Can Machines Learn Weak Signals?*

Zhouyu Shen† Dacheng Xiu‡

Booth School of Business
University of Chicago

Booth School of Business
University of Chicago and NBER

December 11, 2024

Abstract

In high-dimensional regression scenarios with low signal-to-noise ratios, we assess the predictive performance of several prevalent machine learning algorithms. Theoretical insights show Ridge regression's superiority in exploiting weak signals, surpassing a zero benchmark. In contrast, Lasso fails to exceed this baseline, indicating its learning limitations. Simulations reveal that Random Forest generally outperforms Gradient Boosted Regression Trees when signals are weak. Moreover, Neural Networks with ℓ_2-regularization excel in capturing nonlinear functions of weak signals. Our empirical analysis across six economic datasets suggests that the weakness of signals, not necessarily the absence of sparsity, may be Lasso's major limitation in economic predictions.

Keywords: Weak Signals, Precise Error, Machine Learning, Bayes Risk

*We benefited tremendously from discussions with Gustavo Greire, Ulrich Muller, Mikkel Plagborg-Moller, Alberto Quaini, Pragya Sur, as well as seminar and conference participants at Aarhus University, Duke University, Tsinghua University, Bates White LLC, ESIF Economics and AI+ML Meeting, Triangle Econometrics Conference, Applied Machine Learning, Economics, and Data Science Webinar, the Stevanovich Center Conference on Big Data and Machine Learning in Econometrics, Finance, and Statistics, the fifth International Workshop in Financial Econometrics, and Young Econometricians in Asia-Pacific Annual Meeting.

†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.

1 Introduction

In regression analysis, covariates with non-zero coefficients are recognized as true signals, while those with zero coefficients are considered false signals. In a population model, this distinction is clear-cut, resembling a "black and white" scenario. However, in finite samples, the presence of minuscule non-zero coefficients introduces a "gray" area, blurring the lines between true and false signals.[1] This gray area represents weak signals—covariates that exert negligible influence on the outcome variable.

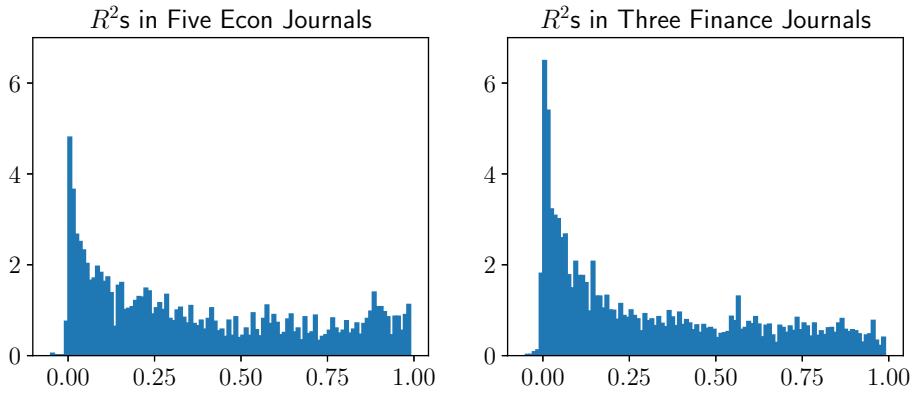
The investigation of weak signals holds tangible implications for economic and financial decision-making. Often, it is the collective impact of these weak signals that drives the outcomes in these fields. Supporting this, Figure 1 offers an empirical perspective, showcasing R^2 values gathered from a compendium of Economics and Finance journal articles published in 2022. The 25% quantiles of these R^2 values stand at 9.7% for economics and 5.8% for finance, suggesting that models in these disciplines frequently rely on covariates with modest explanatory power. Moreover, Figure 1 focuses solely on published papers, which likely skews towards studies with higher R^2 values due to selection bias. This suggests that the presence of weak signals may be even more widespread than the data here indicates.

The decision to incorporate weak signals into a regression model is fraught with the peril of overfitting, which can undermine predictive performance. This issue arises when the errors associated with estimating the coefficients of these weak signals outweigh the benefits of reducing bias that their inclusion offers. To include these variables or not thereby hinges on a trade-off between bias and variance. Compounding this challenge is the increasing prevalence of high-dimensional covariates in data-rich environments, a scenario frequently encountered recently, which can further exacerbate prediction errors due to the scarcity in terms of the sample size relative to the dimensionality of covariates.

Machine learning methods, known for their emphasis on variable selection and dimension reduction, have proven effective in mitigating overfitting and detecting true signals from false ones, particularly when the true signals are strong. These methods employ regularization techniques, such as penalizing the ℓ_1 or ℓ_2 norms of model parameters, to achieve this objective. A pivotal question arises: Can machines learn weak signals, or in other words, can they surpass the naive zero-estimator? The zero estimator, designed to ignore all covariates, serves as a passive baseline in the context of weak signals. If an estimator manages to surpass this baseline, it implies that it has effectively learned valuable signals. Conversely, failing to

[1] The comparison of the magnitudes of regression coefficients becomes meaningful only when the predictor variables have been normalized. This premise is implicitly assumed in our subsequent discussion.

Figure 1: Histograms of R^2s in Selected Economics and Finance Journals



Note: The histograms depict R^2s manually collected from published papers in a selection of Economics and Finance journals in 2022. This collection comprises data from five Economics journals (left) and three Finance journals (right). Specifically, there are a total of 411 papers published in the American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies, resulting in $8,129$ R^2 observations. In addition, there are 380 papers from the Journal of Finance, Journal of Financial Economics, and Review of Financial Studies, contributing $12,198$ R^2 observations.

do so indicates a deficiency in its learning capacity.

In view of these considerations, we shift our focus to evaluate the relative performance of regularization techniques, particularly Ridge and Lasso estimators, against the zero estimator within high-dimensional regression contexts, where the dependent variable is driven by predictors that exhibit weak correlations with it.

In scenarios with sufficiently strong signals, both Lasso and Ridge estimators are expected to outperform the zero estimator by effectively capturing and utilizing at least some of these signals. Hence, accurately defining the notion of "weak" signals is crucial at the outset of our investigation. This definition serves a dual purpose: it prevents the scenario from defaulting to trivial comparisons akin to strong signal cases and ensures practical relevance to finite sample scenarios. We characterize a weak signal scenario as one in which the zero estimator, which always predicts the value zero, achieves the minimal Bayes prediction risk asymptotically. This setting turns out to encompass a sufficiently wide class of data generating processes (DGPs), and it approximates a finite sample reality in which the zero predictor serves as a competitive benchmark. If the zero predictor were not optimal, there would exist an estimator that dominates it, having identified some of the useful signals. In such instances, we would classify these scenarios as cases of strong signals.

In the defined weak signal scenarios, conventional error-bound analyses are insufficient in distinguishing the performance of different estimators. All three estimators — Lasso, Ridge, and the zero estimator — can attain the optimal Bayes prediction error, rendering them asymptotically indistinguishable. Intuitively, this occurs as the tuning parameters for both Ridge and Lasso estimators approach infinity, where Ridge tends toward zero, while Lasso becomes equivalent to zero.

To compare the predictive performance of these estimators, we employ a precise error analysis approach, which enables us to zoom in and explicitly characterize the *relative* differences in the asymptotic behavior of zero, Ridge, and Lasso. This nuanced analysis reveals that with an appropriately chosen tuning parameter, Ridge asymptotically outperforms the zero estimator. In contrast, Lasso does not surpass the zero estimator's performance, regardless of its tuning parameter choice. This finding aligns with the intuition that shrinkage methods like Ridge are more effective in environments with more homogenous signal strength. On the other hand, selection methods like Lasso are preferable in scenarios where there is a clear distinction between true and false signals. In weak signal contexts where this distinction is blurred, the advantage of Lasso tends to wane.

The assumptions about the DGPs in our analysis are sufficiently versatile to encompass a range of scenarios, including spike-and-slab models (George and McCulloch (1993)), where traditionally, Lasso has been the method of choice. Moreover, our analysis permits the cases where regression coefficients follow a Gaussian distribution, a scenario where Ridge regression represents the posterior mean from a Bayesian perspective. This generality in DGP ensures that our findings do not inherently favor one estimator over the other, providing a balanced evaluation of their respective capabilities and limitations in weak signal scenarios.

Our study further emphasizes the validity of the cross-validation algorithm in identifying the optimal tuning parameter for Ridge regression, even in contexts with weak signals. This suggests that cross-validation remains a robust tool for model tuning, resilient to variations in signal strength. Moreover, we find that in the optimal Ridge regression, the out-of-sample R^2, a metric frequently used to evaluate the performance of different estimators on unseen data, proves a relevant indicator of the signal-to-noise ratio in the DGP, despite a notable gap between its asymptotic limit and the population R^2 inherent to the underlying regression.

In the final aspect of our theoretical analysis, we expand our framework to include models featuring a mix of signal strengths. This section specifically addresses scenarios in which a benchmark model contains potentially strong signals. We then shift our focus to assess benefit in harnessing predictive power from the weak signals that remain. To this end, we

derive ordinary least squares (OLS) residuals, from which the impact of potentially strong covariates in the benchmark model has been removed. Consistent with our earlier findings, applying Ridge regression to these residuals, using the remaining covariates, enhances predictive performance compared to a baseline estimator that ignores these additional covariates.

Our simulation analysis corroborates our theoretical findings: the Ridge estimator surpasses zero, which in turn edges out Lasso, especially in DGPs characterized by low R^2 values. Moving to more sophisticated machine learning techniques, we find that Random Forest (RF), yielding a dense model with almost all variables included, outperforms the zero estimator, which itself surpasses Gradient Boosted Regression Trees (GBRT). Resembling Lasso, GBRT tends to produce more sparse models in these scenarios of weak signal strength. Furthermore, Neural Networks (NNs), when paired with the ℓ_2-norm regularization, can yield superior predictions. In contrast, applying an ℓ_1-penalty in these networks does not yield comparable results.

From an empirical standpoint, our analysis covers six datasets derived from macroeconomics, microeconomics, and finance. Five of these datasets are in line with those used by Giannone et al. (2022), and one is sourced from Gu et al. (2020). Our finance examples delve into predicting market returns using financial and economic indicators, as well as firm-level return prediction based on their specific characteristics. In the macroeconomic context, we examine time-series predictions of industrial production using macroeconomic indicators, and a cross-country GDP growth prediction, utilizing socio-economic, institutional, and geographical factors. Our microeconomic studies focus on crime rate predictions and pro-plaintiff appellate decisions in takings law rulings.

The relevance of weak signals in datasets is contingent on the choice of benchmark models. For instance, when compared to a constant benchmark model, weak signals are revealed in four out of six datasets. Further benchmarking against covariates informed by economic theory reveals weak signals across all datasets, making them particularly well-suited for the application of our asymptotic theory. Drawing from their empirical analysis of these datasets, Giannone et al. (2022) argue that sparsity may be an illusion, as optimal predictive models often rely on a large number of covariates. Our collective theoretical and empirical evidence points to signal weakness as a key factor in the underperformance of Lasso. As our results suggest, even in cases where the majority of signals have zero coefficients in the true DGP, Ridge may still outperform Lasso if the true signals are weak. This comparative analysis of their performances thus does not necessarily offer insights into the nature of the DGP itself.

In light of these findings, we recommend a cautious approach to employing Lasso in

economic and financial settings. Despite its popularity as a modern counterpart to OLS, Lasso's effectiveness may be compromised in scenarios characterized by weak signals. Our study complements the findings of Kolesár et al. (2024), who highlight issues with sparsity-based estimators, such as their lack of invariance to reparametrization and sensitivity to normalizations that are otherwise innocuous to ordinary least squares.

Our paper is closely related to the literature on the theoretical performance of Ridge and Lasso, with two main threads being particularly relevant. The first focuses on error-bound analysis. For Ridge, Hoerl and Kennard (1970) show that the prediction error decreases at a rate of p/n, where p is the number of covariates and n the sample size, with its magnitude tied to the eigen-structure of the design matrix. For Lasso, the prediction error vanishes if $s \log p/n \to 0$, where s is the number of non-zero parameters (Zou, 2006, Zhao and Yu, 2006, Zhang and Huang, 2008, Bickel et al., 2009). However, we consider an asymptotic setting where these error bounds fail to distinguish Ridge and Lasso from the zero estimator, as their leading-order prediction errors are identical. This motivates a more granular, higher-order analysis of prediction errors.

The second, more recent strand of research focuses on determining the precise probability limit of the prediction error for Ridge and Lasso.[2] Bayati and Montanari (2012) employ approximate message passing algorithms to link them with the Lasso estimator and derive its error limit. Alternatively, Thrampoulidis et al. (2015) use the Convex Gaussian Minimax Theory (CGMT) to simplify Lasso's optimization problem, enabling precise error derivation. For Ridge regression, Dicker (2016) provides analogous insights into its prediction error. However, these precise error analyses often rely on stringent parametric assumptions, such as independently Gaussian-distributed design matrix elements. Dobriban and Wager (2018) extend Dicker (2016)'s work by accommodating dependent covariates and non-Gaussian predictors, leveraging universality results from random matrix theory.

This paper is organized as follows. Section 2 presents the main theoretical results regarding Ridge and Lasso regressions. Section 3 conducts simulations to illuminate our theoretical predictions while also expanding the analysis to assess the performance of advanced machine learning methods under weak signals. Lastly, Section 4 provides empirical results supporting

[2]The technique of precise error analysis has provided valuable insights into various machine learning methods. For example, Liang and Sur (2022) use CGMT to examine the properties of minimum ℓ_1-norm interpolation and boosting in linear models. Miolane and Montanari (2021) explore cross-validation for Lasso, while Hastie et al. (2022) investigate minimum ℓ_2-norm interpolation, shedding light on the double-descent phenomenon in neural networks and the benefits of overparameterization. Regarding variable selection, Su et al. (2017) study the false discovery rate of the Lasso path, and Wang et al. (2020) compare the variable selection properties of bridge estimators.

the practical relevance of our theoretical results.

Notation: For any $x \in \mathbb{R}$, we refer to $\max(x, 0)$ as x_+. For any vector x, $\|x\|_0$, $\|x\|_1$, $\|x\|$ and $\|x\|_\infty$ represent its ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms, respectively. For a real matrix A, we use $\|A\|$ and $\|A\|_{\mathrm{F}}$ to denote its spectral norm (or ℓ_2 norm), and the Frobenius norm, that is, $\sqrt{\lambda_{\max}(A^\top A)}$, and $\sqrt{\mathrm{Tr}(A^\top A)}$, respectively. In the case where A is a $p \times p$ matrix, $\lambda_i(A)$ denotes its i-th largest eigenvalue, for $1 \leq i \leq p$. We use the notation $x_n \lesssim y_n$ when there exists a constant C such that $x_n \leq C y_n$ holds for sufficiently large n. Similarly, we use $x_n \lesssim_{\mathrm{P}} y_n$ to denote $x_n = O_{\mathrm{P}}(y_n)$. If $x_n \lesssim y_n$ and $y_n \lesssim x_n$, we write $x_n \asymp y_n$ for short. Similarly, we use $x_n \asymp_{\mathrm{P}} y_n$ if $x_n \lesssim_{\mathrm{P}} y_n$ and $y_n \lesssim_{\mathrm{P}} x_n$.

2 Theoretical Results

2.1 Model Setup

We start with the following linear regression model:

$$y = X\beta_0 + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta_0 \in \mathbb{R}^p$ and $\varepsilon \in \mathbb{R}^n$. Throughout our discussion, X, β_0, and ε are treated as random variables, mutually independent of one another. Central to our analysis is the calculation of the probability limit of the prediction errors, which necessities assuming a (prior) probability distribution on the coefficients. This setting aligns with standard practices in the literature on precise errors, which also connects our analysis of prediction error with Bayes risk.

Our objective is to investigate the predictive performance of machine learning techniques in the presence of weak signals.[3] To accomplish this, we focus on a high-dimensional regression setting characterized by an increasing number of predictors, that is, $p \to \infty$. In such a context, regularization techniques become not just relevant but often necessary due to the challenges posed by the curse of dimensionality.

Moreover, our specific focus is on situations where the signals are weak, characterized by the condition: $\|\beta_0\|^2 \asymp_{\mathrm{P}} \tau \to 0$. The choice to use $\|\beta_0\|$ as the metric for characterizing weak signals is due to its close relationship with the widely-adopted R^2 metric in regression analysis, which provides a familiar and intuitive understanding of signal strength.

[3]Several studies, including Donoho and Jin (2004), Hall and Jin (2010), and Jin and Ke (2016), have explored variable selection in the context of rare and weak signals. However, our focus lies on prediction, particularly in asymptotic settings where identifying non-zero coefficients is infeasible.

Our investigation then delves into an asymptotic analysis within a regime broadly characterized by these two conditions. The exact conditions p, τ, and the sample size n satisfy will be provided in detail once we introduce the baseline predictor.

Now, we proceed to present the assumptions governing the DGP of X:

Assumption 1. *The covariates $X \in \mathbb{R}^{n\times p}$ are generated as $X = \Sigma_1^{1/2} Z \Sigma_2^{1/2}$ for an $n \times p$ matrix Z with i.i.d. standard Gaussian entries, deterministic $n \times n$ and $p \times p$ positive definite matrices Σ_1 and Σ_2.*[4] *In addition, there exist positive constants c_1, C_1, c_2, C_2 such that $c_1 \leq \lambda_i(\Sigma_1) \leq C_1$, $i = 1, 2, \ldots, n$ and $c_2 \leq \lambda_i(\Sigma_2) \leq C_2$, $i = 1, 2, \ldots, p$.*

This assumption effectively accommodates both time series and cross-sectional dependence among the covariates in X, with Σ_1 capturing both heteroskedasticity and autocorrelations, while Σ_2 characterizes cross-sectional correlations. The constraints placed on the eigenvalues of Σ_1 and Σ_2 serve a dual purpose. First, the upper bounds on these eigenvalues eliminate strong time series and cross-sectional dependencies within X. Second, the lower bounds prevent multicollinearity and the scenario where an observation at a particular time point is linearly dependent on observations from other times.

Moreover, since we focus on prediction rather than variable selection, the dependence structure among X does not adversely impact Lasso's predictive performance, even though its variable selection properties are sensitive to strong dependence among covariates. In fact, Lasso can achieve a small prediction error even when the signals are highly correlated. For further discussion, see Section 7.4 of Wainwright (2019). Additionally, our comparison between Ridge and Lasso remains valid even when the covariates of X are independent.

While the Gaussian assumption for X is integral to our use of Gordon's inequality (Gordon (1988)) for Gaussian processes in the proof, it does raise concerns regarding the robustness of our findings when this assumption is not met in real-world scenarios. Our simulation results indicate that the Gaussian assumption appears non-essential and our asymptotic theory approximates finite sample behavior even with relatively small sizes, typically a few hundred observations. This observation aligns with similar findings in random matrix theory, where asymptotic properties initially derived for Gaussian ensembles were subsequently shown to extend to a wider spectrum of random matrices — a phenomenon referred to as the universality, as also noted by Bayati and Montanari (2012). In fact, in cases where Σ_1 is the identity matrix, Dobriban and Wager (2018) demonstrate the feasibility of conducting precise error analysis using random matrix theory, thus bypassing the Gaussian assumption.

[4] We are also able to accommodate random Σ_1 and Σ_2, with an additional assumption that their entries are mutually independent and also independent of Z.

However, their technique appears to be only applicable to the Ridge estimator. Our aim is to conduct a comparative analysis of both Ridge and Lasso under a unified framework.

Next, we specify the assumptions regarding ε. Similar to the case of X, we introduce Σ_ε to account for heteroskedasticity and autocorrelations in the noise.

Assumption 2. *Let $\varepsilon = \Sigma_\varepsilon^{1/2} z$, where z comprises i.i.d. variables with mean zero, variance one and finite fourth moment and Σ_ε is a positive definite matrix satisfying $c_\varepsilon \leq \lambda_i(\Sigma_\varepsilon) \leq C_\varepsilon$, $i = 1, 2, \ldots, n$, for some fixed positive constants c_ε and C_ε.*

If Σ_ε is a diagonal matrix, its spectral norm is evidently bounded under the condition that every element along the diagonal is bounded, i.e., that ε has finite variances. In the appendix, we further establish that even if the noise follows a stationary process characterized by exponentially decaying autocorrelations, the spectral norm of Σ_ε remains bounded.

Under Assumptions 1 and 2, it follows that $\|X\beta_0\| \asymp_{\mathrm{P}} \sqrt{n}\,\|\beta_0\|$ and $\|\varepsilon\| \asymp_{\mathrm{P}} \sqrt{n}$. This indicates that the magnitude of each entry in matrix X and the error term ε neither explode nor vanish asymptotically. Consequently, the magnitude of the signal-to-noise ratio (or prediction R^2) is entirely dictated by $\|\beta_0\|$. Next, we impose an assumption that governs the properties of a large number of parameters collected in β_0:

Assumption 3. *The vector $b_0 = \sqrt{p\tau^{-1}}\beta_0$ comprises i.i.d. random variables, each following a prior probability distribution F belonging to the class $\mathcal{F}$. The class $\mathcal{F}$ is defined such that any included random variable can be represented as $q^{-1/2} b_1 b_2$, where b_1 and b_2 are independent, b_1 follows a binomial distribution $B(1, q)$, and b_2 is a sub-exponential random variable with a mean of zero and a variance denoted as σ_β^2.*

Without loss of generality, we use the term $\sqrt{p\tau^{-1}}$ as the normalization factor, ensuring that $\|\beta_0\|^2 \asymp_{\mathrm{P}} \tau$. This choice of normalization facilitates a clearer interpretation of our results. While the i.i.d. assumption may seem strong, it offers greater transparency by simplifying more complex technical assumptions necessary to derive essential probability bounds. In particular, this assumption allows for important classes of models, such as a spike-and-slab prior for b_0, extensively studied by Giannone et al. (2022) to examine the empirical relevance of sparsity in economic datasets. Each element of b_0 follows a mixed distribution, such as when $q = 1$, with b_2 modeled by $(1 - \upsilon)\psi_0 + \upsilon\psi_1$, where υ, a fixed constant within $[0, 1]$, modulates the mix between the spike (ψ_0) and slab (ψ_1) components of the prior. These components may assume the form of Gaussian distributions, as suggested by George and McCulloch (1993), or Laplace distributions, as explored in Rovcková and George (2018). More generally, the formulation $q^{-1/2} b_1 b_2$ accommodates a spike-and-slab model

with more extreme sparsity, facilitating a scenario where sparsity, q, can vanish ($q \to 0$) through the component $q^{-1/2}b_1$. This scaling, $q^{-1/2}$, ensures the variance of $q^{-1/2}b_1b_2$ remains finite and non-vanishing. Essentially, q dictates the sparsity of β_0: when $\mathrm{P}(b_2 = 0) = 0$, $\|\beta_0\|_0 \asymp_{\mathrm{P}} pq$. In scenarios with strong signals ($\tau = 1$), a DGP with q nearing zero typically favors the Lasso estimator, whereas a q closer to one suggests a preference for the Ridge estimator. Therefore, this framework does not inherently privilege any particular estimator. Our theoretical exploration considers the case when $\tau \to 0$. We will apply this spike-and-slab model in our simulations to validate our theoretical findings.

The underlying assumptions that justify Ridge and Lasso are notably distinct, particularly in the context of error-bound analysis. For instance, the analysis of Lasso often requires the approximate sparsity condition and the restricted eigenvalue condition (see, e.g., Belloni et al. (2013b) and Bickel et al. (2009)). On the other hand, the convergence rate of Ridge's prediction error requires intricate conditions on the eigenvalue structure of the design matrix, as discussed in Tsigler and Bartlett (2023). In contrast, our analysis here compares the asymptotic properties of different estimators within a common framework.

2.2 Estimators

We now turn our attention to the discussion of the estimators. In scenarios involving weak signals, characterized by $\|\beta_0\| \to 0$, a straightforward and natural baseline estimator emerges, that is, the naive zero estimator. This estimator is clearly consistent in terms of the ℓ_2-loss of the estimation error, because the error reduces to $\|\beta_0\|$, which vanishes in this context.

The zero estimator functions as a passive baseline, serving as a benchmark for a scenario where no learning occurs. To surpass the performance of the zero estimator, any alternative estimator must harness some of the available weak signals. This indicates the alternative estimator's capability to successfully identify and leverage predictive signals, even when they are weak. Therefore, to address the earlier question of whether machines can learn weak signals, we need to compare the machine learning method's performance with that of the zero estimator. Only if they can do so can they outperform the naive zero estimator.

In our study, we consider Ridge and Lasso as contenders that leverage machine learning techniques. These methods are widely used benchmarks in practice, owing to their simplicity and universality. An in-depth analysis of these estimators provides valuable insights into their specific regularization techniques, which can be extended to more advanced models.

The Ridge estimator, denoted as $\hat{\beta}_r$, is the solution to the following optimization problem:

$$\hat{\beta}_r(\lambda_n) := \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2, \tag{2}$$

where λ_n is its tuning parameter governing the strength of the regularization. In contrast, the Lasso estimator, denoted as $\hat{\beta}_l$, is defined as:

$$\hat{\beta}_l(\lambda_n) := \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|^2 + \frac{\lambda_n}{\sqrt{n}}\|\beta\|_1. \tag{3}$$

By convention, and without loss of generality, the terms involving penalties are typically scaled by p/n in the case of the Ridge estimator and by $1/\sqrt{n}$ in the case of Lasso.

In addition to Ridge and Lasso, our theoretical results also encompass the ordinary least squares (OLS) estimator and the Ridgeless estimator, both of which correspond to special cases of Ridge when the tuning parameter λ_n is set to zero. When $p \leq n$, the least squares problem yields a unique solution, which is the OLS estimator. However, when $p > n$, the least squares problem has an infinite number of solutions. Among these solutions, the Ridgeless estimator can be regarded as a minimum-norm interpolating linear predictor, aiming to minimize the ℓ_2-norm of β:

$$\hat{\beta}_r(0) = \arg\min_{\beta} \|\beta\|, \quad \text{s.t.} \quad X\beta = y, \tag{4}$$

as noted by Bartlett et al. (2020). It is also possible to explore other interpolators, such as the minimum ℓ_1-norm interpolator studied by Liang and Sur (2022). Future research might extend our analysis to other penalized linear estimators, such as Elastic Net, as introduced by Zou and Hastie (2005), or SCAD by Fan and Li (2001).

2.3 Bayes Risk

With β_0 estimated, it is straightforward to construct corresponding linear predictors. Now, we proceed to define the metric by which we assess various predictors. For any predictor, our primary interest is its Bayes prediction risk. This risk is related to the expected squared prediction error evaluated at a new, independent data point $(x^{\text{new}}, y^{\text{new}})$. In the case of a linear model, we can write the prediction error explicitly as:

$$\begin{aligned}\mathbb{E}_F(y^{\text{new}} - \hat{y}^{\text{new}})^2 &= \sigma_\varepsilon^2 + \mathbb{E}_F\left[(x^{\text{new}})^\top(\hat{\beta} - \beta_0)\right]^2 = \sigma_\varepsilon^2 + \mathbb{E}_F\left\{\mathbb{E}\left[((x^{\text{new}})^\top(\hat{\beta} - \beta_0))^2 | X, y, \beta_0\right]\right\} \\ &= \sigma_\varepsilon^2 + \mathbb{E}_F\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2, \end{aligned} \tag{5}$$

where the subscript in the expectation operator $\mathbb{E}_F(\cdot)$ emphasizes the fact that the expectation is taken with respect to the prior distribution of $b_0 = \sqrt{p\tau^{-1}}\beta_0$. Given that σ_ε^2 does not depend on the estimator, it is the second term in (5) that governs the relative predictive performance of different estimators. Barring $\|\Sigma_2\|$, the prediction risk is closely tied to the estimation error of $\hat{\beta}$, i.e., $\|\hat{\beta} - \beta_0\|$.

