The Statistical Limit of Arbitrage*

Rui Da[†]
University of Chicago

Stefan Nagel[‡]
University of Chicago, NBER, and CEPR

Dacheng Xiu[§]
University of Chicago

This version: June 29, 2023

Abstract

When alphas are weak and rare, and arbitrageurs have to learn about alphas from historical data, there is a gap between Sharpe ratio that is feasible for them to achieve and the infeasible Sharpe ratio that could be obtained with perfect knowledge of parameters in the return generating process. This statistical limit to arbitrage widens the bounds within which alphas can survive in equilibrium relative to the arbitrage pricing theory (APT) in which arbitrageurs are endowed with perfect knowledge. We derive the optimal Sharpe ratio achievable by any feasible arbitrage strategy, and illustrate in a simple model how this Sharpe ratio varies with the strength and sparsity of alpha signals, which characterize the difficulty of arbitrageurs' learning problem. Furthermore, we design an "all-weather" arbitrage strategy that achieves this optimal Sharpe ratio regardless of the conditions of alpha signals. Our empirical analysis of equity returns shows that this optimal strategy, along with other feasible strategies based on multiple-testing, LASSO, and Ridge methods, achieve a moderately low Sharpe ratio out of sample, in spite of a considerably higher infeasible Sharpe ratio, consistent with absence of feasible near-arbitrage opportunities and relevance of statistical limits to arbitrage.

Keywords: Learning about Alphas, Rational Expectation, Portfolio Choice, Rare and Weak Signal, False Negatives, Testing APT, Machine Learning

*We are grateful for comments from Andreas Neuhierl, Ye Luo, and seminar and conference participants at Princeton University, University of Oxford, EPFL, Gregory Chow Seminar Series in Econometrics and Statistics, University of Liverpool, Sao Paulo School of Economics, Peking University, Tsinghua University, Shanghai University of Finance and Economics, University of Gothenburg, Stockholm Business School, Indiana University, NBER Summer Institute, and China International Conference in Finance.

[†]Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: rui.da@chicagobooth.edu.

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

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

1 Introduction

It is a fundamental underlying principle of most asset pricing theories, including the Arbitrage Pricing Theory (APT), that investment opportunities with extremely high ratios of reward to risk do not exist in financial markets. Implicitly, these theories rest on the premise that such near-arbitrage opportunities would attract arbitrageurs who exploit and thereby eliminate these opportunities. An important assumption in these theories is that parameters in the data generating process (DGP) of returns are known to arbitrageurs. Therefore, near-arbitrage opportunities in the DGP of returns are ruled out.

In practice, however, sophisticated investors searching for near-arbitrage opportunities do not know the true parameters. Instead, they commonly conduct statistical analyses to learn about the existence of such opportunities from historical returns data. As a consequence, they face statistical uncertainty. In some settings, such as in some derivatives pricing applications, for instance, the statistical uncertainty may be sufficiently small that it is not a significant impediment to arbitrageur activity. But in noisy, high-dimensional settings such as the cross-section of stock returns, statistical uncertainty can be substantial and it can constitute a statistical limit to arbitrage.

To analyze the effects of arbitrageur learning, we consider a setting in which returns follow a statistical linear factor model. Near-arbitrage opportunities are characterized by high Sharpe ratios. To exploit such opportunities, arbitrageurs need knowledge of factor model alphas, but they must learn about these from historical realizations of alpha signals. We derive the optimal Sharpe ratio achievable by any feasible arbitrage trading strategies, 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-near-feasible-arbitrage bound on the Sharpe ratio that accounts for the statistical limit to arbitrage.

The difficulty of the learning problem hinges on the DGP of alpha signals. While our theory generally does not rely on specific cross-sectional distributions of alpha signals, we use simple special cases to demonstrate 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 learn the distribution of alpha perfectly in the limit. But when alpha is weaker and more rare, its inference becomes more challenging and a gap arises 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 stays bounded.

The existence of this statistical limit to arbitrage implies a widening of the bounds in which mispricing can survive in equilibrium compared with a situation in which arbitrageurs know the DGP and its parameters. Some mispricing may survive because it is clouded by too much statistical uncertainty. Empirically therefore, the feasible, not the infeasible, Sharpe ratio tells us about the minimum reward-to-risk compensation that 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 reality arbitrageurs do not know which DGP is a correct description of the observed data. The optimal strategy estimates the empirical distribution of alpha signals and assigns weights based on the relative magnitudes and associated uncertainty of the alpha estimates. Assets with high alpha t-statistics get portfolio weights proportional to their signal strength. Weaker alphas are more difficult to exploit, yet simply ignoring them would lead to a suboptimal trading strategy. The optimal strategy constructs portfolio weights for weak signals by locally smoothing alpha signals cross-sectionally.

To empirically contrast feasible and infeasible Sharpe ratios, we also 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 realized by any feasible portfolio with weights constructed using historical data. The infeasible Sharpe ratio often serves as the building block for tests of APT, see, e.g., Gibbons et al. (1989), Gagliardini et al. (2016), Fan et al. (2015), and Pesaran and Yamagata (2017). While such tests are powerful and may lead to discoveries of alpha signals, they are not relevant for arbitrageurs who are confined to feasible trading strategies. Our effort in constructing the optimal feasible arbitrage portfolio and evaluating its economic performance directly responds to Shanken's call (Shanken (1992)): "... practical content is given to the notion of 'approximate arbitrage,' by characterizing the investment opportunities that are available as a consequence of the observed expected return deviation ... Far more will be learned, I believe, by examining the extent to which we can approximate an arbitrage with existing assets."

