

The Statistical Limit of Arbitrage*

Rui Da[†]

Indiana University

Stefan Nagel[‡]

University of Chicago

NBER and CEPR

Dacheng Xiu[§]

University of Chicago

NBER

This version: October 13, 2024

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 Expectation, Portfolio Choice, Rare and Weak Signal, False Negatives, Empirical Bayes, Testing APT, Machine Learning

*We are grateful for comments from Adem Atmaz, Lars Hansen, Ye Luo, Andreas Neuhierl, Markus Pelger, Seth Pruitt, 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, and Wabash River Finance 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

Full-information rational expectations models assume that economic agents have perfect knowledge of the functional form and parameters of the data-generating process (DGP), thereby bypassing the statistical challenges real-world agents encounter when estimating the law of motion. While this assumption might be reasonable in a low-dimensional environment with ample data for learning, many economic settings are more realistically high-dimensional, with an unknown DGP 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 functional form and the parameters governing the DGP of returns, these arbitrageurs would exploit and, in turn, eliminate such opportunities. In reality, however, sophisticated investors seeking near-arbitrage opportunities lack perfect knowledge of the true functional form and parameters. Instead, they rely on statistical analysis to infer the existence of such opportunities from historical data, introducing statistical uncertainty. In some cases, such as derivatives pricing, this uncertainty 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 uncertainty can be substantial, imposing a meaningful limit on arbitrage.

To analyze the effects of arbitrageur learning, we consider a setting in which returns follow a statistical linear factor model. Near-arbitrage opportunities take the form of factor-neutral trading strategies with high Sharpe ratios. Exploiting these opportunities requires arbitrageurs to have knowledge of factor model alphas, which they must learn from historical return data. We derive the optimal Sharpe ratio achievable by any feasible arbitrage trading strategy, which is strictly dominated by the infeasible optimal Sharpe ratio that arbitrageurs could achieve if they were endowed with perfect knowledge of alphas. This, in turn, provides a new no-feasible-near-arbitrage bound on the Sharpe ratio that accounts for the statistical limit to arbitrage.

The difficulty of the learning problem depends on the DGP of alpha signals. While our theory does not generally rely on specific cross-sectional distributions of alpha signals, we use simple special cases to illustrate how the optimal Sharpe ratio 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 the distri-

bution of alpha. However, when alphas are weaker and more rare, inferring their distribution becomes more challenging, creating a gap between the optimal feasible Sharpe ratio 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 remains bounded.

The existence of this statistical limit to arbitrage implies that the range within which mispricing can survive in equilibrium is broader compared to a scenario where arbitrageurs have full knowledge of the DGP and its parameters. Mispricing can survive because it is obscured by statistical uncertainty. Empirically, it is the feasible Sharpe ratio—not the unattainable, infeasible one—that reveals the minimum reward-to-risk compensation arbitrageurs require.

We further demonstrate how arbitrageurs can construct a feasible trading strategy that achieves the theoretically optimal feasible Sharpe ratio, uniformly over DGPs of alphas, regardless of the strength and sparsity of alphas. This means that the feasible Sharpe ratio bound is in fact sharp. A uniformly valid trading strategy is desirable because, in practice, arbitrageurs do not know which DGP accurately represents the observed data. The optimal strategy involves estimating the empirical distribution of alpha signals and assigning portfolio weights based on the relative magnitudes and associated uncertainty of the alpha estimates. Assets with high alpha t -statistics receive portfolio weights proportional to their signal strength. Although weaker alphas are more difficult to exploit, ignoring them would lead to a suboptimal strategy. Instead, the optimal approach constructs portfolio weights for weak signals by locally smoothing alpha signals cross-sectionally.

To empirically contrast feasible and infeasible Sharpe ratios, we propose an estimator of the infeasible Sharpe ratio that a hypothetical arbitrageur endowed with perfect knowledge of DGP parameters would perceive. While this Sharpe ratio can be estimated consistently, it cannot be achieved by any feasible portfolio with weights constructed using 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 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.

Under certain conditions, the optimal feasible arbitrage strategy can be approximated using alternative approaches such as multiple testing, shrinkage, and selection applied to alphas estimated from cross-sectional regressions. These alternative approaches are used widely by

empirical asset pricing researchers and practitioners, making it valuable to understand how closely they approximate 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. Applying ridge-style shrinkage to portfolio weights yields a portfolio that is equivalent to one constructed with alpha estimates from plain cross-sectional regressions without shrinkage. This approach can achieve optimality when almost all true alphas are either uniformly strong or uniformly weak. Lasso-based selection aims to balance the strengths of the two previous methods, yielding a portfolio that closely approaches the theoretically optimal feasible Sharpe ratio, provided an optimal tuning parameter is chosen.

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 multi-factor 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 \(2020\)](#).¹

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^2 s 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

¹See openassetpricing.com (Stock-level Signal Datasets August 2023 Release).

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. These infeasible Sharpe ratios represent what arbitrageurs could achieve with perfect knowledge of the DGP parameters. 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.

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 DGP parameters. 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 statistical uncertainty, 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 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 an upper bound for all feasible Sharpe ratios without imposing

restrictions on the choice of strategies. Our analytical framework is primarily concerned with arbitrage portfolios, but our results are broadly 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.²

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.³ Our objective here is not on model testing or signal detection. Rather, we strive for the optimal economic performance of 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.

In macroeconomics and finance, rational expectation models in which decision makers are not confronted with statistical uncertainty about the structure of the economy have attracted criticism ([Hansen, 2007](#)). Bayesian learning is one way to expose model decision makers to statistical uncertainty. [Pastor and Veronesi \(2009\)](#) survey the literature on learning in financial markets. In many settings, e.g., [Collin-Dufresne et al. \(2016\)](#), learning can be sufficiently slow such that its effects persist in empirically realistic sample sizes, even though

²On the empirical side, [Guijarro-Ordóñez 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.

³[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.

convergence to rational expectations takes place in the long-run. An exception is [Martin and Nagel \(2021\)](#) where learning effects persist because investors face a high-dimensional inference problem about the process generating firm cash flows. Similarly, arbitrageurs in our model attempt to make inference on a high-dimensional parameter vector with a potentially insufficient sample size, but they learn about returns, not firms' underlying cash flows. We examine different sequences of DGPs and in most scenarios, our learning system does not converge to a rational expectations limit.⁴

Our paper is also related to [Chen et al. \(2021b\)](#) and [Chen et al. \(2021a\)](#) in that they also account for the distinction between beliefs of economic agents and the DGP revealed by empirical evidence. They model belief distortions as a change of measure in moment conditions, use statistical measures of divergence relative to rational expectations to bound the set of subjective probabilities, and seek robust inference with respect to this form of misspecification. In the spirit of [Hansen \(2014\)](#), we calculate an optimal feasible Sharpe ratio for arbitrageurs inside the economic model, which differs from the infeasible one as it appears to an econometrician outside the model. In our setting, the deviation from rational expectations stems naturally from the statistical obstacles faced by economic agents. A more subtle but important point is that the outcomes attainable by economic agents embracing machine learning methods in a high-dimensional environment differ from those that rational expectation agents endowed with knowledge of the economy's structure could achieve asymptotically.

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 unlike the classical empirical Bayes inference, our analysis allows for weak and rare alphas as motivated from economic restrictions, and also considers Sharpe ratios, not only the posterior mean of alphas.

⁴Our analysis is related to a large literature in econometrics and statistics that discuss uniform validity of asymptotic approximations, see, e.g., [Staiger and Stock \(1997\)](#), [Imbens and Manski \(2004\)](#), [Leeb and Pötscher \(2005\)](#), [Andrews et al. \(2020\)](#).

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 - 2.4 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 2.5 constructs a feasible trading strategy that achieves the bound, Section 2.6 proposes an estimator of the infeasible Sharpe ratio, and finally Section 2.7 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:

$$r_t = \alpha + \beta\gamma + \beta v_t + u_t, \quad (1)$$

where β is an $N \times K$ matrix of factor exposures (with the first column being a vector of 1s), 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 (first entry corresponding to the column of 1s is the zero beta rate), and u_t is a vector of idiosyncratic returns with a diagonal covariance matrix Σ_u and zero mean conditional on all the variables before time- t .⁵

We assume at any given time t , arbitrageurs examine a sample of size T , derived from Equation (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.⁶ To facilitate our asymptotic analysis along

⁵While 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.

⁶ This framework presents a slight deviation from the conventional scenario in which arbitrageurs observe

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 parameters. This distinction is made because γ and Σ_v serve as nuisance parameters in our analysis and their dimensions remain fixed. We assume that α has mean zero, and is cross-sectionally independent of β . These conditions are essential for identification of γ in a model that allows for pricing errors. We formalize these conditions below.

To allow for a more general DGP, we may consider a conditional version of (1):

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

where β_t is a vector of time-varying factor loadings and γ_t is a vector of time-varying risk premia.⁷ Despite the conditional model's greater generality, it does not provide additional economic insights in our analysis of the theoretical limits of arbitrage, compared with the unconditional model. While our primary theoretical findings, such as those provided in Theorems 1 and 2, remain applicable when β is substituted with β_t , we concentrate, for clarity, on the more stylized model (1) to illustrate our theory.

There are at least three variations of the factor model, depending on what econometricians assume to be observable. The most common setup in the academic finance literature imposes that factors are observable as, e.g., in Fama and French (1993).⁸ 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, arguably most prevalent among practitioners, is the MSCI Barra model originally proposed by Rosenberg (1974), where factor exposures are equal to stock characteristics and hence observable.⁹ 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. In our empirical work, we prefer to use the MSCI Barra model for its practicality in analyzing individual stock returns,

a sample spanning $t = 1, \dots, T$ and make their investment decision at time $T + 1$. Essentially, our approach highlights the continuous process of making investment decisions, rather than conceiving these decisions as occurring at a single, distant future point, T . That said, this conceptual difference does not result in any tangible difference in our theoretical and empirical results.

⁷This 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).

⁸This is different from saying factor innovations, v_t , are observable. The setting of observable factors typically involves another equation that $f_t = \mu + v_t$, where μ are the population means of the observed factors f_t , which are not necessarily identical to the factor risk premia, γ . Since μ is an unknown parameter, v_t is still not observable.

⁹See Kozak and Nagel (2023) on the conditions on factor construction and return covariance matrix that must hold for characteristics to equal factor loadings.

adhering to (2) but with a constant alpha.¹⁰

The constant alpha assumption in our theoretical analysis may seem restrictive. In practice, arbitrageurs may make use of trading signals that capture time-variation in 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:

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

This transformation converts a conditional model for individual stocks with time-varying alphas into an unconditional latent factor model for managed portfolios, $(c_{t-1}^\top c_{t-1})^{-1} c_{t-1}^\top r_t$, with constant alphas as in (1). Thus, via these managed portfolios, our theoretical framework also applies to a setting in which arbitrageurs make use of signals about time-varying individual stock 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 upon the insight of Ross (1976), Huberman (1982) and Ingersoll (1984) established the concept of near-arbitrage, which can be formalized in a more general setting as below:

Definition 1. A portfolio strategy w at time t is said to generate a near-arbitrage under a sequence of data-generating processes, such as (1), defined in a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_s\}_{s \leq t}, P)$, where \mathcal{F}_t contains information generated by $\alpha, \beta, \{r_s, v_s, u_s\}_{s \leq t}$, if it satisfies $w \in \mathcal{F}_t$, and along some diverging subsequence,¹¹ with probability approaching one,

$$\text{Var}(w^\top r_{t+1} | \mathcal{F}_t) \rightarrow 0, \quad E(w^\top r_{t+1} | \mathcal{F}_t) \geq \delta > 0.$$

Intuitively, no near-arbitrage means there exists no sequence of portfolios that earn positive expected returns with vanishing risks. Ingersoll (1984) established that a sufficient and necessary condition for the absence of near-arbitrage is that¹²

$$S^* := \sqrt{\alpha^\top \Sigma_u^{-1} \alpha} \lesssim_P 1. \quad (4)$$

¹⁰Empirically, we use a moving window method for alpha estimation, which effectively allows alphas to vary slowly over time.

¹¹We adopt the same subsequence definition as that used in Ingersoll (1984). The subsequence typically depends on the count of investment opportunities, i.e., N , though we do not need make this explicit in this definition. For simplicity of notation and without ambiguity, we omit the dependence of w on N and t .

¹²We use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$, $a \lesssim_P b$ to denote $a = O_P(b)$, $a \approx b$ if $a \lesssim b$ and $b \lesssim a$, and use $a \approx_P b$ accordingly. Throughout, c_N represent some sequence vanishing with N .

Here, S^* 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 alphas that are too large, to the extent that S^* explodes.¹³

To achieve this optimal Sharpe ratio, arbitrageurs should hold a factor-neutral portfolio with weights given by $w^* \propto \Sigma_u^{-1}\alpha$, 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 knowledge of parameters. Fundamentally, this requires that the learning problem in the limiting experiment becomes increasingly simpler as the sample size increases.

In the current context, the difficulty of the learning problem also hinges on the number of investment opportunities, N . 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 expectation case. In our setting, it turns out that the rational expectation limit S^* is only relevant for rather restrictive scenarios. In more realistic settings, e.g., N is much larger than T , the optimal Sharpe ratio arbitrageurs can achieve without factor exposures is far smaller than S^* because of their inability to make error-free inference. Therefore, the condition (4) could 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:*

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

¹⁴In Ingersoll (1984), α is defined to be the cross-sectional projection of the expected returns onto β in the population model, such that $\alpha^\top \Sigma_u^{-1} \beta = 0$. As a result, his arbitrage portfolio weights are factor neutral, i.e., $w^* \beta = 0$ 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 $E(\alpha^\top \beta) = 0$. Consequently, the factor-neutral optimal strategy derived in Theorem 1 and illustrated by (10) 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^* asymptotically as N increases.

$$\alpha_i \stackrel{i.i.d.}{\sim} \begin{cases} \mu & \text{with prob. } \rho/2 \\ -\mu & \text{with prob. } \rho/2 \\ 0 & \text{with prob. } 1 - \rho \end{cases}, \quad 1 \leq i \leq N, \quad (5)$$

where $\mu \geq 0$ and $0 \leq \rho \leq 1$, and they potentially vary with N and T . In addition, we also assume $\beta = 0$, $\Sigma_u = \sigma^2 \mathbb{I}_N$, for some $\sigma > 0$. 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. 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 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

$$\mu \approx T^{-1/2} \quad \text{and} \quad \rho \approx N^{-1/2}. \quad (6)$$

This condition (6) implies that the signal strength μ vanishes as the sample size increases ($T \rightarrow \infty$) and the signal percentage count ρ decays as the investment universe expands ($N \rightarrow \infty$). This setup is used to approximate a situation in which only a small portion of assets have a nonzero and small alpha. On the other hand, σ is assumed to be fixed, since in reality idiosyncratic risks never vanish, whereas alphas can be small due to competition among arbitrageurs. 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, $\alpha^\top \alpha$ would explode rather rapidly. Even in the current setting, in light of the fact that $E(\alpha^\top \alpha) = \rho \mu^2 N$, we still have $\alpha^\top \alpha \xrightarrow{P} \infty$ as long as $N^{1/2}/T \rightarrow \infty$. 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

¹⁵Adopting 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 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 $1 \leq i \leq N$,

precisely which assets among all have nonzero alpha.

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

$$\begin{aligned} \mathbb{E}(\sigma^{-2}(\alpha + \bar{u})^\top(\alpha + u_{t+1})|\mathcal{F}_t) &= \sigma^{-2}(\alpha^\top\alpha + \bar{u}^\top\alpha), \\ \text{Var}(\sigma^{-2}(\alpha + \bar{u})^\top(\alpha + u_{t+1})|\mathcal{F}_t) &= \sigma^{-2}(\alpha^\top\alpha + 2\alpha^\top\bar{u} + \bar{u}^\top\bar{u}), \end{aligned}$$

where u_{t+1} denotes a future de-meaned return at $t + 1$, that shares the same distribution as $\{u_s\}_{s \leq t}$, 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:

$$S^2 = \frac{\sigma^{-4}(\alpha^\top\alpha + \bar{u}^\top\alpha)^2}{\sigma^{-2}(\alpha^\top\alpha + 2\alpha^\top\bar{u} + \bar{u}^\top\bar{u})} \lesssim_P T^{-1} \rightarrow 0, \quad (8)$$

where we use the fact that $\bar{u}^\top\bar{u} \asymp_P N/T$ 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 $N, T \rightarrow \infty$. Conversely, the *infeasible* optimal Sharpe ratio, denoted S^* , diverges under the condition that $N^{1/2}/T \rightarrow \infty$. There is a vast disparity, as shown by this example, between S^{OPT} and S^* .

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 clarify, we formalize the definitions as follows:

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

where σ_i^2 is the i th 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 $(1 + \gamma^\top(\Sigma_v)^{-1}\gamma)$ disappears.

¹⁷The 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.

¹⁸For 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 \hat{w} , which we will denote by \hat{w}^{CSR} , fails to achieve the optimal Sharpe ratio in all scenarios. We will discuss the optimal strategy in Section 2.5.

Definition 2. *A strategy is deemed feasible when it relies solely on observable data, in stark contrast to an infeasible strategy that presupposes complete knowledge of the DGP, encapsulated by the full information set, \mathcal{F}_t .*

Recognizing the distinction between feasible and infeasible strategies, we proceed to investigate the maximum Sharpe ratio attainable by feasible strategies. This exploration necessitates defining the decision-making problem faced by arbitrageurs, which we turn to next.

2.3 Arbitrageurs' Decision Problem and Feasible Sharpe Ratio Bound

In the broader context of the DGP, given by equation (1), the information set \mathcal{F}_t not only includes α and $\{r_s\}_{s \leq t}$ but also other variables such as Σ_u , β , and $\{v_s\}_{s \leq t}$, alongside unknown parameters like Σ_v and γ . For a strategy to be considered feasible, it must refrain from leveraging these unknown variables and parameters. This constraint introduces complexity to the arbitrageurs' decision-making process by intertwining optimization with statistical techniques, resulting in solutions that are not uniquely defined.

Recognizing that the disparity in Sharpe ratios between infeasible and feasible strategies primarily stems from the challenges of learning about α , we seek to navigate this complexity by considering a broader set of trading strategies. These strategies deliberately exclude reliance on information about α , although they may not all be feasible. Given this constraint, we derive an upper bound for the Sharpe ratio that applies to all such strategies, and by extension, to all feasible strategies, since they similarly avoid using information on α . However, this upper bound may not be achievable by a feasible strategy, as the derivation does not preclude the use of potentially infeasible information on other variables, such as Σ_u . In Section 2.5, we demonstrate that it is indeed possible to attain this upper bound with a well-crafted feasible trading strategy. Consequently, even without perfect knowledge about other variables and parameters, arbitrageurs can effectively use statistical inference to estimate these unknowns, with the statistical error of such estimates diminishing asymptotically. This highlights that the main hurdle for arbitrageurs is not in learning about factors, volatilities, or risk premia, but in acquiring knowledge about α .

We are now ready to characterize the decision problem arbitrageurs confront. Operating under a certain information set, \mathcal{G} , a subset of \mathcal{F}_t , arbitrageurs at time t are tasked with solving a mean-variance optimization problem expressed as:

$$\tilde{w} = \arg \max_{w \in \mathcal{G}: w^\top \beta = 0} U(w), \quad \text{where} \quad U(w) = \mathbb{E}(w^\top r_{t+1} | \mathcal{G}) - \frac{\kappa}{2} \text{Var}(w^\top r_{t+1} | \mathcal{G}),$$

and κ measures the degree of risk aversion. Our objective is to identify the optimal strategy subject to the measurability constraint: $w \in \mathcal{G}$. Given the necessity for statistical arbitrage

to be factor-neutral, the strategy w must be orthogonal to β , with β being predetermined, i.e., $\beta \in \mathcal{G}$.

To streamline the portfolio formation process, we exclude transaction cost considerations, which yields a static, single-period optimization problem. Thus, we omit subscript t whenever possible. This approach aligns with the APT framework introduced by Ross (1976), where the arbitrageurs' chosen strategy aims to eliminate exposure to systematic factors, concentrating instead on capitalizing on alphas and balancing idiosyncratic risks. Arbitrageurs' utility is modeled as a Bayesian posterior expected loss, a concept discussed in Berger (1985). The next theorem presents a solution with minimal conditions on \mathcal{G} .

Theorem 1. *Suppose that r_t follows (1) and that β is \mathcal{G} -measurable. It holds that*

$$\tilde{w} = \frac{1}{\kappa} \tilde{\Sigma}_u^{-1/2} \mathbb{M}_{\tilde{\Sigma}_u^{-1/2} \beta} \tilde{\Sigma}_u^{-1/2} \tilde{\alpha}, \quad (9)$$

where $\tilde{\alpha} := E(\alpha|\mathcal{G})$ and $\tilde{\Sigma}_u := E(\Sigma_u|\mathcal{G}) + \text{Var}(\alpha|\mathcal{G})$.¹⁹

This result associates the optimal strategy with the posterior summary statistics of α and Σ_u . Essentially, the choice of the optimal \mathcal{G} -measurable strategy is determined by arbitrageurs' best parameter estimates based on their information set, rather than the true values of the unknown parameters. In the special scenario where $\mathcal{G} = \mathcal{F}_t$, meaning arbitrageurs possess complete information up to time t , we have $\tilde{\alpha} = \alpha$, and $\tilde{\Sigma}_u = \Sigma_u$. Under these conditions, the optimal infeasible strategy, denoted by w^* , is given by

$$w^* = \frac{1}{\kappa} \Sigma_u^{-1/2} \mathbb{M}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2} \alpha. \quad (10)$$

Moving forward, our attention is directed towards a specific choice of \mathcal{G} as generated by $\{(r_s, \beta, v_s, \Sigma_u) : t - T + 1 \leq s \leq t\}$.²⁰ Accordingly, \tilde{w} is assumed associated with this particular choice by default hereafter.

Next, we examine the investment performance of arbitrageurs' optimal strategy, \tilde{w} . It is important to note that in evaluating a strategy, the conditional Sharpe ratio is calculated based on the comprehensive information set, \mathcal{F}_t , rather than the arbitrageurs' specific information set, \mathcal{G} . For any given strategy w , its conditional Sharpe ratio is defined as:

$$S(w) := E(w^\top r_{t+1} | \mathcal{F}_t) / \text{Var}(w^\top r_{t+1} | \mathcal{F}_t)^{1/2}.$$

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

²⁰The nuisance parameters γ and Σ_v are deterministic constants and, as such, are encompassed within any information set. Although they are part of \mathcal{G} , their presence does not influence the optimal strategy of arbitrageurs.

To describe the asymptotic behavior of a strategy, we need to impose assumptions on the return generating process:

Assumption 1. *For each $N \geq 1$, the following conditions hold:*

- (a) *(α_i, u_i) is i.i.d. across i , and satisfies $E(\alpha_i | (\Sigma_u)_{i,i}) = 0$ and $E\|\alpha\|_{\text{MAX}}^2 = o(1)$. Moreover, it holds that $1 \lesssim_P \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) \lesssim_P 1$.²¹*
- (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 the alphas in our model are inherently weak; as the number of assets, N , increases, their magnitudes diminish.²² Moreover, this condition ensures that volatilities 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).

Given this assumption, the subsequent theorem illustrates that the optimal strategy \tilde{w} yields a Sharpe ratio that is nearly at its maximum.

Theorem 2. *Suppose that r_t follows (1) and that Assumption 1 holds. Let \mathcal{G} represent the information set generated by $\{(r_s, \beta, v_s, \Sigma_u) : t - T + 1 \leq s \leq t\}$. For any strategy w that is both \mathcal{G} -measurable and factor-neutral, it follows that as $N \rightarrow \infty$,*

$$S(w) \leq S(\mathcal{G}) + o_P(1 + S(\mathcal{G})), \quad (11)$$

where $S(\mathcal{G})^2 = \tilde{\alpha}^\top \Sigma_u^{-1} \tilde{\alpha}$, and the equality in (11) holds if $w = \tilde{w}$. Moreover, the optimal strategy for arbitrageurs, \tilde{w} , has a simple approximation:

$$\left\| \tilde{w} - \frac{1}{\kappa} \mathbb{M}_\beta \Sigma_u^{-1} \tilde{\alpha} \right\| = o_P(1 + S(\mathcal{G})). \quad (12)$$

Accordingly, their optimal utility satisfies $\sqrt{2\kappa U(\tilde{w})} = S(\mathcal{G}) + o_P(1 + S(\mathcal{G}))$.

Theorem 2 derives an upper bound, denoted as $S(\mathcal{G})$, on the Sharpe ratio that any \mathcal{G} -measurable strategy can attain. This upper bound also connects with arbitrageurs' optimal

²¹For a matrix A , we use $\|A\|$ and $\|A\|_{\text{MAX}} = \max_{i,j} |a_{ij}|$ 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 .

²²By Assumption 1(a), $\text{Var}(\alpha_i) = E(\alpha_i^2) = o(1)$. Referencing our earlier discussion (footnote 13), a diminishing variance in α is essential for precluding near-arbitrage opportunities within Ross' APT framework.

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

$$E(S(\mathcal{G})^2) \leq E(\alpha^\top \Sigma_u^{-1} \alpha), \quad (13)$$

where the equality holds only when $\tilde{\alpha} = \alpha$ almost surely. The right-hand side corresponds to the infeasible scenario in which arbitrageurs have perfect knowledge of α , which echoes (4).

Furthermore, the result demonstrates that the optimal strategy, as articulated in Theorem 1, can reach this upper bound with a negligible approximation error. In scenarios where $S(\mathcal{G})$ is finite, that is $O_P(1)$, the error term $o_P(1 + S(\mathcal{G}))$ reduces to $o_P(1)$ and vanishes in the limit. The theorem additionally addresses situations where $S(\mathcal{G})$ may diverge. In this case, the approximation error is reduced to $o_P(S(\mathcal{G}))$, remaining inconsequentially small relative to $S(\mathcal{G})$ itself.

Also, Equation (12) offers a more straightforward and intuitive formula for approximating the optimal feasible strategy, \tilde{w} . Additionally, as a byproduct of Theorem 2, the optimal infeasible strategy in Equation (10) can also be approximated in a simpler form:

$$\left\| w^* - \frac{1}{\kappa} \mathbb{M}_\beta \Sigma_u^{-1} \alpha \right\| = o_P(1 + S^*), \quad (14)$$

which represents a special case of Equation (12) where \mathcal{G} is set as \mathcal{F}_t . At its core, it is $\tilde{\alpha}$, the posterior estimate of the pricing errors that dictates the optimal Sharpe ratio achievable by arbitrageurs, rather than the true α . Intuitively, part of the construction in (12), $\Sigma_u^{-1} \tilde{\alpha}$, is the optimal allocation to the ex-factor returns, $\alpha + u_t = r_t - \beta(\gamma + v_t)$, based on a simple mean-variance analysis, except for the use of $\tilde{\alpha}$, because arbitrageurs do not observe true alphas in the DGP. Pre-multiplying $\Sigma_u^{-1} \tilde{\alpha}$ by \mathbb{M}_β simply eliminates the portfolio's factor exposures, because $(\mathbb{M}_\beta \Sigma_u^{-1} \tilde{\alpha})^\top \beta = 0$.

Theorems 1 and 2 are concerned with factor-neutral strategies, as our main results focus on the limits of arbitrage. Proposition B1 in our appendix broadens this scope to encompass any strategy, revealing that any \mathcal{G} -measurable strategy, w , adheres to a similar upper limit:

$$S(w) \leq (S(\mathcal{G})^2 + \gamma^\top \Sigma_v^{-1} \gamma)^{1/2} + o_P(1 + S(\mathcal{G})), \quad (15)$$

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

This result addresses 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^* . 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 only mild restrictions on the DGP of returns, equation (15) identifies an upper bound for all feasible Sharpe ratios without limiting the choice of strategies.

The result in Theorem 2 appears to require that arbitrageurs rely on the information set \mathcal{G} , which embodies perfect knowledge of factors, v_t , and their exposures, β , in addition to past asset returns, r_t . Moreover, arbitrageurs appear to have perfect knowledge of the covariance matrix of idiosyncratic errors, Σ_u . In fact, the upper bound in (15) still holds if arbitrageurs are endowed with less information, because for any information sets \mathcal{G}' and \mathcal{G} such that $\mathcal{G}' \subseteq \mathcal{G}$, we have $E(S(\mathcal{G}')^2) \leq E(S(\mathcal{G})^2)$. Since all feasible strategies are measurable with respect to \mathcal{G} , $S(\mathcal{G})$ establishes an upper bound for the Sharpe ratio applicable to this set of strategies, although it may not be sharp. In light of this and Definitions 1 and 2, we immediately obtain a sufficient condition of the absence of near-arbitrage with feasible strategies:

Corollary 1. *Suppose the same assumptions as in Theorem 2 hold. For any given return-generating process satisfying (1), there exists no feasible factor-neutral strategy that leads to a near-arbitrage, if*

$$S(\mathcal{G}) \lesssim_P 1, \quad \text{as } N \rightarrow \infty. \quad (16)$$

2.4 Bayes Correction for Selection Bias

While the previous discussion underscored the role of the posterior mean of α , denoted as $\tilde{\alpha} = E(\alpha|\mathcal{G})$, on the optimal \mathcal{G} -measurable portfolio, this expression remains implicit and not directly actionable. Our focus now shifts to reducing the information set \mathcal{G} to its sufficient statistics for α . This step is crucial for understanding why $\tilde{\alpha}$ leads to enhanced portfolio performance. To facilitate this analysis, we introduce further assumptions:

Assumption 2. *For each $N \geq 1$, the following conditions hold:*

- (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 DGP described in Equation (1) and given the information set \mathcal{G} , Assumptions 1 and 2(a) together ensure that the key summary statistics for α_i are the volatility σ_i and the sample average of the ex-factor returns, expressed as $\check{\alpha}_i := \bar{r}_i - \beta_i(\gamma + \bar{v}) = \alpha_i + \bar{u}_i$. In other words, $E(\alpha_i|\mathcal{G}) = E(\alpha_i|\check{\alpha}_i, \sigma_i)$. 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 .

In light of this, we introduce Assumption 2(b), which allows us to further express $E(\alpha_i|\check{\alpha}_i, \sigma_i) = \sigma_i E(s_i|\check{\alpha}_i, \sigma_i) = \sigma_i E(s_i|\check{\alpha}_i/\sigma_i)$.²³ This leads to $\tilde{s}_i := E(s_i|\mathcal{G}) = E(s_i|\check{s}_i)$, where $\check{s}_i := \check{\alpha}_i/\sigma_i$. 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 2, $\tilde{\alpha}$, \tilde{w} , and $S(\mathcal{G})^2$ can all be represented in relation to \tilde{s} .²⁴

$$\tilde{\alpha} = \Sigma_u^{-1/2}\tilde{s}, \quad \tilde{w} \propto \mathbb{M}_\beta \Sigma_u^{-1/2}\tilde{s}, \quad S(\mathcal{G})^2 = \tilde{s}^\top \tilde{s}. \quad (17)$$

Finally, note that

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

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 \check{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 \check{s} . This task is complicated by what Efron (2011) described as selection bias or “the winner’s curse”: A high observed signal, \check{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 \tilde{s} , which accounts for this bias. To see this, we present an explicit formula for \tilde{s} :

Theorem 3. *Suppose that r_t follows (1) and Assumptions 1 and 2 hold. We introduce a function $\psi(a) = E(s_i|\check{s}_i = a)$, with which we have $\tilde{\alpha}_i = \sigma_i \tilde{s}_i$, where $\tilde{s}_i = \psi(\check{s}_i)$. Moreover, it holds that*

$$\psi(a) = a + \frac{1}{T} \frac{d}{da} \log p(a), \quad (18)$$

where $p(a) = E(\phi_{1/T}(a - s_i))$ is the probability distribution function of \check{s}_i .²⁵

The preceding discussion directly leads to the first assertion of the theorem. Equation (18) is rooted in Tweedie’s formula (Robbins, 1956), which establishes a connection between the posterior mean of s , given $\check{s} = a$, denoted $\psi(a)$, and the posterior probability distribution of

²³This equality relies on the result that conditional on $\hat{\alpha}_i/\sigma_i$, α_i/σ_i is independent of σ_i . We impose this condition primarily for clarity of exposition and simplicity of Algorithm 1 below.

²⁴Given 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.

²⁵ $\phi_x(a)$ is the distribution function for $\mathcal{N}(0, x)$.

š. The formula’s second component, $T^{-1}d \log p(a)/da$, plays a crucial role in adjusting for the selection bias in the observed signal, \check{s}_i , which introduces several intriguing 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, under mild assumptions about the prior distribution of s_i , such as being symmetric and unimodal, the posterior mean \tilde{s} induces a shrinkage effect towards the prior mean, which, in our context, is zero.

The posterior shrinkage on α , or equivalently on s , consequently induces a “shrinkage” effect on the optimal Sharpe ratio achievable by arbitrageurs, as demonstrated by the inequality (13). Based on the results of Theorem 3, and under a technical condition concerning the tail behavior of s_i , we can obtain a more explicit formula for $S(\mathcal{G})$:

Corollary 2. *Under the same assumptions outlined in Theorem 3, and with the additional condition that $E(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) = o(N^{-1})$, we arrive at the conclusion that:*

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

This result enables us to compute S^{OPT} in various examples. For instance, by utilizing this result, we can compare S^{OPT} with S^* in Example 1. This comparison illuminates the feasibility of attaining non-vanishing Sharpe ratios and how different DGP conditions influence the attainability.

