The Statistical Limit of Arbitrage[*]

Rui Da[†]
Indiana University

Stefan Nagel[‡]
University of Chicago
NBER and CEPR

Dacheng Xiu[§]
University of Chicago
NBER

Abstract

We investigate the economic consequences of statistical learning for arbitrage pricing in a high-dimensional setting. Arbitrageurs learn about alphas from historical data. When alphas are weak and rare, estimation errors hinder arbitrageurs—even those employing optimal machine learning techniques—from fully exploiting all true pricing errors. This statistical limit to arbitrage widens the equilibrium bounds of alphas beyond what traditional arbitrage pricing theory predicts, leading to a significant divergence between the feasible Sharpe ratio achievable by arbitrageurs and the unattainable theoretical maximum under perfect knowledge of alphas.

Keywords: Learning about Alphas, Rational Expectations, Portfolio Choice, Rare and Weak Signal, False Negatives, Empirical Bayes, Testing APT, Machine Learning, Decision-making under Uncertainty

[*] We are grateful for comments from Adem Atmaz, Lars Hansen, Ye Luo, Andreas Neuhierl, Markus Pelger, Seth Pruitt, Yizhou Xiao, Yao Zeng, and seminar and conference participants at Columbia University, Indiana University, Louisiana State University, Princeton University, ITAM Business School, University of Cincinnati, University of Chicago, Yale School of Management, EPFL, ETH Zurich, Stockholm Business School, University of Gothenburg, University of Liverpool, University of Oxford, City University of Hong Kong, Gregory Chow Seminar Series in Econometrics and Statistics, HKUST, KAIST, Peking University, Shanghai University of Finance and Economics, Southern University of Science and Technology, Tsinghua University, University of Melbourne, Sao Paulo School of Economics, American Finance Association Annual Meetings, NBER Summer Institute, Jacobs Levy Center Frontiers in Quantitative Finance Conference, Stanford Institute for Theoretical Economics, SFS Cavalcade North America, Annual SoFiE Conference, Global AI Finance Research Conference in Singapore, China International Conference in Finance, Wabash River Finance Conference, and SAIF Annual Research Conference.

[†] Address: 1275 E 10th St, Bloomington, IN 47405. E-mail: ruida@iu.edu.

[‡] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637. E-mail: stefan.nagel@chicagobooth.edu.

[§] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637. E-mail: dacheng.xiu@chicagobooth.edu.

1 Introduction

The full-information rational expectations hypothesis assumes that economic agents possess perfect knowledge of the economic model, thereby bypassing the statistical challenges real-world agents encounter when estimating its parameters. While this assumption may be plausible in a low-dimensional environment with ample data for learning, many economic settings are more realistically high-dimensional, with unknown model parameters and limited data. This raises important questions: How should agents optimally learn in such environments, and what are the implications for equilibrium outcomes? In this paper, we explore these questions in the context of asset market arbitrage.

The absence of near-arbitrage opportunities is a fundamental principle in most asset pricing theories, including the Arbitrage Pricing Theory (APT). These theories implicitly assume that if investment opportunities with exceptionally high reward-to-risk ratios were to exist, they would attract arbitrageurs. Armed with knowledge of the investment opportunities (i.e., alphas and covariances), these arbitrageurs would exploit and, in turn, eliminate such opportunities. In reality, however, sophisticated investors seeking rewarding investment opportunities lack perfect knowledge. Instead, they rely on statistical analysis to infer the existence of such opportunities from historical data, introducing a statistical challenge. In some cases, such as derivatives pricing, this challenge may be small enough that it does not significantly hinder arbitrage activity. However, in noisier, high-dimensional settings like the cross-section of stock returns, statistical challenge can be substantial, imposing a formidable limit on arbitrage.

To analyze the effects of arbitrageur learning, we consider a setting in which returns follow a statistical linear factor model, with the high-dimensional alphas and volatilities drawn from a prior distribution.[1] Arbitrageurs evaluate trading strategies based on mean-variance utility, seeking factor-neutral arbitrage portfolios that optimally capture factor model alphas while balancing risks. Importantly, these strategies must be feasible, relying solely on observable data such as historical returns. This feasibility constraint reflects the reality that arbitrageurs do not have direct access to the realizations of alphas and volatilities; instead, they must infer them from observable data.

Incorporating this constraint, we derive upper bounds on the achievable utility and Sharpe ratio for any feasible arbitrage portfolio, given a particular prior distribution. These upper bounds are strictly dominated by the infeasible optimal values that arbitrageurs could attain if they had perfect knowledge of the realizations of alphas and volatilities. Moreover, for each

[1] Throughout, we adopt a Bayesian perspective, treating (high-dimensional) parameters as realizations of random variables. Nuisance (low-dimensional) parameters are treated as constants, that is, as having degenerate (singleton) distributions.

prior distribution considered, we construct a feasible strategy that achieves these bounds.

Moreover, the difficulty of the arbitrageurs' learning problem depends primarily on the prior distribution of alphas, rather than on volatilities, as the latter are much easier to learn in our setting. We use simple special cases to illustrate how the Sharpe ratio bound on feasible portfolios varies with the strength and sparsity of alphas. When alphas are strong and not too rare relative to the dimensionality of the cross-section and the sample size, arbitrageurs can, in the limit, perfectly learn alphas. However, when alphas are weaker and more rare, inferring about them becomes more challenging, creating a gap between the feasible Sharpe ratio bound and the infeasible Sharpe ratio that requires perfect knowledge of alphas. For instance, the infeasible Sharpe ratio may explode asymptotically, while the feasible Sharpe ratio bound remains finite.

The optimal strategy leverages posterior inference, which in turn depends on the prior distribution of alphas. Under different priors, approaches such as multiple testing, shrinkage, and selection correspond to the optimal strategy for their respective prior assumptions. These alternative approaches are widely used by empirical asset pricing researchers and practitioners, making it valuable to understand the prior scenarios under which they achieve the optimal feasible strategy. A multiple-testing procedure as in Benjamini and Hochberg (1995) (BH) seeks to prevent false alpha discoveries by applying a p-value threshold that controls the false discovery rate (FDR), setting alphas that do not pass this threshold to zero. The BH procedure achieves optimal performance only when a small number of true alpha signals are strong; otherwise, it tends to be overly conservative, falling short of optimality. In contrast, Ridge shrinkage can achieve optimality when almost all true alphas are either uniformly strong or uniformly weak. Lasso-based selection and shrinkage of alphas aims to balance the strengths of these two approaches, with a tuning parameter that can adapt to different scenarios.

Yet none of these methods attain the upper bounds on utility and Sharpe ratio across all possible prior distributions. These strategies are only optimal if the prior implied by their regularizations matches the true prior, about which arbitrageurs generally lack precise knowledge. To address this challenge, we seek a uniformly optimal strategy—one that outperforms others in terms of utility across all plausible priors. In other words, when arbitrageurs face ambiguity about the true prior, striving for uniformity offers the most robust solution. The aforementioned strategies are feasible but not uniformly optimal.

In fact, we demonstrate that a uniformly optimal feasible strategy does exist in our setting. Specifically, arbitrageurs can construct a feasible trading strategy that achieves the aforementioned upper bounds uniformly across all prior distributions of alphas, regardless of their strength and sparsity. Under this strategy, portfolio weights are determined by the relative magnitudes and associated uncertainty of the alpha estimates, following Bayes' rule. Assets with high alpha t-statistics receive portfolio weights proportional to their t-statistics, while those with weaker estimates are assigned appropriately shrunk weights. As alluded to previously, the prior of alphas, unknown to arbitrageurs, plays a key role, because it directly impacts the portfolio weights via the Bayes' rule. Borrowing insights from the empirical Bayes method, our strategy fully recovers the impact of the prior from the empirical distribution of alpha estimates, ensuring its optimality.

The uniformly optimal feasible utility and Sharpe ratio precisely characterize the magnitude of the statistical limit to arbitrage. The gap between them and their infeasible counterparts implies that the amount of mispricing that can survive in equilibrium is more sizable compared to a scenario where arbitrageurs have exact knowledge of alphas and volatilities. Empirically, it is the feasible Sharpe ratio—not the unattainable, infeasible one—that reveals the minimum reward-to-risk compensation arbitrageurs require.

To empirically contrast feasible and infeasible Sharpe ratios, we propose an estimator for the infeasible Sharpe ratio. While this Sharpe ratio can be consistently estimated, it is not attainable by any feasible portfolio with weights derived from historical data. The infeasible Sharpe ratio often serves as the foundation for tests of the APT, see, e.g., Gibbons et al. (1989), Gagliardini et al. (2016), Fan et al. (2015), and Pesaran and Yamagata (2017). While these tests are powerful and can lead to the discovery of alpha signals, they are not relevant for arbitrageurs limited to feasible trading strategies. Our effort in constructing the uniformly optimal feasible arbitrage portfolio and evaluating its economic performance directly responds to Shanken's call to empirically assess the APT by "characterizing the investment opportunities that are available as a consequence of the observed expected return deviation" and "examining the extent to which we can approximate an arbitrage with existing assets" (Shanken, 1992). Here we do so taking into account the statistical limits to arbitrage.

Finally, we demonstrate the empirical implications of the statistical limits of arbitrage by examining monthly equity returns in US stock market from 1965 to 2020. Our empirical analysis has two parts. In the first part, we examine individual stock returns. We construct a multifactor model that uses observed stock characteristics as risk exposures. The characteristics include 16 commonly used return predictors and 11 industry indicators. These characteristics and industry dummies capture similar equity factors in the MSCI Barra model widely-used among practitioners. In the second part, we analyze returns on 1,273 characteristics-sorted portfolios and 49 industry portfolios, for which we construct latent factor models via singular value decomposition, following Giglio and Xiu (2021). Data on characteristics and portfolio returns in our empirical analysis is from Chen and Zimmermann (2020).[2]

[2] See openassetpricing.com (Stock-level Signal Datasets August 2023 Release).

For individual stocks, we estimate alphas as the averaged residuals in cross-sectional regressions of stock returns on characteristics in 60-month rolling windows. The cross-sectional R^2s are relatively low, averaging around 8% over our sample period. This suggests that learning alphas from realized returns is difficult, as the alphas are obscured by large amounts of noise. Accordingly, only 7.58% and 1.12% of individual stocks' alpha estimates have t-statistics greater than 2.0 and 3.0, respectively, in absolute values. Even without accounting for multiple testing, these estimates indicate that non-zero alphas are both rare and weak.

One may be concerned that these results may overstate the difficulty of learning alphas because they use only realized returns as the only signal of alpha. Our analysis of portfolio returns addresses these concerns. The large number of characteristics used to construct these portfolios can be interpreted as signals of alphas, and the portfolios as managed investment strategies that exploit the information in these signals. Although a latent factor model accounts for a larger portion (around 35%) of the cross-sectional variation in realized portfolio returns than the factor model for individual stock returns, we find that portfolio alphas still exhibit notable rarity and weakness. This is particularly evident once we adjust for publication years, ensuring that portfolios are considered only after their sorting characteristics were published in the previous years.

We then find, across both individual stocks and portfolios and using various methods, that the optimal feasible arbitrage portfolios yield moderately low annualized Sharpe ratios below 0.7. In contrast, the infeasible Sharpe ratios are considerably higher, averaging more than 4.8 and reaching as high as 16 in some sample periods for individual stocks, and ranging from 5 to 20 for portfolios. The large gap between feasible and infeasible Sharpe ratios highlights the empirical significance of the statistical limits to arbitrage. When these statistical limits to arbitrage are taken into account, the empirical facts are in line with the implication of the APT that near-arbitrage opportunities should be absent.

In macroeconomics and finance, rational expectations models—where decision makers are not confronted with parameter uncertainty and prior ambiguity—have attracted criticism (Hansen, 2007, 2014). Hansen (2014) emphasizes the importance of both types of uncertainty. He notes, "To confront model ambiguity, we may assign subjective probabilities across models, including the unknown parameters." Given a prior, much of the literature has adopted Bayesian learning as a way to study how decision makers act in the presence of parameter uncertainty. See, for example, Pastor and Veronesi (2009) for a survey on learning in financial markets. In many settings—particularly when learning a low-dimensional parameter, as in Collin-Dufresne et al. (2016)—learning can be sufficiently slow for its effects to persist within empirically realistic sample sizes, even though convergence to rational expectations takes place in the long run. An exception is Martin and Nagel (2022), where learning effects persist because investors face a high-dimensional inference problem regarding the process generating firm cash flows. Similarly, arbitrageurs in our model attempt to infer a high-dimensional vector with a potentially insufficient sample size, though they focus on learning about alphas rather than the underlying cash flow processes of firms. Our results therefore highlight the statistical limits of arbitrageurs' ability to counteract mispricing that could arise from the asset demand of less sophisticated or liquidity-driven investors.

Hansen (2014) further notes that "there are multiple reasons to consider a family of priors...This family...could also capture the ambiguity to a single decision maker struggling with which prior should be used." Concerns about prior ambiguity are formalized in earlier work as well. For instance, agents in Gilboa and Schmeidler (1989) and Maccheroni et al. (2006) address prior ambiguity by ranking decision rules according to a max-min utility criterion, where the minimization is over the set of potential priors. Equivalently, they choose a decision rule that performs best under the "worst-case" prior. In contrast, our arbitrageurs are able to achieve utility and Sharpe ratio outcomes that match those attainable by an agent who knows the actual prior, uniformly across all potential priors.[3] They accomplish this by leveraging the empirical Bayes method to learn about the prior from observable data. Despite this virtually perfect learning about the prior, however, the outcomes attainable by agents employing the best machine learning methods in a high-dimensional environment still fall short of those achievable by rational expectations agents in the asymptotic limit.

Our paper builds on a large literature on the arbitrage pricing theory (APT) developed by Ross (1976) and later refined by Huberman (1982), Chamberlain and Rothschild (1983), and Ingersoll (1984). As in these foundational studies, we employ asymptotic arguments that avoid imposing strong structural assumptions about the economy. The results of this analysis should be interpreted as an asymptotic approximation for a more realistic setting with a finite number of assets. The key point is that weak economic restrictions rule out Sharpe ratios far above the Sharpe ratios of diversified factor portfolios. Different from earlier work on the APT, we show in this paper that the statistical limit to arbitrage dramatically widens this Sharpe ratio bound compared with an economy in which arbitrageurs are endowed with perfect knowledge of investment opportunities. In this regard, our paper is also related to the literature on the limits of arbitrage reviewed in Gromb and Vayanos (2010). Complementary to this literature, the arbitrage limit in our setting stems from model ambiguity, rather than from risk, costs, frictions, and other constraints faced by arbitrageurs.

Our paper provides a solution to a long standing problem in optimal portfolio choice under parameter uncertainty. Historically, the plug-in mean-variance portfolio, which is based on

[3] In this regard, our analysis is related to a substantial literature in econometrics and statistics that discusses the uniform validity of asymptotic approximations; see, e.g., Staiger and Stock (1997), Imbens and Manski (2004), Leeb and Pötscher (2005), Andrews et al. (2020).

sample means and covariance matrices, has been criticized for underperforming. Assuming normally distributed returns, Kan and Zhou (2007) study the expected performance of this portfolio and observe that its Sharpe ratio is below its theoretical value. Further studies by Tu and Zhou (2010) and Kan et al. (2022) expand their analysis by exploring alternative portfolio rules that aim to minimize utility loss under estimation uncertainty. However, these studies only identify optimal solutions within a predetermined class of strategies, which does not reflect the flexibility investors have in practice to choose any perceived superior strategy. In contrast, our study identifies a uniformly optimal solution that achieves optimal feasible Sharpe ratios without imposing restrictions on the choice of strategies. Our analytical framework is primarily concerned with arbitrage portfolios, but our results are directly applicable to optimal portfolios in general, thus providing a solution to this classic problem.

A closely related paper to ours is Kim et al. (2020), which proposes a characteristics-based factor model to construct feasible arbitrage portfolios. Their asymptotic theory does not preclude arbitrage opportunities with a theoretically infinite Sharpe ratio, which implies a rather strong signal-to-noise ratio in their alpha signals. Our setting is considerably different in that the premise of our framework rules out infinite feasible Sharpe ratios, which enforces weak and rare signals. In our setting, alphas cannot possibly be recovered with certainty even when the sample size is large.[4]

Our paper also contributes to the evolving literature on applications of statistical and machine learning in asset pricing, and in particular on the topic of testing the APT, e.g., Gibbons et al. (1989), Gagliardini et al. (2016), and Fan et al. (2015), as well as on testing for alphas, e.g., Barras et al. (2010), Harvey and Liu (2020), and Giglio et al. (2021). The first literature focuses on testing a null that all alphas are equal to zero. This is certainly an interesting null hypothesis. However, as we emphasize in this paper, an economically sensible interpretation of the APT should allow for statistical limits to arbitrage. This means that the APT does allow for non-zero alphas as long as they do not induce an explosive feasible Sharpe ratio. The second literature focuses on detecting strong alphas, applying multiple testing methods, such as the BH method, or extensions thereof, to control the FDR. In contrast, we allow for rare and weak alpha signals such that any procedure aiming to control the FDR is too conservative with too few or no discoveries.[5] Our objective here is not on model testing or signal detection. Rather, we strive for the optimal economic performance of

[4] On the empirical side, Guijarro-Ordonez et al. (2022) propose a deep learning approach to statistical arbitrage that achieves a sizable out-of-sample Sharpe ratio. The profits of their trading strategy stem from generalized return reversals at daily to weekly frequencies, potentially due to liquidity provision and other microstructure channels. Our empirical analysis is not targeted towards characterizing the reward-to-risk ratios for high frequency traders, nor for traders that turnover a large portion of their portfolios daily.

[5] Donoho and Jin (2004) adopt the so-called higher criticism approach, dating back to Tukey (1976), to detect rare and weak signals in a stylized multiple testing problem.

arbitrage portfolios. We show that even if signals were so weak that they are undetectable by multiple testing methods, they may lead to a portfolio with a considerable Sharpe ratio.

The literature on portfolio choice through Bayesian and economic decision-theoretical perspectives is extensive, as outlined in the survey by Avramov and Zhou (2010). Early works like those by Jorion (1986) and Frost and Savarino (1986) highlight the advantages of using Bayes-Stein shrinkage solutions. Further contributions in this topic incorporate informative and economically motivated priors, as in Black and Litterman (1992), Pastor (2000), and Pastor and Stambaugh (2000). In our context, the optimal portfolio weights are proportional to the posterior mean of alpha. This resembles the classical normal mean problem in empirical Bayes, dating back to Robbins (1956), where the unknown parameters (alphas) are regarded as random draws from some common distribution, and only a noisy version (realized ex-factor returns) is observed. Our nonparametric approach thereby shares the same spirit of nonparametric empirical Bayes, see, e.g., Johns (1957), Zhang (1997), and Brown and Greenshtein (2009). Yet, our focus is on utility and Sharpe ratios rather than the mean-squared error of parameter estimation, as in classical empirical Bayes inference. The latter is unsuitable in our context also because, under the economically relevant scenario where the feasible optimal Sharpe ratio diminishes relative to the infeasible one, the mean-squared error of the optimal alpha estimator asymptotically converges to that of a naive zero estimator, rendering the mean-squared error criterion uninformative.

Our paper proceeds as follows. Section 2 develops our main result on the statistical limit to arbitrage. Specifically, Section 2.1 sets up the model, Section 2.2 motivates and then defines the feasibility constraint facing arbitrageurs, Sections 2.3 - 2.5 specify arbitrageurs' decision problem, derive the optimal strategy, illustrate the Bayes correction for alpha, and demonstrate the gap between feasible and infeasible Sharpe ratios, Section 2.6 constructs a uniformly optimal and feasible trading strategy that achieves the bound, Section 2.7 proposes an estimator of the infeasible Sharpe ratio, and finally Section 2.8 analyzes alternative strategies. Section 3 provides simulation evidence, followed by an empirical analysis in Section 4. Section 5 concludes. The appendix provides additional empirical and theoretical results, as well as all the technical details.

2 Main Theoretical Results

We start by revisiting the arbitrage pricing theory framework developed by Ross (1976). This theory is primarily based on a reduced-form statistical model for asset returns, which, despite its stylized nature, offers significant theoretical insights and remains relevant for guiding empirical investment decisions.

2.1 Factor Model Setup

The factor economy has N assets in the investment universe. The $N \times 1$ vector of excess returns r_t follows a reduced-form linear factor model:

$$r_t = \alpha + \beta\gamma + \beta v_t + u_t, \tag{1}$$

where β is an $N \times K$ matrix of factor exposures, α is an $N \times 1$ vector of pricing errors, v_t is a $K \times 1$ vector of zero-mean factor innovations with covariance matrix Σ_v, γ is a $K \times 1$ vector of risk premia, and u_t is a vector of idiosyncratic returns which, conditional on all the variables before time-t, has a diagonal covariance matrix Σ_u and zero mean.[6]

To facilitate our asymptotic analysis along the cross-sectional dimension, N, we regard high dimensional objects such as α, β, and Σ_u as random variables drawn from some cross-sectional distributions, whereas γ and Σ_v are regarded as deterministic constants. This distinction allows us to apply large-sample approximations to analyze the former, whereas the latter function as nuisance parameters in our analysis. We assume that α has zero mean, and is cross-sectionally independent of β.[7] These conditions are essential for identification of γ in a model that allows for pricing errors. We formalize these conditions later.

A more general version of this model allows for time-varying parameters:

$$r_t = \alpha_{t-1} + \beta_{t-1}\gamma_{t-1} + \beta_{t-1}v_t + u_t, \tag{2}$$

where β_t is a vector of time-varying factor loadings and γ_t is a vector of time-varying risk premia.[8] Despite its greater generality, this specification does not materially change the economic insights regarding the theoretical limits of arbitrage. For clarity, our theoretical analysis focuses on the more stylized unconditional model in (1), although our results, including Theorem 1, continue to hold under (2) with time-varying β_t.

There are at least three variations of (1), depending on what is assumed to be observable to arbitrageurs. The most common setup in the academic finance literature imposes that

[6] While approximate factor models become more prevalent following Chamberlain and Rothschild (1983), allowing for off-diagonal entries in the covariance matrix Σ_u would introduce additional statistical obstacles due to the estimation of large covariance matrix for inference on alpha and for building optimal portfolios. For simplicity, we illustrate the economic consequences of statistical limits to arbitrage using a strict factor model. We discuss violations of these model assumptions later.

[7] The zero-mean assumption on α is not restrictive, in that we can impose the first column of β as a vector of ones, with the first entry in v_t set to zero, so that the first entry in γ represents the zero-beta rate or the cross-sectional mean of α.

[8] This model is overly parametrized that parameters are not identifiable without additional restrictions. Some examples of parsimonious conditional factor models include Connor et al. (2012), Gagliardini et al. (2016), and Kelly et al. (2019).

factors are observable as, e.g., in Fama and French (1993).[9] The second setting, which has gained more popularity recently since its debut in Connor and Korajczyk (1986), assumes that factors are latent. The third setting is the MSCI Barra model originally proposed by Rosenberg (1974), where factor exposures are equal to stock characteristics and hence observable.[10] One notable advantage of this model is that it sidesteps the statistically and computationally demanding task of estimating a large number of potentially time-varying stock-level factor exposures. Since the MSCI Barra model is arguably the most prevalent approach among practitioners, we assume that β is observable to arbitrageurs and, in our empirical work, use this approach in analyzing individual stock returns. Specifically, we follow the conditional model in (2), making use of time-varying observed β's. We also employ a moving window method to estimate α, which effectively accounts for slow-varying α's.

In practice, arbitrageurs may make use of firm-level characteristics to predict individual stock's alphas. Our framework can in fact easily accommodate this if we assume that (2) holds with $\alpha_t = c_t\alpha$ and $\beta_t = c_t\beta$ for certain characteristics c_t that represent arbitrageurs' signals, then projecting returns onto lagged characteristics results in:

$$(c_{t-1}^\intercal c_{t-1})^{-1}c_{t-1}^\intercal r_t = \alpha + \beta(\gamma_{t-1} + v_t) + (c_{t-1}^\intercal c_{t-1})^{-1}c_{t-1}^\intercal u_t. \tag{3}$$

This transformation converts a conditional model for individual stocks with time-varying alphas into an unconditional latent factor model for managed portfolios, $(c_{t-1}^\intercal c_{t-1})^{-1}c_{t-1}^\intercal r_t$, with time-invariant alphas as in (1). Thus, via these managed portfolios, our theoretical framework also applies to a setting in which arbitrageurs use signals to predict time-varying alphas. Empirically, we therefore also examine returns on portfolios formed on a large number of stock characteristics, in addition to the individual stock analysis.

2.2 Feasible Near-Arbitrage Opportunities

Building on the insights of Ross (1976), Huberman (1982), and Ingersoll (1984), the concept of near-arbitrage can be formalized as follows. We adopt the subsequence framework used in Ingersoll (1984), where the subsequence typically depends on the number of investment opportunities, denoted by N. For each N, all random variables are defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_s\}_{s\leq t}, \mathrm{P})$, where $\mathcal{F}_t$ contains the information generated by $\alpha, \beta, \{r_s, v_s, u_s\}_{s\leq t}$. The probability measure P represents the data-generating process (DGP) and may vary with

[9] This differs from saying that the factor innovations in v_t are observable. In the observable factors setting, we typically write $f_t = \mu + v_t$, where μ is a vector of the population means of f_t; importantly, μ does not necessarily equal the factor risk premia γ. Since μ is unknown, v_t remains unobservable.

[10] See Kozak and Nagel (2023) on the conditions on factor construction and return covariance matrix that must hold for characteristics to equal factor loadings.

N. Accordingly, we consider a sequence of DGPs $\{\mathrm{P}_N\}_{N\geq 1}$, and the expectation and variance operators, each defined with respect to P_N. For simplicity and without loss of clarity, we will omit explicit reference to N and simply write P, $\mathrm{E}(\cdot)$, and $\mathrm{Var}(\cdot)$.

Definition 1. *A portfolio strategy w at time t is said to generate a near-arbitrage under a sequence of DGPs*[11]*, if it satisfies $w \in \mathcal{F}_t$ and there exists a constant $\delta > 0$ such that, for any $\epsilon > 0$, the following holds with probability approaching one:*

$$\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t) \leq \epsilon, \quad \mathrm{E}(w^\intercal r_{t+1}|\mathcal{F}_t) \geq \delta > 0.$$

Intuitively, no near-arbitrage means there exists no sequence of portfolios that earn positive expected returns with vanishing risks. Ingersoll (1984) establishes that a sufficient and necessary condition for the absence of near-arbitrage is that[12]

$$S^\star := \sqrt{\alpha^\intercal \Sigma_u^{-1} \alpha} \lesssim_{\mathrm{P}} 1. \tag{4}$$

Here, $S^\star$ is the theoretically optimal Sharpe ratio arbitrageurs can achieve in this economy using a portfolio strategy that has zero exposure to factor risks, namely, a "statistical arbitrage" strategy in the jargon of practitioners. This result suggests that moderate mispricing in the form of nonzero alphas is permitted in an economy without near-arbitrage opportunities, but there cannot be too many large alphas that would make $S^\star$ explode.[13]

To achieve this optimal Sharpe ratio, arbitrageurs should hold a factor-neutral portfolio with weights given by $w^\star \propto \Sigma_u^{-1}\alpha$, according to Ingersoll (1984).[14] His analysis assumes that arbitrageurs know the true (population) parameters: α and Σ_u. In reality, however, the true parameters are unknown to arbitrageurs as they only have a finite sample of data to learn about these parameters. The consequences of such parameter uncertainty can be minor in some contexts if the true parameters are revealed asymptotically and model predictions converge to those of a rational expectations model in which agents are endowed with perfect

[11] We omit the dependence of w on N and t.

[12] We use the notation $a_N \lesssim b_N$ to denote $a_N = O(b_N)$. Similarly, we write $a_N \lesssim_{\mathrm{P}} b_N$ to denote $a_N = O_{\mathrm{P}}(b_N)$. We use $a_N \asymp b_N$ if $a_N \lesssim b_N$ and $b_N \lesssim a_N$, and define $a_N \asymp_{\mathrm{P}} b_N$ analogously. The subscript N is omitted whenever there is no ambiguity.

[13] Assuming α_i is *i.i.d.* and $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$, by equation (4), we have $\alpha^\intercal\alpha \lesssim_{\mathrm{P}} \alpha^\intercal \lambda_{\min}(\Sigma_u^{-1})\alpha \lesssim_{\mathrm{P}} \alpha^\intercal \Sigma_u^{-1}\alpha \lesssim_{\mathrm{P}} 1$, so that $\mathrm{E}(\alpha_i^2) = o(1)$ by the law of large numbers.

[14] In Ingersoll (1984), α is defined to be the cross-sectional projection of the expected returns onto β in the population model, such that $\alpha^\intercal \Sigma_u^{-1}\beta = 0$. As a result, his arbitrage portfolio weights are factor neutral, i.e., $w^{\star\intercal}\beta = 0$ by construction. In contrast, our paper sets forth a predetermined DGP as specified in (1), where α is modeled as a random variable with the property $\mathrm{E}(\alpha^\intercal\beta) = 0$. Consequently, the factor-neutral optimal strategy illustrated by (9) does not exactly align with the formula presented by Ingersoll (1984). Under our DGP assumptions, Ingersoll's portfolio weights are not strictly factor-neutral. Nonetheless, we can show that our optimal portfolio weights remain factor-neutral, and achieve the same Sharpe ratio $S^\star$ asymptotically as N increases.

knowledge of parameters. Fundamentally, this requires that the learning problem in the limiting experiment becomes increasingly simpler as the sample size increases.

We assume at any given time t, arbitrageurs examine a sample of size T, derived from (1), spanning from $t-T+1$ to t. Throughout we will consider asymptotic limits as N and T increase while K and t are fixed. In this context, the difficulty of the learning problem also hinges on the number of investment opportunities, N, besides T. As N increases, for a given sample size T, it becomes increasingly difficult for arbitrageurs to determine which assets truly have nonzero alphas. If the learning problem remains difficult as N and T increase, the learning effect persists, which could lead to limiting implications that differ from the rational expectations case. Within this framework, the rational expectations limit $S^\star$ is only relevant for rather restrictive scenarios. By contrast, in more realistic cases—such as when N greatly exceeds T—the optimal Sharpe ratio arbitrageurs can achieve without factor exposures is far smaller than $S^\star$, due to their inability to make error-free inference. Therefore, condition (4) may be excessively restrictive in such scenarios.

To illustrate this intuition, we consider a simple and specific example.

Example 1. *Suppose the cross-section of alphas is drawn from the following distribution:*

$$\alpha_i \overset{i.i.d.}{\sim} \begin{cases} \mu & \textit{with prob. } \rho/2 \\ -\mu & \textit{with prob. } \rho/2 \\ 0 & \textit{with prob. } 1-\rho \end{cases}, \quad 1 \le i \le N, \tag{5}$$

where $\mu \ge 0$ and $0 \le \rho \le 1$, and they potentially vary with N and T. In addition, we also assume $\beta = 0$, $u \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_N)$ for some $\sigma > 0$, and u is independent of α. Consequently, equation (1) *becomes $r_t = \alpha + u_t$.*

In this example, μ dictates the strength of alphas, ρ describes how rare alphas are, whereas σ is a nuisance parameter. To emphasize the role of signal strength and count, we impose in this example that all assets share the same alpha distribution and the same idiosyncratic variance. Now suppose, more specifically, that the magnitude of (μ, ρ) satisfies

$$\mu \precsim T^{-1/2} \quad \text{and} \quad \rho \precsim N^{-1/2}. \tag{6}$$

By modeling parameters μ and ρ as functions of the sample size and dimension of the investment set, we introduce greater flexibility in depicting the challenges arbitrageurs encounter in finite sample situations.[15] This condition (6) implies that the signal strength μ vanishes

[15] Adopting a drifting sequence for parameters is a common trick in econometrics to provide more accurate finite sample approximations. As Bekker (1994) put, "in evaluating the results, it is important to keep in mind

as the sample size increases ($T \to \infty$) and the signal percentage count ρ decays as the investment universe expands ($N \to \infty$). This setup is intended to approximate a situation in which only a small subset of assets have nonzero, and typically small, alpha. The objective is to characterize a setting in which alphas can survive in the presence of arbitrageurs. On the other hand, σ is assumed to be fixed, since arbitrage activity does not affect the magnitude of idiosyncratic risks.

This model resides in an uncommon territory in the existing literature of asset pricing: weak and rare alphas. In fact, the classical no-near-arbitrage condition (4) imposes, implicitly, weakness or rareness on alphas; otherwise, if alphas were strong and dense, $\alpha^\intercal\alpha$ would explode rather rapidly. Even in the current setting, in light of the fact that $\mathrm{E}(\alpha^\intercal\alpha) = \rho\mu^2 N$, we still have $\alpha^\intercal\alpha \xrightarrow{\mathrm{P}} \infty$ as long as $N^{1/2}/T \to \infty$. In other words, a near-arbitrage opportunity arises according to (4), with a strategy $w = \sigma^{-2}\alpha$.

However, statistical uncertainty prevents arbitrageurs from having this "free lunch." In general, it is only possible to recover any element of alpha up to some estimation error of magnitude $T^{-1/2}$.[16] Since, by design, the true alpha is of the same order of magnitude as its level of statistical uncertainty, i.e., $\mu \approx T^{-1/2}$, it is impossible for arbitrageurs to determine precisely which assets among all have nonzero alpha.

For illustration purpose, suppose that arbitrageurs adopt the strategy $\widehat{w} = \sigma^{-2}\widehat{\alpha}$,[17] replacing α in w with $\widehat{\alpha} = \bar{r} = \alpha + \bar{u}$.[18] Out of sample, this portfolio's conditional expected return and conditional variance can be written as:

$$\mathrm{E}\left(\sigma^{-2}\left(\alpha + \bar{u}\right)^\intercal\left(\alpha + u_{t+1}\right)|\mathcal{F}_t\right) = \sigma^{-2}(\alpha^\intercal\alpha + \bar{u}^\intercal\alpha),$$
$$\mathrm{Var}\left(\sigma^{-2}\left(\alpha + \bar{u}\right)^\intercal\left(\alpha + u_{t+1}\right)|\mathcal{F}_t\right) = \sigma^{-2}(\alpha^\intercal\alpha + 2\alpha^\intercal\bar{u} + \bar{u}^\intercal\bar{u}),$$

where u_{t+1} denotes a future de-meaned return at $t+1$, that shares the same distribution

that the parameter sequence is designed to make the asymptotic distribution fit the finite sample distribution better. It is completely irrelevant whether or not further sampling will lead to samples conforming to this sequence or not."

[16] Giglio et al. (2021) develop the asymptotic normality result for alpha estimates via a Fama-MacBeth procedure in various scenarios, in which factors are (partially) observable or latent whereas β is unknown. The CLTs in these scenarios share the same form: for any $1 \le i \le N$,

$$\sqrt{T}(\widehat{\alpha}_i - \alpha_i) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(1 + \gamma^\intercal(\Sigma_v)^{-1}\gamma)), \tag{7}$$

where σ_i^2 is the ith entry of Σ_u. In the case that β is observable (but factors are not), we can show that the CLT has a similar form except that the scalar $(1 + \gamma^\intercal(\Sigma_v)^{-1}\gamma)$ disappears.

[17] The knowledge of σ is ultimately inconsequential for our purposes, as we will demonstrate subsequently for a more general setting. Despite σ being known, this strategy fails to yield any positive Sharpe ratio.

[18] For any time series of random vector a_t, we use $\bar{a}$ to denote its sample average. As we will point out later in the paper, this strategy $\widehat{w}$, which we will denote by $\widehat{w}^{\mathrm{CSR}}$, fails to achieve the optimal Sharpe ratio in all scenarios. We will discuss this in detail in Section 2.6.

as $\{u_s\}_{s\le t}$, but is independent of $\bar{u}$ which belongs to the information set up to t, $\mathcal{F}_t$. The resulting squared conditional Sharpe ratio is given by:

$$S^2 = \frac{\sigma^{-4}(\alpha^\intercal\alpha + \bar{u}^\intercal\alpha)^2}{\sigma^{-2}(\alpha^\intercal\alpha + 2\alpha^\intercal\bar{u} + \bar{u}^\intercal\bar{u})} \lesssim_{\mathrm{P}} T^{-1} \to 0, \tag{8}$$

where we use the fact that $\bar{u}^\intercal\bar{u} \eqsim_{\mathrm{P}} N/T$ when u_t is i.i.d.. In other words, this portfolio achieves a Sharpe ratio equal to zero asymptotically.

Is there a superior trading strategy capable of maintaining a non-vanishing Sharpe ratio? The straightforward answer is no. Our discussion below will elucidate that, within this context, namely, (6) holds true, the highest achievable Sharpe ratio for all *feasible* trading strategies employed by arbitrageurs, represented as S^{OPT}, vanishes asymptotically as $N, T \to \infty$. Conversely, the *infeasible* optimal Sharpe ratio, denoted $S^\star$, diverges under the condition that $N^{1/2}/T \to \infty$. There is a vast disparity, as shown by this example, between S^{OPT} and $S^\star$. The difference between feasible and infeasible strategies is primarily driven by the information set accessible to arbitrageurs when they implement their trading strategies. In this example, the infeasible strategy assumes that arbitrageurs have access to a comprehensive information set, $\mathcal{F}_t$, that encompasses the knowledge of α. This knowledge proves to be extremely valuable when α is both rare and weak, which creates significant gap between this strategy and others that lack access to such information.

To formalize this, we introduce the concept of feasibility:

Definition 2. *Let $\mathcal{G}$ denote the information set generated by the data observable to arbitrageurs. A strategy w is said to be feasible (or infeasible) if it is (or is not) $\mathcal{G}$-measurable.*

By modeling the unknown parameters as random variables that lie outside the information set $\mathcal{G}$, we explicitly acknowledge the presence of *parameter uncertainty*. Arbitrageurs must therefore rely on observed data within $\mathcal{G}$ to learn about these unknowns. Within a Bayesian framework, they update their priors using historical data, forming posterior beliefs that serve as the basis for inference.

However, just as parameters can take many possible values, arbitrageurs in practice also face uncertainty about which of the many possible priors the parameters are drawn from—that is, they face *prior ambiguity*. Prior ambiguity breaks the feasibility of strategies based on posterior beliefs, because the posterior—conditional on the same observed data—can still vary with the choice of prior. Since feasibility requires strategies to be $\mathcal{G}$-measurable—that is, entirely determined by observable quantities—the dependence of the posterior on an unobservable prior renders such strategies infeasible. To illustrate this point, consider Example 1, where the prior distribution of α is characterized by two unknown parameters: μ and ρ. In this setting, $\widehat{\alpha}$ serves as the sufficient statistic summarizing the information set

$\mathcal{G}$. Yet, the posterior mean of α_i is positive for all observed $\widehat{\alpha}_i > 0$, provided that both $\mu > 0$ and $\rho > 0$, but is zero conditioning on the same positive $\widehat{\alpha}_i$ when $\mu = 0$. The key issue is that the posterior depends not only on the observed data $\widehat{\alpha}$, but also on the values of the unobserved prior parameters μ and ρ. As a result, even when the observable input remains the same, the posterior belief about α_i can vary depending on the prior. Since μ and ρ are not part of $\mathcal{G}$, the posterior is not $\mathcal{G}$-measurable. Thus, in the presence of prior ambiguity, feasibility requirement generally excludes strategies that directly uses posterior beliefs.

Recognizing the distinction between feasible and infeasible strategies, we proceed to investigate impact of the feasibility constraint on arbitrageurs' behavior and the consequences thereof. This exploration necessitates defining the decision-making problem faced by arbitrageurs, which we turn to next.

2.3 Arbitrageurs' Decision Problem

At time t, arbitrageurs aim to maximize mean-variance utility by selecting a strategy. To ensure factor neutrality, they restrict attention to strategies orthogonal to the observed β. This setup follows the APT framework of Ross (1976), where arbitrageurs eliminate exposure to systematic factors, exploiting alphas while managing idiosyncratic risk.

The utility function is formulated as:

$$U(w) = \mathrm{E}\left(w^\intercal r_{t+1} \middle| \mathcal{F}_t\right) - \frac{\kappa}{2}\mathrm{Var}\left(w^\intercal r_{t+1} \middle| \mathcal{F}_t\right) = w^\intercal \alpha - \frac{\kappa}{2} w^\intercal \Sigma_u w,$$

where κ represents the degree of risk aversion. To focus on our main objective, we exclude transaction costs and dynamic considerations, reducing the problem to a static, single-period optimization. Accordingly, we suppress the time subscript t when possible.

Without the feasibility constraint, a unique strategy $w^\star$ maximizes $U(w)$:[19]

$$w^\star = \frac{1}{\kappa}\Sigma_u^{-1/2}\mathbb{M}_{\Sigma_u^{-1/2}\beta}\Sigma_u^{-1/2}\alpha, \tag{9}$$

since the optimization problem $\max_{w: w^\intercal \beta = 0} U(w)$ is quadratic with a linear constraint. Under rational expectations—where the parameters α and Σ_u are known and fixed constants—arbitrageurs face no obstacle in implementing $w^\star$.

In contrast, to confront parameter uncertainty—that is, the presence of other plausible values for α and Σ_u—we model these parameters as random variables that lie outside the arbitrageurs' information set $\mathcal{G}$. Yet, arbitrageurs are constrained to access only strategies within the set

[19] $\mathbb{M}_A = \mathbb{I} - A(A^\intercal A)^{-1}A^\intercal$ for any matrix A.

$$\mathcal{W} = \{w \in \mathcal{G} : w^\intercal \beta = 0\},$$

namely, the set of $\mathcal{G}$-measurable strategies that are orthogonal to β. This feasibility constraint limits arbitrageurs to capturing investment opportunities based solely on observable information, thereby excluding $w^\star$.

In this context, no feasible strategy w can almost surely maximize $U(w)$, because the arbitrageur's objective becomes a stochastic function of historical data rather than a deterministic function of α and Σ_u. Regardless of the distribution of the realized observed data, there is always a non-trivial—though small—probability (i.e., bad luck) that the first asset has a large alpha while all others have zero. In such cases, any sophisticated strategy that diversifies risk across assets will underperform a naive strategy that concentrates on the first asset. The possibility of such rare but extreme events implies that no feasible strategy can maximize utility almost surely. Instead, arbitrageurs must pursue a more conservative objective: a strategy is considered desirable if it achieves higher utility than alternatives with high probability. Formally, a strategy $w \in \mathcal{W}$ is said to outperform $w' \in \mathcal{W}$ if, for all small fixed $\epsilon > 0$ and with probability approaching one,

$$U(w') \leq_\epsilon U(w).^{20} \tag{10}$$

With this criterion, arbitrageurs seek a feasible strategy w that outperforms all other feasible strategies.

Just as arbitrageurs face parameter uncertainty, they also confront prior ambiguity—that is, uncertainty about which (prior) distribution correctly characterizes the law of these random variables (e.g., α and Σ_u). This ambiguity is critical because the probability of a strategy outperforming another is determined by a practically unknown prior specified in the DGP. In response, we consider arbitrageurs who seek strategies that outperform all alternatives under any admissible prior. To formalize this, we allow for a general class of priors over α, Σ_u (and other nuisance parameters), while assuming that the return model (1) is correctly specified.[21] We denote by $\mathbb{P}$ the set of all DGPs consistent with these priors—representing the full scope of DGPs that arbitrageurs may plausibly face.

Given this collection of possible DGPs, arbitrageurs seek uniformly optimal and feasible strategies—if such strategies exist. A strategy $w \in \mathcal{W}$ is said to be *uniformly optimal* if, for every alternative strategy $w' \in \mathcal{W}$ and for every fixed $\epsilon > 0$,

[20] We use $a \leq_\epsilon b$ to denote that $a \leq b + \epsilon + \epsilon|b|$. Similarly, we write $a =_\epsilon b$ if $a \leq_\epsilon b$ and $b \leq_\epsilon a$. The notation accommodates the case where b may be large, reducing the total error to around $\epsilon|b|$, which remains small relative to b itself.

[21] Hansen (2014) considers a third source of uncertainty—model misspecification—in addition to parameter uncertainty and prior ambiguity. We discuss this point briefly in the conclusion.

$$\inf_{\mathrm{P}\in\mathbb{P}} \mathrm{P}\Big(U(w') \leq_\epsilon U(w)\Big) \to 1.$$

That is, w asymptotically outperforms all other feasible strategies under every DGP in $\mathbb{P}$, up to a small slack ϵ.

In addition to utility, we also evaluate investment performance using the conditional Sharpe ratio, defined as

$$S(w) := \mathrm{E}(w^\intercal r_{t+1}|\mathcal{F}_t)/\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t)^{1/2}.$$ [22]

Thus far, we have fully delineated the objectives (uniform optimality) and the constraints (feasibility) faced by arbitrageurs in the presence of parameter uncertainty and prior ambiguity. The following subsections focus on developing the corresponding solutions.

2.4 Feasible Utility and Sharpe Ratio Bounds

We begin by establishing upper bounds on the utility and investment outcomes of all feasible strategies. These bounds are economically important, as they reveal a substantial gap between feasible and infeasible strategies. We will later show that these bounds are in fact sharp and can be achieved by a uniformly optimal strategy that we construct.

We impose the following conditions, which must be satisfied by every DGP in the collection $\mathbb{P}$ that arbitrageurs may potentially face:

Assumption 1. *We have (1) and the following:*

(a) (α_i, u_i) *is i.i.d. across* i,[23] *and satisfies* $\mathrm{E}(\alpha_i|(\Sigma_u)_{i,i}) = 0$ *and* $\mathrm{E}\|\alpha\|^2_{\mathrm{MAX}} = o(1)$.[24] *Moreover, it holds that* $1 \lesssim_{\mathrm{P}} \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$.

(b) The pricing errors α, *factors* v_t, *factor loadings* β, *and idiosyncratic errors* u_t *are, conditionally on* Σ_u, *mutually independent.*

Condition (a) suggests that our model focuses on weak alphas; as the number of assets, N, increases, their magnitudes diminish.[25] Moreover, this condition ensures that volatilities

[22] In the presence of prior ambiguity, we consider a collection, $\mathbb{P}$, of DGPs, each defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_s)_{s\leq t}, \mathrm{P})$. Expectations $\mathrm{E}(\cdot)$ and variances $\mathrm{Var}(\cdot)$ are taken with respect to the particular $\mathrm{P} \in \mathbb{P}$ and are therefore P-specific, though we suppress this dependence for notational simplicity. The same applies to $U(\cdot)$ and $S(\cdot)$, both of which depend on $\mathrm{E}(\cdot)$ and $\mathrm{Var}(\cdot)$.

[23] The i.i.d. assumption on $u_{i,t}$ does not imply that Σ_u is a scalar multiple of the identity matrix; rather, Σ_u is a diagonal matrix with each diagonal entry following the same distribution.

[24] For a matrix A, we use $\|A\|$ and $\|A\|_{\mathrm{MAX}} = \max_{i,j} |a_{ij}|$ to denote the operator norm (or ℓ_2 norm) and the ℓ_∞ norm of A on the vector space. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of A.

[25] By Assumption 1(a), $\mathrm{Var}(\alpha_i) = \mathrm{E}(\alpha_i^2) = o(1)$. Referencing our earlier discussion (footnote 13), a diminishing variance in α is essential for precluding near-arbitrage opportunities within Ross' APT framework.

remain within upper and lower bounds. Condition (a) therefore implies that learning about alpha is a more arduous task than learning about volatilities. Condition (b) is imperative for the model's identification. For instance, assuming independence between α and β is key to identify the risk premia, γ, see, e.g., Giglio et al. (2022).

We now derive upper bounds on the utility and investment performance of feasible arbitrage strategies. To do so, we adopt an expanded information set $\mathcal{G}$, generated by $\{r_s, \beta, v_s, \Sigma_u\}_{s=t-T+1}^{t}$. While empirical applications typically rely only on past returns and betas—as discussed in Section 2.1—this broader information set serves two purposes. First, it isolates the central challenge: learning α accounts for most of the performance gap between infeasible and feasible strategies. Second, since any strategy feasible under a smaller information set remains feasible under a larger one, the bounds derived under $\mathcal{G}$ are valid—though potentially loose—for all feasible strategies based on subsets of $\mathcal{G}$.

In Section 2.6, we construct a feasible strategy using only observed returns and betas, estimating v_s and Σ_u to approach these bounds. This reinforces the conclusion that the main difficulty lies in learning α, rather than volatilities or factor exposures. We now present the formal results.

Theorem 1. *Suppose Assumption 1 holds and $\mathcal{G}$ is generated by $\{r_s, \beta, v_s, \Sigma_u\}_{s=t-T+1}^{t}$. We define $\widetilde{w} := \kappa^{-1}\mathbb{M}_\beta\Sigma_u^{-1}\widetilde{\alpha}$, where $\widetilde{\alpha} := \mathrm{E}(\alpha|\mathcal{G})$. Then, for any $w \in \mathcal{W}$ and all small fixed $\epsilon > 0$,*

$$\inf_{\mathrm{P}\in\mathbb{P}} \mathrm{P}\Big(U(w) \leq_\epsilon U(\widetilde{w}) =_\epsilon (2\kappa)^{-1}S(\mathcal{G})^2\Big) \to 1, \text{ and } \inf_{\mathrm{P}\in\mathbb{P}} \mathrm{P}\Big(S(w) \leq_\epsilon S(\widetilde{w}) =_\epsilon S(\mathcal{G})\Big) \to 1, \quad (11)$$

where $S(\mathcal{G}) := \sqrt{\widetilde{\alpha}^\intercal\Sigma_u^{-1}\widetilde{\alpha}}$.

The theorem proposes a strategy $\widetilde{w}$ that, under all DGPs admissible by Assumption 1, outperforms all feasible arbitrage strategies in terms of utility and, correspondingly, attains Sharpe ratios that are at least as large, up to negligible errors. Accordingly, the utility and Sharpe ratio achieved by $\widetilde{w}$ serve as upper bounds for what any feasible strategy can achieve.

By definition, the upper bound $S(\mathcal{G})$ satisfies

$$\mathrm{E}\big(S(\mathcal{G})^2\big) \leq \mathrm{E}\big(\alpha^\intercal\Sigma_u^{-1}\alpha\big), \quad (12)$$

with equality if and only if $\widetilde{\alpha} = \alpha$ almost surely. In this special case, $\widetilde{w}$—aside from the projection adjustment via $\mathbb{M}_\beta$—coincides with the optimal mean-variance allocation with perfect knowledge of α, and we have $w^\star \approx \widetilde{w}$, as $N \to \infty$. The right-hand side of (12) reflects the infeasible benchmark where arbitrageurs have perfect knowledge of α, as in equation (4). This inequality highlights the gap between feasible and infeasible Sharpe ratios, which in turn translates into a gap in the utilities they can achieve.

Notably, the strategy $\widetilde{w}$ relies on arbitrageurs' best prediction of alphas, $\mathrm{E}(\alpha|\mathcal{G})$, formed using their information set $\mathcal{G}$ and a given prior distribution P. This makes the strategy infeasible in practice: it depends on unknown quantities such as Σ_u, v_s, and γ (implicitly assumed constant within $\mathcal{G}$). Moreover, under prior ambiguity, P represents only one among many plausible DGPs in $\mathbb{P}$. Thus, Theorem 1 characterizes optimal behavior in an idealized setting where arbitrageurs face no uncertainty other than over α and know the correct prior. The remaining question is whether arbitrageurs can accommodate uncertainty of the remaining parameters and prior ambiguity at virtually no cost and still construct a feasible strategy that performs as well as $\widetilde{w}$ across all DGPs. We address this issue in the next two subsections. For clarity, we henceforth assume that $\widetilde{w}$, $\widetilde{\alpha}$, and $S(\mathcal{G})$ are defined with respect to the information set $\mathcal{G}$ specified in Theorem 1, unless otherwise noted.