Formally, we define the Bayes prediction risk associated with an estimator $\hat{\beta}$ as

$$\mathcal{R}(\hat{\beta}, F) := \mathbb{E}_F \|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2.$$

The predictor that minimizes this Bayes risk is termed the Bayes predictor. In our framework, it is straightforward to show (see Chapter 4 of Berger (1985)) that the Bayes predictor corresponds with the predictor that is derived from the posterior mean of β_0, which is represented as $\mathbb{E}_F(\beta_0|X, y)$. We denote its Bayes risk as $\mathcal{R}(F)$.[5]

Under strong signals, i.e., $\tau = 1$, and suppose that Σ_1, Σ_2 and Σ_ε are identity matrices, $p/n \to c_0 \in \mathbb{R}^+$, significant progress has been made in understanding the asymptotic behavior of Bayes risk. For Ridge regression, notable studies by Dicker (2016) and Dobriban and Wager (2018) have derived the asymptotic limit of the Bayes risk.[6] Similarly, in the case of Lasso, several studies, such as those conducted by Bayati and Montanari (2012) and Thrampoulidis et al. (2018), have established its asymptotic Bayes risk limit.[7]

In Figure 2, we present two heatmaps, with the y and x axes representing various values of p/n and $\|\beta_0\|^2 = \tau\sigma_\beta^2$. The left heatmap illustrates the ratio of Bayes risk between optimal

[5]There exists an extensive body of literature focused on empirical Bayes methods, which explores feasible approaches for implementing $\mathbb{E}_F(\beta_0|X, y)$, in cases where F is unknown, see, e.g., Robbins (1964), Efron (2012), Brown and Greenshtein (2009), and Jiang and Zhang (2009).

[6]The exact form of the limit is given by

$$\lim_{n\to\infty} \mathcal{R}(\hat{\beta}_r(\lambda_n), F) = c_0 m(-\lambda, c_0) + \lambda(\lambda\sigma_\beta^2 - c_0)m'(-\lambda, c_0), \tag{6}$$

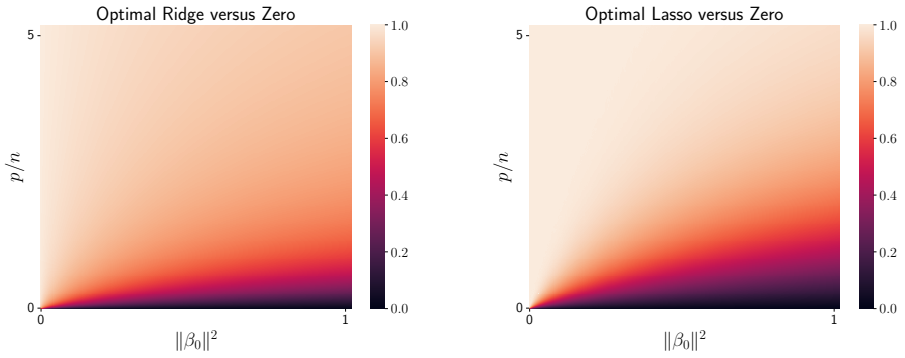
where $\lambda = \lim_{n\to\infty} c_0\lambda_n$ and $m(-\lambda, c_0) = \left(-(1 - c_0 + \lambda) + \sqrt{(1 - c_0 + \lambda)^2 + 4c_0\lambda}\right)/2c_0\lambda$.

[7]The limit in the case of Lasso can be explicitly written as follows: $\lim_{n\to\infty} \mathcal{R}(\hat{\beta}_l(\lambda_n), F) = (\alpha^*)^2$, where

$$\alpha^* = \arg\min_{\alpha\geq 0} \left\{ \inf_{\tau_g>0} \sup_{\substack{\beta\geq 0 \\ \tau_h>0}} \frac{\beta\tau_g}{2} + \frac{1}{c_0} L\left(\alpha, \frac{\tau_g}{\beta}\right) - \frac{\alpha\tau_h}{2} - \frac{\alpha\beta^2}{2\tau_h} + \lambda G\left(\frac{\alpha\beta}{\tau_h}, \frac{\alpha\lambda}{\tau_h}\right) \right\}. \tag{7}$$

Here $\lambda = \lim_{n\to\infty} \lambda_n/\sqrt{c_0}$, $L(c, \tau) := \mathbb{E}[e_{x^2}(cZ + \varepsilon, \tau) - \varepsilon^2]$, and $G(c, \tau) := \mathbb{E}[e_{|x|}(cZ + X_b, \tau) - |X_b|]$, with $\mathrm{e}_f(y, \tau) := \min_v (y - v)^2/2\tau + f(v)$, where random variables Z, ε, and X_b follow a standard normal distribution ($Z \sim \mathcal{N}(0, 1)$), the distribution of the noise, and the distribution F, respectively.

Figure 2: Comparison of Prediction Errors: Optimal Ridge and Lasso vs. the Zero Estimator



Note: The left panel illustrates the ratio of prediction error between the optimal Ridge and the baseline zero estimator. Conversely, the right panel presents a similar comparison for the optimal Lasso estimator against the zero estimator. Both axes, y and x, depict a range of p/n ratios and $\|\beta_0\|^2$ values, corresponding to data generated in accordance with the model described in (1), with Σ_1, Σ_2, and Σ_ε in Assumptions 1 and 2 set as identity matrices. We set b_0 as a Dirac-spike and a Gaussian slab with $q = 1/5$. In this context of strong signals, the prediction errors for both optimal Ridge and Lasso are calculated using tuning parameters that are optimally selected to minimize the expected prediction errors' probability limits, as given by (6) and (7).

Ridge and the zero estimator, while the right heatmap represents the ratio of optimal Lasso against the zero estimator. For both Ridge and Lasso, their optimal tuning parameters are selected by minimizing the probability limits of their Bayes risk given by (6) and (7), respectively. A prediction error ratio below 1 within these visualizations suggests that the zero estimator is outperformed.

The heatmaps, as anticipated, clearly demonstrate that both Ridge and Lasso estimators surpass the performance of the zero estimator. This superiority is particularly pronounced in scenarios involving strong signals and relatively lower dimensions. Notably, the disparity between these estimators becomes less pronounced as the norm of $\|\beta_0\|$ approaches zero and the ratio p/n increases, indicating a shift towards scenarios characterized by weaker signals. The existing result on precise error analysis is primarily built upon the assumption of strong signals, where $\tau = 1$. To discern the performance of various estimators under weak signal conditions, a more intricate analysis in the limiting case ($\tau \to 0$ and $p/n \not\to 0$) is necessary.

2.4 Zero's Optimality and Relative Prediction Error

Figure 2 also indicates that our attention should be directed towards a regime where the zero estimator exhibits meaningful competitiveness. Otherwise, we may question the appropriateness of our definition of "weak" signals if some machine learning approaches can

obviously outperform it by a wide margin.

One might be tempted to define "weak signals" as instances where the signal strength falls below a certain "detection boundary," thereby becoming indiscernible through hypothesis testing.[8] Our primary focus is on prediction, rather than signal detection. This distinction is key because, even when signals are undetectable by hypothesis testing, their collective contribution to prediction can still outperform the zero predictor. The zero predictor serves as a natural benchmark for demonstrating the capacity of machine learning to utilize weak signals.

To motivate our concept of weak signals, we analyze a regime where the zero estimator achieves certain notion of optimality, indicated by its Bayes risk being identical to that of the Bayes predictor. This scenario is delineated more precisely by the assumption below:

Assumption 4. $n^{-1}p \to c_0 \in (0, \infty]$, $n^{-1}\tau p(\log p)^4 \to 0$, $n\tau p^{-2/3}(\log p)^{-4} \to \infty$, *and* $n^{-1}pq\tau^{-1}(\log p)^{-4} \to \infty$.

Assumption 4 covers a wide spectrum of signal strengths and counts, while simultaneously imposing constraints to prevent an excessively large p/n ratio and overly rapid vanishing of τ. The first two constraints imply $\tau \to 0$, while the third imposes a lower bound on τ. Together, these constraints require that τ is bounded below by $n^{-1/3}$.

The final constraint addresses cases of extreme sparsity in β_0. It becomes redundant when q does not vanish, as it is already implied by the first two constraints. Collectively, these conditions imply that $(pq)\log p/n$ is bounded below by τ, and consequently by $n^{-1/3}$ (up to a logarithmic factor). Importantly, these constraints do not entirely exclude the sparsity assumptions commonly adopted in the literature when using Lasso: $\|\beta_0\|_0 \log p/n \to 0$, where $\|\beta_0\|_0 \asymp_{\mathrm{P}} pq$.

These constraints serve to exclude edge cases where the relative performance of different estimators cannot be conclusively determined using our proof technique. In the appendix, we investigate scenarios outside the scope of these constraints, such as cases of extreme sparsity with only one true signal in the DGP. By employing an alternative proof method that leverages the closed-form solution of Lasso in a special case, we demonstrate that these constraints are not necessary for arriving at our conclusions.

To facilitate the discussion of the optimal estimator in our context, we refer to the definition provided by Robbins (1964).

[8]The "detection boundary" in this scenario represents the threshold level of signal strength at which statistical tests can reliably discern the presence of a signal amidst noise. Relevant tests include Ingster et al. (2010), Cui et al. (2018), and Li et al. (2020). However, this boundary generally hinges on the chosen alternative and the maintained hypotheses, presenting a challenge in establishing a unified benchmark.

Definition 1. *We say $\hat{\beta}$ is asymptotically optimal relative to F, if it satisfies*

$$\lim_{n\to\infty} \frac{\mathcal{R}(\hat{\beta}, F)}{\mathcal{R}(F)} = 1.^{9}$$

Theorem 1. *Assume that Assumptions 1–4 hold. Furthermore, assume that the error term ε in Assumption 2 follows a Gaussian distribution.*[10] *Under these conditions, the zero estimator is asymptotically optimal relative to any distribution $F \in \mathcal{F}$.*

This theorem suggests that regardless of the unknown prior distribution, the prediction risk associated with the zero estimator is asymptotically identical to that of the Bayes predictor. Notably, while the Bayes predictor requires knowledge of the prior distribution and is thus infeasible in many practical scenarios, the zero estimator achieves the same level of prediction risk without requiring such information. From this perspective, the zero estimator is both feasible and optimal.

It is noteworthy that the zero estimator can be considered as a particular case of both Ridge and Lasso estimators when a sufficiently large tuning parameter is chosen. Given this perspective, and in accordance with the insights of Theorem 1, the relative Bayes risk of the optimal Ridge and Lasso estimators, in comparison to the zero estimator, is expected to asymptotically approach one. This result suggests that under conditions of weak signals, merely comparing their Bayes risk ratios may not be an effective approach to tell any differences among these estimators.

As such, we shift our attention to the relative prediction error between any estimator $\hat{\beta}$ and the zero estimator, defined as follows, in absolute difference rather than their ratio, in the spirit of Bayesian regret. To ensure a meaningful scale in the limit, we multiply the relative error by $pn^{-1}\tau^{-2}$, and adopt the following metric for comparison:

$$\Delta(\hat{\beta}) = pn^{-1}\tau^{-2}(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2). \tag{8}$$

$\Delta(\hat{\beta})$ magnifies the relative prediction error, a measure that predominantly depends on higher order differences for estimators we consider. Based on this definition, if $\Delta(\hat{\beta}) > 0$ holds with probability approaching one, it indicates that the estimator $\hat{\beta}$ exhibits inferior prediction

[9] The definition of asymptotic optimality is provided in terms of a ratio to accommodate more general scenarios where $\mathcal{R}(F)$ varies with the sample size n.

[10] The Gaussian assumption on ε is only used to facilitate considerations of optimality, which is a standard assumption in the empirical Bayes literature, e.g., Jiang and Zhang (2009). While this assumption motivates our characterization of weak signal regimes, it is not utilized in follow-up analysis of the estimators.

performance compared to zero. Conversely, if $\Delta(\hat{\beta}) < 0$ holds with probability approaching one, it implies that the estimator $\hat{\beta}$ outperforms zero.

Before we proceed to present our main results in the following section, we need to provide technical conditions governing the limiting behavior of Σ_1, Σ_2, and Σ_ε:

Assumption 5. *For matrices Σ_1, Σ_2 and Σ_ε:*

$$\frac{1}{n}\operatorname{Tr}(\Sigma_1) = 1 + O(n^{-1/2}), \quad \frac{1}{p}\operatorname{Tr}(\Sigma_2) = \sigma_x^2 + O(p^{-1/2}), \quad \frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon) = \sigma_\varepsilon^2 + O(n^{-1/2}).$$

Additionally, there exist constants θ_1 to θ_4 such that

$$\begin{aligned}
&\frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1) = \sigma_\varepsilon^2\theta_1 + o(n\tau/p), \\
&\frac{1}{p}\operatorname{Tr}(\Sigma_2^2) = \sigma_x^4\theta_2 + o(1), \quad \frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1^2) = \sigma_\varepsilon^2\theta_3 + o(n/p), \quad \frac{1}{n}\operatorname{Tr}(\Sigma_1^2) = \theta_4 + o(n/p).
\end{aligned}$$

As Σ_1, Σ_2, and Σ_ε are positive definite, all of these constants θ_i, where $i = 1, 2, 3,$ and 4, are positive. The condition concerning Σ_2 can be verified through a more primitive condition often found in the literature—namely, the existence of the limit of Σ_2's empirical spectral distribution, as assumed by Dobriban and Wager (2018). Regarding the conditions concerning Σ_1 and Σ_ε, we establish in the appendix (Lemma 23) that when the time series of covariates and noise are stationary with exponentially decaying correlations, these conditions hold. In situations where all three matrices reduce to identity matrices, which is a common scenario in the literature on precise error analysis, Assumption 5 holds trivially.

2.5 Analysis of the Ridge Estimator

In this section, we present the results of the Ridge estimator in the context of weak signals. We begin by presenting the relative error of Ridge for any tuning parameter value:

Theorem 2. *Assuming that Assumptions 1–5 hold, and setting $\lambda_n = \tau^{-1}\lambda$, we establish the following convergence result:*

$$\Delta(\hat{\beta}_r(\lambda_n)) \xrightarrow{\mathrm{P}} \alpha^* := 2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2} - \frac{\sigma_\beta^2}{\lambda}\right).$$

This theorem yields several important findings. First, by minimizing α^* with respect to λ, we can determine the optimal tuning parameter value: $\lambda_n^{opt} = \tau^{-1}\sigma_\varepsilon^2\theta_1/\sigma_\beta^2$. Furthermore, with

the optimal tuning parameter in place, α^* is negative, indicating that Ridge can effectively learn weak signals when the tuning parameter is chosen appropriately.

Second, when we set $\lambda \to \infty$ (equivalently, $\tau\lambda_n \to \infty$), the value of α^* converges to zero. This outcome is expected, as the use of a large tuning parameter makes the estimator's performance increasingly resemble that of the zero estimator. Nevertheless, it is noteworthy that $\alpha^* \to 0^-$; in other words, as λ increases, the Ridge estimator consistently outperforms the zero estimator until it gradually becomes indistinguishable from it in the limit.

Third, as $\lambda \to 0$, in which case $\lambda_n = o(\tau^{-1})$, the corresponding value of α^* tends to positive infinity. This indicates that Ridge's performance deteriorates to the point where the Ridgeless estimator (corresponding to $\lambda = 0$) is surpassed by the zero estimator. This is a significant departure from the strong signal setup in which Ridgeless can still outperform the zero estimator, as demonstrated by Hastie et al. (2022). It is important to note that the Ridgeless estimator, defined in the form of no regularization ($\lambda = 0$), is not completely devoid of regularization. It incorporates implicit regularization by yielding the interpolator that achieves the minimum ℓ_2 norm. This inherent form of regularization enables the Ridgeless estimator to effectively control variance, particularly in situations where the number of predictors p exceeds the sample size n, thus ensuring desirable performance in strong signal scenarios. In contrast, under conditions of weak signals, this implicit regularization is insufficient for effective variance control. This inadequacy results in the estimated $\|\hat{\beta}\|$ being substantially larger than $\|\beta_0\|$, leading to the poor performance of $\hat{\beta}$.

Furthermore, given that the Ridgeless estimator is defined as the interpolator that minimizes $\|\hat{\beta}\|$, it follows that all linear interpolators, including, for instance, the one that minimizes the ℓ_1-norm, result in even larger values of $\|\hat{\beta}\|$. Consequently, these interpolators also fail to outperform the zero estimator in contexts with weak signals.

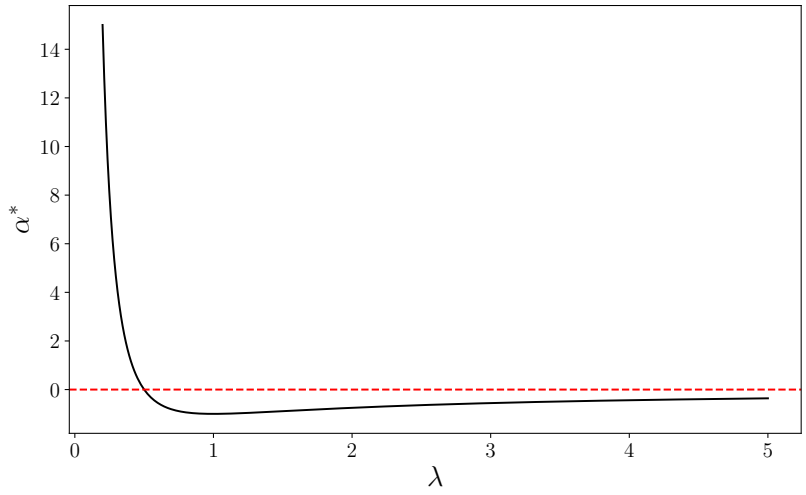
Figure 3 provides an illustrative example of the relationship between the relative error of the Ridge estimator and the tuning parameter λ, showcasing the theoretical insights we have discussed. Corollary 1 below summarizes the result on the Ridgeless estimator:

Corollary 1. *Under the same assumptions as in Theorem 2, the Ridgeless estimator, defined by (4), satisfies:*

$$\Delta(\hat{\beta}_r(0)) \xrightarrow{\mathrm{P}} \infty.$$

Given Ridge regression's ability to effectively learn weak signals with an appropriately tuned parameter, the data-dependent selection of this parameter becomes crucial. A paradigmatic approach for this purpose is K-fold cross-validation (CV).

Figure 3: Ridge vs. Zero Estimator's Relative Precise Error



Note: In this plot, the black curve represents the probability limit of $\Delta(\hat{\beta}_r(\lambda_n))$, denoted as α^*, as a function of the tuning parameter λ, defined in Theorem 2, in the context of weak signals. To create this plot, we set all parameters $(c_0, \theta_1, \theta_2, \sigma_x, \sigma_\beta, \sigma_\varepsilon)$ to one for simplicity.

In cases where the signals are strong, Hastie et al. (2022) demonstrate the effectiveness of CV for Ridge. Specifically, the cross-validated tuning parameter converges in probability to the optimal value within a pre-specified interval. The fact that this optimal value lies in some known interval simplifies the derivation of the theoretical properties of CV. In scenarios with weak signals, however, the optimal tuning parameter tends to diverge as the sample size increases. The rate of divergence depends on the unknown strength of the weak signal, τ. As we show next, CV remains a valid and useful tool in this case. To narrow our focus to the matter of weak signals without delving into a complicated CV procedure, we consider the case where both Σ_ε and Σ_1 are identity matrices. This assumption of no temporal dependence in the data facilitates a more straightforward CV procedure for i.i.d. data.

To determine the optimal tuning parameter using K-fold CV, denoted as $\hat{\lambda}^{K-CV}$, we begin by partitioning the rows of the design matrix X into K distinct subsets, labeled as $X_{(1)}, \cdots, X_{(K)}$. For each index $i \in \{1, \cdots, K\}$, we define $X_{(-i)}$ as the submatrix of X obtained by excluding the rows corresponding to $X_{(i)}$. Similarly, we have associated subvectors $y_{(i)}$, $\varepsilon_{(i)}$, as well as $y_{(-i)}$, $\varepsilon_{(-i)}$. We next define $\hat{\beta}_r^i(\lambda_n)$ for each λ_n as the solution to the Ridge optimization problem for each index $i = 1, \cdots, K$:

$$\hat{\beta}_r^i(\lambda_n) = \arg\min_{\beta} \left\{ \frac{1}{n} \|y_{(-i)} - X_{(-i)}\beta\|^2 + \frac{p\lambda_n}{n} \|\beta\|^2 \right\}.$$

Consequently, the tuning parameter selected by K-fold CV is given by

$$\hat{\lambda}_n^{K-CV} = \underset{\lambda_n \in [\epsilon, \infty)}{\arg\min} \frac{1}{n} \sum_{i=1}^{K} \|y_{(i)} - X_{(i)} \hat{\beta}_r^i(\lambda_n)\|^2,$$

where $\epsilon > 0$ is an arbitrary small constant. The following theorem provides justification for the validity of this CV procedure in the context of weak signals:

Theorem 3. *Under the same assumptions as in Theorem 2, if we also assume that* $\Sigma_1 = \mathbb{I}$, $\Sigma_\varepsilon = \sigma_\varepsilon^2 \mathbb{I}$, ε *follows a sub-exponential distribution, and that* $q^{-1}\tau^{-1}n^{-1/2}\log(p)$, $q^{-1/2}\tau^{-3/2}n^{-1/2}\log(p) \to 0$, *then we can establish that:*

$$\tau \hat{\lambda}_n^{K-CV} \xrightarrow{\text{P}} \lambda^{opt} = \sigma_\varepsilon^2 / \sigma_\beta^2.$$

This theorem justifies the use of $\hat{\lambda}_n^{K-CV}$ as an approximation for the optimal tuning parameter $\lambda_n^{opt} = \lambda^{opt}/\tau$ ($\theta_1 = 1$ in this case) for Ridge. Importantly, this result does not require prior knowledge of τ, making the CV approach directly applicable in practical scenarios. The additional constraints on q become relevant only when q vanishes; otherwise, they naturally follow from Assumption 4. These conditions ensure uniform convergence across the spectrum of tuning parameter values, a prerequisite for the results of Theorem 3. With our analysis of Ridge concluded, we will now turn our attention to Lasso.

2.6 Analysis of the Lasso Estimator

Unlike Ridge, the analysis of Lasso is more intricate, primarily because the Lasso estimator lacks a closed-form formula. In the special case where Σ_1, Σ_2, and Σ_ε are identity matrices, several studies, including Bayati and Montanari (2012) and Thrampoulidis et al. (2018), have established Lasso's precise error given by (7). Additionally, based on (7), Wang et al. (2020) conducted a small-signal Taylor expansion of α^* with respect to σ_β^2, which affects α^* through the prior distribution F. They concluded that the optimal Lasso estimator fails to outperform optimal Ridge.[11] In the general case we consider, pinpointing the exact precise error appears a daunting task. Instead, we seek probability bounds that allow us to characterize the location of the limit. This turns out sufficient for us to conclude that Lasso cannot outperform zero for all values of its tuning parameter in the context of weak signals. The next theorem summarizes our main findings:

[11] Their analysis does not address the scenario of Lasso with an arbitrary tuning parameter, nor does it elucidate its relative performance compared to zero.

Theorem 4. *Assume that Assumptions 1–5 are satisfied and the tuning parameter λ_n is chosen such that the following equation holds for some $C_\lambda > 0$:*

$$pn^{-2}\tau^{-2}\mathbb{E}_{U\sim\mathcal{N}(0,\Sigma_2)}\left\|(2\sigma_\varepsilon\sqrt{\theta_1}|U| - \lambda_n)_+\right\|^2 = C_\lambda.^{12} \tag{9}$$

Then, with probability approaching one, we have $c_\alpha \leq \Delta(\hat{\beta}_l(\lambda_n)) \leq C_\alpha$, where c_α and C_α are the solutions to the following equation in terms of x:

$$x - \sqrt{\frac{2C_\lambda}{c_2}x} = -\frac{C_\lambda}{100C_2}, \tag{10}$$

where c_2 and C_2 are constants defined in Assumption 1.

Equation (9) implicitly determines the rate at which λ_n diverges to infinity. For any fixed $C_\lambda > 0$, we can solve for the tuning parameter λ_n from (9), and derive the upper and lower bounds, C_α and c_α, from equation (10). Furthermore, equation (10) directly implies that C_α and c_α are non-negative, indicating that Lasso does not outperform the zero estimator for any given tuning parameter value in the context of weak signals.

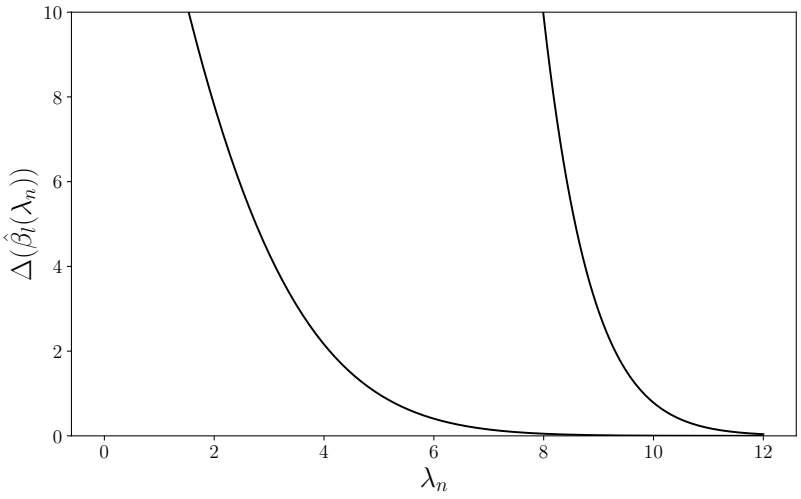
Moreover, as λ_n approaches zero, C_λ diverges to infinity, leading to a simultaneous divergence of both c_α and C_α. This suggests that the Lasso estimator behaves increasingly worse compared to the zero estimator. Conversely, a larger tuning parameter λ_n causes C_λ to converge to zero from the positive side. As a result, both c_α and C_α converge to zero while remaining non-negative. This implies that the performance of the Lasso estimator improves but remains inferior to the zero estimator, until they become equivalent in the limit. Figure 4 visually represents upper and lower bounds for the relative error of Lasso in comparison to the zero estimator across various tuning parameter values.

Intuitively, Lasso's underperformance with weak signals arises from its challenge in differentiating genuine from spurious signals. Its failure to detect true weak signals marginally affects its performance relative to the zero estimator, which disregards such signals entirely. The core issue with Lasso is its inability to effectively eliminate irrelevant signals. While a sufficiently large tuning parameter could address this problem, our theory suggests that only when the penalty is so substantial that Lasso becomes identical to the zero estimator does it enforce an appropriate penalty.

In light of Lasso's underwhelming performance, it is evident that in situations involving weak signals, the elastic net estimator, which combines ℓ_1 and ℓ_2 norms of the parameters

[12] When applied to a vector, $|\cdot|$ and $(\cdot)_+$ represent element-wise operations.