Corollary 3. *Suppose that the same assumptions as in Corollary 2 hold. In addition, we assume alpha follows (5) as in Example 1. Then we have $S^* = \sigma^{-1} \mu(\rho N)^{1/2} + o_P(1)$. Further, assuming that $\sigma^{-1} \mu(\rho N)^{1/2}$ does not vanish, then it holds that $S^{\text{OPT}} \leq (1 - \epsilon) \sigma^{-1} \mu(\rho N)^{1/2}$ for some $\epsilon > 0$, if and only if*

$$T^{1/2} \mu / \sigma - \sqrt{-2 \log \rho} \lesssim 1. \tag{19}$$

Corollary 3 suggests that when $T^{1/2} \mu / \sigma$ is large enough that the constraint (19) is violated, $S^* \approx_P S^{\text{OPT}}$ —that is, in the limit, learning does not play any role—arbitrageurs in this 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^* .

To give a concrete example of Corollary 3, consider an alternative DGP assumption instead of (6):²⁶

$$\mu \asymp N^{-\eta} \quad \text{and} \quad \rho > 0 \quad \text{is fixed.} \quad (20)$$

In this scenario, $(S^*)^2 \asymp_P N^{1-2\eta}$, which explodes unless $\eta > 1/2$. If we further assume that $N/T \rightarrow \psi > 0$, then the left-hand-side of condition (19) is of order $N^{1/2-\eta} \vee 1$, so that (19) holds if and only if $\eta \geq 1/2$. Therefore, $\eta < 1/2$ is not consistent with absence of (feasible) near arbitrage because S^* explodes and, since $S^{\text{OPT}} = S^*$ by Corollary 3, S^{OPT} explodes, too. If $\eta > 1/2$, S^* (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 (20), 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 3 using the DGP specified in Example 1. Figure 1 reports the Sharpe ratio, S^{OPT} , of optimal 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.

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 optimal Sharpe ratio 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 optimal Sharpe ratios drop to near zero.

²⁶It is easy to show that the setup (20) satisfies all assumptions of Corollary 2 for all fixed $\eta > 0$.

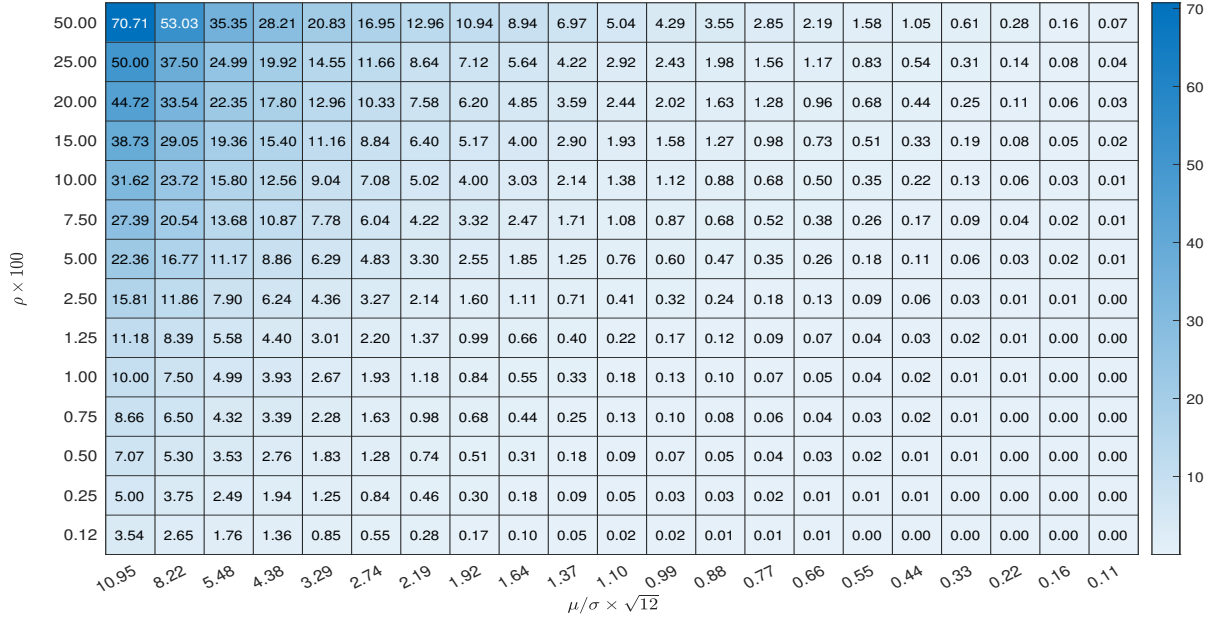


Figure 1: Optimal Sharpe Ratios (S^{OPT}) of Feasible Arbitrage Portfolios

Note: The figure reports optimal Sharpe ratios 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, $S^* = \sqrt{\alpha^\top (\Sigma_u)^{-1} \alpha} = \mu/\sigma \sqrt{\rho N}$, as shown by Figure 2. The pattern we see from Figure 2 agrees with theoretical predictions of Corollary 3. 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^{OPT}/S^* 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^* and S^{OPT} widens. In almost all empirically relevant scenarios, S^* is much larger than the feasible Sharpe ratio than arbitrageurs can actually earn with an optimal strategy.

2.5 Constructing the Optimal Arbitrage Portfolio

In our previous discussion, Theorem 2 establishes that the optimal Sharpe ratio attainable by any feasible strategy is capped by $S(\mathcal{G})$, which, by Corollary 2, is approximately equal to S^{OPT} under additional conditions. Following this, Corollary 3 illustrates the dependence of S^{OPT} on the unknown parameters in the DGP of alpha, i.e., (5). In addition, (17) and Theorem 3 characterize the strategy \tilde{w} that achieves the optimal Sharpe ratio attainable by any feasible strategy.

However, this characterization assumes access to the information set \mathcal{G} , which includes

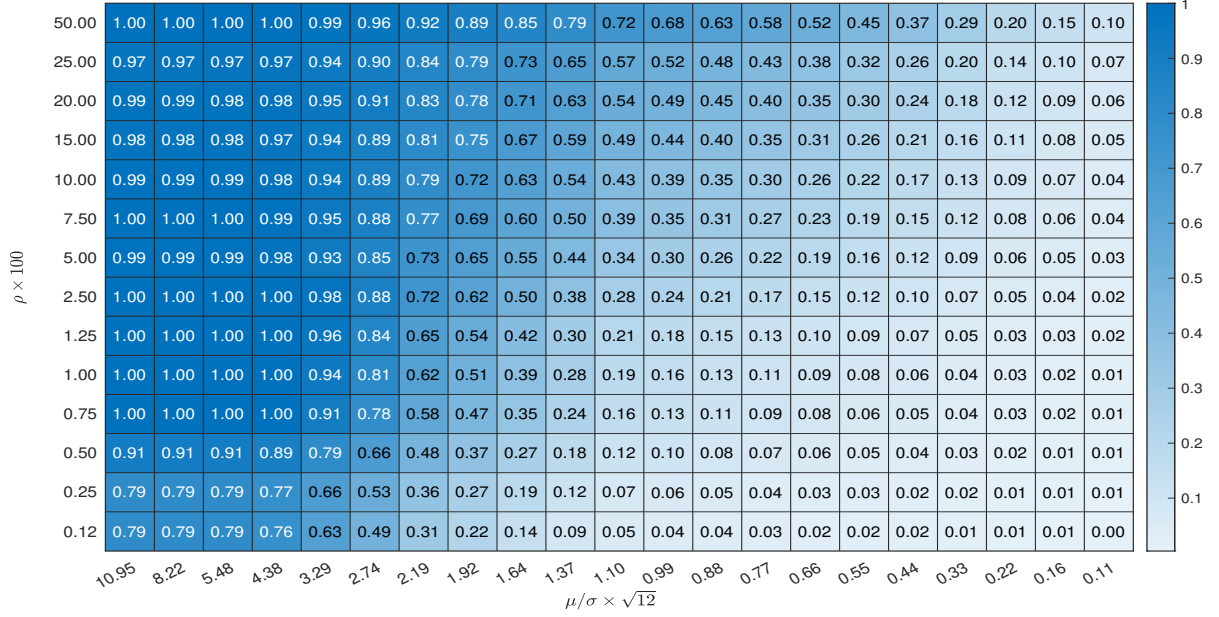


Figure 2: Ratios between S^{OPT} and S^*

Note: The figure reports the ratios of optimal Sharpe ratios between feasible and infeasible arbitrage portfolios. 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 $\mu/\sigma \times \sqrt{12}$.

not only past returns, r_t , but also potentially unknown variables, such as β , v_t , and Σ_u . Furthermore, it relies on knowledge of γ and the probability distribution function of \check{s} , i.e., $p(a)$. As a result, while the Sharpe ratio bound remains a valid upper ceiling—since having less information generally leads to lower achievable Sharpe ratios—the strategy \tilde{w} is infeasible. Its reliance on Σ_u and \tilde{s} , which in turn depend on ex-factor returns, $\bar{r}_i - \beta_i(\gamma + \bar{v})$, and the probability distribution of \check{s} via (18), makes its implementation impractical. Therefore, arbitrageurs encounter additional challenges in this context. They must conduct statistical inference concerning unknown parameters and variables to shape their investment strategies. Moreover, the devised optimal strategy must be flexible and adapt to the nuances of different DGPs.

Despite these challenges, arbitrageurs can indeed construct a uniformly optimal strategy that attains S^{OPT} across a broad spectrum of DGPs, even without access to all information in \mathcal{G} and perfect knowledge of γ and the probability distribution of \check{s} . In fact, this strategy relies only on observable data, making it entirely feasible. Specifically, we formulate this strategy within a framework where factors remain latent, but factor exposures are observable—a scenario that aligns with our empirical analysis.

We describe this portfolio strategy as “all weather,” signifying its adaptability to all considered DGP scenarios. Moreover, the fact that this strategy achieves S^{OPT} implies that

the feasible Sharpe ratio upper bound we derive is sharp. We outline the following algorithm to construct this strategy:

Algorithm 1 (Optimal Arbitrage Portfolio via Empirical Bayes).

Inputs: r_t , $t \in \mathcal{T} = \{t - T + 1, \dots, t\}$ and β .

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

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

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

$$\hat{p}(a) = \frac{1}{N k_N} \sum_i \phi_{1/T}\left(\frac{\hat{s}_i - a}{k_N}\right).$$

S3. Estimate \tilde{s} by Tweedie's formula (18) and isotonic regression:

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

S4. Construct the arbitrage portfolio weights as $\hat{w}^{\text{OPT}} \propto \mathbb{M}_{\beta} \hat{\Sigma}_u^{-1/2} \hat{\psi}$.

Outputs: \hat{w}^{OPT} .

Step S1 of Algorithm 1 provides feasible estimates of $\hat{\alpha}$ and $\hat{\Sigma}_u = \text{Diag}(\hat{\sigma}_i^2)$, which in turn leads to the sufficient statistic, \hat{s} . The motivation behind Steps S2 and S3 stems from Tweedie's formula (18); 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 $\hat{p}(a)$. To enhance the nonparametric estimator's performance in finite samples, enforcing monotonicity on $\psi(\cdot)$ 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, \hat{w}^{OPT} , following (17).

An essential step towards achieving optimality involves aggregating information from assets with comparable \hat{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 following theorem demonstrates the optimality of \widehat{w}^{OPT} :

Theorem 4. *Let \mathbb{P} denote the collection of all data-generating processes under which the same assumptions outlined in Corollary 2 hold. Moreover, we assume $\|\beta\|_{\text{MAX}} \lesssim_{\mathbb{P}} 1$, $\lambda_{\min}(\beta^\top \beta) \gtrsim_{\mathbb{P}} N$, and that the distribution of s_i is symmetric. In addition, suppose that $N^d \lesssim T \lesssim N^{d'}$ for fixed constants $d > 1/2$ and $d' < 1$. We denote the Sharpe ratio achieved by \widehat{w}^{OPT} as $\widehat{S}^{\text{OPT}} := E(r_{t+1}^\top \widehat{w}^{\text{OPT}} | \mathcal{F}_t) / \text{Var}(r_{t+1}^\top \widehat{w}^{\text{OPT}} | \mathcal{F}_t)^{1/2}$. Then it holds that \widehat{w}^{OPT} achieves, asymptotically, the upper bound S^{OPT} , uniformly over all sequences of data-generating processes. That is, for any $\epsilon > 0$,*

$$\lim_{N, T \rightarrow \infty} \sup_{\mathbb{P} \in \mathbb{P}} P(|\widehat{S}^{\text{OPT}} - S^{\text{OPT}}| \geq \epsilon S^{\text{OPT}} + \epsilon) = 0.$$

Theorem 4 imposes a mild assumption on $\lambda_{\min}(\beta^\top \beta)$, which requires that all factors are pervasive.²⁷ 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, s_i 's. Instead, we estimate these signals via cross-sectional regressions. This step results in our estimators, \widehat{s}_i 's, being inevitably polluted by estimation errors. The additional conditions imposed by Theorem 4 ensure that such estimation errors become asymptotically negligible.

Theorem 4 concludes that in the context of a linear factor model, arbitrageurs can construct this strategy, without any knowledge besides past returns and risk exposures (beta), to achieve the maximal Sharpe ratio over all feasible trading strategies that have zero exposure to factor risks. This Sharpe ratio precisely characterizes the limit of feasible arbitrages in economic terms. 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 supported; alphas are expected to be small and rare, a result of the competitive nature of arbitrage activities, in contrast to the consistent presence of idiosyncratic risks. 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 of β , the problem of learning about v_t and γ is negligible as well.

²⁷See, 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.

The term $\epsilon S^{\text{OPT}} + \epsilon$ accommodates both small and large values of S^{OPT} . If $S^{\text{OPT}} \lesssim 1$, then ϵ dominates and the estimation error inside the probability is characterized by the absolute difference between \widehat{S}^{OPT} and S^{OPT} . Otherwise, if $S^{\text{OPT}} \rightarrow \infty$, the estimation error is described in percentage terms. This is necessary because we simultaneously consider a large class of models where the Sharpe ratios may become extremely high or approach zero.

With Theorem 4, we establish the necessity for the no near-arbitrage condition given by (16).

Corollary 4. *Suppose the same assumptions as in Theorem 4 hold. The portfolio weights \widehat{w}^{OPT} yield a near-arbitrage strategy under any sequences of data-generating processes for which condition (16) does not hold.*

The ability of arbitrageurs to construct an optimal strategy that realizes S^{OPT} carries important economic implications. 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}^{OPT} can be interpreted as an empirical estimate of this equilibrium compensation, which we seek to pin down empirically.

2.6 Estimating Optimal Infeasible Sharpe Ratio

We are also interested in estimating the optimal infeasible Sharpe ratio, S^* . 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^* (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 S^{OPT} tackles this issue, providing an economically more meaningful evaluation of the APT.

To provide a benchmark for \widehat{S}^{OPT} , we propose an estimator for S^* inspired by its sample analog:

$$\widetilde{S}^* = \left(\bar{r}^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{r} \right)^{1/2}. \quad (21)$$

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:

$$\widehat{S}^* = \left(\bar{r}^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{r} - N/T \right)^{1/2}. \quad (22)$$

The next proposition summarizes the asymptotic properties of both estimators.

Proposition 1. *Suppose that the same assumptions outlined in Corollary 2 hold. Moreover, we assume $\|\beta\|_{\text{MAX}} \lesssim_P 1$ and $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$. We also impose $T \lesssim N$ and $T^{-1}N^{1/2} \log N \leq c_N$. Then we have*

$$\begin{aligned} \left| \hat{S}^* - S^* \right| / (1 + S^*) &= o_P \left(T^{-1/2} N^{1/4} \sqrt{\log N} \right), \\ \left| \tilde{S}^* - \left((S^*)^2 + NT^{-1} \right)^{1/2} \right| / (1 + S^*) &= o_P \left(T^{-1} N^{1/2} \log N \right). \end{aligned}$$

Similar to Theorem 4, the estimation error is relative when S^* dominates 1.0 asymptotically, and in absolute terms if S^* is dominated by 1.0.²⁸ This accommodates a large class of models, some of which have an exploding or a shrinking S^* . While it is possible to estimate S^* , it is not possible to build a portfolio that realizes it, unless the signal-to-noise ratio is sufficiently large such that $S^* = S^{\text{OPT}}$. Empirically, the difference between \hat{S}^* and \hat{S}^{OPT} thereby informs us about the economic consequences of learning.

2.7 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.

2.7.1 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 \hat{w}^{CSR} :

$$\hat{w}^{\text{CSR}} \propto \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \hat{\alpha}. \quad (23)$$

This is effectively a sample analogue of the approximated infeasible strategy given by (14). 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 \hat{w}^{CSR} , and S^{OPT} across a spectrum of DGPs. According to Proposition B2 in the online appendix, S^{CSR}

²⁸Obviously, the threshold 1.0 can be replaced by any fixed constant.

tends to be dominated by S^{OPT} in scenarios when alpha signals are sparse ($\sqrt{\rho T}\mu/\sigma$ is not 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.

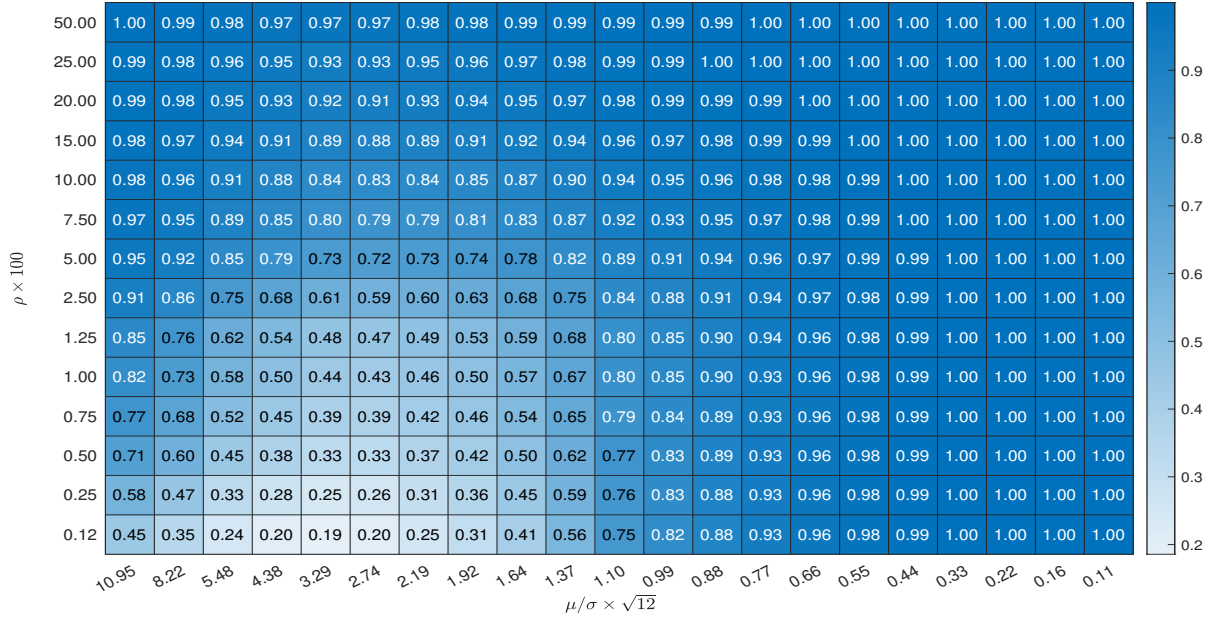


Figure 3: Ratios between S^{CSR} and S^{OPT}

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 \(1995\)](#).

2.7.2 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 , $\sqrt{T}\hat{s}_i$, for $i = 1, 2, \dots, N$. 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 $p_{(1)} \leq \dots \leq p_{(N)}$. Identify a critical index, \hat{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 \hat{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 $p_{(\hat{k})}$) 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:

$$\hat{w}^{\text{BH}}(\tau) \propto \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \hat{\alpha}^{\text{BH}}(\tau), \text{ where } \hat{\alpha}_i^{\text{BH}}(\tau) = \hat{\alpha}_i \mathbb{1}_{\{p_i \leq p_{(\hat{k})}\}}. \quad (24)$$

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.

2.7.3 Shrinkage Approaches

In the approximated infeasible strategy given by (14), 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.²⁹ Multiplying the ex-factor return weights by \mathbb{M}_β 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:

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

²⁹Regularizing those weights amounts to incorporating priors onto the alpha estimates.

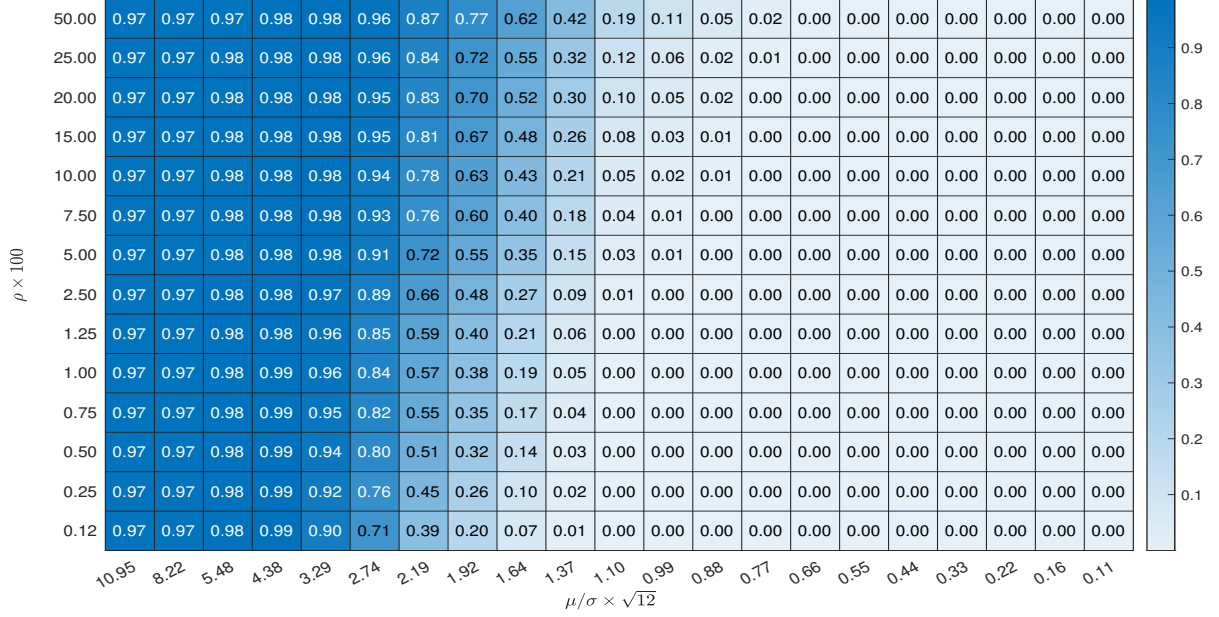


Figure 4: Ratios between S^{BH} and S^{OPT}

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 $\mu/\sigma \times \sqrt{12}$.

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

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

This leads to the optimal portfolio weight on r_t :³⁰

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

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

³⁰An 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.

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 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 weak, ridge (and hence CSR) approach optimality, and Lasso can replicate their performance by setting the tuning 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}_{β} 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

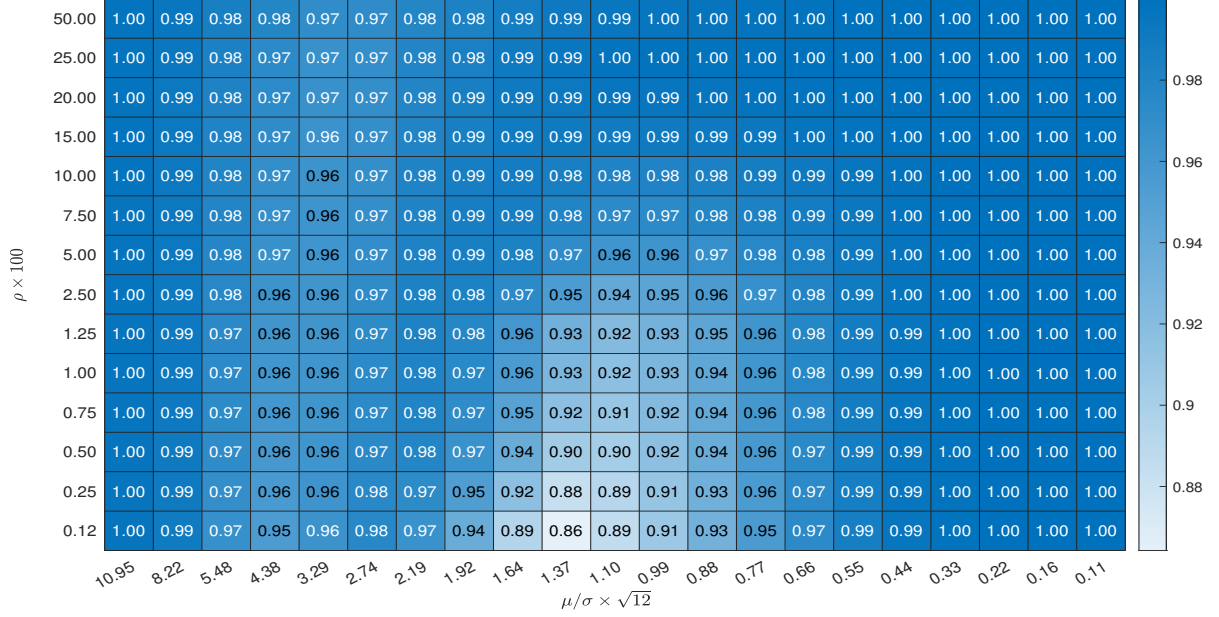


Figure 5: Ratios between $S_{\lambda}^{\text{Lasso}}$ and S^{OPT}

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 $S_{\lambda}^{\text{Lasso}}$ 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 $S_{\lambda}^{\text{Lasso}}$.

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.

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, \hat{w}^A , where A represents OPT, CSR, BH, or Lasso, and compute the corresponding theoretical Sharpe ratio: $\hat{S}^A = \hat{w}^{A\top} \mu / \sqrt{\hat{w}^{A\top} \Sigma_u^{-1} \hat{w}^A}$.

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.³¹

In light of Theorem 4, the following performance measure is sensible:

³¹The specific grid we use for tuning our method consists of the set $\{0.25, 0.5, \dots, 4\} \times (\log N)^{-1}$. For the Lasso method, the parameter λ is selected from the set $\{2^{-10}, 2^{-9}, \dots, 2^{-1}\}$. For our method, theoretical guidance from Theorem 4 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.

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

where the explicit dependence of \widehat{S}^A and S^{OPT} on μ/σ and ρ is not specified, and $\widehat{\mathbb{E}}(\cdot)$ 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., $o_P(1)$), the error is gauged in absolute terms. This approach ensures meaningful error measurement 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:

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

where $\widehat{\text{Var}}(\cdot)$ 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 $|S^A - S^{\text{OPT}}| / (1 + S^{\text{OPT}})$, 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 $|\widehat{S}^* - S^*| / (1 + S^*)$ 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^* is large or moderate (significantly greater than 1). Conversely, near the bottom right corner of Figure 6, where S^* is essentially zero as shown by Figures 1 and 2, the estimation proves more challenging. In this scenario, the error, becoming absolute ($S^* \ll 1$), 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

	$N = 1,000$				$N = 3,000$			
	$T = 2$	$T = 5$	$T = 10$	$T = 20$	$T = 2$	$T = 5$	$T = 10$	$T = 20$
	$\sup_{\mu/\sigma, \rho} \text{Gap}^A(\mu/\sigma, \rho)$							
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
	$\sup_{\mu/\sigma, \rho} \text{RMSE}^A(\mu/\sigma, \rho)$							
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 $\sup_{\mu/\sigma, \rho} \text{Gap}^A(\mu/\sigma, \rho)$, 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 $\sup_{\mu/\sigma, \rho} \text{RMSE}^A(\mu/\sigma, \rho)$. 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.

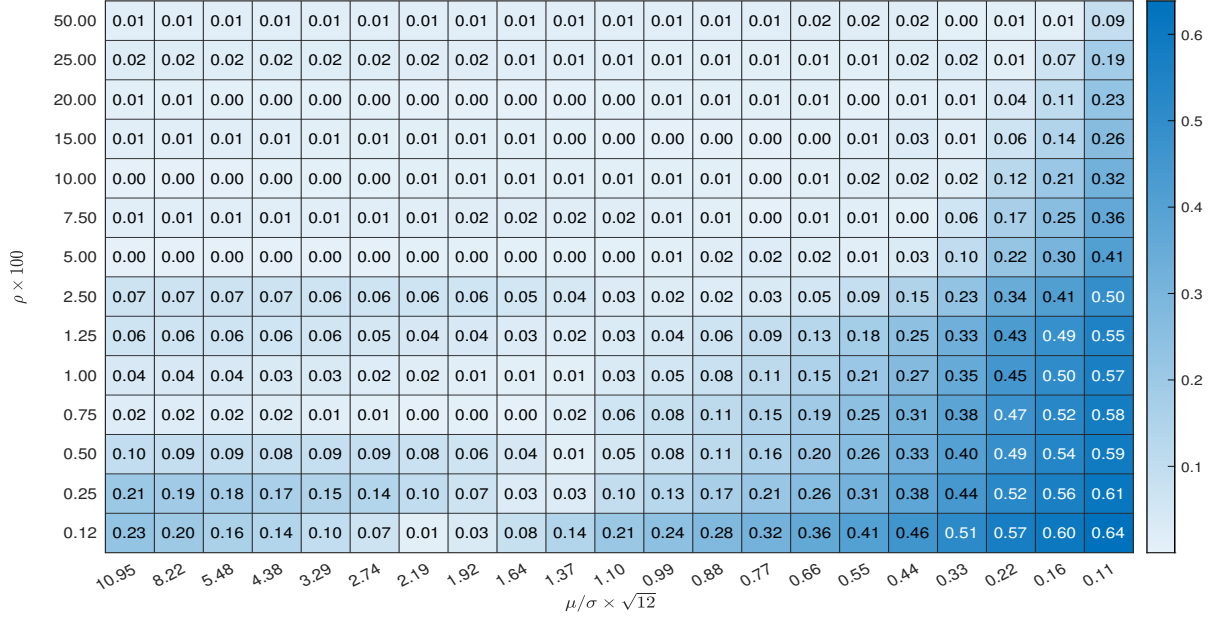


Figure 6: Comparison between \hat{S}^* and S^*

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

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

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.³²

Figure 7 plots the time series of the cross-sectional regression R^2 s 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^2 s 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.

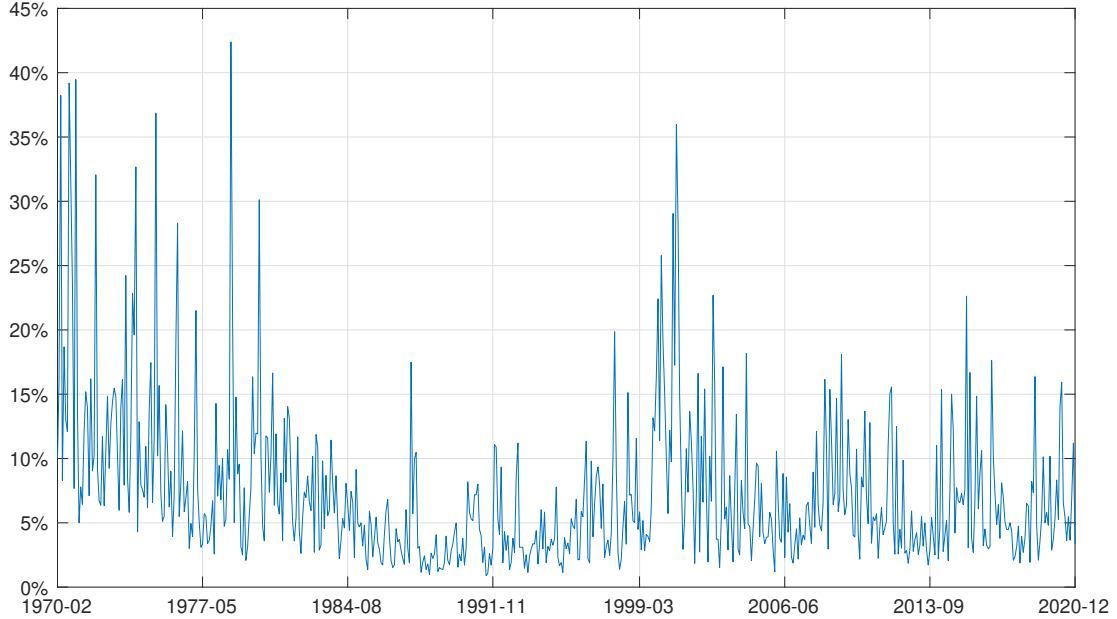


Figure 7: Time-series of the Cross-sectional R^2 s

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

³²For 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.