The projection matrix $\mathbb{M}_\beta$ in $\widetilde{w}$ ensures that it is factor-neutral, as required in Theorem 1, since $(\mathbb{M}_\beta \Sigma_u^{-1}\widetilde{\alpha})^\intercal \beta = 0$. While Theorem 1 restricts attention to factor-neutral strategies to reflect arbitrage-based considerations, Proposition B1 in the appendix extends the analysis to general $\mathcal{G}$-measurable strategies. It shows that any such strategy w satisfies the bound:

$$S(w) \leq \left(S(\mathcal{G})^2 + \gamma^\intercal \Sigma_v^{-1} \gamma\right)^{1/2} + o_{\mathrm{P}}(1 + S(\mathcal{G})), \tag{13}$$

where $\gamma^\intercal \Sigma_v^{-1}\gamma$ is the squared optimal Sharpe ratio earned from the factor portfolios.

2.5 Bayes Correction for Selection Bias

A central component of $\widetilde{w}$ is the posterior mean of α, $\widetilde{\alpha} = \mathrm{E}(\alpha \mid \mathcal{G})$. However, this expression is implicit and not directly implementable. We now focus on reducing the information set $\mathcal{G}$ to its sufficient statistics for α. This simplification is crucial for three reasons: it explains why $\widetilde{\alpha}$ achieves optimal utility and Sharpe ratios; it helps quantify the gap between feasible and infeasible strategies; and it informs the design of a uniformly optimal feasible strategy. To proceed, we introduce additional assumptions.

Assumption 2. *For each $N \geq 1$, it holds that*

(a) $u_{i,t} = \sigma_i \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ follows a standard normal distribution, and is i.i.d. across (i, t) and independent of Σ_u.

(b) $s_i := \alpha_i / \sigma_i$ is independent of σ_i.

Based on the return generating process described in Equation (1) and given the information set $\mathcal{G}$, Assumptions 1 and 2(a) together ensure that the summary statistics for α_i are the volatility σ_i and the sample average of the ex-factor returns, expressed as

$\breve{\alpha}_i := \bar{r}_i - \beta_i(\gamma + \bar{v}) = \alpha_i + \bar{u}_i$. In other words, $\mathrm{E}(\alpha_i|\mathcal{G}) = \mathrm{E}(\alpha_i|\breve{\alpha}_i, \sigma_i)$. Consequently, this assumption simplifies the conditional information set $\mathcal{G}$ in the posterior distribution of α to merely a two-dimensional vector comprising these conditioning variables. To evaluate this conditional expectation, it becomes necessary to assume a specific form of dependence between α_i and σ_i.

Assumption 2(b) further refines this by allowing the conditional expectation to be expressed as $\mathrm{E}(\alpha_i|\breve{\alpha}_i, \sigma_i) = \sigma_i \mathrm{E}(s_i|\breve{\alpha}_i, \sigma_i) = \sigma_i \mathrm{E}(s_i|\breve{\alpha}_i/\sigma_i)$.[26] This leads to $\widetilde{s}_i := \mathrm{E}(s_i|\mathcal{G}) = \mathrm{E}(s_i|\breve{s}_i)$, where $\breve{s}_i := \breve{\alpha}_i/\sigma_i$. Consequently, in terms of the scaled version of α, namely, s, the conditioning information set is now a single scalar variable, which simplifies the estimation problem later. Further, in light of Theorem 1, $\widetilde{\alpha}$, $\widetilde{w}$, and $S(\mathcal{G})^2$ can all be represented in relation to $\widetilde{s}$:[27]

$$\widetilde{\alpha} = \Sigma_u^{-1/2}\widetilde{s}, \quad \widetilde{w} \propto \mathbb{M}_\beta \Sigma_u^{-1/2}\widetilde{s}, \quad S(\mathcal{G})^2 = \widetilde{s}^\intercal \widetilde{s}. \tag{14}$$

Finally, note that

$$\breve{s}_i = s_i + \bar{\varepsilon}_i \sim \mathcal{N}(s_i, T^{-1}), \quad \text{conditional on } s_i.$$

This formulation casts the original posterior inference problem into the framework of the classical Gaussian sequence model for recovering a high-dimensional mean vector, s, from noisy observations $\breve{s}$, which has been extensively studied in the statistics literature (see, e.g., Robbins (1956), Efron (2011), and Efron (2019)). Although the assumption that ε_i follows a Gaussian distribution may seem restrictive, this framework is sufficiently versatile to accommodate a wide range of distributions for s_i.

Arbitrageurs face the challenge of identifying the true underlying signal, s, from an observed noisy signal $\breve{s}$. This task is complicated by what Efron (2011) described as selection bias or "the winner's curse": A high observed signal, $\breve{s}_i$, could reflect a high s_i, or it could be the result of "luck", with an unusually large noise realization $\bar{\varepsilon}_i$. Consequently, arbitrageurs must carefully adjust their investment strategies to mitigate this potential bias. The correction involves relying on the posterior of s, $\widetilde{s}$, which accounts for this bias. To see this, we present an explicit formula for $\widetilde{s}$:

Theorem 2. *Suppose Assumptions 1 and 2 hold and define $\psi(a) := \mathrm{E}\left(s_i|\breve{s}_i = a\right)$. Then we have $\widetilde{\alpha}_i = \sigma_i \widetilde{s}_i$, where $\widetilde{s}_i = \psi(\breve{s}_i)$. Moreover, it holds that*

$$\psi(a) = a + \frac{1}{T}\frac{d}{da}\log p(a), \tag{15}$$

[26] This equality relies on the result that conditional on $\widehat{\alpha}_i/\sigma_i$, α_i/σ_i is independent of σ_i. We impose this condition primarily for clarity of exposition and simplicity of Algorithm 1 below.

[27] Given that risk aversion does not impact the out-of-sample Sharpe ratio, we will use $\propto$ to substitute for κ^{-1} in our subsequent discussions on the portfolio strategy.

where $p(a) = \mathrm{E}\big(\phi_{1/T}(a - s_i)\big)$ is the probability distribution function of $\check{s}_i$.[28]

The preceding discussion directly leads to the first assertion of the theorem. Equation (15) is based on Tweedie's formula (Robbins, 1956), which links the posterior mean of s, given $\check{s} = a$, to the marginal distribution of $\check{s}$. The adjustment term, $T^{-1} d \log p(a)/da$, plays a critical role in correcting for selection bias in the observed signal $\check{s}_i$, and gives rise to several notable properties. An example of these properties, as shown by Andrews et al. (1972), is that $\psi(a)$ is a nondecreasing function of a. This property ensures that the relative magnitude of various signals is preserved post-correction. Moreover, the posterior mean $\widetilde{s}$ on average induces a shrinkage effect towards the prior mean, which, in our context, is zero.

The posterior shrinkage on s, or equivalently on α, consequently induces a "shrinkage" effect on the maximal Sharpe ratio and utility achievable by arbitrageurs represented by $S(\mathcal{G})$, as demonstrated by the inequality (12). Based on the results of Theorem 2, and under a technical condition concerning the tail behavior of s_i, we can obtain a more explicit formula for $S(\mathcal{G})$:

Corollary 1. *Suppose Assumptions 1 and 2, and the additional condition $\mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) = o(N^{-1})$ hold. We arrive at the conclusion that:*[29]

$$S(\mathcal{G}) = S^{\mathrm{OPT}} + o_{\mathrm{P}}(1), \quad \textit{with} \quad S^{\mathrm{OPT}} = \left(N \int \psi(a)^2 p(a) da \right)^{1/2}.$$

This result enables an explicit computation of $S(\mathcal{G})$ and links it directly to the marginal density of $\check{s}$, $p(\cdot)$. Since different DGPs (priors) imply different marginal densities, it clarifies how $S(\mathcal{G})$ depends on the prior and informs the feasibility of attaining non-vanishing Sharpe ratios under varying DGPs. The following corollary provides illustrative calculations based on Example 1, where each DGP is fully specified by parameters μ/σ and ρ.

Corollary 2. *Suppose that the same assumptions as in Corollary 1 hold. In addition, we assume alpha follows (5) as in Example 1. Then we have $S^\star = \sigma^{-1}\mu(\rho N)^{1/2} + o_{\mathrm{P}}(1)$. Further, assuming that $\sigma^{-1}\mu(\rho N)^{1/2}$ does not vanish, then it holds that $S^{\mathrm{OPT}} \leq (1-\epsilon)\sigma^{-1}\mu(\rho N)^{1/2}$ for some $\epsilon > 0$, if and only if*

$$T^{1/2}\mu/\sigma - \sqrt{-2\log\rho} \lesssim 1. \tag{16}$$

Corollary 2 suggests that when $T^{1/2}\mu/\sigma$ is large enough that the constraint (16) is violated, $S^\star \approx_{\mathrm{P}} S^{\mathrm{OPT}}$—that is, in the limit, learning does not play any role—arbitrageurs in this

[28] $\phi_x(a)$ is the distribution function for $\mathcal{N}(0, x)$.

[29] Throughout, c_N represent some sequence vanishing with N.

scenario achieve the same Sharpe ratio as under perfect knowledge of alphas. Furthermore, the rareness parameter ρ does not make much difference if $T^{1/2}\mu/\sigma$ gets sufficiently large. That said, if ρ approaches to zero so fast that $\sqrt{-2\log\rho}$ dominates $T^{1/2}\mu/\sigma$—that is, alpha is extremely rare and sufficiently weak—the learning problem becomes rather challenging and the Sharpe ratio attainable by arbitrageurs, S^{OPT}, is dominated, in the limit, by the infeasible Sharpe ratio $S^{\star}$.

To give a concrete example of Corollary 2, consider an alternative DGP assumption instead of (6):[30]

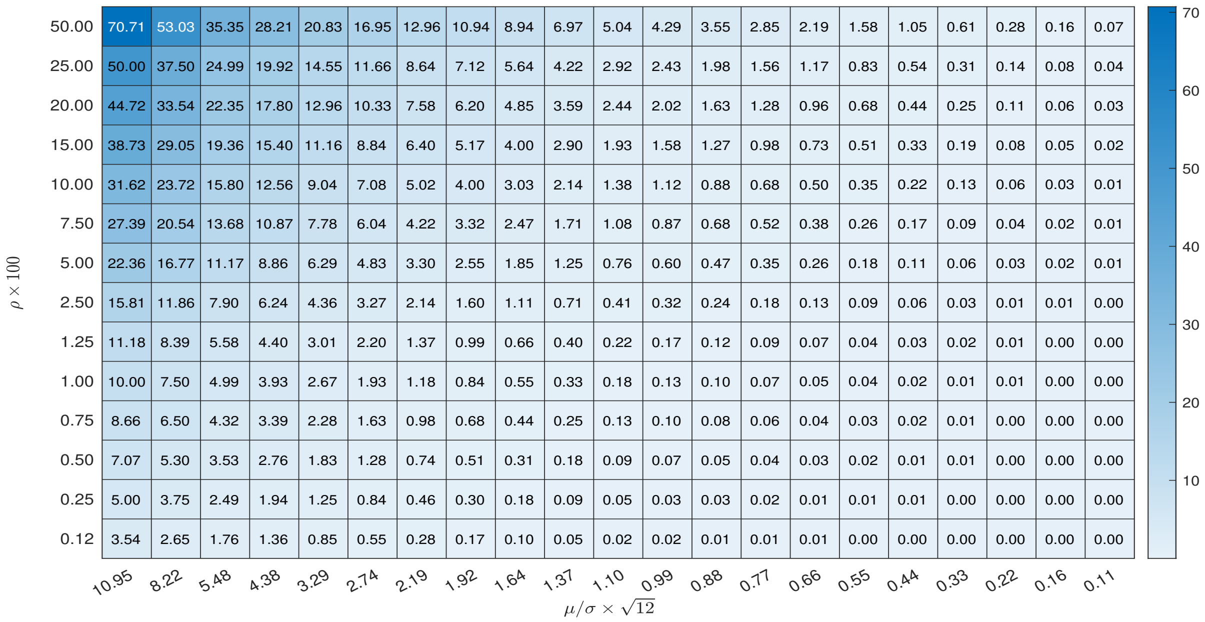
$$\mu \precsim N^{-\eta} \quad \text{and} \quad \rho > 0 \quad \text{is fixed.} \tag{17}$$

In this scenario, $(S^{\star})^2 \precsim_{\text{P}} N^{1-2\eta}$, which explodes unless $\eta > 1/2$. If we further assume that $N/T \to \psi > 0$, then the left-hand-side of condition (16) is of order $\max\{N^{1/2-\eta}, 1\}$, so that (16) holds if and only if $\eta \geq 1/2$. Therefore, $\eta < 1/2$ is not consistent with absence of (feasible) near arbitrage because $S^{\star}$ explodes and, since $S^{\text{OPT}} \precsim_{\text{P}} S^{\star}$ by Corollary 2, S^{OPT} explodes, too. If $\eta > 1/2$, $S^{\star}$ (and hence S^{OPT}) vanishes, which does not seem like an economically plausible case. If we think that asset demand distortions are sufficiently big so that arbitrageur activity is required to prevent substantial mispricing, and that arbitrageurs require some compensation for holding arbitrage positions, then a setting where true alphas disappear asymptotically is not plausible. This suggests that under the DGP (17), the only economically plausible case with absence of near-arbitrage, but not absence of mispricing, is $\eta = 1/2$. That is, η can be thought as determined in equilibrium. Mispricing should be big enough that arbitrageurs are active, earning a non-vanishing Sharpe ratio, and at the same time small enough that near-arbitrage opportunities do not exist.

We now illustrate the behavior of S^{OPT} numerically and verify the theoretical predictions of Corollary 2 using the setting of Example 1. Figure 1 reports the Sharpe ratio upper bound, S^{OPT}, of feasible arbitrage portfolios for a range of μ/σ and ρ values in the case of $N = 1,000$ and $T = 20$ years. Recall that according to model (5), a fraction ρ of assets have alphas with a Sharpe ratio μ/σ. That is, ρ characterizes the rareness of the alpha signal, whereas μ/σ captures its strength. We intentionally choose a wide range of μ/σ (with annualized Sharpe ratios from 0.11 to 10.95) and ρ (from 0.12% to 50%) to shed light on the dependence of Sharpe ratios on signal weakness and rareness, although some of the resulting portfolio Sharpe ratios (the top left corner of Figure 1) are unrealistically high. Note that when $\mu/\sigma \times \sqrt{12}$ hits 0.44, its corresponding t-statistic based on a 20-year sample exceeds 1.96, the typical t-hurdle for a standard student-t test.

[30] It is easy to show that the setup (17) satisfies all assumptions of Corollary 1 for all fixed $\eta > 0$.

The pattern of Sharpe ratios agrees with our intuition and theoretical predictions. For any fixed ρ, as the alpha signal weakens (i.e., μ/σ decreases), the feasible Sharpe ratio bound drops. The same is true if we decrease the signal count (i.e., ρ vanishes), for any fixed value of μ/σ. The arbitrageur's learning problem is easiest when signal is strong and count is large (top left corner), and most challenging towards the right bottom corner, where the feasible Sharpe ratio bound drop to near zero.



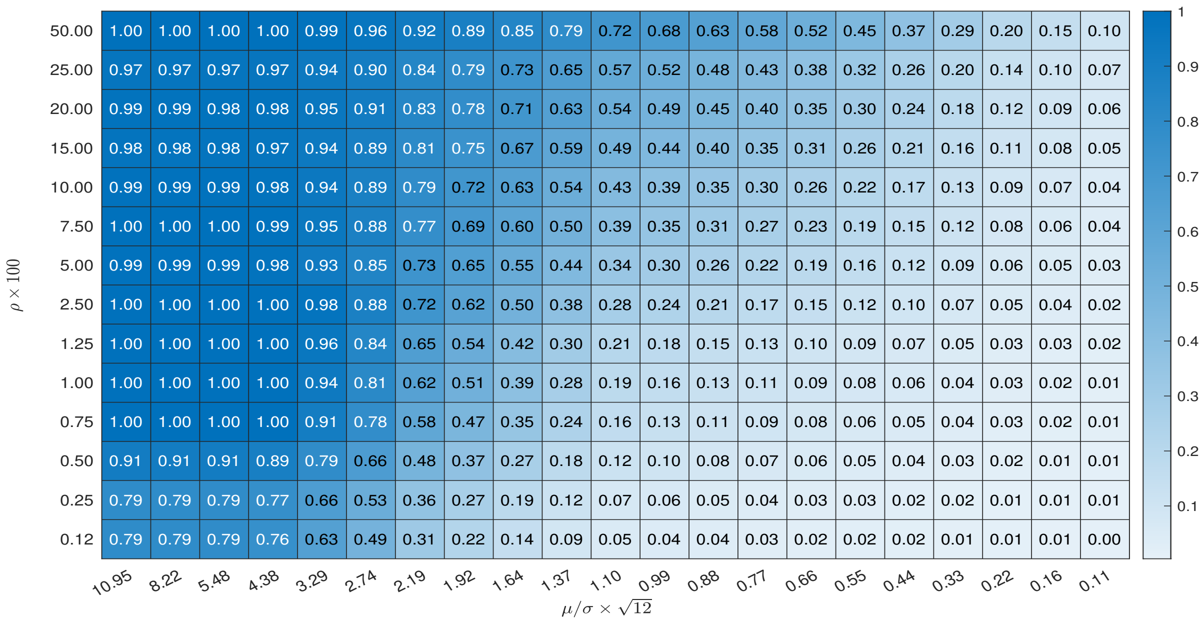
$\rho\times 100$	10.95	8.22	5.48	4.38	3.29	2.74	2.19	1.92	1.64	1.37	1.10	0.99	0.88	0.77	0.66	0.55	0.44	0.33	0.22	0.16	0.11
50.00	70.71	53.03	35.35	28.21	20.83	16.95	12.96	10.94	8.94	6.97	5.04	4.29	3.55	2.85	2.19	1.58	1.05	0.61	0.28	0.16	0.07
25.00	50.00	37.50	24.99	19.92	14.55	11.66	8.64	7.12	5.64	4.22	2.92	2.43	1.98	1.56	1.17	0.83	0.54	0.31	0.14	0.08	0.04
20.00	44.72	33.54	22.35	17.80	12.96	10.33	7.58	6.20	4.85	3.59	2.44	2.02	1.63	1.28	0.96	0.68	0.44	0.25	0.11	0.06	0.03
15.00	38.73	29.05	19.36	15.40	11.16	8.84	6.40	5.17	4.00	2.90	1.93	1.58	1.27	0.98	0.73	0.51	0.33	0.19	0.08	0.05	0.02
10.00	31.62	23.72	15.80	12.56	9.04	7.08	5.02	4.00	3.03	2.14	1.38	1.12	0.88	0.68	0.50	0.35	0.22	0.13	0.06	0.03	0.01
7.50	27.39	20.54	13.68	10.87	7.78	6.04	4.22	3.32	2.47	1.71	1.08	0.87	0.68	0.52	0.38	0.26	0.17	0.09	0.04	0.02	0.01
5.00	22.36	16.77	11.17	8.86	6.29	4.83	3.30	2.55	1.85	1.25	0.76	0.60	0.47	0.35	0.26	0.18	0.11	0.06	0.03	0.02	0.01
2.50	15.81	11.86	7.90	6.24	4.36	3.27	2.14	1.60	1.11	0.71	0.41	0.32	0.24	0.18	0.13	0.09	0.06	0.03	0.01	0.01	0.00
1.25	11.18	8.39	5.58	4.40	3.01	2.20	1.37	0.99	0.66	0.40	0.22	0.17	0.12	0.09	0.07	0.04	0.03	0.02	0.01	0.00	0.00
1.00	10.00	7.50	4.99	3.93	2.67	1.93	1.18	0.84	0.55	0.33	0.18	0.13	0.10	0.07	0.05	0.04	0.02	0.01	0.01	0.00	0.00
0.75	8.66	6.50	4.32	3.39	2.28	1.63	0.98	0.68	0.44	0.25	0.13	0.10	0.08	0.06	0.04	0.03	0.02	0.01	0.00	0.00	0.00
0.50	7.07	5.30	3.53	2.76	1.83	1.28	0.74	0.51	0.31	0.18	0.09	0.07	0.05	0.04	0.03	0.02	0.01	0.01	0.00	0.00	0.00
0.25	5.00	3.75	2.49	1.94	1.25	0.84	0.46	0.30	0.18	0.09	0.05	0.03	0.03	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.00
0.12	3.54	2.65	1.76	1.36	0.85	0.55	0.28	0.17	0.10	0.05	0.02	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Figure 1: Sharpe Ratio Bound (S^{OPT}) of Feasible Arbitrage Portfolios

Note: The figure reports Sharpe ratio bound of feasible arbitrage portfolios in model (5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

The reported Sharpe ratios in Figure 1 are only a fraction of the corresponding (infeasible) Sharpe ratios, $S^{\star} = \sqrt{\alpha^{\intercal}(\Sigma_u)^{-1}\alpha} = \mu/\sigma\sqrt{\rho N}$, as shown by Figure 2. The pattern we see from Figure 2 agrees with theoretical predictions of Corollary 2. When the annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$ is larger than 2.74, regardless of the values of ρ, the signal-to-noise ratio of the learning problem is sufficiently strong that the statistical limit to arbitrage does not matter much, and hence $S^{\text{OPT}}/S^{\star}$ is close to 1. Nonetheless, this regime is irrelevant in practice, since it is mostly associated with unrealistically high Sharpe ratios (see Figure 1). In contrast, as μ/σ diminishes, the gap between $S^{\star}$ and S^{OPT} widens. In almost all empirically relevant scenarios, $S^{\star}$ is much larger than the maximal Sharpe ratio than arbitrageurs can actually earn with a feasible strategy.

The example above illustrates how $S(\mathcal{G})$ depends on the underlying prior, which arbitrageurs do not observe. To build a feasible strategy that adapts to any admissible prior, one must infer the underlying prior from observable data—a task complicated by the possi-



	10.95	8.22	5.48	4.38	3.29	2.74	2.19	1.92	1.64	1.37	1.10	0.99	0.88	0.77	0.66	0.55	0.44	0.33	0.22	0.16	0.11
50.00	1.00	1.00	1.00	1.00	0.99	0.96	0.92	0.89	0.85	0.79	0.72	0.68	0.63	0.58	0.52	0.45	0.37	0.29	0.20	0.15	0.10
25.00	0.97	0.97	0.97	0.97	0.94	0.90	0.84	0.79	0.73	0.65	0.57	0.52	0.48	0.43	0.38	0.32	0.26	0.20	0.14	0.10	0.07
20.00	0.99	0.99	0.98	0.98	0.95	0.91	0.83	0.78	0.71	0.63	0.54	0.49	0.45	0.40	0.35	0.30	0.24	0.18	0.12	0.09	0.06
15.00	0.98	0.98	0.98	0.97	0.94	0.89	0.81	0.75	0.67	0.59	0.49	0.44	0.40	0.35	0.31	0.26	0.21	0.16	0.11	0.08	0.05
10.00	0.99	0.99	0.99	0.98	0.94	0.89	0.79	0.72	0.63	0.54	0.43	0.39	0.35	0.30	0.26	0.22	0.17	0.13	0.09	0.07	0.04
7.50	1.00	1.00	1.00	0.99	0.95	0.88	0.77	0.69	0.60	0.50	0.39	0.35	0.31	0.27	0.23	0.19	0.15	0.12	0.08	0.06	0.04
5.00	0.99	0.99	0.99	0.98	0.93	0.85	0.73	0.65	0.55	0.44	0.34	0.30	0.26	0.22	0.19	0.16	0.12	0.09	0.06	0.05	0.03
2.50	1.00	1.00	1.00	1.00	0.98	0.88	0.72	0.62	0.50	0.38	0.28	0.24	0.21	0.17	0.15	0.12	0.10	0.07	0.05	0.04	0.02
1.25	1.00	1.00	1.00	1.00	0.96	0.84	0.65	0.54	0.42	0.30	0.21	0.18	0.15	0.13	0.10	0.09	0.07	0.05	0.03	0.03	0.02
1.00	1.00	1.00	1.00	1.00	0.94	0.81	0.62	0.51	0.39	0.28	0.19	0.16	0.13	0.11	0.09	0.08	0.06	0.04	0.03	0.02	0.01
0.75	1.00	1.00	1.00	1.00	0.91	0.78	0.58	0.47	0.35	0.24	0.16	0.13	0.11	0.09	0.08	0.06	0.05	0.04	0.03	0.02	0.01
0.50	0.91	0.91	0.91	0.89	0.79	0.66	0.48	0.37	0.27	0.18	0.12	0.10	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.01
0.25	0.79	0.79	0.79	0.77	0.66	0.53	0.36	0.27	0.19	0.12	0.07	0.06	0.05	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0.01
0.12	0.79	0.79	0.79	0.76	0.63	0.49	0.31	0.22	0.14	0.09	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.00

Figure 2: Ratios between S^{OPT} and $S^{\star}$

Note: The figure reports the ratios of feasible Sharpe ratio bound and optimal infeasible Sharpe ratio. The simulation setting is based on model (5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

bility that alphas are rare and weak. Theorem 2 shows that the influence of the prior on the posterior mean $\widetilde{s}$ is fully captured by the marginal distribution of the observable statistic $\check{s}$. Leveraging this insight, the next subsection develops a uniformly optimal feasible strategy using a nonparametric estimate of this marginal density.

2.6 Constructing Uniformly Optimal Arbitrage Portfolio

As established after Theorem 1, the key challenge is to approximate the infeasible benchmark $\widetilde{w}$ uniformly across all DGPs using only observable data. Although Theorem 2 provides an explicit form for $\widetilde{w}$, its implementation depends on the marginal density $p(\cdot)$ and on the ex-factor returns $\bar{r}_i - \beta_i(\gamma + \bar{v})$, which in turn require knowledge of v_t and γ—quantities that are not directly observed.

Nonetheless, arbitrageurs can construct a uniformly optimal feasible strategy. We do so under a framework in which factor exposures are observable, but factors themselves remain latent—a setting consistent with our empirical analysis.

The following algorithm outlines the construction of this strategy:

Algorithm 1 (Optimal Arbitrage Portfolio via Empirical Bayes)**.**
Inputs: r_s, $s \in \mathcal{T} = \{t - T + 1, \ldots, t\}$ *and* β.

S1. Construct cross-sectional regression estimates of alpha and idiosyncratic volatilities, for each $i = 1, 2, \ldots, N$:

$$\widehat{\alpha} = T^{-1}\sum_{s\in\mathcal{T}}\mathbb{M}_\beta r_s, \qquad \widehat{\sigma}_i^2 = T^{-1}\sum_{s\in\mathcal{T}}\big((\mathbb{M}_\beta r_s)_i - \widehat{\alpha}_i\big)^2, \quad \text{and} \quad \widehat{s}_i := \widehat{\alpha}_i/\widehat{\sigma}_i.$$

S2. Construct a nonparametric estimate of the marginal density of $\widehat{s}$ using Gaussian kernel function $\phi_{1/T}(x)$ and bandwidth $k_N \precsim (\log N)^{-1}$:

$$\widehat{p}(a) = \frac{1}{Nk_N}\sum_i \phi_{1/T}\Big(\frac{\widehat{s}_i - a}{k_N}\Big).$$

S3. Estimate $\widetilde{s}$ by Tweedie's formula (15) and isotonic regression:

$$\begin{aligned}\breve{\psi}(a) =& a + \frac{1+k_N^2}{T}\frac{d}{da}\log\widehat{p}(a).\\ \widehat{\psi} =& \arg\min_{x\in\mathbb{R}^N}\|x-\breve{\psi}\|^2, \text{ s.t. } x_i \leq x_j \text{ if } \widehat{s}_i \leq \widehat{s}_j, \text{ for } 1 \leq i, j \leq N, \text{ where } \breve{\psi}_i := \breve{\psi}(\widehat{s}_i).\end{aligned}$$

S4. Construct the arbitrage portfolio weights as $\widehat{w}^{\mathrm{OPT}} = \kappa^{-1}\mathbb{M}_\beta\widehat{\Sigma}_u^{-1/2}\widehat{\psi}$.

Outputs: $\widehat{w}^{\mathrm{OPT}}$.

Step S1 of Algorithm 1 provides feasible estimates of $\widehat{\alpha}$ and $\widehat{\Sigma}_u = \mathrm{Diag}(\widehat{\sigma}_i^2)$, which in turn leads to the sufficient statistic, $\widehat{s}$. The motivation behind Steps S2 and S3 stems from Tweedie's formula (15); here, we employ a nonparametric empirical Bayes method for estimating the posterior mean function using kernel density estimation, as suggested by Brown and Greenshtein (2009). The incorporation of the factor $(1 + k_N^2)$ serves to adjust for finite sample biases introduced by estimation errors in $\widehat{p}(a)$. As discussed earlier, $\psi(\cdot)$ is nondecreasing. To enhance the nonparametric estimator's performance in finite samples, enforcing monotonicity on the estimate of $\psi(\cdot)$ proves beneficial. For this purpose, isotonic regression is utilized (see Robertson et al. (1988)), yielding a monotonic piece-wise linear approximation of $\psi(\cdot)$. Step S4 constructs the optimal portfolio weights, $\widehat{w}^{\mathrm{OPT}}$, following (14).

An essential step towards achieving optimality involves aggregating information from assets with comparable $\widehat{s}_i$, as done in Step S2. This strategy outperforms the alternatives, some of which directly use estimated $\widehat{s}$ as if these estimates are not susceptible to estimation errors even when they are rather weak, or simply ignore the contribution of all weaker signals. Like any machine learning method, the proposed approach requires a tuning parameter k_N, which can be selected in a validation sample. The next theorem demonstrates the success of $\widehat{w}^{\mathrm{OPT}}$:

Theorem 3. *Let $\mathbb{P}$ denote the collection of all DGPs under which the assumptions in Corollary 1 hold, and assume that $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$, $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$, and that the distribution of s_i is symmetric. Further suppose that $N^d \lesssim T \lesssim N^{d'}$ for fixed constants $d > 1/2$ and $d' < 1$. Then, for all small fixed $\epsilon > 0$,*

$$\inf_{\mathrm{P}\in\mathbb{P}} \Big(U(\widehat{w}^{\mathrm{OPT}}) =_\epsilon U(\widetilde{w}) \Big) \to 1, \quad \textit{and} \quad \inf_{\mathrm{P}\in\mathbb{P}} \Big(S(\widehat{w}^{\mathrm{OPT}}) =_\epsilon S(\widetilde{w}) \Big) \to 1.$$

Thus, $\widehat{w}^{\mathrm{OPT}}$ is uniformly optimal and feasible with respect to the information set generated by $\{r_s, \beta\}_{s=t-T+1}^t$.[31]

Theorem 3 imposes a mild assumption on $\lambda_{\min}(\beta^\intercal\beta)$, which requires that all factors are pervasive.[32] This condition is frequently utilized in the factor model literature and is particularly relevant here, given our assumption that the factors within our model are latent. A significant difference between our context and the traditional empirical Bayes literature (e.g., Brown and Greenshtein (2009)) lies in our lack of direct access to the signals, $\check{s}_i$'s. Instead, we estimate these signals via cross-sectional regressions in Algorithm 1. This step results in our estimators, $\widehat{s}_i$'s, being inevitably polluted by estimation errors. The additional conditions imposed by Theorem 3 ensure that such estimation errors become asymptotically negligible.

Theorem 3 shows that, within a linear factor model, arbitrageurs with access only to past returns and risk exposures—and facing uncertainty over a broad class of DGPs—can construct $\widehat{w}^{\mathrm{OPT}}$ to attain the maximal Sharpe ratio among all feasible strategies with zero factor exposure. Remarkably, they match the performance of arbitrageurs with additional access to latent factors, risk premia, idiosyncratic volatility, or the true prior. This implies that the economic cost of feasibility arises solely from the inability to observe alphas; uncertainty about the prior or lack of information about factors and volatility does not hinder optimality. The resulting Sharpe ratio precisely quantifies the economic limit of feasible arbitrage.

A key reason for this result is that idiosyncratic variances represented by Σ_u remain bounded both from below and above as N and T increase, unlike alphas. This difference is empirically plausible: alphas must be small and rare to persist in the presence of arbitrageurs, whereas there is no analogous constraint on the magnitude of idiosyncratic risk. Consequently, uncertainty about Σ_u becomes negligible in the context of statistical arbitrage limitations. Furthermore, given the low-dimensional nature of v_t and γ and arbitrageurs' knowledge knowledge of β, the problem of learning about v_t and γ is negligible as well. Finally, the influence of the unknown prior is effectively eliminated by the availability of a large

[31] The uniform optimality remains valid for any information set that contains $\{r_s, \beta\}_{s=t-T+1}^t$ but is a subset of $\mathcal{G}$ as defined in Theorem 1.

[32] See, e.g., Assumption I.1 of Giglio and Xiu (2021). While our theoretical results may extend to certain weak factor settings, this is not our emphasis here.

cross-section of assets.

Our results also address a longstanding challenge in optimal portfolio choice under parameter uncertainty. Kan and Zhou (2007) examine the expected performance of the plug-in mean-variance portfolio and find that its Sharpe ratio is smaller than the infeasible Sharpe ratio, $S^{\star}$. To construct a solution, they consider a limited class of trading strategies that does not accommodate the broad range of strategies investors might consider in real-world scenarios. In contrast, by imposing mild restrictions on the DGP of returns, we identify the optimal ones among all feasible strategies.

More broadly, important economic implications are carried by the fact that every uniformly optimal feasible strategy adopted by arbitrageurs realizes S^{OPT}. In an economy with statistical limits to arbitrage, the equilibrium compensation that arbitrageurs demand for executing such a strategy equals S^{OPT}. If it were otherwise, arbitrageurs would continue trading until all profit opportunities were exhausted. Therefore, $\widehat{S}^{\mathrm{OPT}} := S(\widehat{w}^{\mathrm{OPT}})$, the Sharpe ratio generated by $\widehat{w}^{\mathrm{OPT}}$, can be interpreted as an empirical estimate of this equilibrium compensation, which we seek to pin down empirically.

2.7 Estimating Optimal Infeasible Sharpe Ratio

We are also interested in estimating the optimal infeasible Sharpe ratio, $S^{\star}$. This Sharpe ratio can be estimated by an econometrician ex-post and in-sample, but it cannot be achieved by any feasible portfolio. The existing literature on testing the APT often constructs test statistics in the spirit of Gibbons et al. (1989), which are effectively based on $S^{\star}$ (see, e.g., Pesaran and Yamagata (2017) and Fan et al. (2015)). While such tests are powerful and may lead to detection of alphas, their relevance for arbitrageurs might be limited. The challenge for arbitrageurs lies in translating statistical evidence of alpha discoveries into a feasible portfolio strategy that realizes profits. Our proposed $\widehat{S}^{\mathrm{OPT}}$ tackles this issue, providing an economically more meaningful evaluation of the APT.

To compare with $\widehat{S}^{\mathrm{OPT}}$, we propose an estimator for $S^{\star}$ inspired by its sample analog:

$$\widetilde{S}^{\star} = \left(\bar{r}^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_{u}^{-1}\mathbb{M}_{\beta}\bar{r}\right)^{1/2}. \tag{18}$$

Unfortunately, this estimator has a non-vanishing asymptotic bias for certain DGPs we consider, as we will show later. To fix this issue, we propose a new estimator that is uniformly consistent:

$$\widehat{S}^{\star} = \left(\bar{r}^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_{u}^{-1}\mathbb{M}_{\beta}\bar{r} - N/T\right)^{1/2}. \tag{19}$$

The next proposition summarizes the asymptotic properties of both estimators.

Proposition 1. *Suppose that the same assumptions outlined in Corollary 1 hold. Moreover, we assume* $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$ *and* $\lambda_{\min}(\beta^{\intercal}\beta) \gtrsim_{\mathrm{P}} N$. *We also impose* $T \lesssim N$ *and* $T^{-1}N^{1/2}\log N \to 0$. *Then we have*

$$\begin{aligned}\left|\widehat{S}^{\star} - S^{\star}\right| \Big/ (1+S^{\star}) &= o_{\mathrm{P}}\Big(T^{-1/2}N^{1/4}\sqrt{\log N}\Big),\\ \left|\widetilde{S}^{\star} - \Big((S^{\star})^2 + NT^{-1}\Big)^{1/2}\right| \Big/ (1+S^{\star}) &= o_{\mathrm{P}}\Big(T^{-1}N^{1/2}\log N\Big).\end{aligned}$$

Similar to Theorem 3, the estimation error is relative when $S^{\star}$ dominates 1.0 asymptotically, and in absolute terms if $S^{\star}$ is dominated by 1.0.[33] This accommodates a large class of models, some of which have an exploding or a shrinking $S^{\star}$. While it is possible to estimate $S^{\star}$, it is not possible to build a portfolio that realizes it, unless the signal-to-noise ratio is sufficiently large such that $S^{\star} = S^{\mathrm{OPT}}$. Empirically, the difference between $\widehat{S}^{\star}$ and $\widehat{S}^{\mathrm{OPT}}$ thereby informs us about the economic consequences of learning.

2.8 Alternative Strategies for Arbitrage Portfolios

Algorithm 1 introduces a nuanced approach that allows arbitrageurs to achieve feasible optimality. In practice, many empirical asset pricing researchers and practitioners often rely on simpler strategies. Given the widespread adoption of these alternative approaches, it is useful to assess how closely they approximate the optimal feasible strategy. In this section, we examine their strengths and weaknesses in different DGP scenarios.

2.8.1 Cross-Sectional Regression

One of the most common strategies involves forming portfolios based directly on the cross-sectional regression estimates of α, obtained in Step S1 of Algorithm 1. This method is referred to as CSR, with its portfolio represented by $\widehat{w}^{\mathrm{CSR}}$:

$$\widehat{w}^{\mathrm{CSR}} \propto \mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\widehat{\alpha}. \tag{20}$$

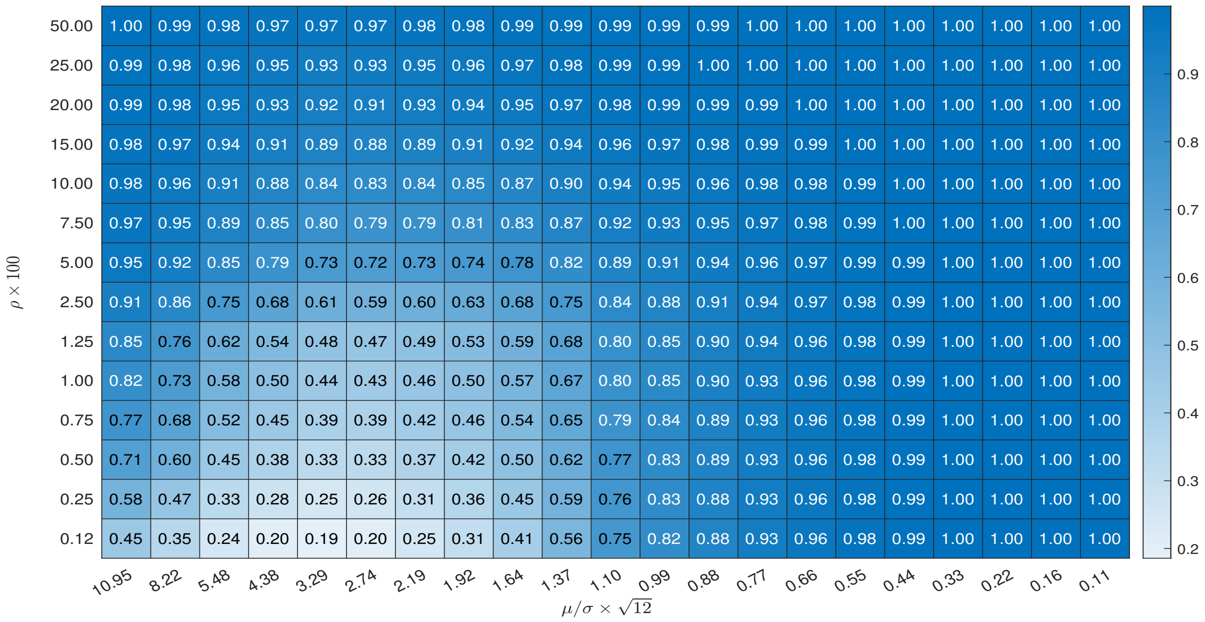
This is effectively a sample analogue of the strategy $\widetilde{w}$ given by Theorem 1. A specific instance of this strategy was discussed following Example 1.

We now exploit Example 1 to illustrate the pros and cons of the CSR strategy. Figure 3 illustrates the relationship between S^{CSR}, the theoretical Sharpe ratio achieved by $\widehat{w}^{\mathrm{CSR}}$, and S^{OPT} across a spectrum of DGPs. According to Proposition B2 in the online appendix, S^{CSR} tends to be dominated by S^{OPT} in scenarios when alpha signals are sparse ($\sqrt{\rho T}\mu/\sigma$ is not

[33] Obviously, the threshold 1.0 can be replaced by any fixed constant.

excessively large) and simultaneously strong ($\sqrt{T}\mu/\sigma$ is not exceedingly small). This specific regime of dominance is clearly marked with black numerals within the heatmap in Figure 3. As the ratio $\mu/\sigma \times \sqrt{12}$ edges closer to 1.0—either moving towards this vertical threshold from the right hand side or descending from the upper left corner—the gap between S^{CSR} and S^{OPT} enlarges increasingly.

The CSR approach takes all signal estimates directly, without differentiating the significant ones from the insignificant ones. Consequently, even fake signals (pure noise) are assigned non-zero weights. This hurts the portfolio's performance. On the other hand, the CSR strategy can achieve optimality when the strong signals are abundant (so that portfolio weights allocated to noise are inconsequential) or when all signals are weak (so that they do not differ too much from fake ones). The latter case is interesting, as it also suggests that simply ignoring weaker signals is not optimal. That said, Figure 1 shows that the DGPs for which the cross-sectional regression approach is strongly dominated by our optimal strategy are associated with realistic Sharpe ratios.



	10.95	8.22	5.48	4.38	3.29	2.74	2.19	1.92	1.64	1.37	1.10	0.99	0.88	0.77	0.66	0.55	0.44	0.33	0.22	0.16	0.11
50.00	1.00	0.99	0.98	0.97	0.97	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25.00	0.99	0.98	0.96	0.95	0.93	0.93	0.95	0.96	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20.00	0.99	0.98	0.95	0.93	0.92	0.91	0.93	0.94	0.95	0.97	0.98	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15.00	0.98	0.97	0.94	0.91	0.89	0.88	0.89	0.91	0.92	0.94	0.96	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
10.00	0.98	0.96	0.91	0.88	0.84	0.83	0.84	0.85	0.87	0.90	0.94	0.95	0.96	0.98	0.98	0.99	1.00	1.00	1.00	1.00	1.00
7.50	0.97	0.95	0.89	0.85	0.80	0.79	0.79	0.81	0.83	0.87	0.92	0.93	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00
5.00	0.95	0.92	0.85	0.79	0.73	0.72	0.73	0.74	0.78	0.82	0.89	0.91	0.94	0.96	0.97	0.99	0.99	1.00	1.00	1.00	1.00
2.50	0.91	0.86	0.75	0.68	0.61	0.59	0.60	0.63	0.68	0.75	0.84	0.88	0.91	0.94	0.97	0.98	0.99	1.00	1.00	1.00	1.00
1.25	0.85	0.76	0.62	0.54	0.48	0.47	0.49	0.53	0.59	0.68	0.80	0.85	0.90	0.94	0.96	0.98	0.99	1.00	1.00	1.00	1.00
1.00	0.82	0.73	0.58	0.50	0.44	0.43	0.46	0.50	0.57	0.67	0.80	0.85	0.90	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00
0.75	0.77	0.68	0.52	0.45	0.39	0.39	0.42	0.46	0.54	0.65	0.79	0.84	0.89	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00
0.50	0.71	0.60	0.45	0.38	0.33	0.33	0.37	0.42	0.50	0.62	0.77	0.83	0.89	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00
0.25	0.58	0.47	0.33	0.28	0.25	0.26	0.31	0.36	0.45	0.59	0.76	0.83	0.88	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00
0.12	0.45	0.35	0.24	0.20	0.19	0.20	0.25	0.31	0.41	0.56	0.75	0.82	0.88	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00

Figure 3: Ratios between S^{CSR} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the OLS based portfolio and the feasible optimal arbitrage portfolio. The theoretical Sharpe ratio achieved by CSR is denoted as S^{CSR}, given explicitly by Proposition B2. The simulation setting is based on model (5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

The CSR approach is a simple benchmark as it does not rely on any advanced statistical techniques to detect signals or distinguish their strength. The strategy we discuss next is more advanced, in that it controls false discoveries among selected strong signals using the

B-H procedure proposed by Benjamini and Hochberg (1995).

2.8.2 False Discovery Rate Control

To address the aforementioned selection bias in identifying profitable alpha signals, an alternative methodology conceptualizes the search for alpha as a multiple testing problem. In this case, with N assets that are each potentially associated with a nonzero α_i, we can establish for each asset i a null hypothesis $\mathbb{H}_0^i : \alpha_i = 0$. Rejection of this null hypothesis leads to the discovery of a non-zero alpha. Rather than focusing on the significance level of individual tests, a more appropriate strategy involves controlling the FDR, an approach recommended by Barras et al. (2010), Bajgrowicz and Scaillet (2012), and Harvey et al. (2016) in various asset pricing contexts. Giglio et al. (2021) have proved the validity of the B-H procedure for FDR control in a general factor model setting for alpha detection. Below we describe the necessary steps to prepare alpha estimates for constructing an arbitrage portfolio.

Begin with a series of p-values, p_i, where each is the result of a t-test on the cross-sectional regression estimate of α_i, $\sqrt{T}\widehat{s}_i$, for $i = 1, 2, \ldots, N$. These p-values assess the significance of each α_i's deviation from zero. Arrange these p-values in ascending order, from the smallest to the largest, resulting in a sorted sequence $p_{(1)} \leq \ldots \leq p_{(N)}$. Identify a critical index, $\widehat{k}$, defined as the maximum i such that $p_{(i)} \leq \tau i/N$ where τ is a predetermined significance level, commonly set at 5%.

The threshold $\widehat{k}$ is selected such that, on average, at least a fraction $(1 - \tau)$ of the alpha estimates identified as significant (i.e., those with p-values smaller than $p_{(\widehat{k})}$) are truly non-zero. This B-H criterion proves effective in guarding against false discoveries irrespective of the overall proportion of non-zero alphas present in the DGP. The selected alpha estimates are then used as inputs for constructing an arbitrage portfolio, as illustrated by the following equation:

$$\widehat{w}^{\mathrm{BH}}(\tau) \propto \mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\widehat{\alpha}^{\mathrm{BH}}(\tau), \text{ where } \widehat{\alpha}_i^{\mathrm{BH}}(\tau) = \widehat{\alpha}_i \mathbb{1}_{\{p_i \leq p_{(\widehat{k})}\}}. \tag{21}$$

This strategy introduces a hard-thresholding mechanism to the alpha estimates, effectively nullifying the impact of alphas deemed insignificant.

Controlling the FDR on top of the CSR estimates is intuitively appealing, but doing so incurs a potential loss of power that may hurt investment performance. Indeed, Proposition B3 in the appendix shows, in the context of Example 1, arbitrageurs who adopt the B-H based-trading strategy do not achieve the optimal portfolio for a large class of DGP sequences.

As shown by Proposition B3 and illustrated numerically by Figure 4, the discrepancy between the optimal Sharpe ratio and that achieved through the B-H method is largely determined by signal strength. The B-H procedure nears optimality when the signal strength, quantified by $\sqrt{T}\mu/\sigma$, is substantial—exceeding the threshold of $\sqrt{-2\log\rho}$. The instances where the B-H strategy reaches optimality are depicted by the white values on Figure 4, with the boundary of this optimal region approximated by the line where $\mu/\sigma\sqrt{12} = 2.19$. This demonstrates that the B-H method excels in identifying strong signals, leading to near-optimal portfolios when signals are strong. Conversely, in the presence of weak signals, the B-H procedure, which amounts to hard-thresholding, tends to underperform compared to CSR.

This point is further elaborated in Figure 1, which underscores that even if individual alphas are weak, their aggregated effect on a portfolio's Sharpe ratio can be non-trivial. The B-H approach takes a notably conservative stance towards signal selection, especially in contexts where signals are weak. This cautious approach ensures the reliability of selected alphas by focusing on those that are truly significant. However, this method might not fully capitalize on the potential cumulative impact of weaker signals. In contrast, our optimal arbitrage portfolio leverages the full spectrum of alpha estimates, including false positives, extending beyond the significant selections made through the B-H procedure.

At its core, this delineates a subtle yet critical divergence between two objectives: alpha testing and portfolio construction. Alpha testing prioritizes the identification of statistically significant alphas while controlling FDR, whereas portfolio construction concentrates on using all available information to optimize performance. These objectives do not always align.

The CSR and the B-H approaches represent two typical strategies in practice. The former trades all signals without distinguishing their strength, whereas the latter only trades the stronger signals. Neither approach always achieves optimality.

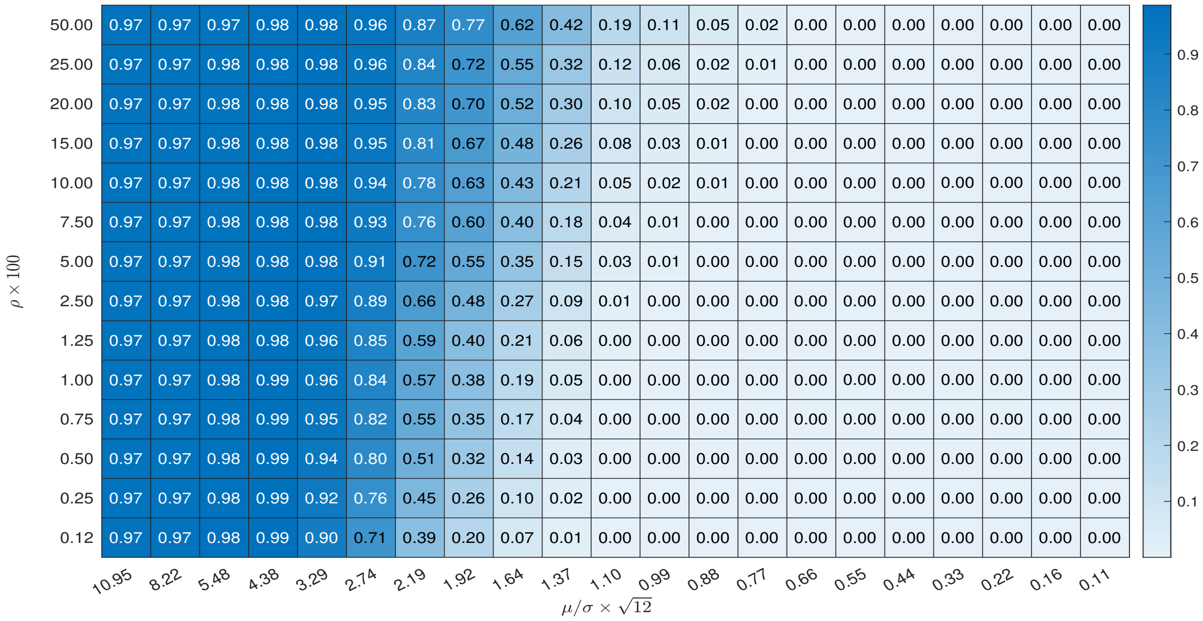
2.8.3 Shrinkage Approaches

In the strategy $\widetilde{w}$ given by Theorem 1, the portfolio weights with respect to ex-factor returns $(\mathbb{M}_\beta r_{t+1} \approx \alpha + u_t)$ are $\Sigma_u^{-1}\alpha$. The CSR approach replaces the weights $\Sigma_u^{-1}\alpha$ with the sample analogue $\widehat{\Sigma}_u^{-1}\widehat{\alpha}$, and the B-H approach additionally imposes hard-thresholding regularization on those weights.[34] Multiplying the ex-factor return weights by $\mathbb{M}_\beta$ yields the portfolio weights in terms of raw returns (r_{t+1}). Besides the hard-thresholding of B-H, we can consider shrinkage-type regularization:

$$\max_w \{w^\intercal\widehat{\alpha} - \frac{1}{2}w^\intercal\widehat{\Sigma}_u w - p_\lambda(\widehat{\Sigma}_u^{1/2}w)\},$$

where $p_\lambda(x) = \lambda\|x\|_1$ or $\lambda\|x\|_2^2$, for some $\lambda > 0$. Since $\widehat{\Sigma}_u$ is diagonal, this optimization problem has a closed-form solution of weights with respect to ex-factor returns: $\breve{w}_{q,i}(\lambda) =$

[34] Regularizing those weights amounts to incorporating priors onto the alpha estimates.