Figure 4: Lasso vs. Zero Estimator: Relative Precise Error Bounds



Note: In this plot, the black curves represent the lower and upper probability bounds on $\Delta(\hat{\beta}_l(\lambda_n))$, i.e., c_α and C_α, as a function of the tuning parameter λ_n in the context of weak signals. In this setup, we fix $n = p = 2,000$. We assign the elements of b_0 to follow a standard Gaussian distribution and set Σ_2 as the identity matrix $\mathbb{I}$. As in Figure 3, we set all parameters $(c_0, \theta_1, \theta_2, \sigma_x, \sigma_\beta, \sigma_\varepsilon)$ to one. Finally, we select $\tau = 0.001$, which results in a population R^2 around 0.1%.

in its penalty function, is unlikely to outperform Ridge.[13]

In contemporary regression analysis, particularly in the context of a large number of covariates, Lasso has gained prominence as a valuable tool, often regarded as the modern counterpart to OLS. Yet, our analysis reveals a critical caveat: in situations where the signals are weak, Lasso, regardless of its tuning parameter choice, becomes an unsuitable option unless the data are extremely sparse, to a degree that violates Assumption 4. The necessity for sparsity increases with the weakness of the signal, positioning Lasso as a bet for extreme sparsity. This finding has significant implications, especially in areas like economics and finance, where large-scale regression analyses are commonplace, and signal-to-noise ratios tend to be low. Conversely, our results strongly advocate for the use of Ridge regression in scenarios characterized by weak signals. This insight underscores the importance of assessing the data's specific characteristics and the strength of the underlying signals when making decisions regarding the most suitable regularization technique.

[13] Although a formal justification for this observation can be provided in the setting of $\Sigma_2 = \mathbb{I}$, we omit it here due to space constraints.

2.7 Assessing Signal-to-Noise Ratio

In line with this perspective, we delve into the assessment of the signal-to-noise ratio, a measure that can offer valuable insights into the viability of different machine learning techniques. Our preceding analyses provide initial but indirect guidance in this regard. Specifically, if Lasso underperforms the zero estimator, it implies a potential issue with the strength of the signal in the data.

A more conventional and direct approach to evaluating the signal-to-noise ratio is through the goodness-of-fit measure known as R^2. However, in-sample R^2 is prone to overfitting, and as such, out-of-sample R^2 is commonly used in machine learning. This metric essentially involves the comparison of mean-squared errors between two predictors. For our specific application, we have chosen to use zero as the reference predictor and define this metric for a given estimator $\hat{\beta}$ as follows:

$$R^2_{\text{oos}}(\hat{\beta}) = 1 - \frac{\sum_{i\in\text{OOS}}(y_i - X_i\hat{\beta})^2}{\sum_{i\in\text{OOS}} y_i^2}, \tag{11}$$

where "OOS" represents the out-of-sample data.

Since a model's predictive performance hinges on the signal-to-noise ratio, it is reasonable to employ this metric to evaluate the signal-to-noise ratio inherent in the DGP. In situations characterized by strong signals, it is expected that the out-of-sample R^2 serves as a consistent estimator for the signal-to-noise ratio as measured by the population R^2. This holds true irrespective of the specific estimator $\hat{\beta}$ employed, as long as it is consistent with respect to β_0, i.e., $\|\hat{\beta} - \beta_0\| = o_{\text{P}}(1)$. However, in situations characterized by weak signals, the outcome depends critically on the choice of estimator beyond the signal-to-noise ratio itself. As an illustration, both Lasso and Ridge are consistent in the sense that their prediction errors asymptotically diminish, yet the out-of-sample R^2 for Lasso can turn non-positive, indicating either no improvement or underperformance compared to the zero estimator, as we have shown in Theorem 4.

In the context of weak signals, the next proposition provides a theoretical justification for the relevance of optimal Ridge's R^2_{oos} in assessing the signal-to-noise ratio in the data.

Proposition 1. *Under the same assumptions as Theorem 3, and assuming that the out-of-sample data follows the same DGP as the in-sample data, if $n_{oos}p^{-2}n^2\tau^2 \to \infty$, where n_{oos} is the size of the out-of-sample data, then for the optimal Ridge estimator, it holds that*

$$R^2_{oos}(\hat{\beta}_r(\lambda_n^{opt})) = p^{-1}n\theta_2(R^2)^2\left(1 + o_{\text{P}}(1)\right),$$

where R^2 denotes the population R-squared, given by $\tau\sigma_x^2\sigma_\beta^2/(\tau\sigma_x^2\sigma_\beta^2+\sigma_\varepsilon^2)$ in this context.

Interestingly, when the size of out-of-sample data is sufficiently large, to the extent that the estimation error in R^2_{oos} does not mask the performance differential between the optimal Ridge and the zero estimator, R^2_{oos} is approximately proportional to the squared population R^2. While it does not exactly mirror the population R^2, R^2_{oos} still serves as an indicator of signal strength in the data. The reason for their discrepancy is that in the weak signal case, the numerator of the R^2_{oos}—which reflects the relative prediction error between the two estimators—decreases more rapidly than the numerator of R^2. Therefore, the numerator of R^2_{oos} only provides a higher-order characterization of signal strength.

2.8 Mixed Signal Strengths and Alternative Benchmarks

In the preceding sections, our analysis primarily focuses on scenarios where all signals are weak, leading us to consider the zero estimator as our natural benchmark. This section, however, expands our analysis to include models where potentially strong signals serve as benchmarks. Consider another DGP:

$$y = W\gamma_0 + X\beta_0 + \varepsilon, \tag{12}$$

where $W \in \mathbb{R}^{n\times d}$ represents a predefined set of covariates. These covariates include potentially strong signals and form the basis of the benchmark model. We allow the dimension d to increase to ∞, however, it does so at a slower rate compared to n, ensuring that OLS of y against W remains a viable method for estimation.

In many cases, W could simply be a vector of ones, allowing us to remain agnostic about the magnitude of the regression's intercept. In our empirical analysis, W can be motivated from economic theory, whose impact on the response variable is of central interest. Alternatively, W can encompass lagged values of y, thereby facilitating the inclusion of temporal dependence in the benchmark model. This setup is particular relevant when using an autoregressive model as a benchmark for forecasting economic variables. Exploring the possibility of a data-driven selection of W is an intriguing direction for future research.

Building on these considerations, our focus now shifts to evaluating and comparing the performance against the OLS benchmark with covariates in W. In this context, the OLS benchmark predictor, $\hat{y}_b^{new}$, for a new observation (w^{new}, x^{new}) is defined as follows:

$$\hat{y}_b^{new} = (w^{new})^\top\hat{\gamma}, \quad \text{where} \quad \hat{\gamma} = (W^\top W)^{-1}W^\top y. \tag{13}$$

The inclusion of W leads us to explore the following Ridge estimator with regularization only imposed on coefficients of X:

$$\begin{aligned}\hat{\beta}(\lambda_n) &:= \arg\min_{\beta} \left\{ \min_{\gamma} \left(\frac{1}{n}\|y - W\gamma - X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2 \right) \right\} \\ &= \arg\min_{\beta} \left\{ \frac{1}{n}\|\mathcal{M}_W y - \mathcal{M}_W X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2 \right\}, \end{aligned} \tag{14}$$

where $\mathcal{M}_W = \mathbb{I} - W(W^\top W)^{-1}W^\top$. Consequently, the estimator for γ is thus given by

$$\hat{\gamma}(\lambda_n) = (W^\top W)^{-1}W^\top(y - X\hat{\beta}(\lambda_n)). \tag{15}$$

The construction for Lasso is similar. Therefore, utilizing the estimated parameters $(\hat{\beta}(\lambda_n), \hat{\gamma}(\lambda_n))$ we are able to formulate a predictor for y as

$$\hat{y}^{new} = (w^{new})^\top \hat{\gamma}(\lambda_n) + (x^{new})^\top \hat{\beta}(\lambda_n) = \hat{y}_b^{new} + (\hat{x}^{new})^\top \hat{\beta}(\lambda_n), \tag{16}$$

where $\hat{x}^{new} = x^{new} - X^\top W(W^\top W)^{-1}w^{new}$.

Notably, equation (16) illuminates the role of the second term, $(\hat{x}^{new})^\top \hat{\beta}(\lambda_n)$, specifically highlighting the contribution of weak signals in contrast to the OLS benchmark, $\hat{y}_b^{new}$. Moreover, a comparison with the zero-benchmark scenario, previously analyzed, reveals a notable distinction in the modification of the regressor and covariates in equation (14). In this instance, our approach involves a regression of $\mathcal{M}_W y$ against $\mathcal{M}_W X$, which intuitively means predicting the residuals of the benchmark model using covariates that have been adjusted to remove the dependence on W. While our earlier conclusions are likely still valid, the inclusion of generated variables in regressions brings an additional layer of statistical error that warrants careful examination. The forthcoming theorem will elucidate that this extra error does not compromise our prior conclusion.

Given this context and our previous comparative analysis, we focus on the optimal Ridge estimator in this scenario. This is because the performance of Ridgeless, OLS, or Lasso is unlikely to show improvement with the incorporation of additional estimation error.

Theorem 5. *Let the design matrix X be generated as $X = W\eta_0 + U$. Assume that the triplet $(U, \beta_0, \varepsilon)$ follows the same distribution as $(X, \beta_0, \varepsilon)$ in Theorem 2. Additionally, the matrix W is independent from U, β_0, and ε. Each covariate within W is assumed to have a finite second moment. Furthermore, we assume that $d = o(n^2 p^{-1}\tau)$, and the eigenvalues of $n^{-1}W^\top W$ are lower bounded by some positive constant. Given these assumptions, the predictor, $\hat{y}^{new}$,*

as defined in (16) *and based on the Ridge estimator from* (14) *with* $\lambda_n = \tau^{-1}\lambda$*, and the benchmark predictor* $\hat{y}_b^{new}$ *from* (13)*, satisfy the following:*

$$pn^{-1}\tau^2\Big(\mathbb{E}_F\left[(\hat{y}^{new} - y^{new})^2 \mid \mathcal{I}\right] - \mathbb{E}_F\left[(\hat{y}_b^{new} - y^{new})^2 \mid \mathcal{I}\right]\Big) \xrightarrow{\text{P}} \alpha^*, \tag{17}$$

where $\mathcal{I}$ *denotes the information set generated by* $(W, X, y, \gamma_0, \beta_0)$*,* α^* *is defined in Theorem 2, and the tuple* $(y^{new}, w^{new}, x^{new})$ *satisfies* (12)*.*

This result indicates that a Ridge-augmented benchmark model demonstrates superior performance compared to the benchmark model alone. In essence, this suggests that Ridge estimator's predictive performance retains its superiority over the zero estimator. Notably, the error arising from the initial estimation of the benchmark model does not influence the comparative performance between the Ridge and zero estimators.

3 Monte Carlo Simulations

In this section, we conduct simulation experiments to assess the finite sample performance of our asymptotic theory. We begin by establishing a linear model setup and evaluate the performance of Ridge and Lasso estimators.

3.1 Ridge and Lasso for Linear Models

Now, we provide details of the DGP given by (1) for the first simulation exercise. We set $(\Sigma_1)_{ij} = 2^{-|i-j|}$ for $1 \leq i, j \leq n$. We construct Σ_ε as a diagonal matrix with i.i.d. entries sampled from the uniform distribution $U(0.5, 1.5)$. The eigenvalues of Σ_2 are also simulated from $U(0.5, 1.5)$, with corresponding eigenvectors from a randomly generated orthogonal matrix, forming Σ_2. These matrices are generated once and then fixed throughout simulations. By direct calculations, we have $\theta_1 = 1$, $\theta_2 = 13/12$, and $\theta_3 = \theta_4 = 5/3$.

We experiment with $n = 500$ and $p = 300$, dimensions that align closely with the first microeconomic example studied below. For each simulated sample, we construct β_0 as $\sqrt{p^{-1}\tau}b_0$, with b_0 drawn from a spike-and-slab distribution: $(1-q)\delta_0 + q\mathcal{N}(0, q^{-1}\sigma_\beta^2)$. Here, δ_0 represents the Dirac delta function, and we set $\sigma_\beta^2 = 1$. The error term ε is sampled from $\mathcal{N}(0, 1)$, while the design matrix X is drawn from $\mathcal{N}(0, 1)$ and subsequently transformed by multiplication with $\Sigma_1^{1/2}$ and post-multiplication with $\Sigma_2^{1/2}$. We consider two cases for the sparsity parameter, $q = 0.2$ and $q = 0.8$. To represent weak and strong signal scenarios, we calibrate two values of τ to achieve $R^2 = 5\%$ and 50%, respectively. The parameters q and

τ (through R^2) are varied as they are critical to the asymptotic performance, as highlighted in Assumption 4. A total of $1,000$ Monte Carlo repetitions are conducted.

For each simulated sample, we compute the relative prediction error, $\Delta(\hat{\beta}(\hat{\lambda}_n^{K-CV}))$, as defined in equation (8), with the tuning parameter for each method selected via cross-validation. The histograms of these errors are presented in Figure 5. Additionally, Table 1 provides summary statistics, including quantiles and the percentage of values classified as "zeros." (where "zero" is defined as within machine precision).

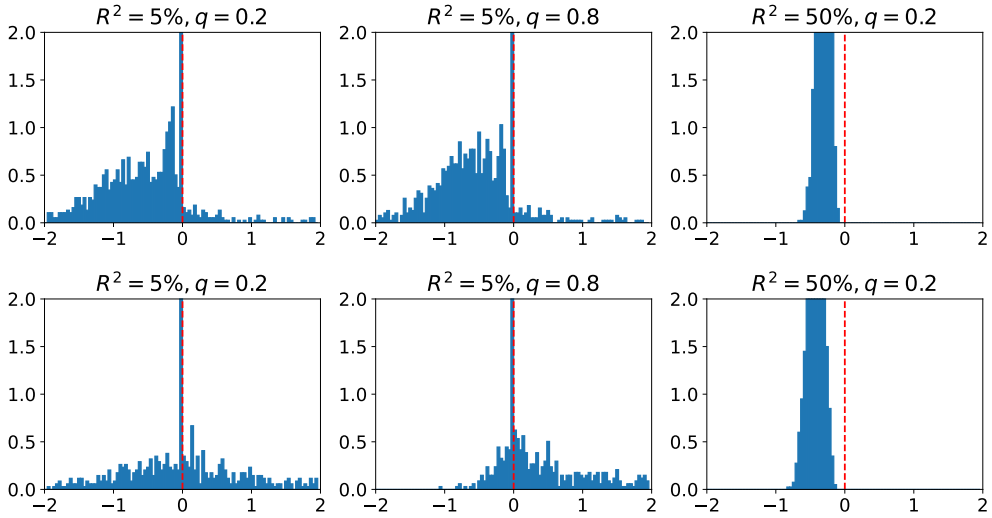
The histograms reveal notable differences in the performance of Ridge and Lasso when R^2 is low, while showing that both methods outperform the zero estimator when R^2 is high. Ridge displays a clear probability mass on the negative side of the error distribution, even in weak signal settings, underscoring its ability to outperform the zero estimator in most cases. Notably, Ridge's performance appears largely unaffected by changes in sparsity levels.

In contrast, Lasso struggles to capture weak signals effectively, as evidenced by a substantial probability mass on the positive side of the y-axis. While increasing the sparsity level (i.e., lowering q) leads to a modest performance improvement for Lasso, it remains inferior to Ridge overall. Table 1 further reveals that, in finite samples with high sparsity levels, some probability mass for Lasso can fall on the negative side, as indicated by the first quantile. Nevertheless, Lasso also shows a heavier probability mass at zero compared to Ridge, consistent with the theoretical prediction that optimal Lasso solution collapses to zero when signals are weak.

The above results use tuning parameters selected via 10-fold cross-validation. To validate our theoretical findings independently of tuning parameter selection, Appendix A.1 presents experiments with fixed tuning parameters. Additionally, Further simulations in Appendix A.2 support our theoretical predictions regarding the R^2_{oos} of the optimal Ridge. Appendix A.3 presents evidence indicating that Type I error primarily influences Lasso's performance relative to the zero estimator. As λ increases, Type I error diminishes, leading to an improvement in Lasso's performance, which ultimately becomes identical to that of the zero estimator, while Type II error persists at a high level.

To examine the robustness of our theory in cases of extreme sparsity, Appendix A.4 examines the effect of further reducing Lasso's sparsity level to $q = 0.1$, 0.05, and 0.02. The results reveal that for each sparsity level, as R^2 decreases, Ridge's performance improves, whereas Lasso's performance deteriorates. This suggests that even under extreme sparsity conditions, the relative performance is dictated by the strength of the signal. Ridge continues to outperform both the zero estimator and Lasso when the signal is sufficiently weak.

Figure 5: Simulation Results for Ridge and Lasso in Linear DGPs



Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_r(\hat{\lambda}_n^{K-CV}))$ (top) and $\Delta(\hat{\beta}_l(\hat{\lambda}_n^{K-CV}))$ (bottom) following equation (8) across 1,000 Monte Carlo samples. We analyze three setups of (R^2, q), where $(R^2, q) = (5\%, 0.2)$, $(5\%, 0.8)$, and $(50\%, 0.2)$.

3.2 Advanced Machine Learning Methods for Nonlinear Models

Appendix A.4 also demonstrates the robustness of our theoretical predictions under other deviations from model assumptions. In this section, we extend our investigation to an important form of deviation: the application of nonlinear machine learning methodologies, including RF, GBRT, and NNs, through simulation experiments. While providing a precise theoretical analysis of errors for these algorithms remains challenging—and this part therefore involves some degree of speculation—we draw on insights from linear models to interpret and contextualize our simulation findings.

We simulate the following DGP, expressed explicitly in element-wise form:

$$y_i = \sum_{j=1}^{p} \beta_{0,j} f(Z_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{18}$$

where y_i denotes the ith observation of the response variable, $\beta_{0,j}$ represents the coefficient associated with a function $f(\cdot)$ of the predictor variable Z_{ij}.

In order to exploit insights from the prior simulation exercise within this new context, we adopt the following procedure for simulating this model: We generate Z_{ij} by applying an inverse transform to X_{ij}, which was previously simulated in Section 3.1. Specifically,

Table 1: Summary Statistics for Ridge and Lasso in Linear DGPs

		Lasso				Ridge			
q	R^2 (%)	Q1	Q2	Q3	#Zero	Q1	Q2	Q3	#Zero
0.2	5%	-0.127	0.000	0.521	360	-0.992	-0.501	-0.129	97
0.8	5%	0.000	0.000	0.975	400	-0.918	-0.541	-0.169	88
0.2	50%	-0.508	-0.414	-0.331	0	-0.375	-0.300	-0.233	0

Note: The table illustrate the summary statistics (quantiles and the percentage of zeros) of relative prediction error $\Delta(\hat{\beta}(\hat{\lambda}_n^{K-CV}))$, for Ridge and Lasso, based on 1,000 Monte Carlo samples. The analysis considers three setups of (R^2, q): $(R^2, q) = (5\%, 0.2)$, $(5\%, 0.8)$, and $(50\%, 0.2)$.

Z_{ij} is defined as $f^{-1}(X_{ij})$, where the design matrix X is constructed using the identical DGP as previously outlined. Additionally, both the coefficients β_0 and the error term ε_i follow the same baseline DGP as previously described. This methodology guarantees the replication of the exact simulation results observed when regressing y on X. Nevertheless, our primary focus now shifts to predicting y based on nonlinear models of Z without prior knowledge of $f(\cdot)$. The effective signal-to-noise ratio diminishes relative to the linear case due to the added complexity of learning an unknown function $f(\cdot)$. Needless to say, the machine learning models we explore in the subsequent experiments are capable of handling more general DGPs than the one given by Eq. (18).

3.2.1 Simulations with Tree Algorithms

Tree algorithms are essential in machine learning for handling complex DGPs with discrete variables, nonlinearities, and intricate interactions. However, single tree models often underperform, prompting the use of ensemble methods to enhance predictions.

Two popular ensemble techniques are RF and GBRT. RF uses bagging, where multiple trees are trained independently on bootstrap samples, and their predictions are averaged to improve performance. In contrast, GBRT employs boosting, iteratively fitting residuals from prior trees to build a strong ensemble from weak learners.

Since trees are invariant to monotonic transformations, it suffices to report their prediction results for the linear DGP, as these are identical to those for the nonlinear DGP under consideration.[14] However, tree methods may underperform Ridge and Lasso, partly due to an additional approximation error from using piecewise constant functions to approximate the linear DGP. Therefore, the primary focus here should not be on comparing tree methods with linear models but rather on assessing the effectiveness of different ensemble techniques

[14]Separate simulation experiments, not included here, confirm this observation.

in capturing weak signals and comparing their performance to the zero predictor.

Our implementation of RF involves three tuning parameters: the depth of each individual tree, ranging from 3 to 15; the number of randomly selected features used in each tree split, varying from 10 to 300; and the ratio of bootstrapped samples, ranging from 0.1 to 0.2. The total number of trees in the RF ensemble is fixed at 5,000, as increasing it to 10,000 yields no significant improvement. For GBRT, we also consider three tuning parameters: the depth of the trees, the number of trees, and the learning rate. The learning rate is adjusted between 0.001 and 0.5, while the depth of each tree varies from 1 to 6, reflecting GBRT's preference for shallower trees compared to RF. The maximum number of trees is set to 100, with most experiments halting training well before reaching this limit.

We repeat the experiments from Section 3.1, this time generating an additional set of n_{oos} out-of-sample observations to evaluate predictive performance.[15] As illustrated in Figure 6, RF demonstrates its ability to learn weak signals when $R^2 = 5\%$, with more than half of the probability mass located to the left of the y-axis. In contrast, GBRT struggles at this signal strength level. Sparsity does not appear to significantly impact either method. Nonetheless, both methods markedly outperform the zero predictor as R^2 increases to 50%.

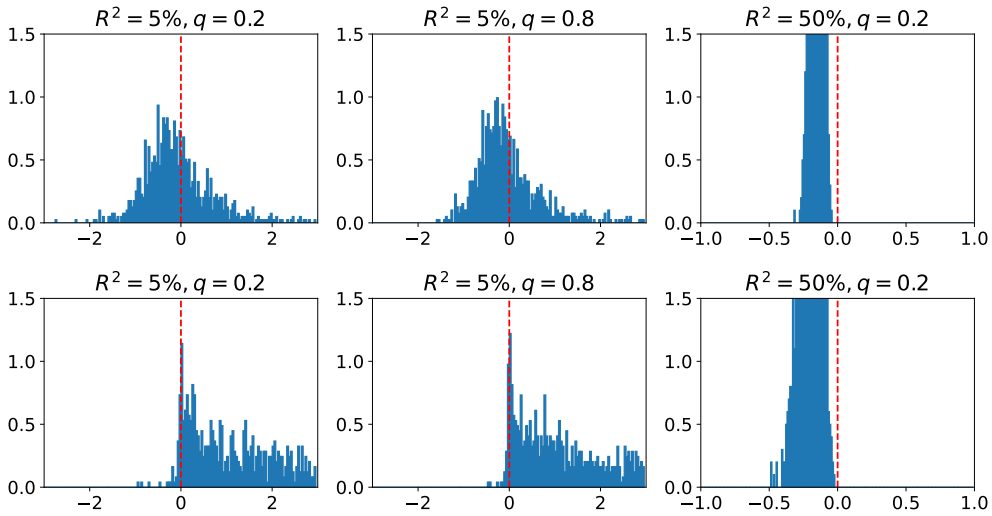
A possible explanation for GBRT's performance pattern could be an inherent ℓ_1-like regularization in its boosting approach. This hypothesis draws support from the work of Efron et al. (2004), which demonstrates a parallel between boosting and the Lasso path in linear regressions. To substantiate this conjecture, we examine the number of active variables (those with a non-zero importance score) from both tree methods under the benchmark case. For RF, the average count of active variables is 300. This mirrors that of Ridge regression. Conversely, GBRT demonstrates a significantly lower count of active variables, with an average of 27.9, aligning more with the variable selection feature of Lasso. Based on our theoretical findings that Ridge outperforms Lasso in weak signal scenarios, we can infer that RF is more adept than GBRT in settings characterized by low signal strengths.

3.2.2 Simulations with Neural Networks

Next, we study fully-connected feed-forward NNs, following the stylized architectures as outlined by Gu et al. (2020). For this exploration, we revisit the benchmark case where $n = 500$, $p = 300$, $R^2 = 5\%$, and $q = 0.2$. Such parameters lead to an input layer in the NN configured with 300 neurons. Our specifically chosen architecture features a neural network with a single hidden layer, which includes 16 neurons.

[15] We set $n_{oos} \approx [\tau^{-3}]$ to meet the assumption in Proposition 1.

Figure 6: Simulation Results for RF and GBRT in Linear DGPs



Note: The histograms depict the relative prediction error, $pn^{-1}\tau^{-2}n_{oos}^{-1}\sum_{i\in\text{OOS}}((y_i-\hat{y}_i)^2-y_i^2)$, across $1,000$ Monte Carlo samples. We consider RF and GBRT under the setting $n = 500$, $q = 300$, $n_{oos} = 10,000$, with $(R^2, q) = (5\%, 0.2), (5\%, 0.8)$, and $(50\%, 0.2)$. The red dashed line marks the y axis for reference.

The training process of these NNs often incorporates a sophisticated mix of optimization and regularization techniques, crucial for enhancing performance.[16] To specifically assess the influence of ℓ_1 and ℓ_2 regularization on NN performance, we will minimize potential interference from other factors. Therefore, our implementation will involve using plain SGD as the sole optimization technique, coupled exclusively with either ℓ_1 or ℓ_2 penalties, and deliberately avoiding additional optimization enhancements.[17] This approach is designed to isolate and clarify the specific contributions of these regularization techniques to NN perfor-

[16]Key methods include stochastic gradient descent (SGD) with Adam (Kingma and Ba (2014)), which expedites the optimization process through an adaptive learning rate. Early stopping, as discussed in Goodfellow et al. (2016), is employed to prevent overfitting by halting training when validation performance starts to decline. Dropout (Srivastava et al. (2014)) is utilized for better generalization, achieved by randomly deactivating neurons. Batch normalization (Ioffe and Szegedy (2015)) aids in stabilizing the training process. Moreover, ensembling over various random seeds is implemented to reduce the variances in model outputs. Furthermore, the integration of ℓ_1 and ℓ_2 penalties with these techniques helps regulate the NN parameters.

[17]Consequently, the training process hinges primarily on two tuning parameters: the learning rate and the regularization parameter. The learning rate is set to fluctuate within the interval $[0.001, 0.015]$, a range established based on insights from the validation sample. To efficiently control computational costs, we have adopted a strategy of jointly tuning the learning rate and the number of epochs, while keeping the product of these two factors constant and fixing the batch size in SGD at 100. This method is designed to achieve a balanced compromise, optimizing both the efficiency of the learning process and the stability of the resulting model. On the other hand, the choice of regularization parameter is contingent upon the DGPs and the specific regularization technique employed.

mance, albeit at the expense of not fully exploiting the NN's potential. Additionally, given the conceptual similarities between early stopping and shrinkage methods—specifically, early stopping effectively shrinks parameter values towards their initial, smaller magnitudes—we will also conduct a comparative analysis of the effects of early stopping and ℓ_2-regularization.