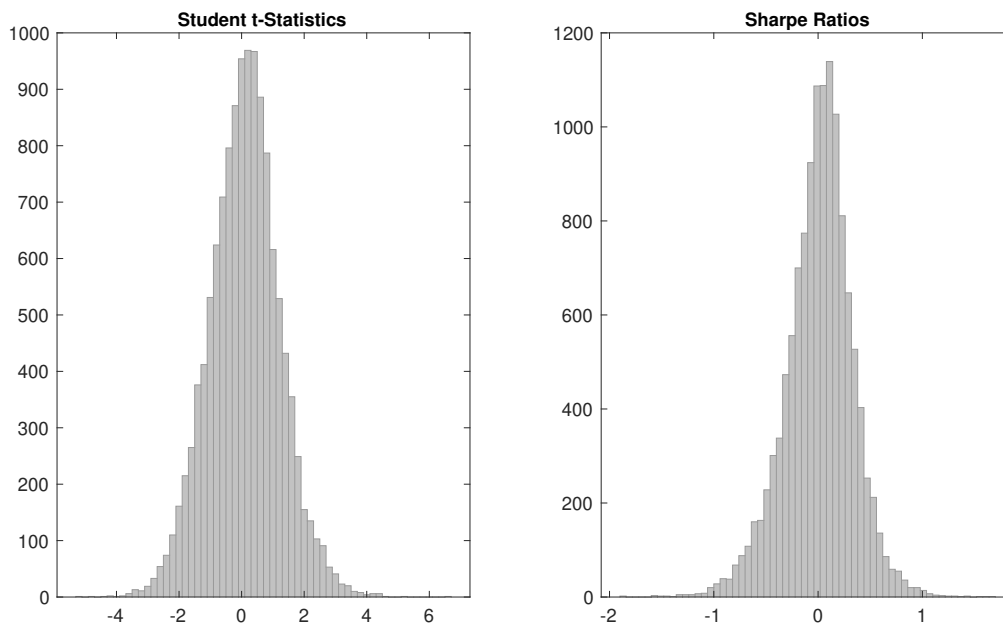


Figure 8: Histograms of the t-Statistics and Sharpe Ratios of Estimated Alphas

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 2.7. 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.³³ 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 real market. 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 we lived in a market where its DGP makes 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 perceived or infeasible Sharpe ratios using (21) within the same 60-month rolling window. Our estimate, \tilde{S}^* , averages 4.81 (with negative estimates truncated

³³Empirical results based on 36- and 84-month windows are reported in Table A2 in the appendix for robustness checks.

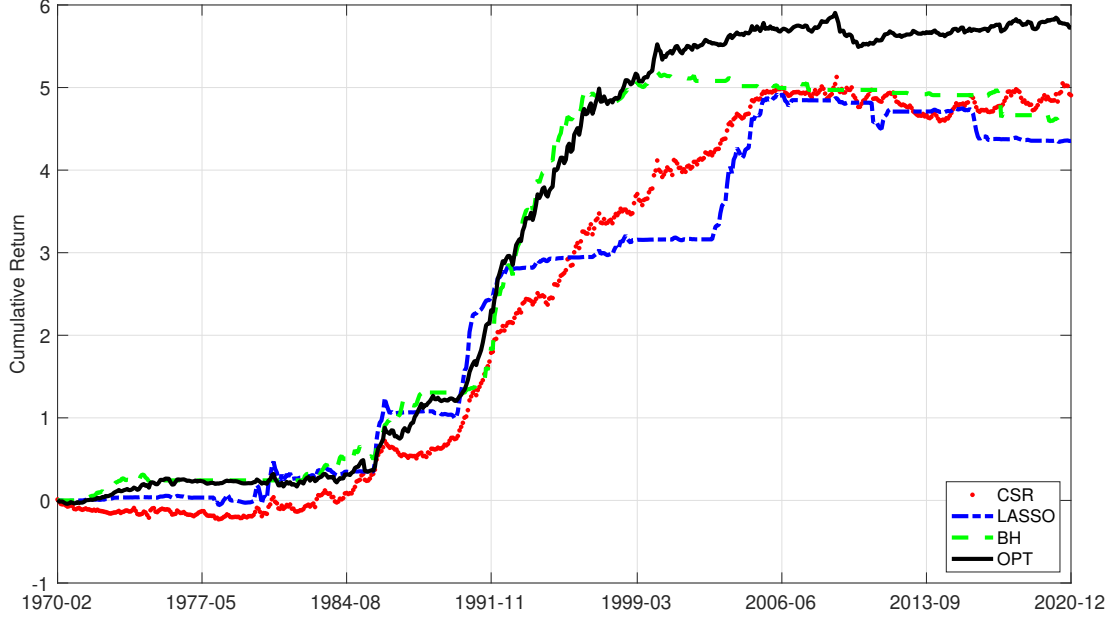


Figure 9: Normalized Cumulative Returns of Arbitrage Portfolios

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.

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.³⁴ 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 the residuals, which serve as estimates for α 's. It is important to note that these loadings

³⁴Results 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.

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,³⁵ accounting for an average of approximately 48% of the cross-section variation in average returns across all rolling windows, with 10 factors.³⁶ 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 \(2006\)](#), gross profits to total assets analyzed by [Novy-Marx \(2013\)](#), and industry return of big firms

³⁵This 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.

³⁶Results for models with varying numbers of factors are included in Tables A1 and A2 in the appendix for comparison.

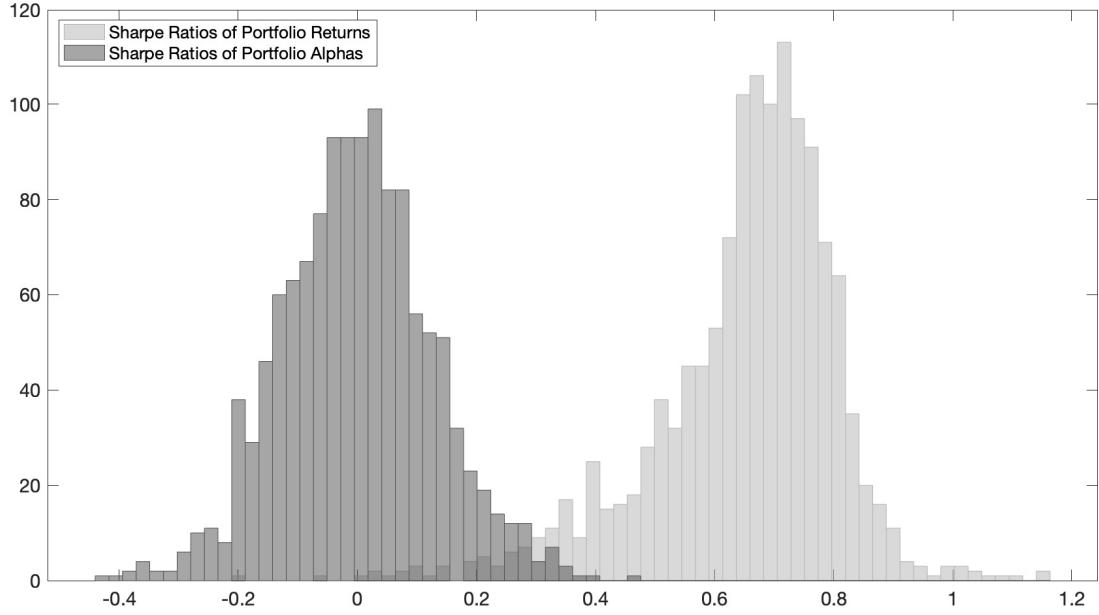


Figure 10: Sharpe Ratios of Portfolio Returns and Alphas

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

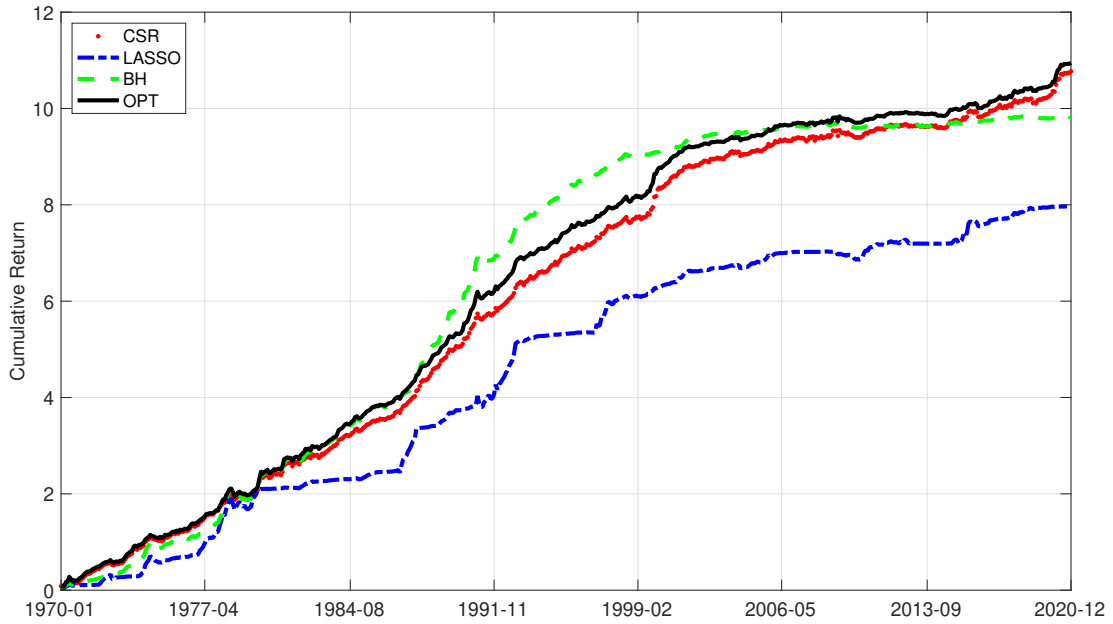


Figure 11: Normalized Cumulative Returns of Arbitrage Portfolios

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.

studied in [Hou \(2007\)](#). Conversely, the characteristics with the most negative weights in-

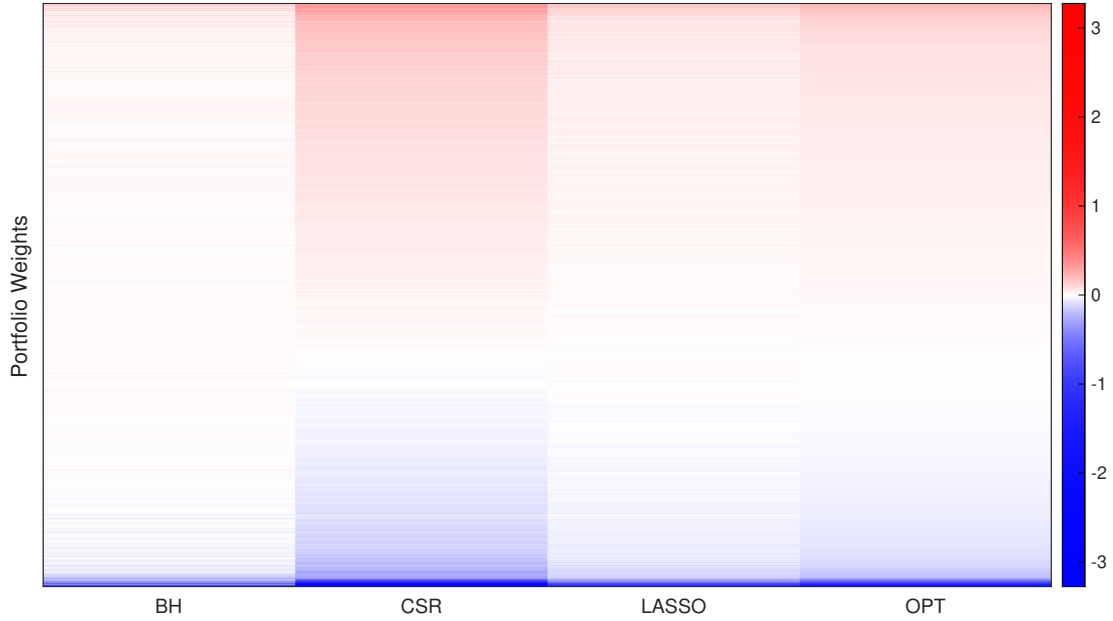


Figure 12: Average Weights Assigned to Portfolios by Different Methods

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.

clude 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.

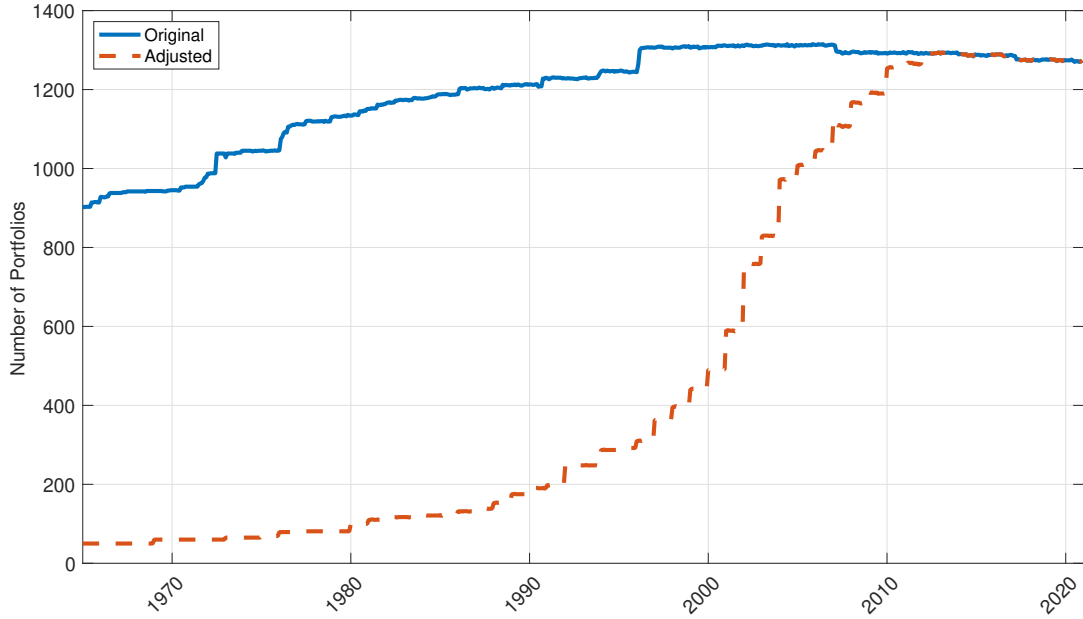


Figure 13: Number of Portfolios Before and After Adjusting for Publication Years

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.

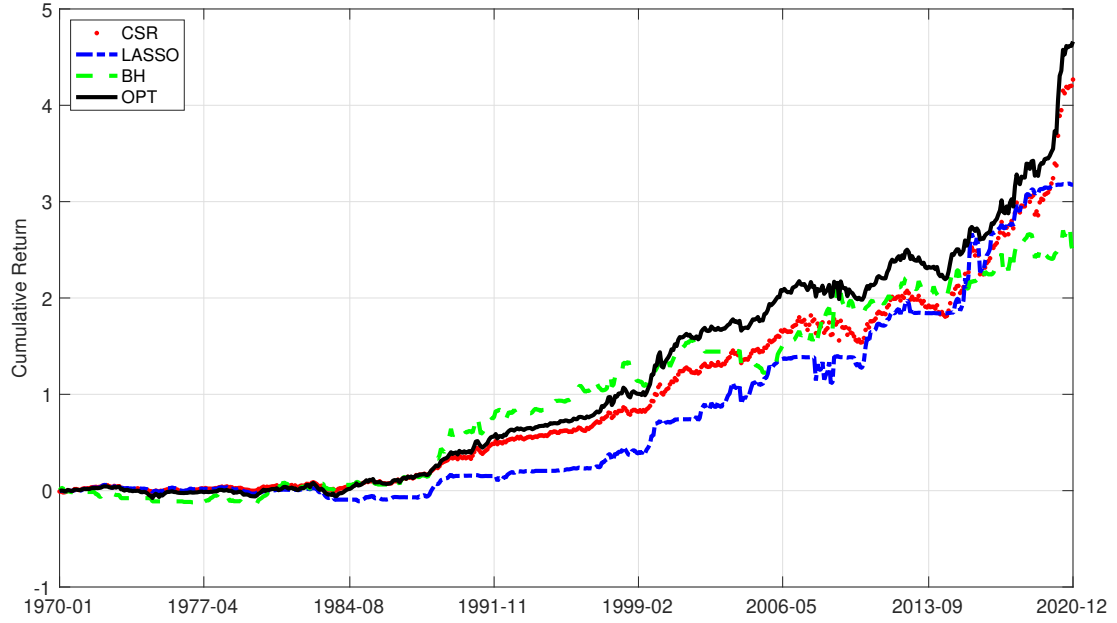


Figure 14: Normalized Cumulative Returns of Arbitrage Portfolios

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.

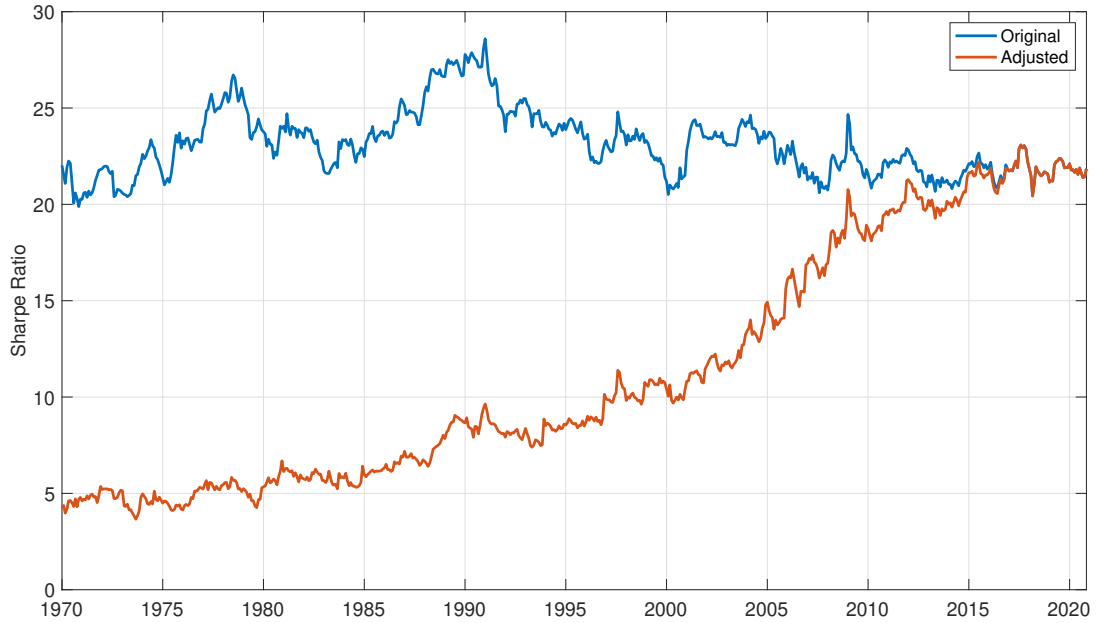


Figure 15: Time Series Plots of Infeasible Sharpe Ratios

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 DGP parameters, 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 can attain 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 the specific context of arbitrage pricing theory. In an equilibrium model, this upper bound in turn would be tied to the 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.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.
- 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*.
- Chen, X., L. P. Hansen, and P. G. Hansen (2021a). Robust identification of investor beliefs. *Proceedings of the National Academy of Sciences* 117(52), 33130–33140.
- Chen, X., L. P. Hansen, and P. G. Hansen (2021b). Robust inference for moment condition models without rational expectations. *Journal of Econometrics*, *forthcoming*.
- 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. Kelly, 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.
- Gromb, D. and D. Vayanos (2010). Limits of arbitrage. *Annual Review of Financial Economics* 2, 251–275.
- Guijarro-Ordóñez, J., M. Pelger, and G. Zanolini (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. *Econometric Theory* 21(1), 21–59.
- Martin, I. and S. Nagel (2021). Market efficiency in the age of big data. *Journal of Financial Economics*, *forthcoming*.
- 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

This version: October 13, 2024

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^*) corresponding to the portfolio alphas, and the distributions of portfolio alphas' t -statistics.

T	# of Factors	R^2	S^*	$ t\text{-stat} > 2$	$ t\text{-stat} > 3$	R^2	S^*	$ t\text{-stat} > 2$	$ t\text{-stat} > 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^*), 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^2 s 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^2 s 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^*) 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 Portfolios

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

¹The 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 2. Moreover, we assume that (a) $\|\beta\|_{\text{MAX}} \lesssim_{\text{P}} 1$ and $\lambda_{\min}(\beta^\top \beta) \gtrsim_{\text{P}} N$; (b) v_t is i.i.d. across t , $\mathbb{E}(v_t) = 0$, 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,*

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

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 (23), as \hat{S}^{CSR} . Then it satisfies $\hat{S}^{\text{CSR}} - S^{\text{CSR}} = o_{\text{P}}(1)$, 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 $c_N \rightarrow 0$, 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 (24) denoted by \hat{S}_τ^{BH} , satisfies $\hat{S}_\tau^{\text{BH}} = S_\tau^{\text{BH}} + o_{\text{P}}(1 + S^{\text{OPT}})$, where²*

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

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

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

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

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

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

such that, as $N, T \rightarrow \infty$, $S_\tau^{\text{BH}} > (1 - \epsilon)S^{\text{OPT}}$, 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 $\hat{w}_q(\lambda)$, denoted as $\hat{S}_{q,\lambda}$ for $q = 1, 2$, satisfies $\hat{S}_{1,\lambda} - S_\lambda^{\text{Lasso}} = o_P(1 + S^{\text{OPT}})$ and $\hat{S}_{2,\lambda} - S^{\text{CSR}} = o_P(1 + S^{\text{OPT}})$, where

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

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

Furthermore, if S^{OPT} does not vanish, then $S_\lambda^{\text{Lasso}} = S^{\text{OPT}}(1 + o(1))$, under some deterministic non-negative sequence of λ , if and only if, for some $c_N \rightarrow 0$, 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_λ^{Lasso} approaches S^{OPT} if and only if λ satisfies $T^{1/2}\lambda \rightarrow 0$. When $\sqrt{T}\mu/\sigma - \sqrt{-2\log\rho} \geq c_N^{-1}$, S_λ^{Lasso} approaches S^{OPT} if and only if λ satisfies $\sqrt{T}(\mu/\sigma - \lambda) \rightarrow \infty$ and $\frac{\phi(\sqrt{T}\lambda)}{\rho(1+T\lambda^2)T(\mu/\sigma - \lambda)^2} \rightarrow 0$.

Appendix C Mathematical Proofs

C.1 Proof of Theorem 1

Since β is \mathcal{G} -measurable, we can write, for all w satisfying $w^\top\beta = 0$,

$$w^\top r_{t+1} = w^\top(\alpha + u_{t+1}).$$

Because $E(u_{t+1}|\mathcal{F}_t) = 0$ according to the specification of (1) and \mathcal{G} is a subset of \mathcal{F}_t , we have $E(u_{t+1}|\mathcal{G}) = 0$. Moreover, since α is \mathcal{F} -measurable, it holds that $E(u_{t+1}\alpha^\top|\mathcal{G}) = E(E(u_{t+1}\alpha^\top|\mathcal{F})|\mathcal{G}) = 0$, i.e., u_{t+1} and α are \mathcal{G} -conditionally uncorrelated. Therefore, $\text{Var}(\alpha + u_{t+1}|\mathcal{G}) = \text{Var}(\alpha|\mathcal{G}) + \text{Var}(u_{t+1}|\mathcal{G})$. On the other hand, since $E(u_{t+1}|\mathcal{G}) = 0$, we have $\text{Var}(u_{t+1}|\mathcal{G}) = E(u_{t+1}u_{t+1}^\top|\mathcal{G}) = E(E(u_{t+1}u_{t+1}^\top|\mathcal{F})|\mathcal{G}) = E(\Sigma_u|\mathcal{G})$. Finally, we obtain $\text{Var}(\alpha + u_{t+1}|\mathcal{G}) = \tilde{\Sigma}_u$. Then, it holds that, for all $w \in \mathcal{G}$ satisfying $w^\top\beta = 0$,

$$U(w) = E(w^\top r_{t+1}|\mathcal{G}) - \frac{\kappa}{2}\text{Var}(r_{t+1}|\mathcal{G}) = w^\top\tilde{\alpha} - \frac{\kappa}{2}w^\top\tilde{\Sigma}_u w. \quad (\text{C.3})$$

We first find w that maximize $w^\top\tilde{\alpha} - \frac{\kappa}{2}w^\top\tilde{\Sigma}_u w$ only under the constraint $w^\top\beta = 0$. Next, we will verify that the solution is unique and is indeed \mathcal{G} -measurable, which would prove the theorem. The first step is equivalent with maximizing

$$w^\top \tilde{\alpha} - \frac{\kappa}{2} w^\top \tilde{\Sigma}_u w + \lambda w^\top \beta,$$

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

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

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

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

C.2 Proof of Theorem 2

To simplify the notation, we omit the dependence of β , Σ on N , and w on N and T . All limits are taken as $N \rightarrow \infty$. The derivation applies to either fixed T or $T \rightarrow \infty$ together with 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 \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 $\mathcal{E} = \mathbb{E}(w^\top r_{t+1} | \mathcal{F}_t) - \mathbb{E}(w^\top r_{t+1} | \mathcal{G})$. By the definition of $S(\hat{w})$, we have

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

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

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

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

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

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

$$|\mathcal{E}|/\text{Var}(w^\top r_{t+1}|\mathcal{F}_t)^{1/2} \lesssim_P |\mathcal{E}|/\|w\| = o_P(1). \quad (\text{C.7})$$

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

$$S(w) = w^\top \mathbb{E}(r_{t+1}|\mathcal{G})(w^\top \Sigma w)^{-1/2} + o_P(1). \quad (\text{C.8})$$

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

$$S(w) = w^\top \tilde{\alpha}(w^\top \Sigma_u w)^{-1/2} + o_P(1). \quad (\text{C.9})$$

Applying Cauchy-Schwarz inequality, we obtain

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

which, combined with (C.9), proves the inequality in (11).

We move on to the remaining results: the equality in (11), (12), and the optimal utility. They all hinge on the property of \tilde{w} . As a first step, we introduce short-hand notation

$$\check{w} := \frac{1}{\kappa} \Sigma_u^{-1/2} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2} \tilde{\alpha}, \quad \check{w}' := \frac{1}{\kappa} \mathbb{P}_\beta \Sigma_u^{-1} \tilde{\alpha}.$$

Given the specification of \mathcal{G} , we have $\tilde{\Sigma}_u = \Sigma_u$ and thereby

$$\tilde{w} = \frac{1}{\kappa} \Sigma_u^{-1} \tilde{\alpha} - \check{w}, \quad \tilde{w} - \frac{1}{\kappa} \mathbb{M}_\beta \Sigma_u^{-1} \tilde{\alpha} = \check{w}' - \check{w}. \quad (\text{C.10})$$

To establish the equality in (11), (12), and the optimal utility, we now prove that both $\|\check{w}\|$ and $\|\check{w}'\|$ are $o_P(1)$, which will then quickly leads to those results.

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

$$\mathbb{E}(\tilde{\alpha}|\beta, \Sigma_u) = \mathbb{E}(\alpha|\beta, \Sigma_u) = \mathbb{E}(\alpha|\Sigma_u) = 0. \quad (\text{C.11})$$

The first equality comes from that $\tilde{\alpha} := E(\alpha|\mathcal{G})$ 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), $\tilde{\alpha}_i$ 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,

$$\text{Cov}(\tilde{\alpha}_i, \tilde{\alpha}_j | \Sigma_u, \beta) = \text{Cov}(\tilde{\alpha}_i, \tilde{\alpha}_j | \Sigma_u) = \delta_{i,j} \text{Var}(\tilde{\alpha}_i | \Sigma_u). \quad (\text{C.12})$$

The second equality comes from $\text{Cov}(\tilde{\alpha}_i, \tilde{\alpha}_j | \Sigma_u) = 0$ 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

$$\text{Cov}(\tilde{\alpha} | \Sigma_u, \beta) \leq E(\|\alpha\|_{\text{MAX}}^2 | \Sigma_u) \mathbb{I}_N \quad (\text{C.13})$$

It hence holds that

$$\begin{aligned} E(\|\tilde{w}\|^2 | \Sigma_u, \beta) &= \text{Tr}(\Sigma_u^{-1/2} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2} \text{Cov}(\tilde{\alpha} | \Sigma_u, \beta) \Sigma_u^{-1/2} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2}) \\ &\leq E(\|\alpha\|_{\text{MAX}}^2 | \Sigma_u) \text{Tr}(\Sigma_u^{-1/2} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2}). \end{aligned} \quad (\text{C.14})$$

The equality comes from (C.11). The inequality comes from (C.12), (C.13), 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

$$\begin{aligned} \text{Tr}(\Sigma_u^{-1/2} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2}) &\lesssim_P \text{Tr}(\Sigma_u^{-1/2} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1/2}) \\ &= \text{Tr}(\mathbb{P}_{\Sigma_u^{-1/2} \beta} \Sigma_u^{-1} \mathbb{P}_{\Sigma_u^{-1/2} \beta}) \lesssim_P \text{Tr}(\mathbb{P}_{\Sigma_u^{-1/2} \beta}) = \text{rank}(\beta) = K. \end{aligned} \quad (\text{C.15})$$

Both inequalities come from $\Sigma_u \gtrsim_P \mathbb{I}_N$ by Assumption 1 (a). Substituting (C.15) into (C.14), we obtain

$$E(\|\tilde{w}\|^2 | \Sigma_u, \beta) \lesssim_P E(\|\alpha\|_{\text{MAX}}^2 | \Sigma_u) = o_P(1), \quad (\text{C.16})$$

where the last bound comes from Assumption 1 (a). Symmetrically, we have

$$\begin{aligned} E(\|\tilde{w}'\|^2 | \Sigma_u, \beta) &= \text{Tr}(\mathbb{P}_\beta \Sigma_u^{-1} \text{Cov}(\tilde{\alpha} | \Sigma_u, \beta) \Sigma_u^{-1} \mathbb{P}_\beta) \leq E(\|\alpha\|_{\text{MAX}}^2 | \Sigma_u) \text{Tr}(\mathbb{P}_\beta \Sigma_u^{-2} \mathbb{P}_\beta) \\ &\lesssim_P E(\|\alpha\|_{\text{MAX}}^2 | \Sigma_u) \text{Tr}(\mathbb{P}_\beta) = E(\|\alpha\|_{\text{MAX}}^2 | \Sigma_u) K = o_P(1). \end{aligned} \quad (\text{C.17})$$

Applying Chebyshev's inequality to (C.16) and (C.17), we obtain

$$\|\tilde{w}\| + \|\tilde{w}'\| = o_P(1). \quad (\text{C.18})$$

Combining (C.18) with the second part of (C.10), we establish (12) using the triangle inequality.

To prove the equality in (11), we note that, using the first part of (C.10),

$$\tilde{w}^\top \tilde{\alpha} = \frac{1}{\kappa} \tilde{\alpha}^\top \Sigma_u^{-1} \tilde{\alpha} - \check{w}^\top \tilde{\alpha}, \quad \tilde{w}^\top \Sigma_u \tilde{w} = \frac{1}{\kappa^2} \tilde{\alpha}^\top \Sigma_u^{-1} \tilde{\alpha} + \check{w}^\top \Sigma_u \check{w} - \frac{2}{\kappa} \check{w}^\top \tilde{\alpha}. \quad (\text{C.19})$$

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

$$\check{w}^\top \Sigma_u \check{w} \lesssim_P \|\check{w}\| = o_P(1), \quad |\check{w}^\top \tilde{\alpha}| \leq \sqrt{\check{w}^\top \Sigma_u \check{w}} S(\mathcal{G}) \lesssim_P \|\check{w}\| S(\mathcal{G}) = o_P(S(\mathcal{G})). \quad (\text{C.20})$$

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

$$\tilde{w}^\top \tilde{\alpha} = \frac{1}{\kappa} S(\mathcal{G})^2 + o_P(S(\mathcal{G})), \quad \tilde{w}^\top \Sigma_u \tilde{w} = \frac{1}{\kappa^2} S(\mathcal{G})^2 + o_P(1 + S(\mathcal{G})). \quad (\text{C.21})$$

Suppose $S(\mathcal{G}) \gtrsim_P 1$. Then, if $w = \tilde{w}$, the equality in (11) indeed follows from (C.21). Suppose $S(\mathcal{G}) = o_P(1)$. Then $S(\tilde{w}) \leq o_P(1)$ and $-S(\tilde{w}) = S(-\tilde{w}) \leq o_P(1)$ according to (the inequality part of) (11) as \tilde{w} is both \mathcal{G} -measurable and factor-neutral. In other words, we have $S(\tilde{w}) = o_P(1)$ and that the equality in (11) also holds. Applying the classic subsequence argument, we establish that the equality in (11) holds if $w = \tilde{w}$.

Finally, substituting (C.21) into (C.3), we obtain (recall $\tilde{\Sigma}_u = \Sigma_u$ here)

$$U(\tilde{w}) = \tilde{w}^\top \tilde{\alpha} - \frac{\kappa}{2} \tilde{w}^\top \Sigma_u \tilde{w} = \frac{1}{2\kappa} S(\mathcal{G})^2 + o_P(1 + S(\mathcal{G})).$$

Taking square root, we establish the optimal utility result.