$\rho \times 100$ \ $\mu/\sigma \times \sqrt{12}$	10.95	8.22	5.48	4.38	3.29	2.74	2.19	1.92	1.64	1.37	1.10	0.99	0.88	0.77	0.66	0.55	0.44	0.33	0.22	0.16	0.11
50.00	0.97	0.97	0.97	0.98	0.98	0.96	0.87	0.77	0.62	0.42	0.19	0.11	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25.00	0.97	0.97	0.98	0.98	0.98	0.96	0.84	0.72	0.55	0.32	0.12	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20.00	0.97	0.97	0.98	0.98	0.98	0.95	0.83	0.70	0.52	0.30	0.10	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15.00	0.97	0.97	0.98	0.98	0.98	0.95	0.81	0.67	0.48	0.26	0.08	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10.00	0.97	0.97	0.98	0.98	0.98	0.94	0.78	0.63	0.43	0.21	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7.50	0.97	0.97	0.98	0.98	0.98	0.93	0.76	0.60	0.40	0.18	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5.00	0.97	0.97	0.98	0.98	0.98	0.91	0.72	0.55	0.35	0.15	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2.50	0.97	0.97	0.98	0.98	0.97	0.89	0.66	0.48	0.27	0.09	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.25	0.97	0.97	0.98	0.98	0.96	0.85	0.59	0.40	0.21	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.97	0.97	0.98	0.99	0.96	0.84	0.57	0.38	0.19	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.75	0.97	0.97	0.98	0.99	0.95	0.82	0.55	0.35	0.17	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.97	0.97	0.98	0.99	0.94	0.80	0.51	0.32	0.14	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.25	0.97	0.97	0.98	0.99	0.92	0.76	0.45	0.26	0.10	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.12	0.97	0.97	0.98	0.99	0.90	0.71	0.39	0.20	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 4: Ratios between S^{BH} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the multiple testing based portfolio (via B-H procedure) and the feasible optimal arbitrage portfolio. The theoretical Sharpe ratio achieved by B-H is denoted as S^{BH}, given explicitly by Proposition B3. The simulation setting is based on model (5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

$\widehat{\sigma}_i^{-1}\psi_q(\widehat{s}_i, \lambda)$, for $i = 1, 2, \ldots, N$, where $q = 1$ corresponds to the Lasso penalty and $q = 2$ the ridge, and $\psi_q(s, \lambda)$ is

$$\psi_1(s, \lambda) = \operatorname{sgn}(s)(|s| - \lambda)_+, \quad \psi_2(s, \lambda) = (1 + 2\lambda)^{-1}s.$$

This leads to the optimal portfolio weight on r_t:[35]

$$\widehat{w}_q(\lambda) \propto \mathbb{M}_\beta \breve{w}_q(\lambda), \quad q = 1, 2.$$

Depending on the magnitude of λ, the Lasso approach replaces all smaller signals (i.e., $\widehat{s}_i$) by zero and shrinks the larger ones by λ in absolute terms. In other words, the Lasso approach is the soft-thresholding alternative to the B-H method. In contrast, the ridge penalty shrinks all signals proportionally. Since proportional scaling of portfolio weights does not affect the Sharpe ratio, this means that ridge is equivalent to CSR! This "embedded" shrinkage effect of CSR explains why it performs well in the case of small signals.

Proposition B4, along with Proposition B2 in the online appendix, offers explicit formulae

[35] An alternative strategy is to impose sparsity directly on the portfolio weights with respect to raw returns. While this approach might be appealing from the transaction cost point of view, it does not associate with an explicit prior on alpha, hence is more difficult to interpret.

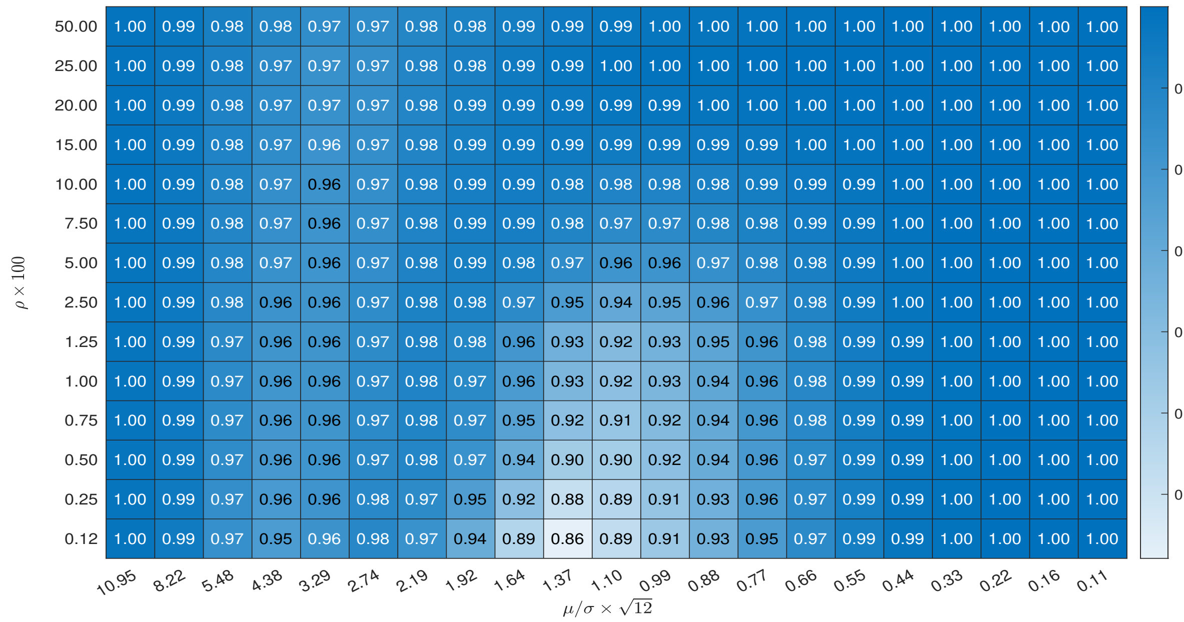
for the optimal feasible Sharpe ratios in the shrinkage case. The Sharpe ratio of ridge is not affected by the tuning parameter, but Lasso's performance is contingent on its tuning parameter. However, even with the optimal choice of tuning parameter, it cannot achieve the optimal feasible Sharpe ratio in all DGPs.

Figure 5 compares $S_{\lambda}^{\text{Lasso}}$ with S^{OPT}, where $S_{\lambda}^{\text{Lasso}}$ denotes the theoretical Sharpe ratio achieved by Lasso for a given tuning parameter choice λ. In practice, the tuning parameter choice would require a cross-validation procedure. Here we adopt the theoretically optimal tuning parameter that maximizes $S_{\lambda}^{\text{Lasso}}$ in Figure 5. This strategy, unlike the B-H approach where the tuning parameter τ is selected based on a clear statistical criterion, is not fully feasible in practice, since the optimal tuning depends on the DGP, which is not fully known. Hence, Figure 5 should be understood as showcasing the optimal performance of a class of trading strategies indexed by the tuning parameter of Lasso, and only reports the best performance within this class. This is a fair comparison to OPT, which represents the optimal performance of all strategies, beyond the Lasso class, as the implementation of our algorithm to achieve S^{OPT} would also require choice of a tuning parameter. Although Proposition B4 suggests that Lasso is not uniformly optimal, it performs remarkably well, achieving the optimal Sharpe ratio in almost all regimes. Intuitively, when signals are very strong, Lasso behaves like a hard-thresholding selector, as shrinkage has minimal effect. Conversely, when signals are week, ridge (and hence CSR) approach optimality, and Lasso can replicate their performance by setting the tunning parameter close to zero.

3 Simulation Evidence

This section illustrates the applicability of our theory through simulations and assesses the finite sample performance of our proposed portfolio strategies.

For simplicity and clarity, we simulate a one-factor model of returns as specified by (1). We set the factor risk premium at 5% per annum with an annualized volatility of 25%. The distribution of beta across assets is modeled as a normal distribution with a mean and variance of one. Given our focus on arbitrage portfolios, the parameters of the factor component (including the number of factors) are inconsequential, because factors are eliminated by $\mathbb{M}_{\beta}$ in the initial steps of constructing these trading strategies. In addition, we use the model in (5) from Example 1 for the cross-sectional distribution of alpha. We maintain a constant idiosyncratic volatility, denoted as σ for all assets, because the ratio α/σ determines signal strength, and so there is no need to vary both α and σ in the cross section. While our Monte Carlo experiment is stylized, it effectively demonstrates the impact of estimation errors in factors and volatilities and the finite sample performance of our theoretical predictions.



$\rho \times 100$	10.95	8.22	5.48	4.38	3.29	2.74	2.19	1.92	1.64	1.37	1.10	0.99	0.88	0.77	0.66	0.55	0.44	0.33	0.22	0.16	0.11
50.00	1.00	0.99	0.98	0.98	0.97	0.97	0.98	0.98	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25.00	1.00	0.99	0.98	0.97	0.97	0.97	0.98	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20.00	1.00	0.99	0.98	0.97	0.97	0.97	0.98	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15.00	1.00	0.99	0.98	0.97	0.96	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10.00	1.00	0.99	0.98	0.97	0.96	0.97	0.98	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00
7.50	1.00	0.99	0.98	0.97	0.96	0.97	0.98	0.99	0.99	0.98	0.97	0.97	0.98	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00
5.00	1.00	0.99	0.98	0.97	0.96	0.97	0.98	0.99	0.98	0.97	0.96	0.96	0.97	0.98	0.98	0.99	1.00	1.00	1.00	1.00	1.00
2.50	1.00	0.99	0.98	0.96	0.96	0.97	0.98	0.98	0.97	0.95	0.94	0.95	0.96	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00
1.25	1.00	0.99	0.97	0.96	0.96	0.97	0.98	0.98	0.96	0.93	0.92	0.93	0.95	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00
1.00	1.00	0.99	0.97	0.96	0.96	0.97	0.98	0.97	0.96	0.93	0.92	0.93	0.94	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00
0.75	1.00	0.99	0.97	0.96	0.96	0.97	0.98	0.97	0.95	0.92	0.91	0.92	0.94	0.96	0.98	0.99	0.99	1.00	1.00	1.00	1.00
0.50	1.00	0.99	0.97	0.96	0.96	0.97	0.98	0.97	0.94	0.90	0.90	0.92	0.94	0.96	0.97	0.99	0.99	1.00	1.00	1.00	1.00
0.25	1.00	0.99	0.97	0.96	0.96	0.98	0.97	0.95	0.92	0.88	0.89	0.91	0.93	0.96	0.97	0.99	0.99	1.00	1.00	1.00	1.00
0.12	1.00	0.99	0.97	0.95	0.96	0.98	0.97	0.94	0.89	0.86	0.89	0.91	0.93	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00

Figure 5: Ratios between S_λ^{Lasso} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the Lasso based portfolio and the feasible optimal arbitrage portfolio. The theoretical Sharpe ratio achieved by Lasso is treated as the maximum of S_λ^{Lasso} over all λ, given explicitly by Proposition B4. The simulation setting is based on model (5), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$. The tuning parameter λ is selected to maximize S_λ^{Lasso}.

We compare the finite sample performance of our portfolio estimators across various DGPs. We examine a broad spectrum of signal strength (μ/σ) and sparsity (ρ) values to cover a range of potential empirical scenarios. For each set of parameters $(\mu/\sigma, \rho)$, we simulate the corresponding DGP, construct portfolio weights, $\widehat{w}^{\text{A}}$, where A represents OPT, CSR, BH, or Lasso, and compute the corresponding theoretical Sharpe ratio: $\widehat{S}^{\text{A}} = \widehat{w}^{\text{A}\intercal}\mu / \sqrt{\widehat{w}^{\text{A}\intercal}\Sigma_u^{-1}\widehat{w}^{\text{A}}}$.

Our algorithm requires a tuning parameter. We choose it with a validation procedure that divides the in-sample data into two parts: the first 80% is used for training, and the remaining 20% is reserved for validation. The optimal parameter maximizes the sample Sharpe ratio in this validation subset. Once identified, this tuning parameter is then used with the entire in-sample (training and validation) data to estimate portfolio weights. We use the same validation approach to determine the tuning parameter for the Lasso method.[36]

In light of Theorem 3, the following performance measure is sensible:

$$\text{Gap}^{\text{A}}(\mu/\sigma, \rho) = \widehat{\text{E}}\left(|\widehat{S}^{\text{A}} - S^{\text{OPT}}| / (1 + S^{\text{OPT}}) \right),$$

[36] The specific grid we use for tuning our method consists of the set $\{0.25, 0.5, \ldots, 4\} \times (\log N)^{-1}$. For the Lasso method, the parameter λ is selected from the set $\{2^{-10}, 2^{-9}, \ldots, 2^{-1}\}$. For our method, theoretical guidance from Theorem 3 helps — we know a tuning parameter $\simeq (\log N)^{-1}$ would be optimal. For Lasso, however, the optimal rate depends on the DGP, and we lack simple theoretical results to guide us.

where the explicit dependence of $\widehat{S}^{\mathrm{A}}$ and S^{OPT} on μ/σ and ρ is not specified, and $\widehat{\mathrm{E}}(\cdot)$ denotes the sample average computed over Monte Carlo simulations. This formula measures error as relative percentages of S^{OPT} when S^{OPT} is significantly large (i.e., significantly greater than 1). Conversely, when S^{OPT} is small (i.e., $o_{\mathrm{P}}(1)$), the error is gauged in absolute terms. This approach ensures meaningful error measurment if S^{OPT} diverges or diminishes depending on the parameters involved.

To evaluate the asymptotic behavior of various estimators, we also report a measure of the estimation error:

$$\mathrm{RMSE}^{\mathrm{A}}(\mu/\sigma,\rho)=\left(\left(\widehat{\mathrm{E}}(\widehat{S}^{\mathrm{A}}-S^{\mathrm{A}})/(1+S^{\mathrm{OPT}})\right)^2+\widehat{\mathrm{Var}}\left(\widehat{S}^{\mathrm{A}}/(1+S^{\mathrm{OPT}})\right)\right)^{1/2},$$

where $\widehat{\mathrm{Var}}(\cdot)$ denotes the sample variance across Monte Carlo repetitions.

Table 1 presents the maximum performance gap and estimation error across all values of μ/σ and ρ. The results consistently show that OPT has the smallest gap to optimal performance in comparison to CSR, BH, or Lasso. Lasso ranks second, while CSR and BH exhibit significantly poorer performance. Notably, as T increases from 2 years to 20 years, the maximum gap of OPT decreases from 0.381 to 0.152 for $N=1,000$, and from 0.276 to 0.127 for $N=3,000$. In contrast, as N increases, the performance of the latter three methods tends to deteriorate. This trend can be attributed to an increasing theoretical performance gap, quantified as $|S^{\mathrm{A}}-S^{\mathrm{OPT}}|/(1+S^{\mathrm{OPT}})$, that plays a dominating role. Indeed, the bottom panel of the table illustrate that the estimation error asymptotically diminishes with increasing N.

Finally, Figure 6 reports the sample average of estimation error $\left|\widehat{S}^\star-S^\star\right|/\left(1+S^\star\right)$ over Monte Carlo simulations for each value of μ/σ and ρ. The results confirm the consistency result in Proposition 1. The relative error remains small when $S^\star$ is large or moderate (significantly greater than 1). Conversely, near the bottom right corner of Figure 6, where $S^\star$ is essentially zero as shown by Figures 1 and 2, the estimation proves more challenging. In this scenario, the error, becoming absolute ($S^\star << 1$), is notably larger due to a significant upward bias that is difficult to mitigate in finite sample.

4 Empirical Analysis of US Equities

To demonstrate the empirical relevance of the statistical limit of arbitrage, we analyze US monthly equity returns. Our analysis is divided into two parts: the first focuses on individual equity returns, and the second utilizes portfolios as test assets. Employing portfolios allows us to explore scenarios where alphas are linear functions of characteristics, aligning our approach with the common empirical practice of predicting alphas based on such characteristics. We

	$N = 1,000$				$N = 3,000$			
	$T = 2$	$T = 5$	$T = 10$	$T = 20$	$T = 2$	$T = 5$	$T = 10$	$T = 20$
	$\sup_{\mu/\sigma,\rho} \text{Gap}^{\text{A}}(\mu/\sigma, \rho)$							
OPT	0.381	0.243	0.171	0.152	0.276	0.228	0.143	0.127
CSR	0.622	0.586	0.532	0.477	0.462	0.604	0.610	0.562
BH	0.739	0.759	0.743	0.704	0.781	0.822	0.814	0.791
Lasso	0.396	0.259	0.176	0.154	0.298	0.265	0.159	0.172
	$\sup_{\mu/\sigma,\rho} \text{RMSE}^{\text{A}}(\mu/\sigma, \rho)$							
OPT	0.521	0.445	0.441	0.446	0.333	0.346	0.273	0.236
CSR	128	0.165	0.211	0.264	0.093	0.075	0.086	0.105
BH	0.492	0.436	0.440	0.442	0.359	0.347	0.263	0.248
Lasso	0.521	0.435	0.439	0.444	0.336	0.356	0.302	0.241

Table 1: Simulation Results

Note: The top panel displays the maximum Sharpe Ratio gap, defined as $\sup_{\mu/\sigma,\rho} \text{Gap}^{\text{A}}(\mu/\sigma, \rho)$, across all values of μ/σ and ρ shown in Figure 1. Similarly, the lower panel details the maximum root-mean-squared error, defined as $\sup_{\mu/\sigma,\rho} \text{RMSE}^{\text{A}}(\mu/\sigma, \rho)$. Here, A represents the methods OPT, CSR, BH, or Lasso for various combinations of N and T (measured in years). The OPT and Lasso methods employ a validation procedure to determine the optimal tuning parameters. The BH method ensures control of the false discovery rate at a 5% level.

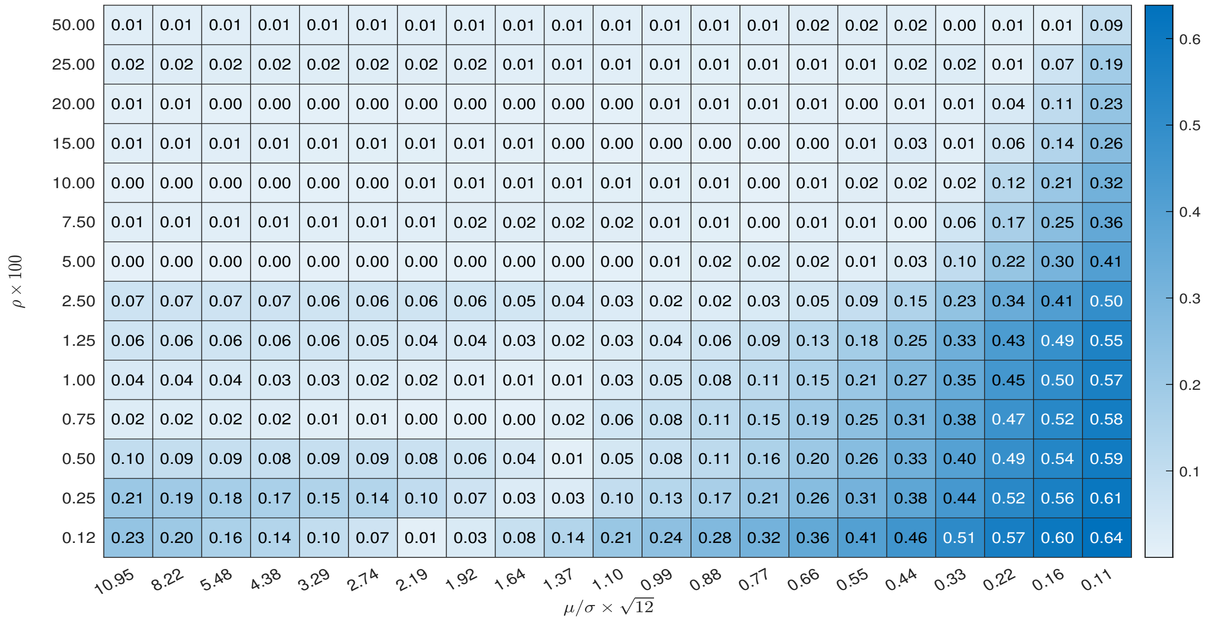


Figure 6: Comparison between $\widehat{S}^{\star}$ and $S^{\star}$

Note: The figure reports the sample average of $\left|\widehat{S}^{\star} - S^{\star}\right| / \left(1 + S^{\star}\right)$ over Monte Carlo repetitions. The simulation setting is based on model (5), in which a $100 \times \rho\%$ of assets have αs that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$. In this experiment, $N = 1,000$ and $T = 20$ years.

begin our discussion with a description of the datasets.

4.1 US Equity Data

Our monthly equity sample covers the period from January 1965 to December 2020. For individual equities, we utilize a multi-factor model with observable factor loadings, closely resembling the widely used MSCI Barra model. Specifically, we use 16 characteristics and 11 GICS sectors, drawing on both empirical insights from existing asset pricing literature and industry practice. The selected characteristics include market beta, size, operating profits/book equity, book equity/market equity, asset growth, momentum, short-term reversal, industry momentum, illiquidity, leverage, return seasonality, sales growth, accruals, dividend yield, tangibility, and idiosyncratic risk, which are downloaded directly from the website openassetpricing.com (Stock-level Signal Datasets August 2023 Release). Details on the construction of these characteristics can be found in Chen and Zimmermann (2020).

We download monthly return data for individual equities from CRSP and apply several preprocessing steps. First, in the case of delistings, we use the delisting return as the final return in the delisting month. Next, we merge the returns data with the aforementioned characteristics database using permnos, which yields an average of 6,540 unique permnos per month. We then apply standard filters (require share codes 10 and 11, and exchange codes 1, 2, and 3) to refine the dataset. This process selectively removes returns for certain months from stocks that fail to meet these criteria during those periods. After applying these filters, the average number of stocks remaining per month totals 4,756.

We address missing characteristics in our dataset with an approach that avoids forward-looking bias. If GICS codes are missing, we use the most recent records available before the data is missing, but we do not copy records backwards in time. Observations that do not have a GICS code after this procedure are excluded, which predominantly affects records before 1990. With GICS codes in place, we implement a two-step procedure to address other missing characteristics. We fill each missing characteristic with the sector-wise median for that characteristic each month. If a characteristic's values are missing for an entire sector in a given month, we use the cross-sectional median from all stocks for that month. After these preprocessing steps, the average number of stocks per month is reduced to 4,095.

Characteristics-sorted portfolio returns are directly downloaded from the website openassetpricing.com under the "Portfolio Return Datasets" section. The dataset comprises 1,322 portfolios, spanning the years from 1965 to 2020. The portfolios are long-only and sorted based on 212 characteristics, including 49 industry indicators. The number of portfolios per characteristic varies depending on the specifications in the original papers that introduced each characteristic. Additionally, some portfolios may contain missing values due to a lack of observations in certain categories. We use value-weighted portfolios to avoid high exposure to illiquid small stocks.

4.2 Analysis of Individual Equity Returns

We consider a specific version of the model as defined in (2), setting α_t as constant and β_t as observable. Each month, we regress next month's returns on the 27 predictors, using all stocks present in the current month's cross-section, and including an intercept. We normalize the 16 characteristics within each cross-section to mitigate the impact of extreme outliers. This normalization transforms the characteristics to follow a normal distribution.[37]

Figure 7 plots the time series of the cross-sectional regression R^2s over time. The R^2 has declined since the beginning of the sample until the 1990s. This coincides with an increase in the number of stocks in US equity markets. The R^2s are moderately low, with an average of 8.07%, which suggests that a substantial portion of cross-sectional variation of individual equity returns is idiosyncratic. Therefore, learning alphas from residuals of the factor model is a difficult statistical task.

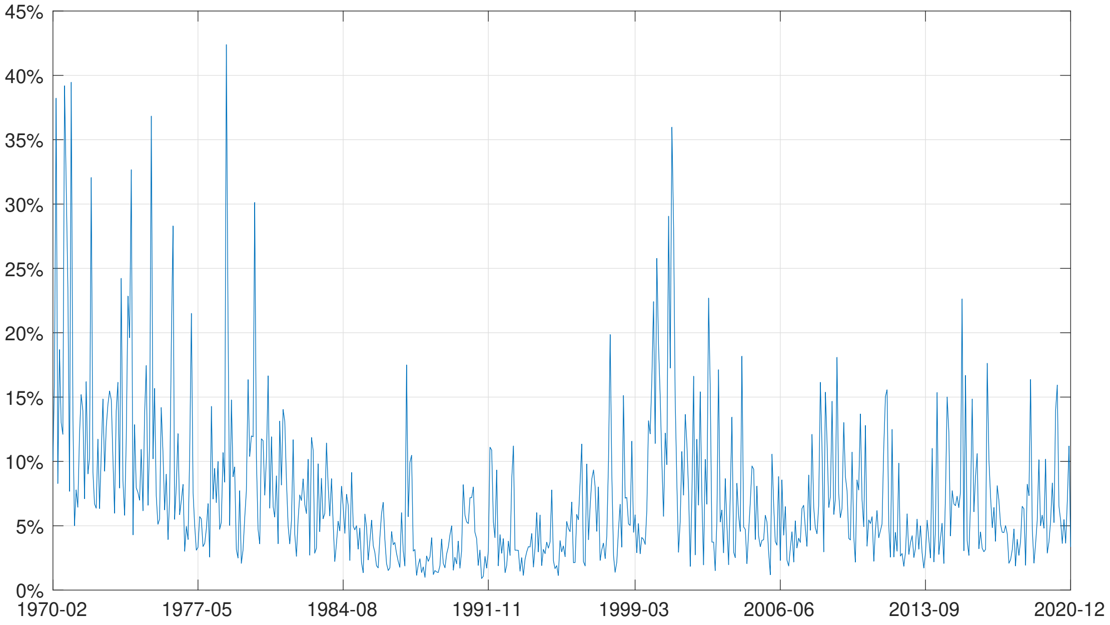


Figure 7: Time-series of the Cross-sectional R^2s

Note: The figure provides a time series of the cross-sectional R^2s for individual equity returns, derived from regressing next month's returns against 27 firm-level characteristics in cross-sectional regressions.

[37] For each characteristic $c_{i,t}$, the normalization applies the functional form $\Phi^{-1}(\text{rank}(c_{i,t}))$, where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution. Compared to uniform normalization, this approach leads to slightly improved cross-sectional R^2 values, thereby providing a more accurate representation of realized returns. It is important to note, however, that the specific form of normalization does not significantly impact subsequent results on arbitrage Sharpe ratios.

4.2.1 Rare and Weak Alphas

We now study the statistical properties of individual equity alphas using the full sample data. For each stock, we collect its regression residuals and take their average as an estimate for its alpha. We require at least 60 observations. This ensures a sufficiently large sample size for inference on alpha, although the empirical distribution of alphas' t-statistics turns out be insensitive to this requirement. Figure 8 provides histograms of the t-statistics and Sharpe ratios for alphas of all 12,734 stocks in our sample that meet this criterion. Because these stocks have different sample sizes, the histograms of the Sharpe ratios are not simply the scaled version of the histogram of the t-statistics.

Only 7.58% of the t-statistics exceed 2.0 in magnitude, and more than 1.12% exceed 3.0. This suggests that truly significant alphas are extremely rare. Moreover, the largest Sharpe ratio among all individual stocks' alphas is approximately 1.86. Notably, only 0.71% of the alphas have a Sharpe ratio greater than 1.0. These summary statistics suggest that rare and weak alpha is perhaps the most relevant scenario in practice.

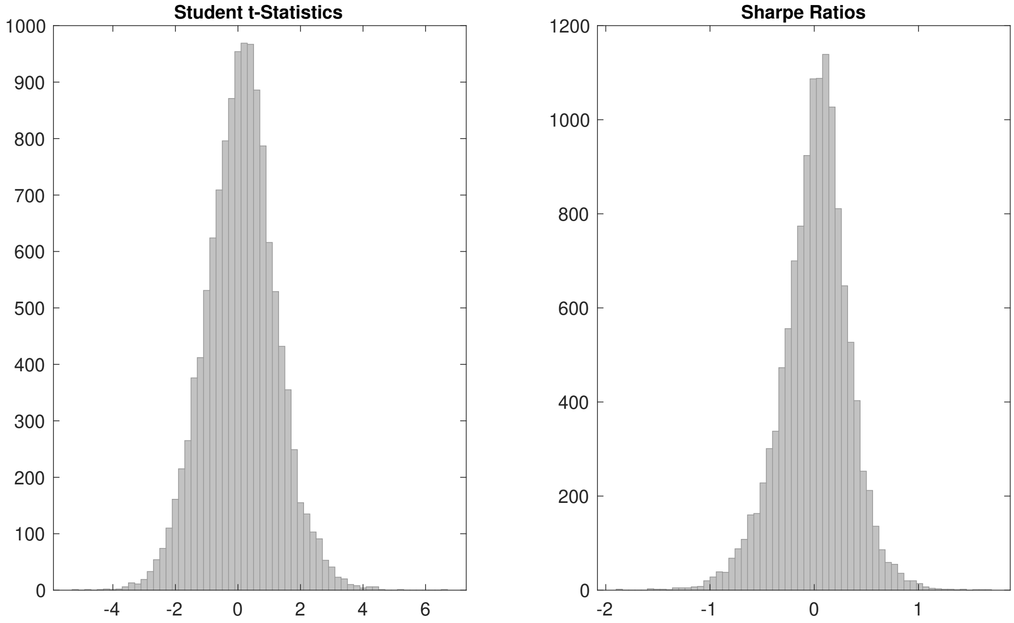


Figure 8: Histograms of the t-Statistics and Sharpe Ratios of Estimated Alphas

Note: The figure provides the histograms of the Student t-statistics (left) and Sharpe ratios (right) of estimated alphas for all tickers in our sample with at least 60 months of data. The total number of tickers available is 12,734.

Alphas are meaningless without reference to a specific factor model. While our analysis includes only 27 firm-level characteristics, constructing a factor model with additional characteristics would transform "alpha" into risk premia. Put it differently, extracting additional "factors" from returns would result in even rarer and weaker alphas.

4.2.2 Modest Feasible vs. Large Infeasible Sharpe Ratios

We now compare arbitrage portfolios based on various strategies discussed in Section 2.8. At the end of each month, we construct optimal portfolio weights using these strategies. These weights are estimated monthly with a 60-month rolling window, and the portfolios are rebalanced accordingly.[38] Both Lasso and OPT methods require a tuning parameter, which is selected annually. For this purpose, the final year of the rolling window is set aside as the validation sample to optimize the tuning parameter selection.

All these strategies yield modest Sharpe ratios. OPT reaches the top of the chart, yielding 0.82, followed by CSR at 0.70. The BH and Lasso approaches obtain Sharpe ratios of 0.65 and 0.62, respectively. To compare cumulative returns, we normalize all strategies to have the same (ex-post) volatility. The resulting time-series of normalized cumulative returns are shown in Figure 14. Notably, most returns are generated during the late 1980s to early 2000s. After 2005, the cumulative returns of all strategies tend to plateau, indicating a decrease in profitability of statistical arbitrage.

A detailed examination of these strategies offers further insights. The BH strategy is highly conservative; over 51 years of out-of-sample trading (from January 1970 to December 2020), there are 284 months with zero holdings. The maximum number of stocks held in any month is 32, with an average of only 6 stocks during months when holdings are non-zero. In contrast, the CSR and OPT strategies hold all stocks that meet the sample criteria, averaging 2,833 stocks per month. The number of stocks held by the Lasso strategy is notably volatile, ranging from none to almost all stocks in a given month, with an average of 722 stocks per month. This volatility reflects the underlying weakness in their alphas, whose estimates are prone to fluctuate with different samples.

In contrast to the simulation study where performance of strategies are evaluated over a large class of DGPs, results presented in the empirical analysis here all originate from a single DGP, i.e., the one that generated the empirically observed stock returns. Given the strong performance of CSR and the weaker performance of BH, it is likely that the alphas in reality are rare and weak to the extent that CSR is rather close to the optimal strategy. In this case, Lasso could only perform equally well as CSR if we were able to select the optimal tuning, i.e., zero, (no selection or shrinkage), which cross-validation cannot reliably achieve in all rolling windows. If empirically observed returns were generated by a DGP that made CSR suboptimal in many rolling windows, Lasso would have a chance to outperform, given its ability to achieve optimality within a larger class of DGPs.

We also estimate the infeasible Sharpe ratios using (18) within the same 60-month rolling

[38] Empirical results based on 36- and 84-month windows are reported in Table A2 in the appendix for robustness checks.

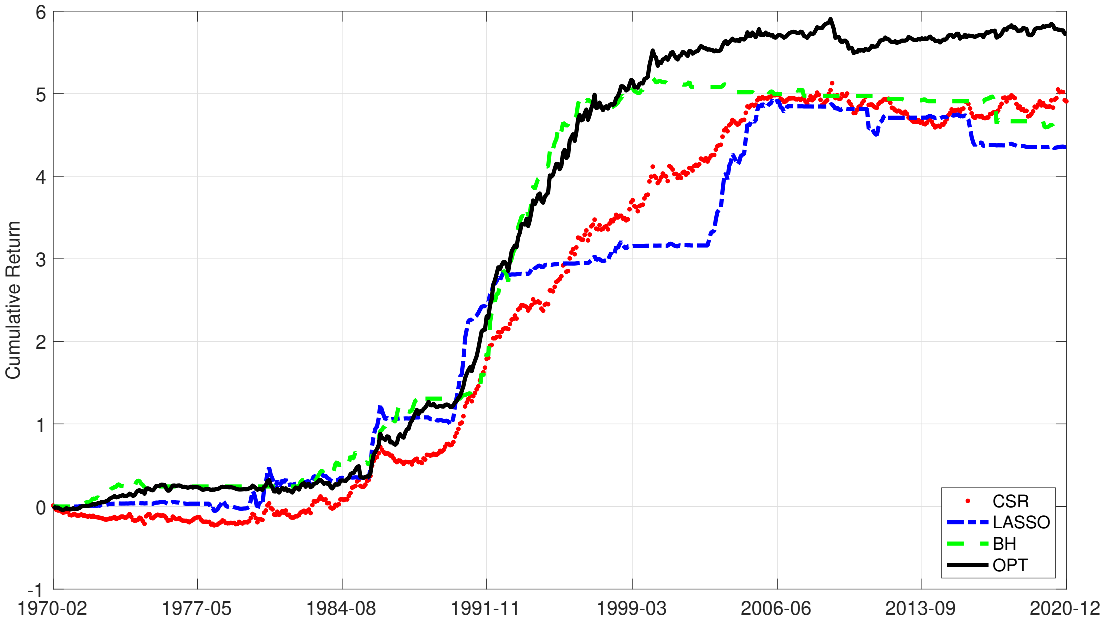


Figure 9: Normalized Cumulative Returns of Arbitrage Portfolios

Note: This figure compares the cumulative returns of OPT (black solid), CSR (red dotted), BH (green dashed), and Lasso (blue dot-dashed) strategies. We normalize all returns by their realized volatilities calculated by the square root of the sum of the squared returns over the entire sample, only for comparison purpose.

window. Our estimate, $\widetilde{S}^{\star}$, averages 4.81 (with negative estimates truncated at 0), but it can occasionally exceed 16.0. These estimates are much higher than the feasible Sharpe ratios ($0.6 \sim 0.8$) we obtain for any of these strategies. This supports our theoretical prediction that the impact of statistical learning alone can eliminate a substantial portion of investment opportunities.

4.3 Analysis of Portfolio Returns

When portfolios are chosen as test assets, we opt for latent factor models due to the absence of natural choices of their factors or factor exposures. This approach aligns with the return generating process presented in (1) but with unknown β and v_t. Our factor model estimation step follows Algorithm 4 outlined in Giglio et al. (2021). We estimate this model using a 60-month rolling window, consistent with the main empirical analysis above.[39] Using the estimated factor loadings from each rolling window, we then conduct a cross-sectional regression of the average returns over this window against these loadings, augmented with a column of ones, to determine the risk premia of the latent factors, the zero-beta rate, and

[39] Results based on alternative 3-year and 7-year windows, which show similar outcomes, are reported in Tables A1 and A2 in the appendix for reference.

the residuals, which serve as estimates for α's. It is important to note that these loadings are identifiable only up to a rotation; hence, we report only those results that are invariant to such rotations, for instance, R^2 values and summary statistics for α's.

4.3.1 Portfolio Alphas

These latent factor models demonstrate substantial in-sample cross-sectional explanatory power for expected returns,[40] accounting for an average of approximately 48% of the cross-section variation in average returns across all rolling windows, with 10 factors.[41] This results in notably smaller Sharpe ratios for the alphas compared to the portfolios' own Sharpe ratios. Figure 10 contrasts the histograms of their respective Sharpe ratios. Notably, the distribution of the alphas' Sharpe ratios in darker gray exhibits thinner tails compared to those of the portfolios' Sharpe ratios in lighter gray.

We then apply the same algorithms to the portfolios' alphas that we previously used for individual stocks. The normalized cumulative returns are displayed in Figure 11. All strategies earn substantially higher Sharpe ratios than those achieved with individual stocks. Notably, the OPT strategy once again leads, yielding a Sharpe ratio of 1.71, closely followed by CSR at 1.67. Meanwhile, the BH and Lasso strategies register Sharpe ratios of 1.49 and 1.18, respectively.

Next, we visualize the average portfolio weights over time using a heatmap in Figure 12. The portfolios in this figure are sorted by the average weight they receive with the OPT method, with the highest positive weights on top and the biggest negative weights at the bottom. The figure shows that both OPT and CSR exhibit similar allocation patterns, with non-zero weights assigned to many portfolios. However, OPT, which applies shrinkage to portfolio alphas, tends to assign smaller weights than CSR. Due to the absence of shrinkage, CSR has consistently higher weights than all other methods. As previously mentioned, CSR achieves the same Sharpe ratio as ridge, because uniform shrinkage of all assets' alphas does not affect the Sharpe ratio. BH selects a narrower range of portfolios, most of which have zero or near-zero weights, indicating a highly selective strategy. Lasso, although less conservative than BH, also shows lighter colors on the heatmap, reflecting a significant level of selection and shrinkage in the portfolio weights.

Among all characteristics, those assigned higher positive weights by these methods include portfolios sorted by dividend seasonality identified by Hartzmark and Solomon (2013), price as discussed in Blume and Husic (1973), intangible return from Daniel and Titman (2006),

[40] This is different from the cross-section R^2 with observable characteristics reported for individual equities, which represents the (in-sample) explanatory power for realized returns.

[41] Results for models with varying numbers of factors are included in Tables A1 and A2 in the appendix for comparison.

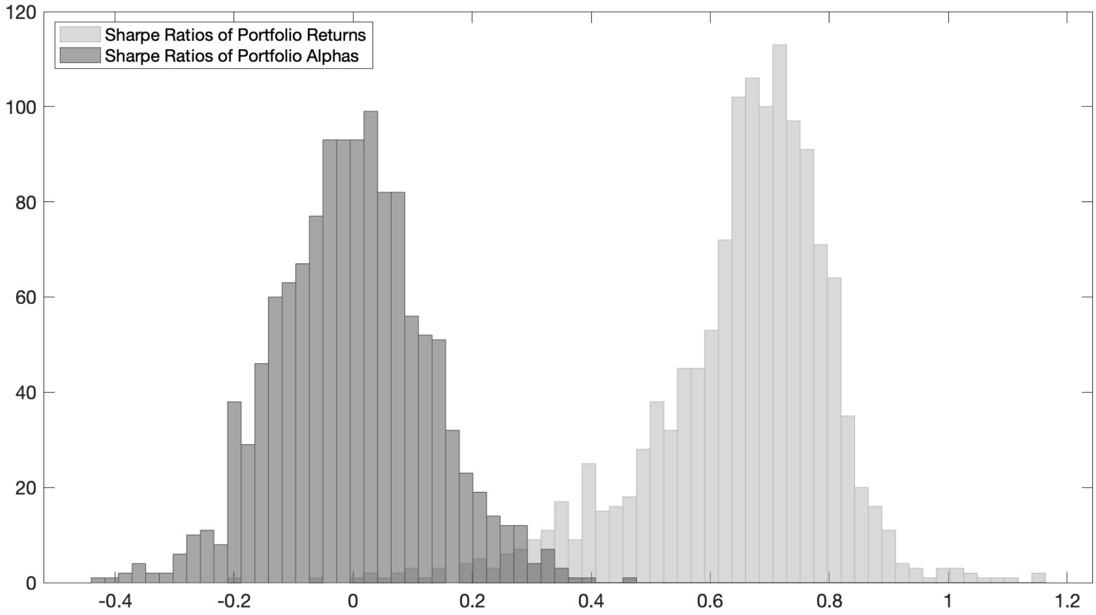


Figure 10: Sharpe Ratios of Portfolio Returns and Alphas

Note: This figure compares the histograms of the Sharpe ratios of portfolios' returns and their alphas.

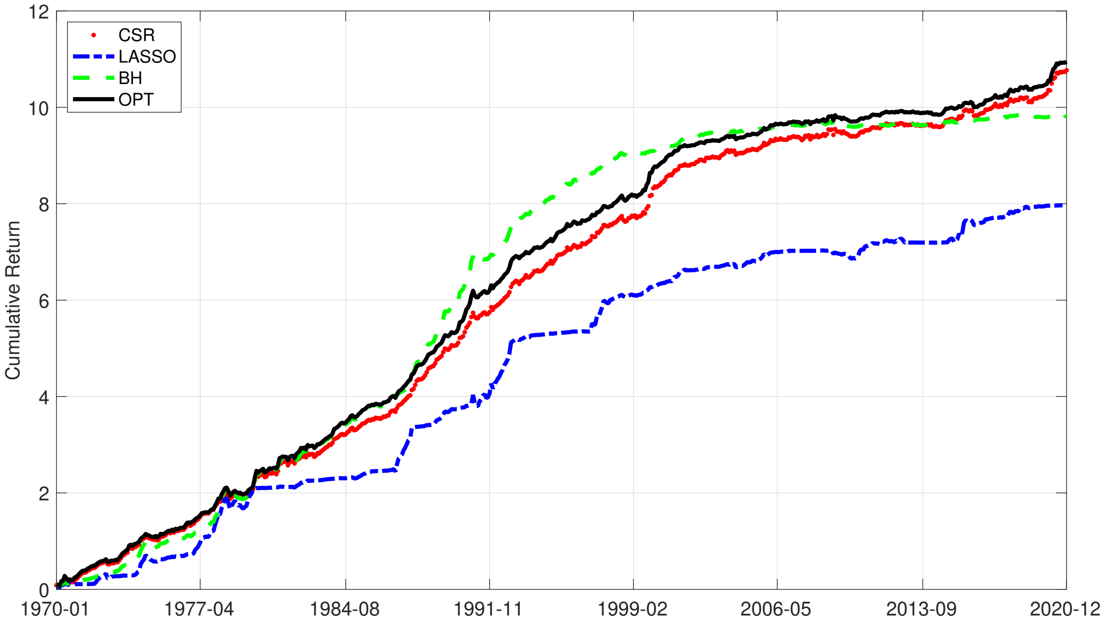


Figure 11: Normalized Cumulative Returns of Arbitrage Portfolios

Note: This figure compares the cumulative returns of OPT (black solid), CSR (red dotted), BH (green dashed), and Lasso (blue dot-dashed) strategies. We normalize all returns by their realized volatilities calculated by the square root of the sum of the squared returns over the entire sample, only for comparison purpose. The base assets for each strategy are all portfolios sorted by characteristics.

gross profits to total assets analyzed by Novy-Marx (2013), and industry return of big firms

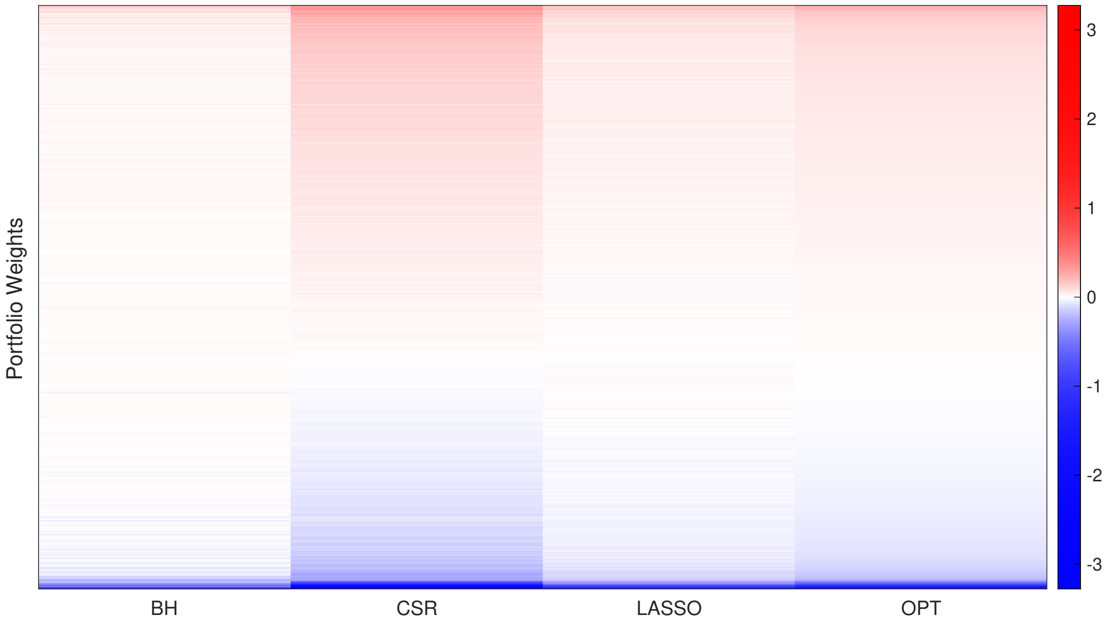


Figure 12: Average Weights Assigned to Portfolios by Different Methods

Note: This figure displays a heatmap that shows the average portfolio weights over time for portfolios sorted by characteristics, formed using BH, CSR, Lasso, and OPT methods.

studied in Hou (2007). Conversely, the characteristics with the most negative weights include portfolios sorted by spinoffs documented by Cusatis et al. (1993), past trading volume from Brennan et al. (1998), initial public offerings from Ritter (1991), size from Dharan and Ikenberry (1995), and exchange switch as in Banz (1981).

4.3.2 Accounting for Publication Effects

Many of the characteristics used in the construction of the portfolio return data set were introduced to the academic literature later in the sample period. For this reason, the above analysis may overstate achievable Sharpe ratios. If arbitrageurs were not aware of the predictive power of these characteristics in the early parts of the sample, these Sharpe ratios were not truly achievable. Moreover, academics' discovery of the predictive power of these characteristics could potentially reflect ex-post selection bias, rendering the in-sample predictive power spurious. To address this issue, we revise our portfolio analysis to only include characteristics from the year following their publication. Except for the 49 industry portfolios (assumed) available since 1965, few portfolios were introduced in earlier years. As illustrated in Figure 13, a significant number of these characteristics were introduced after 2000.

With this adjusted approach, the Sharpe ratios for the OPT and CSR are now substantially lower at 0.61 and 0.66, respectively. These magnitudes align more closely with the Sharpe ratios we obtained from individual stocks. The Lasso and BH strategies now achieve only 0.45 and 0.35, respectively. Consistent with McLean and Pontiff (2016), this suggests that some of the initial discoveries of signals may have been "lucky," performing well in-sample but then failing to replicate this success out-of-sample, or that competition among arbitrageurs shrinks the alphas once the discoveries become public.

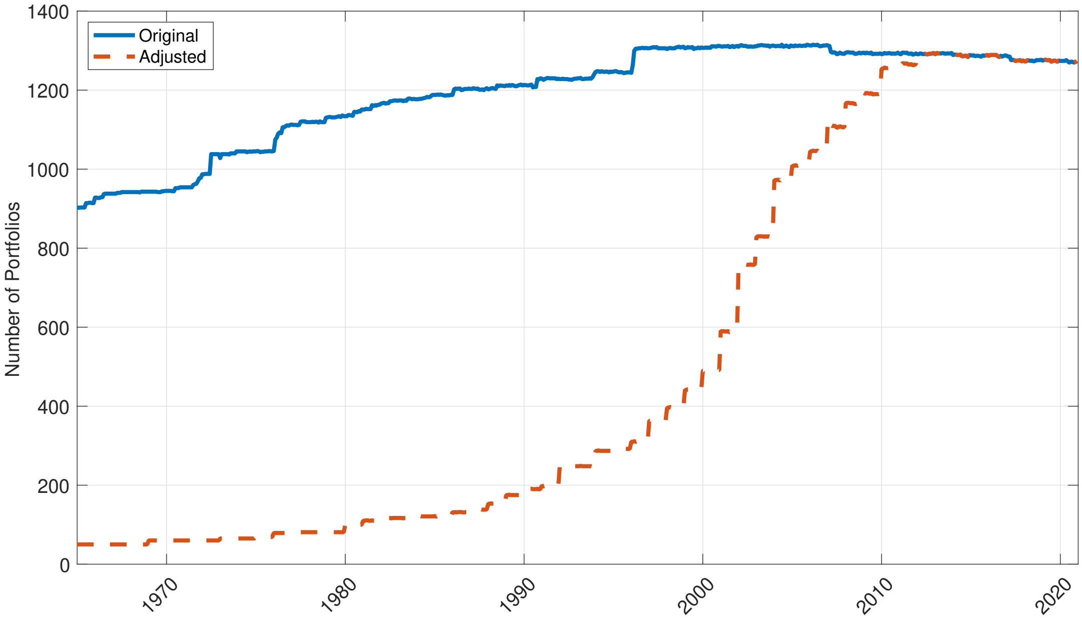


Figure 13: Number of Portfolios Before and After Adjusting for Publication Years

Note: This figure plots two curves: one showing the total monthly count of characteristics-sorted portfolios and the other displaying the count adjusted for only those portfolios whose characteristics have been previously published.

Figure 15 shows infeasible Sharpe ratios estimated both before and after adjusting for publication effects. Before this adjustment, the infeasible Sharpe ratio is consistently above 20, but after this adjustment, including characteristics only after their publication year, it is much lower in the early part of the sample, starting around 5 in the early 1970s. As more signals are discovered over time, the infeasible Sharpe ratio increases, eventually reaching levels similar to those estimated without adjustment.

Despite these adjustments, a significant gap still exists between the theoretical infeasible Sharpe ratios that range from 5 to 20 and the achievable Sharpe ratios, which are smaller than 0.7. This highlights the quantitative relevance of statistical limits of arbitrage in real-world scenarios. The large reduction in Sharpe ratios suggests that statistical limits may play an even more critical role than publication effects for characterizing the investment opportunities that are actually feasible for arbitrageurs.