To evaluate the performance of NNs, we analyze their behavior with three specific monotonic nonlinear functions: Tangent ($\tan(x)$), Cubic (x^3), and Sinh ($(e^x + e^{-x})/2$). The histograms in Figure 7 display the relative prediction errors of NNs when they are applied with various regularization techniques, including early stopping, ℓ_2 regularization, and ℓ_1 regularization. It is observed that both ℓ_2 regularization and early stopping are effective in detecting and leveraging weak signals, evidenced by the majority of the probability mass of their histograms being positioned on the negative side of the y-axis. In contrast, under ℓ_1 regularization, there is a notable decline in performance. This observation aligns with expectations based on our theoretical findings concerning linear models.

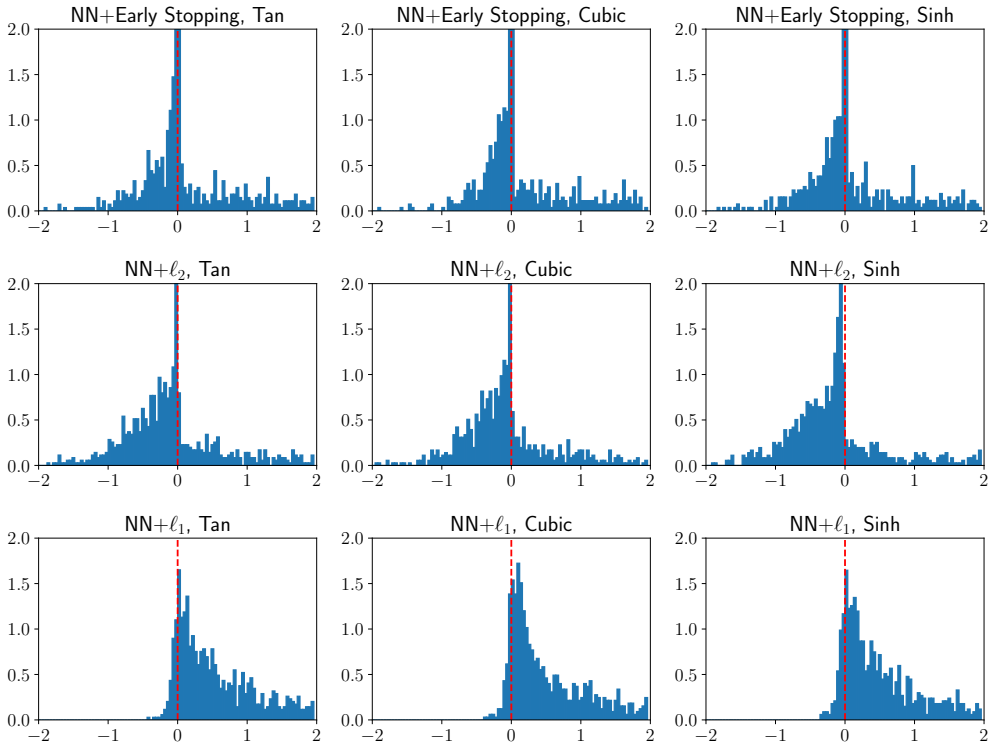
4 Empirical Analysis of Six Economic Datasets

In this section, we demonstrate the practical relevance of our theoretical insights by applying seven machine learning methods—Ridge, Lasso, OLS/Ridgeless, RF, GBRT, NNs with both ℓ_1 and ℓ_2 penalties—across six datasets. These datasets are sourced from three fields: microeconomics, macroeconomics, and finance, with two datasets representing each field. Five of these datasets are similar to those utilized by Giannone et al. (2022), with updates to the latest available data wherever feasible. Additionally, we incorporate an updated dataset from Gu et al. (2020) for our second finance example, which offers a more comprehensive coverage of firm characteristics than the analogous example discussed by Giannone et al. (2022). There is a notable difference between our empirical strategy and that of Giannone et al. (2022), which focuses on estimating a parametric model using a Spike-and-Slab prior within a Bayesian framework. In contrast, our study, aligning more closely with Gu et al. (2020), places a greater emphasis on the comparative analysis of various methods.

At the outset of each empirical exercise, we face a variety of decisions regarding our implementation strategy. These include defining the in-sample and out-of-sample periods, opting for either a rolling window or an expanding window approach, selecting a cross-validation procedure, and deciding on the normalization of covariates.[18] We delegate these

[18]It is crucial to normalize covariates before employing machine learning methods. This step standardizes the scales of covariates, facilitating regularization and enhancing the convergence of optimization algorithms. To avoid forward-looking bias, we ensure that the normalization of covariates is conducted using their

Figure 7: Simulation Results for NNs in Nonlinear DGPs



Note: The histograms depict the relative prediction error, $pn^{-1}\tau^{-2}n_{oos}^{-1}\sum_{i\in\text{OOS}}((y_i-\hat{y}_i)^2-y_i^2)$, across 1,000 Monte Carlo samples. These histograms pertain to NNs in scenarios where $n=500$, $p=300$, $q=0.2$, $n_{oos}=10,000$, and $R^2=5\%$. The focus is on three different regularization techniques: early stopping, ℓ_1-penalty, and ℓ_2-penalty, and the experiments encompass three nonlinear models: Tan, Cubic, and Sinh.

choices to the frameworks established by Giannone et al. (2022) and Gu et al. (2020), with the intention of minimizing degrees of freedom to enhance the robustness, comparability, and reproducibility of our findings. In the application of each machine learning method, the selection of an appropriate grid for tuning parameters is essential. This crucial step requires balancing performance optimization with computational efficiency. Finer and wider grids, while potentially enhancing performance, also increase computational demands. Appendix B provides details regarding our model configuration and tuning parameter selection.

Below we present the empirical findings derived from six distinct datasets, each analyzed and reported separately. The primary summary statistics, R^2_{oos}s, are collected in Table 2. Additionally, we include variable importance plots in Figure 8 as supplementary evidence

respective in-sample mean and standard deviation, thereby maintaining the validity and integrity of our predictive analysis.

to decode the performance of different methods. The notion of variable importance is not universally established and varying across different contexts. Our approach diverges from the well-known method associated with RF, originally presented in Breiman (2001). In our analysis, variable importance is quantified as the reduction in R^2_{oos} resulting from setting each variable, one at a time, to zero (its mean value post-normalization), with this metric normalized across all variables. For each method, the most significant variable, as per this definition, is assigned a value of one, and a color gradient is employed to visually represent the relative importance of each variable.

Table 2: Out-of-sample R-squared Values in Empirical Studies

	Ridge	Lasso	OLS/Ridgeless	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
Finance 1	0.80	−12.19	−81.08	4.38	−14.21	1.41	−10.31
Finance 2	0.19	0.10	−1.25	0.10	−0.30	0.26	0.14
Macro 1a	15.29	15.40	−1375	24.37	16.44	16.94	19.09
Macro 1b	3.49	3.69	−2939	8.45	1.11	7.09	5.39
Macro 2	6.58	−14.58	−837	9.65	1.28	4.00	1.92
	(4.83)	(43.74)	(854)	(9.19)	(14.04)	(18.42)	(13.36)
Micro 1	0.48	−1.01	−13198	−9.53	−5.07	0.49	−6.77
	(0.84)	(2.01)	(12479)	(3.82)	(6.60)	(0.27)	(17.87)
Micro 2a	26.27	20.37	−12729	27.80	16.44	23.87	23.37
	(7.50)	(6.41)	(9213)	(5.27)	(3.40)	(10.07)	(10.09)
Micro 2b	1.89	-3.43	-14724	0.81	-6.45	1.11	-1.73
	(3.09)	(5.25)	(10506)	(2.43)	(6.83)	(2.20)	(5.09)

Note: This table reports R^2_{oos} values, presented in percentages, for Ridge, Lasso, OLS/Ridgeless, RF, GBRT, and NNs with respective ℓ_1 and ℓ_2 penalties, across six empirical studies spanning Finance, Macroeconomics, Microeconomics. For the first example in Macroeconomics and the second example in Microeconomics, two benchmark models are considered for comparison. Where standard deviations are applicable, they are provided in parentheses.

4.1 Finance 1: Market Equity Premium

In the first analysis, we focus on predicting market equity returns using a dataset of financial and macroeconomic indicators compiled by Welch and Goyal (2007).[19] This dataset comprises 16 predictors and includes 74 annual observations, covering a period from 1948 to 2021. Despite Welch and Goyal (2007) reporting a consistently negative R^2_{oos} for this dataset, several other studies, such as those by Campbell and Thompson (2007), Ferreira

[19]The data was sourced from Amit Goyal's website, accessible at https://sites.google.com/view/agoyal145, and processed using the methodology provided by Giannone et al. (2022).

Figure 8: Variable Importance Plots

Note: This figure illustrates the variable importance across six empirical studies, using color gradients to show the relative reductions in R^2_{oos} by each covariate. For the first example in Macroeconomics and the second example in Microeconomics, we only present the cases with a more complex benchmark model.

and Santa-Clara (2011), Rapach et al. (2010), Kelly and Pruitt (2013), and Kelly et al. (2023), have developed forecasting strategies resulting in economically meaningful R^2_{oos} values. These modest R^2_{oos}s translate into significant economic gains through simple market timing strategies, as extensively discussed in these aforementioned studies.

We revisit this exercise. The procedure involves selecting an expanding window of in-sample data, using cross-validation for optimal tuning parameter selection, and then refitting the model for predictions on a test sample. Following the empirical framework of Giannone et al. (2022), the initial training set spans from 1948 to 1964, with the model's performance evaluated using the 1965 data. Subsequently, data from 1965 is added to the training set, and the process is repeated, with the next model tested on the 1966 data. This procedure is conducted 57 times in total, progressively incorporating an additional year's data into the training set and shifting the test sample forward by one year each time.

We evaluate R^2_{oos} from 57 different predictions.[20] Despite these predictions stemming from 57 distinct models, the empirical results, as detailed in Table 2, corroborate our theoretical predictions. Specifically, Ridge regression records an R^2_{oos} of 0.80%, significantly outperforming Lasso's -12.19%. With the smallest sample size being 17—just sufficient to run OLS with 16 predictors—OLS produces a highly negative R^2_{oos} of -81.08%. In comparison, the NN with an ℓ_2 penalty, NN(ℓ_2), attains an R^2_{oos} of 1.41%, whereas its counterpart with an ℓ_1 penalty, NN(ℓ_1), shows a lower R^2_{oos} of -10.31%. Our RF achieves the highest R^2_{oos} of 4.38% while GBRT show performance similar to Lasso, with an R^2_{oos} of -14.21%. As indicated in Figure 8, Ridge and NN(ℓ_2) appear to assign similar weights to these covariates. The leading covariate, eqis—the equity issuing activity ratio —is closely followed in importance by the dividend-price ratio, d/p.

4.2 Finance 2: Cross-Section of Expected Returns

In our second analysis, we build upon the predictors utilized by Gu et al. (2020) for predicting equity returns, extending the data up to December 2021. We analyze monthly total individual equity returns from the CRSP database for all firms listed on the NYSE, AMEX, and NASDAQ. The average number of stocks analyzed per month exceeds 6,200. Our dataset starts from March 1957, and we compile a total of 920 covariates to predict future returns.[21]

[20] In this example, $R^2_{\text{oos}} = 1 - \sum_{t=1965}^{2021}(y_t - \hat{y}_t)^2 / \sum_{t=1965}^{2021}(y_t - \bar{y}_t)^2$, where $\bar{y}_t = \sum_{s=1948}^{t-1} y_t/(t-1948)$.

[21] Due to the large scale of the dataset and the resulting computational limitations, we have adopted a two-fold cross-validation approach for this analysis. Additionally, a more restrictive grid selection is employed for the Lasso and Ridge models. Specifically, for Ridge, $\log(\lambda)$ is set to range between 6 and 7, while for lasso, it varies from -3.4 to -2.4. The optimal tuning parameters fall within the central range of these specified grids.

Following Gu et al. (2020), our initial training phase utilizes data from 1957 to 1986, followed by performance evaluation using 1987 data. This process is repeated 35 times, with each iteration expanding the training sample by an additional year and shifting the evaluation period forward by one year.

Table 2 compares the model performance in terms of R^2_{oos}, where the zero estimator serves as the benchmark, according to the recommendation by Gu et al. (2020).[22] In this case, NN(ℓ_2) emerges as the leading model, closely followed by Ridge, achieving R^2_{oos} of 0.26% and 0.19%, respectively. These models dominate NN(ℓ_1) and Lasso, which records an R^2_{oos} of 0.14% and 0.10%. The performance of the OLS continues to be underwhelming in this exercise. As for the tree-based models, RF demonstrates superior performance compared to GBRT. The former achieves a slightly positive R^2_{oos}, whereas the latter exhibits a slight negative R^2_{oos}. Despite their inherent ability to capture non-linear relationships and interactions, the effectiveness of tree models is somewhat limited in scenarios characterized by low signal strength. Figure 8 reveals an intriguing pattern: there appears to be a relationship between the relatively stronger performance among these pairs — Ridge vs Lasso, RF vs GBRT, and NN(ℓ_2) vs NN(ℓ_1) — and their respective patterns of sparsity in variable importance plots. Models with denser variable weights outperform those with sparser ones

To illustrate the economic significance of these relatively low R^2_{oos}s, we adopt the approach outlined by Gu et al. (2020) and devise a stock selection portfolio strategy. This strategy involves going long on the top 10% and shorting the bottom 10% of stocks, sorted based on their predicted returns for the upcoming month, with equal weighting applied to each stock every month. In terms of performance, NN(ℓ_2) achieves the highest Sharpe ratio at 2.13, indicating its superior risk-adjusted returns. This is closely followed by Ridge regression with a Sharpe ratio of 1.64. NN(ℓ_1) also demonstrates commendable performance, yielding a Sharpe ratio of 1.55. GBRT exhibits the least impressive performance, with the lowest Sharpe ratio of 0.80, which aligns with its underwhelming predictive performance.

4.3 Macro 1: Macroeconomic Forecasting

The prediction of US macroeconomic activity using a wide range of predictors has been a topic of significant interest since its initial exploration by Stock and Watson (2002). In our current study, we utilize the FRED-MD dataset, compiled by McCracken and Ng (2016), to forecast the monthly growth rate of US industrial production (IP). This dataset includes 119 potential predictors, covering a diverse array of macroeconomic indicators, and extends from

[22] In this pooled regression setting, we define $R^2_{\text{oos}} = 1 - \sum_{i,t\in\text{OOS}}(y_{i,t} - \hat{y}_{i,t})^2 / \sum_{i,t\in\text{OOS}} y^2_{i,t}$.

February 1960 to December 2019. Our evaluation methodology aligns with the prediction procedure outlined by Giannone et al. (2022). We begin by training these machine learning models using data from February 1960 to December 1974 and then evaluate their performance on data from the subsequent year. This process is repeated 45 times, with each iteration expanding the training dataset by one year (12 observations) and similarly shifting the evaluation period forward. We adhere to the guidelines set by McCracken and Ng (2016) for transforming the covariates. Additionally, we follow their prescribed approach for managing data quality issues, which involves the removal of outliers and the filling of missing data.

We initiate our analysis with a benchmark model that includes only an intercept term. In this scenario, all machine learning models significantly outperform this benchmark, achieving R^2_{oos} values ranging from 14.13% to 24.37%.[23] On the other hand, the OLS model, somewhat expectedly, overfits the data, resulting in a negative R^2_{oos} of -16. This outcome suggests the presence of strong signal strength within a high-dimensional set of covariates. The benchmark model's lack of competitiveness aligns with our expectations, particularly when considering the temporal dependence prevalent in macroeconomic time series. Therefore, a more suitable benchmark model should incorporate lagged values of IP growth.

We thereby propose an alternative benchmark that incorporates an Autoregressive (AR) component. Within each training sample, we fit an AR model to the IP growth, selecting its order based on the AIC. The residuals from this model then serve as our prediction target. As discussed in Section 2.8, this approach effectively combines the predictions from the AR model with those from our machine learning models, yielding a hybrid output out-of-sample.[24] Consequently, the new benchmark for comparison becomes the direct use of the AR model's predictions, where adding zero implies no alteration to the prediction. In this alternative setup, the comparison of R^2_{oos} values reveals a pattern somewhat associated with scenarios of weak signal strength: NN(ℓ_2) and RF emerge as the top performers, achieving R^2_{oos} values of 7.09% and 8.45%, respectively. Following closely are NN(ℓ_1) at 5.39%, while GBRT lags with a considerably lower R^2_{oos} of 1.11%. This disparity in performance appears associated with the findings in Figure 8, which illustrates GBRT's tendency towards sparser models in comparison to their counterparts. In this case, linear models, specifically

[23] In this case, the definition of R^2_{oos} is similar to how it is defined in the Finance 1 case.

[24] In implementing advanced machine learning models with a hybrid benchmark linear component $W\gamma$, we adopt a methodology that parallels the one used in Eq. (14). This approach entails a DGP assumption that $\mathcal{M}_W y$ is a general function of $\mathcal{M}_W X$. This assumption plays a critical role in streamlining the implementation of these machine learning methods, ensuring that the results are directly comparable to those obtained in linear settings. However, it is important to note that this DGP assumption is generally not equivalent to the assumption that $y - W\gamma$ is a function of X.

Ridge and Lasso, demonstrate comparable performance, achieving R^2_{oos} values of 3.49% and 3.69%, respectively. This pattern suggests that the primary challenges associated with signal weakness are most evident from nonlinear features.

4.4 Macro 2: Economic Growth Across Countries

Next, we explore a dataset originally compiled by Barro and Lee (1994), which includes 60 socio-economic, institutional, and geographical covariates across 90 countries. This dataset is utilized for predicting long-term economic growth, specifically measured by the growth rate of GDP per capita from 1960 to 1985. A pivotal aspect of this analysis involves testing a key prediction of the classical Solow-Swan-Ramsey growth model, which concerns the effect of an initial (lagged) GDP per capita level on subsequent growth rates. By incorporating the logarithm of each country's GDP per capita in 1960 alongside a constant term, our prediction model includes a total of 62 potential covariates.

Belloni et al. (2013b) implement the Square-root-Lasso technique in their regression, anticipating sparsity among the control variables. This methodology results in a remarkable sparse model, characterized by the inclusion of a singular control variable: the log of the black market premium, a measure of trade openness. In contrast, Giannone et al. (2022) employ a Bayesian approach with a spike-and-slab prior, concluding that a dense model, which includes all covariates, yields the best log-predictive score.

In our predictive analysis, we adopt the same empirical methodology outlined by Giannone et al. (2022). We begin by randomly selecting half of the data samples for model estimation and then proceed to assess the performance of these models using the remaining samples. This process is repeated 100 times. The average out-of-sample R^2_{oos} from these 100 repetitions, in comparison to a benchmark model that includes only the intercept, is presented in Table 2, accompanied by their standard deviations, provided in parentheses.[25][26]

Our empirical findings align with those of Giannone et al. (2022), indicating a similar pattern that dense models, specifically Ridge, RF, and NN(ℓ_2), exhibit superior performance compared to their sparse counterparts, such as Lasso, GBRT, and NN(ℓ_1). The limited sample size appears to disadvantage complex NN models, rendering them less effective than the simpler Ridge regression. RF demonstrates strong performance, achieving an R^2_{oos} of

[25] Here $R^2_{\text{oos}} = 1 - \sum_{i\in\text{OOS}} (y_i - \hat{y}_i)^2 / \sum_{i\in\text{OOS}} (y_i - \bar{y})^2$, where $\bar{y}$ is in-sample average of y_i.

[26] We may also consider a benchmark model with GDP per capita in 1960 included, as predicted by theory. Interestingly, enforcing the inclusion of this variable in the model leads to a reduction in predictive performance across all models. In essence, adding this variable leads to a negative R^2_{oos} compared to the model that includes only an intercept.

9.65%, although it concurrently introduces a twofold increase in the variability of R^2_{oos} values compared to those based on Ridge regression. The variance of Lasso is pronounced, driven by a handful of extreme values; excluding these anomalies, its R^2_{oos} improves to -0.56%. Across all evaluated models, the black market premium consistently emerges as the most influential variable in Figure 8, aligning with the sole variable selected by Belloni et al. (2013b).

4.5 Micro 1: Crime Rates across US States

Our first microeconomic case revisits the study by Donohue and Levitt (2001), which analyzes the effect of the legalization of abortion following the Roe vs. Wade decision in 1973 on the decline in crime rates. Their dependent variable is the change in log per-capita murder rates from 1986 to 1997 across 48 states, with a total of 576 observations. This variable is then regressed on the effective abortion rate. To account for potential confounding factors, Belloni et al. (2013a) expanded the control set used by Donohue and Levitt (2001) by including interactions and higher-order terms, resulting in a comprehensive set of 284 variables.

When Belloni et al. (2013a) employ the Lasso method for the selection of control variables for murder rate in their analysis, they discover that none of the control variables were selected.[27] In a similar vein, Giannone et al. (2022) observe from their Bayesian analysis of this regression that the posterior density is concentrated on very low probability values of the slab component, which suggests that the regression model is sparse with high likelihood. In a recent study, Guo and Toulis (2023) employ a randomization test to assess the null hypothesis that all regression coefficients are zero. Their test fails to reject this hypothesis.

We employ the same benchmark model and sample splitting strategy outlined by Giannone et al. (2022). For the initial estimation, we use data spanning from 1986 to 1989, covering all states. Additionally, we incorporate data from a randomly selected 50% of the states for the period from 1990 to 1997. The remaining 50% of the states from 1990 to 1997 are set aside for evaluating the model. This procedure is iterated 8 times, with each iteration expanding the training sample to include one additional year of data, starting from 1990, while correspondingly adjusting the evaluation sample to reflect this change. [28] The entire sequence is carried out 13 times in total, yielding $8 \times 13 = 104$ distinct training and evaluation samples. We report the mean and standard deviation of R^2_{oos}s in Table 2.

[27] Belloni et al. (2013a) proposes a double-Lasso estimator to make inference on the effect of abortion on murder rate. Part of their procedure involves a Lasso regression of murder rate on control variables. It is important to note that they use differences as the dependent variable, but observe no substantial changes when using levels instead.

[28] Here and after, we calculate R^2_{oos} in the same way as Finance 2 case.

Our findings reveal that NN(ℓ_2) exhibits a slight edge over Ridge regression, attaining a marginally superior R^2_{oos} of 0.49%, compared to Ridge's 0.48%. Apart from these two models, all other models tested demonstrate negative R^2_{oos} values, which suggests the presence of very weak signals in the data. Our argument posits that the scarcity of significant signals observed in the existing literature is likely attributable to signal weakness. The empirical evidence does not definitively categorize the underlying DGP as either dense or sparse. It may simply be that no individual signals are particularly strong. While in such scenarios, a sparse model could seem like a reasonable approximation, our results reveal that the cumulative predictive power of weak signals, though individually insubstantial, is collectively non-negligible. This is corroborated by Figure 8, where we observe that both Ridge and NN(ℓ_2) assign small weights to nearly all covariates. Although RF follows a similar pattern, our simulations indicate that it is more adversely affected by the weak signals compared to the other models.

4.6 Micro 2: Eminent Domain and Economic Outcomes

In our final study, we concentrate on a regression setting pertinent to eminent domain. Previous research by Chen and Yeh (2012), and subsequently Belloni et al. (2012), employ instrumental variable regressions to understand the impact of eminent domain on economic outcomes. Differing from their broader focus, our study aligns closely with Giannone et al. (2022), who concentrate on the first stage of this regression. This involves predicting pro-plaintiff decisions in takings law cases based on the characteristics of judicial panels. Their dataset includes 138 potential covariates and a total of 312 observations.

Adopting their strategy, we estimate the model using data spanning from 1979 to 1984 for all circuits. This is augmented with data from 1985 to 2004, selected randomly for 50% of the circuits. We assess the model's performance with 1985 data from the circuits not included in the training set. Since the period from 1985 to 2004 encompasses 20 years, we repeat this procedure 20 times. Each repetition involves a new random selection of half of the circuits and the sequential addition of one year's data to the training set, while correspondingly updating the evaluation set. This entire process is independently executed five times, resulting in a total of $20 \times 5 = 100$ distinct training and evaluation datasets.

In our analysis, we initially consider a benchmark model that includes only an intercept. In this setting, all machine learning models successfully identify predictive signals, as evidenced by significant R^2_{oos}s. RF emerges as the top performer with an R^2_{oos} of 27.63%, with other models also showing strong results, albeit GBRT being an exception. However, the scenario shifts markedly when the benchmark model is expanded to include not only

the intercept but also additional variables. These include a dummy variable for the absence of cases in a given circuit-year and the number of takings appellate decisions, bringing the total to three covariates. Against this simple benchmark, the incremental predictive power contributed by the remaining covariates diminishes dramatically. Ridge's R^2_{oos} falls to 1.89%, that of RF to 0.81%, and NN(ℓ_2) to 1.11%, with the R^2_{oos} of all other methods turning negative. Intriguingly, as highlighted in Figure 8, these results seem to associate with the distinct approaches these methods take in weighting covariates. Ridge, RF, and NN(ℓ_2) assign small weights uniformly across all covariates. Meanwhile, NN(ℓ_1) also opts for a model with a considerable number of coefficients, resulting in a performance that slightly surpasses both Lasso and GBRT, which favor more sparse models in this case.

5 Conclusion

In this paper, we scrutinize the performance of machine learning techniques in contexts characterized by low signal-to-noise ratios, a situation frequently observed in economics and finance. Our theoretical analysis indicates that while Lasso is often considered a modern alternative to traditional ordinary least squares, its application in these areas should be approached cautiously, primarily due to its lessened effectiveness with weak signals.

Our research complements and expands upon the arguments made by Giannone et al. (2022), who cast doubt on the prevalence of sparsity in economic datasets. We take this debate further by showing that it is signal weakness, not necessarily the absence of sparsity, that more significantly contributes to the observed limitations of Lasso in economic applications. Furthermore, the lack of significant variables in empirical studies may be attributed more to signal weakness than to the sparse nature of the underlying DGP.

Our analysis also reveals a marked difference in the performance of Ridge regression. Notably, Ridge demonstrates superior resilience and effectiveness in these environments. Our theoretical findings are further substantiated by simulation studies encompassing a range of advanced machine learning techniques, including trees and neural networks. These experiments consistently reveal that algorithms designed to exploit sparsity tend to underperform in environments where signals are inherently weak. Broadly, our findings emphasize the importance of a nuanced, context-sensitive application of machine learning techniques, adapting to the distinctive data characteristics encountered across various domains.

References

Barro, R. J. and J.-W. Lee (1994). Sources of economic growth. *Carnegie-Rochester Conference Series on Public Policy 40*, 1–46.

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

Bayati, M. and A. Montanari (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory 58*(4), 1997–2017.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2013a). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies 81*(2), 608–650.

Belloni, A., V. Chernozhukov, and C. B. Hansen (2013b). *Inference for High-Dimensional Sparse Econometric Models*, Volume 3 of *Econometric Society Monographs*, pp. 245–295. Cambridge University Press.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Breiman, L. (2001). Random forests. In *Machine Learning*, pp. 5–32.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics 37*(4), 1685 – 1704.

Campbell, J. Y. and S. B. Thompson (2007). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies 21*(4), 1509–1531.

Chen, D. L. and S. Yeh (2012). Growth under the shadow of expropriation? the economic impacts of eminent domain. Mimeo, Toulouse School of Economics.