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

Proof of Theorem 3. We have established in the beginning of the proof of Theorem 2 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 $\{r_{i,s}^* := \alpha_i + u_{i,s}, t - T + 1 \leq s \leq t\}$ given s_i and σ_i , denoted by $p(r_i^* | s_i, \sigma_i)$, is

$$p(r_i^* | s_i, \sigma_i) = \prod_{t-T+1 \leq s \leq t} \sigma_i^{-1} \phi(\sigma_i^{-1} r_{i,s}^* - s_i) = \phi(T^{1/2}(\check{s}_i - s_i)) f(r_i^*). \quad (\text{C.22})$$

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

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

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.22). The second equality in the last line comes from the Bayes' theorem and that $\phi(T^{1/2}(\tilde{s}_i - x))$ is, up to a constant, the density of \tilde{s}_i conditional on $s_i = x$. We hence establish the first statement in the theorem.

Next, we note that $\psi(a) = \mathbb{E}(s_i | \tilde{s}_i = a)$ and, conditional on $s_i = a$, $\tilde{s}_i \approx \mathcal{N}(a, T^{-1})$. The marginal density of \tilde{s}_i is then indeed $\mathbb{E}(\phi_{1/T}(a - s_i))$. Therefore, (18) directly comes from the Tweedie's formula, see, e.g., Robbins (1956). The proof ends. ■

Proof of Corollary 2. Apparently, $\tilde{s}_i = s_i + \sigma_i^{-1} \bar{u}_i$ is i.i.d. across i . By direction calculation and the definition of S^{OPT} in the statement of the corollary, we have

$$\mathbb{E}(\tilde{\alpha}^\top \Sigma_u^{-1} \tilde{\alpha}) = \sum_i \mathbb{E}(\mathbb{E}(s_i | \mathcal{G})^2) = N \int \psi(a)^2 p(a) da = (S^{\text{OPT}})^2. \quad (\text{C.23})$$

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

$$\begin{aligned}
&\mathbb{E}(|\mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2 - \mathbb{E}(s_i | \mathcal{G})^2|) \\
&\leq \mathbb{E}(\mathbb{E}(s_i \mathbb{1}_{\{|s_i| > c_N\}} | \mathcal{G})^2) + 2\mathbb{E}(|\mathbb{E}(s_i | \mathcal{G}) \mathbb{E}(s_i \mathbb{1}_{\{|s_i| > c_N\}} | \mathcal{G})|) \\
&\leq \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| > c_N\}}) + 2\sqrt{\mathbb{E}(\mathbb{E}(s_i | \mathcal{G})^2) \mathbb{E}(\mathbb{E}(s_i \mathbb{1}_{\{|s_i| > c_N\}} | \mathcal{G})^2)} \\
&\leq \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| > c_N\}}) + 2\sqrt{\mathbb{E}(\mathbb{E}(s_i | \mathcal{G})^2) \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| > c_N\}})} \\
&\leq c_N N^{-1} + \sqrt{\mathbb{E}(\mathbb{E}(s_i | \mathcal{G})^2) c_N N^{-1}},
\end{aligned} \quad (\text{C.24})$$

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

$$\mathbb{E} \left(\left| \sum_i \mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2 - \sum_i \mathbb{E}(s_i | \mathcal{G})^2 \right| \right)$$

$$\begin{aligned}
&\leq \sum_i \mathbb{E}(|\mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2 - \mathbb{E}(s_i | \mathcal{G})^2|) \\
&\leq c_N + \sqrt{\sum_i \mathbb{E}(\mathbb{E}(s_i | \mathcal{G})^2) c_N} = o(1 + S^{\text{OPT}}),
\end{aligned} \tag{C.25}$$

where the second inequality is a direct result of (C.24), and the last estimate is given by (C.23). From (C.25) and (C.23), it follows, respectively, using Markov's inequality and triangle inequality that

$$\sum_i \mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2 = \sum_i \mathbb{E}(s_i | \mathcal{G})^2 + o_P(1 + S^{\text{OPT}}), \tag{C.26}$$

$$\mathbb{E} \left(\sum_i \mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2 \right) = (S^{\text{OPT}})^2 + o(1 + S^{\text{OPT}}). \tag{C.27}$$

Further, we have

$$\begin{aligned}
\text{Var} \left(\sum_i \mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2 \right) &= \sum_i \text{Var}(\mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2) \\
&\leq c_N^2 \sum_i \mathbb{E}(\mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})^2) = o(1 + (S^{\text{OPT}})^2).
\end{aligned} \tag{C.28}$$

For the first line, we use that $\mathbb{E}(s_i \mathbb{1}_{\{|s_i| \leq c_N\}} | \mathcal{G})$ 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.27). Combining (C.27) and (C.28), we obtain

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

Along with (C.26), we obtain

$$\sum_i \mathbb{E}(s_i | \mathcal{G})^2 = (S^{\text{OPT}})^2 + o_P(1 + S^{\text{OPT}}).$$

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

$$((S^{\text{OPT}})^2 + o_P(1 + S^{\text{OPT}}))^{1/2} = S^{\text{OPT}} + o_P(1),$$

we conclude the proof. ■

Proof of Corollary 3. Because of the tail condition $\mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) \leq c_N N^{-1}$ for some sequence $c_N \rightarrow 0$ and that σ_i is constant across i , we have

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

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

$$\alpha^\top \alpha = \sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}} + o_P(1), \quad \mathbb{E} \left(\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}} \right) = \mu^2 \rho N. \quad (\text{C.29})$$

On the other hand, it holds that

$$\text{Var} \left(\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}} \right) \leq \sum_i \mathbb{E}(\alpha_i^4 \mathbb{1}_{\{|\alpha_i| < c_N\}}) \leq c_N^2 \sum_i \mathbb{E}(\alpha_i^2) = c_N^2 \mu^2 \rho N. \quad (\text{C.30})$$

Combining (C.29) and (C.30), we obtain

$$\alpha^\top \alpha = \mu^2 \rho N + o_P(1 + \mu \sqrt{\rho N}).$$

As a result, it holds that

$$S^* = \sigma^{-1} \sqrt{\alpha^\top \alpha} = \sigma^{-1} \mu (\rho N)^{1/2} + o_P(1). \quad (\text{C.31})$$

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

$$\psi(a) = \frac{\mu \rho \phi(a - T^{1/2} \mu / \sigma) - \mu \rho \phi(a + T^{1/2} \mu / \sigma)}{(2 - 2\rho) \phi(a) + \rho \phi(a - T^{1/2} \mu / \sigma) + \rho \phi(a + T^{1/2} \mu / \sigma)}, \quad (\text{C.32})$$

$$(S^{\text{OPT}})^2 = \frac{\mu \rho N}{2\sigma^2} \int \psi(a) (\phi(a - T^{1/2} \mu / \sigma) - \phi(a + T^{1/2} \mu / \sigma)) da. \quad (\text{C.33})$$

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

$$\begin{aligned} \sup_{a \geq C} \frac{\rho \phi(a)}{\phi(a - T^{1/2} \mu / \sigma)} &= \exp \left(\log \rho + T^{1/2} \mu \sigma^{-1} \left(\frac{1}{2} T^{1/2} \mu \sigma^{-1} - C \right) \right) \\ &\leq \exp \left(\log \rho + \frac{1}{2} \left(\sqrt{-2 \log \rho} + C \right) \left(\sqrt{-2 \log \rho} - C \right) \right) \leq 1 \end{aligned} \quad (\text{C.34})$$

On the other hand, in light of (C.32) and (C.33), we have

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

We hence obtain from (C.34) that, for N sufficiently large,

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

This proves the “if” part, given (C.31) and that $\mu^2 \rho N / \sigma^2$ does not vanish. Now suppose $T^{1/2} \mu \sigma^{-1} - \sqrt{-2 \log \rho} \rightarrow \infty$. Then, for all fixed $x > 0$, we have, for sufficiently large N ,

$$\begin{aligned} \sup_{a: |a| \leq x} \frac{\phi(a + T^{1/2} \mu / \sigma)}{\rho \phi(a)} &= \exp \left(-\log \rho - T^{1/2} \mu \sigma^{-1} \left(\frac{1}{2} T^{1/2} \mu \sigma^{-1} - x \right) \right) \\ &\leq \exp \left(-\log \rho - \frac{1}{2} \left(\sqrt{-2 \log \rho} + c_N^{-1} \right) \left(\sqrt{-2 \log \rho} + c_N^{-1} \right) \right) \\ &\leq \exp \left(-c_N^{-2} / 2 \right) \rightarrow 0, \end{aligned} \quad (\text{C.35})$$

$$\sup_{a: |a| \leq x} \frac{\phi(a + 2T^{1/2} \mu / \sigma)}{\phi(a)} = \exp \left(-2T^{1/2} \mu \sigma^{-1} (T^{1/2} \mu \sigma^{-1} - x) \right) \rightarrow 0. \quad (\text{C.36})$$

Given (C.32), it holds that

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

Substituting (C.36) into the numerator, and (C.35) and (C.36) into the denominator, we obtain that, for all fixed $x > 0$,

$$\sup_{a: |a| \leq x} \left| \mu^{-1} \psi(a + T^{1/2} \mu / \sigma) - 1 \right| \rightarrow 0. \quad (\text{C.37})$$

Since the integrand of (C.33) is always positive and even in a , it holds that, for all fixed $x > 0$,

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

Here the second inequality comes from (C.36), the third inequality is a result of (C.37), 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 4

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 $\psi := \Sigma_u^{-1/2} \mathbb{E}(\alpha | \mathcal{G}) = (\psi(\check{s}_1), \dots, \psi(\check{s}_N))^\top$, and the estimate $\check{\psi} := (\check{\psi}(\hat{s}_1), \dots, \check{\psi}(\hat{s}_N))^\top$, measured by L2 norm, is small compared to S^{OPT} . Based on the result, step 5 shows the gap between monotonicity-preserving estimate $\hat{\psi}$ and ψ in L2 norm is also small compared to S^{OPT} . This leads to that the Sharpe ratio generated by $\hat{w}^{\text{OPT}} = \mathbb{M}_\beta \hat{\Sigma}_u^{-1/2} \hat{\psi}$ converges to S^{OPT} , proved in the last step.

We note that, because of the rare and weak nature of alphas, $\mathbb{E}(\alpha_i | \mathcal{G})$ 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 $\hat{\psi}$ to converge to zero.

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

$$\check{s}_i = T^{-1} \sum_{s \in \mathcal{T}} (s_i + \varepsilon_{i,s}), \quad \hat{s}_i = \hat{\alpha}_i / \hat{\sigma}_i, \quad p(a) = \mathbb{E}(\phi_{1/T}(a - s_i)), \quad \psi(a) = \mathbb{E}(s_i | \check{s}_i = a). \quad (\text{C.38})$$

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 $B = \{i \leq N : |\check{s}_i| \leq \tilde{k}_N T^{-1/2}\}$ to separate the two cases, where $\tilde{k}_N = k_N^{-2}$. Moreover, we set $\check{\psi}$ and ψ as the N -dimensional vectors with entries $\check{\psi}_i := \check{\psi}(\hat{s}_i)$ and $\psi_i := \psi(\check{s}_i)$. It holds that

$$\|\check{\psi} - \psi\|^2 \leq \sum_{i \in B} (\check{\psi}_i - \psi_i)^2 + \sum_{i \in B^c} (\check{\psi}_i - \psi_i)^2. \quad (\text{C.39})$$

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

$$\|\check{\psi} - \psi\|^2 = o_P(1 + (S^{\text{OPT}})^2). \quad (\text{C.40})$$

The last two steps prove optimality of our portfolio strategy based on the above result. We end this step by noting that Corollary 2 states

$$\|\psi\| = S(\mathcal{G}) = S^{\text{OPT}} + o_{\text{P}}(1). \quad (\text{C.41})$$

Step 2. This step control the magnitude of $\sum_{i \in B} (\check{\psi}_i - \psi_i)^2$ of (C.39). It does so by showing

$$\sum_{i \in B^c} (\psi_i - \check{s}_i)^2 = o_{\text{P}}(1 + (S^{\text{OPT}})^2) \quad \text{and} \quad \sum_{i \in B^c} (\check{\psi}_i - \check{s}_i)^2 = o_{\text{P}}(1 + (S^{\text{OPT}})^2). \quad (\text{C.42})$$

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

$$\psi(a) - a = T^{-1} \frac{p'(a)}{p(a)}. \quad (\text{C.43})$$

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

$$\begin{aligned} |p'(a)| &\leq T \int |x - a| \phi_{1/T}(a - x) p_s(x) dx \\ &\leq b_N T \int_{|x-a| \leq b_N} \phi_{1/T}(a - x) p_s(x) dx + \sup_{x: |x-a| > b_N} T |x - a| \phi_{1/T}(a - x) \\ &\leq b_N T p(a) + \sup_{y: |y| > b_N} T^{3/2} |y| \exp(-T y^2 / 2). \end{aligned} \quad (\text{C.44})$$

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 $b_N = o(T^{-1/2} \tilde{k}_N)$, which is always possible, we obtain, for all a ,

$$|p'(a)| \leq c_N T^{1/2} \tilde{k}_N p(a) + c_N T N^{-2}. \quad (\text{C.45})$$

It hence holds that

$$\max_i \frac{|p'(\check{s}_i)|}{p(\check{s}_i)} \lesssim_{\text{P}} \sup_a \frac{|p'(a)|}{p(a)} \mathbb{1}_{\{p(a) \geq T^{1/2} N^{-3/2}\}} \leq c_N T^{1/2} \tilde{k}_N. \quad (\text{C.46})$$

The first inequality comes from (D.13) of Lemma D3. The second directly follows from (C.45). Combining (C.46) and (C.43), we obtain

$$\text{P}((\check{s}_i - \psi(\check{s}_i))^2 \leq c_N T^{-1} \tilde{k}_N^2, \forall i \leq N) \geq 1 - c_N. \quad (\text{C.47})$$

As a result,

$$\sum_{i \in B^c} (\check{s}_i - \psi(\check{s}_i))^2 \lesssim_{\text{P}} c_N T^{-1} \tilde{k}_N^2 |B^c| \leq c_N \sum_{i \in B^c} \check{s}_i^2 \lesssim_{\text{P}} c_N \sum_{i \in B^c} \psi(\check{s}_i)^2. \quad (\text{C.48})$$

Here the first inequality is simply (C.47), the second holds since $\check{s}_i^2 \geq T^{-1}\tilde{k}_N^2$ for all $i \in B^c$ by definition, and the last inequality is a direct implication of the first two. Given (C.48), we obtain the first part of (C.42) by noting $\sum_{i \in B^c} \psi(\check{s}_i)^2 \lesssim_P (S^{\text{OPT}})^2 + 1$ due to (C.41).

Now we establish the second part of (C.42). By construction we have

$$\check{\psi}(a) - a = \frac{1 + k_N^2}{T} \frac{\hat{p}'(a)}{\hat{p}(a)}, \quad \text{with} \quad \hat{p}(a) = \frac{1}{Nk_N} \sum_i \phi_{1/T} \left(\frac{\hat{s}_i - a}{k_N} \right). \quad (\text{C.49})$$

Similar to (C.44), we have, for all positive sequence b_N and all a ,

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

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

$$|\hat{p}'(a)| \leq c_N T^{1/2} \tilde{k}_N \hat{p}(a) + c_N T^{1/2} N^{-2}. \quad (\text{C.50})$$

Therefore, it holds that

$$\max_i \frac{|\hat{p}'(\hat{s}_i)|}{\hat{p}(\hat{s}_i)} \leq c_N T^{1/2} \tilde{k}_N, \quad (\text{C.51})$$

which comes from (C.50) and that $\hat{p}(\hat{s}_i) \geq \frac{1}{Nk_N} \phi_{1/T}(0) \gtrsim \frac{\sqrt{T}}{Nk_N}$ for all i . As a result, we obtain the second part of (C.42):

$$\sum_{i \in B^c} (\check{\psi}_i - \check{s}_i)^2 \leq c_N T^{-1} |B^c| \tilde{k}_N^2 + |B^c| \max_{i \leq N} |\hat{s}_i - \check{s}_i|^2 \leq c_N T^{-1} |B^c| \tilde{k}_N^2 \lesssim_P c_N (S^{\text{OPT}})^2 + c_N.$$

Here the first inequality is simply substituting (C.51) into (C.49), the second inequality comes from $\max_{i \leq N} |\hat{s}_i - \check{s}_i| \leq c_N T^{-1/2} \tilde{k}_N$ by Lemma D2, the last inequality holds by (C.41) and (the last two inequalities of) (C.48).

Step 3. To analyze $\sum_{i \in B} (\hat{\psi}_i - \psi_i)^2$ of (C.39), we introduce an auxiliary function:

$$\bar{\psi}(a) = \frac{\int x \phi_{v^2/T}(a - x) p_s(x) dx}{\int \phi_{v^2/T}(a - x) p_s(x) dx}, \quad \text{with} \quad v := \sqrt{1 + k_N^2}. \quad (\text{C.52})$$

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

$$\sum_{i \in B} (\psi_i - \bar{\psi}(\check{s}_i))^2 = o_P(1 + (S^{\text{OPT}})^2) \quad \text{and} \quad \sum_{i \in B} (\check{\psi}_i - \bar{\psi}(\check{s}_i))^2 = o_P(1 + (S^{\text{OPT}})^2). \quad (\text{C.53})$$

Then the triangle inequality would give us the desired bound on $\sum_{i \in B} (\check{\psi}_i - \psi_i)^2$. 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.52), 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 ,

$$\begin{aligned} |\phi_{v^2}(y) - \phi(y)| &\leq \sup_{y: |y| \leq k_N^{-1}} |\phi_{v^2}(y) - \phi(y)| + \sup_{y: |y| > k_N^{-1}} |\phi_{v^2}(y) + \phi(y)| \\ &\leq c_N k_N^{-1} \phi(y) \sup_{y: |y| \leq k_N^{-1}} |y/v - y| + c_N N^{-2} \leq c_N \phi(y) + c_N N^{-2}. \end{aligned} \quad (\text{C.54})$$

Here we use (D.17) of Lemma D4 (choose $j = 0$) and that $|v^{-1} - 1| \approx k_N^2$. We note $\phi_{1/T}(y) = \sqrt{T} \phi(\sqrt{T}y)$ (and symmetrically for $\phi_{v^2/T}(y)$). Hence, using (C.54), we directly obtain that, for all a ,

$$\begin{aligned} |\bar{p}(a) - p(a)| &\leq \int |\phi_{v^2/T}(a-x) - \phi_{1/T}(a-x)| p_s(x) dx \\ &\leq c_N \int \phi_{1/T}(a-x) p_s(x) dx + c_N \sqrt{T} N^{-2} = c_N p(a) + c_N \sqrt{T} N^{-2}. \end{aligned} \quad (\text{C.55})$$

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

$$\pi(a) = \int_0^\infty x \bar{\phi}(|a|, x) p_s(x) dx, \quad \text{and} \quad \bar{\pi}(a) = \int_0^\infty x \bar{\phi}(a/v, x/v) p_s(x) dx, \quad (\text{C.56})$$

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.54) and direct calculations that, for all $a \geq 0$ and $x \geq 0$,

$$|\bar{\phi}(a/v, x/v) - \bar{\phi}(a, x)| \leq c_N \bar{\phi}(a, x) + c_N \sqrt{T} N^{-2}. \quad (\text{C.57})$$

Substituting (C.57) into (C.56), we obtain, for all $a \geq 0$,

$$|\bar{\pi}(a) - \pi(a)| \leq c_N |\pi(a)| + c_N \sqrt{T} N^{-2} \int_0^\infty x p_s(s) dx \leq |\pi(a)| + c_N \sqrt{T} N^{-2}. \quad (\text{C.58})$$

Here the last inequality holds by $\mathbb{E}(|s|) \leq \sqrt{\mathbb{E}(s^2)} \leq c_N$ 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 an even function of x , (C.58) apparently holds for all a .

To establish from (C.55) and (C.58) that $\bar{\psi}(a)$ and $\psi(a)$ are close, we set $A := \{a : |a| \leq \tilde{k}_N T^{-1/2}, p(a) \geq \sqrt{T} N^{-3/2}\}$. Then we obtain that, for all $a \in A$,

$$\begin{aligned} |\bar{\psi}(a) - \psi(a)| &= \left| \frac{\bar{\pi}(a)}{\bar{p}(a)} - \frac{\pi(a)}{p(a)} \right| + \left| \frac{\pi(a)}{\bar{p}(a)} - \frac{\pi(a)}{p(a)} \right| \\ &\leq (1 + c_N) \frac{|\bar{\pi}(a) - \pi(a)|}{p(a)} + c_N \frac{|\pi(a)|}{p(a)} \leq c_N \frac{N^{-2}}{p(a)} + c_N \psi(a). \end{aligned} \quad (\text{C.59})$$

Here the first equality is obvious, the first inequality comes from the lower bound of $p(a)$ (by the definition of A) and (C.55), the second inequality is a result of (C.58). From (C.59), it follows that, for all a satisfying $a \in A$,

$$|\bar{\psi}(a) - \psi(a)|^2 \leq c_N \frac{N^{-2}}{p(a)} + c_N \psi(a)^2. \quad (\text{C.60})$$

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

$$N \int_A |\bar{\psi}(a) - \psi(a)|^2 p(a) da \leq c_N + c_N N \int_{-\infty}^\infty \psi(a)^2 p(a) da \leq c_N + c_N (S^{\text{OPT}})^2, \quad (\text{C.61})$$

which comes from (C.60) and that $\int_A da \leq 2\tilde{k}_N$. Therefore, using Chebyshev's inequality and comparing the definitions of sets A and B , we obtain

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

where the last inequality holds by (C.61). Given (D.13) of Lemma D3, we obtain the first part of (C.53).

Step 4. This step proves the second part of (C.53), i.e., we bound $\sum_{i \in B} (\check{\psi}_i - \bar{\psi}(\tilde{s}_i))^2$. We introduce $\tilde{p}(a)$ and $\tilde{\psi}(a)$ that mimic $\hat{p}(a)$ and $\hat{\psi}(a)$ by replacing the data input \tilde{s}_i with \hat{s}_i :

$$\tilde{p}(a) = \frac{1}{Nk_N} \sum_i \phi_{1/T} \left(\frac{\check{s}_i - a}{k_N} \right), \quad \text{and} \quad \tilde{\psi}(a) = a + \frac{v^2}{T} \frac{\tilde{p}'(a)}{\tilde{p}(a)}. \quad (\text{C.62})$$

Then we can decompose the quantity of interest:

$$\sum_{i \in B} (\check{\psi}_i - \bar{\psi}(\check{s}_i))^2 \leq \sum_{i \in B} (\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2 + \sum_{i \in B} (\check{\psi}_i(\hat{s}_i) - \tilde{\psi}(\check{s}_i))^2. \quad (\text{C.63})$$

We first show that $\sum_{i \in B} (\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2$ is small. Since we have $\tilde{p}(\check{s}_i) \geq \frac{1}{Nk_N} \phi_{1/T}(0) \gtrsim \frac{\sqrt{T}}{Nk_N}$ for all i , symmetric to the derivation of (C.51), we have

$$\max_i \frac{\tilde{p}'(\check{s}_i)}{\tilde{p}(\check{s}_i)} \leq c_N T^{1/2} \tilde{k}_N. \quad (\text{C.64})$$

On the other hand, symmetric to the derivation of (C.46), we obtain

$$\max_i \frac{\bar{p}'(\check{s}_i)}{\bar{p}(\check{s}_i)} \lesssim_P \max_i \frac{\bar{p}'(a)}{\bar{p}(a)} \mathbb{1}_{\{p(a) \geq N^{-3/2}\}} \lesssim c_N T^{1/2} \tilde{k}_N. \quad (\text{C.65})$$

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.64) and (C.65) into the definitions of $\tilde{\psi}(a)$ and $\bar{\psi}(a)$ ((C.62) and (C.52)), we obtain

$$\max_i |\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i)| \lesssim_P c_N \tilde{k}_N T^{-1/2}. \quad (\text{C.66})$$

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 $d > 0$,

$$\mathbb{E} \left(\sum_i T (\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2 \right) \lesssim N^d.$$

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 $\max_{i \leq N} \sqrt{T} \bar{\psi}(\check{s}_i) = o(N^d)$ for every $d > 0$. In the absence of this additional condition, a weaker result holds: for every $d > d' > 0$,

$$\mathbb{E} \left(\sum_i \min \{ T (\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2, N^{d'} \} \mathbb{1}_{\{\sqrt{T} |\check{s}_i| \leq N^{d'}, p(\check{s}_i) \geq \sqrt{T} N^{d'-1}\}} \right) \lesssim N^d. \quad (\text{C.67})$$

(C.67) turns out sufficient for establishing a desired bound on $\sum_{i \in B} (\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2$, which we demonstrate now. Then we have, for every $d > d' > 0$,

$$\begin{aligned}
& \sum_{i \in B} T(\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2 \lesssim_P \sum_{i \in B} \min\{T(\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2, N^{d'}\} \\
& \lesssim_P N^d + \sum_{i \in B} \min\{T(\tilde{\psi}(\check{s}_i) - \bar{\psi}(\check{s}_i))^2, N^{d'}\} \mathbb{1}_{\{p(\check{z}_i) \geq \sqrt{T}N^{d'-1}\}} \lesssim N^d. \tag{C.68}
\end{aligned}$$

Here the first inequality comes from (C.66), the second inequality comes from that $\mathbb{E} \left(\sum_{i \in B} \mathbb{1}_{\{p(\check{s}_i) < \sqrt{T}N^{d'-1}\}} \right) \lesssim \tilde{k}_N N^{d'}$ (by the definition of set B), and the last is simply (C.67).

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

$$\max_{i \in B} |\hat{s}_i - \check{s}_i| \lesssim_P \chi_N := N^{-1/2}(\epsilon_N + \mathbb{E}(s_j^2)^{1/2}), \quad \text{with } \epsilon_N := k_N^5. \tag{C.69}$$

Since $\check{\psi}(\hat{s}_i)$ and $\tilde{\psi}(\check{s}_i)$ depends on $\{\hat{s}_j\}$ 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 $i \in B$,

$$\begin{aligned}
& |\hat{p}(\hat{s}_i) - \tilde{p}(\check{s}_i)| \\
& \leq \frac{1}{Nk_N} \sum_j \left| \phi_{1/T} \left(\frac{\hat{s}_j - \hat{s}_i}{k_N} \right) - \phi_{1/T} \left(\frac{\check{s}_j - \check{s}_i}{k_N} \right) \right| \lesssim_P \sqrt{T} \chi_N k_N^{-2} \tilde{p}(\check{s}_i) + \sqrt{T} N^{-2} k_N^{-1}, \tag{C.70} \\
& |\hat{p}'(\hat{s}_i) - \tilde{p}'(\check{s}_i)| \\
& \leq \frac{T}{Nk_N^3} \sum_j \left| (\hat{s}_j - \hat{s}_i) \phi_{1/T} \left(\frac{\hat{s}_j - \hat{s}_i}{k_N} \right) - (\check{s}_j - \check{s}_i) \phi_{1/T} \left(\frac{\check{s}_j - \check{s}_i}{k_N} \right) \right| \\
& \lesssim_P T \chi_N k_N^{-4} \tilde{p}(\check{z}_i) + T N^{-2} k_N^{-1}. \tag{C.71}
\end{aligned}$$

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.69) into (D.17) of Lemma D4. Since $\tilde{p}(\check{s}_i) \geq \frac{1}{Nk_N} \phi_{1/T}(0) \gtrsim \frac{\sqrt{T}}{Nk_N}$ by definition, we obtain from (C.70) and (C.71) that

$$\max_{i \in B} \frac{|\hat{p}(\hat{s}_i) - \tilde{p}(\check{s}_i)|}{\tilde{p}(\check{s}_i)} \lesssim_P \sqrt{T} \chi_N k_N^{-2} + N^{-1} \lesssim \sqrt{T} \chi_N k_N^{-2}, \tag{C.72}$$

$$\max_{i \in B} \frac{|\hat{p}'(\hat{s}_i) - \tilde{p}'(\check{s}_i)|}{\tilde{p}(\check{s}_i)} \lesssim_P T \chi_N k_N^{-4} + \sqrt{T} N^{-1} \lesssim T \chi_N k_N^{-4}. \tag{C.73}$$

Then we have

$$\begin{aligned}
& \max_{i \in B} \left| \frac{\hat{p}'(\hat{s}_i)}{\hat{p}(\hat{s}_i)} - \frac{\tilde{p}'(\check{s}_i)}{\tilde{p}(\check{s}_i)} \right| \\
& \leq \max_{i \in B} \frac{\tilde{p}(\check{s}_i)}{\hat{p}(\hat{s}_i)} \frac{|\hat{p}'(\hat{s}_i) - \tilde{p}'(\check{s}_i)|}{\tilde{p}(\check{s}_i)} + \max_{i \in B} \frac{\tilde{p}(\check{s}_i)}{\hat{p}(\hat{s}_i)} \frac{\tilde{p}'(\check{s}_i)}{\tilde{p}(\check{s}_i)} \frac{|\hat{p}(\hat{s}_i) - \tilde{p}(\check{s}_i)|}{\tilde{p}(\check{s}_i)} \lesssim_P T \chi_N k_N^{-4}. \tag{C.74}
\end{aligned}$$

The first inequality is direct algebra. Substituting (C.64), (C.72), and (C.73) into the right-hand-side of the first inequality, we obtain the second inequality. Combining (C.69) and (C.74) with the definitions of $\check{\psi}$ and $\tilde{\psi}$ ((C.49) and (C.62)), we obtain

$$\sum_{i \in B} (\check{\psi}(\hat{s}_i) - \tilde{\psi}(\check{s}_i))^2 \leq N \max_{i \in B} |\hat{s}_i - \check{s}_i|^2 + \frac{N}{T^2} \max_{i \in B} \left| \frac{\check{p}'(\hat{s}_i)}{\check{p}(\hat{s}_i)} - \frac{\check{p}'(\check{s}_i)}{\check{p}(\check{s}_i)} \right|^2 \lesssim_P k_N^{-8} (\epsilon_N^2 + E(s_j^2)). \quad (\text{C.75})$$

The goal is to show $\sum_{i \in B} (\check{\psi}(\hat{s}_i) - \tilde{\psi}(\check{s}_i))^2 = o_P(1 + (S^{\text{OPT}})^2)$, which is apparently true from (C.75) if $E(s_j^2) \leq \epsilon_N^2$. For the case $E(s_j^2) > \epsilon_N^2$, we observe

$$E(s_j^2) = E(s_j^2 \mathbb{1}_{\{\epsilon_N/2 < |s_i| \leq 1\}}) + E(s_j^2 \mathbb{1}_{\{|s_i| \leq \epsilon_N/2\}}) + E(s_j^2 \mathbb{1}_{\{|s_i| > 1\}}) \leq P(|s_i| > \epsilon_N/2) + \epsilon_N^2/4 + c_N N^{-1},$$

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

$$N \epsilon_N^4 \lesssim_P \sum_i \epsilon_N^2 \mathbb{1}_{\{|s_i| \geq \epsilon_N/2\}} \lesssim_P \sum_i \check{s}_i^2 \mathbb{1}_{\{|\check{s}_i| \geq \epsilon_N/4\}} \lesssim \sum_{i \in B^c} \check{s}_i^2 \lesssim_P 1 + (S^{\text{OPT}})^2. \quad (\text{C.76})$$

Here the second inequality comes from $\check{s}_i - s_i = \bar{\epsilon}_i$ and $\max_i |\bar{\epsilon}_i| \lesssim \sqrt{(\log N)/T}$ by the uniform bound on i.i.d normal variables. The third inequality holds by the definition of B , and the last inequality can be established from holds by (C.41) and (the last two inequalities of) (C.48). Since $E(s_j^2) \leq 1 + E(s_j^2 \mathbb{1}_{\{|s_i| > 1\}}) \lesssim 1$ by condition (a) of Assumption 2, it follows from (C.76) that $k_N^{-8} E(s_j^2) = o_P(1 + (S^{\text{OPT}})^2)$. Given (C.75), we prove $\sum_{i \in B} (\check{\psi}(\hat{s}_i) - \tilde{\psi}(\check{s}_i))^2 = o_P(1 + (S^{\text{OPT}})^2)$. Substituting this result and (C.68) into (C.63), we obtain $\sum_{i \in B} (\check{\psi}(\hat{s}_i) - \tilde{\psi}(\check{s}_i))^2 = o_P(1 + (S^{\text{OPT}})^2)$, i.e., the second part of (C.53). Substituting (C.42) and (C.53) into (C.39), we finally establish (C.40).