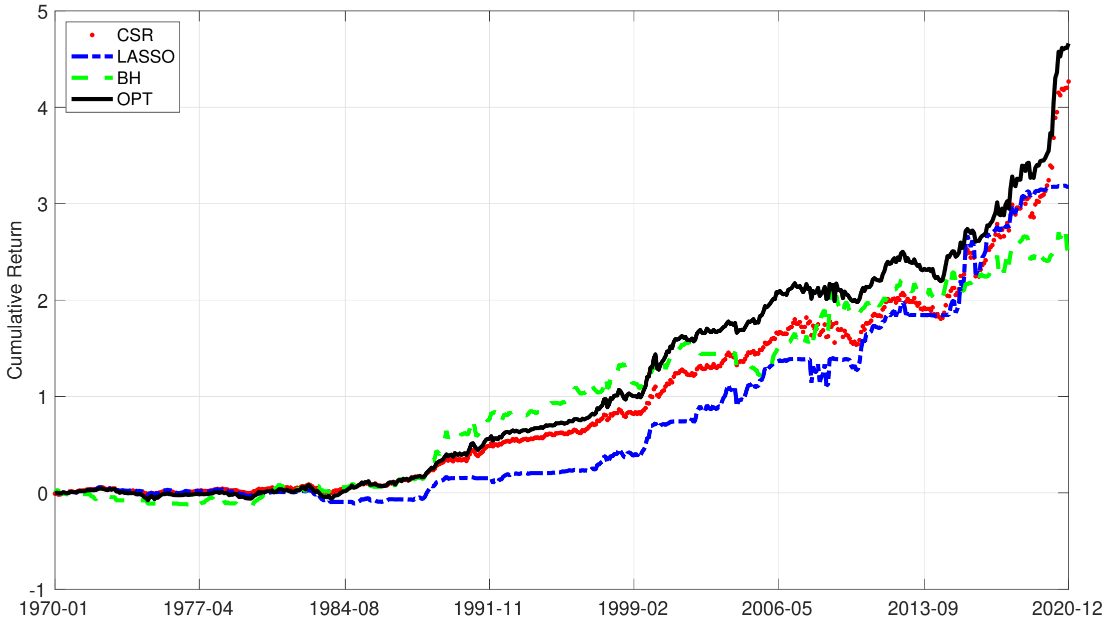


Figure 14: Normalized Cumulative Returns of Arbitrage Portfolios

Note: This figure compares the cumulative returns of OPT (black solid), CSR (red dotted), BH (green dashed), and Lasso (blue dot-dashed) strategies. We normalize all returns by their realized volatilities calculated by the square root of the sum of the squared returns over the entire sample, only for comparison purpose. The base assets only include portfolios sorted by characteristics that were previously published.

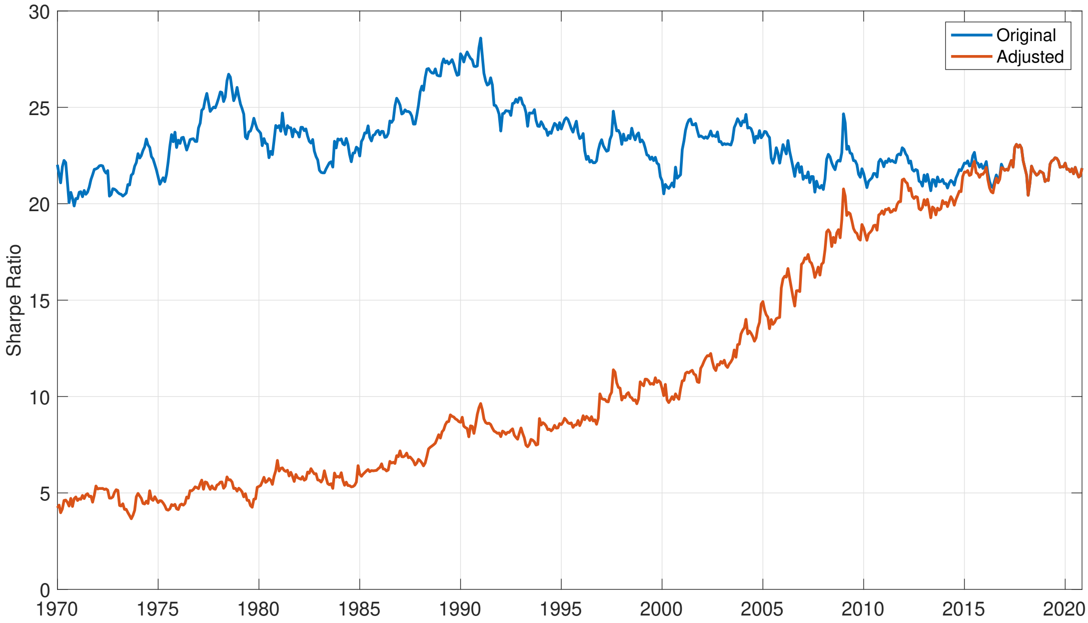


Figure 15: Time Series Plots of Infeasible Sharpe Ratios

Note: This figure compares the Infeasible Sharpe ratios on the basis of portfolios before and after adjusting for publication years.

5 Conclusion

Taking stock, our paper provides a new theoretical framework for understanding the implications of statistical learning in asset pricing. Rather than endowing arbitrageurs with perfect knowledge of the economic model, we expose them to the challenge of learning about alphas. Arbitrageurs in reality arguably face a high-dimensional world, with many assets and many alpha signals, which can make statistical learning difficult. These difficulties give rise to statistical limits of arbitrage, which manifest as a gap between the infeasible Sharpe ratio that arbitrageurs could earn with perfect knowledge of alphas and the feasible Sharpe ratio that arbitrageurs attain in equilibrium with statistical learning.

While high-dimensional, the linear factor model environment that arbitrageurs face in our analysis is still quite simple. The gap between feasible and infeasible Sharpe ratios would further increase if arbitrageurs faced additional statistical challenges, e.g., model misspecification, omitted factors, weak factors, or a large non-sparse idiosyncratic covariance matrix.

Our approach to characterizing the feasible Sharpe ratio is in important ways different from other analyses of statistical learning in asset pricing. Many papers applying machine learning methods in the construction of trading strategies have documented impressive Sharpe ratios. Such strategies often rely on ad-hoc model design (e.g., a neural network with a specific architecture) and tuning parameters selection. In this regard, the empirical analysis can at best provide a "lower bound" on the performance of machine learning strategies. Our paper provides a theoretical framework to understand the "upper bound" on the performance of any statistical learning strategy in a specific context. The decision rule of arbitrageurs in turn ties this upper bound to the equilibrium compensation that arbitrageurs require in equilibrium, giving it an important economic interpretation.

In addition to the optimal strategy that attains the feasible Sharpe ratio, we also examine, theoretically and empirically, the performance of other portfolio construction approaches that rely multiple testing correction, variable selection, and shrinkage. While these approaches are well motivated statistically, they do not perform as well as the optimal strategy in economic terms, as measured by the Sharpe ratio. This finding highlights that good statistical properties, such as, for example, with regards to Type I and Type II errors or the false discovery rate, do not necessarily translate into good economic performance. For instance, a statistical procedure that guards against false discoveries may be overly conservative for investment purposes. This divergence between statistical and economic objectives suggests some caution when applying statistical tools imported from other areas—as in the current machine learning literature in asset pricing—without proper consideration of the economic objectives.

References

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2020). Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics 218*(2), 496–531.

Andrews, R. L., J. C. Arnold, and R. G. Krutchkoff (1972). Shrinkage of the posterior mean in the normal case. *Biometrika 59*(3), 693–695.

Avramov, D. and G. Zhou (2010). Bayesian portfolio analysis. *Annual Review of Financial Economics 2*(1), 25–47.

Bajgrowicz, P. and O. Scaillet (2012, December). Technical trading revisited False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics 106*(3), 473–491.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of financial economics 9*(1), 3–18.

Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance 65*(1), 179–216.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*(3), 657–681.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Black, F. and R. Litterman (1992). Global portfolio optimization. *Financial Analysts Journal 48*(5), 28–43.

Blume, M. E. and F. Husic (1973). Price, beta, and exchange listing. *The Journal of Finance 28*(2), 283–299.

Brennan, M. J., T. Chordia, and A. Subrahmanyam (1998). Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of financial Economics 49*(3), 345–373.

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.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–1304.

Chen, A. Y. and T. Zimmermann (2020). Open source cross-sectional asset pricing. *Available at SSRN*.

Collin-Dufresne, P., M. Johannes, and L. A. Lochstoer (2016). Parameter learning in general equilibrium: The asset pricing implications. *American Economic Review 106*(3), 664–698.

Connor, G., M. Hagmann, and O. Linton (2012). Efficient semiparametric estimation of the fama-french model and extensions. *Econometrica 80*(2), 713–754.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Cusatis, P. J., J. A. Miles, and J. R. Woolridge (1993). Restructuring through spinoffs: The stock market evidence. *Journal of financial economics 33*(3), 293–311.

Daniel, K. and S. Titman (2006). Market reactions to tangible and intangible information. *The Journal of Finance 61*(4), 1605–1643.

Dharan, B. G. and D. L. Ikenberry (1995). The long-run negative drift of post-listing stock returns. *The Journal of Finance 50*(5), 1547–1574.

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

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. (2019). Bayes, Oracle Bayes and Empirical Bayes. *Statistical Science 34*(2), 177 – 201.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*(4), 1497–1541.

Frost, P. A. and J. E. Savarino (1986). An empirical bayes approach to efficient portfolio selection. *The Journal of Financial and Quantitative Analysis 21*(3), 293–305.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity datasets. *Econometrica 84*(3), 985–1046.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989). A test of the efficiency of a given portfolio. *Econometrica 29*, 1121–1152.

Giglio, S., B. Kellly, and D. Xiu (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics 14*, 337–368.

Giglio, S., Y. Liao, and D. Xiu (2021). Thousands of alpha tests. *Review of Financial Studies 34*(7), 3456–3496.

Giglio, S. and D. Xiu (2021). Asset pricing with omitted factors. *Journal of Political Economy 129*(7), 1947–1990.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics 18*(2), 141–153.

Gromb, D. and D. Vayanos (2010). Limits of arbitrage. *Annual Review of Financial Economics 2*, 251–275.

Guijarro-Ordonez, J., M. Pelger, and G. Zanotti (2022). Deep learning statistical arbitrage. Technical report, Stanford University.

Hansen, L. P. (2007). Beliefs, doubts, and learning: Valuing macroeconomic risk. *American Economic Review 97*(2), 1–30.

Hansen, L. P. (2014). Nobel lecture: Uncertainty outside and inside economic models. *Journal of Political Economy 122*(51), 945–987.

Hartzmark, S. M. and D. H. Solomon (2013). The dividend month premium. *Journal of Financial Economics 109*(3), 640–660.

Harvey, C. R. and Y. Liu (2020). False (and missed) discoveries in financial economics. *Journal of Finance, forthcoming*.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *Review of Financial Studies 29*(1), 5–68.

Hou, K. (2007). Industry information diffusion and the lead-lag effect in stock returns. *The review of financial studies 20*(4), 1113–1138.

Huberman, G. (1982). A simple approach to arbitrage pricing theory. *Journal of Economic Thoery 28*(1), 183–191.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Ingersoll, J. E. (1984). Some results in the theory of arbitrage pricing. *Journal of Finance 39*(4), 1021–1039.

Johns, M. V. (1957). Non-parametric empirical bayes procedures. *Annals of Mathematical Statistics 28*, 649–669.

Jorion, P. (1986). Bayes-stein estimation for portfolio analysis. *The Journal of Financial and Quantitative Analysis 21*(3), 279–292.

Kan, R., X. Wang, and G. Zhou (2022). Optimal portfolio choice with estimation risk: No risk-free asset case. *Management Science, forthcoming.*

Kan, R. and G. Zhou (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis 42*(3), 621–656.

Kelly, B., S. Pruitt, and Y. Su (2019). Some characteristics are risk exposures, and the rest are irrelevant. *Journal of Financial Economics, forthcoming.*

Kim, S., R. Korajczyk, and A. Neuhierl (2020). Arbitrage portfolios. *Review of Financial Studies, forthcoming.*

Kozak, S. and S. Nagel (2023). When do cross-sectional asset pricing factors span the stochastic discount factor? Technical report, National Bureau of Economic Research.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometic Theory 21*(1), 21–59.

Maccheroni, F., M. Marinacci, and A. Rustichini (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica 74*(6), 1447–1498.

Martin, I. W. and S. Nagel (2022). Market efficiency in the age of big data. *Journal of financial economics 145*(1), 154–177.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of financial economics 108*(1), 1–28.

Pastor, L. (2000). Portfolio selection and asset pricing models. *The Journal of Finance 55*(1), 179–223.

Pastor, L. and R. F. Stambaugh (2000). Comparing asset pricing models: An investment perspective. *Journal of Financial Economics 56*, 335–381.

Pastor, L. and P. Veronesi (2009). Learning in financial markets. *Annual Review of Financial Economics 1*(1), 361–381.

Pesaran, H. and T. Yamagata (2017). Testing for alpha in linear factor pricing models with a large number of securities. Technical report.

Ritter, J. R. (1991). The long-run performance of initial public offerings. *The journal of finance 46*(1), 3–27.

Robbins, H. (1956). An empirical bayes approach to statistics. *Berkeley Symposium on Mathematical Statistics and Probability 3*, 157–163.

Robertson, T., R. Dykstra, and F. Wright (1988). Order restricted statistical inference. *New York: Wiley*.

Rosenberg, B. (1974). Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis 9*(2), 263–274.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*(3), 341–360.

Shanken, J. (1992). The current state of the arbitrage pricing theory. *Journal of Finance 47*(4), 1569–1574.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Tu, J. and G. Zhou (2010). Incorporating economic objectives into bayesian priors: Portfolio choice under parameter uncertainty. *The Journal of Financial and Quantitative Analysis 45*(4), 959–986.

Tukey, J. W. (1976). The higher criticism. course notes, statistics 411. Technical report, Princeton University.

Zhang, C.-H. (1997). Empirical bayes and compound estimation of normal means. *Statistica Sinica* 7(1), 181–193.

Online Appendix for
The Statistical Limit of Arbitrage

Rui Da[*]
Indiana University

Stefan Nagel[†]
University of Chicago
NBER and CEPR

Dacheng Xiu[‡]
University of Chicago
NBER

Abstract

The appendix contains additional empirical and theoretical results, as well as all the mathematical proofs.

[*] Address: 1275 E 10th St, Bloomington, IN 47405. E-mail: ruida@iu.edu.

[†] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637. E-mail: stefan.nagel@chicagobooth.edu.

[‡] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637. E-mail: dacheng.xiu@chicagobooth.edu.

Appendix A Additional Empirical Results

In this appendix we expand our empirical analysis. We first estimate latent factor models of portfolio returns, outlined in Section 4.3, under various alternative estimation windows and numbers of factors. The estimation is conducted for both the original portfolios and the ones adjusted for publication effects. We report the results in Table A1, which include the cross-sectional in-sample explanatory power (R^2) of expected returns of the factor models, the infeasible Sharpe ratios ($S^\star$) corresponding to the portfolio alphas, and the distributions of portfolio alphas' t-statistics.

	# of	R^2	$S^\star$	$\lvert t\text{-stat}\rvert$	$\lvert t\text{-stat}\rvert$	R^2	$S^\star$	$\lvert t\text{-stat}\rvert >$	$\lvert t\text{-stat}\rvert$
T	Factors			> 2	> 3			> 2	> 3
		Original Portfolios				Adjusted for Publication Effects			
3 yrs	5	0.41	27.58	12.17	2.63	0.48	13.47	9.93	1.67
	10	0.46	33.04	17.90	5.19	0.58	16.31	15.53	3.91
	15	0.50	39.69	24.44	8.98	0.63	19.77	21.78	7.29
5 yrs	5	0.41	20.60	11.54	2.37	0.47	10.05	9.11	1.37
	10	0.47	23.11	14.78	3.68	0.59	11.25	12.06	2.45
	15	0.51	25.53	17.73	5.07	0.66	12.46	14.94	3.55
7 yrs	5	0.38	17.62	12.49	2.76	0.43	8.56	9.57	1.45
	10	0.46	19.01	14.15	3.45	0.59	9.14	11.12	1.99
	15	0.50	20.44	16.02	4.39	0.65	9.93	13.27	2.81

Table A1: Latent Factor Models of Portfolio Returns

Note: The left panel displays the results of estimating latent factor models with the original portfolio returns as test assets. Across estimation windows and numbers of factors, the panel reports the cross-sectional R^2, the infeasible Sharpe ratios ($S^\star$), and the percentages of portfolio alphas with large t-statistics. Similarly, the right panel details the result when the test assets are portfolio returns adjusted for publication effects.

The cross-sectional R^2s by these models are similar across estimation windows, and moderately increase with more factors included and publication effects adjusted for. Indeed, in both cases, a larger proportion of cross-sectional variation of expected returns will be attributed to risk premia. Considering that these R^2s are in-sample ones, we choose to focus on the 10-factor model in Section 4. Regarding the strength and rareness of alphas, both the percentage of portfolio alphas with large t-statistics and the infeasible Sharpe ratio ($S^\star$) increase with the number of factors. It suggests that, when we control for more factors, the removal of factor risks is perhaps more prominent than the attribution of expected returns to risk premia (thereby less to alphas), which ultimately allows alphas to stand out and to be profited from more easily. On the other hand, with longer estimation window, we see both the infeasible Sharpe ratio and the percentage of large t-statistics reduce. This result could originate from that many alphas are only moderately persistent and is behind our choice of 5-year window in Section 4. Lastly, since fewer alphas are left after the publication effects are adjusted for, we observe increased cross-section R^2, and both the infeasible Sharpe ratios and the proportion of strong alpha signals shrink.

With the portfolio alphas obtained from the factor models, we show in Table A2 the investment performance of the four different trading strategies: OPT method, CSR, BH, and Lasso, measured by the average Sharpe ratios between January 1970 and December 2020.[1] It complements the analysis of Section 4.3, which focuses on 5-year estimation window and 10-factor specification. As a robustness check for the analysis on individual equity returns by Section 4.2, we also report in Table A2 the average Sharpe ratios generated by applying the four trading strategies to individual equity alphas.

T	# of Factors		5	10	15	5	10	15
		Individual Equities	Original Portfolios			Adjusted for Publication Effects		
3 yrs	OPT	0.71	1.23	1.33	1.25	0.54	0.59	0.54
	CSR	0.65	1.22	1.30	1.27	0.52	0.57	0.56
	BH	0.49	1.06	1.22	1.25	0.25	0.43	0.57
	Lasso	0.56	0.87	1.08	1.09	0.45	0.56	0.47
5 yrs	OPT	0.71	1.50	1.71	1.71	0.51	0.66	0.67
	CSR	0.70	1.47	1.67	1.70	0.55	0.61	0.66
	BH	0.65	1.52	1.49	1.67	0.25	0.35	0.48
	Lasso	0.62	0.95	1.18	1.23	0.48	0.45	0.47
7 yrs	OPT	0.81	1.71	1.95	2.01	0.72	0.73	0.79
	CSR	0.90	1.64	1.93	1.98	0.77	0.74	0.76
	BH	0.38	1.68	1.75	1.87	0.42	0.60	0.48
	Lasso	0.83	1.10	1.32	1.54	0.66	0.60	0.74

Table A2: Sharpe Ratios of Arbitrage Portfiolios

Note: The table reports the average Sharpe ratios generated by arbitrage portfolios constructed using OPT method, CSR, BH, and Lasso, with estimation windows set as 3, 5, and 7 years. The left panel displays the results when we invest in individual equity alphas. The center panel provides the Sharpe ratios generated when the investment universe is composed of portfolio alphas. The right panel displays the results when the publication effects are adjusted for when using portfolio alphas. The center and right panels include outputs under various numbers of factors in the latent factor model estimation.

Consistent with the evidence on the strength and rareness of alphas from Table A1, the investment gains from trading portfolio alphas significantly decline after the adjustment for publication effects. OPT method and CSR yield leading performance under most model specifications and for both individual equity alphas and portfolio alphas. Lasso generates

[1] The Sharpe ratio is the average between January 1972 and December 2020 under the 7-year estimation window, since our return data start from at January 1965.

second-tier outputs, due to the challenge of locating the optimal tuning parameter based on highly noisy return data. Notably, for portfolio alphas, we generally capture smaller Sharpe ratios when we control for fewer factors, which echoes the suggestive evidence from Table A1 on the impact of factor risks.

Appendix B Additional Theoretical Results

Proposition B1. *Suppose the same assumptions and information set $\mathcal{G}$ as in Theorem 1. Moreover, we assume that (a) $\|\beta\|_{\text{MAX}} \lesssim_{\text{P}} 1$ and $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\text{P}} N$; (b) v_t is i.i.d. across t, $\text{E}(v_t) = 0$, and its covariance matrix Σ_v satisfies $1 \lesssim \lambda_{\min}(\Sigma_v) \leq \lambda_{\max}(\Sigma_v) \lesssim 1$. Then it holds that, for any strategy w that is $\mathcal{G}$-measurable,*

$$S(w) \leq (S(\mathcal{G})^2 + \gamma^\intercal\Sigma_v^{-1}\gamma)^{1/2} + o_{\text{P}}(1 + S(\mathcal{G})).$$

Proposition B2. *Suppose that r_t follows (1), Assumption 1 holds, $u_{i,t} \sim \mathcal{N}(0, \sigma^2)$, and α following (5) as in Example 1. We also assume $\mu \lesssim N^{-d}$, $N^d \lesssim T \lesssim N^{1-d}$, and $N^{-d'} \lesssim \rho \lesssim N^{-d}$ for some fixed $d' > d > 0$. We denote the Sharpe ratio of the arbitrage portfolio given by (20), as $\widehat{S}^{\text{CSR}}$. Then it satisfies $\widehat{S}^{\text{CSR}} - S^{\text{CSR}} = o_{\text{P}}(1)$, where*

$$S^{\text{CSR}} = \frac{N^{1/2}\rho\mu^2\sigma^{-2}}{(T^{-1} + \rho\mu^2\sigma^{-2})^{1/2}}.$$

Suppose further that S^{OPT} does not vanish. It follows that $S^{\text{CSR}} = S^{\text{OPT}}(1 + o(1))$, if and only if there exists $c_N \to 0$, such that $\sqrt{T}\mu/\sigma \leq c_N$ or $\sqrt{\rho T}\mu/\sigma \geq c_N^{-1}$ holds for all large N.

Proposition B3. *Suppose the same assumptions as in Proposition B2 hold. For any given τ, the Sharpe ratio of the arbitrage portfolio with weights given by (21) denoted by $\widehat{S}^{\text{BH}}_\tau$, satisfies $\widehat{S}^{\text{BH}}_\tau = S^{\text{BH}}_\tau + o_{\text{P}}(1 + S^{\text{OPT}})$, where*[2]

$$S^{\text{BH}}_\tau = \frac{N^{1/2}\text{E}(\alpha_i\check{s}_i\mathbb{1}_{\{|\check{s}_i|\geq s^*\}})}{\sigma\sqrt{\text{E}(\check{s}_i^2\mathbb{1}_{\{|\check{s}_i|\geq s^*\}})}} = \frac{N^{1/2}\rho\mu\int_{-\infty}^{\infty} x\mathbb{1}_{\{|x|\geq s^*\}}\phi_{1/T}(x - \mu\sigma^{-1})dx}{\sigma\sqrt{\int_{-\infty}^{\infty} x^2\mathbb{1}_{\{|x|\geq s^*\}}((1-\rho)\phi_{1/T}(x) + \rho\phi_{1/T}(x - \mu\sigma^{-1}))dx}},$$

$\check{s}_i = (\alpha_i + \bar{u}_i)/\sigma_i$, and s^* is the smallest positive solution of the equation

$$2(1-\tau)\Phi(-T^{1/2}s) = \tau\rho\Phi(T^{1/2}(\mu/\sigma - s)).^{3} \tag{B.1}$$

Suppose further that S^{OPT} does not vanish. Then there exists, for any fixed $\epsilon > 0$, some $\tau > 0$

[2] If $\widehat{w}^{\text{BH}} = 0$, i.e., no asset is selected, we set $\widehat{S}^{\text{BH}} = 0$ by convention.

[3] $\Phi(\cdot)$ is the standard normal cumulative distribution function.

such that, as $N, T \to \infty$, $S_\tau^{\mathrm{BH}} > (1-\epsilon) S^{\mathrm{OPT}}$, if and only if $\sqrt{T}\mu/\sigma \geq \sqrt{-2\log\rho}(1+o(1))$.

Proposition B4. *Suppose the same assumptions as in Proposition B2 hold. The Sharpe ratio of the arbitrage portfolio with weights given by $\widehat{w}_q(\lambda)$, denoted as $\widehat{S}_{q,\lambda}$ for $q = 1, 2$, satisfies $\widehat{S}_{1,\lambda} - S_\lambda^{\mathrm{Lasso}} = o_{\mathrm{P}}(1 + S^{\mathrm{OPT}})$ and $\widehat{S}_{2,\lambda} - S^{\mathrm{CSR}} = o_{\mathrm{P}}(1 + S^{\mathrm{OPT}})$, where*

$$S_\lambda^{\mathrm{Lasso}} = \frac{N^{1/2}\mathrm{E}(\alpha_i \psi_1(\check{s}_i, \lambda))}{\sigma\sqrt{\mathrm{E}(\psi_1(\check{s}_i,\lambda)^2)}} = \frac{N^{1/2}\rho\mu\int_{-\infty}^{\infty}\mathrm{sgn}(x)(|x|-\lambda)_+\phi_{1/T}(\mu\sigma^{-1}-x)dx}{\sigma\sqrt{\int_{-\infty}^{\infty}((|x|-\lambda)_+)^2((1-\rho)\phi_{1/T}(x)+\rho\phi_{1/T}(\mu\sigma^{-1}-x))dx}}, \tag{B.2}$$

and S^{CSR} is defined in Proposition B2.

Furthermore, if S^{OPT} does not vanish, then $S_\lambda^{\mathrm{Lasso}} = S^{\mathrm{OPT}}(1+o(1))$, under some deterministic non-negative sequence of λ, if and only if, for some $c_N \to 0$, either $\sqrt{T}\mu/\sigma \leq c_N$ or $\sqrt{T}\mu/\sigma - \sqrt{-2\log\rho} \geq c_N^{-1}$ holds for all large N.

Moreover, when $\sqrt{T}\mu/\sigma \leq c_N$, $S_\lambda^{\mathrm{Lasso}}$ approaches S^{OPT} if and only if λ satisfies $T^{1/2}\lambda \to 0$. When $\sqrt{T}\mu/\sigma - \sqrt{-2\log\rho} \geq c_N^{-1}$, $S_\lambda^{\mathrm{Lasso}}$ approaches S^{OPT} if and only if λ satisfies $\sqrt{T}(\mu/\sigma - \lambda) \to \infty$ and $\frac{\phi(\sqrt{T}\lambda)}{\rho(1+T\lambda^2)T(\mu/\sigma-\lambda)^2} \to 0$.

Appendix C Mathematical Proofs

C.1 Proof of Equation (9)

The constrained maximization problem is equivalent with maximizing

$$w^\intercal\alpha - \frac{\kappa}{2}w^\intercal\Sigma_u w + \lambda w^\intercal\beta,$$

where λ is the Lagrange multiplier such that the solution w satisfies $w^\intercal\beta = 0$. As the objective function is clearly strictly concave, we only need to look at the first-order condition, which reads

$$w = \frac{1}{\kappa}\Sigma_u^{-1}(\alpha + \lambda\beta).$$

We can pin down λ by requiring $\beta^\intercal\Sigma_u^{-1}(\alpha+\lambda\beta) = 0$, which gives $\lambda = (\beta^\intercal\Sigma_u^{-1}\beta)^{-1}\beta^\intercal\Sigma_u^{-1}\alpha$ and thereby

$$w = \frac{1}{\kappa}\Sigma_u^{-1}(\mathbb{I}_N + \beta(\beta^\intercal\Sigma_u^{-1}\beta)^{-1}\beta^\intercal\Sigma_u^{-1})\alpha = \frac{1}{\kappa}\Sigma_u^{-1/2}\mathbb{M}_{\Sigma_u^{-1/2}\beta}\Sigma_u^{-1/2}\alpha.$$

C.2 Proof of Theorem 1

To simplify the notation, we omit the dependence of β, Σ on N, and w on N and T. All limits are taken as $N \to \infty$. The derivation applies to either fixed T or $T \to \infty$ together with N. We first show that, under any sequence of probability measures $\{\mathrm{P}_N\}$ that satisfies Assumption 1, and for all fixed positive ε,

$$\mathrm{P}_N(U(w) \leq_\varepsilon (2\kappa)^{-1} S(\mathcal{G})^2 =_\varepsilon U(\widetilde{w})) \to 1, \quad \text{and} \quad \mathrm{P}_N(S(w) \leq_\varepsilon S(\mathcal{G}) =_\varepsilon S(\widetilde{w})) \to 1. \quad \text{(C.3)}$$

The following derivation proceeds under any such sequence of probability measures and we omit the subscript N. We first note that, given (1), conditioning on $\mathcal{G}$ is equivalent to conditioning on the information set generated by

$$\{(\alpha_i + u_{i,s}, \beta_i, v_s, \sigma_i) : t - T + 1 \leq s \leq t, i \leq N\}.$$

According to Assumption 1, conditionally on Σ_u, $\{(\alpha_i, \alpha_i + u_{i,s}) : t - T + 1 \leq s \leq t\}$ is independent of $\{(\alpha_j + u_{j,s}, \beta_{j'}, v_s) : t - T + 1 \leq s \leq t, j, j' \leq N, j \neq i\}$. Therefore, the $\mathcal{G}$-conditional distribution of α_i is the same as the distribution of α_i conditional on $\{\alpha_i + u_{i,s} : t - T + 1 \leq s \leq t\}$ and Σ_u. Because σ_j is independent with (α_i, u_i) for $j \neq i$, the $\mathcal{G}$-conditional distribution of α_i is the same as the the $\mathcal{G}_i$-conditional distribution of α_i, where $\mathcal{G}_i$ is the information set generated by $\{(\alpha_i + u_{i,s}, \sigma_i) : t - T + 1 \leq s \leq t\}$. Since $\mathcal{G}_i$ is independent across i by Assumption 1, we conclude that, conditionally on $\mathcal{G}$, α_i remains independent across i.

Now define $\mathcal{E} = \mathrm{E}(w^\intercal r_{t+1}|\mathcal{F}_t) - \mathrm{E}(w^\intercal r_{t+1}|\mathcal{G})$. By the definition of $S(w)$, we have

$$S(w) = \mathrm{E}(w^\intercal r_{t+1}|\mathcal{G})/\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t)^{1/2} + \mathcal{E}/\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t)^{1/2}. \quad \text{(C.4)}$$

Since w is $\mathcal{G}$-measurable, it follows that $\mathcal{E} = w^\intercal(\alpha - \mathrm{E}(\alpha|\mathcal{G}))$ and that $\mathrm{E}(\mathcal{E}^2|\mathcal{G}) = w^\intercal \mathrm{Var}(\alpha|\mathcal{G}) w$. Then, using Chebyshev's inequality, we have, for all positive fixed ϵ,

$$\mathrm{P}(|\mathcal{E}|/\|w\| \geq \epsilon) \leq \mathrm{E}(\mathcal{E}^2/\|w\|^2)/\epsilon^2 = \mathrm{E}(w^\intercal \mathrm{Var}(\alpha|\mathcal{G}) w/\|w\|^2)/\epsilon^2. \quad \text{(C.5)}$$

Because conditionally on $\mathcal{G}$, α_i is independent across i, we have $\mathrm{Var}(\alpha|\mathcal{G})_{i,j} = \delta_{i,j}\mathrm{Var}(\alpha_i|\mathcal{G})$. It thereby follows that

$$\mathrm{E}(w^\intercal \mathrm{Var}(\alpha|\mathcal{G}) w/\|w\|^2) \leq \mathrm{E}(\max_{i \leq N} \mathrm{Var}(\alpha_i|\mathcal{G})) \leq \mathrm{E}(\max_{i \leq N} \alpha_i^2) = o(1), \quad \text{(C.6)}$$

where the last step comes from condition (a) of Assumption 1. Combining (C.5) and (C.6), and using $\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t) = w^\intercal \Sigma w \geq \lambda_{\min}(\Sigma_u)\|w\|^2 \gtrsim_{\mathrm{P}} \|w\|^2$, we obtain

$$|\mathcal{E}|/\mathrm{Var}(w^\intercal r_{t+1}|\mathcal{F}_t)^{1/2} \lesssim_{\mathrm{P}} |\mathcal{E}|/\|w\| = o_{\mathrm{P}}(1). \tag{C.7}$$

(C.7) and (C.4) lead to

$$S(w) = w^\intercal \mathrm{E}(r_{t+1}|\mathcal{G})(w^\intercal \Sigma w)^{-1/2} + o_{\mathrm{P}}(1). \tag{C.8}$$

Furthermore, if $w^\intercal \beta = 0$, then it follows that $w^\intercal r_t = w^\intercal(\alpha + u_t)$ and $w^\intercal \Sigma w = w^\intercal \Sigma_u w$. Equation (C.8) then becomes

$$S(w) = w^\intercal \widetilde{\alpha}(w^\intercal \Sigma_u w)^{-1/2} + o_{\mathrm{P}}(1). \tag{C.9}$$

Applying Cauchy-Schwarz inequality, we obtain

$$|w^\intercal \widetilde{\alpha}|^2 (w^\intercal \Sigma_u w)^{-1} \leq \widetilde{\alpha}^\intercal \Sigma_u^{-1} \widetilde{\alpha} = S(\mathcal{G})^2,$$

which, combined with (C.9), proves the second inequality in (C.3).

Next, we have that, for all w that is $\mathcal{G}$-measurable and satisfies $\beta^\intercal w = 0$, and for all positive fixed ϵ and with probability approaching one,

$$U(w) = w^\intercal \alpha - \frac{\kappa}{2} w^\intercal \Sigma_u w = \mathcal{E} + w^\intercal \widetilde{\alpha} - \frac{\kappa}{2} w^\intercal \Sigma_u w \leq w^\intercal \widetilde{\alpha} - \frac{\kappa}{2}(1-\varepsilon) w^\intercal \Sigma_u w, \tag{C.10}$$

where the last inequality comes from (C.7). Cauchy-Schwarz inequality leads to

$$w^\intercal \widetilde{\alpha} - \frac{\kappa}{2}(1-\varepsilon) w^\intercal \Sigma_u w \leq \frac{1}{2\kappa(1-\varepsilon)} \widetilde{\alpha}^\intercal \Sigma_u \widetilde{\alpha}.$$

Substituting this result into (C.10), we obtain the first inequality in (C.3).

We move on to the remaining results: the two equalities in (C.3). They all hinge on the property of $\widetilde{w}$. As a first step, we introduce short-hand notation

$$\breve{w} := \frac{1}{\kappa} \mathbb{P}_\beta \Sigma_u^{-1} \widetilde{\alpha}, \quad \text{with} \quad \mathbb{P}_\beta := \mathbb{I}_N - \mathbb{M}_\beta.$$

We can then write

$$\widetilde{w} = \frac{1}{\kappa} \Sigma_u^{-1} \widetilde{\alpha} - \breve{w}. \tag{C.11}$$

To establish the two equalities in (C.3), we now prove that $\|\breve{w}\|$ is $o_{\mathrm{P}}(1)$, which will then quickly leads to those results.

To this end, we start by analyzing $\mathrm{E}(\widetilde{\alpha}|\beta, \Sigma_u)$ and $\mathrm{Cov}(\widetilde{\alpha}|\Sigma_u, \beta)$. We note that

$$\mathrm{E}(\widetilde{\alpha}|\beta, \Sigma_u) = \mathrm{E}(\alpha|\beta, \Sigma_u) = \mathrm{E}(\alpha|\Sigma_u) = 0. \tag{C.12}$$

The first equality comes from that $\widetilde{\alpha} := \mathrm{E}(\alpha|\mathcal{G})$ and that β and Σ_u are $\mathcal{G}$-measurable. The second equality comes from Assumption 1 (b). The last equality comes from Assumption 1 (a). Moreover, from the analysis above (C.4), $\widetilde{\alpha}_i$ is a function of $\{(\alpha_i + u_{i,s}, \sigma_i) : t-T+1 \leq s \leq t\}$. According to Assumption 1 (b), $\{(\alpha_i + u_{i,s}, \sigma_i) : t - T + 1 \leq s \leq t\}$ and β are, conditionally on Σ_u, independent. Therefore,

$$\mathrm{Cov}(\widetilde{\alpha}_i, \widetilde{\alpha}_j|\Sigma_u, \beta) = \mathrm{Cov}(\widetilde{\alpha}_i, \widetilde{\alpha}_j|\Sigma_u) = \delta_{i,j}\mathrm{Var}(\widetilde{\alpha}_i|\Sigma_u). \tag{C.13}$$

The second equality comes from $\mathrm{Cov}(\widetilde{\alpha}_i, \widetilde{\alpha}_j|\Sigma_u) = 0$ for $i \neq j$, which is because (α_i, u_i) is i.i.d. across i per Assumption 1 (a). Therefore, it holds that

$$\mathrm{Cov}(\widetilde{\alpha}|\Sigma_u, \beta) \leq \mathrm{E}(\|\alpha\|^2_{\mathrm{MAX}}|\Sigma_u)\mathbb{I}_N \tag{C.14}$$

It hence holds that

$$\begin{aligned} \mathrm{E}(\|\breve{w}\|^2|\Sigma_u, \beta) &= \mathrm{Tr}(\mathbb{P}_\beta\Sigma_u^{-1}\mathrm{Cov}(\widetilde{\alpha}|\Sigma_u, \beta)\Sigma_u^{-1}\mathbb{P}_\beta) \\ &\leq \mathrm{E}(\|\alpha\|^2_{\mathrm{MAX}}|\Sigma_u)\mathrm{Tr}(\mathbb{P}_\beta\Sigma_u^{-2}\mathbb{P}_\beta). \end{aligned} \tag{C.15}$$

The equality comes from (C.12). The inequality comes from (C.13), (C.14), and the well-known result on Loewner order (see, e.g., Theorem 7.7.2 at Horn and Johnson (2012)). On the other hand, it holds that

$$\mathrm{Tr}(\mathbb{P}_\beta\Sigma_u^{-2}\mathbb{P}_\beta) \lesssim_{\mathrm{P}} \mathrm{Tr}(\mathbb{P}_\beta) = \mathrm{rank}(\beta) = K. \tag{C.16}$$

The inequality comes from $\Sigma_u \gtrsim_{\mathrm{P}} \mathbb{I}_N$ by Assumption 1 (a). Substituting (C.16) into (C.15), we obtain

$$\mathrm{E}(\|\breve{w}\|^2|\Sigma_u, \beta) \lesssim_{\mathrm{P}} \mathrm{E}(\|\alpha\|^2_{\mathrm{MAX}}|\Sigma_u) = o_{\mathrm{P}}(1), \tag{C.17}$$

where the last bound comes from Assumption 1 (a). Applying Chebyshev's inequality to (C.17), we obtain

$$\|\breve{w}\| = o_{\mathrm{P}}(1). \tag{C.18}$$

To prove the equalities in (C.3), we note that, using (C.11),

$$\widetilde{w}^\intercal\widetilde{\alpha} = \frac{1}{\kappa}\widetilde{\alpha}^\intercal\Sigma_u^{-1}\widetilde{\alpha} - \breve{w}^\intercal\widetilde{\alpha}, \quad \widetilde{w}^\intercal\Sigma_u\widetilde{w} = \frac{1}{\kappa^2}\widetilde{\alpha}^\intercal\Sigma_u^{-1}\widetilde{\alpha} + \breve{w}^\intercal\Sigma_u\breve{w} - \frac{2}{\kappa}\breve{w}^\intercal\widetilde{\alpha}. \tag{C.19}$$

On the other hand, applying that $\Sigma_u \lesssim_{\mathrm{P}} \mathbb{I}_N$ by Assumption 1 (a) and (C.18), we obtain

$$\breve{w}^\intercal \Sigma_u \breve{w} \lesssim_{\mathrm{P}} \|\breve{w}\| = o_{\mathrm{P}}(1), \quad |\breve{w}^\intercal \widetilde{\alpha}| \leq \sqrt{\breve{w}^\intercal \Sigma_u \breve{w}} S(\mathcal{G}) \lesssim_{\mathrm{P}} \|\breve{w}\| S(\mathcal{G}) = o_{\mathrm{P}}(S(\mathcal{G})). \tag{C.20}$$

Substituting (C.19) into (C.20), we establish

$$\widetilde{w}^\intercal \widetilde{\alpha} = \frac{1}{\kappa} S(\mathcal{G})^2 + o_{\mathrm{P}}(S(\mathcal{G})), \quad \widetilde{w}^\intercal \Sigma_u \widetilde{w} = \frac{1}{\kappa^2} S(\mathcal{G})^2 + o_{\mathrm{P}}(1 + S(\mathcal{G})). \tag{C.21}$$

Suppose $S(\mathcal{G}) \gtrsim_{\mathrm{P}} 1$. Then, the second equality in (C.3) indeed follows from (C.21). Suppose $S(\mathcal{G}) = o_{\mathrm{P}}(1)$. Then $S(\widetilde{w}) \leq o_{\mathrm{P}}(1)$ and $-S(\widetilde{w}) = S(-\widetilde{w}) \leq o_{\mathrm{P}}(1)$ according to the first inequality of (C.3) as $\widetilde{w}$ is both $\mathcal{G}$-measurable and factor-neutral. In other words, we have $S(\widetilde{w}) = o_{\mathrm{P}}(1)$ and that the second equality in (C.3) also holds. Applying the classic subsequence argument, we establish that the second equality in (C.3) holds.

Given (C.21) and (C.7), we obtain the first equality in (C.3) as

$$U(\widetilde{w}) = \mathcal{E} + \widetilde{w}^\intercal \widetilde{\alpha} - \frac{\kappa}{2} \widetilde{w}^\intercal \Sigma_u \widetilde{w} = \widetilde{w}^\intercal \widetilde{\alpha} - \frac{\kappa}{2}(1 + o_{\mathrm{P}}(1)) \widetilde{w}^\intercal \Sigma_u \widetilde{w} = \frac{1}{2\kappa} S(\mathcal{G})^2 + o_{\mathrm{P}}(1 + S(\mathcal{G})).$$

Now we prove (11) using (C.3). First, from the definition of $\leq_\varepsilon$ and $=_\varepsilon$, (C.3) holds under any sequences of probability measuring satisfying Assumption 1 and for all fixed $\varepsilon > 0$ is equivalent with that, under any such sequence and for all fixed $\varepsilon > 0$,

$$\mathrm{P}_N(U(w) \leq_\varepsilon U(\widetilde{w}) =_\varepsilon (2\kappa)^{-1} S(\mathcal{G})^2) \to 1, \quad \text{and} \quad \mathrm{P}_N(S(w) \leq_\varepsilon S(\widetilde{w}) =_\varepsilon S(\mathcal{G})) \to 1. \tag{C.22}$$

Next, suppose the first inequality of (11) does not hold. Then we have, for some fixed $\varepsilon > 0$,

$$\liminf_{N \to 0} \inf_{\mathrm{P} \in \mathbb{P}} \mathrm{P}(U(w) \leq_\varepsilon U(\widetilde{w})) < 1.$$

Hence there is a subsequence $\{\mathrm{P}_{N_k}\}_{k \geq 1}$ with $\mathrm{P}_{N_k} \in \mathbb{P}$ for each $k \geq 1$ such that, for some fixed $\varepsilon > 0$,

$$\lim_{k \to \infty} \mathrm{P}_{N_k}(U(w) \leq_\varepsilon U(\widetilde{w})) < 1. \tag{C.23}$$

Since $\mathbb{P}$ only collects probability measures under which Assumption 1 holds, (C.23) gives a contradiction to (C.22), which means that the first inequality of (11) must hold. Symmetric arguments establish the second inequality and the two equalities of (11).

C.3 Proof of Theorem 2, Corollary 1, and Corollary 2

Proof of Theorem 2. We have established in the beginning of the proof of Theorem 1 that the $\mathcal{G}$-conditional distribution of α_i is the same as the $\mathcal{G}_i$-conditional distribution of α_i, where $\mathcal{G}_i$ is the information set generated by $\{(\alpha_i + u_{i,s}) : t - T + 1 \leq s \leq t\}$ and σ_i. Note that $u_{i,s}$ is centered normal, we have that the conditional probability density of $\{r^*_{i,s} :=$

$\alpha_i + u_{i,s}, t-T+1 \le s \le t\}$ given s_i and σ_i, denoted by $p(r_i^*|s_i,\sigma_i)$, is

$$p(r_i^*|s_i,\sigma_i) = \prod_{t-T+1\le s\le t} \sigma_i^{-1}\phi(\sigma_i^{-1}r_{i,s}^* - s_i) = \phi(T^{1/2}(\check{s}_i - s_i))f(r_i^*). \tag{C.24}$$

Here $f(r_i^*)$ is a function of r_i^* and σ_i that does not depend on s_i. Hence, applying Bayes' theorem, we have

$$\begin{aligned}
\widetilde{\alpha}_i = \mathrm{E}(\alpha_i|\mathcal{G}_i) = \sigma_i\mathrm{E}(s_i|\mathcal{G}_i) &= \sigma_i\int xp(s_i = x|r_i^*,\sigma_i)dx \\
&= \sigma_i\int x\frac{p(r_i^*|s_i = x,\sigma_i)p(s_i = x|\sigma_i)}{\int p(r_i^*|s_i = x',\sigma_i)p(s_i = x'|\sigma_i)dx'}dx \\
&= \sigma_i\int x\frac{p(r_i^*|s_i = x,\sigma_i)p_s(x)}{\int p(r_i^*|s_i = x',\sigma_i)p_s(x')dx'}dx \\
&= \sigma_i\int x\frac{\phi(T^{1/2}(\check{s}_i - x))p_s(x)}{\int \phi(T^{1/2}(\check{s}_i - x'))p_s(x')dx'}dx = \sigma_i\mathrm{E}(s_i|\check{s}_i),
\end{aligned}$$

The second line comes from the Bayes' theorem. In the third line, $p_s(\cdot)$ is the marginal density of s_i that is invariant across i, and we use the fact that s_i and σ_i are independent, given by condition (a) of Assumption 2. The first equality in the last line comes from (C.24). The second equality in the last line comes from the Bayes' theorem and that $\phi(T^{1/2}(\check{s}_i - x))$ is, up to a constant, the density of $\check{s}_i$ conditional on $s_i = x$. We hence establish the first statement in the theorem.