Cui, H., W. Guo, and W. Zhong (2018). Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *The Annals of Statistics 46*(3), 958 – 988.

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

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

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics 32*(3), 962 – 994.

Donohue, John J., I. and S. D. Levitt (2001). The Impact of Legalized Abortion on Crime. *The Quarterly Journal of Economics 116*(2), 379–420.

Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407 – 499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Ferreira, M. A. and P. Santa-Clara (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics 100*(3), 514–537.

George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

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

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Gordon, Y. (1988). On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}_n$. In *Geometric Aspects of Functional Analysis*, Berlin, Heidelberg, pp. 84–106. Springer Berlin Heidelberg.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

Guo, W. and P. Toulis (2023). Invariance-based inference in high-dimensional regression with finite-sample guarantees.

Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics 38*(3), 1686 – 1732.

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

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Ingster, Y. I., A. B. Tsybakov, and N. Verzelen (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics 4*(none), 1476 – 1526.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 448–456. JMLR.org.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics 37*(4), 1647 – 1684.

Jin, J. and Z. T. Ke (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statistica Sinica 26*(1), 1–34.

Kelly, B. and S. Pruitt (2013). Market expectations in the cross-section of present values. *The Journal of Finance 68*(5), 1721–1756.

Kelly, B. T., S. Malamud, and K. Zhou (2023). The virtue of complexity in return prediction. *The Journal of Finance*. Forthcoming.

Kingma, D. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolesár, M., U. K. Müller, and S. T. Roelsgaard (2024). The fragility of sparsity.

Li, Y., I. Kim, and Y. Wei (2020). Randomized tests for high-dimensional regression: A more efficient and powerful solution. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc.

Liang, T. and P. Sur (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. *The Annals of Statistics 50*(3), 1669 – 1695.

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Miolane, L. and A. Montanari (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics 49*(4), 2313 – 2335.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies 23*(2), 821–862.

Robbins, H. (1964). The Empirical Bayes Approach to Statistical Decision Problems. *The Annals of Mathematical Statistics 35*(1), 1 – 20.

Rovcková, V. and E. I. George (2018). The spike-and-slab lasso. *Journal of the American Statistical Association 113*(521), 431–444.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*(56), 1929–1958.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Su, W., M. Bogdan, and E. Candès (2017). False discoveries occur early on the lasso path. *The Annals of Statistics 45*(5), 2133–2150.

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 P. Grünwald, E. Hazan, and S. Kale (Eds.), *Proceedings of The 28th Conference on Learning Theory*, Volume 40 of *Proceedings of Machine Learning Research*, Paris, France, pp. 1683–1709. PMLR.

Tsigler, A. and P. L. Bartlett (2023). Benign overfitting in ridge regression. *Journal of Machine Learning Research 24*(123), 1–76.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, S., H. Weng, and A. Maleki (2020). Which bridge estimator is the best for variable selection? *The Annals of Statistics 48*(5), 2791 – 2823.

Welch, I. and A. Goyal (2007). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies 21*(4), 1455–1508.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics 36*(4), 1567 – 1594.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(2), 301–320.

Online Appendix of Can Machines Learn Weak Signals?

Zhouyu Shen* Dacheng Xiu†

Booth School of Business University of Chicago

Booth School of Business University of Chicago

December 11, 2024

Abstract

Appendix A presents additional results from Monte Carlo simulations. Appendix B provides an in-depth discussion on the selection of tuning parameters. Appendix C explores the theoretical properties of Lasso and Ridge in the context of extreme sparsity. Appendix D contains the mathematical proofs of the main theorems presented in the paper. Appendix E is devoted to the exposition of technical lemmas along with their corresponding proofs.

A Supplemental Simulation Results

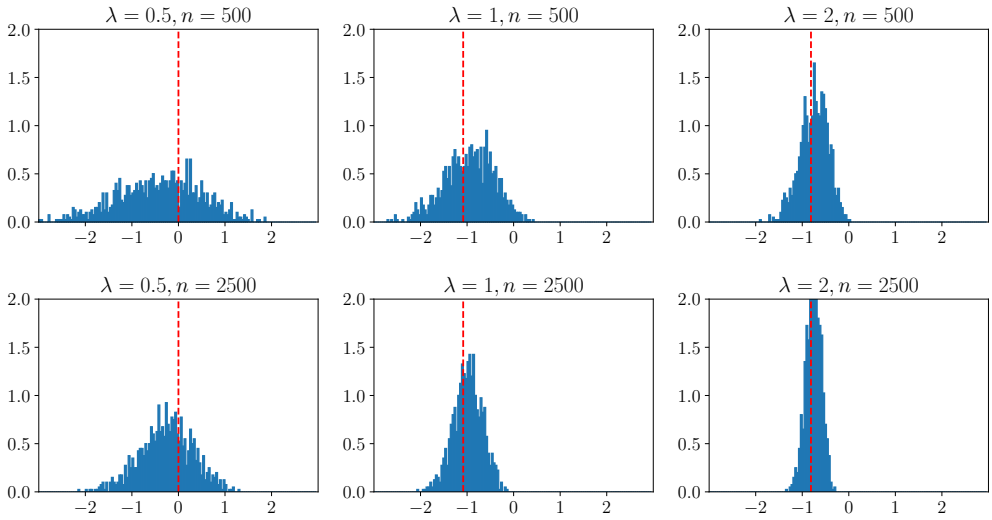
A.1 Additional Simulations with Fixed Tunings

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.

*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

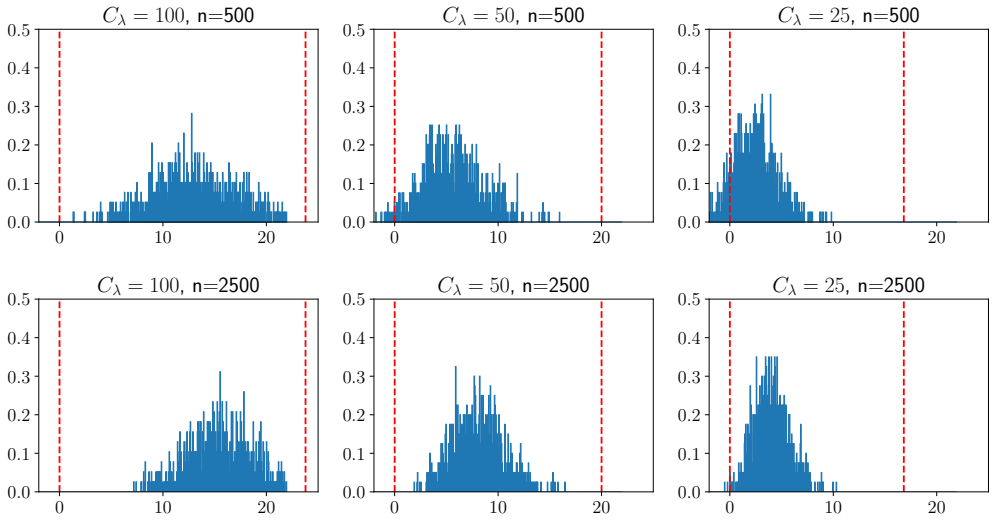


Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_r(\lambda_n))$ following equation (8) across $1,000$ Monte Carlo samples. We consider two different sample sizes ($n = 500$ and $n = 2,500$) and examine three different values of λ, where $\lambda = 0.5, 1$, and 2. Notably,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

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 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

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).

parameters and sample sizes, accompanied by two red vertical lines in each plot representing the lower and upper bounds, c_α and C_α.

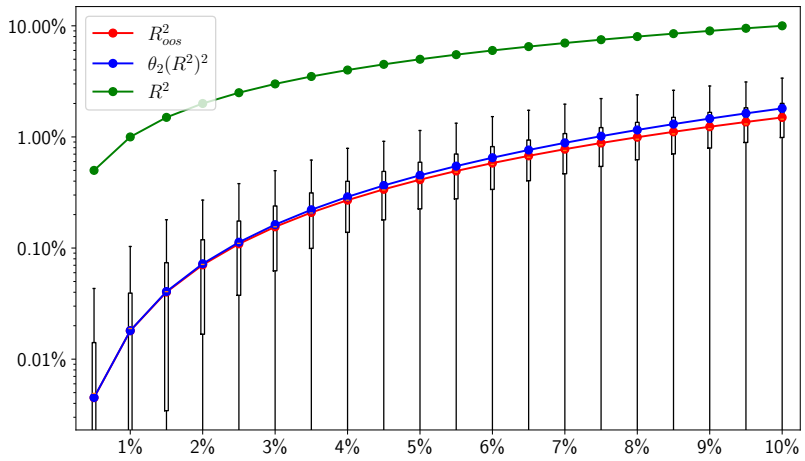
These plots yield several interesting findings. First, as the sample size increases, we observe that the probability mass becomes more concentrated and largely falls within the intervals defined by the bounds. Second, regardless of the tuning parameter values, Lasso consistently underperforms the zero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



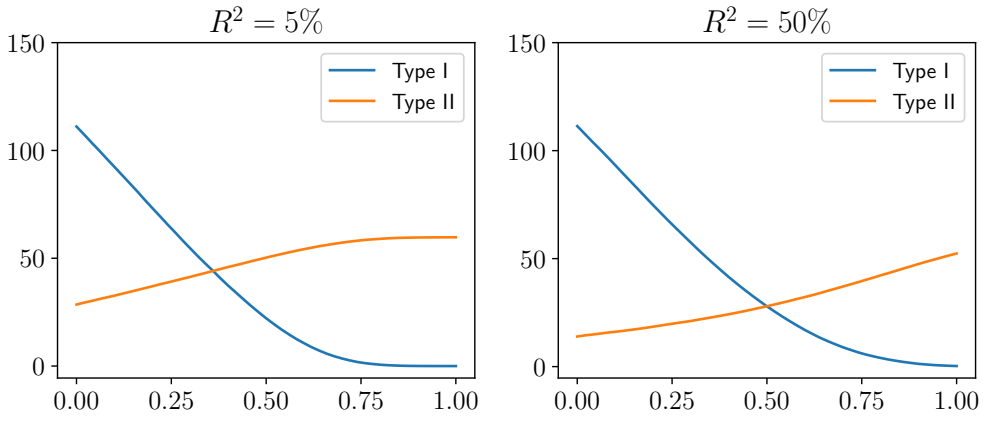
Note: The figure presents boxplots showing the distributions of R^2_{oos} for optimal Ridge regression ($\lambda = 1$) over 1,000 Monte Carlo repetitions, with $n = 500$, $p = 300$, $q = 0.2$, and $n_{oos} = 10,000$. We explore a range of population R^2 values, from 0.5% to 10% in increments of 0.5% 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 Appendix C. In Case II, we observe the increased ratio of p/n does not affect our conclusion. Case III demonstrate 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
	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
Case I	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

Table B2 provides details regarding model configuration and tuning parameters. For Ridge and Lasso methods, which each involve only one tuning parameter, we employ the glmnet package. This package effectively determines the optimal tuning parameter through a default ten-fold cross-validation process. The process is conducted on an adaptively selected grid, ensuring efficient and effective selection of the optimal tuning parameter. Regarding our implementation of RF, GBRT, and NNs, we follow the protocol outlined in the simulation section. 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.

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=0.25∼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)} $\log(\lambda) \in [-2, 1]$	architecture∼{16,4,1} batch size=16 (lr,epochs)={(0.4,1), (0.08,5), (0.02,20)} $\log(\lambda) \in [-2, 1]$
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)} $\log(\lambda) \in [-4, 0]$	architecture∼{920,32,1} batch size=10000 (lr,epochs)={(0.5,2), (0.2,5), (0.067,15), (0.05,20),(0.04,25)} $\log(\lambda) \in [-5, -3]$
Macro 1	depth=5∼50 #trees=500 #features=2∼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)} $\log(\lambda) \in [-2, 2]$	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.05,2), (0.02,5),(0.01,10), (0.005,20), (0.002,50)} $\log(\lambda) \in [-10.5, 1.5]$
Macro 1b	depth=5∼40 #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)} $\log(\lambda) \in [-1, -0.5]$	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.004,25), (0.002,50),(0.001,100), (0.0005,200)} $\log(\lambda) \in [2, 3]$
Macro 2	depth=1∼5 #trees=500 #features=1∼5 %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)} $\log(\lambda) \in [-3, -3]$	architecture∼{61,8,1} batch size=16 (lr,epochs)={(0.05,20), (0.02,50), (0.002,500)} $\log(\lambda) \in [-7, 4]$
Micro 1	depth=1∼60 #trees=500 #features=1∼20 %samples=0.25∼1	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)} $\log(\lambda) \in [-11, -7]$	architecture∼{297,16,1} batch size=16 (lr,epochs)={(0.4,1), (0.2,2), (0.08,5), (0.04,10), (0.02,20)} $\log(\lambda) \in [-8, 0]$
Micro 2	depth=1∼20 #trees=500 #features=2∼30 %samples=0.2∼0.5	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)} $\log(\lambda) \in [-10, -6]$	architecture∼{217,16,1} batch size=16 (lr,epochs)={(0.1,1),(0.01,10), (0.001,100), (0.0001,1000)} $\log(\lambda) \in [-12, -9]$
Micro 2b	depth=1∼5 #trees=500 #features=1∼5 %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)} $\log(\lambda) \in [0, 2]$	architecture∼{215,16,1} batch size=16 (lr,epochs)={(0.01,50), (0.005,100), (0.0025,200)} $\log(\lambda) \in [0, 2]$

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 ($\log(\lambda)$).

C Lasso and Ridge in the Extremely Sparse Setting

In this section, we show that the restriction imposed on our asymptotic settings is primarily for technical reasons. Even in the extremely sparse scenario outside the scope of our main analysis, Lasso fails to outperform zero as long as the signal remains sufficiently weak, whereas Ridge is capable of learning weak signals with a non-negligible probability. The proof relies on a distinct approach, utilizing the closed-form formula of the Lasso estimator in a simplified setting. However, this proof does not generalize to the broader case considered earlier.

Consider the Gaussian sequence model where $y = \beta_0 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \mathbb{I})$. For this model, $X = \mathbb{I}$ and $n = p$. We let $\beta_0 = (\sqrt{n\tau}, 0, \ldots, 0)^\top$, so that $s = \|\beta_0\|_0 = 1$ and $R^2 = \frac{\|X\beta_0\|^2}{\|X\beta_0+\varepsilon\|^2} \asymp \tau \to 0$. We prove that as long as $\tau = o(s\log(p)/n)$ (up to some log factors), Lasso cannot learn weak signals.

Proposition C1. *Assume $\tau \le n^{-1}\log(n)/100$. There exists $n_0 > 0$, when $n > n_0$,*

$$\mathrm{P}\left(\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \ge 0, \forall \lambda \ge 0\right) \ge 1 - n^{-1/2}.$$

Proof. Let $\tau_0 = \sqrt{n\tau} \le \sqrt{\log n}/10$ and $\lambda_0 = \sqrt{n}\lambda/2$. Note that when $X = \mathbb{I}$, Lasso has a closed form solution:

$$\hat{\beta}_l(\lambda) = ((|\varepsilon_1 + \tau_0| - \lambda_0)_+ \mathrm{sgn}(\varepsilon_1 + \tau_0), (|\varepsilon_2| - \lambda_0)_+ \mathrm{sgn}(\varepsilon_2), \ldots, (|\varepsilon_n| - \lambda_0)_+ \mathrm{sgn}(\varepsilon_n))^\top.$$

Therefore,

$$\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 = ((|\varepsilon_1 + \tau_0| - \lambda_0)_+ - \tau_0)^2 - \tau_0^2 + \sum_{i=2}^{n}((|\varepsilon_i| - \lambda_0)_+)^2.$$

Assume for now that $\max\limits_{1\le i\le n} |\varepsilon_i| \ge |\varepsilon_1| + 2\tau_0$. Let $i_0 = \arg\max |\varepsilon_i|$, then we have

$$\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \ge ((|\varepsilon_1 + \tau_0| - \lambda_0)_+ - \tau_0)^2 - \tau_0^2 + ((|\varepsilon_{i_0}| - \lambda_0)_+)^2 := Q$$

Consider the following three cases for ε_1:

(i) $\varepsilon_1 \in [-\lambda_0 - \tau_0, \lambda_0 - \tau_0]$: $Q = 0 + ((|\varepsilon_{i_0}| - \lambda_0)_+)^2 \ge 0$.

(ii) $\varepsilon_1 \in (-\infty, -\lambda_0 - \tau_0)$: $Q \ge -\tau_0^2 + ((|\varepsilon_1| + 2\tau_0 - \lambda_0)_+)^2 \ge -\tau_0^2 + 9\tau_0^2 \ge 0$.

(iii) $\varepsilon_1 \in (\lambda_0 - \tau_0, \infty)$: $Q \geq -\tau_0^2 + ((|\varepsilon_1| + 2\tau_0 - \lambda_0)_+)^2 \geq -\tau_0^2 + \tau_0^2 = 0$.

Therefore, under the event that $\max_{1\leq i\leq n} |\varepsilon_i| \geq |\varepsilon_1| + 2\tau_0$, it holds that

$$\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \geq 0.$$

Now we evaluate the probability of this event. Note that

$$\mathbb{P}(\max_{1\leq i\leq n} |\varepsilon_i| \leq u) = (\mathbb{P}(|\varepsilon_i| \leq u))^n = \left(\operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)\right)^n \leq \exp\left\{-\frac{n}{2}\exp\left\{-\frac{2}{\pi}u^2\right\}\right\},$$

where $\operatorname{erf}(\cdot)$ represents the Gauss error function. The last inequality uses the fact that $(\operatorname{erf}(x))^2 \leq 1 - \exp(-4x^2/\pi)$ and $1 + x \leq e^x$.

Reparametrizing u in terms of δ by solving $\delta = \exp\left\{-\frac{n}{2}\exp\left\{-\frac{2}{\pi}u^2\right\}\right\}$, we obtain that, with probability at least $1 - \delta$:

$$\max_{1\leq i\leq n} |\varepsilon_i| \geq \sqrt{\frac{\pi}{2}\log\frac{n}{2} - \frac{\pi}{2}\log\log\frac{1}{\delta}}. \tag{C1}$$

Choosing $\delta = n^{-1}$ in (C1), then the following event happens with probability at least $1 - \frac{1}{n}$:

$$\mathcal{C} = \left\{\max_{1\leq i\leq n} |\varepsilon_i| \geq \sqrt{\frac{\pi}{2}\log\frac{n}{2} - \frac{\pi}{2}\log\log n}\right\}.$$

Setting $u = \sqrt{\frac{\pi}{2}\log\frac{n}{2} - \frac{\pi}{2}\log\log n} - 2\tau_0$. There exists $n_0 \in \mathbb{N}$, when $n \geq n_0$, $u \geq \sqrt{1.7\log n} - 0.2\sqrt{\log n} \geq \sqrt{\log n}$. By Mills' inequalities, when $n \geq n_0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq i\leq n} |\varepsilon_i| - 2\tau_0 \geq |\varepsilon_1| \middle| \mathcal{C}\right) \geq \mathbb{P}\left(|\varepsilon_1| \leq \sqrt{\frac{\pi}{2}\log\frac{n}{2} - \frac{\pi}{2}\log\log n} - 2\tau_0\right) \\
=&\mathbb{P}\left(|\varepsilon_1| \leq u\right) \geq 1 - \sqrt{\frac{2}{\pi}}\frac{\exp\left(-u^2/2\right)}{u} \geq 1 - \sqrt{\frac{2}{\pi\log n}}n^{-1/2}
\end{aligned}$$

There exists $n_1 \geq 1$, when $n \geq n_1$, $(1 - \sqrt{\frac{2}{\pi\log n}}n^{-1/2})(1 - \frac{1}{n}) \geq 1 - n^{-1/2}$. Hence when $n \geq \max(n_0, n_1)$:

$$\mathbb{P}\left(\max_{1\leq i\leq n} |\varepsilon_i| - 2\tau_0 \geq |\varepsilon_1|\right) \geq \mathbb{P}(\mathcal{C}) \cdot \mathbb{P}\left(\max_{1\leq i\leq n} |\varepsilon_i| - 2\tau_0 \geq |\varepsilon_1| \middle| \mathcal{C}\right) \geq 1 - n^{-1/2}. \qquad \square$$

In contrast, under the same scenario, we demonstrate that Ridge is capable of learning weak signals with a non-negligible probability.

Proposition C2. *Under the same conditions as Proposition C1, there exists $n_0 > 0$, when $n > n_0$,*

$$\mathrm{P}\left(\exists \lambda > 1 \ s.t. \ \|\hat{\beta}_r(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 < 0\right) \geq 0.5.$$

Proof. Since $\hat{\beta}_r(\lambda) = (1 + n\lambda)^{-1} y$, it holds that

$$\|\hat{\beta}_r(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 = \frac{-2n\lambda\varepsilon_1\tau_0 - (1 + 2n\lambda)\tau_0^2 + \sum_{i=1}^n \varepsilon_i^2}{(1 + n\lambda)^2}.$$

Observe that with probability equal to 0.5, $\varepsilon_1\tau_0 \geq 0$. Under this event,

$$\|\hat{\beta}_r(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \leq \frac{-(1 + 2n\lambda)\tau_0^2 + \sum_{i=1}^n \varepsilon_i^2}{(1 + n\lambda)^2}.$$

Therefore, as long as $\lambda > (2n)^{-1}(-1 + \tau_0^{-2}\sum_{i=1}^n \varepsilon_i^2)$, we have $\|\hat{\beta}_r(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 < 0$. □

D Mathematical Proofs

D.1 Proof of Theorem 1

Proof. Throughout the proof, we employ the shorthand notation "w.a.p.1" to denote "with probability approaching one." For two random variables X and Y, we write $X \perp Y$ when they are independent and $X \overset{d}{=} Y$ when they have the same distribution.

For convenience, we omit the subscript F from the expectation operator $\mathbb{E}_F(\cdot)$. Our objective is to demonstrate that

$$\frac{\mathbb{E}\|\Sigma_2^{1/2}(\mathbb{E}(\beta_0|X, y) - \beta_0)\|^2}{\mathbb{E}\|\Sigma_2^{1/2}\beta_0\|^2} \to 1, \text{ as } n \to \infty.$$

This can be shown by proving $\mathbb{E}\|\Sigma_2^{1/2}\mathbb{E}(\beta_0|X, y)\|^2 = o(\mathbb{E}\|\Sigma_2^{1/2}\beta_0\|^2)$. Given that the eigenvalues of Σ_2 are bounded away from zero and positive infinity, it suffices to establish that $\mathbb{E}\|\mathbb{E}(\beta_0|X, y)\|^2 = o(\tau)$. Therefore, we need to prove for all $1 \leq i \leq p$, $\mathbb{E}(\mathbb{E}(\beta_{0,i}|X, y))^2 = o(p^{-1}\tau)$, or, equivalently, $\mathbb{E}(\mathbb{E}(b_{0,i}|X, y))^2 = o(1)$.

By the inequality $\mathbb{E}(\mathbb{E}(A|\mathcal{F})^2) \leq \mathbb{E}(\mathbb{E}(A|\mathcal{G})^2)$ for $\mathcal{F} \subset \mathcal{G}$, and that β_0 is i.i.d., we have

$$\mathbb{E}(\mathbb{E}(b_{0,i}|X, y))^2 \leq \mathbb{E}(\mathbb{E}(b_{0,i}|X, y, \beta_{0,-i}))^2 = \mathbb{E}(\mathbb{E}(b_{0,i}|X_{\cdot,i}, \beta_{0,i}\Sigma_\varepsilon^{-1/2}X_{\cdot,i} + z))^2,$$

where z is defined in Assumption 2, $X_{\cdot,i}$ represents the i-th column of X, and $\beta_{0,-i}$ de-

notes the subvector of β without the ith entry. Denote the information set generated by $\{X_{\cdot,i}, \beta_{0,i}\Sigma_\varepsilon^{-1/2}X_{\cdot,i} + z\}$ as $\mathcal{G}_i$. By Assumption 3, $b_{0,i}$ can be written as $q^{-1/2}b_{1i}b_{2i}$ where $b_{1i} \sim B(1,q)$ and b_{2i} is a sub-exponential random variable with mean zero and variance σ_β^2, whose distribution function is denoted by F_{b_2}. For any $M_1 < 0$, find M_2 (a function of M_1) such that $\mathbb{E}b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]} = q^{1/2}\mathbb{E}b_{2i}\mathbf{1}_{b_{2i}\in[M_1,M_2]} = 0$. This is always feasible because $\mathbb{E}b_{2i} = 0$. By Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}(\mathbb{E}(b_{0,i}|\mathcal{G}_i))^2 \leq & 3\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i)\big)^2 + 3\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}>M_2}|\mathcal{G}_i)\big)^2 \\ & + 3\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}<M_1}|\mathcal{G}_i)\big)^2 =: 3S_{1n} + 3S_{2n} + 3S_{3n}. \end{aligned} \tag{D1}$$

Now we prove for any given M_1, $\lim_{n\to\infty} S_{1n} = 0$. 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

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

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

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

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

$$A_n := \left\{\left|p^{-1/2}\tau^{1/2}\sum_{k=1}^n\tilde{y}_k\tilde{x}_k\right| \leq \tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p) \text{ and } p^{-1}\tau\sum_{k=1}^n\tilde{x}_k^2 \leq \tilde{C}p^{-1}n\tau\right\} \tag{D2}$$

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

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

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 F_{b_2} is a sub-exponential distribution, the integration of both terms on the right-hand-side converges to zero as $n \to \infty$. Therefore, we conclude that for any $\epsilon > 0$, there exists n_0 such that when $n > n_0$, $|Q_{1n}| \leq \epsilon$ under the event A_n. Similarly, it can be proven there exists n_1 such that when $n > n_1$, $Q_{2n} \geq 1/2$ under the event A_n. Hence we have

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

where we use Lemma 13 in the last inequality. Since ϵ is arbitrary, we have $\lim_{n\to\infty} S_{1n} = 0$.