Step 5. The goal of this step is to establish

$$\|\hat{\psi} - \psi\| = o_P(1 + S^{\text{OPT}}). \quad (\text{C.77})$$

We start by defining

$$l_j = \arg \min_{i \leq N} |\check{s}_i - \hat{s}_j|. \quad (\text{C.78})$$

Suppose $\check{s}_{l_i} > \check{s}_{l_j}$ and $\hat{s}_i \leq \hat{s}_j$ for some (i, j) . According to (C.78), we have $|\check{s}_{l_i} - \hat{s}_i| \leq |\check{s}_{l_j} - \hat{s}_i|$ and $|\check{s}_{l_j} - \hat{s}_j| \leq |\check{s}_{l_i} - \hat{s}_j|$. The former would lead to $\check{s}_{l_i} + \check{s}_{l_j} < 2\hat{s}_i$, and the latter would lead to $\check{s}_{l_i} + \check{s}_{l_j} > 2\hat{s}_j$. They together contradict $\hat{s}_i \leq \hat{s}_j$. Hence we have $\check{s}_{l_i} \leq \check{s}_{l_j}$ if $\hat{s}_i \leq \hat{s}_j$ 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

$$\|\widehat{\psi} - \check{\psi}\| \leq \|\psi^* - \check{\psi}\| \leq \|\psi^* - \psi\| + o_P(1 + S^{\text{OPT}}). \quad (\text{C.79})$$

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.40) proved in steps 2 - 4.

Given (C.40) and (C.79), to obtain (C.77) we only need to bound $\|\psi^* - \psi\|$. For this purpose, we note that (C.78) leads to $|\check{s}_{l_i} - \widehat{s}_i| \leq |\check{s}_i - \widehat{s}_i|$ for all i . Applying Lemma D2, we obtain

$$\max_{1 \leq i \leq N} |\check{s}_{l_i} - \check{s}_i| \lesssim_P c_N \widetilde{k}_N T^{-1/2}, \quad \max_{i \in B} |\check{s}_{l_i} - \check{s}_i| \lesssim_P \chi_N, \quad (\text{C.80})$$

where $\chi_N = N^{-1/2}(\epsilon_N + E(s_j^2)^{1/2})$ with $\epsilon_N := k_N^5$. Similar to (C.48), we have

$$\sum_{i \in B^c} ((\check{s}_{l_i} - \psi(\check{s}_{l_i}))^2 + (\check{s}_{l_i} - \check{s}_i)^2) \lesssim_P c_N T^{-1} \widetilde{k}_N^2 |B^c| \leq c_N \sum_{i \in B^c} \psi(\check{s}_i)^2 \lesssim_P o_P(1 + (S^{\text{OPT}})^2). \quad (\text{C.81})$$

Here the first inequality comes from (C.47) and the first part of (C.80). The second inequality is already shown in (C.48). The last inequality is established right below (C.48). Combining (C.81) and the first part of (C.42), and applying the triangle inequality, we obtain

$$\sum_{i \in B^c} (\psi(\check{s}_{l_i}) - \psi(\check{s}_i))^2 = o_P(1 + (S^{\text{OPT}})^2). \quad (\text{C.82})$$

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

$$\begin{aligned} |p(\check{s}_{l_i}) - p(\check{s}_i)| &\leq \int |\phi_{1/T}(\check{s}_{l_i} - x) - \phi_{1/T}(\check{s}_i - x)| p_s(x) ds \\ &\lesssim (T \log N)^{1/2} |\check{s}_{l_i} - \check{s}_i| p(\check{s}_i) + c_N \sqrt{T} N^{-2} \\ &\lesssim_P (T \log N)^{1/2} \chi_N p(\check{s}_i) + c_N \sqrt{T} N^{-2}. \end{aligned} \quad (\text{C.83})$$

$$\begin{aligned} |p(\check{s}_{l_i})' - p(\check{s}_i)'| &\leq T \int |(\check{s}_{l_i} - x) \phi_{1/T}(\check{s}_{l_i} - x) - (\check{s}_i - x) \phi_{1/T}(\check{s}_i - x)| p_s(x) ds \\ &\lesssim (T \log N) |\check{s}_{l_i} - \check{s}_i| p(\check{s}_i) + c_N T N^{-2} \\ &\lesssim_P (T \log N) \chi_N p(\check{s}_i) + c_N T N^{-2}. \end{aligned} \quad (\text{C.84})$$

For both (C.83) and (C.84), 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.80). Then we have, uniformly over i for which $\check{s}_i \in A$ (A is defined above (C.59)),

$$\begin{aligned}
|\psi(\check{s}_{l_i}) - \psi(\check{s}_i)| &\leq |\check{s}_{l_i} - \check{s}_i| + \frac{1}{T} \frac{|p(\check{s}_{l_i})' - p(\check{s}_i)'|}{p(\check{s}_{l_i})} + \frac{|p(\check{s}_i)'|}{T} \left| \frac{1}{p(\check{s}_{l_i})} - \frac{1}{p(\check{s}_i)} \right| \\
&\lesssim_P \chi_N + \chi_N + \frac{\chi_N}{\sqrt{T}} \frac{p(\check{s}_i)'}{p(\check{s}_i)} \lesssim \chi_N \left(\tilde{k}_N + \sqrt{T} |\psi(\check{s}_i)| \right). \tag{C.85}
\end{aligned}$$

The first inequality is obvious. The second inequality comes from the second part of (C.80), (C.83), (C.84), 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.43) and $|\check{s}_i| \leq \tilde{k}_N$ as $\check{s}_i \in A$. Therefore, we have

$$\begin{aligned}
&\sum_{i \in B} (\psi(\check{s}_{l_i}) - \psi(\check{s}_i))^2 \mathbb{1}_{\{p(\check{s}_i) \geq \sqrt{T}N^{-3/2}\}} \\
&= \sum_{i: \check{s}_i \in A} (\psi(\check{s}_{l_i}) - \psi(\check{s}_i))^2 \lesssim_P N \chi_N^2 \tilde{k}_N^2 + T \chi_N^2 \sum_{i \in B} \psi(\check{z}_i)^2 \\
&\lesssim_P (\epsilon_N^2 + \mathbb{E}(s_j^2)) \tilde{k}_N^2 + o_P(1 + (S^{\text{OPT}})^2) = o_P(1 + (S^{\text{OPT}})^2). \tag{C.86}
\end{aligned}$$

The equality is obvious and the first inequality comes from (C.85). The second inequality comes from (C.41) and that $T \chi_N^2 = o(1)$ due to $\mathbb{E}(s_j^2) = o(1)$ and $T \lesssim N^{d'}$ for fixed $d' < 1$. The last inequality is established in the analysis after (C.75). Combining (C.86), Lemma D3, and (C.82), we establish $\|\psi^* - \psi\| = o_P(1 + S^{\text{OPT}})$. Substituting this into (C.79) and using (C.40) lead to (C.77).

Step 6. This step combines (C.77) with (C.41) to prove the theorem, i.e., that the Sharpe ratio of the strategy \hat{w}^{OPT} we construct achieves $(S^{\text{OPT}})^2$ asymptotically. Since Sharpe ratio is invariant to scaling of weights, we can treat \hat{w}^{OPT} as $\hat{w}^{\text{OPT}} = \mathbb{M}_\beta \check{w}$, where $\check{w} := \hat{\Sigma}_u^{-1/2} \hat{\psi}$.

Using condition (a) of Assumption 1 and (D.1) of Lemma D1, we have $\max_{i \leq N} |\hat{\sigma}_i / \sigma_i - 1| \lesssim_P c_N$. As a result, we obtain $\|\Sigma^{1/2} \check{w} - \hat{\psi}\| \lesssim_P c_N \|\hat{\psi}\|$. Then, it follows from (C.77) and (C.41) that

$$\|\Sigma_u^{1/2} \check{w} - \psi\| \leq \|\Sigma^{1/2} \check{w} - \hat{\psi}\| + \|\hat{\psi} - \psi\| \lesssim_P c_N \|\psi\| + \|\hat{\psi} - \psi\| = o_P(1 + S^{\text{OPT}}). \tag{C.87}$$

Hence we have

$$|(\check{w}^\top \Sigma_u^{1/2} - \psi^\top) \psi| \leq \|\Sigma_u^{1/2} \check{w} - \psi\| \|\psi\| = o_P(1 + (S^{\text{OPT}})^2), \tag{C.88}$$

$$|\check{w}^\top \Sigma_u \check{w} - \psi^\top \psi| \leq \|\Sigma_u^{1/2} \check{w} - \psi\|^2 + 2 \|\Sigma_u^{1/2} \check{w} - \psi\| \|\psi\| = o_P(1 + (S^{\text{OPT}})^2). \tag{C.89}$$

Here for both (C.88) and (C.89), the first inequalities come from Cauchy-Schwarz, whereas the last equalities come from (C.87) and (C.41). Further, substituting (C.41) into (C.88) and

(C.89), we obtain

$$\check{w}^\top \Sigma_u^{1/2} \psi = (S^{\text{OPT}})^2 + o_P(1 + (S^{\text{OPT}})^2), \quad \check{w}^\top \Sigma_u \check{w} = (S^{\text{OPT}})^2 + o_P(1 + (S^{\text{OPT}})^2). \quad (\text{C.90})$$

Next, we note that

$$\begin{aligned} \|\mathbb{P}_\beta \check{w}\| &\leq \|\mathbb{P}_\beta \Sigma_u^{-1/2} \psi\| + \|\mathbb{P}_\beta (\check{w} - \Sigma_u^{-1/2} \psi)\| \\ &= \|\mathbb{P}_\beta \Sigma_u^{-1} \tilde{\alpha}\| + (\check{w} - \Sigma_u^{-1/2} \psi)^\top \mathbb{P}_\beta (\check{w} - \Sigma_u^{-1/2} \psi) \\ &\leq \|\mathbb{P}_\beta \Sigma_u^{-1} \tilde{\alpha}\| + \|\check{w} - \Sigma_u^{-1/2} \psi\| = o_P(1 + S^{\text{OPT}}). \end{aligned} \quad (\text{C.91})$$

To obtain the last equality, we note that the first term can be bounded using (C.18) as $\check{w}' := \frac{1}{\kappa} \mathbb{P}_\beta \Sigma_u^{-1} \tilde{\alpha}$ and that the second term can be bounded using (C.87) and $1 \lesssim_P \lambda_{\min}(\Sigma_u)$. As a result, we obtain

$$\check{w}^\top \mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta \check{w} = \check{w}^\top \Sigma_u \check{w} + \check{w}^\top \mathbb{P}_\beta \Sigma_u \mathbb{P}_\beta \check{w} - 2\check{w}^\top \Sigma_u \mathbb{P}_\beta \check{w} = \check{w}^\top \Sigma_u \check{w} + o_P(1 + (S^{\text{OPT}})^2). \quad (\text{C.92})$$

For the last equality, we use (C.89), (C.91), and $\lambda_{\max}(\Sigma_u) \lesssim_P 1$ by condition (a) of Assumption 1. Similarly, we write

$$\check{w}^\top \mathbb{M}_\beta \tilde{\alpha} = \check{w}^\top \tilde{\alpha} - \check{w}^\top \mathbb{P}_\beta \tilde{\alpha} = \check{w}^\top \tilde{\alpha} + o_P(1 + (S^{\text{OPT}})^2), \quad (\text{C.93})$$

where for the last equality, we use (C.91) and $\|\tilde{\alpha}\|^2 \lesssim_P \tilde{\alpha}^\top \Sigma_u^{-1} \tilde{\alpha} = (S^{\text{OPT}})^2$.

We now conclude that, when S^{OPT} does not vanish,

$$\begin{aligned} \hat{S}^{\text{OPT}} &= \frac{(\hat{w}^{\text{OPT}})^\top \alpha}{\sqrt{(\hat{w}^{\text{OPT}})^\top \Sigma \hat{w}^{\text{OPT}}}} = \frac{\check{w}^\top \mathbb{M}_\beta \alpha}{\sqrt{\check{w}^\top \mathbb{M}_\beta \Sigma \mathbb{M}_\beta \check{w}}} = \frac{\check{w}^\top \mathbb{M}_\beta \tilde{\alpha}}{\sqrt{\check{w}^\top \mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta \check{w}}} + o_P(1) \\ &= \frac{\check{w}^\top \tilde{\alpha}}{\sqrt{\check{w}^\top \Sigma_u \check{w}}} + o_P(1 + S^{\text{OPT}}) = S^{\text{OPT}} + o_P(1 + S^{\text{OPT}}). \end{aligned} \quad (\text{C.94})$$

The first two equalities hold by definition. The third one comes (C.9) and that $\mathbb{M}_\beta \check{w}$ is \mathcal{G} -measurable and factor-neutral. The fourth one comes from (C.92), (C.93), and the second part of (C.90). The last equality comes from (C.90) ($\tilde{\alpha} = \Sigma_u^{1/2} \psi$). Because \hat{w}^{OPT} is \mathcal{G} -measurable and $\beta^\top \hat{w}^{\text{OPT}} = 0$, Theorem 2 and Corollary 2 apply. We hence have $\hat{S}^{\text{OPT}} \leq S(\mathcal{G}) + o_P(1) = S^{\text{OPT}} + o_P(1)$. Because $-\hat{S}^{\text{OPT}}$ is the Sharpe ratio generated by $-\hat{w}^{\text{OPT}}$, we also have $-\hat{S}^{\text{OPT}} \leq S^{\text{OPT}} + o_P(1)$. As a result, when S^{OPT} does vanish, we have $\hat{S}^{\text{OPT}} = o_P(1)$. Therefore, given (C.94) and using the subsequence argument (see, e.g., Andrews and Cheng (2012)), we have

$$\hat{S}^{\text{OPT}} = S^{\text{OPT}} + o_P(1 + S^{\text{OPT}}).$$

In other words, we have, for all $P \in \mathbb{P}$,

$$\lim_{N, T \rightarrow \infty} P(|\hat{S}^{\text{OPT}} - S^{\text{OPT}}| \geq \epsilon S^{\text{OPT}} + \epsilon) = 0. \quad (\text{C.95})$$

Suppose the theorem does not hold, then there is a sequence of data-generating processes P_k with $P_k \in \mathbb{P}$ for each $k \in \{1, 2, \dots\}$ such that

$$\limsup_{N, T \rightarrow \infty} \lim_{k \rightarrow \infty} P_k(|\hat{S}^{\text{OPT}} - S^{\text{OPT}}| \geq \epsilon S^{\text{OPT}} + \epsilon) > 0.$$

This contradicts (C.95), and the theorem is proved.

C.5 Proof of Proposition 1

By definition we have

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

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

$$\|\hat{\Sigma}_u - \Sigma_u\|_{\text{MAX}} \lesssim_P \sqrt{T^{-1} \log N}. \quad (\text{C.96})$$

As a result, noting $P(0 \leq \mathbb{M}_\beta \leq \mathbb{I}_N) \rightarrow 1$ and $P(\Sigma_u \approx \mathbb{I}_N) \rightarrow 1$ by the assumption $\|\beta\|_{\text{MAX}} \lesssim_P 1$ and $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$, and recalling $(S^*)^2 = \alpha^\top \Sigma_u^{-1} \alpha$, we have

$$|\alpha^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha - \alpha^\top \mathbb{M}_\beta \Sigma_u^{-1} \mathbb{M}_\beta \alpha| \lesssim_P \sqrt{T^{-1} \log N} (S^*)^2. \quad (\text{C.97})$$

On the other hand, it holds that

$$\begin{aligned} |\alpha^\top \mathbb{M}_\beta \Sigma_u^{-1} \mathbb{M}_\beta \alpha - \alpha^\top \Sigma_u^{-1} \alpha| &\leq \alpha^\top \mathbb{P}_\beta \Sigma_u^{-1} \mathbb{P}_\beta \alpha + 2\sqrt{(\alpha^\top \Sigma_u^{-1} \alpha)(\alpha^\top \mathbb{P}_\beta \Sigma_u^{-1} \mathbb{P}_\beta \alpha)} \\ &\lesssim_P \alpha^\top \mathbb{P}_\beta \alpha + \sqrt{(\alpha^\top \Sigma_u^{-1} \alpha)(\alpha^\top \mathbb{P}_\beta \alpha)} \\ &\lesssim_P N^{-1} \|\alpha^\top \beta\|^2 + \sqrt{N^{-1}(\alpha^\top \Sigma_u^{-1} \alpha) \|\alpha^\top \beta\|^2} \\ &\lesssim_P E(s_i^2) + S^* E(s_i^2)^{1/2}. \end{aligned} \quad (\text{C.98})$$

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

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

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

which, by Markov's inequality, leads to

$$\alpha^\top \Sigma_u^{-1} \alpha = \sum_i s_i^2 \mathbb{1}_{\{|s_i| < c_N\}} + o_P(1).$$

Moreover, it holds that

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

Using Chebyshev's inequality, we obtain

$$(S^*)^2 = \alpha^\top \Sigma_u^{-1} \alpha \geq \sum_i s_i^2 \mathbb{1}_{\{|s_i| < c_N\}} \gtrsim_P \sum_i \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| < c_N\}}) \geq N \mathbb{E}(s_i^2) + o(1).$$

Combining this result with (C.98) and (C.97), we have

$$\alpha^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha = (S^*)^2 + o_P \left(\sqrt{T^{-1} \log N} ((S^*)^2 + 1) \right). \quad (\text{C.99})$$

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

$$\alpha^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \lesssim \alpha^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha \sqrt{\bar{u}^\top \bar{u}} \lesssim \alpha^\top \alpha \sqrt{\bar{u}^\top \bar{u}} = O_P(((S^*)^2 + 1)T^{-1/2}). \quad (\text{C.100})$$

The first inequality comes from Cauchy-Schwarz. The second inequality holds because $\mathbb{P}(\mathbb{M}_\beta \approx \mathbb{I}_N) \rightarrow 1$, $\mathbb{M}_\beta^2 = \mathbb{M}_\beta$, and $\mathbb{P}(\widehat{\Sigma}_u \approx \mathbb{I}_N) \rightarrow 1$ due to $\mathbb{P}(\Sigma_u \approx \mathbb{I}_N) \rightarrow 1$ and (C.96). The third inequality holds by $\mathbb{P}(\widehat{\Sigma}_u \approx \mathbb{I}_N) \rightarrow 1$ as well.

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

$$\begin{aligned} N = \text{tr}(\widehat{\Sigma}_u^{-1} \widehat{\Sigma}_u) &= \sum_{i \leq N} (\widehat{\Sigma}_u^{-1})_{i,i} \left(T^{-1} \sum_{s \in T} (\mathbb{M}_\beta u_s)_i^2 - (\mathbb{M}_\beta \bar{u})_i^2 \right) \\ &= T^{-1} \sum_{s \in T} u_s^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_s - \bar{u}^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \\ &= T^{-1} \sum_{s \in T} u_s^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_s + O_P(N/T). \end{aligned} \quad (\text{C.101})$$

The last line comes from $\bar{u}^\top \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \lesssim_P \bar{u}^\top \bar{u}$ because of $\mathbb{M}_\beta^2 = \mathbb{M}_\beta$ and $\mathbb{P}(\widehat{\Sigma}_u \approx \mathbb{I}_N) \rightarrow 1$.

Furthermore, I have

$$\begin{aligned} \bar{u}^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - \bar{u}^\top \hat{\Sigma}_u^{-1} \bar{u} &\leq 2|\bar{u}^\top \hat{\Sigma}_u^{-1} \mathbb{P}_\beta \bar{u}| + \bar{u}^\top \mathbb{P}_\beta \hat{\Sigma}_u^{-1} \mathbb{P}_\beta \bar{u} \\ &\lesssim_P \sqrt{\bar{u}^\top \bar{u}} \sqrt{\bar{u}^\top \mathbb{P}_\beta \bar{u}} + \bar{u}^\top \mathbb{P}_\beta \bar{u} \lesssim_P N^{1/2}/T. \end{aligned} \quad (\text{C.102})$$

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

$$\begin{aligned} &T^{-2} \sum_{t \in T} (u_t^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta u_t - u_t^\top \hat{\Sigma}_u^{-1} u_t) \\ &= T^{-2} \sum_{t \in T} \left(2\sqrt{u_t^\top u_t} \sqrt{u_t^\top \mathbb{P}_\beta \hat{\Sigma}_u^{-1} \mathbb{P}_\beta u_t} + u_t^\top \mathbb{P}_\beta \hat{\Sigma}_u^{-1} \mathbb{P}_\beta u_t \right) \\ &\lesssim_P T^{-2} \sum_{t \in T} \left(\sqrt{u_t^\top u_t} \sqrt{u_t^\top \mathbb{P}_\beta u_t} + u_t^\top \mathbb{P}_\beta u_t \right) \lesssim_P \frac{N^{1/2}}{T}. \end{aligned} \quad (\text{C.103})$$

From (C.101), (C.102), and (C.103), it directly follows

$$\bar{u}^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1}N = \bar{u}^\top \hat{\Sigma}_u^{-1} \bar{u} - T^{-2} \sum_{s \in T} u_s^\top \hat{\Sigma}_u^{-1} u_s + O_P(N^{1/2}/T + N/T^2). \quad (\text{C.104})$$

On the other hand, we have

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

It then follows from (C.96) and $P(\Sigma_u \approx \mathbb{I}_N) \rightarrow 1$ that

$$\begin{aligned} \bar{u}^\top \hat{\Sigma}_u^{-1} \bar{u} &= -\bar{u}^\top \Sigma_u^{-2}(\hat{\Sigma}_u - 2\Sigma_u)\bar{u} + O_P(T^{-1}(\log N)\bar{u}^\top \bar{u}) \\ &= -\bar{u}^\top \Sigma_u^{-2}(\hat{\Sigma}_u - 2\Sigma_u)\bar{u} + O_P(T^{-2}N \log N). \end{aligned} \quad (\text{C.105})$$

Similarly, we have

$$T^{-2} \sum_{t \in T} u_t^\top \hat{\Sigma}_u^{-1} u_t = -T^{-2} \sum_{t \in T} u_t^\top \Sigma_u^{-2}(\hat{\Sigma}_u - 2\Sigma_u)u_t + O_P(T^{-2}N \log N). \quad (\text{C.106})$$

Substituting (C.105) and (C.106) into (C.104), we have

$$\begin{aligned} \bar{u}^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1}N &= -\bar{u}^\top \Sigma_u^{-2}(\hat{\Sigma}_u - 2\Sigma_u)\bar{u} + T^{-2} \sum_{t \in T} u_t^\top \Sigma_u^{-2}(\hat{\Sigma}_u - 2\Sigma_u)u_t \\ &\quad + O_P(T^{-1}N^{1/2} + T^{-2}N \log N). \end{aligned} \quad (\text{C.107})$$

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

$$\begin{aligned} (\widehat{\Sigma}_u)_{i,i} &= (T^{-1}uu^\top)_{i,i} + (\mathbb{M}_\beta \bar{u}\bar{u}^\top \mathbb{M}_\beta)_{i,i} \\ &\quad + (\mathbb{P}_\beta T^{-1}uu^\top)_{i,i} + (T^{-1}uu^\top \mathbb{P}_\beta)_{i,i} + (\mathbb{P}_\beta T^{-1}uu^\top \mathbb{P}_\beta)_{i,i}. \end{aligned} \quad (\text{C.108})$$

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

$$\|\mathbb{M}_\beta \bar{u}\bar{u}^\top \mathbb{M}_\beta\|_{\text{MAX}} \lesssim_P \|\bar{u}\bar{u}^\top\|_{\text{MAX}} = \|\bar{u}\|_{\text{MAX}}^2 \lesssim T^{-1} \log N.$$

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

$$\sum_{i \leq N} |\bar{u}_i^2 (\Sigma_u^{-2})_{i,i} (\mathbb{M}_\beta \bar{u}\bar{u}^\top \mathbb{M}_\beta)_{i,i}| \lesssim_P T^{-1} (\log N) \sum_{i \leq N} \bar{u}_i^2 \lesssim_P T^{-2} N \log N, \quad (\text{C.109})$$

and

$$T^{-2} \sum_{t \in T} \sum_{i \leq N} |u_{i,t}^2 (\Sigma_u^{-2})_{i,i} (\mathbb{M}_\beta \bar{u}\bar{u}^\top \mathbb{M}_\beta)_{i,i}| \lesssim_P T^{-3} (\log N) \sum_{t \in T} \sum_{i \leq N} u_{i,t}^2 \lesssim_P T^{-2} N \log N. \quad (\text{C.110})$$

Further, we obtain

$$\begin{aligned} &\sum_{i \leq N, j \leq K} \mathbb{E}(|(T^{-1}uu^\top \beta)_{i,j}| | \beta, \Sigma_u) \\ &\leq \sum_{i \leq N, j \leq K} \sqrt{\mathbb{E}((T^{-1}uu^\top)_{i,i} | \Sigma_u) \mathbb{E}((T^{-1}\beta^\top uu^\top \beta)_{j,j} | \beta, \Sigma_u)} \\ &= \sum_{i \leq N, j \leq K} \sqrt{(\Sigma_u)_{i,i} (\beta^\top \Sigma_u \beta)_{j,j}} \leq N^{3/2} K \|\beta\|_{\text{MAX}} \lambda_{\max}(\Sigma_u) \lesssim_P N^{3/2}. \end{aligned} \quad (\text{C.111})$$

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

$$\begin{aligned} &\sum_{j \leq K, k \leq K} \mathbb{E}(|(T^{-1}\beta uu^\top \beta)_{j,k}| | \beta, \Sigma_u) \\ &\leq \sum_{j \leq K, k \leq K} \sqrt{\mathbb{E}((T^{-1}\beta^\top uu^\top \beta)_{j,j} | \beta, \Sigma_u) \mathbb{E}((T^{-1}\beta^\top uu^\top \beta)_{k,k} | \beta, \Sigma_u)} \\ &= \sum_{j \leq K, k \leq K} \sqrt{(\beta^\top \Sigma_u \beta)_{j,j} (\beta^\top \Sigma_u \beta)_{k,k}} \leq K N \|\beta\|_{\text{MAX}} \lambda_{\max}(\Sigma_u) \lesssim_P N. \end{aligned} \quad (\text{C.112})$$

From (C.111) and (C.112), it directly follows

$$\sum_{i \leq N, j \leq K} |(T^{-1}uu^\top \beta)_{i,j}| \lesssim_P N^{3/2}, \quad \sum_{j \leq K, k \leq K} |(T^{-1}\beta uu^\top \beta)_{j,k}| \lesssim_P N. \quad (\text{C.113})$$

Using (C.113), and noting $\max_{i \leq N} |\bar{u}_i| \lesssim_P \sqrt{T^{-1} \log N}$ from the uniform bound on i.i.d. random variables and $\|(\beta^\top \beta)^{-1} \beta\|_{\text{MAX}} \lesssim N^{-1}$ by assumption, we obtain

$$\begin{aligned} & \sum_{i \leq N} |\bar{u}_i^2 (T^{-1}uu^\top \mathbb{P}_\beta)_{i,i}| \\ & \leq \sum_{i \leq N, j \leq K} |\bar{u}_i^2 (T^{-1}uu^\top \beta)_{i,j}| \|(\beta^\top \beta)^{-1} \beta\|_{\text{MAX}} \\ & \lesssim_P N^{-1} T^{-1} \log N \sum_{i \leq N, j \leq K} |(T^{-1}uu^\top \beta)_{i,j}| \lesssim_P N^{1/2} T^{-1} \log N, \end{aligned} \quad (\text{C.114})$$

and

$$\begin{aligned} & \sum_{i \leq N} |\bar{u}_i^2 (\mathbb{P}_\beta (T^{-1}uu^\top) \mathbb{P}_\beta)_{i,i}| \\ & \leq \sum_{i \leq N, j \leq K, k \leq K} \bar{u}_i^2 |(T^{-1}\beta uu^\top \beta)_{j,k}| \|(\beta^\top \beta)^{-1} \beta\|_{\text{MAX}}^2 \\ & \lesssim_P N^{-1} T^{-1} \log N \sum_{j \leq K, k \leq K} |(T^{-1}\beta uu^\top \beta)_{j,k}| \lesssim_P T^{-1} \log N. \end{aligned} \quad (\text{C.115})$$

Symmetric reasoning leads to

$$\frac{1}{T^2} \sum_{s \in T} \sum_{i \leq N} |u_{i,s}^2 (T^{-1}uu^\top \mathbb{P}_\beta)_{i,i}| \lesssim_P N^{1/2} T^{-1}, \quad (\text{C.116})$$

$$\frac{1}{T^2} \sum_{s \in T} \sum_{i \leq N} |u_{i,s}^2 (\mathbb{P}_\beta (T^{-1}uu^\top) \mathbb{P}_\beta)_{i,i}| \lesssim_P T^{-1}. \quad (\text{C.117})$$

Substituting (C.109), (C.110), (C.114), (C.116), (C.115), and (C.117) into (C.107) and (C.108), we obtain

$$\bar{u}^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1} N = -T^{-2} \sum_{i: i \leq N} A_i + O_P(T^{-1} N^{1/2} \log N + T^{-2} N \log N). \quad (\text{C.118})$$

Here and only here we use short-hand notation

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

Since A_i is i.i.d. across i , we only need to analyze it for a single i . It obviously holds that $E(A_i | \Sigma_u) = 0$. We also note $E(((T^{-1}uu^\top - 2\Sigma_u)_{i,i})^2 u_{i,t} u_{i,t'} u_{i,s} u_{i,s'} | \Sigma_u) = 0$ unless two elements

of $\{t, t', s, s'\}$ are the same, and $E(((T^{-1}uu^\top - 2\Sigma_u)_{i,i})^2 u_{i,t} u_{i,t'} u_{i,s} u_{i,s'} | \Sigma_u) \lesssim T^{-2} E(u_{i,t}^8 | \Sigma_u)$ unless elements of $\{t, t', s, s'\}$ only take two different values. Then we obtain

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

It hence follows that

$$T^{-2} \sum_{i:i \leq N} A_i \lesssim_P T^{-1} N^{1/2} E(u_{i,t}^8 (\Sigma_u^{-4})_{i,i}) \lesssim T^{-1} N^{1/2}, \quad (\text{C.119})$$

where the last inequality comes from that $\varepsilon_{i,t}$ has finite eighth moment by condition (a) of Assumption 2. Substituting (C.119) into (C.118), we obtain

$$\bar{u}^\top \mathbb{M}_\beta \hat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1} N = O_P(T^{-1} N^{1/2} \log N + T^{-2} N \log N). \quad (\text{C.120})$$

Combining (C.99), (C.100), and (C.120), and noting $N^{1/2} T \leq c_N$ and $T \lesssim N$ by assumption, we obtain

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

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

$$(\hat{S}^\star)^2 = (S^\star)^2 (1 + o_P(T^{-1} N^{1/2} \log N)), \implies \frac{\hat{S}^\star - S^\star}{S^\star} = o_P(T^{-1} N^{1/2} \log N).$$

And, under $S^\star \leq c_N$, we have

$$(\hat{S}^\star)^2 = (S^\star)^2 + o_P(T^{-1} N^{1/2} \log N), \implies \hat{S}^\star - S^\star = o_P\left(\sqrt{T^{-1} N^{1/2} \log N}\right).$$

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

$$(\hat{S}^\star)^2 = (S^\star)^2 + \frac{N}{T} + (S^\star)^2 o_P\left(\frac{\sqrt{N} \log N}{T}\right) \implies \frac{\hat{S}^\star - \sqrt{(S^\star)^2 + N/T}}{S^\star} = o_P\left(\frac{\sqrt{N} \log N}{T}\right).$$

Similarly, under $S^\star \leq c_N$, we have

$$(\hat{S}^\star)^2 = (S^\star)^2 + \frac{N}{T} + o_P\left(\frac{\sqrt{N} \log N}{T}\right) \implies \hat{S}^\star - \sqrt{(S^\star)^2 + N/T} = o_P\left(\frac{\sqrt{N} \log N}{T}\right).$$

The proof concludes.