Next, we note that $\psi(a) = \mathrm{E}(s_i|\check{s}_i = a)$ and, conditional on $s_i = a$, $\check{s}_i \approx \mathcal{N}(a, T^{-1})$. The marginal density of $\check{s}_i$ is then indeed $\mathrm{E}(\phi_{1/T}(a - s_i))$. Therefore, (15) directly comes from the Tweedie's formula, see, e.g., Robbins (1956) . The proof ends. ∎

Proof of Corollary 1. Apparently, $\check{s}_i = s_i + \sigma_i^{-1}\bar{u}_i$ is i.i.d. across i. By direction calculation and the definition of S^{OPT} in the statement of the corollary, we have

$$\mathrm{E}(\widetilde{\alpha}^{\intercal}\Sigma_u^{-1}\widetilde{\alpha}) = \sum_i \mathrm{E}(\mathrm{E}(s_i|\mathcal{G})^2) = N\int \psi(a)^2p(a)da = (S^{\mathrm{OPT}})^2. \tag{C.25}$$

Now we study $S(\mathcal{G}) = \widetilde{\alpha}^{\intercal}\Sigma_u^{-1}\widetilde{\alpha} = \sum_i \mathrm{E}(s_i|\mathcal{G})^2$. Using the fact that $a^2 - b^2 = (a-b)^2 + 2b(a-b)$, we have

$$\begin{aligned}
&\mathrm{E}(|\mathrm{E}(s_i\mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2 - \mathrm{E}(s_i|\mathcal{G})^2|) \\
\le\ &\mathrm{E}(\mathrm{E}(s_i\mathbb{1}_{\{|s_i|>c_N\}}|\mathcal{G})^2) + 2\mathrm{E}(|\mathrm{E}(s_i|\mathcal{G})\mathrm{E}(s_i\mathbb{1}_{\{|s_i|>c_N\}}|\mathcal{G})|) \\
\le\ &\mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|>c_N\}}) + 2\sqrt{\mathrm{E}(\mathrm{E}(s_i|\mathcal{G})^2)\mathrm{E}(\mathrm{E}(s_i\mathbb{1}_{\{|s_i|>c_N\}}|\mathcal{G})^2)}
\end{aligned}$$

$$\begin{aligned}&\leq \mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|>c_N\}}) + 2\sqrt{\mathrm{E}(\mathrm{E}(s_i|\mathcal{G})^2)\mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|>c_N\}})}\\ &\leq c_N N^{-1} + \sqrt{\mathrm{E}(\mathrm{E}(s_i|\mathcal{G})^2)c_N N^{-1}},\end{aligned} \tag{C.26}$$

where the last step holds by the assumption $\mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|\geq c_N\}}) = o(N^{-1})$. Then we have

$$\begin{aligned}&\mathrm{E}\left(|\sum_i \mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2 - \sum_i \mathrm{E}(s_i|\mathcal{G})^2|\right)\\ \leq\ & \sum_i \mathrm{E}(|\mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2 - \mathrm{E}(s_i|\mathcal{G})^2|)\\ \leq\ & c_N + \sqrt{\sum_i \mathrm{E}(\mathrm{E}(s_i|\mathcal{G})^2)c_N} = o(1+S^{\mathrm{OPT}}),\end{aligned} \tag{C.27}$$

where the second inequality is a direct result of (C.26), and the last estimate is given by (C.25). From (C.27) and (C.25), it follows, respectively, using Markov's inequality and triangle inequality that

$$\sum_i \mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2 = \sum_i \mathrm{E}(s_i|\mathcal{G})^2 + o_{\mathrm{P}}(1+S^{\mathrm{OPT}}), \tag{C.28}$$

$$\mathrm{E}\left(\sum_i \mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2\right) = (S^{\mathrm{OPT}})^2 + o(1+S^{\mathrm{OPT}}). \tag{C.29}$$

Further, we have

$$\begin{aligned}\mathrm{Var}\left(\sum_i \mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2\right) &= \sum_i \mathrm{Var}(\mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2)\\ &\leq c_N^2 \sum_i \mathrm{E}(\mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2) = o(1+(S^{\mathrm{OPT}})^2).\end{aligned} \tag{C.30}$$

For the first line, we use that $\mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})$ is independent across i. The second line is obvious as $|s_i|\mathbb{1}_{\{|s_i|\leq c_N\}} \leq c_N$. The last line comes from (C.29). Combining (C.29) and (C.30), we obtain

$$\sum_i \mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2 = (S^{\mathrm{OPT}})^2 + o(1+S^{\mathrm{OPT}}) + o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)^{1/2}.$$

Along with (C.28), we obtain

$$\sum_i \mathrm{E}(s_i|\mathcal{G})^2 = (S^{\mathrm{OPT}})^2 + o_{\mathrm{P}}(1+S^{\mathrm{OPT}}).$$

In light of the definition of $S(\mathcal{G})$, and the fact that

$$((S^{\mathrm{OPT}})^2 + o_{\mathrm{P}}(1 + S^{\mathrm{OPT}}))^{1/2} = S^{\mathrm{OPT}} + o_{\mathrm{P}}(1),$$

we conclude the proof. ∎

Proof of Corollary 2. Because of the tail condition $\mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) \leq c_N N^{-1}$ for some sequence $c_N \to 0$ and that σ_i is constant across i, we have

$$\mathrm{E}\left|\alpha^{\intercal}\alpha - \sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}}\right| = \mathrm{E}\left|\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| \geq c_N\}}\right| = o(1),$$

which, by Markov's inequality and triangle inequality, respectively, leads to

$$\alpha^{\intercal}\alpha = \sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}} + o_{\mathrm{P}}(1), \quad \mathrm{E}\left(\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}}\right) = \mu^2 \rho N. \tag{C.31}$$

On the other hand, it holds that

$$\mathrm{Var}\left(\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}}\right) \leq \sum_i \mathrm{E}(\alpha_i^4 \mathbb{1}_{\{|\alpha_i| < c_N\}}) \leq c_N^2 \sum_i \mathrm{E}(\alpha_i^2) = c_N^2 \mu^2 \rho N. \tag{C.32}$$

Combining (C.31) and (C.32), we obtain

$$\alpha^{\intercal}\alpha = \mu^2 \rho N + o_{\mathrm{P}}(1 + \mu\sqrt{\rho N}).$$

As a result, it holds that

$$S^{\star} = \sigma^{-1}\sqrt{\alpha^{\intercal}\alpha} = \sigma^{-1}\mu(\rho N)^{1/2} + o_{\mathrm{P}}(1). \tag{C.33}$$

Further, in light of the explicit distribution of α in Example 1, we have

$$\begin{aligned} \psi(a) &= \frac{\mu\rho\phi(a - T^{1/2}\mu/\sigma) - \mu\rho\phi(a + T^{1/2}\mu/\sigma)}{(2-2\rho)\phi(a) + \rho\phi(a - T^{1/2}\mu/\sigma) + \rho\phi(a + T^{1/2}\mu/\sigma)}, && \text{(C.34)} \\ \left(S^{\mathrm{OPT}}\right)^2 &= \frac{\mu\rho N}{2\sigma^2}\int \psi(a)(\phi(a - T^{1/2}\mu/\sigma) - \phi(a + T^{1/2}\mu/\sigma))da. && \text{(C.35)} \end{aligned}$$

Suppose that $T^{1/2}\mu\sigma^{-1} - \sqrt{-2\log\rho} \leq C < \infty$. Then we have

$$\sup_{a \geq C} \frac{\rho\phi(a)}{\phi(a - T^{1/2}\mu/\sigma)} = \exp\left(\log\rho + T^{1/2}\mu\sigma^{-1}\left(\frac{1}{2}T^{1/2}\mu\sigma^{-1} - C\right)\right)$$

$$\leq \quad \exp\left(\log\rho + \frac{1}{2}\left(\sqrt{-2\log\rho}+C\right)\left(\sqrt{-2\log\rho}-C\right)\right) \leq 1 \tag{C.36}$$

On the other hand, in light of (C.34) and (C.35), we have

$$\begin{aligned}
\left(S^{\mathrm{OPT}}\right)^2 &= \frac{\mu\rho N}{\sigma^2}\int \psi(a)\phi(a-T^{1/2}\mu/\sigma)da \\
&\leq \frac{\mu\rho N}{\sigma^2}\int \frac{\mu\rho\phi(a-T^{1/2}\mu/\sigma)}{(2-2\rho)\phi(a)+\rho\phi(a-T^{1/2}\mu/\sigma)}\phi(a-T^{1/2}\mu/\sigma)da \\
&= \frac{\mu^2\rho N}{\sigma^2}\int \frac{\rho\phi(a)}{(2-2\rho)\phi(a-T^{1/2}\mu/\sigma)+\rho\phi(a)}\phi(a)da.
\end{aligned}$$

We hence obtain from (C.36) that, for N sufficiently large,

$$\left(S^{\mathrm{OPT}}\right)^2 \leq \frac{\mu^2\rho N}{\sigma^2}\left(\int_{a\geq C}\frac{1}{3-2\rho}\phi(a)da + \int_{a\leq C}\phi(a)da\right) \leq \frac{\mu^2\rho N}{\sigma^2}\left(1-\frac{1}{2}\Phi(-C)\right).$$

This proves the "if" part, given (C.33) and that $\mu^2\rho N/\sigma^2$ does not vanish. Now suppose $T^{1/2}\mu\sigma^{-1}-\sqrt{-2\log\rho}\to\infty$. Then, for all fixed $x>0$, we have, for sufficiently large N,

$$\begin{aligned}
\sup_{a:|a|\leq x}\frac{\phi(a+T^{1/2}\mu/\sigma)}{\rho\phi(a)} &= \exp\left(-\log\rho - T^{1/2}\mu\sigma^{-1}\left(\frac{1}{2}T^{1/2}\mu\sigma^{-1}-x\right)\right) \\
&\leq \exp\left(-\log\rho-\frac{1}{2}\left(\sqrt{-2\log\rho}+c_N^{-1}\right)\left(\sqrt{-2\log\rho}+c_N^{-1}\right)\right) \\
&\leq \exp\left(-c_N^{-2}/2\right)\to 0, && \text{(C.37)} \\
\sup_{a:|a|\leq x}\frac{\phi(a+2T^{1/2}\mu/\sigma)}{\phi(a)} &= \exp\left(-2T^{1/2}\mu\sigma^{-1}(T^{1/2}\mu\sigma^{-1}-x)\right)\to 0. && \text{(C.38)}
\end{aligned}$$

Given (C.34), it holds that

$$\psi\left(a+T^{1/2}\mu/\sigma\right) = \mu\frac{1-\frac{\phi(a+2T^{1/2}\mu/\sigma)}{\phi(a)}}{1+\frac{(2-2\rho)\phi(a+T^{1/2}\mu/\sigma)}{\rho\phi(a)}+\frac{\phi(a+2T^{1/2}\mu/\sigma)}{\phi(a)}}.$$

Substituting (C.38) into the numerator, and (C.37) and (C.38) into the denominator, we obtain that, for all fixed $x>0$,

$$\sup_{a:|a|\leq x}\left|\mu^{-1}\psi\left(a+T^{1/2}\mu/\sigma\right)-1\right|\to 0. \tag{C.39}$$

Since the integrand of (C.35) is always positive and even in a, it holds that, for all fixed $x>0$,

$$\begin{aligned}\left(S^{\mathrm{OPT}}\right)^2 &\geq \frac{\mu\rho N}{\sigma^2}\int_{|a-T^{1/2}\mu/\sigma|\leq x}\psi(a)(\phi(a-T^{1/2}\mu/\sigma)-\phi(a+T^{1/2}\mu/\sigma))da\\ &\geq \frac{\mu\rho N}{\sigma^2}\int_{|a-T^{1/2}\mu/\sigma|\leq x}\psi(a)\phi(a-T^{1/2}\mu/\sigma)(1-c_N)da\\ &\geq \frac{\mu\rho N}{\sigma^2}\int_{|a-T^{1/2}\mu/\sigma|\leq x}\mu\phi(a-T^{1/2}\mu/\sigma)(1-c_N)da\\ &\geq \frac{\mu^2\rho N}{\sigma^2}(1-c_N-2\Phi(-x)).\end{aligned}$$

Here the second inequality comes from (C.38), the third inequality is a result of (C.39), and the last inequality is obvious. Because this result holds for all fixed $x > 0$, the "only if" part is proved. ∎

C.4 Proof of Theorem 3

Given the length of the proof, a briefly explanation is warranted to clarify the key ideas and structure.

The whole proof is organized into 5 steps. Steps 1 - 4 demonstrate that the distance between the conditional expectation vector $\psi := \Sigma_u^{-1/2}\mathrm{E}(\alpha|\mathcal{G}) = (\psi(\check{s}_1),\ldots,\psi(\check{s}_N))^\intercal$, and the estimate $\breve{\psi} := (\breve{\psi}(\widehat{s}_1),\ldots,\breve{\psi}(\widehat{s}_N))^\intercal$, measured by L2 norm, is small compared to S^{OPT}. Based on the result, step 5 shows the gap between monotonicity-preserving estimate $\widehat{\psi}$ and ψ in L2 norm is also small compared to S^{OPT}. This leads to that the utility and Sharpe ratio generated by $\widehat{w}^{\mathrm{OPT}} = \kappa^{-1}\mathbb{M}_\beta\widehat{\Sigma}_u^{-1/2}\widehat{\psi}$ converges to, respectively, $(2\kappa)^{-1}\left(S^{\mathrm{OPT}}\right)^2$ and S^{OPT}, proved in the last step.

We note that, because of the rare and weak nature of alphas, $\mathrm{E}(\alpha_i|\mathcal{G})$ converges to zero in probability for each individual i, despite their large collective contribution to Sharpe ratio. Therefore, we need instead the L2 norm of errors involved in $\widehat{\psi}$ to converge to zero.

Step 1. Throughout the proof, we use the following notation, introduced in the main text of the paper,

$$\check{s}_i = T^{-1}\sum_{s\in\mathcal{T}}(s_i+\varepsilon_{i,s}),\quad \widehat{s}_i = \widehat{\alpha}_i/\widehat{\sigma}_i,\quad p(a) = \mathrm{E}(\phi_{1/T}(a-s_i)),\quad \psi(a) = \mathrm{E}(s_i|\check{s}_i = a).\ \text{(C.40)}$$

As in that statement, $p(a)$ is the density of $\check{s}_i$, and $\psi(a)$ is the expectation of s_i, conditional on $\check{s}_i = a$.

Intuitively, for assets with large $\check{s}_i$, $\check{s}_i$ is a relatively precisely estimate the true s_i. In contrast, for assets with small $\check{s}_i$, more likely $\check{s}_i$ is driven by noise. As a result, we introduce $B = \{i \leq N : |\check{s}_i| \leq \widetilde{k}_N T^{-1/2}\}$ to separate the two cases, where $\widetilde{k}_N = k_N^{-2}$. Moreover, we set

$\breve{\psi}$ and ψ as the N-dimensional vectors with entries $\breve{\psi}_i := \breve{\psi}(\widehat{s}_i)$ and $\psi_i := \psi(\check{s}_i)$. It holds that

$$\|\breve{\psi} - \psi\|^2 \leq \sum_{i \in B} (\breve{\psi}_i - \psi_i)^2 + \sum_{i \in B^c} (\breve{\psi}_i - \psi_i)^2. \tag{C.41}$$

The majority of the proof (steps 2 - 4) is to establish that $\widehat{\psi}$ constructed by us estimates conditional expectation vector ψ sufficiently precisely in the following sense:

$$\|\breve{\psi} - \psi\|^2 = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.42}$$

The last two steps prove optimality of our portfolio strategy based on the above result. We end this step by noting that Corollary 1 states

$$\|\psi\| = S(\mathcal{G}) = S^{\mathrm{OPT}} + o_{\mathrm{P}}(1). \tag{C.43}$$

Step 2. This step control the magnitude of $\sum_{i \in B} (\breve{\psi}_i - \psi_i)^2$ of (C.41). It does so by showing

$$\sum_{i \in B^c} (\psi_i - \check{s}_i)^2 = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2) \quad \text{and} \quad \sum_{i \in B^c} (\breve{\psi}_i - \check{s}_i)^2 = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.44}$$

Since $\psi_i := \psi(\check{s}_i)$, to bound $\sum_{i \in B^c} (\psi_i - \check{s}_i)^2$, we show that $|\psi(a) - a|$ is small. On the other hand, Tweedie's formula reads

$$\psi(a) - a = T^{-1} \frac{p'(a)}{p(a)}. \tag{C.45}$$

Moreover, we have, for all positive sequence b_N and all a,

$$\begin{aligned} |p'(a)| &\leq T \int |x - a| \phi_{1/T}(a - x) p_s(x) dx \\ &\leq b_N T \int_{|x-a| \leq b_N} \phi_{1/T}(a - x) p_s(x) dx + \sup_{x : |x - a| > b_N} T |x - a| \phi_{1/T}(a - x) \\ &\leq b_N T p(a) + \sup_{y : |y| > b_N} T^{3/2} |y| \exp(-T y^2 / 2). \end{aligned} \tag{C.46}$$

The second inequality comes from the $p_s(x)$, as a density, integrates to one. Then, choosing b_N that satisfies $b_N \gtrsim T^{-1/2} (\log N)^d$ with $d > 1/2$ and $b_N = o(T^{-1/2} \widetilde{k}_N)$, which is always possible, we obtain, for all a,

$$|p'(a)| \leq c_N T^{1/2} \widetilde{k}_N p(a) + c_N T N^{-2}. \tag{C.47}$$

It hence holds that

$$\max_i \frac{|p'(\check{s}_i)|}{p(\check{s}_i)} \lesssim_{\mathrm{P}} \sup_a \frac{|p'(a)|}{p(a)} \mathbb{1}_{\{p(a) \geq T^{1/2} N^{-3/2}\}} \leq c_N T^{1/2} \widetilde{k}_N. \tag{C.48}$$

The first inequality comes from (D.13) of Lemma D3. The second directly follows from (C.47). Combining (C.48) and (C.45), we obtain

$$\mathrm{P}((\check{s}_i - \psi(\check{s}_i))^2 \leq c_N T^{-1} \widetilde{k}_N^2, \forall i \leq N) \geq 1 - c_N. \tag{C.49}$$

As a result,

$$\sum_{i \in B^c} (\check{s}_i - \psi(\check{s}_i))^2 \lesssim_{\mathrm{P}} c_N T^{-1} \widetilde{k}_N^2 |B^c| \leq c_N \sum_{i \in B^c} \check{s}_i^2 \lesssim_{\mathrm{P}} c_N \sum_{i \in B^c} \psi(\check{s}_i)^2. \tag{C.50}$$

Here the first inequality is simply (C.49), the second holds since $\check{s}_i^2 \geq T^{-1} \widetilde{k}_N^2$ for all $i \in B^c$ by definition, and the last inequality is a direct implication of the first two. Given (C.50), we obtain the first part of (C.44) by noting $\sum_{i \in B^c} \psi(\check{s}_i)^2 \lesssim_{\mathrm{P}} (S^{\mathrm{OPT}})^2 + 1$ due to (C.43).

Now we establish the second part of (C.44). By construction we have

$$\breve{\psi}(a) - a = \frac{1 + k_N^2}{T} \frac{\widehat{p}'(a)}{\widehat{p}(a)}, \quad \text{with} \quad \widehat{p}(a) = \frac{1}{N k_N} \sum_i \phi_{1/T} \left(\frac{\widehat{s}_i - a}{k_N} \right). \tag{C.51}$$

Similar to (C.46), we have, for all positive sequence b_N and all a,

$$\begin{aligned} |\widehat{p}'(a)| &\leq \frac{T}{N k_N^2} \sum_i \frac{|\widehat{s}_i - a|}{k_N} \phi_{1/T} \left(\frac{\widehat{s}_i - a}{k_N} \right) \\ &\leq \frac{T}{N k_N^2} \sum_{i : |\widehat{s}_i - a|/k_N \leq b_N} \frac{|\widehat{s}_i - a|}{k_N} \phi_{1/T} \left(\frac{\widehat{s}_i - a}{k_N} \right) + \frac{T}{k_N^2} \sup_{i : |\widehat{s}_i - a|/k_N > b_N} \frac{|\widehat{s}_i - a|}{k_N} \phi_{1/T} \left(\frac{\widehat{s}_i - a}{k_N} \right) \\ &\leq \frac{T b_N}{k_N} \widehat{p}(a) + \frac{T}{k_N^2} \sup_{y : |y| > b_N} |y| \exp(-T y^2 / 2). \end{aligned}$$

Choosing b_N that satisfies $b_N \gtrsim T^{-1/2} (\log N)^d$ with $d > 1/2$ and $b_N = o(T^{-1/2} \widetilde{k}_N k_N)$, which is always possible, we obtain, for all a,

$$|\widehat{p}'(a)| \leq c_N T^{1/2} \widetilde{k}_N \widehat{p}(a) + c_N T^{1/2} N^{-2}. \tag{C.52}$$

Therefore, it holds that

$$\max_i \frac{|\widehat{p}'(\widehat{s}_i)|}{\widehat{p}(\widehat{s}_i)} \leq c_N T^{1/2} \widetilde{k}_N, \tag{C.53}$$

which comes from (C.52) and that $\widehat{p}(\widehat{s}_i) \geq \frac{1}{N k_N} \phi_{1/T}(0) \gtrsim \frac{\sqrt{T}}{N k_N}$ for all i. As a result, we obtain the second part of (C.44):

$$\sum_{i\in B^c}(\breve{\psi}_i - \breve{s}_i)^2 \le c_N T^{-1}|B^c|\widetilde{k}_N^2 + |B^c|\max_{i\le N}|\widehat{s}_i - \breve{s}_i|^2 \le c_N T^{-1}|B^c|\widetilde{k}_N^2 \lesssim_{\mathrm{P}} c_N(S^{\mathrm{OPT}})^2 + c_N.$$

Here the first inequality is simply substituting (C.53) into (C.51), the second inequality comes from $\max_{i\le N}|\widehat{s}_i - \breve{s}_i| \le c_N T^{-1/2}\widetilde{k}_N$ by Lemma D2, the last inequality holds by (C.43) and (the last two inequalities of) (C.50).

Step 3. To analyze $\sum_{i\in B}(\widehat{\psi}_i - \psi_i)^2$ of (C.41), we introduce an auxiliary function:

$$\bar{\psi}(a) = \frac{\int x\phi_{v^2/T}(a-x)p_s(x)dx}{\int \phi_{v^2/T}(a-x)p_s(x)dx}, \quad \text{with} \quad v := \sqrt{1+k_N^2}. \tag{C.54}$$

$\bar{\psi}(a)$ is essentially the expectation of s_i, conditional on $\breve{s}_i' = a$, where $\breve{s}_i' \eqsim \mathcal{N}(s_i, v^2)$, i.e., $\breve{s}_i'$ has slightly more noisy than $\breve{s}_i$. The goal is to establish

$$\sum_{i\in B}(\psi_i - \bar{\psi}(\breve{s}_i))^2 = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2) \quad \text{and} \quad \sum_{i\in B}(\breve{\psi}_i - \bar{\psi}(\breve{s}_i))^2 = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.55}$$

Then the triangle inequality would give us the desired bound on $\sum_{i\in B}(\breve{\psi}_i - \psi_i)^2$. The current step proves the first part, whereas the next step will be devoted to show the second part.

We use $\bar{p}(a)$ and $\bar{\pi}(a)$ to denote the denominator and numerator of $\bar{\psi}(a)$ as in (C.54), and use $\pi(a)$ to denote $\int x\phi_{1/T}(a-x)p_s(x)dx$ so that $\psi(a) = \pi(a)/p(a)$. The goal is to show that $\bar{p}(a)$ and $\bar{\pi}(a)$ are, respectively, close to $p(a)$ and $\pi(a)$. We first note that $\phi_{v^2}(y)$ and $\phi(y)$ are close in that, for all y,

$$\begin{aligned} |\phi_{v^2}(y) - \phi(y)| &\le \sup_{y:|y|\le k_N^{-1}} |\phi_{v^2}(y) - \phi(y)| + \sup_{y:|y|>k_N^{-1}} |\phi_{v^2}(y) + \phi(y)| \\ &\le c_N k_N^{-1}\phi(y) \sup_{y:|y|\le k_N^{-1}} |y/v - y| + c_N N^{-2} \le c_N\phi(y) + c_N N^{-2}. \end{aligned} \tag{C.56}$$

Here we use (D.17) of Lemma D4 (choose $j = 0$) and that $|v^{-1} - 1| \eqsim k_N^2$. We note $\phi_{1/T}(y) = \sqrt{T}\phi\left(\sqrt{T}y\right)$ (and symmetrically for $\phi_{v^2/T}(y)$). Hence, using (C.56), we directly obtain that, for all a,

$$\begin{aligned} |\bar{p}(a) - p(a)| &\le \int |\phi_{v^2/T}(a-x) - \phi_{1/T}(a-x)|p_s(x)dx \\ &\le c_N \int \phi_{1/T}(a-x)p_s(x)dx + c_N\sqrt{T}N^{-2} = c_N p(a) + c_N\sqrt{T}N^{-2}. \end{aligned} \tag{C.57}$$

Now we bound the difference $|\pi(a) - \bar{\pi}(a)|$. Because $p_s(x)$ is a even function, we note that, for all $a \ge 0$,

$$\pi(a) = \int_0^\infty x\bar{\phi}(|a|, x)p_s(x)dx, \quad \text{and} \quad \bar{\pi}(a) = \int_0^\infty x\bar{\phi}(a/v, x/v)p_s(x)dx, \tag{C.58}$$

where

$$\bar{\phi}(a, x) := \phi_{1/T}(a - x) - \phi_{1/T}(a + x) = \phi_{1/T}(a - x)(1 - e^{-2Txa}).$$

Since $|(1 - e^{-y}) - (1 - e^{-y/v^2})| \leq c_N(1 - e^y)$ for all $y \geq 0$, it follows from (C.56) and direct calculations that, for all $a \geq 0$ and $x \geq 0$,

$$|\bar{\phi}(a/v, x/v) - \bar{\phi}(a, x)| \leq c_N\bar{\phi}(a, x) + c_N\sqrt{T}N^{-2}. \tag{C.59}$$

Substituting (C.59) into (C.58), we obtain, for all $a \geq 0$,

$$|\bar{\pi}(a) - \pi(a)| \leq c_N|\pi(a)| + c_N\sqrt{T}N^{-2}\int_0^\infty xp_s(s)dx \leq |\pi(a)| + c_N\sqrt{T}N^{-2}. \tag{C.60}$$

Here the last inequality holds by $\mathrm{E}(|s|) \leq \sqrt{\mathrm{E}(s^2)} \leq c_N$ due to condition (a) of Assumption 2. Because $\pi(a)$ and $\bar{\pi}(a)$ are both odd functions in a due to that $p_s(x)$ is a even function of x, (C.60) apparently holds for all a.

To establish from (C.57) and (C.60) that $\bar{\psi}(a)$ and $\psi(a)$ are close, we set $A := \{a : |a| \leq \widetilde{k}_N T^{-1/2}, p(a) \geq \sqrt{T}N^{-3/2}\}$. Then we obtain that, for all $a \in A$,

$$\begin{aligned} |\bar{\psi}(a) - \psi(a)| &= \left|\frac{\bar{\pi}(a)}{\bar{p}(a)} - \frac{\pi(a)}{\bar{p}(a)}\right| + \left|\frac{\pi(a)}{\bar{p}(a)} - \frac{\pi(a)}{p(a)}\right| \\ &\leq (1 + c_N)\frac{|\bar{\pi}(a) - \pi(a)|}{p(a)} + c_N\frac{|\pi(a)|}{p(a)} \leq c_N\frac{N^{-2}}{p(a)} + c_N\psi(a). \end{aligned} \tag{C.61}$$

Here the first equality is obvious, the first inequality comes from the lower bound of $p(a)$ (by the definition of A) and (C.57), the second inequality is a result of (C.60). From (C.61), it follows that, for all a satisfying $a \in A$,

$$|\bar{\psi}(a) - \psi(a)|^2 \leq c_N\frac{N^{-2}}{p(a)} + c_N\psi(a)^2. \tag{C.62}$$

where we use Cauchy-Schwarz inequality and the lower bound of $p(a)$. Therefore, we arrive at

$$N\int_A |\bar{\psi}(a) - \psi(a)|^2 p(a)da \leq c_N + c_N N\int_{-\infty}^\infty \psi(a)^2 p(a)da \leq c_N + c_N(S^{\mathrm{OPT}})^2, \tag{C.63}$$

which comes from (C.62) and that $\int_A da \leq 2\widetilde{k}_N$. Therefore, using Chebyshev's inequality and comparing the definitions of sets A and B, we obtain

$$\begin{aligned}\sum_{i\in B}(\psi_i-\bar{\psi}(\check{s}_i))^2\mathbb{1}_{\{p(\check{s}_i)\geq\sqrt{T}N^{-3/2}\}} &\lesssim_{\mathrm{P}} \mathrm{E}\sum_{i\in B}(\psi_i-\bar{\psi}(\check{s}_i))^2\mathbb{1}_{\{p(\check{s}_i)\geq\sqrt{T}N^{-3/2}\}}\\ &= N\int_A|\bar{\psi}(a)-\psi(a)|^2p(a)da\leq c_N+c_N(S^{\mathrm{OPT}})^2,\end{aligned}$$

where the last inequality holds by (C.63). Given (D.13) of Lemma D3, we obtain the first part of (C.55).

Step 4. This step proves the second part of (C.55), i.e., we bound $\sum_{i\in B}(\breve{\psi}_i-\bar{\psi}(\check{s}_i))^2$. We introduce $\widetilde{p}(a)$ and $\widetilde{\psi}(a)$ that mimick $\widehat{p}(a)$ and $\widehat{\psi}(a)$ by replacing the data input $\check{s}_i$ with $\widehat{s}_i$:

$$\widetilde{p}(a)=\frac{1}{Nk_N}\sum_i\phi_{1/T}\left(\frac{\check{s}_i-a}{k_N}\right),\quad\text{and}\quad\widetilde{\psi}(a)=a+\frac{v^2}{T}\frac{\widetilde{p}'(a)}{\widetilde{p}(a)}. \tag{C.64}$$

Then we can decompose the quantity of interest:

$$\sum_{i\in B}(\breve{\psi}_i-\bar{\psi}(\check{s}_i))^2\leq\sum_{i\in B}(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2+\sum_{i\in B}(\breve{\psi}_i(\widehat{s}_i)-\widetilde{\psi}(\check{s}_i))^2. \tag{C.65}$$

We first show that $\sum_{i\in B}(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2$ is small. Since we have $\widetilde{p}(\check{s}_i)\geq\frac{1}{Nk_N}\phi_{1/T}(0)\gtrsim\frac{\sqrt{T}}{Nk_N}$ for all i, symmetric to the derivation of (C.53), we have

$$\max_i\frac{\widetilde{p}'(\check{s}_i)}{\widetilde{p}(\check{s}_i)}\leq c_NT^{1/2}\widetilde{k}_N. \tag{C.66}$$

On the other hand, symmetric to the derivation of (C.48), we obtain

$$\max_i\frac{\bar{p}'(\check{s}_i)}{\bar{p}(\check{s}_i)}\lesssim_{\mathrm{P}}\max_i\frac{\bar{p}'(a)}{\bar{p}(a)}\mathbb{1}_{\{p(a)\geq N^{-3/2}\}}\lesssim c_NT^{1/2}\widetilde{k}_N. \tag{C.67}$$

where for the second inequality we note $\bar{p}(a)\gtrsim p(a)$ for all a due to $1\leq v\lesssim 1$. Substituting (C.66) and (C.67) into the definitions of $\widetilde{\psi}(a)$ and $\bar{\psi}(a)$ ((C.64) and (C.54)), we obtain

$$\max_i|\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i)|\lesssim_{\mathrm{P}}c_N\widetilde{k}_NT^{-1/2}. \tag{C.68}$$

According to Lemma 3 of Brown and Greenshtein (2009), with the additional condition that $\max_{i\leq N}\sqrt{T}|s_i|=o(N^{d'})$ for every $d'>0$, we have (in our notation) that, for every $d>0$,

$$\mathrm{E}\left(\sum_iT(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2\right)\lesssim N^d.$$

A scrutiny of their proof of the lemma reveals that this additional condition is only indispensable (a) to derive three equalities: (48), (59), and (62) (the way it is used is similar across the three), and (b) to guarantee that $\max_{i\le N}\sqrt{T}\bar{\psi}(\check{s}_i)=o(N^d)$ for every $d>0$. In the absence of this additional condition, a weaker result holds: for every $d>d'>0$,

$$\mathrm{E}\left(\sum_i \min\{T(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2,N^{d'}\}\mathbb{1}_{\{\sqrt{T}|\check{s}_i|\le N^{d'},p(\check{s}_i)\ge\sqrt{T}N^{d'-1}\}}\right)\lesssim N^d. \tag{C.69}$$

(C.69) turns out sufficient for establishing a desired bound on $\sum_{i\in B}(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2$, which we demonstrate now. Then we have, for every $d>d'>0$,

$$\begin{aligned}\sum_{i\in B}T(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2 &\lesssim_{\mathrm{P}}\sum_{i\in B}\min\{T(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2,N^{d'}\}\\ \lesssim_{\mathrm{P}}\quad N^d+\sum_{i\in B}&\min\{T(\widetilde{\psi}(\check{s}_i)-\bar{\psi}(\check{s}_i))^2,N^{d'}\}\mathbb{1}_{\{p(\widetilde{z}_i)\ge\sqrt{T}N^{d'-1}\}}\lesssim N^d.\end{aligned} \tag{C.70}$$

Here the first inequality comes from (C.68), the second inequality comes from that $\mathrm{E}\left(\sum_{i\in B}\mathbb{1}_{\{p(\check{s}_i)<\sqrt{T}N^{d'-1}\}}\right)\lesssim\widetilde{k}_N N^{d'}$ (by the definition of set B), and the last is simply (C.69).

Next, we show that $\sum_{i\in B}(\breve{\psi}(\widehat{s}_i)-\widetilde{\psi}(\check{s}_i))^2$ is small. Lemma D2 states that

$$\max_{i\in B}|\widehat{s}_i-\check{s}_i|\lesssim_{\mathrm{P}}\chi_N:=N^{-1/2}(\epsilon_N+\mathrm{E}(s_j^2)^{1/2}),\quad\text{with}\quad\epsilon_N:=k_N^5. \tag{C.71}$$

Since $\breve{\psi}(\widehat{s}_i)$ and $\widetilde{\psi}(\check{s}_i)$ depends on $\{\widehat{s}_j\}$ and $\{\check{s}_j\}$ in the exactly same way, we can obtain the desired result by exploiting that such dependence is sufficiently "continuous". Concretely, we write, uniformly over $i\in B$,

$$\begin{aligned}&|\widehat{p}(\widehat{s}_i)-\widetilde{p}(\check{s}_i)|\\ \le\quad&\frac{1}{Nk_N}\sum_j\left|\phi_{1/T}\left(\frac{\widehat{s}_j-\widehat{s}_i}{k_N}\right)-\phi_{1/T}\left(\frac{\check{s}_j-\check{s}_i}{k_N}\right)\right|\lesssim_{\mathrm{P}}\sqrt{T}\chi_N k_N^{-2}\widetilde{p}(\check{s}_i)+\sqrt{T}N^{-2}k_N^{-1},\end{aligned} \tag{C.72}$$

$$\begin{aligned}&|\widehat{p}'(\widehat{s}_i)-\widetilde{p}'(\check{s}_i)|\\ \le\quad&\frac{T}{Nk_N^3}\sum_j\left|(\widehat{s}_j-\widehat{s}_i)\phi_{1/T}\left(\frac{\widehat{s}_j-\widehat{s}_i}{k_N}\right)-(\check{s}_j-\check{s}_i)\phi_{1/T}\left(\frac{\check{s}_j-\check{s}_i}{k_N}\right)\right|\\ \lesssim_{\mathrm{P}}\quad&T\chi_N k_N^{-4}\widetilde{p}(\widetilde{z}_i)+TN^{-2}k_N^{-1}.\end{aligned} \tag{C.73}$$

Here the first inequalities for both results hold by definition (note $\phi_{1/T}(a)=\sqrt{T}\phi(\sqrt{T}a)$ and $\phi'(a)=-a\phi(a)$). The second inequalities for both results comes from substituting (C.71) into (D.17) of Lemma D4. Since $\widetilde{p}(\check{s}_i)\ge\frac{1}{Nk_N}\phi_{1/T}(0)\gtrsim\frac{\sqrt{T}}{Nk_N}$ by definition, we obtain from (C.72) and (C.73) that

$$\max_{i\in B}\frac{|\widehat{p}(\widehat{s}_i)-\widetilde{p}(\breve{s}_i)|}{\widetilde{p}(\breve{s}_i)} \lesssim_{\mathrm{P}} \sqrt{T}\chi_N k_N^{-2}+N^{-1}\lesssim\sqrt{T}\chi_N k_N^{-2}, \tag{C.74}$$

$$\max_{i\in B}\frac{|\widehat{p}'(\widehat{s}_i)-\widetilde{p}'(\breve{s}_i)|}{\widetilde{p}(\breve{s}_i)} \lesssim_{\mathrm{P}} T\chi_N k_N^{-4}+\sqrt{T}N^{-1}\lesssim T\chi_N k_N^{-4}. \tag{C.75}$$

Then we have

$$\begin{aligned}&\max_{i\in B}\left|\frac{\widehat{p}'(\widehat{s}_i)}{\widehat{p}(\widehat{s}_i)}-\frac{\widetilde{p}'(\breve{s}_i)}{\widetilde{p}(\breve{s}_i)}\right|\\ \le\ &\max_{i\in B}\frac{\widetilde{p}(\breve{s}_i)}{\widehat{p}(\widehat{s}_i)}\frac{|\widehat{p}'(\widehat{s}_i)-\widetilde{p}'(\breve{s}_i)|}{\widetilde{p}(\breve{s}_i)}+\max_{i\in B}\frac{\widetilde{p}(\breve{s}_i)}{\widehat{p}(\widehat{s}_i)}\frac{\widetilde{p}'(\breve{s}_i)}{\widetilde{p}(\breve{s}_i)}\frac{|\widehat{p}(\widehat{s}_i)-\widetilde{p}(\breve{s}_i)|}{\widetilde{p}(\breve{s}_i)}\lesssim_{\mathrm{P}} T\chi_N k_N^{-4}.\end{aligned} \tag{C.76}$$

The first inequality is direct algebra. Substituting (C.66), (C.74), and (C.75) into the right-hand-side of the first inequality, we obtain the second inequality. Combining (C.71) and (C.76) with the definitions of $\widehat{\psi}$ and $\widetilde{\psi}$ ((C.51) and (C.64)), we obtain

$$\sum_{i\in B}(\breve{\psi}(\widehat{s}_i)-\widetilde{\psi}(\breve{s}_i))^2\le N\max_{i\in B}|\widehat{s}_i-\breve{s}_i|^2+\frac{N}{T^2}\max_{i\in B}\left|\frac{\widehat{p}'(\widehat{s}_i)}{\widehat{p}(\widehat{s}_i)}-\frac{\widetilde{p}'(\breve{s}_i)}{\widetilde{p}(\breve{s}_i)}\right|^2\lesssim_{\mathrm{P}} k_N^{-8}(\epsilon_N^2+\mathrm{E}(s_j^2)). \tag{C.77}$$

The goal is to show $\sum_{i\in B}(\breve{\psi}(\widehat{s}_i)-\widetilde{\psi}(\breve{s}_i))^2=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$, which is apparently true from (C.77) if $\mathrm{E}(s_j^2)\le\epsilon_N^2$. For the case $\mathrm{E}(s_j^2)>\epsilon_N^2$, we observe

$$\mathrm{E}(s_j^2)=\mathrm{E}(s_j^2\mathbb{1}_{\{\epsilon_N/2<|s_i|\le1\}})+\mathrm{E}(s_j^2\mathbb{1}_{\{|s_i|\le\epsilon_N/2\}})+\mathrm{E}(s_j^2\mathbb{1}_{\{|s_i|>1\}})\le\mathrm{P}(|s_i|>\epsilon_N/2)+\epsilon_N^2/4+c_NN^{-1},$$

where the last step comes from condition (a) of Assumption 2. We hence obtain $\mathrm{P}(|s_i|>\epsilon_N/2)\gtrsim\epsilon_N^2$, which futher indicates $\sum_i\mathbb{1}_{\{|s_i|\ge\epsilon_N/2\}}\gtrsim_{\mathrm{P}}N\epsilon_N^2$ (the sum follows binomial distribution with its standard deviation dominated by its mean). As a result, we write

$$N\epsilon_N^4\lesssim_{\mathrm{P}}\sum_i\epsilon_N^2\mathbb{1}_{\{|s_i|\ge\epsilon_N/2\}}\lesssim_{\mathrm{P}}\sum_i\breve{s}_i^2\mathbb{1}_{\{|\breve{s}_i|\ge\epsilon_N/4\}}\lesssim\sum_{i\in B^c}\breve{s}_i^2\lesssim_{\mathrm{P}}1+(S^{\mathrm{OPT}})^2. \tag{C.78}$$

Here the second inequality comes from $\breve{s}_i-s_i=\bar{\varepsilon}_i$ and $\max_i|\bar{\varepsilon}_i|\lesssim\sqrt{(\log N)/T}$ by the uniform bound on i.i.d normal variables. The thrid inequality holds by the definition of B, and the last inequality can be established from holds by (C.43) and (the last two inequalities of) (C.50). Since $\mathrm{E}(s_j^2)\le1+\mathrm{E}(s_j^2\mathbb{1}_{\{|s_i|>1\}})\lesssim1$ by condition (a) of Assumption 2, it follows from (C.78) that $k_N^{-8}\mathrm{E}(s_j^2)=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$. Given (C.77), we prove $\sum_{i\in B}(\breve{\psi}(\widehat{s}_i)-\widetilde{\psi}(\breve{s}_i))^2=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$. Substituting this result and (C.70) into (C.65), we obtain $\sum_{i\in B}(\breve{\psi}(\widehat{s}_i)-\widetilde{\psi}(\breve{s}_i))^2=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$, i.e., the second part of (C.55). Substituting (C.44) and (C.55) into (C.41), we finally establish (C.42).

Step 5. The goal of this step is to establish

$$\|\widehat{\psi} - \psi\| = o_{\mathrm{P}}(1 + S^{\mathrm{OPT}}). \tag{C.79}$$

We start by defining

$$l_j = \arg\min_{i \leq N} |\check{s}_i - \widehat{s}_j|. \tag{C.80}$$

Suppose $\check{s}_{l_i} > \check{s}_{l_j}$ and $\widehat{s}_i \leq \widehat{s}_j$ for some (i, j). According to (C.80), we have $|\check{s}_{l_i} - \widehat{s}_i| \leq |\check{s}_{l_j} - \widehat{s}_i|$ and $|\check{s}_{l_j} - \widehat{s}_j| \leq |\check{s}_{l_i} - \widehat{s}_j|$. The former would lead to $\check{s}_{l_i} + \check{s}_{l_j} < 2\widehat{s}_i$, and the latter would lead to $\check{s}_{l_i} + \check{s}_{l_j} > 2\widehat{s}_j$. They together contradict $\widehat{s}_i \leq \widehat{s}_j$. Hence we have $\check{s}_{l_i} \leq \check{s}_{l_j}$ if $\widehat{s}_i \leq \widehat{s}_j$ for all (i, j). Since $\psi(a)$ is increasing in a (see, e.g., Efron (2011)), letting $\psi_i^* = \psi(\check{s}_{l_i})$, we have $\psi_i^* \leq \psi_j^*$ if $\widehat{s}_i \leq \widehat{s}_j$ for all (i, j). As a result, we have

$$\|\widehat{\psi} - \breve{\psi}\| \leq \|\psi^* - \breve{\psi}\| \leq \|\psi^* - \psi\| + o_{\mathrm{P}}(1 + S^{\mathrm{OPT}}). \tag{C.81}$$

The first inequality comes from the definition of $\widehat{\psi}$ and that $\psi_i^* \leq \psi_j^*$ if $\widehat{s}_i \leq \widehat{s}_j$ for all (i, j) established above. The second inequality comes from (C.42) proved in steps 2 - 4.

Given (C.42) and (C.81), to obtain (C.79) we only need to bound $\|\psi^* - \psi\|$. For this purpose, we note that (C.80) leads to $|\check{s}_{l_i} - \widehat{s}_i| \leq |\check{s}_i - \widehat{s}_i|$ for all i. Applying Lemma D2, we obtain

$$\max_{1 \leq i \leq N} |\check{s}_{l_i} - \check{s}_i| \lesssim_{\mathrm{P}} c_N \widetilde{k}_N T^{-1/2}, \quad \max_{i \in B} |\check{s}_{l_i} - \check{s}_i| \lesssim_{\mathrm{P}} \chi_N, \tag{C.82}$$

where $\chi_N = N^{-1/2}(\epsilon_N + \mathrm{E}(s_j^2)^{1/2})$ with $\epsilon_N := k_N^5$. Similar to (C.50), we have

$$\sum_{i \in B^c} ((\check{s}_{l_i} - \psi(\check{s}_{l_i}))^2 + (\check{s}_{l_i} - \check{s}_i)^2) \lesssim_{\mathrm{P}} c_N T^{-1} \widetilde{k}_N^2 |B^c| \leq c_N \sum_{i \in B^c} \psi(\check{s}_i)^2 \lesssim_{\mathrm{P}} o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.83}$$

Here the first inequality comes from (C.49) and the first part of (C.82). The second inequality is already shown in (C.50). The last inequality is established right below (C.50). Combining (C.83) and the first part of (C.44), and applying the triangle inequality, we obtain

$$\sum_{i \in B^c} (\psi(\check{s}_{l_i}) - \psi(\check{s}_i))^2 = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.84}$$

Next, it holds that, uniformly over $i \in B$,

$$\begin{aligned}
|p(\check{s}_{l_i}) - p(\check{s}_i)| &\leq \int |\phi_{1/T}(\check{s}_{l_i} - x) - \phi_{1/T}(\check{s}_i - x)| p_s(x) ds \\
&\lesssim (T \log N)^{1/2} |\check{s}_{l_i} - \check{s}_i| p(\check{s}_i) + c_N \sqrt{T} N^{-2} \\
&\lesssim_{\mathrm{P}} (T \log N)^{1/2} \chi_N p(\check{s}_i) + c_N \sqrt{T} N^{-2}. \qquad (\mathrm{C.85})
\end{aligned}$$

$$
\begin{aligned}
|p(\check{s}_{l_i})' - p(\check{s}_i)'| &\leq T\int |(\check{s}_{l_i} - x)\phi_{1/T}(\check{s}_{l_i} - x) - (\check{s}_i - x)\phi_{1/T}(\check{s}_i - x)|p_s(x)ds \\
&\lesssim (T\log N)|\check{s}_{l_i} - \check{s}_i|p(\check{s}_i) + c_N T N^{-2} \\
&\lesssim_{\mathrm{P}} (T\log N)\chi_N p(\check{s}_i) + c_N T N^{-2}. \qquad \text{(C.86)}
\end{aligned}
$$

For both (C.85) and (C.86), the first inequalities holds by definition of $p(a)$, the second inequalities come from (D.17) of Lemma D4 (choose $j = 0$ and $j = 1$ respectively), and the last inequalities come from the second part of (C.82). Then we have, uniformly over i for which $\check{s}_i \in A$ (A is defined above (C.61)),

$$
\begin{aligned}
|\psi(\check{s}_{l_i}) - \psi(\check{s}_i)| &\leq |\check{s}_{l_i} - \check{s}_i| + \frac{1}{T}\frac{|p(\check{s}_{l_i})' - p(\check{s}_i)'|}{p(\check{s}_{l_i})} + \frac{|p(\check{s}_i)'|}{T}\left|\frac{1}{p(\check{s}_{l_i})} - \frac{1}{p(\check{s}_i)}\right| \\
&\lesssim_{\mathrm{P}} \chi_N + \chi_N + \frac{\chi_N}{\sqrt{T}}\frac{p(\check{s}_i)'}{p(\check{s}_i)} \lesssim \chi_N\left(\widetilde{k}_N + \sqrt{T}|\psi(\check{s}_i)|\right). \qquad \text{(C.87)}
\end{aligned}
$$

The first inequality is obvious. The second inequality comes from the second part of (C.82), (C.85), (C.86), that $p(\check{s}_i) \geq \sqrt{T}N^{-3/2}$ by the definition of A, and that $1/\sqrt{NT} \lesssim \chi_N$ as $T \gtrsim N^d$ for fixed $d > 1/2$ by assumption. The last inequality comes from (C.45) and $|\check{s}_i| \leq \widetilde{k}_N$ as $\check{s}_i \in A$. Therefore, we have

$$
\begin{aligned}
& \sum_{i\in B}(\psi(\check{s}_{l_i}) - \psi(\check{s}_i))^2 \mathbb{1}_{\{p(\check{s}_i)\geq\sqrt{T}N^{-3/2}\}} \\
= & \sum_{i:\check{s}_i\in A}(\psi(\check{s}_{l_i}) - \psi(\check{s}_i))^2 \lesssim_{\mathrm{P}} N\chi_N^2\widetilde{k}_N^2 + T\chi_N^2\sum_{i\in B}\psi(\widetilde{z}_i)^2 \\
\lesssim_{\mathrm{P}} & (\epsilon_N^2 + \mathrm{E}(s_j^2))\widetilde{k}_N^2 + o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2) = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \qquad \text{(C.88)}
\end{aligned}
$$

The equality is obvious and the first inequality comes from (C.87). The second inequality comes from (C.43) and that $T\chi_N^2 = o(1)$ due to $\mathrm{E}(s_j^2) = o(1)$ and $T \lesssim N^{d'}$ for fixed $d' < 1$. The last inequality is established in the analysis after (C.77). Combining (C.88), Lemma D3, and (C.84), we establish $\|\psi^* - \psi\| = o_{\mathrm{P}}(1 + S^{\mathrm{OPT}})$. Substituting this into (C.81) and using (C.42) lead to (C.79).

Step 6. This step combines (C.79) with (C.43) to prove that the utility and Sharpe ratio of the strategy $\widehat{w}^{\mathrm{OPT}}$ we construct achieves, respectively, $(2\kappa)^{-1}(S^{\mathrm{OPT}})^2$ and $(S^{\mathrm{OPT}})^2$ asymptotically, which, combined with the two equalities of (C.3) and Corollary 1, proves the theorem. For convenience, we write $\widehat{w}^{\mathrm{OPT}} = \kappa^{-1}\mathbb{M}_\beta\breve{w}$, where $\breve{w} := \widehat{\Sigma}_u^{-1/2}\widehat{\psi}$.

We first show that, under any sequence of probability measures $\{\mathrm{P}_N\}$ that satisfies all the conditions described in the statement of the theorem, and for all fixed positive ε,

$$\mathrm{P}_N(U(\widehat{w}^{\mathrm{OPT}}) =_\varepsilon U(\widetilde{w})) \to 1, \quad \text{and} \quad \mathrm{P}_N(S(\widehat{w}^{\mathrm{OPT}}) =_\varepsilon S(\widetilde{w})) \to 1. \tag{C.89}$$

The following derivation proceeds under any given sequence of probability measures and we omit the subscript N. Using condition (a) of Assumption 1 and (D.1) of Lemma D1, we have $\max_{i\le N} |\widehat{\sigma}_i/\sigma_i - 1| \lesssim_{\mathrm{P}} c_N$. As a result, we obtain $\|\Sigma^{1/2}\breve{w} - \widehat{\psi}\| \lesssim_{\mathrm{P}} c_N\|\widehat{\psi}\|$. Then, it follows from (C.79) and (C.43) that

$$\|\Sigma_u^{1/2}\breve{w} - \psi\| \le \|\Sigma^{1/2}\breve{w} - \widehat{\psi}\| + \|\widehat{\psi} - \psi\| \lesssim_{\mathrm{P}} c_N\|\psi\| + \|\widehat{\psi} - \psi\| = o_{\mathrm{P}}(1 + S^{\mathrm{OPT}}). \tag{C.90}$$

Hence we have

$$\begin{aligned} |(\breve{w}^\intercal\Sigma_u^{1/2} - \psi^\intercal)\psi| &\le \|\Sigma_u^{1/2}\breve{w} - \psi\|\|\psi\| = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2), & \text{(C.91)}\\ |\breve{w}^\intercal\Sigma_u\breve{w} - \psi^\intercal\psi| &\le \|\Sigma_u^{1/2}\breve{w} - \psi\|^2 + 2\|\Sigma_u^{1/2}\breve{w} - \psi\|\|\psi\| = o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). & \text{(C.92)}\end{aligned}$$

Here for both (C.91) and (C.92), the first inequalities come from Cauchy-Schwarz, whereas the last equalities come from (C.90) and (C.43). Further, substituting (C.43) into (C.91) and (C.92), we obtain

$$\breve{w}^\intercal\Sigma_u^{1/2}\psi = (S^{\mathrm{OPT}})^2 + o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2), \quad \breve{w}^\intercal\Sigma_u\breve{w} = (S^{\mathrm{OPT}})^2 + o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.93}$$

Next, we note that

$$\begin{aligned} \|\mathbb{P}_\beta\breve{w}\| &\le \|\mathbb{P}_\beta\Sigma_u^{-1/2}\psi\| + \|\mathbb{P}_\beta(\breve{w} - \Sigma_u^{-1/2}\psi)\| \\ &= \|\mathbb{P}_\beta\Sigma_u^{-1}\widetilde{\alpha}\| + (\breve{w} - \Sigma_u^{-1/2}\psi)^\intercal\mathbb{P}_\beta(\breve{w} - \Sigma_u^{-1/2}\psi) \\ &\le \|\mathbb{P}_\beta\Sigma_u^{-1}\widetilde{\alpha}\| + \|\breve{w} - \Sigma_u^{-1/2}\psi\| = o_{\mathrm{P}}(1 + S^{\mathrm{OPT}}). \end{aligned} \tag{C.94}$$

To obtain the last equality, we note that the first term can be bounded using (C.18) as $\breve{w}' := \frac{1}{\kappa}\mathbb{P}_\beta\Sigma_u^{-1}\widetilde{\alpha}$ and that the second term can be bounded using (C.90) and $1 \lesssim_{\mathrm{P}} \lambda_{\min}(\Sigma_u)$. As a result, we obtain

$$\breve{w}^\intercal\mathbb{M}_\beta\Sigma_u\mathbb{M}_\beta\breve{w} = \breve{w}^\intercal\Sigma_u\breve{w} + \breve{w}^\intercal\mathbb{P}_\beta\Sigma_u\mathbb{P}_\beta\breve{w} - 2\breve{w}^\intercal\Sigma_u\mathbb{P}_\beta\breve{w} = \breve{w}^\intercal\Sigma_u\breve{w} + o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2). \tag{C.95}$$

For the last equality, we use (C.92), (C.94), and $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$ by condition (a) of Assumption 1. Similarly, we write

$$\breve{w}^\intercal\mathbb{M}_\beta\widetilde{\alpha} = \breve{w}^\intercal\widetilde{\alpha} - \breve{w}^\intercal\mathbb{P}_\beta\widetilde{\alpha} = \breve{w}^\intercal\widetilde{\alpha} + o_{\mathrm{P}}(1 + (S^{\mathrm{OPT}})^2), \tag{C.96}$$

where for the last equality, we use (C.94) and $\|\widetilde{\alpha}\|^2 \lesssim_{\mathrm{P}} \widetilde{\alpha}\Sigma_u^{-1}\widetilde{\alpha} = (S^{\mathrm{OPT}})^2$.

Now we obtain the first part of (C.89):

$$\begin{aligned} U(\widehat{w}^{\mathrm{OPT}}) &= \frac{1}{\kappa}\breve{w}^\intercal \mathbb{M}_\beta \alpha - \frac{1}{2\kappa}\breve{w}^\intercal \mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta \breve{w} = \frac{1}{\kappa}\breve{w}^\intercal \widetilde{\alpha} - \frac{1}{2\kappa}\breve{w}^\intercal \Sigma_u \breve{w} + o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2) \\ &= \frac{1}{2\kappa}(S^{\mathrm{OPT}})^2 + o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2) = U(\widetilde{w}) + o_{\mathrm{P}}(1+U(\widetilde{w})). \end{aligned}$$

where the first equality holds by definition, the second comes from (C.95) and (C.96), the third equality comes from (C.93), and for the last one we use (C.3) and Corollary 1. We can further conclude that, when S^{OPT} does not vanish,

$$\begin{aligned} S(\widehat{w}^{\mathrm{OPT}}) &= \frac{(\widehat{w}^{\mathrm{OPT}})^\intercal \alpha}{\sqrt{(\widehat{w}^{\mathrm{OPT}})^\intercal \Sigma \widehat{w}^{\mathrm{OPT}}}} = \frac{\breve{w}^\intercal \mathbb{M}_\beta \alpha}{\sqrt{\breve{w}^\intercal \mathbb{M}_\beta \Sigma \mathbb{M}_\beta \breve{w}}} \\ &= \frac{\breve{w}^\intercal \mathbb{M}_\beta \widetilde{\alpha}}{\sqrt{\breve{w}^\intercal \mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta \breve{w}}} + o_{\mathrm{P}}(1) = \frac{\breve{w}^\intercal \widetilde{\alpha}}{\sqrt{\breve{w}^\intercal \Sigma_u \breve{w}}} + o_{\mathrm{P}}(1+S^{\mathrm{OPT}}) \\ &= S^{\mathrm{OPT}} + o_{\mathrm{P}}(1+S^{\mathrm{OPT}}) = S(\widetilde{w}) + o_{\mathrm{P}}(1+S(\widetilde{w})). \end{aligned} \tag{C.97}$$

The first two equalities hold by definition. The third one comes (C.9) and that $\mathbb{M}_\beta \breve{w}$ is $\mathcal{G}$-measurable and factor-neutral. The fourth one comes from (C.95), (C.96), and the second part of (C.93). The fifth equality comes from (C.93) ($\widetilde{\alpha} = \Sigma_u^{1/2}\psi$). For the last one we use (C.3) and Corollary 1. Because $\widehat{w}^{\mathrm{OPT}}$ is $\mathcal{G}$-measurable and $\beta^\intercal \widehat{w}^{\mathrm{OPT}} = 0$, Theorem 1 applies. We hence have $S(\widehat{w}^{\mathrm{OPT}}) \leq S(\widetilde{w}) + o_{\mathrm{P}}(1+S(\widetilde{w}))$. Because $-S(\widehat{w}^{\mathrm{OPT}})$ is the Sharpe ratio generated by $-\widehat{w}^{\mathrm{OPT}}$, we also have $-S(\widehat{w}^{\mathrm{OPT}}) \leq S(\widetilde{w}) + o_{\mathrm{P}}(1+S(\widetilde{w}))$. As a result, when S^{OPT} does vanish, we have $S(\widehat{w}^{\mathrm{OPT}}) = o_{\mathrm{P}}(1)$. We also have $S(\widetilde{w}) = o_{\mathrm{P}}(1)$ per Corollary 1. Therefore, given (C.97) and using the subsequence argument (see, e.g., Andrews and Cheng (2012)), we have

$$S(\widehat{w}^{\mathrm{OPT}}) = S(\widetilde{w}) + o_{\mathrm{P}}(1+S(\widetilde{w})),$$

which gives us the second part of (C.89).