Observe that $S_{2n} = \mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}>M_2}|\mathcal{G}_i)\big)^2 \leq \mathbb{E}b_{0,i}^2\mathbf{1}_{q^{1/2}b_{0,i}>M_2}$. As a result, (D1) implies

$$\lim_{n\to\infty} \mathbb{E}(\mathbb{E}(b_{0,i}|\mathcal{G}_i))^2 \leq 3\mathbb{E}b_{0,i}^2\mathbf{1}_{q^{1/2}b_{0,i}>M_2} + 3\mathbb{E}b_{0,i}^2\mathbf{1}_{q^{1/2}b_{0,i}<M_1} = 3\mathbb{E}b_{2i}^2\mathbf{1}_{b_{2i}>M_2} + 3\mathbb{E}b_{2i}^2\mathbf{1}_{b_{2i}<M_1}.$$

Since b_{2i} has finite variance, the right-hand-side of the above inequality can be arbitrarily small by letting $M_1 \to -\infty$, which completes the proof. □

D.2 Proof of Theorem 2

Proof. For ease of notation, we let $\hat{\beta} := \hat{\beta}_r(\lambda_n)$ and $c_n := p/n$. Additionally, define $\delta_1^* := 2\sqrt{\sigma_\varepsilon^2\theta_1}$, $\delta_2^* := (2\lambda\sigma_x^2\sigma_\beta^2\theta_4 - 4\sigma_\varepsilon^2\sigma_x^2\theta_3)/\delta_1^*\lambda$, $\mu(\sigma_x\sigma_\beta, \delta_1^*, \delta_2) := (\delta_1^*\delta_2 - 2\sigma_x^2\sigma_\beta^2\theta_4)/4\sigma_\varepsilon^2\theta_3$, and

$$C_n^\phi := c_n\tau^{-1}\sigma_x^2\sigma_\beta^2 - c_n\frac{\tau^{-1}\sigma_x^2(\delta_1^*)^2}{4\lambda} + \frac{c_n\sigma_\varepsilon^2\sigma_x^4\theta_3}{\lambda^2} - \frac{c_n\sigma_x^4\sigma_\beta^2\theta_4}{\lambda}.$$

We first show that it is sufficient to establish that

$$c_n\tau^{-3/2}(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\| - \|\Sigma_2^{1/2}\beta_0\|) \xrightarrow{\mathrm{P}} \alpha_2^* := \theta_2\sigma_x^3\left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2\sigma_\beta} - \frac{\sigma_\beta}{\lambda}\right). \tag{D3}$$

This is because Eq. (D3) implies that

$$\|\Sigma_2^{1/2}(\hat{\beta}-\beta_0)\|^2 = \|\Sigma_2^{1/2}\beta_0\|^2 + 2c_n^{-1}\tau^{3/2}\alpha_2^*\|\Sigma_2^{1/2}\beta_0\| + o_{\mathrm{P}}(c_n^{-1}\tau^2).$$

On the other hand, using Lemma 2 and $q^{-1/2}p^{-1/2} = o(1)$ by Assumption 4, we deduce that

$$\|\Sigma_2^{1/2}\beta_0\| = \tau^{1/2}\sigma_x\sigma_\beta + O_{\mathrm{P}}(q^{-1/2}p^{-1/2}\tau^{1/2}). \tag{D4}$$

The above two equations together yield the desired result of the theorem.

To prove Eq. (D3), by incorporating (D4) and $c_n q^{-1/2}p^{-1/2}\tau^{-1} = o(1)$ by Assumption 4, it reduces to showing that

$$c_n\tau^{-1}(\tau^{-1/2}\|\Sigma_2^{1/2}(\hat{\beta}-\beta_0)\| - \sigma_x\sigma_\beta) \xrightarrow{\mathrm{P}} \alpha_2^*. \tag{D5}$$

Set $w = \tau^{-3/2}\Sigma_2^{1/2}(\beta-\beta_0)$ and $\hat{w} = \tau^{-3/2}\Sigma_2^{1/2}(\hat{\beta}-\beta_0)$. After rewriting Ridge's optimization problem (2), $\hat{w}$ equals

$$\arg\min_w \frac{c_n}{n}\left\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\right\|^2 + c_n^2\lambda\left\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\right\|^2 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \tag{D6}$$

where subtracting $c_n\tau^{-2}\|\varepsilon\|^2/n$ and C_n^ϕ from the objective function does not alter the solution. Using the definition of $\hat{w}$, proving (D5) is equivalent to proving $c_n\|\hat{w}\| - c_n\tau^{-1}\sigma_x\sigma_\beta \xrightarrow{\mathrm{P}} \alpha_2^*$. Equivalently, we need to prove for all $\epsilon > 0$, w.p.a.1,

$$\alpha_2^* - \epsilon \le c_n\|\hat{w}\| - c_n\tau^{-1}\sigma_x\sigma_\beta \le \alpha_2^* + \epsilon. \tag{D7}$$

Next, we note from Lemma 14 that it suffices to prove the above convergence holds true for $\hat{w}^B$, where $\hat{w}^B$ is a solution to

$$\arg\min_{w\in S_w^n} \frac{c_n}{n}\left\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\right\|^2 + c_n^2\lambda\left\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\right\|^2 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \tag{D8}$$

and $S_w^n = \{w\big|c_n\tau^{-1}\sigma_x\sigma_\beta - K_\alpha \le c_n\|w\| \le c_n\tau^{-1}\sigma_x\sigma_\beta + K_\alpha\}$ for some sufficiently large K_α. We'll denote the optimal solution as $\hat{w}$ instead of using $\hat{w}^B$ for simplicity.

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 $\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\|^2$ 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 (D8) as

$$\min_{w\in S_w^n}\max_{\tilde{u},\tilde{v}} \frac{c_n\tau^{1/2}}{\sqrt{n}}\tilde{u}^\top Zw - \frac{c_n\tau^{-1}}{\sqrt{n}}\tilde{u}^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}\tilde{u}\|^2}{4} + c_n^2\lambda\tilde{v}^\top w + c_n^2\lambda\tau^{-3/2}\tilde{v}^\top\Sigma_2^{1/2}\beta_0$$

$$- \frac{c_n^2 \lambda \|\Sigma_2^{1/2} \tilde{v}\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^{\phi}. \tag{D9}$$

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. (D9), 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 $\|w\| \leq \tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}K_\alpha$, $\|\Sigma_1\| \leq C_1$, $\|\Sigma_\varepsilon\| \leq C_\varepsilon$. Furthermore, w.p.a. 1, $\|z\| \leq \sqrt{2n}$ by the law of large numbers, which implies $\|\varepsilon\| \leq \sqrt{2C_\varepsilon n}$. 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:

$$\|\hat{u}\| \leq \frac{2\tau^{1/2}}{\sqrt{n}} \|\Sigma_1 Z w\| + \frac{2}{\sqrt{n}} \|\tau^{-1}\Sigma_1^{1/2}\varepsilon\| \leq 4\tau^{-1}\sqrt{C_1 C_\varepsilon}.$$

Let $S_u^n = \{u \big| \|u\| \leq 4\tau^{-1}\sqrt{C_1 C_\varepsilon}\}$. Based on the above result, w.a.p.1, the following optimization problem is equivalent to (D9):

$$\begin{aligned} \min_{w \in S_w^n} \max_{\substack{u \in S_u^n \\ v}} & \frac{c_n \tau^{1/2}}{\sqrt{n}} u^\top Z w - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4} + c_n^2 \lambda v^\top w + c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 \\ & - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^{\phi}. \end{aligned} \tag{D10}$$

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

$$\begin{aligned} \phi(g,h) = & \max_{\substack{0 \leq \delta \leq 4\tau^{-1}\sqrt{C_1 C_\varepsilon} \\ v}} \min_{w \in S_w^n} \max_{\|u\| = \delta} \mathcal{R}_n(w,v,u), \quad \text{where} \\ \mathcal{R}_n(w,v,u) = & \frac{c_n \tau^{1/2}}{\sqrt{n}} \|w\| g^\top u - \frac{c_n \tau^{1/2}}{\sqrt{n}} \delta h^\top w - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4} \\ & + c_n^2 \lambda v^\top w + c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^{\phi}, \end{aligned} \tag{D11}$$

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 (D11), with w restricted to $S_w^n \cap \tilde{\mathcal{S}}_n^c$.

Lemma 15 characterizes the limiting behavior of the optimal solution to (D10), $\hat{w}$, and in turn, proves the desired (D7), under conditions pertaining to the optimization problem (D11). Therefore, we only need show that conditions outlined in Lemma 15 hold. That is, 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 $\mathcal{R}_n(w, v, u)$ pertaining to 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 extensively studied in, e.g., Gander et al. (1989) and Tao and An (1998). The optimal value associated with this problem is given by the following expression:

$$-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) \quad \text{(D12)}$$

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

$$\frac{1}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-2}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{\delta^2}{4} = 0, \quad \text{(D13)}$$

under the condition that $\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I}$ is positive semidefinite. Using this, Eq. (D11) can be rewritten as the following:

$$\begin{aligned}\max_{0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}}\min_{v\ w\in S_w^n} &-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}\\ &\times(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n\tau^{1/2}}{\sqrt{n}}\delta h^\top w + c_n^2\lambda v^\top w\\ &+ c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.\end{aligned}$$

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

$$\begin{aligned}\max_{0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}}\min_{v\ |c_n\alpha - c_n\tau^{-1}\sigma_x\sigma_\beta|\le K_\alpha} &-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}\\ &\times(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - c_n\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|\alpha \quad \text{(D14)}\\ &+ c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.\end{aligned}$$

By Lemma 16, 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. (D14) becomes

$$\min_{c_n|\alpha-\tau^{-1}\sigma_x\sigma_\beta|\le K_\alpha} \max_{\substack{\gamma>0 \\ 0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}}} \max_{v} -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top$$
$$\times (\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n\gamma}{2} - \frac{c_n\alpha^2}{2\gamma}\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|^2$$
$$+ c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.$$

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

$$\min_{c_n|\alpha-\tau^{-1}\sigma_x\sigma_\beta|\le K_\alpha} \max_{\substack{\gamma>0 \\ 0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}}} -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}$$
$$\times (\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n\gamma}{2} + \frac{c_n^2\lambda^2}{4}\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0 + \frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right)^\top\left(\frac{\lambda}{4}\Sigma_2 + \frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}$$
$$\times\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0 + \frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right) - \frac{c_n\tau\alpha^2\delta^2}{2\gamma n}\|h\|^2 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi. \tag{D15}$$

Denote the objective function as $Q_n(\alpha, \delta, \gamma)$. The optimization problem now reduces to a scalar problem. Let us define $\gamma = \tau^{-1}\gamma_1$, $\delta = \tau^{-1}\delta_1^* + \delta_2^* + c_n^{-1/2}\delta_3$ and $\alpha = \tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}\alpha_2$. Subsequently, we present the modified objective function $\tilde{Q}_n(\alpha_2, \delta_3, \gamma_1)$ as follows:

$$\tilde{Q}_n(\alpha_2, \delta_3, \gamma_1) := Q_n(\tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}\alpha_2, \tau^{-1}\delta_1^* + \delta_2^* + c_n^{-1/2}\delta_3, \tau^{-1}\gamma_1). \tag{D16}$$

It is evident that $\tilde{Q}_n$ remains convex with respect to α_2 and jointly concave with respect to (δ_3, γ_1). Moreover, Lemma 17 establishes the probability limit of $\tilde{Q}_n$:

$$\tilde{Q}_n \xrightarrow{\text{P}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1) := -\frac{\delta_3^2\theta_1}{4\theta_3} + 2\sigma_x\sigma_\beta\alpha_2 - \frac{\gamma_1^2}{4\sigma_x^2\sigma_\beta^2\lambda}\theta_2 - \frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta} + \frac{(\delta_1^*)^2\gamma_1}{8\lambda^2\sigma_x^2\sigma_\beta^2}\theta_2, \tag{D17}$$

as well as the following inequalities: for any $\eta > 0$,

$$\text{(i) } \phi(g,h) < \min_{\alpha_2\in[-K_\alpha,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1)+\eta,$$

$$\text{(ii) } \phi_{\tilde{\mathcal{S}}_n^c}(g,h) > \min_{\alpha_2\in[-K_\alpha,\alpha_2^*-\epsilon]\cup[\alpha_2^*+\epsilon,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1)-\eta, \tag{D18}$$

$$\text{(iii) } \min_{\alpha_2\in[-K_\alpha,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1) < \min_{\alpha_2\in[-K_\alpha,\alpha_2^*-\epsilon]\cup[\alpha_2^*+\epsilon,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1).$$

This establishes the conditions of Lemma 15, and thereby completes the proof. □

D.3 Proof of Theorem 3

Proof. For convenience, we define the shorthand notation $\hat{\beta}^i_{\lambda_n} := \hat{\beta}^i_r(\lambda_n)$. Also, we define

$$\hat{R}^{K-CV}(\lambda_n) := \frac{1}{n}\sum_{i=1}^{K} \|y_{(i)} - X_{(i)}\hat{\beta}^i_r(\lambda_n)\|^2. \tag{D19}$$

By Lemma 6, w.p.a.1, we have

$$\frac{1}{n}\|X_{(-i)}^\top X_{(-i)}\| \le \frac{1}{n}C_2\|Z_{(-i)}^\top Z_{(-i)}\| \le C_2(1+\sqrt{c_n})^2, \quad i=1,\dots,K. \tag{D20}$$

Additionally, by Lemmas 2 and 3, w.p.a.1, we have

$$\frac{1}{n}\|\varepsilon\|^2 \le 2\sigma_\varepsilon^2 \quad \text{and} \quad \frac{1}{n}\|y\|^2 \le 2\sigma_\varepsilon^2. \tag{D21}$$

Under the condition that $\lambda_n \ge \epsilon$, using (D20) and (D21), we have, w.p.a.1,

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

Using a similar argument, we have, w.p.a.1,

$$\begin{aligned}\|\hat{\beta}^i_{\lambda_1} - \hat{\beta}^i_{\lambda_2}\|^2 &= c_n^2(\lambda_1-\lambda_2)^2 \left\| \frac{1}{n}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_2\right)^{-1}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_1\right)^{-1} X_{(-i)}y_{(-i)} \right\|^2 \\ &\le \frac{(\lambda_1-\lambda_2)^2}{n^2c_n^2\lambda_1^2\lambda_2^2}\|X_{(-i)}y_{(-i)}\|^2 \le \frac{2C_2\sigma_\varepsilon^2(1+\sqrt{c_n})^2(\lambda_1-\lambda_2)^2}{c_n^2\epsilon^4}.\end{aligned} \tag{D22}$$

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

$$|\hat{R}^{K-CV}(\lambda_1) - \hat{R}^{K-CV}(\lambda_2)| = \frac{1}{n}\left|\sum_{i=1}^{K}\left(\|y_{(i)} - X_{(i)}\hat{\beta}^i_{\lambda_1}\|^2 - \|y_{(i)} - X_{(i)}\hat{\beta}^i_{\lambda_2}\|^2\right)\right|$$
$$\leq \frac{2}{n}\sum_{i=1}^{K}\|X_{(i)}\|\|\hat{\beta}^i_{\lambda_1} - \hat{\beta}^i_{\lambda_2}\|\left(\|y_{(i)}\| + \frac{2C_2^{1/2}\sigma_\varepsilon(1+\sqrt{c_n})}{c_n\epsilon}\|X_{(i)}\|\right) \leq \tilde{C}|\lambda_1 - \lambda_2|, \tag{D23}$$

where $\tilde{C}$ is some fixed constant. Based on this inequality, Lemma 18 proves

$$\inf_{\lambda\in[\epsilon,\tilde{c}\tau^{-1}]} pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\lambda) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} > 0, \tag{D24}$$

w.p.a.1 for some constant $\tilde{c} > 0$. Additionally, as $n \to \infty$, for any fixed $\lambda > 0$,

$$pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\tau^{-1}\lambda) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} \xrightarrow{\text{P}} \frac{2(K-1)}{K}\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2}{2\lambda^2} - \frac{\sigma_\beta^2}{\lambda}\right). \tag{D25}$$

Using (D25) and the definition of λ^{opt}, we obtain, w.p.a.1,

$$pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\tau^{-1}\lambda^{opt}) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\}$$
$$< 0 < \inf_{\lambda\in[\epsilon,\tilde{c}\tau^{-1}]} pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\lambda) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\}.$$

This suggests that the minimizer of $\hat{R}^{K-CV}(\lambda)$ must satisfy $\hat{\lambda}^{K-CV} \geq \tilde{c}\tau^{-1}$, w.a.p.1, so that, $\hat{\lambda}_n^{K-CV} = \arg\min_{\lambda_n\in[\tilde{c}\tau^{-1},\infty)} \hat{R}^{K-CV}(\lambda_n)$. Moreover, it also implies that $\tau^{-1}\lambda^{opt} \notin [\epsilon, \tilde{c}\tau^{-1}]$, that is, $\lambda^{opt} \geq \tilde{c}$. Next, we re-parametrize the above optimization problem:

$$\widetilde{\mu} = \arg\min_{\mu\in[0,\tilde{c}^{-1}]} \widetilde{R}(\mu), \text{ where } \widetilde{R}(\mu) := \hat{R}^{K-CV}(\tau^{-1}\mu^{-1}),$$

and we extend the domain of $\widetilde{R}(\cdot)$ to include 0: $\widetilde{R}(0) := \lim_{\mu\to 0}\widetilde{R}(\mu) = \|y\|^2/n$. Lemma 19 implies that $pn^{-1}\tau^{-2}\widetilde{R}(\mu)$ satisfies stochastic equicontinuity. Using this fact and Theorem 1 of Newey (1991), the convergence of

$$pn^{-1}\tau^{-2}\left\{\widetilde{R}(\mu) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} \xrightarrow{\text{P}} \frac{2(K-1)}{K}\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2\mu^2}{2} - \sigma_\beta^2\mu\right)$$

holds uniformly over the interval $[0, \tilde{c}^{-1}]$. Since $(\lambda^{opt})^{-1}$ is a unique minimizer of the right-hand-side and is distinct from zero, it follows that $\widetilde{\mu} \xrightarrow{\text{P}} (\lambda^{opt})^{-1}$ and $\widetilde{\mu} = \tau^{-1}(\hat{\lambda}_n^{K-CV})^{-1}$, w.a.p.1, which conclude the proof. □

D.4 Proof of Theorem 4

Proof. For ease of notation, we let $\hat{\beta} := \hat{\beta}_l(\lambda_n)$. We adopt the same notation δ_1^* and $\mu(\sigma_x\sigma_\beta, \delta_1^*, \delta_2)$ as used in the proof of Theorem 2. With respect to δ_2^* and C_n^ϕ, we define them as $2\sigma_x^2\sigma_\beta^2\theta_4/\delta_1^*$ and $c_n\tau^{-1}\sigma_x^2\sigma_\beta^2$, respectively.

Analogous to the proof of Theorem 2, it is essential to establish that the following inequality holds w.a.p.1 for any sufficiently small $\epsilon > 0$:

$$\frac{c_\alpha}{2\sigma_\beta} + \epsilon \leq c_n\tau^{-1}(\tau^{-1/2}\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\| - \sigma_x\sigma_\beta) \leq \frac{C_\alpha}{2\sigma_\beta} - \epsilon. \tag{D26}$$

Define $w = \tau^{-3/2}\Sigma_2^{1/2}(\beta - \beta_0)$. Using this, we can rewrite (3) as the following problem:

$$\hat{w} = \arg\min_w \frac{c_n}{n}\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\|^2 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\|_1 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.$$

Define $S_w^n = \{w \big| c_n\tau^{-1}\sigma_x\sigma_\beta + c_\alpha/4\sigma_\beta \leq c_n\|w\| \leq c_n\tau^{-1}\sigma_x\sigma_\beta + C_\alpha/\sigma_\beta\}$. Analogous to the result proved by Lemma 14, if the solution $\hat{w}^B$ to the following problem

$$\min_{w\in S_w^n} \frac{c_n}{n}\|\tau^{1/2}\Sigma_1^{1/2}Z\tilde{w} - \tau^{-1}\varepsilon\|^2 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\|_1 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi \tag{D27}$$

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, (D27). In light of this, without ambiguity we now directly focus on (D27), 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 $\|x\|_1 = \max_{\|v\|_\infty\leq 1} v^\top x$. 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 (D27) can be reformulated as:

$$\begin{aligned}\min_{w\in S_w^n} \max_{\substack{\tilde{u}\\ \|v\|_\infty\leq 1}} \frac{c_n\tau^{1/2}}{\sqrt{n}}\tilde{u}^\top Zw - \frac{c_n\tau^{-1}}{\sqrt{n}}\tilde{u}^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}\tilde{u}\|^2}{4} + \frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}}v^\top\beta_0 \\ + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}v^\top\Sigma_2^{-1/2}w - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.\end{aligned} \tag{D28}$$

For convenience, we shall continue to employ u in place of $\tilde{u}$ throughout the remainder of the proof. Let $S_u^n = \{u\big|\|u\| \leq 4\tau^{-1}\sqrt{C_1C_\varepsilon}\}$. Similar to the proof of Theorem 2, w.a.p.1, the optimization problem below

$$\min_{w\in S_w^n}\max_{\substack{u\in S_u^n\\ \|v\|_\infty\le 1}} \frac{c_n\tau^{1/2}}{\sqrt{n}}u^\top Zw - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} + \frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}}v^\top\beta_0$$
$$+\frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}v^\top\Sigma_2^{-1/2}w - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi \tag{D29}$$

is equivalent to Eq. (D28). Next, we construct an auxiliary optimization problem:

$$\phi(g,h) = \max_{\substack{0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}\\ \|v\|_\infty\le 1}}\min_{w\in S_w^n}\max_{\|u\|=\delta}\mathcal{R}_n(w,v,u), \quad \text{where}$$
$$\mathcal{R}_n(w,v,u) = \frac{c_n\tau^{1/2}}{\sqrt{n}}\|w\|g^\top u - \frac{c_n\tau^{1/2}}{\sqrt{n}}\|u\|h^\top w - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} \tag{D30}$$
$$+\frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}}v^\top\beta_0 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}v^\top\Sigma_2^{-1/2}w - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi,$$

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 (D30), with $w\in S_w^n\cap\tilde{\mathcal{S}}_n^c$.

Lemma 20 characterizes the limiting behavior of the optimal solution to (D29), $\hat{w}$, and in turn, proves the desired (D26), under conditions pertaining to the optimization problem (D30). Therefore, we only need show that conditions outlined in Lemma 20 hold. That is, 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 Theorem 2, after maximizing over the direction of u and minimizing over the direction of w, Eq. (D30) becomes equivalent to:

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

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 16, 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 $\left\|n^{-1/2}\tau^{1/2}\delta h^\top - n^{-1/2}\tau^{-1/2}\lambda_n v^\top\Sigma_2^{-1/2}\right\|\alpha$, we obtain:

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

By completing the square for terms associated with v, we can rewrite this problem as:

$$\min_{\alpha\in K_\alpha}\max_{\substack{\gamma>0\\ 0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}}} -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}$$
$$\times(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)-\frac{c_n\gamma}{2}+\frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0+\frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0 \tag{D31}$$
$$-\min_{\|v\|_\infty\le 1}\frac{c_n\alpha^2}{2\gamma}\left\|n^{-1/2}\tau^{-1/2}\lambda_n\Sigma_2^{-1/2}v-n^{-1/2}\tau^{1/2}\delta h-\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2^{1/2}\beta_0\right\|^2-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2-C_n^\phi.$$

Denote the objective function as $Q_n(\alpha,\delta,\gamma)$. Similar to Theorem 2, we define $\gamma=\tau^{-1}\gamma_1$, $\delta=\tau^{-1}\delta_1^*+\delta_2^*+c_n^{-1/2}\delta_3$, and $\alpha=\tau^{-1}\sigma_x\sigma_\beta+c_n^{-1}\alpha_2$. We obtain the modified objective function $\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ as follows:

$$\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1):=Q_n(\tau^{-1}\sigma_x\sigma_\beta+c_n^{-1}\alpha_2,\tau^{-1}\delta_1^*+\delta_2^*+c_n^{-1/2}\delta_3,\tau^{-1}\gamma_1). \tag{D32}$$

Note that $\delta_3\in K_{\delta_3}:=[-c_n^{1/2}(\tau^{-1}\delta_1^*+\delta_2^*),4c_n^{1/2}\tau^{-1}\sqrt{C_1C_\varepsilon}-c_n^{1/2}(\tau^{-1}\delta_1^*+\delta_2^*)]$. Finally, Lemma 21 verifies the following inequalities:

$$\begin{aligned}\phi(g,h)&=\min_{\alpha_2\in[\frac{c_\alpha}{4\sigma_\beta},\frac{C_\alpha}{\sigma_\beta}]}\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}}\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)<-\frac{C_\lambda}{8C_2}+\eta,\\ \phi_{\tilde{\mathcal{S}}_n^c}(g,h)&=\min_{\alpha_2\in[\frac{c_\alpha}{4\sigma_\beta},\frac{c_\alpha}{2\sigma_\beta}+\epsilon]\cup[\frac{C_\alpha}{2\sigma_\beta}-\epsilon,\frac{C_\alpha}{\sigma_\beta}]}\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}}\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)>-\frac{C_\lambda}{100C_2}-\eta,\end{aligned} \tag{D33}$$

which hold for sufficiently small $\epsilon>0$ and $\eta>0$. With Lemma 20, we conclude the proof. □

D.5 Proof of Proposition 1

Proof. Since the out-of-sample data are mutually independent, Lemmas 2 and 3 lead to:

$$\sum_{i\in\mathrm{OOS}}y_i^2=n_{\mathrm{oos}}(\sigma_\varepsilon^2+\tau\sigma_x^2\sigma_\beta^2)+o_{\mathrm{P}}(n_{\mathrm{oos}}\tau),$$

$$\sum_{i\in\mathrm{OOS}} y_i^2 - (y_i - X_i\hat{\beta}_r(\lambda_n^{opt}))^2 = -n_{\mathrm{oos}}p^{-1}n\tau^2\Delta(\hat{\beta}_r(\lambda_n^{opt})) + O_{\mathrm{P}}(n_{\mathrm{oos}}^{1/2}\tau)$$
$$= n_{\mathrm{oos}}p^{-1}n\tau^2\theta_2\sigma_x^4\sigma_\beta^4\sigma_\varepsilon^{-2}(1+o_{\mathrm{P}}(1)),$$

where we use $\Delta(\hat{\beta}_r(\lambda_n^{opt})) = -\theta_2\sigma_x^4\sigma_\beta^4\sigma_\varepsilon^{-2} + o_{\mathrm{P}}(1)$ by Theorem 2 and $n_{\mathrm{oos}}p^{-2}n^2\tau^2 \to \infty$ in the last equation. The estimates above offer the key components for deriving the limit of R^2_{oos}:

$$R^2_{\mathrm{oos}} = \frac{\sum_{i\in\mathrm{OOS}} y_i^2 - (y_i - X_i\hat{\beta}_r(\lambda_n^{opt}))^2}{\sum_{i\in\mathrm{OOS}} y_i^2} = p^{-1}n\theta_2(R^2)^2(1+o_{\mathrm{P}}(1)). \qquad \square$$

D.6 Proof of Theorem 5

Proof. For convenience, let $\hat{\beta} := \hat{\beta}_r(\lambda_n)$. We write the prediction error of the benchmark as:

$$\begin{aligned} y^{new} - \hat{y}_b^{new} =& (w^{new})^\top\gamma_0 + (x^{new})^\top\beta_0 + \varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top(W\gamma_0 + X\beta_0 + \varepsilon) \\ =& ((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)\beta_0 + (\varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top\varepsilon). \end{aligned}$$

Similarly, for the Ridge estimator, we have

$$y^{new} - \hat{y}^{new} = ((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)(\beta_0 - \hat{\beta}) + (\varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top\varepsilon).$$

As a result, with simple algebra we obtain

$$\begin{aligned} &\mathbb{E}\left[(y^{new} - \hat{y}^{new})^2|\mathcal{I}\right] - \mathbb{E}\left[(y^{new} - \hat{y}_b^{new})^2|\mathcal{I}\right] \\ =&\mathbb{E}\left[\left((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)(\beta_0 - \hat{\beta})\right)^2|\mathcal{I}\right] \\ &- \mathbb{E}\left[\left((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)\beta_0\right)^2|\mathcal{I}\right] \\ &- 2\mathbb{E}\left[((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)\hat{\beta}(\varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top\varepsilon)|\mathcal{I}\right] \\ :=& S_1 - S_2 - S_3. \end{aligned}$$