C.6 Proof of Proposition B1

We note that all the assumptions in Theorem 2, other than the factor-neutrality, are assumed here as well. Therefore, equation (C.8) in the proof of Theorem 2 stays valid, since factor-neutrality is used only after the derivation of (C.8). Applying Cauchy-Schwarz inequality to (C.8), we obtain

$$|S(w)| \leq \sqrt{\mathbb{E}(r_{t+1}|\mathcal{G})^\top \Sigma^{-1} \mathbb{E}(r_{t+1}|\mathcal{G})} + o_P(1). \quad (\text{C.121})$$

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

$$\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1} \beta (\Sigma_v^{-1} + \beta^\top \Sigma_u^{-1} \beta)^{-1} \beta^\top \Sigma_u^{-1}. \quad (\text{C.122})$$

By direct calculations, we have

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

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

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

Therefore, using the fact that $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$ and that $\lambda_{\max}(\Sigma_u) \lesssim_P 1$ by assumption, we have

$$\lambda_{\max}((\beta^\top \Sigma_u^{-1} \beta)^{-1}) = \lambda_{\min}^{-1}(\beta^\top \Sigma_u^{-1} \beta) \leq \lambda_{\min}^{-1}(\beta^\top \beta) \lambda_{\max}(\Sigma_u) \lesssim_P N^{-1}. \quad (\text{C.123})$$

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

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

we have

$$\|\beta^\top \Sigma^{-1} \beta - \Sigma_v^{-1}\| \lesssim_P N^{-1},$$

which in turn leads to

$$\gamma^\top \beta^\top \Sigma^{-1} \beta \gamma = \gamma^\top \Sigma_v^{-1} \gamma + o_P(1). \quad (\text{C.124})$$

Next, we show

$$\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma^{-1} \beta \gamma = o_P(1). \quad (\text{C.125})$$

Notice that $\mathbb{E}(\mathbb{E}(\alpha|\mathcal{G})|\Sigma, \beta) = \mathbb{E}(\alpha|\Sigma, \beta) = \mathbb{E}(\alpha|\Sigma) = 0$ (by conditions (a) and (b) of Assumption 1), and that, conditionally on (Σ, β) , $\mathbb{E}(\alpha_i|\mathcal{G})$ is independent across i as demonstrated in the beginning of the proof of Theorem 2. Therefore,

$$\mathbb{E}((\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma^{-1} \beta \gamma)^2 | \Sigma, \beta) \leq \sum_{i \leq N} \mathbb{E}(\mathbb{E}(\alpha_i | \mathcal{G})^2 | \Sigma, \beta) \max_{j \leq N} (\gamma^\top \beta^\top \Sigma^{-1})_j^2. \quad (\text{C.126})$$

On the other hand, from (C.122), we obtain

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

Because of $\lambda_{\min}(\Sigma_v) \gtrsim 1$, $\|\beta\|_{\text{MAX}} \lesssim_P 1$, $\|\Sigma_u\|_{\text{MAX}} \leq \|\Sigma_u\| \lesssim_P 1$, $\lambda_{\min}(\Sigma_u) \gtrsim_P 1$, Σ_u is diagonal, and (C.123), we have

$$\|\gamma^\top \beta^\top \Sigma^{-1}\|_{\text{MAX}} \lesssim \|(\Sigma_v^{-1} + \beta^\top \Sigma_u^{-1} \beta)^{-1}\| \|\beta^\top \Sigma_u^{-1}\|_{\text{MAX}} \lesssim_P \lambda_{\max}((\beta^\top \Sigma_u^{-1} \beta)^{-1}) \lesssim_P N^{-1}.$$

Hence, we have, for all positive fixed ϵ ,

$$\mathbb{P}(|\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma^{-1} \beta \gamma| \geq \epsilon | \Sigma, \beta) \leq \mathbb{E}((\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma^{-1} \beta \gamma)^2 | \Sigma, \beta) / \epsilon^2 = o_P(1), \quad (\text{C.127})$$

where the last equality comes from (C.126) and that $\mathbb{E}(\sum_{i \leq N} \mathbb{E}(\mathbb{E}(\alpha_i | \mathcal{G})^2 | \Sigma, \beta)) \leq \sum_{i \leq N} \mathbb{E}(\alpha_i^2) = o(N)$ by condition (a) of Assumption 1. Since $\mathbb{P}(|\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma^{-1} \beta \gamma| \geq \epsilon | \Sigma, \beta) \leq 1$ are uniformly bounded for all N (by definition), we obtain by taking expectations on both sides of (C.127) that, for all positive fixed ϵ ,

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

which is equivalent to (C.125).

Finally, we derive

$$\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma^{-1} \mathbb{E}(\alpha|\mathcal{G}) = \mathbb{E}(\alpha|\mathcal{G})^\top \Sigma_u^{-1} \mathbb{E}(\alpha|\mathcal{G}) + o_P(1). \quad (\text{C.128})$$

Following the same derivation for (C.126), we obtain

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

Because $\|\beta\|_{\text{MAX}} \lesssim_P 1$ and $\lambda_{\min}(\Sigma_u) \gtrsim_P 1$, we have

$$\max_j (\Sigma_u^{-1} \beta \beta^\top \Sigma_u^{-1})_{j,j} \lesssim \|\Sigma_u^{-1} \beta\|_{\text{MAX}}^2 \lesssim_P 1.$$

Then given the above result that $\mathbb{E}(\sum_{i \leq N} \mathbb{E}(\mathbb{E}(\alpha_i | \mathcal{G})^2 | \Sigma, \beta)) = o(N)$, we obtain that $\mathbb{E}(|\mathbb{E}(\alpha|\mathcal{G})^\top \Sigma_u^{-1} \beta|_{\text{F}}^2 | \Sigma, \beta) = o_P(N)$. Therefore, similar to the derivation of (C.125), we ob-

tain

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

On the other hand, using (C.123), we obtain

$$\|(\Sigma_v^{-1} + \beta^\top \Sigma_u^{-1} \beta)^{-1}\| = \lambda_{\min}^{-1}(\Sigma_v^{-1} + \beta^\top \Sigma_u^{-1} \beta) \leq \lambda_{\max}((\beta^\top \Sigma_u^{-1} \beta)^{-1}) \lesssim_{\mathbb{P}} N^{-1}. \quad (\text{C.129})$$

Then, using (C.129), we have

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

and hence, in light of (C.122), we obtain (C.128).

Given that $\mathbb{E}(r_{t+1}|\mathcal{G}) = \mathbb{E}(\alpha|\mathcal{G}) + \beta\gamma$, it follows from (C.124), (C.125), and (C.128) that

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

In light of (C.121), 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

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

ζ represents the signal strength of our alphas. \hat{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:

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

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

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

Further, $\tilde{\psi}_q(\lambda)$ and $\hat{\psi}_q(\lambda)$ to represent, respectively, the vectors $(\tilde{\psi}_q(\check{z}_1, \lambda), \dots, \tilde{\psi}_q(\check{z}_N, \lambda))^\top$ and $(\hat{\psi}_{q,1}(\lambda), \dots, \hat{\psi}_{q,N}(\lambda))^\top$. Based on $\hat{\psi}_q(\lambda)$, we introduce $\hat{w}'_q(\lambda) := \hat{\Sigma}_u^{-1/2} \hat{\psi}_q(\lambda)$. The purpose of doing so is that $\hat{w}'_q(\lambda)$, with appropriate choice of λ , incorporates the weights of Lasso, CSR,

and BH methods all as special cases. Lastly, we define

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

$\widehat{S}'_q(\lambda)$ 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 $\widehat{\Sigma}_u^{-1} \widehat{\alpha} = \check{w}_1(\lambda)$ 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 $\lambda = 0$,

$$S_\lambda^{\text{Lasso}} = N^{1/2} \mathbb{E}(\check{s}_i s_i) \mathbb{E}(\check{s}_i^2)^{-1/2}. \quad (\text{C.130})$$

Moreover, we can write

$$\sqrt{T} \times \mathbb{E}(\check{s}_i s_i) = \frac{\rho \mu}{2\sigma} \int a(\phi(a - \zeta) - \phi(a + \zeta)) da = \frac{\rho \mu \zeta}{\sigma}, \quad (\text{C.131})$$

$$T \times \mathbb{E}(\check{s}_i^2) = \int a^2(\phi(a) + \rho \phi(a - \zeta)) da = 1 + \rho(1 + \zeta^2) = (1 + o(1))(1 + \rho \zeta^2). \quad (\text{C.132})$$

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 $\rho \rightarrow 0$. Therefore, it holds that, choosing $\lambda = 0$,

$$\widehat{S}^{\text{CSR}} = \widehat{S}_1(\lambda) = S_\lambda^{\text{Lasso}} + o_P(1 + S^{\text{OPT}}) = S^{\text{CSR}} + o_P(1 + S^{\text{OPT}}). \quad (\text{C.133})$$

Here the second equality comes from $\widehat{S}_1(\lambda) = S_\lambda^{\text{Lasso}} + o_P(1 + S^{\text{OPT}})$ as in Proposition B4. The third equality comes from (C.130), (C.131), (C.132), $\zeta = \sqrt{T} \mu / \sigma$ by definition, and that $|S^{\text{CSR}}| \leq S^{\text{OPT}} + o_P(1)$ by Theorem 2 and Corollary 2.

Now we verify the sufficient and necessary condition. We utilize the last equality of (C.133), 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 $\sqrt{T} \mu / \sigma \rightarrow 0$ and $\sqrt{\rho T} \mu / \sigma \rightarrow \infty$ separately. The former case is obvious since $\lambda = 0$ certainly satisfies $T^{1/2} \lambda \rightarrow 0$ condition required by Proposition B4. For the latter case, noting that $\sqrt{\rho T} \mu / \sigma \rightarrow \infty$ always leads to $\sqrt{T} \mu / \sigma - \sqrt{-2 \log \rho} \rightarrow \infty$ due to $-\log \rho \approx \log N$ by assumption, we only need to verify that, under

$\lambda = 0$, we have both $\sqrt{T}(\mu/\sigma - \lambda) \rightarrow \infty$ and $\frac{\phi(\sqrt{T}\lambda)}{\rho(1+T\lambda^2)T(\mu/\sigma-\lambda)^2} \rightarrow 0$ as in Proposition B4. Both are obviously true.

Next we show the “only if” part. As $\sqrt{T}\mu/\sigma \rightarrow 0$ is violated, we only need to show that, if $\sqrt{\rho T}\mu/\sigma \rightarrow \infty$ is violated, we would not have $\frac{\phi(\sqrt{T}\lambda)}{\rho(1+T\lambda^2)T(\mu/\sigma-\lambda)^2} \rightarrow 0$ 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 $\widehat{S}_\tau^{\text{BH}} = S_\tau^{\text{BH}} + o_P(1 + S^{\text{OPT}})$. 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 $\check{z} = -\Phi^{-1}(p_{(\widehat{k})}/2)$ and $\check{z}' = -\Phi^{-1}(p_{(\widehat{k}+1)}/2)$, where we recall Φ is the standard normal cdf. In other words, \check{z} and \check{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 \check{z} and \check{z}' , we obtain

$$\frac{2N\Phi(-\check{z})}{\sum_{j \in \{0, +, -\}} m_j(\check{z})} \leq \tau, \quad \frac{2N\Phi(-\check{z}')}{\sum_{j \in \{0, +, -\}} m_j(\check{z}')} > \tau, \quad \sum_{j \in \{0, +, -\}} m_j(\check{z}') = \sum_{j \in \{0, +, -\}} m_j(\check{z}) + 1. \quad (\text{C.134})$$

Given (E.74) and the monotonicity of $m_j(z)$ in z , there exists a positive sequence $b_N = o(1/\sqrt{\log N})$ such that, with high probability, uniformly over x , and for $j \in \{0, +, -\}$,

$$\check{m}_j(x + b_N) \leq m_j(x) \leq \check{m}_j(x - b_N). \quad (\text{C.135})$$

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 $N\Phi(-a_N) \rightarrow \infty$ and $\rho N\Phi(\zeta - a'_N) \rightarrow \infty$,

$$\sup_{0 \leq z \leq a_N} \left| \frac{\check{m}_0(z)}{2N\Phi(-z)} - 1 \right| = o_P(1), \quad \sup_{0 \leq z \leq a'_N} \left| \frac{2\check{m}_\pm(z)}{\rho N\Phi(\zeta - z) + \rho N\Phi(-\zeta - z)} - 1 \right| = o_P(1). \quad (\text{C.136})$$

Using (E.15) and $-\Phi^{-1}(N^{-1}) \asymp \sqrt{\log N}$, we note that, for $b_N = o(1/\sqrt{\log N})$, $\Phi(-z \pm b_N) = \Phi(-z)(1 + o(1))$ uniformly over $|z| \leq -\Phi^{-1}(N^{-1})$. Hence we obtain from (C.135) and (C.136) that, for all deterministic sequences (a_N, a'_N) satisfying $N\Phi(-a_N) \rightarrow \infty$ and $\rho N\Phi(\zeta - a'_N) \rightarrow \infty$,

$$\sup_{0 \leq z \leq a_N} \left| \frac{m_0(z)}{2N\Phi(-z)} - 1 \right| = o_P(1), \quad \sup_{0 \leq z \leq a'_N} \left| \frac{2m_{\pm}(z)}{\rho N\Phi(\zeta - z) + \rho N\Phi(-\zeta - z)} - 1 \right| = o_P(1). \quad (\text{C.137})$$

In addition, it follows from (E.15) and the second part of (E.1) that, uniformly over $z \geq 0$,

$$\frac{\rho\Phi(\zeta - z)}{\Phi(-z)} \approx h(z), \quad \text{with} \quad h(z) := \frac{z}{1 + (z - \zeta)_+} e^{(z - a^*)\zeta}. \quad (\text{C.138})$$

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

$$\inf_{z \geq c_{x'}} h(z) \rightarrow \infty. \quad (\text{C.139})$$

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 $h(z) \gtrsim h(a_N) z^{-1} e^{(z - a_N)\zeta} \rightarrow \infty$ as $\zeta \gtrsim \sqrt{\log N}$. We hence establish (C.139).

We first suppose $z^* \geq c_{d_2}$. From (C.138) and the definition of z^* (B.1) we have $h(z^*) \approx 1$. Then, since $d_2 < d_3$, it follows from (C.139) (by contradiction) that

$$\sup_{0 \leq z \leq c_{d_3}} \frac{\rho\Phi(\zeta - z)}{\Phi(-z)} \rightarrow 0. \quad (\text{C.140})$$

Using (C.140), and applying (C.137) and the monotonicity of $m_j(z)$ and $\Phi(-z)$ in z , it holds that, for all (random) sequence z satisfying $0 \leq z \leq c_{d_3}$,

$$m_0(z) = 2N\Phi(-z)(1 + o_P(1)), \quad m_{\pm}(z) = o_P(N\Phi(-z)). \quad (\text{C.141})$$

Here we also use $\rho\Phi(-\zeta - z) \leq \Phi(-z)$. Comparing (C.141) with (C.134), we have that $P(\tilde{z} \leq c_{d_3}) \rightarrow 0$. We hence have, with high probability,

$$\mu^2 \rho N\Phi(\zeta - \tilde{z}) \leq \mu^2 \rho N\Phi(\zeta - c_{d_3}) \leq \mu^2 N\Phi(-c_{d_3}) = o_P(\mu^2 N^d) = o_P(1). \quad (\text{C.142})$$

Here the second inequality comes from (C.140), the first equality comes from $d_3 < d$, and the last equality comes from $\mu \leq N^{-d}$. (The situation where \tilde{z} does not exist is equivalent with

\check{z} being so large that $\sup_i |\hat{z}_i| < \check{z}$ and is hence a special case.) Given (C.142), Lemma E7 applies and we obtain $\hat{S}_\tau^{\text{BH}} = \tilde{S}_2(\check{z}) = o_P(1)$.

On the other hand, we have

$$\begin{aligned} S_\tau^{\text{BH}} &= \frac{N^{1/2} \mathbb{E}(\tilde{\psi}_{2,i}(z^*) s_i)}{\mathbb{E}(\tilde{\psi}_{2,i}(z^*)^2)^{1/2}} \leq \frac{\mu N^{1/2} \mathbb{E}(|\tilde{\psi}_{2,i}(z^*)| \mathbb{1}_{\{\alpha_i \neq 0, |\tilde{z}_i| \geq z^*\}})}{\sigma \mathbb{E}(\tilde{\psi}_{2,i}(z^*)^2)^{1/2}} \\ &\leq \frac{\mu}{\sigma} \sqrt{N \mathbb{E}(\mathbb{1}_{\{\alpha_i \neq 0, |\tilde{z}_i| \geq z^*\}})} \leq \frac{\mu}{\sigma} \sqrt{2\rho N \Phi(\zeta - z^*)} \lesssim \mu \sqrt{N \Phi(-z^*)} = o(1). \end{aligned}$$

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 $N\Phi(-c_{d_2}) = o_P(N^d)$ as $d_2 < d$. Hence we prove $\hat{S}_\tau^{\text{BH}} - S_\tau^{\text{BH}} = o_P(1)$ under $z^* \geq c_{d_2}$.

Now suppose $z^* \leq c_{d_2}$. Then it follows from (C.138), (B.1) and (C.139) that, as $d_1 < d_2$,

$$\inf_{z \geq c_{d_1}} \frac{\rho \Phi(\zeta - z)}{\Phi(-z)} \rightarrow \infty. \quad (\text{C.143})$$

Since $d_1 > 0$, we have $N\Phi(-c_{d_1}) \rightarrow \infty$. Given (C.143), we further have $\rho N\Phi(\zeta - c_{d_1}) \rightarrow \infty$. Applying (C.137), and using (C.143), it holds that

$$m_0(c_{d_1}) = o_P(\rho \Phi(\zeta - c_{d_1})), \quad m_\pm(c_{d_1}) = \rho \Phi(\zeta - c_{d_1})(1 + o_P(1)). \quad (\text{C.144})$$

Let $\tilde{h}(z) := 2N\Phi(-z) / \sum_{j \in \{0, +, -\}} m_j(z)$. From (C.144), we know that $h(c_{d_1}) = o_P(1)$. Since $\tilde{h}(0) = 1$, we know that with high probability \check{z} exists and satisfy $0 \leq \check{z} \leq c_{d_1}$. Since $\Phi(-z)$ is decreasing in z , we have $N\Phi(-\check{z}) \rightarrow \infty$ and $\rho N\Phi(\zeta - \check{z}) \rightarrow \infty$ in probability. Therefore, we obtain from (C.137) that

$$m_0(\check{z}) = 2N\Phi(-\check{z})(1 + o_P(1)), \quad m_\pm(\check{z}) = \frac{\rho}{2} N(\Phi(\zeta - \check{z}) + \Phi(-\zeta - \check{z}))(1 + o_P(1)). \quad (\text{C.145})$$

Since $\check{z}' \leq \check{z}$, (C.145) would still hold if all \check{z} are replaced by \check{z}' . Hence, substituting (C.145) back into (C.134), and noting $\Phi(-\zeta - z) \leq \Phi(-z)$, we have

$$\frac{2(1 - \tau)\Phi(-\check{z})}{\tau \rho \Phi(\zeta - \check{z})} = 1 + o_P(1). \quad (\text{C.146})$$

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.139)). As a result, using (E.15), and comparing (B.1) and (C.146), we have

$$|\check{z} - z^*| = o_P\left(1/\sqrt{\log N}\right). \quad (\text{C.147})$$

We finally obtain

$$\widehat{S}_\tau^{\text{BH}} = \frac{\widehat{w}'_q(T^{-1/2}\check{z})^\top \mathbb{M}_\beta \alpha}{\sigma \|\mathbb{M}_\beta \widehat{w}'_q(T^{-1/2}\check{z})\|} = S_2(z^*) + o_P(S^{\text{OPT}} + 1) = S_\tau^{\text{BH}} + o_P(S^{\text{OPT}} + 1).$$

where the first equality holds by definition, the second comes from (C.147), $\rho N \Phi(\zeta - z^*) \geq \rho N \Phi(\zeta - c_{d_1}) \rightarrow \infty$ (see after (C.143)), and Lemma E6 (choose $q = 2$ and $(\lambda', \lambda) = (\check{z}, z^*)$), and the last holds by $S_2(z^*) = S_\tau^{\text{BH}}$ due to $\psi_i = \mathbb{E}(s_i|\mathcal{G})$ and that $\widetilde{\psi}_{q,i}(z^*)$ is \mathcal{G} -measurable. We hence prove $\widehat{S}_\tau^{\text{BH}} - S_\tau^{\text{BH}} = o_P(S^{\text{OPT}} + 1)$ 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

$$\frac{S_\tau^{\text{BH}}}{S^{\text{OPT}}} = \frac{S_2(z^*)}{S^{\text{OPT}}} = \frac{\mathbb{E}(\widetilde{\psi}_{2,i}(z^*)\psi_i)}{\mathbb{E}(\widetilde{\psi}_{2,i}(z^*)^2)^{1/2}\mathbb{E}(\psi_i^2)^{1/2}} = \text{Corr}(\psi_i, \widetilde{\psi}_{2,i}(z^*)). \quad (\text{C.148})$$

The first equality holds by $S_2(z^*) = S_\tau^{\text{BH}}$ due to $\psi_i = \mathbb{E}(s_i|\mathcal{G})$ and that $\widetilde{\psi}_{q,i}(z^*)$ is \mathcal{G} -measurable. Since $\rho \rightarrow 0$ by assumption, we know that $z^* \rightarrow 0$ does not hold as it contradicts its definition (B.1). Moreover, we have that, using the first part of (E.1) and that $-\log \rho \approx \log N$ 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.148) 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 ,

$$\frac{\Phi(-z^*)}{\rho \Phi(\zeta - a^*)} \leq \epsilon. \quad (\text{C.149})$$

The rest of the proof is to show (C.149). Rewriting the definition of z^* , and noting τ is fixed, we have

$$\frac{\rho \Phi(\zeta - z^*)}{2\Phi(-z^*)} = \frac{1 - \tau}{\tau} \approx 1. \quad (\text{C.150})$$

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.150) that

$$h(z^*) := \frac{z^*}{1 + (z^* - \zeta)_+} e^{(z^* - a^*)\zeta} \approx 1. \quad (\text{C.151})$$

Here we also introduce $h(z^*)$ as short-hand notation. Suppose $z^* \geq a^* + 1$. Then $h(z^*) \geq e^{(z^* - a^*)\zeta} \rightarrow \infty$. Suppose $a^* \leq z^* \leq a^* + 1$. Then $h(z^*) \geq \frac{1 + a^*}{2 + (a^* - \zeta)_+} \rightarrow \infty$. Suppose $z^* \leq a^* - 1$.

Then $h(z^*) \leq a^* e^{-\zeta} \rightarrow 0$. Hence we conclude $a^* - 1 \leq z^* \leq a^*$ for large N . Therefore, it follows from (C.151) that

$$e^{(a^* - z^*)\zeta} \approx \frac{a^*}{1 + (a^* - \zeta)_+}. \quad (\text{C.152})$$

It then holds that, when $\zeta - \zeta^* \geq -c_N \sqrt{\log N}$,

$$\begin{aligned} a^* - z^* &\lesssim \frac{1}{\zeta} \log \frac{a^*}{1 + (a^* - \zeta)_+} \\ &\lesssim \frac{1}{1 + (a^* - \zeta)_+} \frac{1 + (a^* - \zeta)_+}{a^*} \log \frac{a^*}{1 + (a^* - \zeta)_+} \leq o\left(\frac{1}{1 + (a^* - \zeta)_+}\right). \end{aligned} \quad (\text{C.153})$$

The first inequality follows from (C.152). The second inequality comes from $a^* \asymp \zeta$ due to the first part of (E.1) and $\zeta^* \asymp \sqrt{\log N}$ 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 $\lim_{x \rightarrow \infty} x^{-1} \log x = 0$. We hence obtain

$$\frac{|\Phi(\zeta - z^*) - \Phi(\zeta - a^*)|}{\Phi(\zeta - a^*)} \lesssim (a^* - z^*) \max_{z^* \leq a \leq a^*} \phi(a - \zeta) = o\left(\frac{\phi(a^* - \zeta)}{1 + (a^* - \zeta)_+}\right) = o(\Phi(\zeta - a^*)). \quad (\text{C.154})$$

The first inequality is obvious given that $z^* \leq a^*$ for large N . The first equality comes from (C.153) 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.154) that $\Phi(\zeta - a^*) = \Phi(\zeta - z^*)(1 + o(1))$. Combining this result with (C.150), we further have $2\Phi(-z^*)/(\rho\Phi(\zeta - a^*)) = (1 + o(1))\tau/(1 - \tau)$. We hence prove (C.149) and the proof concludes. ■

C.7.3 Proof of Proposition B4

By definition, we have $\check{w}_{1,i}(\lambda) = \hat{\sigma}_i^{-1} \psi_1(\hat{s}_i, \lambda) = T^{-1/2} \hat{\sigma}_i^{-1} \psi_1(\hat{z}_i, T^{1/2} \lambda) = T^{-1/2} \tilde{w}_1'(T^{1/2} \lambda)$ and $\psi_1(\check{s}_i, \lambda) = T^{-1/2} \tilde{\psi}_1(T^{1/2} \lambda)$. Hence, it holds that

$$\hat{S}_{1,\lambda} = \frac{\check{w}_1(\lambda)^\top \mathbb{M}_\beta \alpha}{\sigma \|\mathbb{M}_\beta \check{w}_1(\lambda)\|} = \hat{S}_1'(T^{1/2} \lambda), \quad S_\lambda^{\text{Lasso}} = \frac{N^{1/2} \mathbb{E}(\psi_i \psi_1(\check{s}_i, \lambda))}{\sqrt{\mathbb{E}(\psi_1(\check{s}_i, \lambda)^2)}} = S_1(T^{1/2} \lambda), \quad (\text{C.155})$$

where the first equality in the second part comes from $\psi_i = \mathbb{E}(s_i | \mathcal{G})$ and that $\psi_1(\check{s}_i, \lambda)$ is \mathcal{G} -measurable.

Suppose λ satisfies $\rho N \Phi(\zeta - T^{1/2} \lambda) \rightarrow \infty$. Then we have

$$\hat{S}_{1,\lambda} = \hat{S}_1'(T^{1/2} \lambda) = S_1(T^{1/2} \lambda) + o_P(S^{\text{OPT}} + 1) = S_\lambda^{\text{Lasso}} + o_P(S^{\text{OPT}} + 1). \quad (\text{C.156})$$

Here the first and last equality comes from (C.155) and the second equality comes from

Lemma E6 as $\rho N\Phi(\zeta - T^{1/2}\lambda) \rightarrow \infty$ (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

$$S_1(T^{1/2}\lambda) = \frac{N^{1/2}\mathbb{E}(\tilde{\psi}_{1,i}(T^{1/2}\lambda)s_i)}{\mathbb{E}(\tilde{\psi}_{1,i}(T^{1/2}\lambda)^2)^{1/2}} \leq \frac{\mu}{\sigma}\sqrt{N\mathbb{E}(\mathbb{1}_{\{|s_i| \geq 0, \alpha_i \neq 0\}})} \leq \frac{\mu}{\sigma}\sqrt{2\rho N\Phi(\zeta - T^{1/2}\lambda)} = o(1). \quad (\text{C.157})$$

Here the first equality comes from $\psi_i = \mathbb{E}(s_i|\mathcal{G})$ and that $\tilde{\psi}_{1,i}(\lambda)$ 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 $\mu^2\rho N\Phi(\zeta - T^{1/2}\lambda) \rightarrow 0$, Lemma E7 applies and hence $\hat{S}'_1(T^{1/2}\lambda) = o_P(1)$. Combining with (C.157), we obtain $\hat{S}'_1(T^{1/2}\lambda) = S_1(T^{1/2}\lambda) + o_P(1)$. Given (C.155), we obtain $\hat{S}_1(\lambda) = S_\lambda^{\text{Lasso}} + o_P(1)$ under $\rho N\Phi(\zeta - T^{1/2}\lambda) \lesssim 1$. Hence we establish $\hat{S}_1(\lambda) = S_\lambda^{\text{Lasso}} + o_P(1)$ with the classic subsequence argument.

Finally, it holds that

$$\frac{S_\lambda^{\text{Lasso}}}{S^{\text{OPT}}} = \frac{S_1(T^{1/2}\lambda)}{S^{\text{OPT}}} = \frac{\mathbb{E}(\tilde{\psi}_{1,i}(T^{1/2}\lambda)\psi_i)}{\mathbb{E}(\tilde{\psi}_{1,i}(T^{1/2}\lambda)^2)^{1/2}\mathbb{E}(\psi_i^2)^{1/2}} = \text{Corr}(\psi_i, \tilde{\psi}_{1,i}(T^{1/2}\lambda)), \quad (\text{C.158})$$

where the first equality comes from (C.155) and the second comes from the definition of S_1 and $S^{\text{OPT}} = N^{1/2}\mathbb{E}(\psi_i^2)^{1/2}$. On the other hand, we have that, using the first part of (E.1) and that $-\log \rho \approx \log N$ by assumption, $\zeta - a^* \rightarrow \infty$ and $\zeta - \zeta^* \rightarrow \infty$ 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.158), 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 $\|\beta\|_{\text{MAX}} \lesssim_P 1$ and $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$. Also suppose $\log N \lesssim T \lesssim N^d$ with fixed $d < 1$. Then it holds that, as $N, T \rightarrow \infty$,*

$$\max_{1 \leq i \leq N} |\hat{\sigma}_i^2 - \sigma_i^2| = O_P\left(\sqrt{(\log N)/T}\right), \quad (\text{D.1})$$

$$\max_{1 \leq i \leq N} |(\mathbb{P}_\beta \bar{u})_i| = O_P\left(1/\sqrt{TN}\right), \quad (\text{D.2})$$

$$\max_{1 \leq i \leq N} |(\mathbb{P}_\beta \alpha)_i| = O_P(N^{-1/2}\mathbb{E}(s_i^2)^{1/2}). \quad (\text{D.3})$$

Proof. We start with (D.1). First of all, we write

$$\begin{aligned}
\max_{1 \leq i \leq N} |\hat{\sigma}_i^2 - \sigma_i^2| &\leq \|\mathbb{M}_\beta(T^{-1}uu^\top - \bar{u}\bar{u}^\top)\mathbb{M}_\beta - \Sigma_u\|_{\text{MAX}} \\
&\leq \|\mathbb{M}_\beta\Sigma_u\mathbb{M}_\beta - \Sigma_u\|_{\text{MAX}} + \|\mathbb{M}_\beta(T^{-1}uu^\top - \bar{u}\bar{u}^\top - \Sigma_u)\mathbb{M}_\beta\|_{\text{MAX}}. \quad (\text{D.4})
\end{aligned}$$

Now we establish the upper bounds of the two terms in the second line. We write

$$\begin{aligned}
\|\mathbb{M}_\beta\Sigma_u\mathbb{M}_\beta - \Sigma_u\|_{\text{MAX}} &\leq \|\mathbb{P}_\beta\Sigma_u\mathbb{P}_\beta\|_{\text{MAX}} + 2\|\mathbb{P}_\beta\Sigma_u\|_{\text{MAX}} \\
&\leq (N\|\mathbb{P}_\beta\|_{\text{MAX}} + 2)\|\mathbb{P}_\beta\|_{\text{MAX}}\|\Sigma_u\|_{\text{MAX}} \lesssim_P N^{-1}. \quad (\text{D.5})
\end{aligned}$$

The last inequality comes from $\|\mathbb{P}_\beta\|_{\text{MAX}} \leq C\|\beta\|_{\text{MAX}}^2\|(\beta^\top\beta)^{-1}\|_{\text{MAX}} \lesssim_P N^{-1}$, which is true by assumption. On the other hand, we have

$$\begin{aligned}
&\|\mathbb{M}_\beta(T^{-1}uu^\top - \bar{u}\bar{u}^\top - \Sigma_u)\mathbb{M}_\beta\|_{\text{MAX}} \\
&\leq \|T^{-1}uu^\top - \bar{u}\bar{u}^\top - \Sigma_u\|_{\text{MAX}}(1 + 2N\|\mathbb{P}_\beta\|_{\text{MAX}} + N^2\|\mathbb{P}_\beta\|_{\text{MAX}}) \lesssim_P \sqrt{(\log N)/T}, \quad (\text{D.6})
\end{aligned}$$

where the last inequality comes from the uniform bound on i.i.d. normal variables, $\log N \lesssim T$ by assumption, and that $\lambda_{\max}(\Sigma_u) \lesssim_P 1$ by condition (a) of Assumption 1. Substituting (D.5) and (D.6) into (D.4), and noting $N^{-1} \leq C\sqrt{(\log N)/T}$ by assumption, we obtain (D.1).

We obtain (D.2) by writing

$$\max_{1 \leq i \leq N} |(\mathbb{P}_\beta\bar{u})_k| \leq C\|\beta\|_{\text{MAX}}\|(\beta^\top\beta)^{-1}\|_{\text{MAX}} \max_{1 \leq k \leq K} |(\beta^\top\bar{u})_k| \lesssim_P \max_{1 \leq k \leq K} N^{-1}|(\beta^\top\bar{u})_k| \lesssim_P 1/\sqrt{TN}.$$

Here the last inequality comes from that K is fixed, $E(\bar{u}_i\bar{u}_j|\beta, \Sigma_u) \lesssim \delta_{i,j}\sigma_i^2T^{-1}$ by conditions (a) and (b) of Assumption 1, and $\lambda_{\max}(\Sigma_u) \lesssim_P 1$.