Now we prove the theorem using (C.89). Suppose the first convergence of the theorem does not hold. Then we have, for some fixed $\varepsilon > 0$,

$$\liminf_{N\to 0} \inf_{\mathrm{P}\in\mathbb{P}} \mathrm{P}(U(w) =_\varepsilon U(\widetilde{w})) < 1.$$

Hence there is a subsequence $\{\mathrm{P}_{N_k}\}_{k\geq 1}$ with $\mathrm{P}_{N_k} \in \mathbb{P}$ for each $k \geq 1$ such that, for some fixed $\varepsilon > 0$,

$$\lim_{k\to\infty} \mathrm{P}_{N_k}(U(w) =_\varepsilon U(\widetilde{w})) < 1. \tag{C.98}$$

Since $\mathbb{P}$ only collects probability measures under which all the conditions described in the statement of the theorem hold, (C.98) gives a contradiction to (C.89), which means that the first convergence of the theorem must be true. A symmetric argument establishes the second convergence and proves the theorem.

C.5 Proof of Proposition 1

By definition we have

$$(\widehat{S}^{\star})^2 = \alpha^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\mathbb{M}_{\beta}\alpha + 2\alpha^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\mathbb{M}_{\beta}\bar{u} + \bar{u}^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\mathbb{M}_{\beta}\bar{u} - T^{-1}N.$$

We start with the analysis of $\alpha^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\mathbb{M}_{\beta}\alpha$. From (D.1) of Lemma D1, it follows

$$\|\widehat{\Sigma}_u - \Sigma_u\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} \sqrt{T^{-1}\log N}. \tag{C.99}$$

As a result, noting $\mathrm{P}(0 \leq \mathbb{M}_{\beta} \leq \mathbb{I}_N) \to 1$ and $\mathrm{P}(\Sigma_u \approx \mathbb{I}_N) \to 1$ by the assumption $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$ and $\lambda_{\min}(\beta^{\intercal}\beta) \gtrsim_{\mathrm{P}} N$, and recalling $(S^{\star})^2 = \alpha^{\intercal}\Sigma_u^{-1}\alpha$, we have

$$|\alpha^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\mathbb{M}_{\beta}\alpha - \alpha^{\intercal}\mathbb{M}_{\beta}\Sigma_u^{-1}\mathbb{M}_{\beta}\alpha| \lesssim_{\mathrm{P}} \sqrt{T^{-1}\log N}(S^{\star})^2. \tag{C.100}$$

On the other hand, it holds that

$$\begin{aligned} |\alpha^{\intercal}\mathbb{M}_{\beta}\Sigma_u^{-1}\mathbb{M}_{\beta}\alpha - \alpha^{\intercal}\Sigma_u^{-1}\alpha| &\leq \alpha^{\intercal}\mathbb{P}_{\beta}\Sigma_u^{-1}\mathbb{P}_{\beta}\alpha + 2\sqrt{(\alpha^{\intercal}\Sigma_u^{-1}\alpha)(\alpha^{\intercal}\mathbb{P}_{\beta}\Sigma_u^{-1}\mathbb{P}_{\beta}\alpha)} \\ &\lesssim_{\mathrm{P}} \alpha^{\intercal}\mathbb{P}_{\beta}\alpha + \sqrt{(\alpha^{\intercal}\Sigma_u^{-1}\alpha)(\alpha^{\intercal}\mathbb{P}_{\beta}\alpha)} \\ &\lesssim_{\mathrm{P}} N^{-1}\|\alpha^{\intercal}\beta\|^2 + \sqrt{N^{-1}(\alpha^{\intercal}\Sigma_u^{-1}\alpha)\|\alpha^{\intercal}\beta\|^2} \\ &\lesssim_{\mathrm{P}} \mathrm{E}(s_i^2) + S^{\star}\mathrm{E}(s_i^2)^{1/2}. \end{aligned} \tag{C.101}$$

Here the first inequality comes from Cauchy-Schwarz inequality. The second comes from $\mathrm{P}(\Sigma_u \approx \mathbb{I}_N) \to 1$ and $\mathbb{P}_{\beta}^2 = \mathbb{P}_{\beta}$. We obtain the third line by using $\lambda_{\min}(\beta^{\intercal}\beta) \gtrsim N$. The last line holds because of Chebyshev's inequality and that $\mathrm{E}(\|\alpha^{\intercal}\beta\|^2|\beta,\Sigma_u) \lesssim N\|\beta\|_{\mathrm{MAX}}^2\mathrm{E}(\alpha_i^2|\Sigma_u) \leq N\|\beta\|_{\mathrm{MAX}}^2\lambda_{\max}(\Sigma_u)\mathrm{E}(s_i^2)$ by condition (a) of Assumption 1 and condition (b) of Assumption 2.

On the other hand, because of the assumption $\mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|\geq c_N\}}) \leq c_N N^{-1}$, we have

$$\mathrm{E}\left|\alpha^{\intercal}\Sigma_u^{-1}\alpha - \sum_i s_i^2\mathbb{1}_{\{|s_i|<c_N\}}\right| = \mathrm{E}\left|\sum_i s_i^2\mathbb{1}_{\{|s_i|\geq c_N\}}\right| = o(1),$$

which, by Markov's inequality, leads to

$$\alpha^{\intercal}\Sigma_u^{-1}\alpha = \sum_i s_i^2\mathbb{1}_{\{|s_i|<c_N\}} + o_{\mathrm{P}}(1).$$

Moreover, it holds that

$$\mathrm{Var}\left|\sum_i s_i^2 \mathbb{1}_{\{|s_i|<c_N\}}\right| \leq \sum_i \mathrm{E}(s_i^4 \mathbb{1}_{\{|s_i|<c_N\}}) \leq c_N^2 \sum_i \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|<c_N\}}).$$

Using Chebyshev's inequality, we obtain

$$(S^\star)^2 = \alpha^\intercal \Sigma_u^{-1}\alpha \geq \sum_i s_i^2 \mathbb{1}_{\{|s_i|<c_N\}} \gtrsim_{\mathrm{P}} \sum_i \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|<c_N\}}) \geq N\mathrm{E}(s_i^2) + o(1).$$

Combining this result with (C.101) and (C.100), we have

$$\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha = (S^\star)^2 + o_{\mathrm{P}}\left(\sqrt{T^{-1}\log N}((S^\star)^2+1)\right). \tag{C.102}$$

Next, we study $\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u}$. It holds that

$$\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \lesssim \alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha \sqrt{\bar{u}^\intercal \bar{u}} \lesssim \alpha^\intercal \alpha \sqrt{\bar{u}^\intercal \bar{u}} = O_{\mathrm{P}}(((S^\star)^2+1)T^{-1/2}). \tag{C.103}$$

The first inequality comes from Cauchy-Schwarz. The second inequality holds because $\mathrm{P}(\mathbb{M}_\beta \precsim \mathbb{I}_N) \to 1$, $\mathbb{M}_\beta^2 = \mathbb{M}_\beta$, and $\mathrm{P}(\widehat{\Sigma}_u \precsim \mathbb{I}_N) \to 1$ due to $\mathrm{P}(\Sigma_u \precsim \mathbb{I}_N) \to 1$ and (C.99). The third inequality holds by $\mathrm{P}(\widehat{\Sigma}_u \precsim \mathbb{I}_N) \to 1$ as well.

Now we analyze $\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1}N$. We write

$$\begin{aligned} N = \mathrm{tr}(\widehat{\Sigma}_u^{-1}\widehat{\Sigma}_u) &= \sum_{i\leq N} (\widehat{\Sigma}_u^{-1})_{i,i}\left(T^{-1}\sum_{s\in T}(\mathbb{M}_\beta u_s)_i^2 - (\mathbb{M}_\beta \bar{u})_i^2\right) \\ &= T^{-1}\sum_{s\in T} u_s^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_s - \bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \\ &= T^{-1}\sum_{s\in T} u_s^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_s + O_{\mathrm{P}}(N/T). \end{aligned} \tag{C.104}$$

The last line comes from $\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \lesssim_{\mathrm{P}} \bar{u}^\intercal \bar{u}$ because of $\mathbb{M}_\beta^2 = \mathbb{M}_\beta$ and $\mathrm{P}(\widehat{\Sigma}_u \precsim \mathbb{I}_N) \to 1$. Furthermore, I have

$$\begin{aligned} \bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - \bar{u}^\intercal \widehat{\Sigma}_u^{-1} \bar{u} &\leq 2|\bar{u}^\intercal \widehat{\Sigma}_u^{-1} \mathbb{P}_\beta \bar{u}| + \bar{u}^\intercal \mathbb{P}_\beta \widehat{\Sigma}_u^{-1} \mathbb{P}_\beta \bar{u} \\ &\lesssim_{\mathrm{P}} \sqrt{\bar{u}^\intercal \bar{u}}\sqrt{\bar{u}^\intercal \mathbb{P}_\beta \bar{u}} + \bar{u}^\intercal \mathbb{P}_\beta \bar{u} \lesssim_{\mathrm{P}} N^{1/2}/T. \end{aligned} \tag{C.105}$$

Here we obtain the second inequality using $\mathrm{P}(\widehat{\Sigma}_u \precsim \mathbb{I}_N) \to 1$ and the last inequality using $\mathrm{P}(\mathbb{P}_\beta \lesssim \mathbb{I}_N) \to 1$. Similarly, it holds that

$$\begin{aligned} & T^{-2}\sum_{t\in T}(u_t^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1}\mathbb{M}_\beta u_t - u_t^\intercal \widehat{\Sigma}_u^{-1} u_t) \\ = \quad & T^{-2}\sum_{t\in T}\left(2\sqrt{u_t^\intercal u_t}\sqrt{u_t^\intercal \mathbb{P}_\beta \widehat{\Sigma}_u^{-1}\mathbb{P}_\beta u_t} + u_t^\intercal \mathbb{P}_\beta \widehat{\Sigma}_u^{-1}\mathbb{P}_\beta u_t\right) \\ \lesssim_{\mathrm{P}} \quad & T^{-2}\sum_{t\in T}\left(\sqrt{u_t^\intercal u_t}\sqrt{u_t^\intercal \mathbb{P}_\beta u_t} + u_t^\intercal \mathbb{P}_\beta u_t\right) \lesssim_{\mathrm{P}} \frac{N^{1/2}}{T}. \end{aligned} \tag{C.106}$$

From (C.104), (C.105), and (C.106), it directly follows

$$\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1}\mathbb{M}_\beta \bar{u} - T^{-1}N = \bar{u}^\intercal \widehat{\Sigma}_u^{-1}\bar{u} - T^{-2}\sum_{s\in T} u_s^\intercal \widehat{\Sigma}_u^{-1} u_s + O_{\mathrm{P}}(N^{1/2}/T + N/T^2). \tag{C.107}$$

On the other hand, we have

$$\widehat{\Sigma}_u^{-1} = -\Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u) + \Sigma_u^{-2}\widehat{\Sigma}_u^{-1}(\widehat{\Sigma}_u - \Sigma_u)^2.$$

It then follows from (C.99) and $\mathrm{P}(\Sigma_u \approx \mathbb{I}_N) \to 1$ that

$$\begin{aligned} \bar{u}^\intercal \widehat{\Sigma}_u^{-1}\bar{u} &= -\bar{u}^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)\bar{u} + O_{\mathrm{P}}(T^{-1}(\log N)\bar{u}^\intercal \bar{u}) \\ &= -\bar{u}^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)\bar{u} + O_{\mathrm{P}}(T^{-2}N\log N). \end{aligned} \tag{C.108}$$

Similarly, we have

$$T^{-2}\sum_{t\in T} u_t^\intercal \widehat{\Sigma}_u^{-1} u_t = -T^{-2}\sum_{t\in T} u_t^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)u_t + O_{\mathrm{P}}(T^{-2}N\log N). \tag{C.109}$$

Substituting (C.108) and (C.109) into (C.107), we have

$$\begin{aligned} \bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1}\mathbb{M}_\beta \bar{u} - T^{-1}N \quad = \quad & -\bar{u}^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)\bar{u} + T^{-2}\sum_{t\in T} u_t^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)u_t \\ & + O_{\mathrm{P}}(T^{-1}N^{1/2} + T^{-2}N\log N). \end{aligned} \tag{C.110}$$

Now we analyze $\widehat{\Sigma}_u$. We write

$$\begin{aligned} (\widehat{\Sigma}_u)_{i,i} \quad = \quad & (T^{-1}uu^\intercal)_{i,i} + (\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta)_{i,i} \\ & + (\mathbb{P}_\beta T^{-1}uu^\intercal)_{i,i} + (T^{-1}uu^\intercal \mathbb{P}_\beta)_{i,i} + (\mathbb{P}_\beta T^{-1}uu^\intercal \mathbb{P}_\beta)_{i,i}. \end{aligned} \tag{C.111}$$

From the uniform bound on i.i.d. random variables and $\|\mathbb{P}_\beta\|_{\mathrm{MAX}} \lesssim N^{-1}$ by the assumption $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$ and $\lambda_{\min}(\beta^\intercal \beta) \gtrsim_{\mathrm{P}} N$, it follows

$$\|\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} \|\bar{u}\bar{u}^\intercal\|_{\mathrm{MAX}} = \|\bar{u}\|_{\mathrm{MAX}}^2 \lesssim T^{-1}\log N.$$

Using $\mathrm{P}(\Sigma_u \approx \mathbb{I}_N) \to 1$, this gives

$$\sum_{i\leq N} |\bar{u}_i^2 (\Sigma_u^{-2})_{i,i} (\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta)_{i,i}| \lesssim_{\mathrm{P}} T^{-1}(\log N) \sum_{i\leq N} \bar{u}_i^2 \lesssim_{\mathrm{P}} T^{-2} N \log N, \tag{C.112}$$

and

$$T^{-2}\sum_{t\in T}\sum_{i\leq N} |u_{i,t}^2 (\Sigma_u^{-2})_{i,i} (\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta)_{i,i}| \lesssim_{\mathrm{P}} T^{-3}(\log N) \sum_{t\in T}\sum_{i\leq N} u_{i,t}^2 \lesssim_{\mathrm{P}} T^{-2} N \log N. \tag{C.113}$$

Further, we obtain

$$\begin{aligned} & \sum_{i\leq N, j\leq K} \mathrm{E}(|(T^{-1}uu^\intercal\beta)_{i,j}||\beta,\Sigma_u) \\ \leq & \sum_{i\leq N, j\leq K} \sqrt{\mathrm{E}((T^{-1}uu^\intercal)_{i,i}|\Sigma_u)\mathrm{E}((T^{-1}\beta^\intercal uu^\intercal\beta)_{j,j}|\beta,\Sigma_u)} \\ = & \sum_{i\leq N, j\leq K} \sqrt{(\Sigma_u)_{i,i}(\beta^\intercal\Sigma_u\beta)_{j,j}} \leq N^{3/2}K\|\beta\|_{\mathrm{MAX}}\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} N^{3/2}. \end{aligned} \tag{C.114}$$

The first inequality comes from Cauchy-Schwarz. The last inequality directly follows from condition (a) of Assumption 1. Similarly,

$$\begin{aligned} & \sum_{j\leq K, k\leq K} \mathrm{E}(|(T^{-1}\beta uu^\intercal\beta)_{j,k}||\beta,\Sigma_u) \\ \leq & \sum_{j\leq K, k\leq K} \sqrt{\mathrm{E}((T^{-1}\beta^\intercal uu^\intercal\beta)_{j,j}|\beta,\Sigma_u)\mathrm{E}((T^{-1}\beta^\intercal uu^\intercal\beta)_{j,j}|\beta,\Sigma_u)} \\ = & \sum_{j\leq K, k\leq K} \sqrt{(\beta^\intercal\Sigma_u\beta)_{j,j}(\beta^\intercal\Sigma_u\beta)_{k,k}} \leq KN\|\beta\|_{\mathrm{MAX}}\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} N. \end{aligned} \tag{C.115}$$

From (C.114) and (C.115), it directly follows

$$\sum_{i\leq N, j\leq K} |(T^{-1}uu^\intercal\beta)_{i,j}| \lesssim_{\mathrm{P}} N^{3/2}, \qquad \sum_{j\leq K, k\leq K} |(T^{-1}\beta uu^\intercal\beta)_{j,k}| \lesssim_{\mathrm{P}} N. \tag{C.116}$$

Using (C.116), and noting $\max_{i\leq N}|\bar{u}_i| \lesssim_{\mathrm{P}} \sqrt{T^{-1}\log N}$ from the uniform bound on i.i.d. random variables and $\|(\beta^\intercal\beta)^{-1}\beta\|_{\mathrm{MAX}} \lesssim N^{-1}$ by assumption, we obtain

$$\sum_{i\leq N} |\bar{u}_i^2 (T^{-1}uu^\intercal\mathbb{P}_\beta)_{i,i}|$$

$$
\begin{aligned}
&\leq \sum_{i\leq N, j\leq K} |\bar{u}_i^2 (T^{-1}uu^\intercal\beta)_{i,j}| \|(\beta^\intercal\beta)^{-1}\beta\|_{\mathrm{MAX}} \\
&\lesssim_{\mathrm{P}} N^{-1}T^{-1}\log N \sum_{i\leq N, j\leq K} |(T^{-1}uu^\intercal\beta)_{i,j}| \lesssim_{\mathrm{P}} N^{1/2}T^{-1}\log N, \qquad \text{(C.117)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{i\leq N} |\bar{u}_i^2 (\mathbb{P}_\beta(T^{-1}uu^\intercal)\mathbb{P}_\beta)_{i,i}| \\
&\leq \sum_{i\leq N, j\leq K, k\leq K} \bar{u}_i^2 |(T^{-1}\beta uu^\intercal\beta)_{j,k}| \|(\beta^\intercal\beta)^{-1}\beta\|_{\mathrm{MAX}}^2 \\
&\lesssim_{\mathrm{P}} N^{-1}T^{-1}\log N \sum_{j\leq K, k\leq K} |(T^{-1}\beta uu^\intercal\beta)_{j,k}| \lesssim_{\mathrm{P}} T^{-1}\log N. \qquad \text{(C.118)}
\end{aligned}
$$

Symmetric reasoning leads to

$$
\begin{aligned}
\frac{1}{T^2}\sum_{s\in T}\sum_{i\leq N} |u_{i,s}^2 (T^{-1}uu^\intercal\mathbb{P}_\beta)_{i,i}| &\lesssim_{\mathrm{P}} N^{1/2}T^{-1}, \qquad \text{(C.119)}\\
\frac{1}{T^2}\sum_{s\in T}\sum_{i\leq N} |u_{i,s}^2 (\mathbb{P}_\beta(T^{-1}uu^\intercal)\mathbb{P}_\beta)_{i,i}| &\lesssim_{\mathrm{P}} T^{-1}. \qquad \text{(C.120)}
\end{aligned}
$$

Substituting (C.112), (C.113), (C.117), (C.119), (C.118), and (C.120) into (C.110) and (C.111), we obtain

$$
\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1}N = -T^{-2}\sum_{i:i\leq N} A_i + O_{\mathrm{P}}(T^{-1}N^{1/2}\log N + T^{-2}N\log N). \qquad \text{(C.121)}
$$

Here and only here we use short-hand notation

$$
A_i = \sum_{t\in T}\sum_{t'\in T:t'\neq t} (\Sigma_u^{-2})_{i,i}(T^{-1}uu^\intercal - 2\Sigma_u)_{i,i} u_{i,t}u_{i,t'}.
$$

Since A_i is i.i.d. across i, we only need to analyze it for a single i. It obviously holds that $\mathrm{E}(A_i|\Sigma_u) = 0$. We also note $\mathrm{E}(((T^{-1}uu^\intercal - 2\Sigma_u)_{i,i})^2 u_{i,t}u_{i,t'}u_{i,s}u_{i,s'}|\Sigma_u) = 0$ unless two elements of $\{t, t', s, s'\}$ are the same, and $\mathrm{E}(((T^{-1}uu^\intercal - 2\Sigma_u)_{i,i})^2 u_{i,t}u_{i,t'}u_{i,s}u_{i,s'}|\Sigma_u) \lesssim T^{-2}\mathrm{E}(u_{i,t}^8|\Sigma_u)$ unless elements of $\{t, t', s, s'\}$ only take two different values. Then we obtain

$$
\mathrm{E}(A_i^2|\Sigma_u) \lesssim T^2(\Sigma_u^{-4})_{i,i}\mathrm{E}(u_{i,t}^8|\Sigma_u).
$$

It hence follows that

$$T^{-2}\sum_{i:i\leq N} A_i \lesssim_{\mathrm{P}} T^{-1}N^{1/2}\mathrm{E}(u_{i,t}^8(\Sigma_u^{-4})_{i,i}) \lesssim T^{-1}N^{1/2}, \tag{C.122}$$

where the last inequality comes from that $\varepsilon_{i,t}$ has finite eighth moment by condition (a) of Assumption 2. Subsituting (C.122) into (C.121), we obtain

$$\bar{u}^{\intercal}\mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\mathbb{M}_{\beta}\bar{u} - T^{-1}N = O_{\mathrm{P}}(T^{-1}N^{1/2}\log N + T^{-2}N\log N). \tag{C.123}$$

Combining (C.102), (C.103), and (C.123), and noting $N^{1/2}T \leq c_N$ and $T \lesssim N$ by assumption, we obtain

$$\begin{aligned}(\widehat{S}^{\star})^2 &= (S^{\star})^2 + o_{\mathrm{P}}\left(T^{-1/2}\sqrt{\log N}((S^{\star})^2+1) + T^{-1}N^{1/2}\log N\right)\\ &= (S^{\star})^2 + o_{\mathrm{P}}(T^{-1}N^{1/2}\log N((S^{\star})^2+1)).\end{aligned}$$

Therefore, we obtain, under $S^{\star} \geq C$,

$$(\widehat{S}^{\star})^2 = (S^{\star})^2(1+o_{\mathrm{P}}(T^{-1}N^{1/2}\log N)) \quad \Longrightarrow \quad \frac{\widehat{S}^{\star}-S^{\star}}{S^{\star}} = o_{\mathrm{P}}(T^{-1}N^{1/2}\log N).$$

And, under $S^{*} \leq c_N$, we have

$$(\widehat{S}^{\star})^2 = (S^{\star})^2 + o_{\mathrm{P}}(T^{-1}N^{1/2}\log N) \quad \Longrightarrow \quad \widehat{S}^{\star} - S^{\star} = o_{\mathrm{P}}\left(\sqrt{T^{-1}N^{1/2}\log N}\right).$$

We note by construction $(\widetilde{S}^{\star})^2 = (\widehat{S}^{\star})^2 + N/T$. Then, under $S^{\star} \geq C$, it holds that

$$(\widehat{S}^{\star})^2 = (S^{\star})^2 + \frac{N}{T} + (S^{\star})^2 o_{\mathrm{P}}\left(\frac{\sqrt{N}\log N}{T}\right) \quad \Longrightarrow \quad \frac{\widehat{S}^{\star} - \sqrt{(S^{\star})^2+N/T}}{S^{\star}} = o_{\mathrm{P}}\left(\frac{\sqrt{N}\log N}{T}\right).$$

Similarly, under $S^{*} \leq c_N$, we have

$$(\widehat{S}^{\star})^2 = (S^{\star})^2 + \frac{N}{T} + o_{\mathrm{P}}\left(\frac{\sqrt{N}\log N}{T}\right) \quad \Longrightarrow \quad \widehat{S}^{\star} - \sqrt{(S^{\star})^2+N/T} = o_{\mathrm{P}}\left(\frac{\sqrt{N}\log N}{T}\right).$$

The proof concludes.

C.6 Proof of Proposition B1

We note that all the assumptions in Theorem 1, other than the factor-neutrality, are assumed here as well. Therefore, equation (C.8) in the proof of Theorem 1 stays valid, since factor-neutrality is used only after the derivation of (C.8). Applying Cauchy-Schwarz inequality to (C.8), we obtain

$$|S(w)| \leq \sqrt{\mathrm{E}(r_{t+1}|\mathcal{G})^\intercal \Sigma^{-1} \mathrm{E}(r_{t+1}|\mathcal{G})} + o_{\mathrm{P}}(1). \tag{C.124}$$

On the other hand, it implies by Woodbury matrix identity and from the fact that $\Sigma = \beta\Sigma_v\beta^\intercal + \Sigma_u$,

$$\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\beta(\Sigma_v^{-1} + \beta^\intercal\Sigma_u^{-1}\beta)^{-1}\beta^\intercal\Sigma_u^{-1}. \tag{C.125}$$

By direct calculations, we have

$$\beta^\intercal\Sigma^{-1}\beta = ((\beta^\intercal\Sigma_u^{-1}\beta)^{-1} + \Sigma_v)^{-1}.$$

Let $H_1 = (\beta^\intercal\Sigma_u^{-1}\beta)^{-1}$ and $H_2 = \Sigma_v$, and using the fact that $(H_1 + H_2)^{-1} - H_2^{-1} = -(H_1 + H_2)^{-1}H_1H_2^{-1}$, we have

$$\beta^\intercal\Sigma^{-1}\beta - \Sigma_v^{-1} = -((\beta^\intercal\Sigma_u^{-1}\beta)^{-1} + \Sigma_v)^{-1}(\beta^\intercal\Sigma_u^{-1}\beta)^{-1}\Sigma_v^{-1}.$$

Therefore, using the fact that $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$ and that $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$ by assumption, we have

$$\lambda_{\max}((\beta^\intercal\Sigma_u^{-1}\beta)^{-1}) = \lambda_{\min}^{-1}(\beta^\intercal\Sigma_u^{-1}\beta) \leq \lambda_{\min}^{-1}(\beta^\intercal\beta)\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} N^{-1}. \tag{C.126}$$

Also, note that $\lambda_{\max}(\Sigma_v^{-1}) = \lambda_{\min}^{-1}(\Sigma_v) \lesssim 1$, and that

$$\lambda_{\max}(((\beta^\intercal\Sigma_u^{-1}\beta)^{-1} + \Sigma_v)^{-1}) = \lambda_{\min}^{-1}((\beta^\intercal\Sigma_u^{-1}\beta)^{-1} + \Sigma_v) \leq \lambda_{\min}^{-1}(\Sigma_v) \lesssim 1,$$

we have

$$\|\beta^\intercal\Sigma^{-1}\beta - \Sigma_v^{-1}\| \lesssim_{\mathrm{P}} N^{-1},$$

which in turn leads to

$$\gamma^\intercal\beta^\intercal\Sigma^{-1}\beta\gamma = \gamma^\intercal\Sigma_v^{-1}\gamma + o_{\mathrm{P}}(1). \tag{C.127}$$

Next, we show

$$\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\beta\gamma = o_{\mathrm{P}}(1). \tag{C.128}$$

Notice that $\mathrm{E}(\mathrm{E}(\alpha|\mathcal{G})|\Sigma, \beta) = \mathrm{E}(\alpha|\Sigma, \beta) = \mathrm{E}(\alpha|\Sigma) = 0$ (by conditions (a) and (b) of Assumption 1), and that, conditionally on (Σ, β), $\mathrm{E}(\alpha_i|\mathcal{G})$ is independent across i as demonstrated in the beginning of the proof of Theorem 1. Therefore,

$$\mathrm{E}((\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\beta\gamma)^2|\Sigma, \beta) \leq \sum_{i\leq N} \mathrm{E}(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma, \beta) \max_{j\leq N}(\gamma^\intercal\beta^\intercal\Sigma^{-1})_j^2. \tag{C.129}$$

On the other hand, from (C.125), we obtain

$$\gamma^\intercal\beta^\intercal\Sigma^{-1} = \gamma^\intercal\Sigma_v^{-1}(\Sigma_v^{-1} + \beta^\intercal\Sigma_u^{-1}\beta)^{-1}\beta^\intercal\Sigma_u^{-1}.$$

Because of $\lambda_{\min}(\Sigma_v) \gtrsim 1$, $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$, $\|\Sigma_u\|_{\mathrm{MAX}} \leq \|\Sigma_u\| \lesssim_{\mathrm{P}} 1$, $\lambda_{\min}(\Sigma_u) \gtrsim_{\mathrm{P}} 1$, Σ_u is diagonal, and (C.126), we have

$$\|\gamma^\intercal\beta^\intercal\Sigma^{-1}\|_{\mathrm{MAX}} \lesssim \|(\Sigma_v^{-1} + \beta^\intercal\Sigma_u^{-1}\beta)^{-1}\|\|\beta^\intercal\Sigma_u^{-1}\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} \lambda_{\max}((\beta^\intercal\Sigma_u^{-1}\beta)^{-1}) \lesssim_{\mathrm{P}} N^{-1}.$$

Hence, we have, for all positive fixed ϵ,

$$\mathrm{P}(|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\beta\gamma| \geq \epsilon|\Sigma,\beta) \leq \mathrm{E}((\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\beta\gamma)^2|\Sigma,\beta)/\epsilon^2 = o_{\mathrm{P}}(1), \tag{C.130}$$

where the last equality comes from (C.129) and that $\mathrm{E}\left(\sum_{i\leq N}\mathrm{E}(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma,\beta)\right) \leq \sum_{i\leq N}\mathrm{E}(\alpha_i^2) = o(N)$ by condition (a) of Assumption 1. Since $\mathrm{P}(|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\beta\gamma| \geq \epsilon|\Sigma,\beta) \leq 1$ are uniformly bounded for all N (by definition), we obtain by taking expectations on both sides of (C.130) that, for all positive fixed ϵ,

$$\mathrm{P}(|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\beta\gamma| \geq \epsilon) = o(1),$$

which is equivalent to (C.128).

Finally, we derive

$$\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma^{-1}\mathrm{E}(\alpha|\mathcal{G}) = \mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\mathrm{E}(\alpha|\mathcal{G}) + o_{\mathrm{P}}(1). \tag{C.131}$$

Following the same derivation for (C.129), we obtain

$$\mathrm{E}(|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\beta|_{\mathrm{F}}^2|\Sigma,\beta) \leq \sum_{i\leq N}\mathrm{E}(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma,\beta)\max_j(\Sigma_u^{-1}\beta\beta^\intercal\Sigma_u^{-1})_{j,j}.$$

Because $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$ and $\lambda_{\min}(\Sigma_u) \gtrsim_{\mathrm{P}} 1$, we have

$$\max_j(\Sigma_u^{-1}\beta\beta^\intercal\Sigma_u^{-1})_{j,j} \lesssim \|\Sigma_u^{-1}\beta\|_{\mathrm{MAX}}^2 \lesssim_{\mathrm{P}} 1.$$

Then given the above result that $\mathrm{E}\left(\sum_{i\leq N}\mathrm{E}(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma,\beta)\right) = o(N)$, we obtain that $\mathrm{E}(|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\beta|_{\mathrm{F}}^2|\Sigma,\beta) = o_{\mathrm{P}}(N)$. Therefore, similar to the derivation of (C.128), we obtain

$$|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\beta|_{\mathrm{F}}^2 = o_{\mathrm{P}}(N).$$

On the other hand, using (C.126), we obtain

$$\|(\Sigma_v^{-1} + \beta^\intercal\Sigma_u^{-1}\beta)^{-1}\| = \lambda_{\min}^{-1}(\Sigma_v^{-1} + \beta^\intercal\Sigma_u^{-1}\beta) \leq \lambda_{\max}((\beta^\intercal\Sigma_u^{-1}\beta)^{-1}) \lesssim_{\mathrm{P}} N^{-1}. \tag{C.132}$$

Then, using (C.132), we have

$$\begin{aligned}&\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\beta(\Sigma_v^{-1}+\beta^\intercal\Sigma_u^{-1}\beta)^{-1}\beta^\intercal\Sigma_u^{-1}\mathrm{E}(\alpha|\mathcal{G})\\ \leq&|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\beta|_{\mathrm{F}}^2\|(\Sigma_v^{-1}+\beta^\intercal\Sigma_u^{-1}\beta)^{-1}\|=o_{\mathrm{P}}(1),\end{aligned}$$

and hence, in light of (C.125), we obtain (C.131).

Given that $\mathrm{E}(r_{t+1}|\mathcal{G})=\mathrm{E}(\alpha|\mathcal{G})+\beta\gamma$, it follows from (C.127), (C.128), and (C.131) that

$$\mathrm{E}(r_{t+1}|\mathcal{G})^\intercal\Sigma^{-1}\mathrm{E}(r_{t+1}|\mathcal{G})=\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\mathrm{E}(\alpha|\mathcal{G})+\gamma^\intercal\Sigma_v^{-1}\gamma+o_{\mathrm{P}}(1).$$

In light of (C.124), we conclude the proof.

C.7 Proof of Propositions B2, B3, and B4

We present the notation that is used throughout Section C.7 and Appendix E to facilitate the exposition of our proofs. We write

$$\zeta:=\sqrt{T}\mu/\sigma,\quad \check{z}_i:=\sqrt{T}\check{s}_i,\quad \widehat{z}_i:=\sqrt{T}\widehat{s}_i.$$

ζ represents the signal strengh of our alphas. $\widehat{z}_i$ is simply the t-statistic and $\check{z}_i$ is hypothetically what the t-statistic would be in the absence of risk factors. Next, we introduce soft- and hard-thresholding functions:

$$\widetilde{\psi}_1(a,\lambda):=\mathrm{sgn}(a)(|a|-\lambda),\quad \widetilde{\psi}_2(a,\lambda):=a\mathbb{1}_{\{|a|\geq\lambda\}}.$$

Then, for $q\in\{1,2\}$ and $i\in\{1,\ldots,N\}$,

$$\widetilde{\psi}_{q,i}(\lambda):=\widetilde{\psi}_q(\check{z}_1,\lambda),\quad \widehat{\psi}_{q,i}(\lambda):=\widetilde{\psi}_q(\widehat{z}_i,\lambda).$$

Further, $\widetilde{\psi}_q(\lambda)$ and $\widehat{\psi}_q(\lambda)$ to represent, repectively, the vectors $(\widetilde{\psi}_q(\check{z}_1,\lambda),\ldots,\widetilde{\psi}_q(\check{z}_N,\lambda))^\intercal$ and $(\widehat{\psi}_{q,1}(\lambda),\ldots,\widehat{\psi}_{q,N}(\lambda))^\intercal$. Based on $\widehat{\psi}_q(\lambda)$, we introduce $\widehat{w}'_q(\lambda):=\widehat{\Sigma}_u^{-1/2}\widehat{\psi}_q(\lambda)$. The purpose of doing so is that $\widehat{w}'_q(\lambda)$, with appropriate choice of λ, incorporates the weights of Lasso, CSR, and BH methods all as special cases. Lastly, we define

$$\widehat{S}'_q(\lambda):=\frac{\widehat{w}'_q(\lambda)\mathbb{M}_\beta\alpha}{\sigma\|\mathbb{M}_\beta\widehat{w}'_q(\lambda)\|},\quad S_q(\lambda):=\frac{N^{1/2}\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)\psi_i)}{\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)^{1/2}}.$$

$\widehat{S}'_q(\lambda)$ is the Sharpe ratio generated by the strategy $\widehat{w}'_q(\lambda)$. $S_q(\lambda)$, again with appropriate choice of λ, incorporates as special cases S^{CSR}, S^{BH}, and S^{LASSO} (the probability limits of

Sharpe ratios) all as special cases.

C.7.1 Proof of Proposition B2

By definition it holds that $\widehat{\Sigma}_u^{-1}\widehat{\alpha} = \breve{w}_1(\lambda)$ under $\lambda = 0$. Hence we have $\widehat{S}^{\text{CSR}} = \widehat{S}_1(\lambda)$, choosing $\lambda = 0$. In other words, CSR is a special case of Lasso. On the other hand, by definition we have, choosing $\lambda = 0$,

$$S_\lambda^{\text{Lasso}} = N^{1/2}\mathrm{E}(\check{s}_i s_i)\mathrm{E}(\check{s}_i^2)^{-1/2}. \tag{C.133}$$

Moreover, we can write

$$\sqrt{T} \times \mathrm{E}(\check{s}_i s_i) = \frac{\rho\mu}{2\sigma}\int a(\phi(a-\zeta) - \phi(a+\zeta))da = \frac{\rho\mu\zeta}{\sigma}, \tag{C.134}$$

$$T \times \mathrm{E}(\check{s}_i^2) = \int a^2(\phi(a) + \rho\phi(a-\zeta))da = 1 + \rho(1+\zeta^2) = (1+o(1))(1+\rho\zeta^2). \tag{C.135}$$

For both lines, the first equalities come from the distribution of α_i imposed by Example 1, $s_i = \alpha_i\sigma^{-1}$ by definition, and $\sqrt{T}(\check{s}_i - s_i)$ is standard normal. The second equalities hold by direct calculations. The last equality of the second line comes from $\rho \to 0$. Therefore, it holds that, choosing $\lambda = 0$,

$$\widehat{S}^{\text{CSR}} = \widehat{S}_1(\lambda) = S_\lambda^{\text{Lasso}} + o_{\mathrm{P}}(1 + S^{\text{OPT}}) = S^{\text{CSR}} + o_{\mathrm{P}}(1 + S^{\text{OPT}}). \tag{C.136}$$

Here the second equality comes from $\widehat{S}_1(\lambda) = S_\lambda^{\text{Lasso}} + o_{\mathrm{P}}(1 + S^{\text{OPT}})$ as in Proposition B4. The third equality comes from (C.133), (C.134), (C.135), $\zeta = \sqrt{T}\mu/\sigma$ by definition, and that $|S^{\text{CSR}}| \leq S^{\text{OPT}} + o_{\mathrm{P}}(1)$ by Theorem 1 and Corollary 1.

Now we verify the sufficient and necessary condition. We utilize the last equality of (C.136), i.e., that CSR is a special case of Lasso, and the optimality conditions for Lasso provided in Proposition B4. We start with the "if" part. Invoking the classic subsequence argument, we only need to consider the two cases $\sqrt{T}\mu/\sigma \to 0$ and $\sqrt{\rho T}\mu/\sigma \to \infty$ separately. The former case is obvious since $\lambda = 0$ certainly satisfies $T^{1/2}\lambda \to 0$ condition required by Proposition B4. For the latter case, noting that $\sqrt{\rho T}\mu/\sigma \to \infty$ always leads to $\sqrt{T}\mu/\sigma - \sqrt{-2\log\rho} \to \infty$ due to $-\log\rho \precsim \log N$ by assumption, we only need to verify that, under $\lambda = 0$, we have both $\sqrt{T}(\mu/\sigma - \lambda) \to \infty$ and $\frac{\phi(\sqrt{T}\lambda)}{\rho(1+T\lambda^2)T(\mu/\sigma-\lambda)^2} \to 0$ as in Proposition B4. Both are obviously true.

Next we show the "only if" part. As $\sqrt{T}\mu/\sigma \to 0$ is violated, we only need to show that, if $\sqrt{\rho T}\mu/\sigma \to \infty$ is violated, we would not have $\frac{\phi(\sqrt{T}\lambda)}{\rho(1+T\lambda^2)T(\mu/\sigma-\lambda)^2} \to 0$ under $\lambda = 0$. This is also obvious. The proof ends.

C.7.2 Proof of Proposition B3

Proof of Proposition B3, Part 1. In this part we prove $\widehat{S}_\tau^{\mathrm{BH}} = S_\tau^{\mathrm{BH}} + o_{\mathrm{P}}(1 + S^{\mathrm{OPT}})$. By assumption there exists fixed $d > 0$ such that $\mu \leq N^{-d}$. Then there exists fixed (d_1, d_2, d_3) such that $0 < d_1 < d_2 < d_3 < \min\{d, 1\}$. We also set $c_x = \sqrt{2(1-x)\log N}$.

We let $\breve{z} = -\Phi^{-1}(p_{(\widehat{k})}/2)$ and $\breve{z}' = -\Phi^{-1}(p_{(\widehat{k}+1)}/2)$, where we recall Φ is the standard normal cdf. In other words, $\breve{z}$ and $\breve{z}'$ are the t-statistics whose p-values, calculated based on standard normal distribution, are $p_{(\widehat{k})}$ and $p_{(\widehat{k}+1)}$. We further define, for $j \in \{0, +, -\}$,

$$m_j(a) = \sum_{i \in H_j} \mathbb{1}_{\{|\widehat{z}_i| \geq a\}}, \quad \check{m}_j(a) = \sum_{i \in H_j} \mathbb{1}_{\{|\check{z}_i| \geq a\}},$$

where $H_0 = \{i \leq N : \alpha_i = 0\}$, $H_+ = \{i \leq N : \alpha_i = \zeta\}$, $H_- = \{i \leq N : \alpha_i = -\zeta\}$, and $\widehat{z}_i = T^{1/2}\widehat{s}_i$ is the t-statistic of stock i. From the definitions of $\breve{z}$ and $\breve{z}'$, we obtain

$$\frac{2N\Phi(-\breve{z})}{\sum_{j \in \{0,+,-\}} m_j(\breve{z})} \leq \tau, \quad \frac{2N\Phi(-\breve{z}')}{\sum_{j \in \{0,+,-\}} m_j(\breve{z}')} > \tau, \quad \sum_{j \in \{0,+,-\}} m_j(\breve{z}') = \sum_{j \in \{0,+,-\}} m_j(\breve{z}) + 1. \tag{C.137}$$

Given (E.74) and the monotonicity of $m_j(z)$ in z, there exists a positive sequence $b_N = o\left(1/\sqrt{\log N}\right)$ such that, with high probability, uniformly over x, and for $j \in \{0, +, -\}$,

$$\check{m}_j(x + b_N) \leq m_j(x) \leq \check{m}_j(x - b_N). \tag{C.138}$$

Further, noting $\check{z}_i$ is i.i.d. across i, we obtain from equation (13) of Liu and Shao (2014) that, for all deterministic sequences (a_N, a'_N) satisfying $N\Phi(-a_N) \to \infty$ and $\rho N\Phi(\zeta - a'_N) \to \infty$,

$$\sup_{0 \leq z \leq a_N} \left| \frac{\check{m}_0(z)}{2N\Phi(-z)} - 1 \right| = o_{\mathrm{P}}(1), \quad \sup_{0 \leq z \leq a'_N} \left| \frac{2\check{m}_\pm(z)}{\rho N\Phi(\zeta - z) + \rho N\Phi(-\zeta - z)} - 1 \right| = o_{\mathrm{P}}(1). \tag{C.139}$$

Using (E.15) and $-\Phi^{-1}(N^{-1}) \asymp \sqrt{\log N}$, we note that, for $b_N = o\left(1/\sqrt{\log N}\right)$, $\Phi(-z \pm b_N) = \Phi(-z)(1 + o(1))$ uniformly over $|z| \leq -\Phi^{-1}(N^{-1})$. Hence we obtain from (C.138) and (C.139) that, for all deterministic sequences (a_N, a'_N) satisfying $N\Phi(-a_N) \to \infty$ and $\rho N\Phi(\zeta - a'_N) \to \infty$,

$$\sup_{0 \leq z \leq a_N} \left| \frac{m_0(z)}{2N\Phi(-z)} - 1 \right| = o_{\mathrm{P}}(1), \quad \sup_{0 \leq z \leq a'_N} \left| \frac{2m_\pm(z)}{\rho N\Phi(\zeta - z) + \rho N\Phi(-\zeta - z)} - 1 \right| = o_{\mathrm{P}}(1). \tag{C.140}$$

In addition, it follows from (E.15) and the second part of (E.1) that, uniformly over $z \geq 0$,

$$\frac{\rho\Phi(\zeta - z)}{\Phi(-z)} \approx h(z), \quad \text{with} \quad h(z) := \frac{z}{1 + (z-\zeta)_+} e^{(z-a^*)\zeta}. \tag{C.141}$$

Here a^* is introduced in the statement of Lemma E1, so is ζ^* which appears below. We now establish that, for all fixed (x, x') satisfying $0 < x' < x \leq 2$, if $\sup_{0 \leq z \leq c_x} h(z) \gtrsim 1$, then

$$\inf_{z \geq c_{x'}} h(z) \to \infty. \tag{C.142}$$

Since $h(z) \lesssim e^{\log z + (z-a^*)\zeta}$ by definition, it holds that, if $\sup_{0 \leq z \leq c_x} h(z) \gtrsim 1$, there exists $0 \leq a_N \leq c_x$ such that $h(a_N) \gtrsim 1$ and thereby

$$e^{\log a_N + (a_N - a^*)\zeta} \gtrsim 1.$$

Because $a^*\zeta - \log a_N \gtrsim \log N$ by the first part of (E.1), $\zeta^* \gtrsim \sqrt{\log N}$ due to our assumption on ρ, and $a_N = o(\log N)$, we have $\zeta \gtrsim a_N^{-1} \log N \gtrsim \sqrt{\log N}$. For all z satisfying $z - a_N \geq c_{x'} - c_x \gtrsim \sqrt{\log N}$, it holds that $h(z) \gtrsim h(a_N) z^{-1} e^{(z-a_N)\zeta} \to \infty$ as $\zeta \gtrsim \sqrt{\log N}$. We hence establish (C.142).

We first suppose $z^* \geq c_{d_2}$. From (C.141) and the definition of z^* (B.1) we have $h(z^*) \approx 1$. Then, since $d_2 < d_3$, it follows from (C.142) (by contradiction) that

$$\sup_{0 \leq z \leq c_{d_3}} \frac{\rho\Phi(\zeta - z)}{\Phi(-z)} \to 0. \tag{C.143}$$

Using (C.143), and applying (C.140) and the monotonicity of $m_j(z)$ and $\Phi(-z)$ in z, it holds that, for all (random) sequence z satisfying $0 \leq z \leq c_{d_3}$,

$$m_0(z) = 2N\Phi(-z)(1 + o_{\mathrm{P}}(1)), \quad m_{\pm}(z) = o_{\mathrm{P}}(N\Phi(-z)). \tag{C.144}$$

Here we also use $\rho\Phi(-\zeta - z) \leq \Phi(-z)$. Comparing (C.144) with (C.137), we have that $\mathrm{P}(\breve{z} \leq c_{d_3}) \to 0$. We hence have, with high probability,

$$\mu^2 \rho N \Phi(\zeta - \breve{z}) \leq \mu^2 \rho N \Phi(\zeta - c_{d_3}) \leq \mu^2 N \Phi(-c_{d_3}) = o_{\mathrm{P}}(\mu^2 N^d) = o_{\mathrm{P}}(1). \tag{C.145}$$

Here the second inequality comes from (C.143), the first equality comes from $d_3 < d$, and the last equality comes from $\mu \leq N^{-d}$. (The situation where $\breve{z}$ does not exists is equivalent with $\breve{z}$ being so large that $\sup_i |\widehat{z}_i| < \breve{z}$ and is hence a special case.) Given (C.145), Lemma E7 applies and we obtain $\widehat{S}_\tau^{\mathrm{BH}} = \widetilde{S}_2(\breve{z}) = o_{\mathrm{P}}(1)$.

On the other hand, we have

$$
\begin{aligned}
S_\tau^{\mathrm{BH}} &= \frac{N^{1/2}\mathrm{E}(\widetilde{\psi}_{2,i}(z^*)s_i)}{\mathrm{E}(\widetilde{\psi}_{2,i}(z^*)^2)^{1/2}} \leq \frac{\mu N^{1/2}\mathrm{E}(|\widetilde{\psi}_{2,i}(z^*)|\mathbb{1}_{\{\alpha_i\neq 0,|\check{z}_i|\geq z^*\}})}{\sigma\mathrm{E}(\widetilde{\psi}_{2,i}(z^*)^2)^{1/2}} \\
&\leq \frac{\mu}{\sigma}\sqrt{N\mathrm{E}(\mathbb{1}_{\{\alpha_i\neq 0,|\check{z}_i|\geq z^*\}})} \leq \frac{\mu}{\sigma}\sqrt{2\rho N\Phi(\zeta - z^*)} \lesssim \mu\sqrt{N\Phi(-z^*)} = o(1).
\end{aligned}
$$

The second inequality comes from Cauchy-Schwarz inequality, the last inequality comes from (B.1), and the last equality comes $z^* \geq c_{d_2}$ and $N\Phi(-c_{d_2}) = o_{\mathrm{P}}(N^d)$ as $d_2 < d$. Hence we prove $\widehat{S}_\tau^{\mathrm{BH}} - S_\tau^{\mathrm{BH}} = o_{\mathrm{P}}(1)$ under $z^* \geq c_{d_2}$.