Below we analyze S_1 to S_3 one by one. We start with S_2. Using the independence of w^{new} with W, U, $\mathcal{I}$, and β_0, the fact that w^{new} has bounded variance, we have

$$\mathbb{E}\left[((w^{new})^\top(W^\top W)^{-1}W^\top U\beta_0)^2|\mathcal{I}\right] \asymp \|(W^\top W)^{-1}W^\top U\beta_0\|^2 = \|(W^\top W)^{-1}W^\top\Sigma_1^{1/2}Z\Sigma_2^{1/2}\beta_0\|^2.$$

Let $x \in \mathbb{R}^n$ be a standard Gaussian vector, independent of (W, Z, β_0). Since $Z\Sigma_2^{1/2}\beta_0$ and $x\|\Sigma_2^{1/2}\beta_0\|$ share the same distribution, $\|(W^\top W)^{-1}W^\top\Sigma_1^{1/2}x\|^2 \asymp_{\mathrm{P}} \|(W^\top W)^{-1}W^\top\Sigma_1^{1/2}\|_F^2$ by Lemma 2, $\mathrm{Tr}(W^\top W)^{-1} = o_{\mathrm{P}}(p^{-1}n\tau)$, and that $\|\Sigma_2^{1/2}\beta_0\|^2 \asymp_{\mathrm{P}} \tau$, it follows that

$$\|(W^\top W)^{-1}W^\top\Sigma_1^{1/2}Z\Sigma_2^{1/2}\beta_0\|^2 \asymp_{\mathrm{P}} \|\Sigma_2^{1/2}\beta_0\|^2\|(W^\top W)^{-1}W^\top\Sigma_1^{1/2}\|_{\mathrm{F}}^2 \asymp_{\mathrm{P}} o_{\mathrm{P}}(p^{-1}n\tau^2). \quad \text{(D34)}$$

Additionally, given that u^{new} is mean zero, independent of the remaining terms, we have $\mathbb{E}\left[(w^{new})^\top(W^\top W)^{-1}W^\top U\beta_0(u^{new})^\top\beta_0|\mathcal{I}\right] = 0$. Therefore, we have established that

$$S_2 = \mathbb{E}\left[((u^{new})^\top\beta_0)^2|\mathcal{I}\right] + o_{\mathrm{P}}(p^{-1}n\tau^2) = \|\Sigma_2^{1/2}\beta_0\|^2 + o_{\mathrm{P}}(p^{-1}n\tau^2).$$

With respect to S_1, we note that $\hat{\beta} = n^{-1}(n^{-1}X^\top\mathcal{M}_W X + n^{-1}p\tau^{-1}\lambda\mathbb{I})^{-1}X^\top\mathcal{M}_W(U\beta_0+\varepsilon)$. Define $R_X = (n^{-1}X^\top\mathcal{M}_W X + n^{-1}p\tau^{-1}\lambda\mathbb{I})^{-1}$. By direct calculations, we have

$$\begin{aligned}
&\mathbb{E}\left[((w^{new})^\top(W^\top W)^{-1}W^\top U\hat{\beta})^2|\mathcal{I}\right] \asymp \|(W^\top W)^{-1}W^\top U\hat{\beta}\|^2 \\
\leq& 2n^{-2}\|(W^\top W)^{-1}W^\top UR_X X^\top\mathcal{M}_W U\beta_0\|^2 + 2n^{-2}\|(W^\top W)^{-1}W^\top UR_X X^\top\mathcal{M}_W\varepsilon\|^2. \quad \text{(D35)}
\end{aligned}$$

For the second term in (D35), we first note that for any constant $\lambda > 0$,

$$\|R_X X^\top\mathcal{M}_W XR_X\| = \|R_U U^\top\mathcal{M}_W UR_U\| = \frac{n\lambda_1(n^{-1}U^\top\mathcal{M}_W U)}{(\lambda_1(n^{-1}U^\top\mathcal{M}_W U) + n^{-1}p\tau^{-1}\lambda)^2} \asymp_{\mathrm{P}} p^{-1}n^2\tau^2,$$

since $\|n^{-1}U^\top\mathcal{M}_W U\| \lesssim_{\mathrm{P}} n^{-1}\|U\|^2 \lesssim_{\mathrm{P}} 1 + c_n = o(n^{-1}p\tau^{-1})$ by Lemma 6. Therefore, by Lemma 2 and using inequality $\mathrm{Tr}(AB) \leq \|A\|\,\mathrm{Tr}(B)$, for any $A = A^\top$ and $B \geq 0$, we have

$$\begin{aligned}
&n^{-2}\|(W^\top W)^{-1}W^\top UR_X X^\top\mathcal{M}_W\varepsilon\|^2 \\
\asymp_{\mathrm{P}}& n^{-2}\,\mathrm{Tr}((W^\top W)^{-1}W^\top UR_X X^\top\mathcal{M}_W XR_X U^\top W(W^\top W)^{-1}) \\
=& n^{-2}\,\mathrm{Tr}(R_X X^\top\mathcal{M}_W XR_X U^\top W(W^\top W)^{-2}W^\top U) \\
\leq& n^{-2}\|R_X X^\top\mathcal{M}_W XR_X\|\,\mathrm{Tr}(U^\top W(W^\top W)^{-2}W^\top U) \\
\lesssim_{\mathrm{P}}& p^{-1}\tau^2\,\mathrm{Tr}(U^\top W(W^\top W)^{-2}W^\top U) \leq p^{-1}\tau^2\|U^\top U\|\,\mathrm{Tr}((W^\top W)^{-1}) = o_{\mathrm{P}}(p^{-1}n\tau^3).
\end{aligned}$$

Similarly, we can prove that the first term in (D35) is of order $o_{\mathrm{P}}(p^{-1}n\tau^3)$. Therefore, we have

$$\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1} W^\top U\hat{\beta})^2 | \mathcal{I}\right] = o_{\mathrm{P}}(p^{-1}n\tau^3). \tag{D36}$$

With (D34) and (D35), we have

$$\begin{aligned}
&\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1} W^\top U(\beta_0 - \hat{\beta}))^2 | \mathcal{I}\right] \\
\leq& 2\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1} W^\top U\beta_0)^2 | \mathcal{I}\right] + 2\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1} W^\top U\hat{\beta})^2 | \mathcal{I}\right] = o_{\mathrm{P}}(p^{-1}n\tau^2).
\end{aligned}$$

In addition, since u^{new} is independent of $\mathcal{I}$ and w^{new}, we have $\mathbb{E}\left[(u^{new})^\top ((\beta_0 - \hat{\beta})((w^{new})^\top (W^\top W)^{-1} W^\top U(\beta_0 - \hat{\beta}) | \mathcal{I}\right] = 0$. Therefore, we conclude

$$S_1 = \mathbb{E}\left[((u^{new})^\top (\beta_0 - \hat{\beta}))^2 | \mathcal{I}\right] + o_{\mathrm{P}}(p^{-1}n\tau^2) = \|\Sigma_2^{1/2}(\beta_0 - \hat{\beta})\|^2 + o_{\mathrm{P}}(p^{-1}n\tau^2).$$

Finally we bound S_3. Since u^{new} and ε^{new} are mean zero, mutually independent, and independent of $\mathcal{I}$, along with Eq. (D36), Lemma 2 and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
|S_3| &= \left|2\mathbb{E}\left[(w^{new})^\top (W^\top W)^{-1} W^\top U\hat{\beta}(w^{new})^\top (W^\top W)^{-1} W^\top \varepsilon | \mathcal{I}\right]\right| \\
&\leq 2\left(\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1} W^\top U\hat{\beta})^2 | \mathcal{I}\right]\right)^{1/2} \left(\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1} W^\top \varepsilon)^2 | \mathcal{I}\right]\right)^{1/2} \\
&\asymp_{\mathrm{P}} o_{\mathrm{P}}(p^{-1/2}n^{1/2}\tau^{3/2})(\mathrm{Tr}(W(W^\top W)^{-2}W^\top))^{1/2} = o_{\mathrm{P}}(p^{-1}n\tau^2).
\end{aligned}$$

In total, we conclude that

$$S_1 - S_2 - S_3 = \|\Sigma_2^{1/2}(\beta_0 - \hat{\beta})\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 + o_{\mathrm{P}}(p^{-1}n\tau^2).$$

Now we prove that,

$$pn^{-1}\tau^{-2}(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2) \xrightarrow{\mathrm{P}} \alpha^*. \tag{D37}$$

By Theorem 2, $\widetilde{\beta} := n^{-1}\widetilde{R}_U U^\top (U\beta_0 + \varepsilon)$ satisfies (D37) with $\hat{\beta}$ being replaced by $\widetilde{\beta}$, where $\widetilde{R}_U := (n^{-1}U^\top U + n^{-1}p\tau^{-1}\lambda\mathbb{I})^{-1}$. Given that $\|\Sigma_2^{1/2}(\beta_0 - \hat{\beta})\|^2 = \|\Sigma_2^{1/2}(\beta_0 - \widetilde{\beta}) + \Sigma_2^{1/2}(\hat{\beta} - \widetilde{\beta})\|^2$ and that $\|\Sigma_2^{1/2}(\beta_0 - \widetilde{\beta})\| \asymp_{\mathrm{P}} \|\Sigma_2^{1/2}\beta_0\| \asymp_{\mathrm{P}} \tau^{1/2}$, it is easy to verify that (D37) follows from $\|\hat{\beta} - \widetilde{\beta}\|^2 = o(n^2p^{-2}\tau^3)$. Below we show this bound is satisfied. Note that

$$\begin{aligned}
\|\hat{\beta} - \widetilde{\beta}\|^2 &\leq \frac{2}{n^2}\|R_U U^\top \mathcal{M}_W(U\beta_0 + \epsilon) - R_U U^\top (U\beta_0 + \epsilon)\|^2 + \frac{2}{n^2}\|(R_U - \widetilde{R}_U)U^\top (U\beta_0 + \epsilon)\|^2 \\
&\leq \frac{4}{n^2}\|R_U U^\top \mathcal{M}_W \varepsilon - R_U U^\top \varepsilon\|^2 + \frac{4}{n^2}\|R_U U^\top \mathcal{M}_W U\beta_0 - R_U U^\top U\beta_0\|^2
\end{aligned}$$

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

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

$$\begin{aligned}&\frac{4}{n^2}\|R_UU^\top\mathcal{M}_W\varepsilon-R_UU^\top\varepsilon\|^2\asymp_{\mathrm{P}}\frac{1}{n^2}\operatorname{Tr}((\mathcal{M}_WU-U)R_U^2(U^\top\mathcal{M}_W-U^\top))\\ =&\frac{1}{n^2}\operatorname{Tr}(U^\top W(W^\top W)^{-1}W^\top UR_U^2)\le\frac{1}{n^2}\|R_U^2\|\|U^\top U\|\operatorname{Tr}(W(W^\top W)^{-1}W^\top)\lesssim_{\mathrm{P}}p^{-1}\tau^2\operatorname{rank}(W).\end{aligned}$$

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 $\operatorname{Tr}(AB)\le\|A\|\operatorname{Tr}(B)$ and $\|\widetilde{R}_U\|\lesssim_{\mathrm{P}} np^{-1}\tau$, we have

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

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 $\operatorname{rank}(W)=o(n^2p^{-1}\tau)$. □

E 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 $\psi:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$.

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 $\mathrm{P}(\Phi(G)<c)\le 2\mathrm{P}(\phi(g,h)\le c)$, $\forall c\in\mathbb{R}$. 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 $\mathrm{P}(\Phi(G)>c)\le 2\mathrm{P}(\phi(g,h)\ge c)$, $\forall c\in\mathbb{R}$.*

The next lemma follows from Lemma B.26 from Bai and Silverstein (2009), which addresses the convergence of a quadratic form concerning a random vector with i.i.d. entries.

Lemma 2. *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 $\mathbb{E}x_i^4 \le v_4$. Then, for any $A \in \mathbb{R}^{n\times n}$, it holds that $x^\top Ax - \operatorname{Tr}(A) = O_{\mathrm{P}}\big(\sqrt{v_4 \operatorname{Tr}(AA^\top)}\big)$.*

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 $x^\top Ay = O_{\mathrm{P}}\big(\sqrt{\operatorname{Tr}(AA^\top)}\big)$.*

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

The following result pertains to the Neumann series. A detailed proof and further discussion are available in Meyer (2000).

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, $\|(\mathbb{I} - A)^{-1} - \sum_{k=0}^{\ell} A^k\| \le \sum_{k=\ell+1}^{\infty} \|A\|^k = \|A\|^{\ell+1}/(1 - \|A\|)$.*

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*

$$\mathrm{P}\left(|x^\top Ax - \mathbb{E}x^\top Ax| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|A\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|A\|^{1/2}}\right\}\right), \tag{E1}$$

$$\mathrm{P}\left(|x^\top By| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|B\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|B\|^{1/2}}\right\}\right). \tag{E2}$$

Proof. Inequality (E1) is given by Proposition 1.1 presented in Götze et al. (2021) for the case of symmetric A. To prove it for the asymmetric case, we use the fact that $x^\top Ax = x^\top(A + A^\top)x/2$, so that we can apply (E1) to $(A + A^\top)/2$. Using triangle inequalities, we have $\big\|(A + A^\top)/2\big\|_{\mathrm{F}}^2 \le \|A\|_{\mathrm{F}}^2$ and $\big\|(A + A^\top)/2\big\|^{1/2} \le \|A\|^{1/2}$, (E1) holds for asymmetric A. To prove (E2), let $z = (x^\top, y^\top)^\top$ and $C = \begin{pmatrix} 0_{n\times n} & B \\ 0_{p\times n} & 0_{p\times p} \end{pmatrix}$. Applying (E1), we obtain

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

□

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 $n^{-1}Z^\top Z \leq (1+\epsilon)(1+\sqrt{c_n})^2$, 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) and Cor. II.1 of Andersen and Gill (1982). 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 $M(x) > \inf_{y>0} M(y)$, $\forall x \geq z$. 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 $|x| \geq z$. 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 $M(x)$. Assumption (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 $\inf_{\substack{|x|>z \\ x\in[x_n,y_n]}} M_n(x) < M_n(x^*)$. 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

$$\min(M_n(z), M_n(-z)) \leq \theta_n M_n(z_n) + (1-\theta_n) M_n(x^*) < M_n(x^*).$$

By taking limits on both sides, we have $\min(M(z), M(-z)) \leq M(x^*)$, which contradicts Assumption (b). Therefore, w.p.a.1, we have $\inf_{\substack{|x|>z \\ x\in[x_n,y_n]}} M_n(x) \geq M_n(x^*)$. Furthermore, by Lemma 7, for all arbitrarily small $\epsilon > 0$, w.p.a.1, $\sup_{|x|\leq z} |M_n(x) - M(x)| < \epsilon$. In addition, by definition, there exists a sequence of z_n, such that $|z_n| \leq z$ and $\inf_{|x|\leq z} M_n(x) \geq M_n(z_n) - \epsilon$. Combining these two inequalities with the fact that $M(x^\star)$ minimizes M on $\mathbb{R}$ leads to $\inf_{|x|\leq z} M_n(x) \geq M_n(z_n) - \epsilon \geq M(z_n) - 2\epsilon \geq M(x^*) - 2\epsilon$, 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 $\inf_{|x|\le z} M_n(x) \xrightarrow{\mathrm{P}} M(x^*)$. Along with $\inf_{\substack{|x|>z \\ x\in[x_n,y_n]}} M_n(x) \ge M_n(x^*)$ and $M_n(x^\star) \xrightarrow{\mathrm{P}} M(x^*)$ by (a), we have

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

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:*

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

Proof. We only present the proof for the first inequality, noting that the proof for the second inequality follows a parallel methodology. By Lemma 10, with x sufficiently large we have

$$\mathbb{E}(|X| - x)_+^2 = \frac{2\sqrt{2}}{x^3\sqrt{\pi}}\exp\left(-\frac{x^2}{2}\right) + o\left(\frac{1}{x^3}\right)\exp\left(-\frac{x^2}{2}\right).$$

By integration by parts, we arrive at

$$\mathbb{E}|X|(|X| - x)_+^2 = \frac{2\sqrt{2}}{x^2\sqrt{\pi}}\exp\left(-\frac{x^2}{2}\right) + o\left(\frac{1}{x^2}\right)\exp\left(-\frac{x^2}{2}\right).$$

The desired result directly follows from the above two estimates. $\square$

Definition 1. *A centered random variable $X \in SE(\nu^2, \alpha)$ 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 $q^{1/2}b_{0,i} = b_{1i}b_{2i} \in \mathrm{SE}(1,1)$.

Write the (i,j)-th element of Σ_2 as $\Sigma_{2,ij}$. By the properties of sub-exponential variables, we have $\left(\Sigma_2 q^{1/2}b_0\right)_i \in \mathrm{SE}\left(\sum_{j=1}^p \Sigma_{2,ij}^2, \max_j|\Sigma_{2,ij}|\right)$. 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 $\mathrm{P}\left(|(\Sigma_2 q^{1/2}b_0)_i| > x_p\log(p)\right) \le 2\exp\left(-\frac{x_p\log(p)}{2C_2}\right)$. Therefore, applying union bound inequality, we obtain

$$\mathrm{P}\left(\|\Sigma_2 b_0\|_\infty > x_p q^{-1/2}\log(p)\right) \le 2p\exp\left(-\frac{x_p\log(p)}{2C_2}\right) \to 0. \qquad \square$$

Lemma 13. *For the event* A_n *defined in Eq. (D2), we have* $\mathrm{P}(A_n) \ge 1 - p^{-1}$ *as* p *is sufficiently large.*

Proof. We start with the second event in A_n. Note that

$$\begin{aligned}\sum_{k=1}^n \tilde{x}_k^2 =& X_{\cdot,i}^\top \Sigma_\varepsilon^{-1} X_{\cdot,i} \le c_\varepsilon^{-1} X_{\cdot,i}^\top X_{\cdot,i} = c_\varepsilon^{-1} e_i^\top \Sigma_2^{1/2} Z^\top \Sigma_1 Z \Sigma_2^{1/2} e_i \\ \le& c_\varepsilon^{-1} C_1 \|Z\Sigma_2^{1/2}e_i\|^2 \overset{d}{=} c_\varepsilon^{-1} C_1 \|\Sigma_2^{1/2}e_i\|^2 \chi^2(n) \le c_\varepsilon^{-1} C_1 C_2 \chi^2(n),\end{aligned}$$

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 $c > 0$, $\chi^2(n) \le 5n$, which implies the second event.

For the first event, we observe that under the second event,

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

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 $\left|\sum_{k=1}^n \tilde{x}_k z_k\right| \leq \tilde{C}n^{1/2}\log^2(p)/2$, which implies $p^{-1/2}\tau^{1/2}\left|\sum_{k=1}^n \tilde{x}_k z_k\right| \leq \tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p)/2$.

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

Lemma 14. *Under the conditions of Theorem 2, define* $S_w^n := \{w | c_n\tau^{-1}\sigma_x\sigma_\beta - K_\alpha \leq c_n\|w\| \leq c_n\tau^{-1}\sigma_x\sigma_\beta + K_\alpha\}$ *for some* K_α *such that* $|\alpha_2^*| < K_\alpha$. *If the solution* $\hat{w}^B$ *to*

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

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. (D6).*

Proof. The proof of this lemma is almost identical to Lemma 5 of Thrampoulidis et al. (2018) and is therefore omitted here. □

Lemma 15. *Let* $\hat{w}$ *denote an optimal solution of Eq. (D10). Regarding* $\phi(g,h)$ *and* $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$, *as introduced and discussed in relation to Eq. (D11), 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.*

Proof. Denote Φ as the optimal value of the minimization problem in Eq. (D10), and $\Phi_{\tilde{\mathcal{S}}_n^c}$ as the optimal value when we impose the constraint $w \in S_w^n \cap \tilde{\mathcal{S}}_n^c$. It is evident that $\hat{w} \in \tilde{\mathcal{S}}_n$ whenever $\Phi_{\tilde{\mathcal{S}}_n^c} > \Phi$. In what follows, we show the latter statement holds w.a.p.1.

Let $\phi^P := \min_{w\in S_w^n}\max_{\substack{u\in S_u^n \\ v}} \mathcal{R}_n(w,v,u)$ and $\phi^D := \max_{\substack{u\in S_u^n \\ v}}\min_{w\in S_w^n} \mathcal{R}_n(w,v,u)$, where $\mathcal{R}_n(w,v,u)$ is given by Eq. (D11). Using min-max inequality (Lemma 36.1 in Rockafellar (1970)), we have

$$\begin{aligned}
\phi^P_{\tilde{\mathcal{S}}_n^c} &:= \min_{w\in S_w^n\cap\tilde{\mathcal{S}}_n^c}\max_{\substack{u\in S_u^n \\ v}} \mathcal{R}_n(w,v,u) = \min_{w\in S_w^n\cap\tilde{\mathcal{S}}_n^c}\max_{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon}}\max_{\substack{\|u\|=\delta \\ v}} \mathcal{R}_n(w,v,u) \\
&\geq \max_{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon}}\min_{w\in S_w^n\cap\tilde{\mathcal{S}}_n^c}\max_{\substack{\|u\|=\delta \\ v}} \mathcal{R}_n(w,v,u) = \phi_{\tilde{\mathcal{S}}_n^c}(g,h), \quad \text{and} \qquad \text{(E3)} \\
\phi^D &= \max_{\substack{u\in S_u^n \\ v}}\min_{w\in S_w^n} \mathcal{R}_n(w,v,u) = \max_{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon}}\max_{\substack{\|u\|=\delta \\ v}}\min_{w\in S_w^n} \mathcal{R}_n(w,v,u) \\
&\leq \max_{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon}}\min_{w\in S_w^n}\max_{\substack{\|u\|=\delta \\ v}} \mathcal{R}_n(w,v,u) = \phi(g,h). \qquad \text{(E4)}
\end{aligned}$$

Utilizing CGMT (Lemma 1), and in conjunction with Eq. (E3), we have

$$\mathrm{P}\left(\Phi_{\tilde{\mathcal{S}}_n^c} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \frac{\kappa}{3}\right) \leq 2\mathrm{P}\left(\phi^P_{\tilde{\mathcal{S}}_n^c} \leq \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \frac{\kappa}{3}\right) \leq 2\mathrm{P}\left(\phi_{\tilde{\mathcal{S}}_n^c}(g,h) \leq \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \frac{\kappa}{3}\right). \tag{E5}$$

Similarly, employing CGMT along with Eq. (E4), we deduce:

$$\mathrm{P}\left(\Phi > \bar{\phi} + \frac{\kappa}{3}\right) \leq 2\mathrm{P}\left(\phi^D \geq \bar{\phi} + \frac{\kappa}{3}\right) \leq 2\mathrm{P}\left(\phi(g,h) \geq \bar{\phi} + \frac{\kappa}{3}\right). \tag{E6}$$

Under assumptions (a) and (b) in this lemma, the right-hand sides of Eqs. (E5) and (E6) vanish as $p \to \infty$, given the choice of $\eta = \kappa/3$ for $\kappa := \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \bar{\phi}$. Consequently, w.a.p.1, we have: $\Phi_{\tilde{\mathcal{S}}_n^c} \geq \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \kappa/3 > \bar{\phi} + \kappa/3 \geq \Phi$, which concludes the proof. □

Lemma 16. *The objective function of Eq. (D14) is convex in α and jointly concave in (δ, v).*

Proof. First, we prove the objective function is convex in $\alpha = \|w\|$. We revisit Eq. (D11):

$$\max_{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon}} \min_{w\in S_w^n} \max_{\|u\|=\delta} \frac{c_n\tau^{1/2}}{\sqrt{n}}\alpha g^\top u - \frac{c_n\tau^{1/2}}{\sqrt{n}}\|u\|h^\top w - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4}$$
$$+ c_n^2\lambda v^\top w + c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2.$$

Note that the term $f(\alpha,u) := \frac{c_n\tau^{1/2}}{\sqrt{n}}\alpha g^\top u - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4}$ is convex in α. After maximizing over the direction of u, the term remains convex in α since $\max_{\|u\|=\delta} f(\theta\alpha_1+(1-\theta)\alpha_2,u) \leq \max_{\|u\|=\delta}\{\theta f(\alpha_1,u)+(1-\theta)f(\alpha_2,u)\} \leq \theta\max_{\|u\|=\delta} f(\alpha_1,u)+(1-\theta)\max_{\|u\|=\delta} f(\alpha_2,u)$, for $\theta\in(0,1)$. Note that from Eq. (D12), $\max_{\|u\|=\delta} f(\alpha,u) = -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)$, which yield the first two terms in Eq. (D14). 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. (D14) 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 $\alpha\geq 0$. 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

$$-\frac{\delta^2}{4}\mu_n(\alpha,\delta)+\frac{1}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon) \tag{E7}$$

is concave in δ. Let the eigenvalues and corresponding 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 (E7) equals

$$-\frac{\delta^2}{4}\mu_n(\alpha,\delta)+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1/\lambda_i-\mu_n(\alpha,\delta)}w_i^2. \tag{E8}$$

The first order derivative of the function (E8) with respect to δ is

$$-\frac{\delta}{2}\mu_n(\alpha,\delta)-\frac{\delta^2}{4}\partial_\delta\mu_n(\alpha,\delta)+\frac{\partial_\delta\mu_n(\alpha,\delta)}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n(\alpha,\delta))^2}w_i^2=-\frac{\delta}{2}\mu_n(\alpha,\delta), \tag{E9}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n(\alpha,\delta))^2}w_i^2=\frac{\delta^2}{4}. \tag{E10}$$

Further, the second-order derivative of the function (E8) 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. (E10):

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

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 17. *For $\tilde{Q}_n=\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ in Eq. (D16), Eqs. (D17) and (D18) hold.*

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:

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

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

$$
\begin{aligned}
&c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)\\
&\quad-\frac{c_n\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\
&\quad-\frac{c_n\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=O_{\mathrm{P}}(c_n\tau^{-1}n^{-1/2})=o_{\mathrm{P}}(1). \quad \text{(E11)}
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\
=&\frac{1}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\tau^{-1}\Sigma_1^2+2\left(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon\right)\Sigma_1^3\right)\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]+O(\tau)\\
=&\tau^{-1}\sigma_\varepsilon^2\theta_1+2(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+O(\tau)+o(c_n^{-1}), \quad \text{(E12)}
\end{aligned}
$$

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 (E11), we can deduce that:

$$
\begin{aligned}
&c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)-\frac{c_n\delta^2\tau}{4}\\
&=c_n\tau^{-1}\sigma_\varepsilon^2\theta_1+2c_n(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+c_n\sigma_x^2\sigma_\beta^2\theta_4-c_n\frac{\tau^{-1}(\delta_1^*)^2+2\delta_1^*\delta_2}{4}+o_{\mathrm{P}}(1)\\
&=2c_n(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+c_n\sigma_x^2\sigma_\beta^2\theta_4-\frac{c_n\delta_1^*\delta_2}{2}+o_{\mathrm{P}}(1).
\end{aligned}
$$

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

$$
c_n\tau^{-1}\mu_n(\alpha,\delta)-c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\xrightarrow{\mathrm{P}}0. \tag{E13}
$$

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 (E13), 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. (D13),

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

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:

$$
\begin{aligned}
&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)\\
&=\frac{c_n}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1-\mu_n^2(\alpha,\delta)\Sigma_1^3\right)\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)+O_{\mathrm{P}}(c_n\tau)
\end{aligned}
$$

$$= c_n\tau^{-1}\sigma_x^2\sigma_\beta^2 - c_n\sigma_\varepsilon^2\theta_3\mu^2(\sigma_x\sigma_\beta, \delta_1^*, \delta_2) + 2\sigma_x\sigma_\beta\alpha_2 + \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 + o_{\mathrm{P}}(1). \tag{E14}$$

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:

$$\begin{aligned}
&\frac{c_n^2\lambda^2\tau^{-2}\sigma_\beta^2}{4p}\operatorname{Tr}\left[\Sigma_2^{1/2}\left(\frac{\lambda}{4}\Sigma_2 + \frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}\Sigma_2^{1/2}\right] + c_n\tau\operatorname{Tr}\left[\frac{c_n\lambda^2\alpha^4\delta^2}{4n\gamma_h^2}\left(\frac{\lambda}{4}\Sigma_2 + \frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1} - \frac{\alpha^2\delta^2}{2\gamma n}\mathbb{I}\right] \\
&= \frac{c_n^2\lambda^2\tau^{-2}\sigma_\beta^2}{4p}\operatorname{Tr}\left[\frac{2\gamma}{c_n\alpha^2\lambda^2}\Sigma_2 + \frac{\gamma_h^2}{c_n^2\alpha^4\lambda^3}\Sigma_2^2\right] + c_n\tau\operatorname{Tr}\left[-\frac{\delta^2}{4c_n n\lambda}\Sigma_2 + \frac{\delta^2\gamma}{8c_n^2 n\alpha^2\lambda^2}\Sigma_2^2\right] + o_n(1) \\
&= \frac{c_n\gamma n_1}{2}\tau^{-1} - \frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2 - \frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta} - \tau^{-1}\frac{(\delta_1^*)^2\sigma_x^2 c_n}{4\lambda} - \frac{c_n\sigma_x^2\delta_1^*\delta_2}{2\lambda} + \frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2 + o_n(1),
\end{aligned}$$

where we apply Lemma 4 and the same argument in proving Eq. (E12). Combining this estimate with (E14) we conclude that

$$\begin{aligned}
\tilde{Q}_n &= c_n\tau^{-1}\sigma_x^2\sigma_\beta^2 - \tau^{-1}\frac{(\delta_1^*)^2\sigma_x^2 c_n}{4\lambda} - c_n\sigma_\varepsilon^2\theta_3\mu^2(\sigma_x\sigma_\beta, \delta_1^*, \delta_2) + 2\sigma_x\sigma_\beta\alpha_2 - \frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2 - \frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta} \\
&\quad - \frac{c_n\delta_1^*\delta_2\sigma_x^2}{2\lambda} + \frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2 - C_n^\phi + o_{\mathrm{P}}(1) \\
&= -\frac{\delta_3^2\theta_1}{4\theta_3} + 2\sigma_x\sigma_\beta\alpha_2 - \frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2 - \frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta} + \frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2 + o_{\mathrm{P}}(1).
\end{aligned}$$

We now proceed to establish Claims (i) to (ii). Let K_{δ_3} be the interval $[-c_n^{1/2}(\tau^{-1}\delta_1^* + \delta_2^*), 4c_n^{1/2}\tau^{-1}\sqrt{C_1C_\varepsilon} - c_n^{1/2}(\tau^{-1}\delta_1^* + \delta_2^*)]$. It is sufficient to demonstrate that, for any compact set $A \subset [-K_\alpha, K_\alpha]$, the following equation holds:

$$\phi_A(g,h) := \min_{\alpha_2\in A}\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}}\tilde{Q}_n(\alpha_2, \delta_3\gamma_1) \xrightarrow{\mathrm{P}} \min_{\alpha_2\in A}\max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}}\tilde{Q}(\alpha_2, \delta_3, \gamma_1). \tag{E15}$$

This is because based on this result, we can deduce

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

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

which lead to (i) and (ii).

