2C_2^2\sum_{i=1}^p \mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\Bigg\} = O\left(c_n^{1/2}n^{-1/2}\tau^{-1}\log(p)+c_n n^{-1}\tau^{-2}\log^2(p)\right)=o_n(1),$$

where we use $\sum_{j=1}^p \Sigma_{2,ij}^2 \leq C_2^2$ and Cauchy–Schwartz inequality in the second step. This leads to Eq. (C35). Using the same approach, we can prove

$$\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n+\log^{-1}(p)\right)_+\right\|^2-\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n+\log^{-1}(p)\right)_+\right\|^2=o_{\mathrm{P}}(c_n^{-1}n\tau^2),$$
$$\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n-\log^{-1}(p)\right)_+\right\|^2-\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n-\log^{-1}(p)\right)_+\right\|^2=o_{\mathrm{P}}(c_n^{-1}n\tau^2).$$

Now we are ready to establish Eq. (C34). Note that w.p.a.1, we have

$$c_n\tau^{-1}\left\|\left(\left|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h+\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right|-n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2$$
$$\leq 2c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n+\log^{-1}(p)\right)_+\right\|^2,$$

where the inequality is given by Lemma 12 and the facts that $\tau\log^4(p)=o(1)$ and $n^{1/2}\tau^{1/2}p^{-1/2}q^{-1/2}\log^2(p)=o(1)$ by Assumption 4. Similarly, the left-hand-side is lower bounded by $\frac{1}{2}c_n n^{-1}\tau^{-2}\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h|-\lambda_n-\log^{-1}(p)\big)_+\big\|^2$ w.p.a.1. Finally, by Lemma 10 and the fact that $\lambda_n=o(\log(p))$, it is not hard to verify that $\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h|-\lambda_n-\log^{-1}(p)\big)_+\big\|^2=\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h|-\lambda_n\big)_+\big\|^2\big(1+o(1)\big)$ and $\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h|-\lambda_n+\log^{-1}(p)\big)_+\big\|^2=\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h|-\lambda_n\big)_+\big\|^2\big(1+o(1)\big)$. Together with the definition that $C_\lambda=\lim_{n\to\infty}pn^{-2}\tau^{-2}\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h|-\lambda_n\big)_+\big\|^2$ given by Eq. (9), we conclude the proof. □

Lemma 26. *Eq. (43) defined in the proof of Theorem 5 holds as $n\to\infty$.*

Proof. Note that $\|\hat{\beta}-\widetilde{\beta}\|^2$ is upper bounded by

$$\frac{2}{n^2}\|R_U U^\top\mathcal{M}_W(U\beta_0+\epsilon)-R_U U^\top(U\beta_0+\epsilon)\|^2+\frac{2}{n^2}\|(R_U-\widetilde{R}_U)U^\top(U\beta_0+\epsilon)\|^2$$
$$\leq\frac{4}{n^2}\|R_U U^\top\mathcal{M}_W\varepsilon-R_U U^\top\varepsilon\|^2+\frac{4}{n^2}\|R_U U^\top\mathcal{M}_W U\beta_0-R_U U^\top U\beta_0\|^2$$
$$+\frac{4}{n^2}\|(R_U-\widetilde{R}_U)U^\top\varepsilon\|^2+\frac{4}{n^2}\|(R_U-\widetilde{R}_U)U^\top U\beta_0\|^2.$$

For the first term, using $\|R_U\| \lesssim_{\mathrm{P}} np^{-1}\tau$, $\|U^\top U\| \lesssim_{\mathrm{P}} p$, and Lemma 2, we have

$$\frac{4}{n^2}\|R_U U^\top \mathcal{M}_W \varepsilon - R_U U^\top \varepsilon\|^2 \asymp_{\mathrm{P}} \frac{1}{n^2}\operatorname{Tr}((\mathcal{M}_W U - U)R_U^2(U^\top \mathcal{M}_W - U^\top))$$
$$=\frac{1}{n^2}\operatorname{Tr}(U^\top W(W^\top W)^{-1}W^\top U R_U^2) \leq \frac{1}{n^2}\|R_U^2\|\|U^\top U\|\operatorname{Tr}(W(W^\top W)^{-1}W^\top) \lesssim_{\mathrm{P}} p^{-1}\tau^2 \operatorname{rank}(W).$$

Similarly, it can be shown that the second term is of order $O_{\mathrm{P}}(p^{-1}\tau^2 \operatorname{rank}(W))$. In addition, by Lemma 2, and using the fact that $\operatorname{Tr}(AB) \leq \|A\|\operatorname{Tr}(B)$ and $\|\widetilde{R}_U\| \lesssim_{\mathrm{P}} np^{-1}\tau$, we have

$$\frac{4}{n^2}\|(R_U - \widetilde{R}_U)U^\top \varepsilon\|^2 \asymp_{\mathrm{P}} \frac{1}{n^2}\operatorname{Tr}(U(R_U - \widetilde{R}_U)^2 U^\top) \leq \frac{1}{n^2}\|U^\top U\|\operatorname{Tr}((R_U - \widetilde{R}_U)^2)$$
$$\lesssim_{\mathrm{P}} \frac{p}{n^2}\operatorname{Tr}((R_U(\widetilde{R}_U^{-1} - R_U^{-1})\widetilde{R}_U)^2) = \frac{p}{n^4}\operatorname{Tr}((R_U U^\top W(W^\top W)^{-1}W^\top U\widetilde{R}_U)^2)$$
$$\leq \frac{p}{n^4}\|R_U\|^2\|\widetilde{R}_U\|^2\|U^\top U\|^2\operatorname{Tr}(W(W^\top W)^{-1}W^\top) \lesssim_{\mathrm{P}} p^{-1}\tau^4 \operatorname{rank}(W).$$

Similarly, the final term is of order $O_{\mathrm{P}}(p^{-1}\tau^4 \operatorname{rank}(W))$. To sum up, we have $\|\hat{\beta} - \widetilde{\beta}\|^2 = O(p^{-1}\tau^2 \operatorname{rank}(W)) = o(n^2p^{-2}\tau^3)$, since $\operatorname{rank}(W) = o(n^2p^{-1}\tau)$. □

References

Bai, Z. and J. Silverstein (2009). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer New York.

Chen, B. and G. Pan (2012). Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli 18*(4), 1405 – 1420.

Gander, W., G. H. Golub, and U. von Matt (1989). A constrained eigenvalue problem. *Linear Algebra and its Applications 114-115*, 815–839.

Giannone, D., M. Lenza, and G. E. Primiceri (2022). Economic predictions with big data: The illusion of sparsity. *Econometrica 89*(5), 2409–2437.

Götze, F., H. Sambale, and A. Sinulis (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability 26*(none), 1 – 22.

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

Liese, F. and K. Miescke (2008). *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer Series in Statistics. Springer New York.

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*. USA: Society for Industrial and Applied Mathematics.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics 8*(1), 171 – 176.

Thrampoulidis, C., E. Abbasi, and B. Hassibi (2018). Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory 64*(8), 5592–5628.

Thrampoulidis, C., S. Oymak, and B. Hassibi (2015). Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, Volume 40 of *Proceedings of Machine Learning Research*, Paris, France, pp. 1683–1709. PMLR.