Now suppose $z^* \leq c_{d_2}$. Then it follows from (C.141), (B.1) and (C.142) that, as $d_1 < d_2$,

$$
\inf_{z\geq c_{d_1}} \frac{\rho\Phi(\zeta - z)}{\Phi(-z)} \to \infty. \tag{C.146}
$$

Since $d_1 > 0$, we have $N\Phi(-c_{d_1}) \to \infty$. Given (C.146), we further have $\rho N\Phi(\zeta - c_{d_1}) \to \infty$. Applying (C.140), and using (C.146), it holds that

$$
m_0(c_{d_1}) = o_{\mathrm{P}}(\rho\Phi(\zeta - c_{d_1})), \quad m_\pm(c_{d_1}) = \rho\Phi(\zeta - c_{d_1})(1 + o_{\mathrm{P}}(1)). \tag{C.147}
$$

Let $\widetilde{h}(z) := 2N\Phi(-z)/\sum_{j\in\{0,+,-\}} m_j(z)$. From (C.147), we know that $h(c_{d_1}) = o_{\mathrm{P}}(1)$. Since $\widetilde{h}(0) = 1$, we know that with high probability $\breve{z}$ exists and satisfy $0 \leq \breve{z} \leq c_{d_1}$. Since $\Phi(-z)$ is decreasing in z, we have $N\Phi(-\breve{z}) \to \infty$ and $\rho N\Phi(\zeta - \breve{z}) \to \infty$ in probability. Therefore, we obtain from (C.140) that

$$
m_0(\breve{z}) = 2N\Phi(-\breve{z})(1 + o_{\mathrm{P}}(1)), \quad m_\pm(\breve{z}) = \frac{\rho}{2}N(\Phi(\zeta - \breve{z}) + \Phi(-\zeta - \breve{z}))(1 + o_{\mathrm{P}}(1)). \tag{C.148}
$$

Since $\breve{z}' \leq \breve{z}$, (C.148) would still hold if all $\breve{z}$ are replaced by $\breve{z}'$. Hence, substituting (C.148) back into (C.137), and noting $\Phi(-\zeta - z) \leq \Phi(-z)$, we have

$$
\frac{2(1-\tau)\Phi(-\breve{z})}{\tau\rho\Phi(\zeta - \breve{z})} = 1 + o_{\mathrm{P}}(1). \tag{C.149}
$$

Next, using (B.1) and (E.15), we note that $z^* \leq c_{d_2}$ leads to that $\zeta \gtrsim \sqrt{\log N}$ (see a few lines after (C.142)). As a result, using (E.15), and comparing (B.1) and (C.149), we have

$$
|\breve{z} - z^*| = o_{\mathrm{P}}\left(1/\sqrt{\log N}\right). \tag{C.150}
$$

We finally obtain

$$
\widehat{S}_\tau^{\mathrm{BH}} = \frac{\widehat{w}_q'(T^{-1/2}\breve{z})^\intercal \mathbb{M}_\beta \alpha}{\sigma\|\mathbb{M}_\beta \widehat{w}_q'(T^{-1/2}\breve{z})\|} = S_2(z^*) + o_{\mathrm{P}}(S^{\mathrm{OPT}} + 1) = S_\tau^{\mathrm{BH}} + o_{\mathrm{P}}(S^{\mathrm{OPT}} + 1).
$$

where the first equality holds by definition, the second comes from (C.150), $\rho N\Phi(\zeta - z^*) \geq \rho N\Phi(\zeta - c_{d_1}) \to \infty$ (see after (C.146)), and Lemma E6 (choose $q = 2$ and $(\lambda', \lambda) = (\check{z}, z^*)$)), and the last holds by $S_2(z^*) = S_\tau^{\mathrm{BH}}$ due to $\psi_i = \mathrm{E}(s_i|\mathcal{G})$ and that $\widetilde{\psi}_{q,i}(z^*)$ is $\mathcal{G}$-measurable. We hence prove $\widehat{S}_\tau^{\mathrm{BH}} - S_\tau^{\mathrm{BH}} = o_{\mathrm{P}}(S^{\mathrm{OPT}} + 1)$ under $z^* \leq c_{d_2}$. ∎

Proof of Proposition B3, Part 2. In this part we prove that the sufficient and necessary condition is correct. By definition it holds that

$$\frac{S_\tau^{\mathrm{BH}}}{S^{\mathrm{OPT}}} = \frac{S_2(z^*)}{S^{\mathrm{OPT}}} = \frac{\mathrm{E}(\widetilde{\psi}_{2,i}(z^*)\psi_i)}{\mathrm{E}(\widetilde{\psi}_{2,i}(z^*)^2)^{1/2}\mathrm{E}(\psi_i^2)^{1/2}} = \mathrm{Corr}(\psi_i, \widetilde{\psi}_{2,i}(z^*)). \tag{C.151}$$

The first equality holds by $S_2(z^*) = S_\tau^{\mathrm{BH}}$ due to $\psi_i = \mathrm{E}(s_i|\mathcal{G})$ and that $\widetilde{\psi}_{q,i}(z^*)$ is $\mathcal{G}$-measurable. Since $\rho \to 0$ by assumption, we know that $z^* \to 0$ does not hold as it contradicts its definition (B.1). Moreover, we have that, using the first part of (E.1) and that $-\log\rho \precsim \log N$ by assumption, $\zeta - a^* \geq -c_N\sqrt{\log N}$ and $\zeta - \zeta^* \geq -c_N\sqrt{\log N}$ are equivalent. Hence, according to (E.23) and (E.24) of Lemma E3 (choose $q = 2$), using (C.151) and the subsequence argument, and recalling $\zeta = \sqrt{T}\mu/\sigma$ and $\zeta^* = \sqrt{-2\log\rho}$, we readily obtain the "only if" part. Then we study the "if" part, according to (E.25), we only need to show that, for all fixed $\epsilon > 0$, there exists a fixed $\tau > 0$ such that, for large N,

$$\frac{\Phi(-z^*)}{\rho\Phi(\zeta - a^*)} \leq \epsilon. \tag{C.152}$$

The rest of the proof is to show (C.152). Rewritting the definition of z^*, and noting τ is fixed, we have

$$\frac{\rho\Phi(\zeta - z^*)}{2\Phi(-z^*)} = \frac{1-\tau}{\tau} \precsim 1. \tag{C.153}$$

Using (E.15) and the second part of (E.1) (recall $\chi(a)$ and a^* are introduced in the statement of Lemma E1), noting $\chi(a) := \frac{\rho}{2}\phi(a - \zeta)/\phi(a)$, we obtain from (C.153) that

$$h(z^*) := \frac{z^*}{1 + (z^* - \zeta)_+} e^{(z^* - a^*)\zeta} \precsim 1. \tag{C.154}$$

Here we also introduce $h(z^*)$ as short-hand notation. Suppose $z^* \geq a^* + 1$. Then $h(z^*) \geq e^{(z^* - a^*)\zeta} \to \infty$. Suppose $a^* \leq z^* \leq a^* + 1$. Then $h(z^*) \geq \frac{1+a^*}{2+(a^*-\zeta)_+} \to \infty$. Suppose $z^* \leq a^* - 1$. Then $h(z^*) \leq a^* e^{-\zeta} \to 0$. Hence we conclude $a^* - 1 \leq z^* \leq a^*$ for large N. Therefore, it follows from (C.154) that

$$e^{(a^* - z^*)\zeta} \precsim \frac{a^*}{1 + (a^* - \zeta)_+}. \tag{C.155}$$

It then holds that, when $\zeta - \zeta^* \geq -c_N\sqrt{\log N}$,

$$\begin{aligned} a^* - z^* &\lesssim \frac{1}{\zeta}\log\frac{a^*}{1+(a^*-\zeta)_+} \\ &\lesssim \frac{1}{1+(a^*-\zeta)_+}\frac{1+(a^*-\zeta)_+}{a^*}\log\frac{a^*}{1+(a^*-\zeta)_+} \le o\left(\frac{1}{1+(a^*-\zeta)_+}\right). \end{aligned} \tag{C.156}$$

The first inequality follows from (C.155). The second inequality comes from $a^* \approx \zeta$ due to the first part of (E.1) and $\zeta^* \approx \sqrt{\log N}$ by the assumption on ρ. To obtain the last inequality, we utilize that $(a^* - \zeta)_+ = o(a^*)$, again due to the first part of (E.1), and that $\lim_{x\to\infty} x^{-1}\log x = 0$. We hence obtain

$$\frac{|\Phi(\zeta - z^*) - \Phi(\zeta - a^*)|}{\Phi(\zeta - a^*)} \lesssim (a^* - z^*)\max_{z^*\le a\le a^*}\phi(a-\zeta) = o\left(\frac{\phi(a^*-\zeta)}{1+(a^*-\zeta)_+}\right) = o(\Phi(\zeta - a^*)). \tag{C.157}$$

The first inequality is obvious given that $z^* \le a^*$ for large N. The first equality comes from (C.156) and that $\sup_{y:|y|\lesssim(1+|x|)^{-1}}\phi(x+y) \lesssim \phi(x)$ uniformly over x. The last equality is a result of (E.15). It follows from (C.157) that $\Phi(\zeta - a^*) = \Phi(\zeta - z^*)(1+o(1))$. Combining this result with (C.153), we further have $2\Phi(-z^*)/(\rho\Phi(\zeta - a^*)) = (1+o(1))\tau/(1-\tau)$. We hence prove (C.152) and the proof concludes. ∎

C.7.3 Proof of Proposition B4

By definition, we have $\breve{w}_{1,i}(\lambda) = \widehat{\sigma}_i^{-1}\psi_1(\widehat{s}_i,\lambda) = T^{-1/2}\widehat{\sigma}_i^{-1}\psi_1(\widehat{z}_i, T^{1/2}\lambda) = T^{-1/2}\widetilde{w}_1'(T^{1/2}\lambda)$ and $\psi_1(\check{s}_i,\lambda) = T^{-1/2}\widetilde{\psi}_1(T^{1/2}\lambda)$. Hence, it holds that

$$\widehat{S}_{1,\lambda} = \frac{\breve{w}_1(\lambda)^\intercal \mathbb{M}_\beta\alpha}{\sigma\|\mathbb{M}_\beta\breve{w}_1(\lambda)\|} = \widehat{S}_1'(T^{1/2}\lambda), \quad S_\lambda^{\text{Lasso}} = \frac{N^{1/2}\mathrm{E}(\psi_i\psi_1(\check{s}_i,\lambda))}{\sqrt{\mathrm{E}(\psi_1(\check{s}_i,\lambda)^2)}} = S_1(T^{1/2}\lambda), \tag{C.158}$$

where the first equality in the second part comes from $\psi_i = \mathrm{E}(s_i|\mathcal{G})$ and that $\psi_1(\check{s}_i,\lambda)$ is $\mathcal{G}$-measurable.

Suppose λ satisfies $\rho N\Phi(\zeta - T^{1/2}\lambda) \to \infty$. Then we have

$$\widehat{S}_{1,\lambda} = \widehat{S}_1'(T^{1/2}\lambda) = S_1(T^{1/2}\lambda) + o_{\mathrm{P}}(S^{\text{OPT}}+1) = S_\lambda^{\text{Lasso}} + o_{\mathrm{P}}(S^{\text{OPT}}+1). \tag{C.159}$$

Here the first and last equality comes from (C.158) and the second equality comes from Lemma E6 as $\rho N\Phi(\zeta - T^{1/2}\lambda) \to \infty$ (replace both λ' and λ with $T^{1/2}\lambda$ and choose $q = 1$).

Now suppose $\rho N\Phi(\zeta - T^{1/2}\lambda) \lesssim 1$. It holds that

$$S_1(T^{1/2}\lambda) = \frac{N^{1/2}\mathrm{E}(\widetilde{\psi}_{1,i}(T^{1/2}\lambda)s_i)}{\mathrm{E}(\widetilde{\psi}_{1,i}(T^{1/2}\lambda)^2)^{1/2}} \le \frac{\mu}{\sigma}\sqrt{N\mathrm{E}(\mathbb{1}_{\{|\check{s}_i|\ge 0,\alpha_i\ne 0\}})} \le \frac{\mu}{\sigma}\sqrt{2\rho N\Phi(\zeta - T^{1/2}\lambda)} = o(1). \tag{C.160}$$

Here the first equality comes from $\psi_i = \mathrm{E}(s_i|\mathcal{G})$ and that $\widetilde{\psi}_{1,i}(\lambda)$ is $\mathcal{G}$-measurable. The first inequality holds by $|s_i| = \frac{\mu}{\sigma}\mathbb{1}_{\{\alpha_i \neq 0\}}$ and Cauchy-Schwarz inequality. The second inequality is obvious. The last equality comes from $\mu = o(1)$ by assumption. Since $\rho N\Phi(\zeta - T^{1/2}\lambda) \lesssim 1$ leads to $\mu^2\rho N\Phi(\zeta - T^{1/2}\lambda) \to 0$, Lemma E7 applies and hence $\widehat{S}_1'(T^{1/2}\lambda) = o_{\mathrm{P}}(1)$. Combining with (C.160), we obtain $\widehat{S}_1'(T^{1/2}\lambda) = S_1(T^{1/2}\lambda) + o_{\mathrm{P}}(1)$. Given (C.158), we obtain $\widehat{S}_1(\lambda) = S_\lambda^{\mathrm{Lasso}} + o_{\mathrm{P}}(1)$ under $\rho N\Phi(\zeta - T^{1/2}\lambda) \lesssim 1$. Hence we establish $\widehat{S}_1(\lambda) = S_\lambda^{\mathrm{Lasso}} + o_{\mathrm{P}}(1)$ with the classic subsequence argument.

Finally, it holds that

$$\frac{S_\lambda^{\mathrm{Lasso}}}{S^{\mathrm{OPT}}} = \frac{S_1(T^{1/2}\lambda)}{S^{\mathrm{OPT}}} = \frac{\mathrm{E}(\widetilde{\psi}_{1,i}(T^{1/2}\lambda)\psi_i)}{\mathrm{E}(\widetilde{\psi}_{1,i}(T^{1/2}\lambda)^2)^{1/2}\mathrm{E}(\psi_i^2)^{1/2}} = \mathrm{Corr}(\psi_i, \widetilde{\psi}_{1,i}(T^{1/2}\lambda)), \tag{C.161}$$

where the first equality comes from (C.158) and the second comes from the definition of S_1 and $S^{\mathrm{OPT}} = N^{1/2}\mathrm{E}(\psi_i^2)^{1/2}$. On the other hand, we have that, using the first part of (E.1) and that $-\log\rho \approx \log N$ by assumption, $\zeta - a^* \to \infty$ and $\zeta - \zeta^* \to \infty$ are equivalent. Hence, according to (E.23) and (E.24) of Lemma E3 (replace λ with $T^{1/2}\lambda$ and choose $q = 2$), using (C.161), and recalling $\zeta = \sqrt{T}\mu/\sigma$ and $\zeta^* = \sqrt{-2\log\rho}$, we obtain the remaining statements of Proposition B4.

Appendix D Lemmas supporting Section C.4

Lemma D1. *We define $\bar{u}_i = T^{-1}\sum_{s\in\mathcal{T}} u_{i,s}$. Suppose Assumptions 1 and 2 hold and that $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$ and $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$. Also suppose $\log N \lesssim T \lesssim N^d$ with fixed $d < 1$. Then it holds that, as $N, T \to \infty$,*

$$\max_{1\le i\le N} |\widehat{\sigma}_i^2 - \sigma_i^2| = O_{\mathrm{P}}\left(\sqrt{(\log N)/T}\right), \tag{D.1}$$

$$\max_{1\le i\le N} |(\mathbb{P}_\beta\bar{u})_i| = O_{\mathrm{P}}\left(1/\sqrt{TN}\right), \tag{D.2}$$

$$\max_{1\le i\le N} |(\mathbb{P}_\beta\alpha)_i| = O_{\mathrm{P}}(N^{-1/2}\mathrm{E}(s_i^2)^{1/2}). \tag{D.3}$$

Proof. We start with (D.1). First of all, we write

$$\begin{aligned}\max_{1\le i\le N} |\widehat{\sigma}_i^2 - \sigma_i^2| &\le \|\mathbb{M}_\beta(T^{-1}uu^\intercal - \bar{u}\bar{u}^\intercal)\mathbb{M}_\beta - \Sigma_u\|_{\mathrm{MAX}} \\ &\le \|\mathbb{M}_\beta\Sigma_u\mathbb{M}_\beta - \Sigma_u\|_{\mathrm{MAX}} + \|\mathbb{M}_\beta(T^{-1}uu^\intercal - \bar{u}\bar{u}^\intercal - \Sigma_u)\mathbb{M}_\beta\|_{\mathrm{MAX}}. \end{aligned} \tag{D.4}$$

Now we establish the upper bounds of the two terms in the second line. We write

$$\begin{aligned}\|\mathbb{M}_\beta\Sigma_u\mathbb{M}_\beta - \Sigma_u\|_{\text{MAX}} &\le \|\mathbb{P}_\beta\Sigma_u\mathbb{P}_\beta\|_{\text{MAX}} + 2\|\mathbb{P}_\beta\Sigma_u\|_{\text{MAX}} \\ &\le (N\|\mathbb{P}_\beta\|_{\text{MAX}} + 2)\|\mathbb{P}_\beta\|_{\text{MAX}}\|\Sigma_u\|_{\text{MAX}} \lesssim_{\text{P}} N^{-1}. \qquad \text{(D.5)}\end{aligned}$$

The last inequality comes from $\|\mathbb{P}_\beta\|_{\text{MAX}} \le C\|\beta\|^2_{\text{MAX}}\|(\beta^\intercal\beta)^{-1}\|_{\text{MAX}} \lesssim_{\text{P}} N^{-1}$, which is true by assumption. On the other hand, we have

$$\begin{aligned}&\|\mathbb{M}_\beta(T^{-1}uu^\intercal - \bar{u}\bar{u}^\intercal - \Sigma_u)\mathbb{M}_\beta\|_{\text{MAX}} \\ \le\ &\|T^{-1}uu^\intercal - \bar{u}\bar{u}^\intercal - \Sigma_u\|_{\text{MAX}}(1 + 2N\|\mathbb{P}_\beta\|_{\text{MAX}} + N^2\|\mathbb{P}_\beta\|_{\text{MAX}}) \lesssim_{\text{P}} \sqrt{(\log N)/T}, \text{(D.6)}\end{aligned}$$

where the last inequality comes from the uniform bound on i.i.d. normal variables, $\log N \lesssim T$ by assumption, and that $\lambda_{\max}(\Sigma_u) \lesssim_{\text{P}} 1$ by condition (a) of Assumption 1. Substituting (D.5) and (D.6) into (D.4), and noting $N^{-1} \le C\sqrt{(\log N)/T}$ by assumption, we obtain (D.1).

We obtain (D.2) by writing

$$\max_{1\le i\le N} |(\mathbb{P}_\beta\bar{u})_k| \le C\|\beta\|_{\text{MAX}}\|(\beta^\intercal\beta)^{-1}\|_{\text{MAX}} \max_{1\le k\le K} |(\beta^\intercal\bar{u})_k| \lesssim_{\text{P}} \max_{1\le k\le K} N^{-1}|(\beta^\intercal\bar{u})_k| \lesssim_{\text{P}} 1/\sqrt{TN}.$$

Here the last inequality comes from that K is fixed, $\text{E}(\bar{u}_i\bar{u}_j|\beta,\Sigma_u) \lesssim \delta_{i,j}\sigma_i^2T^{-1}$ by conditions (a) and (b) of Assumption 1, and $\lambda_{\max}(\Sigma_u) \lesssim_{\text{P}} 1$.

Finally, we write

$$\max_{1\le i\le N} |(\mathbb{P}_\beta\alpha)_i| \le C\|\beta\|_{\text{MAX}}\|(\beta^\intercal\beta)^{-1}\|_{\text{MAX}} \max_{1\le k\le K} |(\beta^\intercal\alpha)_k| \le CN^{-1} \max_{1\le k\le K} |(\beta^\intercal\alpha)_k|. \qquad \text{(D.7)}$$

On the other hand, from condition (a) and (b) of Assumption 1 and condition (a) of Assumption 2, we have $\text{E}(\alpha_i\alpha_j|\beta,\Sigma_u) = \delta_{i,j}\sigma_i^2\text{E}(s_i^2)$. Therefore, as K is fixed and $\lambda_{\max}(\Sigma_u) \lesssim_{\text{P}} 1$, we have

$$\max_{1\le k\le K} |(\beta^\intercal\alpha)_k| \lesssim_{\text{P}} c_N N^{1/2}\|\beta\|_{\text{MAX}} \lesssim_{\text{P}} N^{1/2}\text{E}(s_i^2)^{1/2}. \qquad \text{(D.8)}$$

Substituting (D.8) into (D.7), we obtain (D.3). ∎

Lemma D2. *Suppose Assumptions 1 and 2 hold and that* $\|\beta\|^2_{\text{MAX}} \lesssim_{\text{P}} 1$ *and* $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\text{P}} N$*. Also assume* $N^d \lesssim T \lesssim N^{d'}$ *with fixed* $d > 0$ *and* $d' < 1$*. Then it holds that, as* $N, T \to \infty$*,*

$$\max_{1\le i\le N} |\widehat{s}_i - \check{s}_i| \lesssim_{\text{P}} c_N T^{-1/2}\widetilde{k}_N. \qquad \text{(D.9)}$$

If we additionally have $T \gtrsim N^d$ *with fixed* $d > 1/2$*, then it holds that, as* $N, T \to \infty$*,*

$$\max_{i\in B} |\widehat{s}_i - \check{s}_i| \gtrsim_{\text{P}} \chi_N. \qquad \text{(D.10)}$$

Here $\chi_N := \sqrt{1/N}(k_N^5 + \mathrm{E}(s_j^2)^{1/2})$, *and set* B *is* $B := \{i \in N : T^{1/2}|\check{s}_i| \leq \widetilde{k}_N\}$, *with* $\widetilde{k}_N := k_N^{-2}$.

Proof. By definition we have

$$\widehat{s}_i - \check{s}_i = -\frac{(\mathbb{P}_\beta \alpha)_i}{\widehat{\sigma}_i} - \frac{(\mathbb{P}_\beta \bar{u})_i}{\widehat{\sigma}_i} + \left(\frac{\sigma_i}{\widehat{\sigma}_i} - 1\right)\check{s}_i. \tag{D.11}$$

Since $T \gtrsim N^d$ with $d > 0$ by assumption, (D.1) of Lemma D1 leads to $\max_{1\leq i\leq N}|\widehat{\sigma}_i^2 - \sigma_i^2| = o_{\mathrm{P}}(1)$. Then, noting $\min_i \sigma_i \gtrsim_{\mathrm{P}} 1$ by condition (d) of Assumption 1, we obtain $\min_i \widehat{\sigma}_i \gtrsim_{\mathrm{P}} 1$. Applying (D.1) again, we have $\max_i \left|\frac{\sigma_i}{\widehat{\sigma}_i} - 1\right| \lesssim_{\mathrm{P}} \sqrt{(\log N)/T}$. Using these two results, and substituting (D.2) and (D.3) of Lemma D1 into (D.11), we obtain

$$\max_{1\leq i\leq N}|\widehat{s}_i - \check{s}_i| \lesssim_{\mathrm{P}} \chi_N + \sqrt{(\log N)/T}\max_{1\leq i\leq N}|\check{s}_i|, \qquad \max_{i\in B}|\widehat{s}_i - \check{s}_i| \lesssim_{\mathrm{P}} \chi_N + \sqrt{(\log N)/T}\max_{i\in B}|\check{s}_i|. \tag{D.12}$$

Since $\mathrm{P}(|s_i| \geq 1) \leq \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|\geq 1\}}) \leq c_N N^{-1}$ by condition (a) of Assumption 2, we have $\mathrm{P}(\max_i |s_i| \geq 1) \leq c_N$. Combining this result with $\max_i |\bar{\varepsilon}_i| \lesssim_{\mathrm{P}} \sqrt{(\log N)/T}$ by the uniform bound on i.i.d. normal variables, we obtain $\max_{1\leq i\leq N}|\check{s}_i| \lesssim_{\mathrm{P}} 1$ (again noting $T \gtrsim N^d$ with $d > 0$ by assumption). Then we have $\sqrt{(\log N)/T}\max_{1\leq i\leq N}|\check{s}_i| \lesssim_{\mathrm{P}} T^{-1/2}\sqrt{\log N}$. Also, we have $\chi_N \leq c_N T^{-1/2}$ since $T = o(N)$ by assumption and $\mathrm{E}(s_j^2) \lesssim 1 + \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|\geq 1\}}) \lesssim 1$ by condition (a) of Assumption 2. Substituting the two bounds into the first part of (D.12), we achieve (D.9).

On the other hand, the definition of set B leads to $\max_{i\in B}|\check{s}_i| \leq T^{-1/2}\widetilde{k}_N$. Then, noting $T \gtrsim N^d$ with $d > 1/2$ by assumption, we have $\sqrt{(\log N)/T}\max_{i\in B}|\check{s}_i| \lesssim \sqrt{1/N}k_N^5 \leq \chi_N$. Given the second part of (D.12), we obtain (D.10). ∎

Lemma D3. *Suppose Assumptions 1 and 2 hold and that* $\|\beta\|_{\mathrm{MAX}}^2 \lesssim_{\mathrm{P}} 1$ *and* $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$. $N^d \lesssim T \lesssim N^{d'}$ *with fixed* $d > 1/2$ *and* $d' < 1$. *Then it holds that, as* $N, T \to \infty$,

$$\mathrm{P}(p(\check{s}_i) \geq T^{1/2}N^{-3/2}, \forall i \leq N) \geq 1 - c_N. \tag{D.13}$$

Proof. Note that when $|x| < 1$, we can find $C > 1$ such that $a \geq C$ implies $|a - x| \geq C - 1$. Therefore, for $|x| < 1$, we have

$$\int_{|a|\geq C}\phi_{1/T}(x-a)da \leq \int_{|a|\geq C-1}\sqrt{T}\exp(-Ta^2/2)da \lesssim \exp(-T) \leq c_N N^{-1}. \tag{D.14}$$

The last step comes from $T \gtrsim N^d$ for some $d > 1/2$ by assumption. Then we can bound

$$\int_{|a|\geq C} p(a)da \;\; \leq \;\; \int_{|x|\geq 1} p_s(x)dx + \int_{|x|<1}\int_{|a|\geq C}\phi_{1/T}(x-a)da p_s(x)dx$$

$$\leq \int_{|x|\geq 1} p_s(x)dx + \sup_{x:|x|<1} \int_{|a|\geq C} \phi_{1/T}(x-a)da \leq c_N N^{-1}. \qquad \text{(D.15)}$$

Here the last inequality comes from (D.14) and $\int_{|x|\geq 1} p_s(x)dx \leq \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|\geq 1\}}) \leq \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|\geq c_N\}}) \leq c_N N^{-1}$ by condition (a) of Assumption 2. It follows from (D.15) that

$$\begin{aligned} \mathrm{P}(p(\check{s}_i) < T^{1/2}N^{-3/2}) &= \int \mathbb{1}_{\{p(a)<T^{1/2}N^{-3/2}\}} p(a)da \\ &\leq c_N N^{-1} + \int_{|a|<C} \mathbb{1}_{\{p(a)<T^{1/2}N^{-3/2}\}} p(a)da \leq c_N N^{-1}. \qquad \text{(D.16)} \end{aligned}$$

The last inequality also uses $T = o(N)$ by assumption. (D.16) proves the lemma by Bonferroni inequalities. ∎

Lemma D4. *It holds that, for $j \in \{0,1\}$ and for all $(a, \bar{a})$ satisfying $|\bar{a} - a| \leq 1/\sqrt{6 \log N}$, as $N \to \infty$,*

$$\begin{aligned} |a^j \phi(a) - \bar{a}^j \phi(\bar{a})| &\lesssim (\log N)^{(j+1)/2} |\bar{a} - a| \phi(a) + c_N N^{-2}, \qquad \text{(D.17)} \\ \phi(\bar{a}) &\lesssim (\log N)^{(j+1)/2} (1 + |a|)^{-(j+1)} (\phi(a) + c_N N^{-2}). \qquad \text{(D.18)} \end{aligned}$$

Proof. We first write that, for all a and for $j \in \{0,1\}$,

$$\begin{aligned} |a^j \phi(a) - \bar{a}^j \phi(\bar{a})| &\leq |\bar{a}^j - a^j| \phi(a) + (|\bar{a}^j - a^j| + |a|^j) |\phi(\bar{a}) - \phi(a)| \\ &\leq |\bar{a} - a| \phi(a) + (|\bar{a} - a| + |a|^j) \phi(a) |e^{-(a^2 - \bar{a}^2)/2} - 1|. \end{aligned}$$

On the other hand, for all sequence b_N satisfying $b_N \geq 1$, and for all $(a, \bar{a})$ satisfying $|a| \leq b_N$ and $|\bar{a} - a| \leq b_N^{-1}$, we have $|e^{-(a^2-\bar{a}^2)/2} - 1| \lesssim |\bar{a} - a| b_N$. As a result, for all such b_N and $(a, \bar{a})$, it holds that, for $j \in \{0,1\}$,

$$|a^j \phi(a) - \bar{a}^j \phi(\bar{a})| \lesssim b_N^{j+1} |\bar{a} - a| \phi(a). \qquad \text{(D.19)}$$

Moreover, it holds that, for $j \in \{0,1,2\}$ and for all b'_N that satisfies $b'_N \geq \sqrt{5 \log N}$,

$$\sup_{a:|a|\geq b'_N} |a^j \phi(a)| \leq c_N N^{-2}. \qquad \text{(D.20)}$$

Then, choosing $b_N = \sqrt{6 \log N}$, which ensures $b_N \geq \sqrt{5 \log N}$ and $b_N - b_N^{-1} \geq \sqrt{5 \log N}$ for large N, we obtain (D.17) by combining (D.19) and (D.20). Further, (D.18) would directly come from (D.17) by choosing $j = 0$ when $|a| \leq \sqrt{6 \log N}$, which guarantees $(\log N)(1 + |a|)^{-2} \gtrsim 1$. For $|a| \geq \sqrt{6 \log N}$, we obtain (D.18) from (D.20) by choosing $j = 2$, because in this case we have $|\bar{a}| \geq \sqrt{5\log N}$ for large N, and $1+|a| \sim 1+|\bar{a}|$. The proof ends. ∎

Appendix E Lemmas supporting Section C.7

Lemma E1. *Suppose that r_t follows (1), Assumption 1 holds, $u_{i,t} \approx \mathcal{N}(0,\sigma^2)$, and α following (5) as in Example 1. Also assume $N^{-d'} \leq \rho \leq N^{-d}$ with fixed $d' > d > 0$. Recall we define $p(a)$ and $\psi(a)$ in Corollary 1 and $\Phi(x)$ is the standard normal cumulative distribution function. Then it holds that (i) when $\zeta - \zeta^* \lesssim -\sqrt{\log N}$,*

$$\int \psi(a)^2 p(a)da \approx (1 \wedge \zeta^2)\mu^2\rho\iota\Phi(a^* - 2\zeta);$$

(ii) when $\zeta - \zeta^* \geq -c_N\sqrt{\log N}$,

$$\int \psi(a)^2 p(a)da \approx \mu^2\rho\Phi(\zeta - a^*).$$

Here $\iota := e^{3\zeta^2/2-\zeta a^*}$, ζ^* and $\chi(a)$ are defined by $\zeta^* := \sqrt{-2\log(\rho/2)}$ and $\chi(a) := \frac{\rho}{2}\phi(a-\zeta)/\phi(a)$, and a^* is the solution of $\chi(a) = 1$.

Proof. It follows from definition that

$$a^* = \frac{\zeta}{2} + \frac{\zeta^{*2}}{2\zeta} \quad \text{and} \quad \chi(a) = \exp((a-a^*)\zeta). \tag{E.1}$$

By definition we have

$$T^{-1/2}p(T^{-1/2}a) = (1-\rho)\phi(a) + \frac{\rho}{2}\phi(|a|-\zeta) + \frac{\rho}{2}\phi(|a|+\zeta).$$

Because $\phi(|a|+\zeta) \leq \phi(a)$ and $\rho \to 0$ we have, uniformly over $a \geq 0$,

$$T^{-1/2}p(T^{-1/2}a) = (1+o(1))\frac{\rho}{2}\phi(a-\zeta)(1+\chi(a)^{-1}). \tag{E.2}$$

We also note $\phi(a+\zeta) = \exp(-2a\zeta)\phi(a-\zeta)$. Therefore, uniformly over $a \geq 0$,

$$\psi(T^{-1/2}a) = (1+o(1))\mu\frac{1-\exp(-2a\zeta)}{1+\chi(a)^{-1}}. \tag{E.3}$$

As a result of (E.2) and (E.3), it holds that

$$\int \psi(a)^2 p(a)da = (1+o(1))\mu^2\rho B^*, \quad \text{with} \quad B^* := \int_0^\infty (1-\exp(-2a\zeta))^2\frac{\chi(a)}{1+\chi(a)}\phi(a-\zeta)da. \tag{E.4}$$

Then the goal becomes characterization of B^* across different cases. For this purpose, we introduce two set sequences $A_1 = (0, a^*)$ and $A_2 = (a^*, \infty)$. Based on these sets, we define, for $j \in \{1, 2\}$,

$$B_j = \int_{A_j} (1-\exp(-2a\zeta))^2 \frac{\chi(a)}{1+\chi(a)} \phi(a-\zeta)da, \quad B'_j := \int_{A_j} (1-\exp(-2a\zeta))^2 \phi(a-2\zeta)da. \tag{E.5}$$

It apparently holds that $B^* = B_1 + B_2$. Furthermore, from the first part of (E.1), we note $\zeta a^* \to \infty$ and thereby $\sup_{a \geq a^*} \exp(-2a\zeta) \to 0$. From the second part of (E.1), $\chi(a) \leq 1$ for all $a \leq a^*$ and $\chi(a) \geq 1$ for all $a \geq a^*$. Hence, it follows from (E.5) that

$$B_1 \approx \int_0^a (1-\exp(-2a\zeta))^2 \chi(a)\phi(a-\zeta)da, \quad B_2 \approx \int_{a^*}^\infty \phi(a-\zeta)da. \tag{E.6}$$

We note $\chi(a)\phi(a-\zeta) = \iota\phi(a-2\zeta)$. Applying this relation to the two parts of (E.6), we obtain, respectively,

$$B_1 \approx \iota B'_1, \quad B_2 \lesssim \iota B'_2. \tag{E.7}$$

where, for the second result, we use that $\inf_{a \geq a^*}(1-\exp(-2a\zeta))^2\chi(a) \gtrsim 1$ due to $\inf_{a \geq a^*} \chi(a) \geq 1$ and $\inf_{a \geq a^*} a\zeta \to \infty$ by the first part of (E.1).

Now we gauge the magnitude of B_1 and B_2 for cases (i) and (ii). We further separate case (i) into two subcases: case (ia) where $\zeta \lesssim 1$ (under which $\zeta - \zeta^* \lesssim -\sqrt{\log N}$ always holds, as $\zeta^* \gtrsim \sqrt{\log N}$ by the assumption $\rho \leq N^{-d}$ with fixed $d > 0$), and case (ib) where $\zeta \to \infty$ and $\zeta - \zeta^* \lesssim -\sqrt{\log N}$. We start with case (ia) where $\zeta \lesssim 1$. We note

$$\phi(a-2\zeta)\exp(-2a\zeta) = \phi(a)\exp(-2\zeta^2), \quad \phi(a-2\zeta)\exp(-4a\zeta) = \phi(a+2\zeta). \tag{E.8}$$

Using (E.8), we obtain

$$B'_1 + B'_2 = \int (1-\exp(-2a\zeta))^2 \phi(a-2\zeta)da = 1-\exp(-2\zeta^2). \tag{E.9}$$

Moreover, it apparently holds

$$B'_2 \leq \int_{A_2} \phi(a-2\zeta)da = \Phi(2\zeta - a^*). \tag{E.10}$$

On the other hand, when $\zeta \lesssim 1$, we have $a^* - 2\zeta \geq \zeta^{*2}/(3\zeta)$ for N sufficiently large. Then, when $\zeta \lesssim 1$

$$\zeta^{-2}\Phi(2\zeta - a^*) \lesssim \frac{1}{\zeta^{*4}} \frac{\zeta^{*2}}{\zeta} \exp(-\zeta^{*4}/(18\zeta^2)) \to 0. \tag{E.11}$$

where the last convergence comes from $\zeta^* \to \infty$ (by the assumption $\rho \le N^{-d}$ with fixed $d > 0$), $\zeta^{*2}/\zeta \to \infty$ (as $\zeta \lesssim 1$), and $\lim_{x\to\infty} x\exp(-x^2) = 0$. Since $1 - \exp(-2\zeta^2) \approx \zeta^2$ under $\zeta \lesssim 1$, combining (E.9), (E.10), and (E.11), we obtain that, when $\zeta \lesssim 1$,

$$B_1' \approx \zeta^2, \quad B_2' = o(\zeta^2). \tag{E.12}$$

Substituting (E.12) into (E.7), we obtain (recall $B^* = B_1 + B_2$) that, when $\zeta \lesssim 1$,

$$B^* \approx \zeta^2 e^{-\zeta a^*}. \tag{E.13}$$

(E.4) and (E.13) together lead to $\int \psi(a)^2 p(a)da \approx \mu^2\zeta^2\rho e^{-\zeta a^*}$. Given $\iota \approx e^{-\zeta a^*}$ and $\Phi(a^* - 2\zeta) \approx 1$ (by the first part of (E.1)) when $\zeta \lesssim 1$, we prove the lemma for case (ia).

Next, we study cases (ib) and (ii). We note in both cases we have $\zeta \to \infty$. Using (E.8) and $\Phi(x) = 1 - \Phi(-x)$, we obtain by direct calculation

$$B_1' = \int_0^{a^*} \phi(a-2\zeta)da = \Phi(a^*-2\zeta) - 2\Phi(-2\zeta) + \Phi(-a^*-2\zeta) - \exp(-2\zeta^2)(2\Phi(a^*)-1). \tag{E.14}$$

We note that, uniformly over x,

$$\Phi(-x) \approx \frac{1}{1+x_+}\phi(x_+), \tag{E.15}$$

where $x_+ = \max\{x, 0\}$. Using (E.15), we obtain $\Phi(a^* - 2\zeta) \ge \Phi(-3\zeta/2) \approx \frac{1}{1+\zeta}\phi(3\zeta/2) \approx \frac{1}{1+\zeta}\exp(-9\zeta^2/8)$. Then we have, $\Phi(-2\zeta) = o(\Phi(a^* - 2\zeta))$ and $\exp(-2\zeta^2) = o(\Phi(a^* - 2\zeta))$ when $\zeta \to \infty$. Since $\Phi(-a^* - 2\zeta) \le \Phi(-2\zeta)$ and $0 \le \Phi(a^*) \le 1$, we obtain from (E.14) that, when $\zeta \to \infty$,

$$B_1' = (1 + o(1))\Phi(a^* - 2\zeta). \tag{E.16}$$

Substituting (E.16) into the first part of (E.7), and noting the second part of (E.6), we obtain that, when $\zeta \to \infty$,

$$B_1 \approx \iota\Phi(a^* - 2\zeta), \quad B_2 \approx \Phi(\zeta - a^*). \tag{E.17}$$

We first investigate case (ib), in which it holds that $a^* - \zeta \gtrsim \sqrt{\log N}$ according to the first part of (E.1) and $\zeta^* \approx \sqrt{\log N}$ by the assumption on ρ. Then from (E.17) and (E.15), it follows that, in case (ib),

$$B_1 \gtrsim \frac{1}{1 + (2\zeta - a^*)_+}\phi(a^* - \zeta), \quad B_2 \approx \frac{1}{1 + (a^* - \zeta)_+}\phi(a^* - \zeta). \tag{E.18}$$

Here the first part comes from $\phi(x_+) \ge \phi(x)$ for all x, and $\phi(2\zeta - a^*)e^{3\zeta^2/2 - \zeta a^*} = \phi(a^* - \zeta)$.

The second part comes from $a^* - \zeta \gtrsim \sqrt{\log N}$. Since $1 + (2\zeta - a^*)_+ \lesssim (a^* - \zeta)_+$ according to the first part of (E.1) and $\zeta^* \approx \sqrt{\log N}$ by the assumption on ρ, we obtain from (E.18) that, in case (ib),

$$B_2 \lesssim B_1. \tag{E.19}$$

Now we study case (ii), in which it holds that $a^* - \zeta \leq c_N \sqrt{\log N}$ by the first part of (E.1) and $\zeta^* \approx \sqrt{\log N}$. Hence, we have $2\zeta - a^* \gtrsim 1$. Therefore, from (E.17) and (E.15), it follows that, in case (ii),

$$B_1 \approx \frac{1}{1 + (2\zeta - a^*)_+} \phi(a^* - \zeta), \quad B_2 \gtrsim \frac{1}{1 + (a^* - \zeta)_+} \phi(a^* - \zeta). \tag{E.20}$$

Here the first part comes from $2\zeta - a^* \gtrsim 1$ and $\phi(2\zeta - a^*) e^{3\zeta^2/2 - \zeta a^*} = \phi(a^* - \zeta)$. The second part comes from $\phi(x_+) \geq \phi(x)$ for all x. Since $1 + (a^* - \zeta)_+ = o((2\zeta - a^*)_+)$, we obtain from (E.20) that, in case (ii),

$$B_1 = o(B_2). \tag{E.21}$$

Combining (E.19) and (E.21) with (E.17), we prove the lemma for cases (ib) and (ii). The proof concludes. ∎

Lemma E2. *Suppose a real-valued function sequence $h_N : R \to R$, a sequnce a_N, and a positive sequence Δ_N satisfy $\Delta_N^{-1/2}(\mathbb{1}_{A_N} - h_N) \to 0$ in L^2 with $A_N := (a_N, a_N + \Delta_N)$. Then there exists a set sequence $\bar{A}_N$ such that*

$$\int_{A_N - \bar{A}_N} da = o(\Delta_N) \quad \textit{and} \quad \sup_{a \in A'_N} |h_N(a) - 1| = o(1).$$

Proof. We define $h_N^*(a) = h_N(a\Delta_N^{-1})$ and $A_N^* = (a_N \Delta_N^{-1}, a_N \Delta_N^{-1} + 1)$. Then clearly $(\mathbb{1}_{A_N^*} - h_N^*) \to 0$ in L^2. This further leads to $h_N^* \to \mathbb{1}_{A_N^*}$ in measure (see, e.g., Theorem 2.15.a of Folland (2009)), i.e., for every $\epsilon > 0$, $\int_{A_N^*} \mathbb{1}_{\{|1 - h_N^*(a)| \geq \epsilon\}} da \to 0$. This is equivalent to that, for every $\epsilon > 0$,

$$\Delta_N^{-1} \int_{A_N} \mathbb{1}_{\{|1 - h_N(a)| \geq \epsilon\}} da \to 0. \tag{E.22}$$

We hence prove the lemma. ∎

Lemma E3. *Suppose the same assumptions as in Lemma E1 and let a^* be as defined therein. Then there exists a deterministic sequence $\lambda \geq 0$ such that* $\operatorname{Corr}(\psi_i, \widetilde{\psi}_{q,i}(\lambda)) \to 1$ *if and only if, for some $c_N \to 0$,*

$$\begin{cases} (\mu, \rho) \in \left\{(\mu, \rho) : \zeta \leq c_N \, or \, \zeta - a^* \geq c_N^{-1}\right\} & \textit{for } q = 1, \\ (\mu, \rho) \in \left\{(\mu, \rho) : \zeta \leq c_N \, or \, \zeta - a^* \geq -c_N \sqrt{\log N}\right\} & \textit{for } q = 2. \end{cases} \tag{E.23}$$

Moreover, if sequence $\lambda \geq 0$ *is determnistic,* $\mathrm{Corr}(\psi_i, \widetilde{\psi}_{q,i}(\lambda)) \to 1$ *if and only if*

$$\begin{cases} \lambda \to 0, & \textit{when } \zeta \to 0 \textit{and } q \in \{1,2\}, \\ \zeta - \lambda \to \infty \textit{and } \frac{\phi(\lambda)}{\rho(1+\lambda^2)(\zeta-\lambda)^2} \to 0, & \textit{when } \zeta - a^* \to \infty \textit{and } q = 1. \end{cases} \tag{E.24}$$

Further, when $\zeta - a^* \geq -c_N\sqrt{\log N}$*, it holds that*

$$\mathrm{Corr}(\psi_i, \widetilde{\psi}_{q,i}(\lambda)) \geq 1 + o(1) + O\left(\frac{\Phi(-\lambda)}{\rho\Phi(\zeta - a^*)}\right). \tag{E.25}$$

Proof. Step 1. As a result of (E.2) and (E.3), it holds that, by the symmetry of $\psi(a)$, $\widetilde{\psi}_q(a, \lambda)$, and $p(a)$,

$$\int \psi(a)^2 p(a)da = (1 + o(1))\mu^2\rho \int_0^\infty (1 - \exp(-2a\zeta))^2 \frac{\chi(a)}{1 + \chi(a)}\phi(a - \zeta)da, \tag{E.26}$$

$$\int \widetilde{\psi}_q(a, \lambda)^2 p(a)da = (1 + o(1))\rho \int_0^\infty \widetilde{\psi}_q(a, \lambda)^2 \frac{1 + \chi(a)}{\chi(a)}\phi(a - \zeta)da, \tag{E.27}$$

$$\int \psi(a)\widetilde{\psi}_q(a, \lambda)p(a)da = (1 + o(1))\mu\rho \int_0^\infty \widetilde{\psi}_q(a, \lambda)(1 - \exp(-2a\zeta))\phi(a - \zeta)da. \tag{E.28}$$

To facilitate the exposition, we introduce short-hand notation

$$f(a) := \sqrt{\frac{\chi(a)}{1 + \chi(a)}\phi(a - \zeta)\mathbb{1}_{\{a \geq 0\}}}, \quad g(a) = (1 - e^{-2a\zeta})f(a), \quad \widetilde{g}_q(a) = \widetilde{\psi}_q(a, \lambda)\frac{1 + \chi(a)}{\chi(a)}f(a).$$

Comparing the definitions of f and g with the right-hand sides of (E.26), (E.27), and (E.28), we obtain, for $q \in \{1, 2\}$,

$$\mathrm{Corr}(\psi_i, \widetilde{\psi}_{q,i}(\lambda)) = (1 + o(1))\theta_q(\lambda), \quad \theta_q(\lambda) := \frac{\langle g, \widetilde{g}_q \rangle}{\|g\|\|\widetilde{g}_q\|}, \tag{E.29}$$

where $\langle g, \widetilde{g}_q \rangle := \int g(a)\widetilde{g}_q(a)da$ and $\|g\| := \sqrt{\langle g, g \rangle}$ stand for the L^2-inner product and L^2-norm. Moreover, we note that $1 - \theta_q(\lambda) = \frac{1}{2}\frac{1}{\|g\|^2}\left| g - \frac{\|g\|}{\|\widetilde{g}_q\|} \times \widetilde{g}_q \right|^2$. Let d^* be the scalar that minimizes $\|g - d \times \widetilde{g}_q\|$. It obviously holds that, for $q \in \{1, 2\}$,

$$1 - \theta_q(\lambda) \geq \frac{1}{2\|g\|^2}\|g - d^* \times \widetilde{g}_q\|^2. \tag{E.30}$$

Moreover, we have that, if $\theta_q(\lambda) \geq 0$ and for $q \in \{1, 2\}$,

$$1 - \theta_q(\lambda) \leq 1 - \theta_q(\lambda)^2 = \frac{1}{\|g\|^2}\|g - d^* \times \widetilde{g}_q\|^2. \tag{E.31}$$

Given (E.29), (E.30), and (E.31), the proof will focus on evaluating $\|g - d^* \times \widetilde{g}_q\|$. In addition, given the definition of $\widetilde{\psi}(a, \lambda)$, we can write, for $q \in \{1, 2\}$,

$$\|g - d^* \times \widetilde{g}_q\| = \int (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 f(a)^2 da, \quad \text{with} \quad h_q(a) = (e^{(a^*-a)\zeta} + 1)\widetilde{\psi}_q(a, \lambda). \tag{E.32}$$

Further, to characterize the magnitude of $\|g\|$, we consider, resepectively, case (ia) where $\zeta \to 0$ (under which $\zeta - \zeta^* \lesssim -\sqrt{\log N}$ always holds, as $\zeta^* \gtrsim \sqrt{\log N}$ by the assumption $\rho \leq N^{-d}$ with fixed $d > 0$), case (ib) where $\zeta \gtrsim 1$ and $\zeta - \zeta^* \lesssim -\sqrt{\log N}$, and case (ii) where $\zeta - \zeta^* \geq -c_N\sqrt{\log N}$. Here and below we adopt the notation introduced in the statement of Lemma E1. Since B^* is simply $\|g\|^2$ here, it follows from Lemma E1 that

$$\|g\|^2 \asymp \begin{cases} (1 \wedge \zeta^2)\iota\Phi(a^* - 2\zeta), & \text{cases (ia) and (ib);} \\ \Phi(\zeta - a^*), & \text{case (ii).} \end{cases} \tag{E.33}$$

By the classic subsequence argument, we can establish the lemma as long as we prove it under each of cases (ia), (ib), and (ii).

Step 2. We start with demonstrating that, for each $q \in \{0, 1\}$, $\theta_q(\lambda) \to 1$ holds in case (ia) if and only if we choose $\lambda \to 0$. Suppose $\lambda \to 0$ and let $\bar{d}^* = 2\zeta e^{-a^*\zeta}$. We obtain that, for each $q \in \{1, 2\}$ and for all $a \geq 0$,

$$\begin{aligned} |1 - e^{-2a\zeta} - \bar{d}^* \times h_q(a)| &= |1 - e^{-2a\zeta} - 2\zeta\psi_q(a, \lambda)(e^{-a\zeta} + e^{-a^*\zeta})| \\ &\leq |1 - e^{-2a\zeta} - 2\zeta a| + 2\zeta\lambda + 2\zeta a(|1 - e^{-a\zeta}| + e^{-a^*\zeta}) \\ &\leq 4\zeta^2 a^2 + 2\zeta\lambda + 2e^{-a^*\zeta}\zeta a. \end{aligned} \tag{E.34}$$

The second inequality comes from that $|\psi_q(a, \lambda)| \leq a$ and $|\psi_q(a, \lambda) - a| \leq \lambda$ for all $a \geq 0$. The last inequality comes from that $-\frac{1}{2}x^2 \leq 1 - e^{-x} - x \leq 0$ and $0 \leq 1 - e^{-x} \leq x$ for all $x \geq 0$. As a result, we have, for each $q \in \{1, 2\}$,

$$\begin{aligned} \|g - d^* \times \widetilde{g}_q\|^2 &\leq \|g - \bar{d}^* \times \widetilde{g}_q\|^2 \lesssim \iota \int_0^\infty (\zeta^4 a^4 + \zeta^2\lambda^2 + e^{-2a^*\zeta}\zeta^2 a^2)\phi(a - 2\zeta) da \\ &\lesssim \iota(\zeta^4 + \zeta^2\lambda^2 + e^{-2a^*\zeta}\zeta^2) = o(\iota\zeta^2). \end{aligned} \tag{E.35}$$

Here the first inequality holds by the definition of d^*. The second inequality comes from (E.32), (E.34), and $f(a)^2 \leq \chi(a)\phi(a - \zeta) = \iota\phi(a - 2\zeta)$. The third inequality comes from that $\int a^j \phi(a - 2\zeta) da = \int (a + 2\zeta)^j \phi(a) da \lesssim \int (a^j + \zeta^j)\phi(a) da \lesssim 1$ for $j \in \{2, 4\}$ as $\zeta \to 0$. The last inequality comes from $\zeta \to 0$, $a^*\zeta \to \infty$, and $\lambda \to 0$.

Now suppose $\lambda \gtrsim 1$. Then it holds that, for each $q \in \{1, 2\}$,

$$\|g - d^* \times \widetilde{g}_q\|^2 \geq \int_0^{\lambda\wedge 1} (1 - e^{-2a\zeta})^2 f(a)^2 da \gtrsim \iota\zeta^2 \int_0^{\lambda \wedge 1} a^2 \phi(a - 2\zeta) \gtrsim \iota\zeta^2. \tag{E.36}$$

The first inequality comes from $\widetilde{g}_q(a) = 0$ for all $a \in (0, \lambda)$. The second inequality comes from $\inf_{a\leq 1} f(a)^2 \geq \frac{1}{2}\chi(a)\phi(a - \zeta)$. According to (E.33), we have $\|g\|^2 \approx \iota\zeta^2$ as $\Phi(a^* - 2\zeta) \approx 1$ (by the first part of (E.1)) when $\zeta \to 0$. Given (E.29), we prove that for each $q \in \{1, 2\}$, $\mathrm{Corr}(\psi_i, \widetilde{\psi}_{q,i}(\lambda)) \to 1$ if (by (E.31) and (E.35)) and only if (by (E.30) and (E.36)) $\lambda \to 0$ when we are in case (ia), i.e., when we have $\zeta \to 0$. We hence establish the first part of (E.24).