Finally, we write

$$\max_{1 \leq i \leq N} |(\mathbb{P}_\beta\alpha)_i| \leq C\|\beta\|_{\text{MAX}}\|(\beta^\top\beta)^{-1}\|_{\text{MAX}} \max_{1 \leq k \leq K} |(\beta^\top\alpha)_k| \leq CN^{-1} \max_{1 \leq k \leq K} |(\beta^\top\alpha)_k|. \quad (\text{D.7})$$

On the other hand, from condition (a) and (b) of Assumption 1 and condition (a) of Assumption 2, we have $E(\alpha_i\alpha_j|\beta, \Sigma_u) = \delta_{i,j}\sigma_i^2E(s_i^2)$. Therefore, as K is fixed and $\lambda_{\max}(\Sigma_u) \lesssim_P 1$, we have

$$\max_{1 \leq k \leq K} |(\beta^\top\alpha)_k| \lesssim_P c_N N^{1/2}\|\beta\|_{\text{MAX}} \lesssim_P N^{1/2}E(s_i^2)^{1/2}. \quad (\text{D.8})$$

Substituting (D.8) into (D.7), we obtain (D.3). ■

Lemma D2. Suppose Assumptions 1 and 2 hold and that $\|\beta\|_{\text{MAX}}^2 \lesssim_P 1$ and $\lambda_{\min}(\beta^\top\beta) \gtrsim_P N$. Also assume $N^d \lesssim T \lesssim N^{d'}$ with fixed $d > 0$ and $d' < 1$. Then it holds that, as $N, T \rightarrow \infty$,

$$\max_{1 \leq i \leq N} |\hat{s}_i - \check{s}_i| \lesssim_P c_N T^{-1/2} \tilde{k}_N. \quad (\text{D.9})$$

If we additionally have $T \gtrsim N^d$ with fixed $d > 1/2$, then it holds that, as $N, T \rightarrow \infty$,

$$\max_{i \in B} |\widehat{s}_i - \check{s}_i| \lesssim_P \chi_N. \quad (\text{D.10})$$

Here $\chi_N := \sqrt{1/N(k_N^5 + \mathbb{E}(s_j^2))^{1/2}}$, 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

$$\widehat{s}_i - \check{s}_i = -\frac{(\mathbb{P}_\beta \alpha)_i}{\widehat{\sigma}_i} - \frac{(\mathbb{P}_\beta \bar{u})_i}{\widehat{\sigma}_i} + \left(\frac{\sigma_i}{\widehat{\sigma}_i} - 1 \right) \check{s}_i. \quad (\text{D.11})$$

Since $T \gtrsim N^d$ with $d > 0$ by assumption, (D.1) of Lemma D1 leads to $\max_{1 \leq i \leq N} |\widehat{\sigma}_i^2 - \sigma_i^2| = o_P(1)$. Then, noting $\min_i \sigma_i \gtrsim_P 1$ by condition (d) of Assumption 1, we obtain $\min_i \widehat{\sigma}_i \gtrsim_P 1$. Applying (D.1) again, we have $\max_i \left| \frac{\sigma_i}{\widehat{\sigma}_i} - 1 \right| \lesssim_P \sqrt{(\log N)/T}$. Using these two results, and substituting (D.2) and (D.3) of Lemma D1 into (D.11), we obtain

$$\max_{1 \leq i \leq N} |\widehat{s}_i - \check{s}_i| \lesssim_P \chi_N + \sqrt{(\log N)/T} \max_{1 \leq i \leq N} |\check{s}_i|, \quad \max_{i \in B} |\widehat{s}_i - \check{s}_i| \lesssim_P \chi_N + \sqrt{(\log N)/T} \max_{i \in B} |\check{s}_i|. \quad (\text{D.12})$$

Since $P(|s_i| \geq 1) \leq \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq 1\}}) \leq c_N N^{-1}$ by condition (a) of Assumption 2, we have $P(\max_i |s_i| \geq 1) \leq c_N$. Combining this result with $\max_i |\bar{\varepsilon}_i| \lesssim_P \sqrt{(\log N)/T}$ by the uniform bound on i.i.d. normal variables, we obtain $\max_{1 \leq i \leq N} |\check{s}_i| \lesssim_P 1$ (again noting $T \gtrsim N^d$ with $d > 0$ by assumption). Then we have $\sqrt{(\log N)/T} \max_{1 \leq i \leq N} |\check{s}_i| \lesssim_P T^{-1/2} \sqrt{\log N}$. Also, we have $\chi_N \leq c_N T^{-1/2}$ since $T = o(N)$ by assumption and $\mathbb{E}(s_j^2) \lesssim 1 + \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq 1\}}) \lesssim 1$ 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 $\|\beta\|_{\text{MAX}}^2 \lesssim_P 1$ and $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$. $N^d \lesssim T \lesssim N^{d'}$ with fixed $d > 1/2$ and $d' < 1$. Then it holds that, as $N, T \rightarrow \infty$,

$$P(p(\check{s}_i) \geq T^{1/2} N^{-3/2}, \forall i \leq N) \geq 1 - c_N. \quad (\text{D.13})$$

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

$$\int_{|a| \geq C} \phi_{1/T}(x - a) da \leq \int_{|a| \geq C-1} \sqrt{T} \exp(-Ta^2/2) da \lesssim \exp(-T) \leq c_N N^{-1}. \quad (\text{D.14})$$

The last step comes from $T \gtrsim N^d$ for some $d > 1/2$ by assumption. Then we can bound

$$\begin{aligned} \int_{|a| \geq C} p(a) da &\leq \int_{|x| \geq 1} p_s(x) dx + \int_{|x| < 1} \int_{|a| \geq C} \phi_{1/T}(x-a) da p_s(x) dx \\ &\leq \int_{|x| \geq 1} p_s(x) dx + \sup_{x: |x| < 1} \int_{|a| \geq C} \phi_{1/T}(x-a) da \leq c_N N^{-1}. \end{aligned} \quad (\text{D.15})$$

Here the last inequality comes from (D.14) and $\int_{|x| \geq 1} p_s(x) dx \leq \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq 1\}}) \leq \mathbb{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) \leq c_N N^{-1}$ by condition (a) of Assumption 2. It follows from (D.15) that

$$\begin{aligned} \mathbb{P}(p(\check{s}_i) < T^{1/2} N^{-3/2}) &= \int \mathbb{1}_{\{p(a) < T^{1/2} N^{-3/2}\}} p(a) da \\ &\leq c_N N^{-1} + \int_{|a| < C} \mathbb{1}_{\{p(a) < T^{1/2} N^{-3/2}\}} p(a) da \leq c_N N^{-1}. \end{aligned} \quad (\text{D.16})$$

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 $N \rightarrow \infty$,*

$$|a^j \phi(a) - \bar{a}^j \phi(\bar{a})| \lesssim (\log N)^{(j+1)/2} |\bar{a} - a| \phi(a) + c_N N^{-2}, \quad (\text{D.17})$$

$$\phi(\bar{a}) \lesssim (\log N)^{(j+1)/2} (1 + |a|)^{-(j+1)} (\phi(a) + c_N N^{-2}). \quad (\text{D.18})$$

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 $j \in \{0, 1\}$,

$$|a^j \phi(a) - \bar{a}^j \phi(\bar{a})| \lesssim b_N^{j+1} |\bar{a} - a| \phi(a). \quad (\text{D.19})$$

Moreover, it holds that, for $j \in \{0, 1, 2\}$ and for all b'_N that satisfies $b'_N \geq \sqrt{5 \log N}$,

$$\sup_{a: |a| \geq b'_N} |a^j \phi(a)| \leq c_N N^{-2}. \quad (\text{D.20})$$

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 2 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

$$a^* = \frac{\zeta}{2} + \frac{\zeta^{*2}}{2\zeta} \quad \text{and} \quad \chi(a) = \exp((a - a^*)\zeta). \quad (\text{E.1})$$

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 $\rho \rightarrow 0$ we have, uniformly over $a \geq 0$,

$$T^{-1/2} p(T^{-1/2} a) = (1 + o(1)) \frac{\rho}{2} \phi(a - \zeta) (1 + \chi(a)^{-1}). \quad (\text{E.2})$$

We also note $\phi(a + \zeta) = \exp(-2a\zeta) \phi(a - \zeta)$. Therefore, uniformly over $a \geq 0$,

$$\psi(T^{-1/2} a) = (1 + o(1)) \mu \frac{1 - \exp(-2a\zeta)}{1 + \chi(a)^{-1}}. \quad (\text{E.3})$$

As a result of (E.2) and (E.3), it holds that

$$\int \psi(a)^2 p(a) da = (1 + o(1)) \mu^2 \rho B^*, \quad \text{with} \quad B^* := \int_0^\infty (1 - \exp(-2a\zeta))^2 \frac{\chi(a)}{1 + \chi(a)} \phi(a - \zeta) da. \quad (\text{E.4})$$

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 $j \in \{1, 2\}$,

$$B_j = \int_{A_j} (1 - \exp(-2a\zeta))^2 \frac{\chi(a)}{1 + \chi(a)} \phi(a - \zeta) da, \quad B'_j := \int_{A_j} (1 - \exp(-2a\zeta))^2 \phi(a - 2\zeta) da. \quad (\text{E.5})$$

It apparently holds that $B^* = B_1 + B_2$. Furthermore, from the first part of (E.1), we note $\zeta a^* \rightarrow \infty$ and thereby $\sup_{a \geq a^*} \exp(-2a\zeta) \rightarrow 0$. 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

$$B_1 \approx \int_0^{a^*} (1 - \exp(-2a\zeta))^2 \chi(a) \phi(a - \zeta) da, \quad B_2 \approx \int_{a^*}^\infty \phi(a - \zeta) da. \quad (\text{E.6})$$

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,

$$B_1 \approx \iota B'_1, \quad B_2 \lesssim \iota B'_2. \quad (\text{E.7})$$

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 $\inf_{a \geq a^*} a\zeta \rightarrow \infty$ 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 $\zeta \rightarrow \infty$ and $\zeta - \zeta^* \lesssim -\sqrt{\log N}$. We start with case (ia) where $\zeta \lesssim 1$. We note

$$\phi(a - 2\zeta) \exp(-2a\zeta) = \phi(a) \exp(-2\zeta^2), \quad \phi(a - 2\zeta) \exp(-4a\zeta) = \phi(a + 2\zeta). \quad (\text{E.8})$$

Using (E.8), we obtain

$$B'_1 + B'_2 = \int (1 - \exp(-2a\zeta))^2 \phi(a - 2\zeta) da = 1 - \exp(-2\zeta^2). \quad (\text{E.9})$$

Moreover, it apparently holds

$$B'_2 \leq \int_{A_2} \phi(a - 2\zeta) da = \Phi(2\zeta - a^*). \quad (\text{E.10})$$

On the other hand, when $\zeta \lesssim 1$, we have $a^* - 2\zeta \geq \zeta^{*2}/(3\zeta)$ for N sufficiently large. Then, when $\zeta \lesssim 1$

$$\zeta^{-2}\Phi(2\zeta - a^*) \lesssim \frac{1}{\zeta^{*4}} \frac{\zeta^{*2}}{\zeta} \exp(-\zeta^{*4}/(18\zeta^2)) \rightarrow 0. \quad (\text{E.11})$$

where the last convergence comes from $\zeta^* \rightarrow \infty$ (by the assumption $\rho \leq N^{-d}$ with fixed $d > 0$), $\zeta^{*2}/\zeta \rightarrow \infty$ (as $\zeta \lesssim 1$), and $\lim_{x \rightarrow \infty} x \exp(-x^2) = 0$. 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 $\zeta \lesssim 1$,

$$B'_1 \approx \zeta^2, \quad B'_2 = o(\zeta^2). \quad (\text{E.12})$$

Substituting (E.12) into (E.7), we obtain (recall $B^* = B_1 + B_2$) that, when $\zeta \lesssim 1$,

$$B^* \approx \zeta^2 e^{-\zeta a^*}. \quad (\text{E.13})$$

(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 $\zeta \rightarrow \infty$. Using (E.8) and $\Phi(x) = 1 - \Phi(-x)$, we obtain by direct calculation

$$B'_1 = \int_0^{a^*} \phi(a - 2\zeta) da = \Phi(a^* - 2\zeta) - 2\Phi(-2\zeta) + \Phi(-a^* - 2\zeta) - \exp(-2\zeta^2)(2\Phi(a^*) - 1). \quad (\text{E.14})$$

We note that, uniformly over x ,

$$\Phi(-x) \approx \frac{1}{1 + x_+} \phi(x_+), \quad (\text{E.15})$$

where $x_+ = \max\{x, 0\}$. Using (E.15), we obtain $\Phi(a^* - 2\zeta) \geq \Phi(-3\zeta/2) \approx \frac{1}{1+\zeta} \phi(3\zeta/2) \approx \frac{1}{1+\zeta} \exp(-9\zeta^2/8)$. Then we have, $\Phi(-2\zeta) = o(\Phi(a^* - 2\zeta))$ and $\exp(-2\zeta^2) = o(\Phi(a^* - 2\zeta))$ when $\zeta \rightarrow \infty$. Since $\Phi(-a^* - 2\zeta) \leq \Phi(-2\zeta)$ and $0 \leq \Phi(a^*) \leq 1$, we obtain from (E.14) that, when $\zeta \rightarrow \infty$,

$$B'_1 = (1 + o(1))\Phi(a^* - 2\zeta). \quad (\text{E.16})$$

Substituting (E.16) into the first part of (E.7), and noting the second part of (E.6), we obtain that, when $\zeta \rightarrow \infty$,

$$B_1 \approx \iota \Phi(a^* - 2\zeta), \quad B_2 \approx \Phi(\zeta - a^*). \quad (\text{E.17})$$

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),

$$B_1 \gtrsim \frac{1}{1 + (2\zeta - a^*)_+} \phi(a^* - \zeta), \quad B_2 \approx \frac{1}{1 + (a^* - \zeta)_+} \phi(a^* - \zeta). \quad (\text{E.18})$$

Here the first part comes from $\phi(x_+) \geq \phi(x)$ 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),

$$B_2 \lesssim B_1. \quad (\text{E.19})$$

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),

$$B_1 \approx \frac{1}{1 + (2\zeta - a^*)_+} \phi(a^* - \zeta), \quad B_2 \gtrsim \frac{1}{1 + (a^* - \zeta)_+} \phi(a^* - \zeta). \quad (\text{E.20})$$

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),

$$B_1 = o(B_2). \quad (\text{E.21})$$

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 $h_N : R \rightarrow R$, a sequence a_N , and a positive sequence Δ_N satisfy $\Delta_N^{-1/2}(\mathbb{1}_{A_N} - h_N) \rightarrow 0$ in L^2 with $A_N := (a_N, a_N + \Delta_N)$. Then there exists a set sequence \bar{A}_N such that*

$$\int_{A_N - \bar{A}_N} da = o(\Delta_N) \quad \text{and} \quad \sup_{a \in A'_N} |h_N(a) - 1| = o(1).$$

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 $(\mathbb{1}_{A_N^*} - h_N^*) \rightarrow 0$ in L^2 . This further leads to $h_N^* \rightarrow \mathbb{1}_{A_N^*}$ in measure (see, e.g., Theorem 2.15.a of Folland (2009)), i.e., for every $\epsilon > 0$, $\int_{A_N^*} \mathbb{1}_{\{|1 - h_N^*(a)| \geq \epsilon\}} da \rightarrow 0$. This is equivalent to that, for every $\epsilon > 0$,

$$\Delta_N^{-1} \int_{A_N} \mathbb{1}_{\{|1 - h_N(a)| \geq \epsilon\}} da \rightarrow 0. \quad (\text{E.22})$$

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 $\text{Corr}(\psi_i, \tilde{\psi}_{q,i}(\lambda)) \rightarrow 1$ if and only if, for some $c_N \rightarrow 0$,*

$$\begin{cases} (\mu, \rho) \in \{(\mu, \rho) : \zeta \leq c_N \text{ or } \zeta - a^* \geq c_N^{-1}\} & \text{for } q = 1, \\ (\mu, \rho) \in \{(\mu, \rho) : \zeta \leq c_N \text{ or } \zeta - a^* \geq -c_N \sqrt{\log N}\} & \text{for } q = 2. \end{cases} \quad (\text{E.23})$$

Moreover, if sequence $\lambda \geq 0$ is deterministic, $\text{Corr}(\psi_i, \tilde{\psi}_{q,i}(\lambda)) \rightarrow 1$ if and only if

$$\begin{cases} \lambda \rightarrow 0, & \text{when } \zeta \rightarrow 0 \text{ and } q \in \{1, 2\}, \\ \zeta - \lambda \rightarrow \infty \text{ and } \frac{\phi(\lambda)}{\rho(1+\lambda^2)(\zeta-\lambda)^2} \rightarrow 0, & \text{when } \zeta - a^* \rightarrow \infty \text{ and } q = 1. \end{cases} \quad (\text{E.24})$$

Further, when $\zeta - a^* \geq -c_N \sqrt{\log N}$, it holds that

$$\text{Corr}(\psi_i, \tilde{\psi}_{q,i}(\lambda)) \geq 1 + o(1) + O\left(\frac{\Phi(-\lambda)}{\rho\Phi(\zeta - a^*)}\right). \quad (\text{E.25})$$

Proof. Step 1. As a result of (E.2) and (E.3), it holds that, by the symmetry of $\psi(a)$, $\tilde{\psi}_q(a, \lambda)$, and $p(a)$,

$$\int \psi(a)^2 p(a) da = (1 + o(1)) \mu^2 \rho \int_0^\infty (1 - \exp(-2a\zeta))^2 \frac{\chi(a)}{1 + \chi(a)} \phi(a - \zeta) da, \quad (\text{E.26})$$

$$\int \tilde{\psi}_q(a, \lambda)^2 p(a) da = (1 + o(1)) \rho \int_0^\infty \tilde{\psi}_q(a, \lambda)^2 \frac{1 + \chi(a)}{\chi(a)} \phi(a - \zeta) da, \quad (\text{E.27})$$

$$\int \psi(a) \tilde{\psi}_q(a, \lambda) p(a) da = (1 + o(1)) \mu \rho \int_0^\infty \tilde{\psi}_q(a, \lambda) (1 - \exp(-2a\zeta)) \phi(a - \zeta) da. \quad (\text{E.28})$$

To facilitate the exposition, we introduce short-hand notation

$$f(a) := \sqrt{\frac{\chi(a)}{1 + \chi(a)}} \phi(a - \zeta) \mathbb{1}_{\{a \geq 0\}}, \quad g(a) = (1 - e^{-2a\zeta}) f(a), \quad \tilde{g}_q(a) = \tilde{\psi}_q(a, \lambda) \frac{1 + \chi(a)}{\chi(a)} f(a).$$

Comparing the definitions of f and g with the right-hand sides of (E.26), (E.27), and (E.28), we obtain, for $q \in \{1, 2\}$,

$$\text{Corr}(\psi_i, \tilde{\psi}_{q,i}(\lambda)) = (1 + o(1)) \theta_q(\lambda), \quad \theta_q(\lambda) := \frac{\langle g, \tilde{g}_q \rangle}{\|g\| \|\tilde{g}_q\|}, \quad (\text{E.29})$$

where $\langle g, \tilde{g}_q \rangle := \int g(a) \tilde{g}_q(a) da$ and $\|g\| := \sqrt{\langle g, g \rangle}$ stand for the L^2 -inner product and L^2 -norm. Moreover, we note that $1 - \theta_q(\lambda) = \frac{1}{2} \frac{1}{\|g\|^2} \left\| g - \frac{\|g\|}{\|\tilde{g}_q\|} \times \tilde{g}_q \right\|^2$. Let d^* be the scalar that minimizes $\|g - d \times \tilde{g}_q\|$. It obviously holds that, for $q \in \{1, 2\}$,

$$1 - \theta_q(\lambda) \geq \frac{1}{2\|g\|^2} \|g - d^* \times \tilde{g}_q\|^2. \quad (\text{E.30})$$

Moreover, we have that, if $\theta_q(\lambda) \geq 0$ and for $q \in \{1, 2\}$,

$$1 - \theta_q(\lambda) \leq 1 - \theta_q(\lambda)^2 = \frac{1}{\|g\|^2} \|g - d^* \times \tilde{g}_q\|^2. \quad (\text{E.31})$$

Given (E.29), (E.30), and (E.31), the proof will focus on evaluating $\|g - d^* \times \tilde{g}_q\|$. In addition, given the definition of $\tilde{\psi}(a, \lambda)$, we can write, for $q \in \{1, 2\}$,

$$\|g - d^* \times \tilde{g}_q\| = \int (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 f(a)^2 da, \quad \text{with} \quad h_q(a) = (e^{(a^* - a)\zeta} + 1) \tilde{\psi}_q(a, \lambda). \quad (\text{E.32})$$

Further, to characterize the magnitude of $\|g\|$, we consider, respectively, case (ia) where $\zeta \rightarrow 0$ (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

$$\|g\|^2 \approx \begin{cases} (1 \wedge \zeta^2) \iota \Phi(a^* - 2\zeta), & \text{cases (ia) and (ib);} \\ \Phi(\zeta - a^*), & \text{case (ii).} \end{cases} \quad (\text{E.33})$$

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 $q \in \{0, 1\}$, $\theta_q(\lambda) \rightarrow 1$ holds in case (ia) if and only if we choose $\lambda \rightarrow 0$. Suppose $\lambda \rightarrow 0$ and let $\bar{d}^* = 2\zeta e^{-a^*\zeta}$. We obtain that, for each $q \in \{1, 2\}$ and for all $a \geq 0$,

$$\begin{aligned} |1 - e^{-2a\zeta} - \bar{d}^* \times h_q(a)| &= |1 - e^{-2a\zeta} - 2\zeta \psi_q(a, \lambda)(e^{-a\zeta} + e^{-a^*\zeta})| \\ &\leq |1 - e^{-2a\zeta} - 2\zeta a| + 2\zeta \lambda + 2\zeta a(|1 - e^{-a\zeta}| + e^{-a^*\zeta}) \\ &\leq 4\zeta^2 a^2 + 2\zeta \lambda + 2e^{-a^*\zeta} \zeta a. \end{aligned} \quad (\text{E.34})$$

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 $q \in \{1, 2\}$,

$$\begin{aligned} \|g - d^* \times \tilde{g}_q\|^2 &\leq \|g - \bar{d}^* \times \tilde{g}_q\|^2 \lesssim \iota \int_0^\infty (\zeta^4 a^4 + \zeta^2 \lambda^2 + e^{-2a^*\zeta} \zeta^2 a^2) \phi(a - 2\zeta) da \\ &\lesssim \iota (\zeta^4 + \zeta^2 \lambda^2 + e^{-2a^*\zeta} \zeta^2) = o(\iota \zeta^2). \end{aligned} \quad (\text{E.35})$$

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 $\zeta \rightarrow 0$. The last inequality comes from $\zeta \rightarrow 0$, $a^* \zeta \rightarrow \infty$, and $\lambda \rightarrow 0$.

Now suppose $\lambda \gtrsim 1$. Then it holds that, for each $q \in \{1, 2\}$,

$$\|g - d^* \times \tilde{g}_q\|^2 \geq \int_0^{\lambda \wedge 1} (1 - e^{-2a\zeta})^2 f(a)^2 da \gtrsim \iota \zeta^2 \int_0^{\lambda \wedge 1} a^2 \phi(a - 2\zeta) \gtrsim \iota \zeta^2. \quad (\text{E.36})$$

The first inequality comes from $\tilde{g}_q(a) = 0$ 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 $\|g\|^2 \asymp \iota \zeta^2$ as $\Phi(a^* - 2\zeta) \asymp 1$ (by the first part of (E.1)) when $\zeta \rightarrow 0$. Given (E.29), we prove that for each $q \in \{1, 2\}$, $\text{Corr}(\psi_i, \tilde{\psi}_{q,i}(\lambda)) \rightarrow 1$ if (by (E.31) and (E.35)) and only if (by (E.30) and (E.36)) $\lambda \rightarrow 0$ when we are in case (ia), i.e., when we have $\zeta \rightarrow 0$. We hence establish the first part of (E.24).

Step 3. Next, we show that, for each $q \in \{1, 2\}$, $\theta_q(\lambda) \rightarrow 1$ 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 ,

$$\phi(x) \lesssim \inf_{a: |a-x| \leq \frac{C}{1+|x|}} \phi(a) \leq \sup_{a: |a-x| \leq \frac{C}{1+|x|}} \phi(a) \lesssim \phi(x). \quad (\text{E.37})$$

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 $q \in \{1, 2\}$,

$$\|g - d^* \times \tilde{g}_q\|^2 \gtrsim \iota \phi((2\zeta) \wedge a^* - 2\zeta) \int_A (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 da. \quad (\text{E.38})$$

According to (E.33) and (E.15), $\|g\|^2 \asymp \iota \phi((2\zeta) \wedge a^* - 2\zeta) \Delta$ in case (ib). Then, combining (E.30) and (E.38), we obtain that, if $\theta(\lambda) \rightarrow 1$, then, for $q \in \{1, 2\}$,

$$\int_A (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 da = o(\Delta). \quad (\text{E.39})$$

It follows from Lemma E2 and (E.39) that, for each $q \in \{1, 2\}$, if $\theta_q(\lambda) \rightarrow 1$, there exists a set sequence \bar{A} such that,

$$\int_{A-\bar{A}} da = o(\Delta) \quad \text{and} \quad \sup_{a \in \bar{A}} |d^* \times h_q(a) - (1 - e^{-2a\zeta})| = o(1). \quad (\text{E.40})$$

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), $h_q(a) = (1 + o(1))e^{a^*\zeta}\tilde{\psi}_q(a, \lambda)$. Apply the triangular inequality to (E.40) leads to $\sup_{a \in \bar{A}} |d^*e^{a^*\zeta} \times \tilde{\psi}_q(a, \lambda) - (1 - e^{-2a\zeta})| = o(1)$, 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 $\zeta \approx 1$.

Then we consider the situation where $\zeta \rightarrow \infty$. It is trivial from (E.39) and that $\tilde{\psi}_q(a, \lambda) = 0$ for $a \leq \lambda$ that $\lambda \leq (2\zeta) \wedge a^* - 7\Delta/2$ for large N if $\theta_q(\lambda) \rightarrow 1$, 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 $\theta_q(\lambda) \rightarrow 1$ and $\zeta \rightarrow \infty$,

$$\frac{\inf_{a \in A^+} h_q(a)}{\sup_{a \in A^-} h_q(a)} \gtrsim \exp(\zeta\Delta) \rightarrow \infty, \quad (\text{E.41})$$

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, $\theta_q(\lambda) \rightarrow 1$ 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 $q \in \{1, 2\}$, $\theta_q(\lambda) \rightarrow 1$ does not hold for any λ sequence if we are in case (ii) and $\Delta^{-1}(a^* - \lambda) \rightarrow \infty$. 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 $q \in \{1, 2\}$,

$$\|g - d^* \times \tilde{g}_q\|^2 \gtrsim \phi((a^* - \zeta)_+) \int_{A^*} (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 da. \quad (\text{E.42})$$

According to (E.33) and (E.15), $\|g\|^2 \approx \phi((a^* - \zeta)_+)\Delta$. Then, combining (E.30) and (E.42), we obtain that, for $q \in \{1, 2\}$, if $\theta_q(\lambda) \rightarrow 1$,

$$\int_{A^*} (1 - e^{-2a\zeta} - d^* \times h_1(a))^2 da = o(\Delta). \quad (\text{E.43})$$

Moreover, it follows from Lemma E2 and (E.43) that, for $q \in \{1, 2\}$, if $\theta_q(\lambda) \rightarrow 1$, there exists a set sequence \bar{A} such that

$$\int_{A^* - \bar{A}} da = o(\Delta) \quad \text{and} \quad \sup_{a \in \bar{A}} |d^* \times h_q(a) - 1| = o(1). \quad (\text{E.44})$$

Here we also use $e^{-a^*\zeta} \rightarrow 0$ (by the first part of (E.1) and $\zeta^* \approx \sqrt{\log N} \rightarrow \infty$ due to the assumption on ρ). Now we prove by contradiction that, for each $q \in \{1, 2\}$, $\theta_q(\lambda) \rightarrow 1$ can not hold if $\Delta^{-1}(a^* - \lambda) \rightarrow \infty$. For this purpose, we introduce a set sequence $A' := (a^* - 2\Delta, a^* - \Delta)$.

Suppose $\Delta^{-1}(a^* - \lambda) \rightarrow \infty$ holds. Then we obtain that, for $q \in \{1, 2\}$, if $\theta_q(\lambda) \rightarrow 1$ and for all $a \in A^*$ and $a' \in A'$,

$$h_q(a') = (1 + o(1)) \frac{\exp((a^* - a')\zeta) + 1}{\exp((a^* - a)\zeta) + 1} \times h_q(a) = (1 + o(1))\chi(a')^{-1} \times h_q(a). \quad (\text{E.45})$$

The first equality comes from $\Delta^{-1}(a^* - \lambda) \rightarrow \infty$ and the definition of $\tilde{\psi}_q$. The last equality comes from the definitions of A^* and A' and that $\exp(\zeta\Delta) \rightarrow \infty$, 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 $a^* - \zeta \asymp \zeta^* - \zeta \geq -c_N\sqrt{\log N}$. Combining the second part of (E.44) and (E.45), we obtain that, for $q \in \{1, 2\}$, if $\theta_q(\lambda) \rightarrow 1$ and for all $a \in A'$,

$$d^* \times h_q(a) = (1 + o(1))\chi(a)^{-1}. \quad (\text{E.46})$$

Since $\inf_{a \in A'} \chi(a)^{-1} \rightarrow \infty$ due to $\exp(\zeta\Delta) \rightarrow \infty$, it follows from (E.46) that, for $q \in \{1, 2\}$,

$$\begin{aligned} \|g - d^* \times \tilde{g}_q\|^2 &\geq \int_{A'} (1 - e^{-2a\zeta} - d^* \times h_q(a))^2 f(a)^2 da \\ &\gtrsim \int_{A'} \chi(a)^{-1} \phi(a - \zeta) da \gtrsim \Delta \inf_{a \in A'} \phi(a - \zeta) \gtrsim \|g\|^2, \end{aligned}$$

where we note that $\inf_{a \in A'} \chi(a)^{-1} \rightarrow \infty$ (by $\chi(a^*) = 1$ and $\exp(\zeta\Delta) \rightarrow \infty$) and $\inf_{a \in A'} \phi(a - \zeta) \gtrsim \phi((a^* - \zeta)_+)$, and we recall $\|g\|^2 \asymp \phi((a^* - \zeta)_+)\Delta$. This contradicts (E.30) and proves that, for $q \in \{1, 2\}$, $\theta_q(\lambda) \rightarrow 1$ can not hold if we are in case (ii) and $\Delta^{-1}(a^* - \lambda) \rightarrow \infty$.

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 previous 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 $\theta_1(\lambda) \rightarrow 1$ if and only if $\zeta - a^* \rightarrow \infty$.

We first show the “only if”. By definition of A^* (see above (E.42)), it holds that $\sup_{a \in A^*} \exp((a^* - a)\zeta) \leq \exp(-\zeta\Delta) \rightarrow 0$. 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 $\theta_1(\lambda) \rightarrow 1$,

$$\Delta^{-1}(a^* \vee \zeta - \lambda) \rightarrow \infty. \quad (\text{E.47})$$

However, according to Step 4 and the subsequence argument, $\theta_1(\lambda) \rightarrow 1$ also leads to $\Delta^{-1}(a^* - \lambda) \lesssim 1$. This bound and (E.47) can simultaneously hold only if $\zeta - a^* \rightarrow \infty$.

Now we show the “if” part. We note

$$\|f - g\|^2 \leq \int_0^\infty e^{-2a\zeta} \phi(a - \zeta) da = \Phi(-\zeta) \rightarrow 0, \quad (\text{E.48})$$

where the first inequality comes from $f(a)^2 \leq \phi(a - \zeta)$. Moreover, letting $\bar{d}^* = (\zeta - \lambda)^{-1}$, we obtain

$$\begin{aligned} & \|f - \bar{d}^* \times \tilde{g}_1\|^2 \\ &= \int_0^\lambda f(a)^2 da + \int_\lambda^\infty \left(1 - (e^{(a^*-a)\zeta} + 1) \times \frac{a - \lambda}{\zeta - \lambda}\right)^2 f(a)^2 da \\ &\leq \int_0^\lambda f(a)^2 da + 2 \int_\lambda^\infty \left(\frac{\zeta - a}{\zeta - \lambda}\right)^2 f(a)^2 da + 2 \int_\lambda^\infty e^{2(a^*-a)\zeta} \left(\frac{a - \lambda}{\zeta - \lambda}\right)^2 f(a)^2 da \\ &\leq \Phi(\lambda - \zeta) + \frac{2}{(\zeta - a^*)^2} + \frac{4}{(\zeta - \lambda)^2 \rho} \int_\lambda^\infty \phi(a)(a - \lambda)^2 da. \end{aligned} \quad (\text{E.49})$$

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 $\Phi(\lambda - \zeta) \rightarrow 0$ if $\lambda \leq a^*$, as $\zeta - a^* \rightarrow \infty$. Hence, given (E.49), and noting $\int_\lambda^\infty \phi(a)(a - \lambda)^2 da \approx \frac{1}{1+\lambda^2} \phi(\lambda)$, we have that $\|f - \bar{d}^* \times \tilde{g}_1\| \rightarrow 0$, when λ satisfies the second part of (E.24). (Such λ always exists, because $\lambda = a^*$ satisfy the condition as $\zeta - a^* \rightarrow \infty$ and $\rho^{-1} \phi(a^*) = \phi(a^* - \zeta)/2 \rightarrow 0$.) Given (E.48), and noting $\|g\|^2 \approx 1$ from (E.33) (note $\zeta - a^* \rightarrow \infty$ can only occur in case (ii) by the first part of (E.1)), we obtain $\|g - \bar{d}^* \times \tilde{g}_1\| = o(\|g\|)$. Since $\|g - d^* \times \tilde{g}_1\| \leq \|g - \bar{d}^* \times \tilde{g}_1\|$ 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 $\zeta - \lambda \lesssim 1$, it holds that, for all d ,

$$\|f - d \times \tilde{g}_1\|^2 \geq \int_{a^*}^\lambda f(a)^2 da \geq \frac{1}{2}(\Phi(\lambda - \zeta) - \Phi(a^* - \zeta)) \gtrsim 1. \quad (\text{E.50})$$

The first inequality comes from $\lambda \geq a^*$ as $\zeta - a^* \rightarrow \infty$, 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 $\zeta - \lambda \lesssim 1$ and $\zeta - a^* \rightarrow \infty$. Combining (E.48) and (E.50), and recalling (E.31) and that $\|g\|^2 \approx 1$ from (E.33) under $\zeta - a^* \rightarrow \infty$, we obtain that the condition $\zeta - \lambda \rightarrow \infty$ is necessary.