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 the Lemma 9, we conclude that $\max_{\delta_3\in K_{\delta_3}} \tilde{Q}_n(\alpha_2,\delta_3,\gamma_1) \xrightarrow{\text{P}} \max_{\delta_3\in\mathbb{R}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. 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 $\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}} \tilde{Q}_n(\alpha_2,\delta_3,\gamma_1) \xrightarrow{\text{P}} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. 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 Eq. (E15) holds. This concludes the proof of Claims (i) and (ii).

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 (iii) holds true, concluding the proof. □

Lemma 18. *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 (D24) holds. In addition, as $n \to \infty$, for any given fixed $\lambda > 0$, Eq. (D25) holds.*

Proof. 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 $n,p\to\infty$.

Let $S := \{\lambda_j = \epsilon + p^{-9}(j-1) : 1 \le j \le 1 + [p^9(\tilde{c}\tau^{-1} - \epsilon)]\}$. Given $\tau^{-1} = o(p)$, the cardinality of the set satisfies $|S| \le p^{10}$. By definition, for any $\lambda \in [\epsilon, \tilde{c}\tau^{-1}]$, there exists a $\lambda_{j^*} \in S$ such that $|\lambda - \lambda_{j^*}| \le p^{-9}$. By Eq. (D23), we have $|\hat{R}^{K-CV}(\lambda) - \hat{R}^{K-CV}(\lambda_{j^*})| \le \tilde{C}|\lambda - \lambda_{j^*}| \le \tilde{C}p^{-9}$. Therefore, if we show that

$$\inf_{\lambda_j\in S}\left\{\hat{R}^{K-CV}(\lambda_j) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} > np^{-1}\tau^2 \tag{E16}$$

holds w.p.a.1, we have

$$\inf_{\lambda\in[\epsilon,\tilde{c}\tau^{-1}]}\left\{\hat{R}^{K-CV}(\lambda_j) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\right\} > np^{-1}\tau^2 - \tilde{C}p^{-9} > \frac{np^{-1}\tau^2}{2}, \tag{E17}$$

which implies Eq. (D24). By Eq. (D19), it is easy to verify that we only need to 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: $n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 \overset{d}{=} n^{-1}K\chi^2\left(K^{-1}n\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2$, where $\chi^2\left(K^{-1}n\right)$ denotes a Chi-squared random variable with $K^{-1}n$ degrees of freedom. Consequently, we can deduce that:

$$\begin{aligned}
&\mathrm{P}\left(\left|n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2\right| \geq \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2\right) \\
&= \mathrm{P}\left(\left|n^{-1}K\chi^2\left(K^{-1}n\right) - 1\right| \geq \frac{\log(p)}{\sqrt{n}}\right) \leq 2\exp(-\tilde{c}_1\log^2(p)),
\end{aligned}$$

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:

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

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

$$\begin{aligned}\|X_{(-i)}^\top y_{(-i)}\|^2 &= \varepsilon_{(-i)}^\top X_{(-i)} X_{(-i)}^\top \varepsilon_{(-i)} + 2\varepsilon_{(-i)}^\top X_{(-i)} X_{(-i)}^\top X_{(-i)}\beta_0 + \beta_0^\top X_{(-i)}^\top X_{(-i)} X_{(-i)}^\top X_{(-i)}\beta_0 \\ &= \sigma_\varepsilon^2 \operatorname{Tr}(X_{(-i)}X_{(-i)}^\top) + p^{-1}\tau\sigma_\beta^2 \operatorname{Tr}(X_{(-i)}^\top X_{(-i)} X_{(-i)}^\top X_{(-i)}) + o_{\mathrm{P}}(n^{-1/2}).\end{aligned}$$

By the fact that $\lambda_{\min}(A)\operatorname{Tr}(B) \le \operatorname{Tr}(AB) \le \lambda_{\max}(A)\operatorname{Tr}(B)$ when A, B are positive semidefinite, we have $c_2 \operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top) \le \operatorname{Tr}(X_{(-i)}X_{(-i)}^\top) \le C_2 \operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top)$, which, along with $(np)^{-1}\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top) \xrightarrow{\mathrm{P}} (K-1)/K$ and Eq. (D20), imply that

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

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

$$\frac{c_2\sigma_\varepsilon^2 pn}{2K} \le \|X_{(-i)}^\top y_{(-i)}\|^2 \le 2C_2\sigma_\varepsilon^2 pn. \tag{E18}$$

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

$$\|\Sigma_2^{1/2}\hat{\beta}_{\lambda_j}^i\|^2 \ge \frac{c_2^2\sigma_\varepsilon^2 p}{2nK}\left(C_2(1+\sqrt{c_n})^2 + c_n\lambda_j\right)^{-2}. \tag{E19}$$

On the other hand, we have

$$\begin{aligned}|\beta_0^\top \Sigma_2 \hat{\beta}_{\lambda_j}^i| \le{}& \frac{1}{n}\left|\beta_0^\top\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\beta_0\right| \\ &+ \frac{1}{n}\left|\varepsilon_{(-i)}^\top X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2\beta_0\right|.\end{aligned} \tag{E20}$$

To bound the first term in (E20), 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 $1 - 2p^{10}\exp(-\tilde{c}_1\log^2(p))$,

$$\begin{aligned}\sup_{\lambda_j\in S}\Bigg|&\frac{1}{n}\beta_0^\top\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\beta_0 \\ &- \frac{p^{-1}\tau}{n}\operatorname{Tr}\left(\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\right)\Bigg| \le q^{-1}p^{-1}\tau n^{1/2}\log(p).\end{aligned} \tag{E21}$$

Moreover, note that $\operatorname{Tr}(AB) \le \|AB\|\operatorname{rank}(AB) \le \|A\|\,\|B\|\operatorname{rank}(B)$, we have

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

where the last inequality uses the fact that $\lambda_1((A+\mathbb{I})^{-1}A)=(\lambda_1(A)+1)^{-1}\lambda_1(A)$ and that $n^{-1}\|X_{(-i)}^\top X_{(-i)}\|\le C_2(1+\sqrt{c_n})^2$. Combining the above inequality with (E21), we have

$$\frac{1}{n}\left|\beta_0^\top\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_j\mathbb{I}\right)^{-1}X_{(-i)}^\top X_{(-i)}\beta_0\right|\le np^{-1}\tau C_2^2\frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2+c_n\lambda_j}$$
$$+q^{-1}p^{-1}\tau n^{1/2}\log(p),\quad\forall\lambda_j\in S.$$

To bound the second term in (E20), we use Lemma 5. By definition, it equals

$$\left|\frac{p^{-1/2}\tau^{1/2}q^{-1/2}}{n}z_{(-i)}^\top X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2(\sqrt{q}b_0)\right|.$$

Using the fact that $\lambda_{\min}(n^{-1}X_{(-i)}^\top X_{(-i)}+c_n\lambda_j\mathbb{I})\ge c_n\lambda_j\ge c_n\epsilon$ and Eq. (D20), we have

$$\left\|X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2\right\|\le C_2c_n^{-1}\epsilon^{-1}\|X_{(-i)}\|\lesssim np^{-1/2}.$$

Furthermore, since $\|A\|_{\mathrm{F}}\le\sqrt{\operatorname{rank}(A)}\,\|A\|$, it follows that

$$\left\|X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2\right\|_{\mathrm{F}}^2\lesssim\operatorname{rank}(X_{(-i)})n^2p^{-1}\lesssim n^3p^{-1}.$$

Therefore, by Lemma 5 and the fact that $\sqrt{q}b_{0,i}$ has bounded sub-exponential norm, it holds that, for some constant $\tilde{c}_1$,

$$\mathrm{P}\left(\frac{1}{n}\left|\varepsilon_{(-i)}^\top X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2\beta_0\right|>q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p)\right)\le 2\exp(-\tilde{c}_1\log^2(p)).$$

As a consequence, with probability at least $1-2p^{10}\exp(-\tilde{c}_1\log^2(p))$, we have

$$\sup_{\lambda_j \in S} \frac{1}{n} \left| \varepsilon_{(-i)}^\top X_{(-i)} \left(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n \lambda_j \mathbb{I} \right)^{-1} \Sigma_2 \beta_0 \right| \leq q^{-1/2} n^{1/2} \tau^{1/2} p^{-1} \log(p). \tag{E22}$$

Therefore, taking all bounds for components of (E20) altogether, we have, w.p.a.1,

$$\begin{aligned} |\beta_0^\top \Sigma_2 \hat{\beta}_{\lambda_j}^i| &\leq np^{-1}\tau C_2^2 \frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j} + p^{-1}q^{-1}\tau n^{1/2}\log(p) + q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p) \\ &\leq 2np^{-1}\tau C_2^2 \frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j}, \end{aligned} \tag{E23}$$

for each $\lambda_j \in S$. In the second inequality, we use the fact that $p^{-1}q^{-1}\tau n^{1/2}\log(p)$ and $q^{-1/2}n^{1/2}\tau^{1/2}p^{-1}\log(p)$ are $o(np^{-1}\tau^2)$ by the assumptions of Theorem 3. With (E19) and (E23), we have

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

This inequality holds w.p.a. 1 as $n, p \to \infty$ for each $\lambda_j \in S$. 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. (D24).

To prove Eq. (D25), note that

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

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

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

Additionally, by Theorem 2, we deduce:

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

Hence, using the fact that $\|\Sigma_2^{1/2}\beta_0\|^2 \asymp \tau$ 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. (D25), 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 19. *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:

$$\begin{aligned}
&\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}}c_n\tau^{-3}\log^{-1}(p)\left\|\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\tau^{-1}\mu^{-1}\mathbb{I}\right)^{-1}-c_n^{-1}\tau\mu\mathbb{I}+\frac{c_n^{-2}\tau^2\mu^2}{n}X_{(-i)}^\top X_{(-i)}\right\|\\
&=\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}}c_n^{-1}\mu^2\tau^{-1}\log^{-1}(p)\left\|\frac{1}{n}X_{(-i)}^\top\left[\left(\mathbb{I}+\frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X_{(-i)}^\top\right)^{-1}-\mathbb{I}\right]X_{(-i)}\right\|\\
&\le\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}}\mu^3c_n^{-2}\log^{-1}(p)\left\|\frac{1}{n}X_{(-i)}^\top X_{(-i)}\right\|^2\xrightarrow{\mathrm{P}}0.
\end{aligned}\tag{E24}$$

The last inequality is a consequence of Eq. (D20) and the fact that

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

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

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

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

$$\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-X_{(i)}\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}\right\|$$
$$\le\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{n^{3/2}}\left\|X_{(i)}\right\|\left\|\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)}+c_n\tau^{-1}\mu_1^{-1}\right)^{-1}-c_n^{-1}\tau\mu_1\mathbb{I}+\frac{c_n^{-2}\tau^2\mu_1^2}{n}X_{(-i)}^\top X_{(-i)}\right\|$$
$$\times\left\|X_{(-i)}^\top y_{(-i)}\right\|=O_{\mathrm{P}}(\tau^3\log(p))=o_{\mathrm{P}}(c_n^{-1/2}\tau),\tag{E25}$$

where we use Eq. (E24), Eq. (D20), and Eq. (E18). Additionally, it is easy to verify that

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

For the first term, by Eq. (E18), we have

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

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

$$\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\|\leq\tilde{C}_1c_n^{-1/2}\tau+o_{\mathrm{P}}(c_n^{-1/2}\tau). \tag{E26}$$

Analogously, we can prove that $\frac{1}{\sqrt{n}}\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\|\leq\tilde{C}_1|\mu_1-\mu_2|c_n^{-1/2}\tau+o_{\mathrm{P}}(c_n^{-1/2}\tau)$ 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: $\sup_{1\leq i\leq K}W_{1i}(\mu_1,\mu_2)\leq\tilde{C}_1^2c_n^{-1}\tau^2|\mu_1-\mu_2|+o_{\mathrm{P}}(c_n^{-1}\tau^2)$ 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. (E25), it holds that

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

Moreover, employing a similar argument to that used in proving Eq. (E26), 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

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

for some constant $\tilde{C}_1$ w.p.a.1. The second term can be shown analogously. As a result, we have $\sup_{1\leq i\leq K}W_{2i}(\mu_1,\mu_2)\leq\tilde{C}_1c_n^{-1}\tau^2|\mu_1-\mu_2|+o_{\mathrm{P}}(c_n^{-1}\tau^2)$ 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 20. *Let $\hat{w}$ denote an optimal solution of Eq. (D29). Regarding $\phi(g,h)$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$, as introduced and discussed in relation to Eq. (D30), 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.*

Proof. The proof closely resembles that of Lemma 15 and is thus omitted. □

Lemma 21. *There exists some sufficiently small $\epsilon > 0$, such that for any $\eta > 0$, w.p.a.1, the inequalities in (D33) hold.*

Proof. By Eq. (E14) in Lemma 17, we have the following result:

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

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 $\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$:

$$
\min_{\|v\|_\infty\le 1}\left\{\frac{c_n\alpha^2}{2\gamma}\left\|n^{-1/2}\tau^{-1/2}\lambda_n\Sigma_2^{-1/2}v-n^{-1/2}\tau^{1/2}\delta h-\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2^{1/2}\beta_0\right\|^2\right\}. \tag{E27}
$$

By using $\|\Sigma_2^{-1}\|\le c_2^{-1}$, we see (E27) 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 $\lambda_{\min}\Sigma_2^{-1}\ge C_2^{-1}$, (E27) is lower bounded by

$$
\frac{c_n\alpha^2}{2\gamma C_2}\left\|\left(\left|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h+\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right|-n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2.
$$

Together with Lemma 22, we deduce that, w.p.a.1, (E27) 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 (D32). We introduce $\tilde{Q}_n^{\mathrm{upper}}(\alpha_2,\delta_3,\gamma_1)$, defined as:

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

and $\tilde{Q}_n^{\text{lower}}(\alpha_2, \delta_3, \gamma_1)$, defined as:

$$-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)$$
$$-\frac{c_n\gamma}{2}-\frac{\sigma_x^2\sigma_\beta^2}{\gamma_1 c_2}C_\lambda+\frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0+\frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2-C_n^\phi.$$

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

$$\frac{C_\lambda}{2}\le c_n\tau^{-1}\left\|\left(\left|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h+\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right|-n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2\le 2C_\lambda. \tag{E28}$$

Proof. We first establish the following:

$$c_n n^{-1}\tau^{-2}\left\{\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n\right)_+\right\|^2-\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n\right)_+\right\|^2\right\}\xrightarrow{\text{P}}0. \tag{E29}$$

Let $\tilde{h} := \Sigma_2^{1/2}h$; we then have $\tilde{h} \sim \mathcal{N}(0, \Sigma_2)$. 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,

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

$$\equiv \mathbb{E}\left\{(|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2 \mathbb{E}\left[\left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}\tilde{h}_i + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2 \right.\right.$$
$$\left.\left. - \left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}g_2 + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2 \middle| \tilde{h}_i\right]\right\},$$

where $g_2 \sim \mathcal{N}(0, \Sigma_{2,ii})$ is independent of both g_1 and $\tilde{h}_i$. It is straightforward to confirm that the following inequality holds true:

$$\begin{aligned}
&\left|\left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}\tilde{h}_i + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2 \right. \\
&\quad \left. - \left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}g_2 + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2\right| \\
&\leq |\delta_1^*\Sigma_{2,ij}\Sigma_{2,ii}^{-1}(\tilde{h}_i - g_2)| \cdot \left|\left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}\tilde{h}_i + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+ \right. \\
&\qquad \left. + \left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}g_2 + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+\right|.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality to the above inequality yields

$$\begin{aligned}
&\mathbb{E}\left[\left|\left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}\tilde{h}_i + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2 \right.\right. \\
&\quad \left.\left. - \left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}g_2 + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2\right| \middle| \tilde{h}_i\right] \\
&\lesssim \left(\mathbb{E}\left(|\Sigma_{2,ij}\Sigma_{2,ii}^{-1}(\tilde{h}_i - g_2)|^2 \middle| \tilde{h}_i\right)\right)^{1/2} \left(\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2 \right. \\
&\left. + \mathbb{E}\left(\left(\left|\delta_1^*\left(\Sigma_{2,ij}\Sigma_{2,ii}^{-1}\tilde{h}_i + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1}\Sigma_{2,ij}^2}g_1\right)\right| - \lambda_n\right)_+^2 \middle| \tilde{h}_i\right)\right)^{1/2} \\
&\lesssim |\Sigma_{2,ij}\Sigma_{2,ii}^{-1}|\sqrt{\Sigma_{2,ii} + \tilde{h}_i^2}\sqrt{\Sigma_{2,ij}^2\Sigma_{2,ii}^{-2}\tilde{h}_i^2 + \mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2} \\
&\lesssim |\Sigma_{2,ij}|(1 + |\tilde{h}_i|)\left(|\Sigma_{2,ij}||\tilde{h}_i| + \sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2}\right),
\end{aligned}$$

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

$$\mathrm{Cov}\left((|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2, (|\delta_1^*\tilde{h}_j| - \lambda_n)_+^2\right)$$

$$\lesssim |\Sigma_{2,ij}|(\lambda_n+1)\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2}\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2+\Sigma_{2,ij}^2(\lambda_n^2+\lambda_n)\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2.$$

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

$$\begin{aligned}
&\mathrm{Var}\left(c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n\right)_+\right\|^2\right) = \sum_{i,j=1}^p c_n^2 n^{-2}\tau^{-4}\,\mathrm{Cov}\left((|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2,(|\delta_1^*\tilde{h}_j|-\lambda_n)_+^2\right)\\
&\lesssim c_n^2 n^{-2}\tau^{-4}\log(p)\sum_{i=1}^p\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\left(\sum_{j=1}^p|\Sigma_{2,ij}|\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2}\right)\\
&\quad + c_n^2 n^{-2}\tau^{-4}(\log(p))^2\sum_{i=1}^p\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\sum_{j=1}^p\Sigma_{2,ij}^2\\
&\le c_n^2 n^{-2}\tau^{-4}\log(p)C_2\sum_{i=1}^p\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\left(\sum_{j=1}^p\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y|-\lambda_n)_+^2\right)^{1/2}\\
&\quad + c_n^2 n^{-2}\tau^{-4}(\log(p))^2C_2^2\sum_{i=1}^p\mathbb{E}(|\delta_1^*\tilde{h}_i|-\lambda_n)_+^2\\
&= O\left(c_n^{1/2}n^{-1/2}\tau^{-1}\log(p)+c_n n^{-1}\tau^{-2}\log^2(p)\right)=o_n(1),
\end{aligned}$$

where we use $\sum_{j=1}^p\Sigma_{2,ij}^2\le C_2^2$ and Cauchy–Schwartz inequality in the second step. This leads to Eq. (E29). Using the same approach, we can prove

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

Now we are ready to establish Eq. (E28). Note that w.p.a.1, we have

$$\begin{aligned}
&c_n\tau^{-1}\left\|\left(\left|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h+\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right|-n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2\\
\le& c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|+\left|\tau\Sigma_2^{1/2}\delta_2 h\right|+\left|n^{1/2}\frac{\gamma}{\alpha^2}\tau^{-1}\Sigma_2\beta_0\right|-\lambda_n\right)_+\right\|^2\\
\le& c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right|-\lambda_n+\log^{-1}(p)\right)_+\right\|^2,
\end{aligned}$$

where the last 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. Therefore, w.p.a.1, we have

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

Similarly, it holds that w.p.a.1,

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

Finally, by Lemma 10 and the fact that $\lambda_n = o(\log(p))$, it is easy to verify that

$$\frac{\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^*h\right|-\lambda_n\right)_+\right\|^2}{\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^*h\right|-\lambda_n-\log^{-1}(p)\right)_+\right\|^2}\to 1 \text{ and } \frac{\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^*h\right|-\lambda_n\right)_+\right\|^2}{\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^*h\right|-\lambda_n+\log^{-1}(p)\right)_+\right\|^2}\to 1.$$

Together with the definition that $C_\lambda = \lim_{n\to\infty} pn^{-2}\tau^{-2}\mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^*h\right|-\lambda_n\right)_+\right\|^2$ given by Eq. (9), we conclude the proof. □

Lemma 23. *Let Σ_1 and Σ_ε denote the covariance matrices of stationary processes with exponentially decaying correlations. As a result, both $\|\Sigma_1\|$ and $\|\Sigma_\varepsilon\|$ are bounded and the conditions in Assumption 5 pertaining to these matrices are satisfied.*

Proof. Given the symmetric Toeplitz structure of the matrices Σ_1 and Σ_ε, the elements of these matrices can be defined as $(\Sigma_1)_{ij} = \sigma_{1,|i-j|}$ and $(\Sigma_\varepsilon)_{ij} = \sigma_{\varepsilon,|i-j|}$, respectively. By assumption, the sequences $\{\sigma_{1,k}\}_{k=1}^\infty$ and $\{\sigma_{\varepsilon,k}\}_{k=1}^\infty$ decay exponentially, i.e., there exist positive constants c and C such that $|\sigma_{1,k}|, |\sigma_{\varepsilon,k}| \leq C\exp(-ck)$, for all k.

We first show that the eigenvalues of Σ_1 are bounded. The proof for Σ_ε is similar. Note that

$$\max_{i=1,\ldots,n}\sum_{j=1}^n|(\Sigma_1)_{ij}| = \max_{i=1,\ldots,n}\sum_{j=1}^n|\sigma_{1,|i-j|}| \leq 2\sum_{k=0}^{n-1}|\sigma_{1,k}|.$$

Since $\{\sigma_{1,k}\}_{k=1}^\infty$ decays exponentially, the right-hand-side is bounded, as $n\to\infty$. The bound on $\|\Sigma_1\|$ thus follows from Gershgorin circle theorem.

Next we establish $\frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1) = \sigma_\varepsilon^2\theta_1 + o(c_n^{-1}\tau)$. The proofs for θ_3 and θ_4 follow analogously. The series $\sigma_{1,0}\sigma_{\varepsilon,0} + 2\sum_{i=1}^{\infty}\sigma_{1,i}\sigma_{\varepsilon,i}$ is convergent, since the series $\{\sigma_{1,k}\}_{k=1}^{\infty}$ and $\{\sigma_{\varepsilon,k}\}_{k=1}^{\infty}$ decay exponentially. We denote the limit as $\sigma_\varepsilon^2\theta_1$. Moreover, given that

$$\frac{1}{n}\left[\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1) - n\sigma_\varepsilon^2\theta_1\right] = \frac{1}{n}\left[-\sum_{i=1}^{n-1} 2i\sigma_{1,i}\sigma_{\varepsilon,i} - 2n\sum_{i=n}^{\infty}\sigma_{1,i}\sigma_{\varepsilon,i}\right],$$

and that $\{\sigma_{1,k}\}_{k=1}^{\infty}$ and $\{\sigma_{\varepsilon,k}\}_{k=1}^{\infty}$ decay exponentially, it follows that $-\sum_{i=1}^{n-1} 2i\sigma_{1,i}\sigma_{\varepsilon,i} - 2n\sum_{i=n}^{\infty}\sigma_{1,i}\sigma_{\varepsilon,i} < \infty$, as $n \to \infty$. We thereby have that

$$\frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1) = \sigma_\varepsilon^2\theta_1 + O\left(n^{-1}\right) = \sigma_\varepsilon^2\theta_1 + o(c_n^{-1}\tau). \qquad \square$$

References

Andersen, P. K. and R. D. Gill (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics 10*(4), 1100 – 1120.

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.

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.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*(4), 1161–1167.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

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

Tao, P. D. and L. T. H. An (1998). A d.c. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization 8*(2), 476–505.

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 P. Grünwald, E. Hazan, and S. Kale (Eds.), *Proceedings of The 28th Conference on Learning Theory*, Volume 40 of *Proceedings of Machine Learning Research*, Paris, France, pp. 1683–1709. PMLR.