Step 3. Next, we show that, for each $q \in \{1, 2\}$, $\theta_q(\lambda) \to 1$ does not hold for any λ sequence in case (ib). Note that we always have $\zeta \gtrsim 1$ and $\zeta - a^* \lesssim 1$ in case (ib), due to $\zeta - a^* \leq \zeta - \zeta^*$ by the first part of (E.1). We introduce a set sequence $A := ((2\zeta) \wedge a^* - 4\Delta, (2\zeta) \wedge a^*)$, with $\Delta := \frac{1}{1+(2\zeta - a^*)_+} \wedge (\zeta/2)$. Because $\phi(x)' = -x\phi(x)$, it holds that, for any real sequence x and any fixed positive constant C,

$$\phi(x) \lesssim \inf_{a:|a-x|\leq \frac{C}{1+|x|}} \phi(a) \leq \sup_{a:|a-x|\leq \frac{C}{1+|x|}} \phi(a) \lesssim \phi(x). \tag{E.37}$$

As a result, we have $\inf_{a\in A} f(a)^2 \gtrsim \iota\phi((2\zeta) \wedge a^* - 2\zeta)$ (by (E.37), the relation right after (E.6), the definition of f, and $A \subset (0, \infty)$), we have from (E.32) that, for $q \in \{1, 2\}$,

$$\|g - d^* \times \widetilde{g}_q\|^2 \gtrsim \iota\phi((2\zeta) \wedge a^* - 2\zeta) \int_A (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 da. \tag{E.38}$$

According to (E.33) and (E.15), $\|g\|^2 \approx \iota\phi((2\zeta) \wedge a^* - 2\zeta)\Delta$ in case (ib). Then, combining (E.30) and (E.38), we obtain that, if $\theta(\lambda) \to 1$, then, for $q \in \{1, 2\}$,

$$\int_A (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 da = o(\Delta). \tag{E.39}$$

It follows from Lemma E2 and (E.39) that, for each $q \in \{1, 2\}$, if $\theta_q(\lambda) \to 1$, there exists a set sequence $\bar{A}$ such that,

$$\int_{A-\bar{A}} da = o(\Delta) \quad \text{and} \quad \sup_{a\in\bar{A}} |d^* \times h_q(a) - (1 - e^{-2a\zeta})| = o(1). \tag{E.40}$$

It is almost obvious that (E.40) can not hold for any $q \in \{1, 2\}$ and for any λ, given the drastic difference between the definition of $h_q(a)$ and $1 - e^{-2a\zeta}$. To see this more clearly, we first consider the situation where $\zeta \approx 1$. We note that, under $\zeta \approx 1$ and uniformly over $a \in A$ (which becomes $(2\zeta - 4\Delta, 2\zeta)$ for large N), $h_q(a) = (1 + o(1))e^{a^*\zeta}\widetilde{\psi}_q(a, \lambda)$. Apply the triangular inequality to (E.40) leads to $\sup_{a\in\bar{A}} |d^* e^{a^*\zeta} \times \widetilde{\psi}_q(a, \lambda) - (1 - e^{-2a\zeta})| = o(1)$, which can not hold for any $q \in \{1, 2\}$ and for any λ sequence, because of the nonlinearity of

$1 - e^{-2a\zeta}$ in a as $\zeta \precsim 1$.

Then we consider the situation where $\zeta \to \infty$. It is trivial from (E.39) and that $\widetilde{\psi}_q(a,\lambda) = 0$ for $a \le \lambda$ that $\lambda \le (2\zeta) \wedge a^* - 7\Delta/2$ for large N if $\theta_q(\lambda) \to 1$, for each $q \in \{1,2\}$. As a direct result, we have, by the definition of $h_q(a)$, that, for each $q \in \{1,2\}$, if $\theta_q(\lambda) \to 1$ and $\zeta \to \infty$,

$$\frac{\inf_{a\in A^+} h_q(a)}{\sup_{a\in A^-} h_q(a)} \gtrsim \exp(\zeta\Delta) \to \infty, \tag{E.41}$$

where $A^+ := ((2\zeta) \wedge a^* - 3\Delta, (2\zeta) \wedge a^* - 2\Delta)$ and $A^- := ((2\zeta) \wedge a^* - \Delta, (2\zeta) \wedge a^*)$. (E.41) and (E.40) clearly contradict as, for large N, any $\bar{A}$ satisfying the first part of (E.40) would overlap with both A^+ and A^-, and hence would violate the second part of (E.40) according to (E.41). In other words, $\theta_q(\lambda) \to 1$ does not hold under any λ sequence and for any $q \in \{1,2\}$ if we are in case (ib).

Step 4. Now we demonstrate that, for each $q \in \{1,2\}$, $\theta_q(\lambda) \to 1$ does not hold for any λ sequence if we are in case (ii) and $\Delta^{-1}(a^* - \lambda) \to \infty$. We introduce a set sequences $A^* := (a^* \vee \zeta + \Delta, a^* \vee \zeta + 2\Delta)$ with $\Delta := \frac{1}{1+(a^*-\zeta)_+}$. We note that $\inf_{a\in A} \phi(a-\zeta) \gtrsim \phi((a^*-\zeta)_+)$ by (E.37). Then, from (E.32) it follows that, for $q \in \{1,2\}$,

$$\|g - d^* \times \widetilde{g}_q\|^2 \gtrsim \phi((a^* - \zeta)_+) \int_{A^*} (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 da. \tag{E.42}$$

According to (E.33) and (E.15), $\|g\|^2 \precsim \phi((a^* - \zeta)_+)\Delta$. Then, combining (E.30) and (E.42), we obtain that, for $q \in \{1,2\}$, if $\theta_q(\lambda) \to 1$,

$$\int_{A^*} (1 - e^{-2a\zeta} - d^* \times h_1(a))^2 da = o(\Delta). \tag{E.43}$$

Moreover, it follows from Lemma E2 and (E.43) that, for $q \in \{1,2\}$, if $\theta_q(\lambda) \to 1$, there exists a set sequence $\bar{A}$ such that

$$\int_{A^*-\bar{A}} da = o(\Delta) \quad \text{and} \quad \sup_{a\in\bar{A}} |d^* \times h_q(a) - 1| = o(1). \tag{E.44}$$

Here we also use $e^{-a^*\zeta} \to 0$ (by the first part of (E.1) and $\zeta^* \precsim \sqrt{\log N} \to \infty$ due to the assumption on ρ). Now we prove by contradiction that, for each $q \in \{1,2\}$, $\theta_q(\lambda) \to 1$ can not hold if $\Delta^{-1}(a^* - \lambda) \to \infty$. For this purpose, we introduce a set sequence $A' := (a^* - 2\Delta, a^* - \Delta)$. Suppose $\Delta^{-1}(a^* - \lambda) \to \infty$ holds. Then we obtain that, for $q \in \{1,2\}$, if $\theta_q(\lambda) \to 1$ and for all $a \in A^*$ and $a' \in A'$,

$$h_q(a') = (1 + o(1))\frac{\exp((a^* - a')\zeta) + 1}{\exp((a^* - a)\zeta) + 1} \times h_q(a) = (1 + o(1))\chi(a')^{-1} \times h_q(a). \tag{E.45}$$

The first equality comes from $\Delta^{-1}(a^* - \lambda) \to \infty$ and the definition of $\widetilde{\psi}_q$. The last equality comes from the definitions of A^* and A' and that $\exp(\zeta\Delta) \to \infty$, which in turn is a result of that in case (ii) we have (a) $\zeta \geq \zeta^* - c_N\sqrt{\log N} \gtrsim \sqrt{\log N}$ and (b) $\Delta^{-1} \lesssim c_N\sqrt{\log N}$ due to $a^* - \zeta \approx \zeta^* - \zeta \geq -c_N\sqrt{\log N}$. Combining the second part of (E.44) and (E.45), we obtain that, for $q \in \{1, 2\}$, if $\theta_q(\lambda) \to 1$ and for all $a \in A'$,

$$d^* \times h_q(a) = (1 + o(1))\chi(a)^{-1}. \tag{E.46}$$

Since $\inf_{a\in A'} \chi(a)^{-1} \to \infty$ due to $\exp(\zeta\Delta) \to \infty$, it follows from (E.46) that, for $q \in \{1, 2\}$,

$$\begin{aligned} \|g - d^* \times \widetilde{g}_q\|^2 &\geq \int_{A'} (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 f(a)^2 da \\ &\gtrsim \int_{A'} \chi(a)^{-1}\phi(a - \zeta)da \gtrsim \Delta \inf_{a\in A'} \phi(a - \zeta) \gtrsim \|g\|^2, \end{aligned}$$

where we note that $\inf_{a\in A'} \chi(a)^{-1} \to \infty$ (by $\chi(a^*) = 1$ and $\exp(\zeta\Delta) \to \infty$) and $\inf_{a\in A'} \phi(a - \zeta) \gtrsim \phi((a^* - \zeta)_+)$, and we recall $\|g\|^2 \approx \phi((a^* - \zeta)_+)\Delta$. This contradicts (E.30) and proves that, for $q \in \{1, 2\}$, $\theta_q(\lambda) \to 1$ can not hold if we are in case (ii) and $\Delta^{-1}(a^* - \lambda) \to \infty$.

Step 5. This step completes the proof regarding condition (E.23) for $q = 1$, and prove the second part of (E.24). Given (E.29) and the privious analysis for cases (ia) and (ib), to show condition (E.23) for $q = 1$ we only need to demonstrate that, in case (ii), there exists some sequence λ such that $\theta_1(\lambda) \to 1$ if and only if $\zeta - a^* \to \infty$.

We first show the "only if". By definition of A^* (see above (E.42)), it holds that $\sup_{a\in A^*} \exp((a^* - a)\zeta) \leq \exp(-\zeta\Delta) \to 0$. Substituting this result into the definition of h_1 given in (E.32) and that $\psi_1(a, \lambda) = (a - \lambda)_+$ for $a \geq 0$, we obtain from the second part of (E.44) that $\sup_{a\in\bar{A}}(a - \lambda)_+ = (1 + o(1)) \inf_{a\in\bar{A}}(a - \lambda)_+$. This results, given the first part of (E.44) and the definition of A^*, translates into $(a^* \vee \zeta + 2\Delta - \lambda)_+ = (1 + o(1))(a^* \vee \zeta + \Delta - \lambda)_+$. Also, it is trivial from (E.43) that $\lambda \leq a^* \vee \zeta + 3\Delta/2$ for large N. Therefore, we have that, if $\theta_1(\lambda) \to 1$,

$$\Delta^{-1}(a^* \vee \zeta - \lambda) \to \infty. \tag{E.47}$$

However, according to Step 4 and the subsequence argument, $\theta_1(\lambda) \to 1$ also leads to $\Delta^{-1}(a^* - \lambda) \lesssim 1$. This bound and (E.47) can simultaneously hold only if $\zeta - a^* \to \infty$.

Now we show the "if" part. We note

$$\|f - g\|^2 \leq \int_0^\infty e^{-2a\zeta}\phi(a - \zeta)da = \Phi(-\zeta) \to 0, \tag{E.48}$$

where the first inequality comes from $f(a)^2 \leq \phi(a - \zeta)$. Moreover, letting $\bar{d}^* = (\zeta - \lambda)^{-1}$, we obtain

$$\begin{aligned}
&\|f - \bar{d}^* \times \widetilde{g}_1\|^2 \\
= \;& \int_0^\lambda f(a)^2 da + \int_\lambda^\infty \left(1 - (e^{(a^*-a)\zeta} + 1) \times \frac{a-\lambda}{\zeta-\lambda}\right)^2 f(a)^2 da \\
\leq \;& \int_0^\lambda f(a)^2 da + 2\int_\lambda^\infty \left(\frac{\zeta-a}{\zeta-\lambda}\right)^2 f(a)^2 da + 2\int_\lambda^\infty e^{2(a^*-a)\zeta} \left(\frac{a-\lambda}{\zeta-\lambda}\right)^2 f(a)^2 da \\
\leq \;& \Phi(\lambda-\zeta) + \frac{2}{(\zeta-a^*)^2} + \frac{4}{(\zeta-\lambda)^2\rho}\int_\lambda^\infty \phi(a)(a-\lambda)^2 da. \qquad \text{(E.49)}
\end{aligned}$$

Here the equality comes from (E.32) and the first inequality is obvious. For the second inequality we use $f(a)^2 \leq \phi(a-\zeta)$, $\int(x-\zeta)^2\phi(x-\zeta) = 1$, and $e^{2(a^*-a)\zeta} f(a)^2 \leq \chi(a)^{-1}\phi(a-\zeta) = \frac{2}{\rho}\phi(a)$. We also note $\Phi(\lambda-\zeta) \to 0$ if $\lambda \leq a^*$, as $\zeta - a^* \to \infty$. Hence, given (E.49), and noting $\int_\lambda^\infty \phi(a)(a-\lambda)^2 da \precsim \frac{1}{1+\lambda^2}\phi(\lambda)$, we have that $\|f - \bar{d}^* \times \widetilde{g}_1\| \to 0$, when λ satisfies the second part of (E.24). (Such λ always exists, because $\lambda = a^*$ satisfy the condition as $\zeta - a^* \to \infty$ and $\rho^{-1}\phi(a^*) = \phi(a^*-\zeta)/2 \to 0$.) Given (E.48), and noting $\|g\|^2 \precsim 1$ from (E.33) (note $\zeta - a^* \to \infty$ can only occur in case (ii) by the first part of (E.1)), we obtain $\|g - \bar{d}^* \times \widetilde{g}_1\| = o(\|g\|)$. Since $\|g - d^* \times \widetilde{g}_1\| \leq \|g - \bar{d}^* \times \widetilde{g}_1\|$ by definition of d^*, the "if" part follows from (E.31). We hence establish condition (E.23) for $q = 1$.

Now we prove the second part of (E.24). The paragraph after (E.49) already proves the "if" part of the second part of (E.24).

We now demonstrate the "only if" part. When $\zeta - \lambda \precsim 1$, it holds that, for all d,

$$\|f - d \times \widetilde{g}_1\|^2 \geq \int_{a^*}^\lambda f(a)^2 da \geq \frac{1}{2}(\Phi(\lambda-\zeta) - \Phi(a^*-\zeta)) \succsim 1. \qquad \text{(E.50)}$$

The first inequality comes from $\lambda \geq a^*$ as $\zeta - a^* \to \infty$, the second inequality comes from that $f(a)^2 \geq \frac{1}{2}\phi(a-\zeta)$ for all $a \geq a^*$, and the last inequality comes from $\zeta - \lambda \precsim 1$ and $\zeta - a^* \to \infty$. Combining (E.48) and (E.50), and recalling (E.31) and that $\|g\|^2 \precsim 1$ from (E.33) under $\zeta - a^* \to \infty$, we obtain that the condition $\zeta - \lambda \to \infty$ is necessary.

Now suppose $\zeta - \lambda \to \infty$ holds but $\rho^{-1}\phi(\lambda) \succsim (1+\lambda^2)(\zeta-\lambda)^2$. This indicates $\rho^{-1}\phi(\lambda) \to \infty$. Since $\rho^{-1}\phi(a^*) \to 0$ (see after (E.49)), we have $(a^*-\lambda)/(1+\lambda) \to \infty$. Therefore, letting $l(a) = \mathbb{1}_{\{\lambda \leq a \leq a^*\}}$, we have, for all d,

$$\|\widetilde{g}_1 \times l\|^2 = \int_\lambda^{a^*} (e^{(a^*-a)\zeta}+1)^2(a-\lambda)^2 f(a)^2 da \geq \frac{2}{\rho}\int_\lambda^{a^*}(a-\lambda)^2\phi(a)da \succsim \frac{\phi(\lambda)}{\rho(1+\lambda^2)} \succsim (\zeta-\lambda)^2. \qquad \text{(E.51)}$$

The second inequality comes from $(e^{(a^*-a)\zeta}+1)^2 f(a)^2 = (1+\chi(a)^{-1})\phi(a-\zeta)$ and $\chi(a) = \frac{\rho}{2}\phi(a-\zeta)/\phi(a)$ by definition. The third inequality comes from $(a^*-\lambda)/(1+\lambda) \to \infty$. The last inequality comes from $\rho^{-1}\phi(\lambda) \gtrsim (1+\lambda^2)(\zeta-\lambda)^2$ that we suppose. Since $\|f \times l\|^2 = \int_{\lambda}^{a^*} f(a)^2 da \le \Phi(a^* - \zeta) \to 0$, given (E.48) and (E.51), we obtain that, for all d,

$$\|g - d \times \widetilde{g}_1\| = \|f - d \times \widetilde{g}_1\| + o(1) \ge \|(f - d \times \widetilde{g}_1) \times l\| + o(1) \gtrsim d \times (\zeta - \lambda) + o(1). \tag{E.52}$$

Given (E.52), and recalling $\|g\|^2 \eqsim 1$ from (E.33), $\|g - d \times \widetilde{g}_1\| = o(1)$ can hold only if $d = o((\zeta-\lambda)^{-1})$. However, for all $d \le \frac{1}{4}(\zeta-\lambda)^{-1}$, it holds that

$$\|f - d \times \widetilde{g}_1\|^2 \ge \int_{a^*}^{\zeta} (1 - d \times (e^{(a^*-a)\zeta} + 1)(a - \lambda))^2 f(a)^2 da \ge \frac{1}{4}\int_{a^*}^{\zeta} f(a)^2 da \gtrsim \frac{1}{2} - \Phi(a^* - \zeta) \gtrsim 1. \tag{E.53}$$

The second inequality comes from that $e^{(a^*-a)\zeta} + 1 \le 2$ and $a - \lambda \le \zeta - \lambda$ for all $a \in (a^*, \zeta)$, the third inequality comes from that $f(a)^2 \ge \frac{1}{2}\phi(a - \zeta)$ for all $a \ge a^*$, and the last inequality comes from $\zeta - a^* \to \infty$. Given (E.48), (E.53) contradicts $\|g - d \times \widetilde{g}_1\| = o(1)$. Therefore, recalling (E.31), we conclude that $\rho^{-1}\phi(\lambda) = o((1+\lambda^2)(\zeta-\lambda)^2)$ is indeed necessary. The second part of (E.24) has been proved.

Step 6. In this step we finish the proof regarding condition (E.23) for $q = 2$. Given (E.29) and the privious analysis for cases (ia) and (ib), we only need to show that, in case (ii), $\theta_2(\lambda) \to 1$ for some sequence λ. Under $\lambda = a^*$ and $\bar{d}^* = \zeta^{-1}$, we obtain

$$\begin{aligned}
\|f - \bar{d}^* \times \widetilde{g}_2\|^2 &= \int_0^{a^*} f(a)^2 da + \int_{a^*}^{\infty} \left(1 - (e^{(a^*-a)\zeta} + 1) \times \frac{a}{\zeta}\right)^2 f(a)^2 da \\
&\le \int_0^{a^*} f(a)^2 da + 2\int_{a^*}^{\infty} \left(\frac{a - \zeta}{\zeta}\right)^2 f(a)^2 da + 2\int_{a^*}^{\infty} e^{2(a^*-a)\zeta} \frac{a^2}{\zeta^2} f(a)^2 da \\
&\le \iota\Phi(2\zeta - a^*) + \frac{2}{\zeta^2}\int_{a^*-\zeta}^{\infty} a^2\phi(a) da + \frac{2}{\zeta^2} e^{2a^*\zeta}\int_{a^*}^{\infty} (a+\zeta)^2 \phi(a+\zeta) da \\
&\lesssim \iota\Phi(2\zeta - a^*) + \frac{2}{\zeta^2}(1 + (a^* - \zeta)_+)^2 \Phi(\zeta - a^*) \\
&\quad + \frac{2}{\zeta^2}(1 + (a^* + \zeta))^2 e^{2a^*\zeta}\Phi(-a^* - \zeta).
\end{aligned} \tag{E.54}$$

Here the equality comes from (E.32) and the first inequality is obvious. For the second inequality we use $f(a)^2 \le \chi(a)\phi(a - \zeta) = \iota\phi(a - 2\zeta)$, $f(a)^2 \le \phi(a - \zeta)$, and $e^{-2a\zeta}\phi(a - \zeta) = \phi(a + \zeta)$. The last inequality comes from that $\int_x^{\infty} a^2\phi(a) da \eqsim (1 + x_+)^2 \Phi(-x)$ uniformly over x. Moreover, we note $(1 + (a^* + \zeta))e^{2a^*\zeta}\Phi(-a^* - \zeta) \eqsim e^{2a^*\zeta}\phi(a^* + \zeta) = \phi(a^* - \zeta) \eqsim (1 + (a^* - \zeta)_+)\Phi(\zeta - a^*)$ using (E.15). Furthermore, since we are in case (ii), it holds that $\iota\Phi(2\zeta - a^*) = o(\Phi(\zeta - a^*))$ according to (E.17) and (E.21), and that $1 + (a^* - \zeta)_+ = o(\zeta)$ (by the first part of (E.1) and $\zeta^* \eqsim \sqrt{\log N}$ due to the assumption on ρ). Substituting these results into (E.54), we have that, under $\lambda = a^*$ and $\bar{d}^* = \zeta^{-1}$,

$$\|f - \bar{d}^* \times \widetilde{g}_2\|^2 = o(\Phi(\zeta - a^*)). \tag{E.55}$$

On the other hand, we note $\|f - g\|^2 \leq \Phi(-\zeta)$ (see below (E.49)) and $\Phi(-\zeta) = o(\Phi(\zeta - a^*))$ due to $1 + (a^* - \zeta)_+ = o(\zeta)$ (see above (E.55)). Combining these results with (E.55), and noting $\|g\|^2 \eqsim \Phi(\zeta - a^*)$ in case (ii) according to (E.33), we finally obtain that, under $\lambda = a^*$ and $\bar{d}^* = \zeta^{-1}$,

$$\|g - \bar{d}^* \times \widetilde{g}_2\| = o(\|g\|). \tag{E.56}$$

Since $\|g - d^* \times \widetilde{g}_1\| \leq \|g - \bar{d}^* \times \widetilde{g}_1\|$ by definition of d^*, we prove $\theta_2(\lambda) \to 1$ in case (ii) for some sequence λ. (E.23) is completely established.

Next, it follows that, under $\lambda \leq a^*$ and letting $\bar{d}^* = \zeta^{-1}$,

$$\|g - d^* \times \widetilde{g}_2\|^2 \tag{E.57}$$

$$\leq \|g - \bar{d}^* \times \widetilde{g}_2\|^2 \leq o(\|g\|^2) + 2\zeta^{-2} \int_{\lambda}^{a^*} a^2 \frac{1 + \chi(a)}{\chi(a)} \phi(a - \zeta) da$$

$$\leq o(\|g\|^2) + \frac{8}{\zeta^2 \rho} \int_{\lambda}^{a^*} a^2 \phi(a) da \leq o(\|g\|^2) + \frac{8a^{*2}}{\zeta^2 \rho} \Phi(-\lambda) \leq o(\|g\|^2) + O(\rho^{-1}\Phi(-\lambda)). \tag{E.58}$$

The second inequality comes from (E.56), the triangle inequality, the definition of $\widetilde{g}_2(a)$ given after (E.28), and that $\widetilde{\psi}_2(a, \lambda) - \widetilde{\psi}_2(a, a^*) = a\mathbb{1}_{\{\lambda \leq a \leq a^*\}}$ for all $a \geq 0$. The third inequality comes from $\frac{1+\chi(a)}{\chi(a)}\phi(a - \zeta) \leq \frac{2}{\chi(a)}\phi(a - \zeta) = \frac{4}{\rho}\phi(a)$ for $a \leq a^*$. The fourth inequality is obvious and the last comes from $a^* \lesssim \zeta$ as we are in case (ii). (E.58), (E.29), and (E.31) together proves (E.25). ∎

Lemma E4. *Suppose the same assumptions as in Lemma E1. Then, for any deterministic positive λ sequence satisfying $\rho N \Phi(\zeta - \lambda) \to \infty$, it holds that, for $q \in \{1, 2\}$,*

$$\|\widetilde{\psi}_q(\lambda)\|^2 = (1 + o_{\mathrm{P}}(1)) N \mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2).$$

Here $\widetilde{\psi}_q(\lambda)$ stands for the N-dimensional vector whose components are $\widetilde{\psi}_{q,i}(\lambda)$.

Proof. Throughout the proof λ is an arbitrary sequence as described in the statement of the lemma. The strategy is to calculate the magnitude of $\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)$ and $\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^4)$, and then establish the probability limit using Chebyshev's inequality. It holds by the definition of $\widetilde{\psi}_{q,i}$ and (E.2) (the condition $\rho \to 0$ it relies on is assumed here too) that, for $q \in \{1, 2\}$,

$$\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^{2q}) = 2(1 + o(1)) \int_0^{\infty} \widetilde{\psi}_q(a, \lambda)^{2q} \left(\phi(a) + \frac{\rho}{2}\phi(a - \zeta)\right) da. \tag{E.59}$$

On the other hand, we can calculate, for $j \in \{2, 4\}$,

$$\int_0^\infty \widetilde{\psi}_1(a,\lambda)^j\phi(a)da = \int_0^\infty a^j\phi(\lambda+a)da \eqsim \iota_{1,1}^j\Phi(-\lambda), \tag{E.60}$$

$$\int_0^\infty \widetilde{\psi}_1(a,\lambda)^j\phi(a-\zeta)da = \int_0^\infty a^j\phi(\lambda-\zeta+a)da \eqsim \iota_{1,2}^j\Phi(\zeta-\lambda), \tag{E.61}$$

where we use short-hand notation $\iota_{1,1} := (1+\lambda)^{-1}$ and $\iota_{1,2} := \frac{1+(\zeta-\lambda)_+}{1+(\lambda-\zeta)_+}$. Similarly, we have, for $j \in \{2,4\}$,

$$\int_\lambda^\infty \widetilde{\psi}_2(a,\lambda)^j\phi(a)da = \int_0^\infty (\lambda+a)^j\phi(\lambda+a)da \eqsim \iota_{2,1}^j\Phi(-\lambda), \tag{E.62}$$

$$\int_\lambda^\infty \widetilde{\psi}_2(a,\lambda)^j\phi(a-\zeta)da = \int_0^\infty (\lambda+a)^j\phi(\lambda-\zeta+a)da \eqsim \iota_{2,2}^j\Phi(\zeta-\lambda). \tag{E.63}$$

where $\iota_{2,1} = \lambda + (1+\lambda)^{-1}$ and $\iota_{2,2} = \lambda + \frac{1+(\zeta-\lambda)_+}{1+(\lambda-\zeta)_+}$. Then it holds that, for $q \in \{1,2\}$,

$$\begin{aligned}\frac{\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^4)}{\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)^2} &\lesssim \frac{\iota_{q,1}^4\Phi(-\lambda)+\iota_{q,2}^4\rho\Phi(\zeta-\lambda)}{\iota_{q,1}^4\Phi(-\lambda)^2+\iota_{q,2}^4\rho^2\Phi(\zeta-\lambda)^2} \le \iota_{q,2}^2\frac{\iota_{q,1}^2\Phi(-\lambda)+\iota_{q,2}^2\rho\Phi(\zeta-\lambda)}{\iota_{q,1}^4\Phi(-\lambda)^2+\iota_{q,2}^4\rho^2\Phi(\zeta-\lambda)^2} \\ &\le \iota_{q,2}^2\frac{2}{\iota_{q,1}^2\Phi(-\lambda)+\iota_{q,2}^2\rho\Phi(\zeta-\lambda)} \le \frac{2}{\rho\Phi(\zeta-\lambda)}.\end{aligned}$$

Here we obtain the first inequality by substituting (E.60), (E.61), (E.62), and (E.63) into (E.59). The second inequality comes from that $\iota_{q,1} \le \iota_{q,2}$ for $q \in \{1,2\}$. Since $\rho N\Phi(\zeta-\lambda) \to \infty$ by assumption, we have, for $q \in \{1,2\}$,

$$\frac{\mathrm{Var}(\|\widetilde{\psi}_q\|^2)}{(\mathrm{E}(\|\widetilde{\psi}_q\|^2))^2} \le \frac{\mathrm{E}(\widetilde{\psi}_{q,i}^4)}{N\mathrm{E}(\widetilde{\psi}_{q,i}^2)^2} \lesssim \frac{1}{N\rho\Phi(\zeta-\lambda)} \to 0. \tag{E.64}$$

Here we suppress the argument λ of $\widetilde{\psi}_{q,i}$ for simplicity and the first inequality comes from that $\widetilde{\psi}_{q,i}$ is i.i.d. across i. The current lemma directly follows from (E.64) and Chebyshev's inequality. ∎

Lemma E5. *Suppose the same assumptions as in Lemma E1. Let* $\psi_{\Delta,1}(a,\lambda,b) := b\mathbb{1}_{\{|a|-\lambda\ge -b\}}$ *and* $\psi_{\Delta,2}(a,\lambda,b) := \lambda\mathbb{1}_{\{||a|-\lambda|\le b\}}$. *Also suppose that* λ_N *and* b_N *are two deterministic positive sequences satisfying* $\rho N\Phi(\zeta-\lambda_N) \to \infty$ *and* $b_N = o\left(1/\sqrt{\log N}\right)$. *Then it holds that, for* $q \in \{1,2\}$,

$$\mathrm{E}(|\psi_{\Delta,q}(\check{z}_i,\lambda_N,b_N)|^2) = o(\mathrm{E}(\widetilde{\psi}_q(\check{z}_i,\lambda_N)^2)).$$

Proof. Throughout the proof λ_N and b_N are two arbitrary sequences as described in the statement of the lemma. For simplicity, we omit the subscript N of λ_N and b_N and also omit the last two arguments of $\psi_{\Delta,q}$. Using (E.2) (the condition $\rho \to 0$ it relies on is assumed here too), we have, for $q \in \{1,2\}$,

$$\mathrm{E}(\psi_{\Delta,q}(\check{z}_i)^2) = 2(1+o(1)) \int_0^\infty \psi_{\Delta,q}(a)^2 \left(\phi(a) + \frac{\rho}{2}\phi(a-\zeta)\right) da. \tag{E.65}$$

To evaluate the right-hand side of (E.65), we write

$$\begin{aligned}
\int \psi_{\Delta,1}(a)^2\phi(a)da &= b^2\int_{\lambda-b}^\infty \phi(a)da = b^2\Phi(b-\lambda), &\text{(E.66)}\\
\int \psi_{\Delta,1}(a)^2\phi(a-\zeta)da &= b^2\int_{\lambda-b}^\infty \phi(a-\zeta)da = b^2\Phi(\zeta+b-\lambda), &\text{(E.67)}\\
\int \psi_{\Delta,2}(a)^2\phi(a)da &= \lambda^2\int_{\lambda-b}^{\lambda+b} \phi(a)da \le 2b\lambda^2\phi((\lambda-b)_+), &\text{(E.68)}\\
\int \psi_{\Delta,2}(a)^2\phi(a-\zeta)da &= \lambda^2\int_{\lambda-b}^{\lambda+b} \phi(a-\zeta)da \le 2b\lambda^2\phi((\lambda-b-\zeta)_+). &\text{(E.69)}
\end{aligned}$$

On the other hand, it holds that, for all sequence $(a,\bar{a})$ satisfying $|\bar{a}-a| \le b$,

$$\begin{aligned}
b^2\Phi(-\bar{a}) &\le \frac{c_N}{(1+a_+)\log N}\phi(\bar{a}_+) \le c_N(1+a_+)^{-3}(\phi(a_+)+c_N N^{-1})\\
&\le c_N(1+a_+)^{-2}(\Phi(-a)+c_N N^{-1}). \qquad \text{(E.70)}
\end{aligned}$$

The first inequality comes from $b = o\left(1/\sqrt{\log N}\right)$ and (E.15). The second comes from (D.18) of Lemma D4. For the last inequality we use (E.15), too. Further, we note that $(1+\lambda_+)^{-1} \le \iota_{1,1} \le \iota_{1,2}$ and $(1+(\lambda-\zeta)_+)^{-1} \le \iota_{1,2}$ (introduced after (E.61)). Then, using (E.70) (choose $(a,\bar{a}) = (\lambda, \lambda-b)$ and $(a,\bar{a}) = (\lambda-\zeta, \lambda-\zeta-b)$ respectively), and noting $\rho \to 0$, $\iota_{1,1} \le \iota_{1,2}$, and $\rho\Phi(\zeta-\lambda) \gtrsim N^{-1}$ by assumption, we obtain

$$b^2\Phi(b-\lambda) + b^2\rho\Phi(\zeta+b-\lambda) = o(\iota_{1,1}^2\Phi(-\lambda) + \iota_{1,2}^2\rho\Phi(\zeta-\lambda)). \tag{E.71}$$

Given (E.65), (E.66), (E.67), (E.59), (E.60), and (E.61), we obtain from (E.71) that $\mathrm{E}(\psi_{\Delta,1}(\check{z}_i)^2) = o(\mathrm{E}(\widetilde{\psi}_{1,i}(\lambda)^2))$, which proves the lemma for $q=1$.

Next, it holds that, for all sequence $(a,\bar{a})$ satisfying $|\bar{a}-a| \le b$,

$$b\phi(\bar{a}_+) \le \frac{c_N}{\sqrt{\log N}}\phi(\bar{a}_+) \le c_N(1+a_+)^{-1}\phi(a_+) + c_N N^{-1} \le c_N\Phi(-a) + c_N N^{-1}. \tag{E.72}$$

We recall $x_+ = \max\{x,0\}$. The first inequality comes from $b = o\left(1/\sqrt{\log N}\right)$. The second comes from (D.18) of Lemma D4. For the last inequality we use (E.15), too. Further, we note that $\lambda \le \iota_{2,1}$ and $\lambda \le \iota_{2,2}$ (introduced after (E.61)). Then, using (E.72) (choose $(a,\bar{a}) = (\lambda, \lambda-b)$ and $(a,\bar{a}) = (\lambda-\zeta, \lambda-\zeta-b)$ respectively), and noting $\rho \to 0$, $\iota_{2,1} \le \iota_{2,2}$, and $\rho\Phi(\zeta-\lambda) \gtrsim N^{-1}$ by assumption, we obtain

$$b\lambda^2\Phi(b-\lambda) + b\lambda^2\rho\Phi(\zeta+b-\lambda) = o(\iota_{1,1}^2\Phi(-\lambda) + \iota_{1,2}^2\rho\Phi(\zeta-\lambda)). \tag{E.73}$$

Given (E.65), (E.68), (E.69), (E.59), (E.62), and (E.63), we obtain from (E.73) that $\mathrm{E}(\psi_{\Delta,2}(\check{z}_i)^2) = o(\mathrm{E}(\widetilde{\psi}_{2,i}(\lambda)^2))$, which proves the lemma for $q=2$. ∎

Lemma E6. *Suppose the same assumptions as in Proposition B2. Suppose λ is a positive deterministic sequence satisfying $\rho N\Phi(\mu-\lambda) \to \infty$. Also suppose λ' is a $\mathcal{G}$-measurable sequence λ' that satisfies $|\lambda'-\lambda| = o_{\mathrm{P}}\left(1/\sqrt{\log N}\right)$. Then it holds that $\widehat{S}'_q(\lambda') = S_q(\lambda) + o_{\mathrm{P}}(S^{\mathrm{OPT}}+1)$ for $q=\{1,2\}$.*

Proof. Throughout the proof λ and λ' are arbitrary sequences as described in the statement of the lemma. Every result holds for $q\in\{1,2\}$.

We first provide a bound on $\|\widehat{\psi}_q(\lambda') - \widetilde{\psi}_q(\lambda)\|^2$. Since $\check{z}_i = \sqrt{T}(s_i+\bar{\varepsilon}_i)$ and we have $\max_i|\bar{\varepsilon}_i| \lesssim_{\mathrm{P}} \sqrt{(\log N)/T}$ by uniform bound on i.i.d. normal variables and $|s_i| \lesssim N^{-d}$ by assumption, we obtain $\sqrt{(\log N)/T}\max_i|\check{z}_i| = o_{\mathrm{P}}\left(1/\sqrt{\log N}\right)$ as $(\log N)^3/T \to 0$ by assumption. Combining this result with the first part of (D.12) of Lemma D2, and noting $T=o(N)$ by assumption, we obtain

$$\sup_{i\le N}|\widehat{z}_i - \check{z}_i| = o_{\mathrm{P}}\left(1/\sqrt{\log N}\right). \tag{E.74}$$

Further, by definition it holds that, for all a, $b\ge 0$, and $\lambda\ge 0$,

$$\sup_{(a',\lambda'):|a'-a|+|\lambda'-\lambda|\le b}|\widetilde{\psi}_q(a',\lambda') - \widetilde{\psi}_q(a,\lambda)| \le \psi_{\Delta,q}(a,\lambda,b) + \mathbb{1}_{\{q=2\}}\psi_{\Delta,1}(a,\lambda,b), \tag{E.75}$$

where $\psi_{\Delta,q}$ is introduced in the statement of Lemma E5. Substituting (E.74) and $|\lambda'-\lambda| = o_{\mathrm{P}}\left(1/\sqrt{\log N}\right)$ (by assumption) into (E.75) (choose $(a',a) = (\widehat{z}_i,\check{z}_i)$), applying Lemma E5, noting $|\widetilde{\psi}_1(a,\lambda)| \le |\widetilde{\psi}_2(a,\lambda)|$ by definition, and using Chebyshev's inequality, we obtain

$$\|\widehat{\psi}_q(\lambda') - \widetilde{\psi}_q(\lambda)\|^2 = o(N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)). \tag{E.76}$$

Using $\max_{i\le N}|\widehat{\sigma}_i/\sigma_i - 1| \lesssim_{\mathrm{P}} c_N$ by (D.1) of Lemma D1 and the triangular inequality, we further have

$$\|\sigma\widehat{w}'_q(\lambda') - \widetilde{\psi}_q(\lambda)\|^2 = o(N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)). \tag{E.77}$$

Next, we write

$$|\sigma\widehat{w}'_q(\lambda')^{\intercal}\psi - \widetilde{\psi}_q(\lambda)^{\intercal}\psi| \le \|\sigma\widehat{w}'_q(\lambda')^{\intercal} - \widetilde{\psi}_q(\lambda)\|\|\psi\| = o_{\mathrm{P}}\left(\sqrt{N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)}(1+S^{\mathrm{OPT}})\right). \tag{E.78}$$

The first inequality comes from Cauchy-Schwarz inequality. For the last equality we utilize $\|\psi\| \lesssim_{\mathrm{P}} 1 + S^{\mathrm{OPT}}$ by Corollary 1 and (E.77). On the other hand, we have

$$\mathrm{Var}(\widetilde{\psi}_q(\lambda)^\intercal \psi) \leq N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2 \psi_i^2) \leq N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2 \alpha_i^2) = o(N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)). \tag{E.79}$$

The first inequality comes from that both $\widetilde{\psi}_{q,i}(\lambda)$ and ψ_i are i.i.d. across i. The second inequality comes from $\psi_i = \mathrm{E}(\alpha_i|\mathcal{G})$ and that $\widetilde{\psi}_{q,i}(\lambda)$ is $\mathcal{G}$-measurable. The last equality comes from $|\alpha_i| \leq \mu = o(1)$. As a result of (E.79) and Chebyshev's inequality, we have

$$\widetilde{\psi}_q(\lambda)^\intercal \psi = N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)\psi_i) + o_{\mathrm{P}}\left(\sqrt{N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)}\right). \tag{E.80}$$

Combining (E.78) and (E.80), we obtain

$$\sigma \widehat{w}_q'(\lambda')^\intercal \psi = N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)\psi_i) + o_{\mathrm{P}}\left(\sqrt{N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)}(1 + S^{\mathrm{OPT}})\right). \tag{E.81}$$

Moreover, it holds that

$$\begin{aligned} \sigma^2 \|\widehat{w}_q'(\lambda')\|^2 &= \|\widetilde{\psi}_q(\lambda)\|^2 + O(\|\sigma \widehat{w}_q'(\lambda') - \widetilde{\psi}_q(\lambda)\|^2 + \|\widetilde{\psi}_q(\lambda)\| \|\sigma \widehat{w}_q'(\lambda') - \widetilde{\psi}_q(\lambda)\|) \\ &= (1 + o_{\mathrm{P}}(1)) N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2). \end{aligned} \tag{E.82}$$

The first inequality comes from Cauchy-Schwarz inequality. The second equality is a direct result of Lemma E4 and (E.77).

Next, we note that $\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)) = 0$ by symmetry of $\widetilde{\psi}(a, \lambda)$ in a. Because $\widetilde{\psi}_{q,i}(\lambda)$ is i.i.d. across i and independent of β, it follows

$$\mathrm{E}(\|\beta^\intercal \widetilde{\psi}_q(\lambda)\|^2 | \beta) \lesssim \mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2) N \|\beta\|_{\mathrm{MAX}}^2 \lesssim_{\mathrm{P}} N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2). \tag{E.83}$$

Then we have

$$\sigma \|\beta^\intercal \widehat{w}_q'(\lambda')\| \leq \|\beta^\intercal \widetilde{\psi}_q(\lambda)\| + \|\beta^\intercal (\sigma \widehat{w}_q'(\lambda') - \widetilde{\psi}_q(\lambda))\| \lesssim_{\mathrm{P}} \sqrt{N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)} + c_N N \sqrt{\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)}. \tag{E.84}$$

where the last inequality comes from (E.83) and (E.77). Therefore, we obtain

$$\|\mathbb{M}_\beta \widehat{w}_q'(\lambda')\|^2 = \|\widehat{w}_q'(\lambda')\|^2 - \widehat{w}_q'(\lambda')^\intercal \mathbb{P}_\beta \widehat{w}_q'(\lambda') = \sigma^2 \|\widehat{w}_q'(\lambda')\|^2 + o_{\mathrm{P}}(N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)), \tag{E.85}$$

$$\widehat{w}_q'(\lambda')^\intercal \mathbb{M}_\beta \psi = \widehat{w}_q'(\lambda')^\intercal \psi + \widehat{w}_q'(\lambda')^\intercal \mathbb{P}_\beta \psi = \widehat{w}_q'(\lambda')^\intercal \psi + o_{\mathrm{P}}\left(\sqrt{N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)}\right). \tag{E.86}$$

The last equalities of both results come from (E.84) and $\lambda_{\min}(\beta^\intercal \beta) \gtrsim_{\mathrm{P}} N$. For the last equality of (E.86), we also use $\|\beta^\intercal\psi\| \lesssim_{\mathrm{P}} N^{1/2}\mathrm{E}(\alpha_i^2)^{1/2} \lesssim_{\mathrm{P}} c_N N^{1/2}$. We now conclude that

$$\begin{aligned}\frac{\widehat{w}_q'(\lambda')^\intercal \mathbb{M}_\beta \psi}{\|\mathbb{M}_\beta \widehat{w}_q'(\lambda')\|} &= \frac{(1+o_{\mathrm{P}}(1))\widehat{w}_q'(\lambda')^\intercal\psi}{\sqrt{N\mathrm{E}(\widetilde{\psi}_{q,i}(\lambda)^2)}} + o_{\mathrm{P}}(1) \\ &= (1+o_{\mathrm{P}}(1))S_q(\lambda) + o_{\mathrm{P}}(1+S^{\mathrm{OPT}}) = S_q(\lambda) + o_{\mathrm{P}}(S^{\mathrm{OPT}}+1). \quad \text{(E.87)}\end{aligned}$$

The first equality comes from (E.85), (E.86), and (E.82). The second comes from (E.81). The last equality comes from $S_q(\lambda) \le N^{1/2}\mathrm{E}(\psi_i^2)^{1/2} = S^{\mathrm{OPT}}$, in which the inequality is just Cauchy-Schwarz and the equality holds by definition.

Finally, (C.5) and (C.6) establishes that, under Assumption 1 and for all $\mathcal{G}$-measurable w, $w^\intercal(\alpha - \mathrm{E}(\alpha|\mathcal{G}))/\|w\| = o_{\mathrm{P}}(1)$. Choosing $w = \widehat{w}_q'(\lambda')^\intercal\mathbb{M}_\beta$, and noting $\psi = \sigma^{-1}\mathrm{E}(\alpha|\mathcal{G})$, we obtain

$$\widehat{S}_q'(\lambda') = \frac{\widehat{w}_q'(\lambda')^\intercal\mathbb{M}_\beta\alpha}{\sigma\|\mathbb{M}_\beta\widehat{w}_q'(\lambda')\|} = \frac{\widehat{w}_q'(\lambda')^\intercal\mathbb{M}_\beta\psi}{\|\mathbb{M}_\beta\widehat{w}_q'(\lambda')\|} + o_{\mathrm{P}}(1). \quad \text{(E.88)}$$

Given (E.87) and (E.88), we prove the lemma. ∎

Lemma E7. *Suppose the same assumptions as in Proposition B2. Then it holds that* $\widehat{S}_q'(\lambda) = o_{\mathrm{P}}(1)$*, for all* $\mathcal{G}$*-measurable positive sequence* λ *satisfying* $\mu^2\rho N\Phi(\zeta-\lambda) = o_{\mathrm{P}}(1)$ *and for* $q \in \{1,2\}$*.*

Proof. Throughout the proof λ is an arbitrary sequence as described in the statement of the lemma. Every result holds for $q \in \{1,2\}$. We let λ^* be a deterministic sequence such that $\lambda \ge \lambda^*$ in probability and $\mu^2\rho N\Phi(\zeta-\lambda^*) \to 0$, which is apparently always feasible.

When $\Phi(\zeta-\lambda^*) \gtrsim 1$, we have $\mu^2\rho N \to 0$ and thereby $(S^{\mathrm{OPT}})^2 = N\mathrm{E}(\psi_i^2) \le N\mathrm{E}(s_i^2) \le \rho N\mu^2 \to 0$. Since $\widehat{S}_q'(\lambda)$ is the Sharpe ratio generated by $\widehat{w}_q'(\lambda)^\intercal\mathbb{M}_\beta$, which is $\mathcal{G}$-measurable and satisfy $\widehat{w}_q'(\lambda)^\intercal\mathbb{M}_\beta\beta = 0$. Then we obtain $\widehat{S}_q'(\lambda) \le S^{\mathrm{OPT}} + o_{\mathrm{P}}(1) = o_{\mathrm{P}}(1)$ from the second part of Theorem 1 and Corollary 1 (the assumptions of the lemma obviously guarantee the prerequisites of both). Since $-\widehat{S}_q'(\lambda)$ is the Sharpe ratio generated by $-\widehat{w}_q'(\lambda)^\intercal\mathbb{M}_\beta$, we obtain $-\widehat{S}_q'(\lambda) \le o_{\mathrm{P}}(1)$, too.

When $\Phi(\zeta-\lambda^*) \to 0$, we have $\Phi(-\lambda^*) \le \Phi(\zeta-\lambda^*) \to 0$. Moreover, it holds that, for all sequence $(a,\bar{a})$ satisfying $|\bar{a}-a| = o\left(1/\sqrt{\log N}\right)$,

$$\Phi(-\bar{a}) \lesssim \frac{1}{1+a_+}\phi(\bar{a}_+) \lesssim \frac{1}{1+a_+}(\phi(a_+) + c_N N^{-2}) \lesssim \Phi(-a) + c_N N^{-2}. \quad \text{(E.89)}$$

The first inequality comes from $|\bar{a}-a| = o\left(1/\sqrt{\log N}\right)$ and (E.15). The second comes from (D.18) of Lemma D4. For the last inequality we use (E.15), too. Then, using (E.89) (choose $a = \lambda^*$ and $a = \lambda^* - \zeta$ respectively), we have that, for all $b_N = o\left(1/\sqrt{\log N}\right)$,

$$\begin{cases} N\Phi(b_N - \lambda^*) \lesssim N\Phi(-\lambda^*) + c_N N^{-1} = o(N), \\ \rho N\Phi(\zeta + b_N - \lambda^*) \lesssim \rho N\Phi(\zeta - \lambda^*) + c_N \rho N^{-1} \lesssim o(\mu^{-2}). \end{cases} \tag{E.90}$$

Here the last inequality comes from $\mu^2 \rho N\Phi(\zeta - \lambda^*) = o(1)$ and $\mu = o(1)$ by assumption. Hence, it holds that, for some $b_N = o\left(1/\sqrt{\log N}\right)$,

$$\begin{aligned} \sum_{i \le N} \mathbb{1}_{\{\widehat{w}'_{q,i}(\lambda) \neq 0\}} &\lesssim_{\mathrm{P}} \sum_{i \le N} \mathbb{1}_{\{|\check{z}_i| \ge \lambda^* - b_N\}} \\ &\lesssim_{\mathrm{P}} N\Phi(b_N - \lambda^*) + \rho N\Phi(\zeta + b_N - \lambda^*) = o(N), \end{aligned} \tag{E.91}$$

$$\sum_{i \le N} \mathbb{1}_{\{\widehat{w}'_{q,i}(\lambda) \neq 0, \alpha_i \neq 0\}} \lesssim_{\mathrm{P}} \sum_{i \le N} \mathbb{1}_{\{|\check{z}_i| \ge \lambda^* - b_N, \alpha_i \neq 0\}} \lesssim_{\mathrm{P}} \rho N\Phi(\zeta + b_N - \lambda^*) \lesssim o(\mu^{-2}). \tag{E.92}$$

For both results, the first inequality comes from (E.74), the second inequality comes from the density of $\check{z}_i$ and Chebyshev's inequality, the last inequality comes from (E.90). As a result, we have

$$\|\beta^\intercal \widehat{w}'_q(\lambda)\| \lesssim \|\beta\|_{\mathrm{MAX}} \|\widehat{w}'_q(\lambda)\| \sqrt{\sum_{i \le N} \mathbb{1}_{\{\widehat{w}'_{q,i}(\lambda) \neq 0\}}} = o_{\mathrm{P}}\left(\sqrt{N} \|\widehat{w}'_q(\lambda)\|\right). \tag{E.93}$$

Here the first inequality comes from Cauchy-Schwarz inequality, and the last equality comes from (E.91). Then we obtain

$$\|\mathbb{M}_\beta \widehat{w}'_q(\lambda)\|^2 = \|\widehat{w}'_q(\lambda)\|^2 - \widehat{w}'_q(\lambda)^\intercal \mathbb{P}_\beta \widehat{w}'_q(\lambda) = (1 + o_{\mathrm{P}}(1)) \|\widehat{w}'_q(\lambda)\|^2. \tag{E.94}$$

For the last equality, we use (E.93) and $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$. Similarly, using (E.93), $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$ and $\|\beta^\intercal \alpha\| \lesssim_{\mathrm{P}} N^{1/2} \mathrm{E}(\alpha_i^2)^{1/2} \lesssim_{\mathrm{P}} c_N N^{1/2}$, we obtain

$$\widehat{w}'_q(\lambda)^\intercal \mathbb{M}_\beta \alpha = \widehat{w}'_q(\lambda)^\intercal \alpha + \widehat{w}'_q(\lambda)^\intercal \mathbb{P}_\beta \alpha = \widehat{w}'_q(\lambda)^\intercal \alpha + o_{\mathrm{P}}(\|\widehat{w}'_q(\lambda)\|). \tag{E.95}$$

Combining (E.85) and (E.86), we now conclude that

$$\widehat{S}'_q(\lambda) = \frac{\widehat{w}'_q(\lambda)^\intercal \mathbb{M}_\beta \alpha}{\|\mathbb{M}_\beta \widehat{w}'_q(\lambda)\|} = (1 + o_{\mathrm{P}}(1)) \frac{\widehat{w}'_q(\lambda)^\intercal \alpha}{\|\widehat{w}'_q(\lambda)\|} + o_{\mathrm{P}}(1). \tag{E.96}$$

On the other hand, it holds that

$$\frac{|\widehat{w}'_q(\lambda)^\intercal \alpha|}{\|\widehat{w}'_q(\lambda)\|} \le \mu \frac{\sum_{i \le N} |\widehat{w}'_{q,i}(\lambda)| \mathbb{1}_{\{\widehat{w}_{q,i}(\lambda) \neq 0, \alpha_i \neq 0\}}}{\|\widehat{w}'_q(\lambda)\|} \le \mu \sqrt{\sum_{i \le N} \mathbb{1}_{\{\widehat{w}'_{q,i}(\lambda) \neq 0, \alpha_i \neq 0\}}} = o_{\mathrm{P}}(1). \tag{E.97}$$

The first inequality holds by $|s_i| = \frac{\mu}{\sigma} \mathbb{1}_{\{\alpha_i \neq 0\}}$, the second by Cauchy-Schwarz inequality, and the last equality holds by (E.92). The lemma follows from (E.96) and (E.97). ∎

References

Andrews, D. W. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*(5), 2153–2211.

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.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Folland, G. (2009). *A Guide to Advanced Real Analysis*. Number 37. MAA.

Horn, R. A. and C. R. Johnson (2012). *Matrix analysis*. Cambridge university press.

Liu, W. and Q.-M. Shao (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics 42*(5), 2003–2025.

Robbins, H. (1956). An empirical bayes approach to statistics. *Berkeley Symposium on Mathematical Statistics and Probability 3*, 157–163.