Now suppose $\zeta - \lambda \rightarrow \infty$ holds but $\rho^{-1} \phi(\lambda) \gtrsim (1 + \lambda^2)(\zeta - \lambda)^2$. This indicates $\rho^{-1} \phi(\lambda) \rightarrow \infty$. Since $\rho^{-1} \phi(a^*) \rightarrow 0$ (see after (E.49)), we have $(a^* - \lambda)/(1 + \lambda) \rightarrow \infty$. Therefore, letting $l(a) = \mathbb{1}_{\{\lambda \leq a \leq a^*\}}$, we have, for all d ,

$$\|\tilde{g}_1 \times l\|^2 = \int_{\lambda}^{a^*} (e^{(a^*-a)\zeta} + 1)^2 (a - \lambda)^2 f(a)^2 da \geq \frac{2}{\rho} \int_{\lambda}^{a^*} (a - \lambda)^2 \phi(a) da \gtrsim \frac{\phi(\lambda)}{\rho(1 + \lambda^2)} \gtrsim (\zeta - \lambda)^2. \quad (\text{E.51})$$

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 $(a^* - \lambda)/(1 + \lambda) \rightarrow \infty$. The last inequality comes from $\rho^{-1}\phi(\lambda) \gtrsim (1 + \lambda^2)(\zeta - \lambda)^2$ that we suppose. Since $\|f \times l\|^2 = \int_{\lambda}^{a^*} f(a)^2 da \leq \Phi(a^* - \zeta) \rightarrow 0$, given (E.48) and (E.51), we obtain that, for all d ,

$$\|g - d \times \tilde{g}_1\| = \|f - d \times \tilde{g}_1\| + o(1) \geq \|(f - d \times \tilde{g}_1) \times l\| + o(1) \gtrsim d \times (\zeta - \lambda) + o(1). \quad (\text{E.52})$$

Given (E.52), and recalling $\|g\|^2 \approx 1$ from (E.33), $\|g - d \times \tilde{g}_1\| = o(1)$ can hold only if $d = o((\zeta - \lambda)^{-1})$. However, for all $d \leq \frac{1}{4}(\zeta - \lambda)^{-1}$, it holds that

$$\|f - d \times \tilde{g}_1\|^2 \geq \int_{a^*}^{\zeta} (1 - d \times (e^{(a^*-a)\zeta} + 1)(a - \lambda))^2 f(a)^2 da \geq \frac{1}{4} \int_{a^*}^{\zeta} f(a)^2 da \gtrsim \frac{1}{2} - \Phi(a^* - \zeta) \gtrsim 1. \quad (\text{E.53})$$

The second inequality comes from that $e^{(a^*-a)\zeta} + 1 \leq 2$ and $a - \lambda \leq \zeta - \lambda$ for all $a \in (a^*, \zeta)$, the third 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 $\zeta - a^* \rightarrow \infty$. Given (E.48), (E.53) contradicts $\|g - d \times \tilde{g}_1\| = o(1)$. 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 previous analysis for cases (ia) and (ib), we only need to show that, in case (ii), $\theta_2(\lambda) \rightarrow 1$ for some sequence λ . Under $\lambda = a^*$ and $\bar{d}^* = \zeta^{-1}$, we obtain

$$\begin{aligned} \|f - \bar{d}^* \times \tilde{g}_2\|^2 &= \int_0^{a^*} f(a)^2 da + \int_{a^*}^{\infty} \left(1 - (e^{(a^*-a)\zeta} + 1) \times \frac{a}{\zeta}\right)^2 f(a)^2 da \\ &\leq \int_0^{a^*} f(a)^2 da + 2 \int_{a^*}^{\infty} \left(\frac{a - \zeta}{\zeta}\right)^2 f(a)^2 da + 2 \int_{a^*}^{\infty} e^{2(a^*-a)\zeta} \frac{a^2}{\zeta^2} f(a)^2 da \\ &\leq \iota\Phi(2\zeta - a^*) + \frac{2}{\zeta^2} \int_{a^*-\zeta}^{\infty} a^2 \phi(a) da + \frac{2}{\zeta^2} e^{2a^*\zeta} \int_{a^*}^{\infty} (a + \zeta)^2 \phi(a + \zeta) da \\ &\lesssim \iota\Phi(2\zeta - a^*) + \frac{2}{\zeta^2} (1 + (a^* - \zeta)_+)^2 \Phi(\zeta - a^*) \\ &\quad + \frac{2}{\zeta^2} (1 + (a^* + \zeta))^2 e^{2a^*\zeta} \Phi(-a^* - \zeta). \end{aligned} \quad (\text{E.54})$$

Here the equality comes from (E.32) and the first inequality is obvious. For the second inequality we use $f(a)^2 \leq \chi(a)\phi(a - \zeta) = \iota\phi(a - 2\zeta)$, $f(a)^2 \leq \phi(a - \zeta)$, and $e^{-2a\zeta}\phi(a - \zeta) = \phi(a + \zeta)$. The last inequality comes from that $\int_x^{\infty} a^2 \phi(a) da \approx (1 + x_+)^2 \Phi(-x)$ uniformly

over x . Moreover, we note $(1 + (a^* + \zeta))e^{2a^*\zeta}\Phi(-a^* - \zeta) \asymp e^{2a^*\zeta}\phi(a^* + \zeta) = \phi(a^* - \zeta) \asymp (1 + (a^* - \zeta)_+)\Phi(\zeta - a^*)$ 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 $\zeta^* \asymp \sqrt{\log N}$ due to the assumption on ρ). Substituting these results into (E.54), we have that, under $\lambda = a^*$ and $\bar{d}^* = \zeta^{-1}$,

$$\|f - \bar{d}^* \times \tilde{g}_2\|^2 = o(\Phi(\zeta - a^*)). \quad (\text{E.55})$$

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 $\|g\|^2 \asymp \Phi(\zeta - a^*)$ in case (ii) according to (E.33), we finally obtain that, under $\lambda = a^*$ and $\bar{d}^* = \zeta^{-1}$,

$$\|g - \bar{d}^* \times \tilde{g}_2\| = o(\|g\|). \quad (\text{E.56})$$

Since $\|g - d^* \times \tilde{g}_1\| \leq \|g - \bar{d}^* \times \tilde{g}_1\|$ by definition of d^* , we prove $\theta_2(\lambda) \rightarrow 1$ in case (ii) for some sequence λ . (E.23) is completely established.

Next, it follows that, under $\lambda \leq a^*$ and letting $\bar{d}^* = \zeta^{-1}$,

$$\begin{aligned} & \|g - d^* \times \tilde{g}_2\|^2 \\ \leq & \|g - \bar{d}^* \times \tilde{g}_2\|^2 \leq o(\|g\|^2) + 2\zeta^{-2} \int_{\lambda}^{a^*} a^2 \frac{1 + \chi(a)}{\chi(a)} \phi(a - \zeta) da \\ \leq & o(\|g\|^2) + \frac{8}{\zeta^2 \rho} \int_{\lambda}^{a^*} a^2 \phi(a) da \leq o(\|g\|^2) + \frac{8a^{*2}}{\zeta^2 \rho} \Phi(-\lambda) \leq o(\|g\|^2) + O(\rho^{-1} \Phi(-\lambda)). \end{aligned} \quad (\text{E.57})$$

The second inequality comes from (E.56), the triangle inequality, the definition of $\tilde{g}_2(a)$ given after (E.28), and that $\tilde{\psi}_2(a, \lambda) - \tilde{\psi}_2(a, a^*) = a \mathbb{1}_{\{\lambda \leq a \leq a^*\}}$ 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 $\rho N \Phi(\zeta - \lambda) \rightarrow \infty$, it holds that, for $q \in \{1, 2\}$,*

$$\|\tilde{\psi}_q(\lambda)\|^2 = (1 + o_P(1)) N E(\tilde{\psi}_{q,i}(\lambda)^2).$$

Here $\tilde{\psi}_q(\lambda)$ stands for the N -dimensional vector whose components are $\tilde{\psi}_{q,i}(\lambda)$.

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 $E(\tilde{\psi}_{q,i}(\lambda)^2)$ and $E(\tilde{\psi}_{q,i}(\lambda)^4)$, and then establish the probability limit using Chebyshev's inequality. It holds by the definition of $\tilde{\psi}_{q,i}$

and (E.2) (the condition $\rho \rightarrow 0$ it relies on is assumed here too) that, for $q \in \{1, 2\}$,

$$\mathbb{E}(\tilde{\psi}_{q,i}(\lambda)^{2q}) = 2(1 + o(1)) \int_0^\infty \tilde{\psi}_q(a, \lambda)^{2q} \left(\phi(a) + \frac{\rho}{2} \phi(a - \zeta) \right) da. \quad (\text{E.59})$$

On the other hand, we can calculate, for $j \in \{2, 4\}$,

$$\int_0^\infty \tilde{\psi}_1(a, \lambda)^j \phi(a) da = \int_0^\infty a^j \phi(\lambda + a) da \approx \iota_{1,1}^j \Phi(-\lambda), \quad (\text{E.60})$$

$$\int_0^\infty \tilde{\psi}_1(a, \lambda)^j \phi(a - \zeta) da = \int_0^\infty a^j \phi(\lambda - \zeta + a) da \approx \iota_{1,2}^j \Phi(\zeta - \lambda), \quad (\text{E.61})$$

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 $j \in \{2, 4\}$,

$$\int_\lambda^\infty \tilde{\psi}_2(a, \lambda)^j \phi(a) da = \int_0^\infty (\lambda + a)^j \phi(\lambda + a) da \approx \iota_{2,1}^j \Phi(-\lambda), \quad (\text{E.62})$$

$$\int_\lambda^\infty \tilde{\psi}_2(a, \lambda)^j \phi(a - \zeta) da = \int_0^\infty (\lambda + a)^j \phi(\lambda - \zeta + a) da \approx \iota_{2,2}^j \Phi(\zeta - \lambda). \quad (\text{E.63})$$

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 $q \in \{1, 2\}$,

$$\begin{aligned} \frac{\mathbb{E}(\tilde{\psi}_{q,i}(\lambda)^4)}{\mathbb{E}(\tilde{\psi}_{q,i}(\lambda)^2)^2} &\lesssim \frac{\iota_{q,1}^4 \Phi(-\lambda) + \iota_{q,2}^4 \rho \Phi(\zeta - \lambda)}{\iota_{q,1}^4 \Phi(-\lambda)^2 + \iota_{q,2}^4 \rho^2 \Phi(\zeta - \lambda)^2} \leq \iota_{q,2}^2 \frac{\iota_{q,1}^2 \Phi(-\lambda) + \iota_{q,2}^2 \rho \Phi(\zeta - \lambda)}{\iota_{q,1}^4 \Phi(-\lambda)^2 + \iota_{q,2}^4 \rho^2 \Phi(\zeta - \lambda)^2} \\ &\leq \frac{2}{\iota_{q,2}^2 \iota_{q,1}^2 \Phi(-\lambda) + \iota_{q,2}^2 \rho \Phi(\zeta - \lambda)} \leq \frac{2}{\rho \Phi(\zeta - \lambda)}. \end{aligned}$$

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 $\iota_{q,1} \leq \iota_{q,2}$ for $q \in \{1, 2\}$. Since $\rho N \Phi(\zeta - \lambda) \rightarrow \infty$ by assumption, we have, for $q \in \{1, 2\}$,

$$\frac{\text{Var}(\|\tilde{\psi}_q\|^2)}{(\mathbb{E}(\|\tilde{\psi}_q\|^2))^2} \leq \frac{\mathbb{E}(\tilde{\psi}_{q,i}^4)}{N \mathbb{E}(\tilde{\psi}_{q,i}^2)^2} \lesssim \frac{1}{N \rho \Phi(\zeta - \lambda)} \rightarrow 0. \quad (\text{E.64})$$

Here we suppress the argument λ of $\tilde{\psi}_{q,i}$ for simplicity and the first inequality comes from that $\tilde{\psi}_{q,i}$ 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 $\psi_{\Delta,1}(a, \lambda, b) := b \mathbb{1}_{\{|a|-\lambda \geq -b\}}$ and $\psi_{\Delta,2}(a, \lambda, b) := \lambda \mathbb{1}_{\{||a|-\lambda| \leq b\}}$. Also suppose that λ_N and b_N are two deterministic positive sequences satisfying $\rho N \Phi(\zeta - \lambda_N) \rightarrow \infty$ and $b_N = o(1/\sqrt{\log N})$. Then it holds that, for*

$q \in \{1, 2\}$,

$$\mathbb{E}(|\psi_{\Delta,q}(\tilde{z}_i, \lambda_N, b_N)|^2) = o(\mathbb{E}(\tilde{\psi}_q(\tilde{z}_i, \lambda_N)^2)).$$

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 $\rho \rightarrow 0$ it relies on is assumed here too), we have, for $q \in \{1, 2\}$,

$$\mathbb{E}(\psi_{\Delta,q}(\tilde{z}_i)^2) = 2(1 + o(1)) \int_0^\infty \psi_{\Delta,q}(a)^2 \left(\phi(a) + \frac{\rho}{2} \phi(a - \zeta) \right) da. \quad (\text{E.65})$$

To evaluate the right-hand side of (E.65), we write

$$\int \psi_{\Delta,1}(a)^2 \phi(a) da = b^2 \int_{\lambda-b}^\infty \phi(a) da = b^2 \Phi(b - \lambda), \quad (\text{E.66})$$

$$\int \psi_{\Delta,1}(a)^2 \phi(a - \zeta) da = b^2 \int_{\lambda-b}^\infty \phi(a - \zeta) da = b^2 \Phi(\zeta + b - \lambda), \quad (\text{E.67})$$

$$\int \psi_{\Delta,2}(a)^2 \phi(a) da = \lambda^2 \int_{\lambda-b}^{\lambda+b} \phi(a) da \leq 2b\lambda^2 \phi((\lambda - b)_+), \quad (\text{E.68})$$

$$\int \psi_{\Delta,2}(a)^2 \phi(a - \zeta) da = \lambda^2 \int_{\lambda-b}^{\lambda+b} \phi(a - \zeta) da \leq 2b\lambda^2 \phi((\lambda - b - \zeta)_+). \quad (\text{E.69})$$

On the other hand, it holds that, for all sequence (a, \bar{a}) satisfying $|\bar{a} - a| \leq b$,

$$\begin{aligned} b^2 \Phi(-\bar{a}) &\leq \frac{c_N}{(1 + a_+) \log N} \phi(\bar{a}_+) \leq c_N (1 + a_+)^{-3} (\phi(a_+) + c_N N^{-1}) \\ &\leq c_N (1 + a_+)^{-2} (\Phi(-a) + c_N N^{-1}). \end{aligned} \quad (\text{E.70})$$

The first inequality comes from $b = o(1/\sqrt{\log N})$ 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 $(1 + \lambda_+)^{-1} \leq \iota_{1,1} \leq \iota_{1,2}$ and $(1 + (\lambda - \zeta)_+)^{-1} \leq \iota_{1,2}$ (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 $\rho \rightarrow 0$, $\iota_{1,1} \leq \iota_{1,2}$, and $\rho \Phi(\zeta - \lambda) \gtrsim N^{-1}$ by assumption, we obtain

$$b^2 \Phi(b - \lambda) + b^2 \rho \Phi(\zeta + b - \lambda) = o(\iota_{1,1}^2 \Phi(-\lambda) + \iota_{1,2}^2 \rho \Phi(\zeta - \lambda)). \quad (\text{E.71})$$

Given (E.65), (E.66), (E.67), (E.59), (E.60), and (E.61), we obtain from (E.71) that $\mathbb{E}(\psi_{\Delta,1}(\tilde{z}_i)^2) = o(\mathbb{E}(\tilde{\psi}_{1,i}(\lambda)^2))$, which proves the lemma for $q = 1$.

Next, it holds that, for all sequence (a, \bar{a}) satisfying $|\bar{a} - a| \leq b$,

$$b\phi(\bar{a}_+) \leq \frac{c_N}{\sqrt{\log N}} \phi(\bar{a}_+) \leq c_N (1 + a_+)^{-1} \phi(a_+) + c_N N^{-1} \leq c_N \Phi(-a) + c_N N^{-1}. \quad (\text{E.72})$$

We recall $x_+ = \max\{x, 0\}$. The first inequality comes from $b = o(1/\sqrt{\log N})$. The second comes from (D.18) of Lemma D4. For the last inequality we use (E.15), too. Further, we note that $\lambda \leq \iota_{2,1}$ and $\lambda \leq \iota_{2,2}$ (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 $\rho \rightarrow 0$, $\iota_{2,1} \leq \iota_{2,2}$, and $\rho\Phi(\zeta - \lambda) \gtrsim N^{-1}$ by assumption, we obtain

$$b\lambda^2\Phi(b - \lambda) + b\lambda^2\rho\Phi(\zeta + b - \lambda) = o(\iota_{1,1}^2\Phi(-\lambda) + \iota_{1,2}^2\rho\Phi(\zeta - \lambda)). \quad (\text{E.73})$$

Given (E.65), (E.68), (E.69), (E.59), (E.62), and (E.63), we obtain from (E.73) that $E(\psi_{\Delta,2}(\tilde{z}_i)^2) = o(E(\tilde{\psi}_{2,i}(\lambda)^2))$, 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 $\rho N\Phi(\mu - \lambda) \rightarrow \infty$. Also suppose λ' is a \mathcal{G} -measurable sequence λ' that satisfies $|\lambda' - \lambda| = o_P(1/\sqrt{\log N})$. Then it holds that $\hat{S}'_q(\lambda') = S_q(\lambda) + o_P(S^{\text{OPT}} + 1)$ 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 $\|\hat{\psi}_q(\lambda') - \tilde{\psi}_q(\lambda)\|^2$. Since $\tilde{z}_i = \sqrt{T}(s_i + \bar{\varepsilon}_i)$ and we have $\max_i |\bar{\varepsilon}_i| \lesssim_P \sqrt{(\log N)/T}$ by uniform bound on i.i.d. normal variables and $|s_i| \lesssim N^{-d}$ by assumption, we obtain $\sqrt{(\log N)/T} \max_i |\tilde{z}_i| = o_P(1/\sqrt{\log N})$ as $(\log N)^3/T \rightarrow 0$ by assumption. Combining this result with the first part of (D.12) of Lemma D2, and noting $T = o(N)$ by assumption, we obtain

$$\sup_{i \leq N} |\hat{z}_i - \tilde{z}_i| = o_P(1/\sqrt{\log N}). \quad (\text{E.74})$$

Further, by definition it holds that, for all $a, b \geq 0$, and $\lambda \geq 0$,

$$\sup_{(a', \lambda') : |a' - a| + |\lambda' - \lambda| \leq b} |\tilde{\psi}_q(a', \lambda') - \tilde{\psi}_q(a, \lambda)| \leq \psi_{\Delta,q}(a, \lambda, b) + \mathbb{1}_{\{q=2\}}\psi_{\Delta,1}(a, \lambda, b), \quad (\text{E.75})$$

where $\psi_{\Delta,q}$ is introduced in the statement of Lemma E5. Substituting (E.74) and $|\lambda' - \lambda| = o_P(1/\sqrt{\log N})$ (by assumption) into (E.75) (choose $(a', a) = (\hat{z}_i, \tilde{z}_i)$), applying Lemma E5, noting $|\tilde{\psi}_1(a, \lambda)| \leq |\tilde{\psi}_2(a, \lambda)|$ by definition, and using Chebyshev's inequality, we obtain

$$\|\hat{\psi}_q(\lambda') - \tilde{\psi}_q(\lambda)\|^2 = o(NE(\tilde{\psi}_{q,i}(\lambda)^2)). \quad (\text{E.76})$$

Using $\max_{i \leq N} |\hat{\sigma}_i/\sigma_i - 1| \lesssim_P c_N$ by (D.1) of Lemma D1 and the triangular inequality, we further have

$$\|\sigma\widehat{w}'_q(\lambda') - \widetilde{\psi}_q(\lambda)\|^2 = o(NE(\widetilde{\psi}_{q,i}(\lambda)^2)). \quad (\text{E.77})$$

Next, we write

$$|\sigma\widehat{w}'_q(\lambda')^\top \psi - \widetilde{\psi}_q(\lambda)^\top \psi| \leq \|\sigma\widehat{w}'_q(\lambda') - \widetilde{\psi}_q(\lambda)\| \|\psi\| = o_P \left(\sqrt{NE(\widetilde{\psi}_{q,i}(\lambda)^2)} (1 + S^{\text{OPT}}) \right). \quad (\text{E.78})$$

The first inequality comes from Cauchy-Schwarz inequality. For the last equality we utilize $\|\psi\| \lesssim_P 1 + S^{\text{OPT}}$ by Corollary 2 and (E.77). On the other hand, we have

$$\text{Var}(\widetilde{\psi}_q(\lambda)^\top \psi) \leq NE(\widetilde{\psi}_{q,i}(\lambda)^2 \psi_i^2) \leq NE(\widetilde{\psi}_{q,i}(\lambda)^2 \alpha_i^2) = o(NE(\widetilde{\psi}_{q,i}(\lambda)^2)). \quad (\text{E.79})$$

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 $\psi_i = E(\alpha_i | \mathcal{G})$ 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

$$\widetilde{\psi}_q(\lambda)^\top \psi = NE(\widetilde{\psi}_{q,i}(\lambda) \psi_i) + o_P \left(\sqrt{NE(\widetilde{\psi}_{q,i}(\lambda)^2)} \right). \quad (\text{E.80})$$

Combining (E.78) and (E.80), we obtain

$$\sigma\widehat{w}'_q(\lambda')^\top \psi = NE(\widetilde{\psi}_{q,i}(\lambda) \psi_i) + o_P \left(\sqrt{NE(\widetilde{\psi}_{q,i}(\lambda)^2)} (1 + S^{\text{OPT}}) \right). \quad (\text{E.81})$$

Moreover, it holds that

$$\begin{aligned} \sigma^2 \|\widehat{w}'_q(\lambda')\|^2 &= \|\widetilde{\psi}_q(\lambda)\|^2 + O(\|\sigma\widehat{w}'_q(\lambda') - \widetilde{\psi}_q(\lambda)\|^2 + \|\widetilde{\psi}_q(\lambda)\| \|\sigma\widehat{w}'_q(\lambda') - \widetilde{\psi}_q(\lambda)\|) \\ &= (1 + o_P(1)) NE(\widetilde{\psi}_{q,i}(\lambda)^2). \end{aligned} \quad (\text{E.82})$$

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 $E(\widetilde{\psi}_{q,i}(\lambda)) = 0$ 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

$$E(\|\beta^\top \widetilde{\psi}_q(\lambda)\|^2 | \beta) \lesssim E(\widetilde{\psi}_{q,i}(\lambda)^2) N \|\beta\|_{\text{MAX}}^2 \lesssim_P NE(\widetilde{\psi}_{q,i}(\lambda)^2). \quad (\text{E.83})$$

Then we have

$$\sigma \|\beta^\top \widehat{w}'_q(\lambda')\| \leq \|\beta^\top \widetilde{\psi}_q(\lambda)\| + \|\beta^\top (\sigma\widehat{w}'_q(\lambda') - \widetilde{\psi}_q(\lambda))\| \lesssim_P \sqrt{NE(\widetilde{\psi}_{q,i}(\lambda)^2)} + c_N N \sqrt{E(\widetilde{\psi}_{q,i}(\lambda)^2)}. \quad (\text{E.84})$$

where the last inequality comes from (E.83) and (E.77). Therefore, we obtain

$$\|\mathbb{M}_\beta \widehat{w}'_q(\lambda')\|^2 = \|\widehat{w}'_q(\lambda')\|^2 - \widehat{w}'_q(\lambda')^\top \mathbb{P}_\beta \widehat{w}'_q(\lambda') = \sigma^2 \|\widehat{w}'_q(\lambda')\|^2 + o_P(NE(\widetilde{\psi}_{q,i}(\lambda)^2)), \quad (\text{E.85})$$

$$\widehat{w}'_q(\lambda')^\top \mathbb{M}_\beta \psi = \widehat{w}'_q(\lambda')^\top \psi + \widehat{w}'_q(\lambda')^\top \mathbb{P}_\beta \psi = \widehat{w}'_q(\lambda')^\top \psi + o_P\left(\sqrt{NE(\widetilde{\psi}_{q,i}(\lambda)^2)}\right). \quad (\text{E.86})$$

The last equalities of both results come from (E.84) and $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$. For the last equality of (E.86), we also use $\|\beta^\top \psi\| \lesssim_P N^{1/2} \mathbb{E}(\alpha_i^2)^{1/2} \lesssim_P c_N N^{1/2}$. We now conclude that

$$\begin{aligned} \frac{\widehat{w}'_q(\lambda')^\top \mathbb{M}_\beta \psi}{\|\mathbb{M}_\beta \widehat{w}'_q(\lambda')\|} &= \frac{(1 + o_P(1)) \widehat{w}'_q(\lambda')^\top \psi}{\sqrt{NE(\widetilde{\psi}_{q,i}(\lambda)^2)}} + o_P(1) \\ &= (1 + o_P(1)) S_q(\lambda) + o_P(1 + S^{\text{OPT}}) = S_q(\lambda) + o_P(S^{\text{OPT}} + 1). \end{aligned} \quad (\text{E.87})$$

The first equality comes from (E.85), (E.86), and (E.82). The second comes from (E.81). The last equality comes from $S_q(\lambda) \leq N^{1/2} \mathbb{E}(\psi_i^2)^{1/2} = S^{\text{OPT}}$, 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 , $w^\top(\alpha - \mathbb{E}(\alpha|\mathcal{G}))/\|w\| = o_P(1)$. Choosing $w = \widehat{w}'_q(\lambda')^\top \mathbb{M}_\beta$, and noting $\psi = \sigma^{-1} \mathbb{E}(\alpha|\mathcal{G})$, we obtain

$$\widehat{S}'_q(\lambda') = \frac{\widehat{w}'_q(\lambda')^\top \mathbb{M}_\beta \alpha}{\sigma \|\mathbb{M}_\beta \widehat{w}'_q(\lambda')\|} = \frac{\widehat{w}'_q(\lambda')^\top \mathbb{M}_\beta \psi}{\|\mathbb{M}_\beta \widehat{w}'_q(\lambda')\|} + o_P(1). \quad (\text{E.88})$$

Given (E.87) and (E.88), we prove the lemma. ■

Lemma E7. *Suppose the same assumptions as in Proposition B2. Then it holds that $\widehat{S}'_q(\lambda) = o_P(1)$, for all \mathcal{G} -measurable positive sequence λ satisfying $\mu^2 \rho N \Phi(\zeta - \lambda) = o_P(1)$ 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 $\lambda \geq \lambda^*$ in probability and $\mu^2 \rho N \Phi(\zeta - \lambda^*) \rightarrow 0$, which is apparently always feasible.

When $\Phi(\zeta - \lambda^*) \gtrsim 1$, we have $\mu^2 \rho N \rightarrow 0$ and thereby $(S^{\text{OPT}})^2 = NE(\psi_i^2) \leq NE(s_i^2) \leq \rho N \mu^2 \rightarrow 0$. Since $\widehat{S}'_q(\lambda)$ is the Sharpe ratio generated by $\widehat{w}'_q(\lambda)^\top \mathbb{M}_\beta$, which is \mathcal{G} -measurable and satisfy $\widehat{w}'_q(\lambda)^\top \mathbb{M}_\beta \beta = 0$. Then we obtain $\widehat{S}'_q(\lambda) \leq S^{\text{OPT}} + o_P(1) = o_P(1)$ from the second part of Theorem 2 and Corollary 2 (the assumptions of the lemma obviously guarantee the prerequisites of both). Since $-\widehat{S}'_q(\lambda)$ is the Sharpe ratio generated by $-\widehat{w}'_q(\lambda)^\top \mathbb{M}_\beta$, we obtain $-\widehat{S}'_q(\lambda) \leq o_P(1)$, too.

When $\Phi(\zeta - \lambda^*) \rightarrow 0$, we have $\Phi(-\lambda^*) \leq \Phi(\zeta - \lambda^*) \rightarrow 0$. Moreover, it holds that, for all sequence (a, \bar{a}) satisfying $|\bar{a} - a| = o(1/\sqrt{\log N})$,

$$\Phi(-\bar{a}) \lesssim \frac{1}{1 + a_+} \phi(\bar{a}_+) \lesssim \frac{1}{1 + a_+} (\phi(a_+) + c_N N^{-2}) \lesssim \Phi(-a) + c_N N^{-2}. \quad (\text{E.89})$$

The first inequality comes from $|\bar{a} - a| = o(1/\sqrt{\log N})$ 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 $b_N = o(1/\sqrt{\log N})$,

$$\begin{cases} N\Phi(b_N - \lambda^*) \lesssim N\Phi(-\lambda^*) + c_N N^{-1} = o(N), \\ \rho N\Phi(\zeta + b_N - \lambda^*) \lesssim \rho N\Phi(\zeta - \lambda^*) + c_N \rho N^{-1} \lesssim o(\mu^{-2}). \end{cases} \quad (\text{E.90})$$

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 $b_N = o(1/\sqrt{\log N})$,

$$\begin{aligned} \sum_{i \leq N} \mathbb{1}_{\{\hat{w}'_{q,i}(\lambda) \neq 0\}} &\lesssim_P \sum_{i \leq N} \mathbb{1}_{\{|\tilde{z}_i| \geq \lambda^* - b_N\}} \\ &\lesssim_P N\Phi(b_N - \lambda^*) + \rho N\Phi(\zeta + b_N - \lambda^*) = o(N), \end{aligned} \quad (\text{E.91})$$

$$\sum_{i \leq N} \mathbb{1}_{\{\hat{w}'_{q,i}(\lambda) \neq 0, \alpha_i \neq 0\}} \lesssim_P \sum_{i \leq N} \mathbb{1}_{\{|\tilde{z}_i| \geq \lambda^* - b_N, \alpha_i \neq 0\}} \lesssim_P \rho N\Phi(\zeta + b_N - \lambda^*) \lesssim o(\mu^{-2}). \quad (\text{E.92})$$

For both results, the first inequality comes from (E.74), the second inequality comes from the density of \tilde{z}_i and Chebyshev's inequality, the last inequality comes from (E.90). As a result, we have

$$\|\beta^\top \hat{w}'_q(\lambda)\| \lesssim \|\beta\|_{\text{MAX}} \|\hat{w}'_q(\lambda)\| \sqrt{\sum_{i \leq N} \mathbb{1}_{\{\hat{w}'_{q,i}(\lambda) \neq 0\}}} = o_P\left(\sqrt{N} \|\hat{w}'_q(\lambda)\|\right). \quad (\text{E.93})$$

Here the first inequality comes from Cauchy-Schwarz inequality, and the last equality comes from (E.91). Then we obtain

$$\|\mathbb{M}_\beta \hat{w}'_q(\lambda)\|^2 = \|\hat{w}'_q(\lambda)\|^2 - \hat{w}'_q(\lambda)^\top \mathbb{P}_\beta \hat{w}'_q(\lambda) = (1 + o_P(1)) \|\hat{w}'_q(\lambda)\|^2. \quad (\text{E.94})$$

For the last equality, we use (E.93) and $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$. Similarly, using (E.93), $\lambda_{\min}(\beta^\top \beta) \gtrsim_P N$ and $\|\beta^\top \alpha\| \lesssim_P N^{1/2} \mathbb{E}(\alpha_i^2)^{1/2} \lesssim_P c_N N^{1/2}$, we obtain

$$\hat{w}'_q(\lambda)^\top \mathbb{M}_\beta \alpha = \hat{w}'_q(\lambda)^\top \alpha + \hat{w}'_q(\lambda)^\top \mathbb{P}_\beta \alpha = \hat{w}'_q(\lambda)^\top \alpha + o_P(\|\hat{w}'_q(\lambda)\|). \quad (\text{E.95})$$

Combining (E.85) and (E.86), we now conclude that

$$\hat{S}'_q(\lambda) = \frac{\hat{w}'_q(\lambda)^\top \mathbb{M}_\beta \alpha}{\|\mathbb{M}_\beta \hat{w}'_q(\lambda)\|} = (1 + o_P(1)) \frac{\hat{w}'_q(\lambda)^\top \alpha}{\|\hat{w}'_q(\lambda)\|} + o_P(1). \quad (\text{E.96})$$

On the other hand, it holds that

$$\frac{|\widehat{w}'_q(\lambda)^\top \alpha|}{\|\widehat{w}'_q(\lambda)\|} \leq \mu \frac{\sum_{i \leq N} |\widehat{w}'_{q,i}(\lambda)| \mathbb{1}_{\{\widehat{w}_{q,i}(\lambda) \neq 0, \alpha_i \neq 0\}}}{\|\widehat{w}'_q(\lambda)\|} \leq \mu \sqrt{\sum_{i \leq N} \mathbb{1}_{\{\widehat{w}'_{q,i}(\lambda) \neq 0, \alpha_i \neq 0\}}} = o_P(1). \quad (\text{E.97})$$

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.