Next, we examine whether alternative strategies that exploit multiple testing, shrinkage, and selection techniques to build arbitrage portfolios can attain the optimal feasible Sharpe ratio. With alphas estimated from cross-sectional regressions, one strategy adopts a multiple-testing (BH) procedure as in Benjamini and Hochberg (1995) on the individual p-values of t-statistics for alpha, in order to guard against potential false discoveries among significant alphas, before building the optimal portfolio weights using selected alphas. Other strategies use either LASSO or Ridge penalties to regularize portfolio weights based on alpha estimates. Such strategies amount to imposing a prior distribution on the alphas. We illustrate with a simple example that these strategies can achieve optimal Sharpe ratio under distinct alpha assumptions. In particular, BH procedure achieves optimal performance only when few true alpha signals are substantially strong. Its failure to achieve optimality is precisely due to its conservativeness nature against the less potent alphas. In contrast, the ridge-based portfolio is equivalent to that constructed by alpha estimates from plain cross-sectional regressions. This approach can achieve optimality when almost all true alphas are either uniformly strong or uniformly weak. The LASSO approach attempts to strike a balance between the aforementioned two methods, with a small gap to achieving the theoretically optimal Sharpe ratio, provided an optimal tuning parameter.

Finally, we demonstrate the empirical implications of the statistical limits of arbitrage by examin-

ing 56 years of monthly individual equity returns in US stock market from 1965 to 2020. The average number of stocks over this period exceeds 4000. We construct residuals via cross-sectional regressions from a multi-factor model that directly uses observed characteristics as risk exposures. These characteristics include market beta (Fama and MacBeth (1973)), size (Banz (1981)), operating profits/book equity (Fama and French (2006)), book equity/market equity (Fama and French (2006)), asset growth (Cooper et al. (2008)), momentum (Jegadeesh and Titman (1993)), short-term reversal (Jegadeesh (1990)), industry momentum (Moskowitz and Grinblatt (1999)), illiquidity (Amihud (2002)), leverage (Bhandari (1988)), return seasonality (Heston and Sadka (2008)), sales growth (Lakonishok et al. (1994)), accruals (Sloan (1996)), dividend yield (Litzenberger and Ramaswamy (1979)), tangibility (Hahn and Lee (2009)), and idiosyncratic risk (Ang et al. (2006)), as well as 11 Global industry Classification Standard (GICS) sectors. These characteristics and industry dummies capture similar equity factors in the MSCI Barra model widely-used among practitioners.

A few interesting findings emerge. First, the cross-sectional R^2s are rather low, with a time-series average 8.25% over our sample period from Jan 1965 to Dec 2020. These R^2s are in similar magnitudes compared to existing estimates in the literature, e.g., 7.8% average R^2s from May 1964 to Dec 2009 reported in Lewellen (2015) using 15 factors that largely overlap with ours, but lower than 12-14% over 1987 - 2016 reported in Gu et al. (2021) based on latent factor models. This indicates that there exists a considerable amount of idiosyncratic noise in the cross-section of individual equities, which makes learning about alphas an arduous statistical task.

Second, we obtain the t-statistics corresponding to alpha estimates of all individual stocks based on their full record in our sample. Among 12,415 test statistics in total, only 6.35% (0.63%) of the t-statistics are greater than 2.0 (3.0) in absolute values. Only 0.505% of these t-statistics translate to Sharpe ratios greater than 1.0 in magnitude. Even without controlling for multiple testing, these estimates suggest that non-zero alphas are rather rare and weak.

Third, we find that the optimal feasible arbitrage portfolio with different methods achieve a moderately low annualized Sharpe ratio, about 0.5, whereas the infeasible Sharpe ratios over time are considerably higher—beyond 2.5—on average, and can reach as high as 7.5 for some sample periods. The estimated infeasible Sharpe ratio is an estimate of what arbitrageurs could attain if they had perfect knowledge of DGP parameters, but it is not attainable by constructing a feasible arbitrage portfolio. The large gap between feasible and infeasible Sharpe ratios suggests the empirical relevance of the statistical limit of arbitrage. Moreover, the fact that the feasible Sharpe ratio is small suggests the empirical success of APT.

Fourth, among all feasible strategies, BH and our optimal strategy achieve the best performance, around a Sharpe ratio of 0.5, followed by CSR (0.450) and LASSO (0.384), though the differences are not substantial. However, the BH approach is overly conservative that it eliminates almost all weak signals and trade less than 10 stocks each month, with zero trading activities for over half of the entire sample. CSR and our strategy exploit weak signals. CSR trades all stocks, but receive a slightly lower Sharpe ratio, potentially due to misallocation of portfolio weights to fake signals. Our optimal strategy trade almost all stocks, but with weights adaptive to the signal strength. LASSO is

most disappointing, but this is due to the uncertainty in optimal tuning parameters. The resulting number of traded stocks per month varies considerably from none to all.

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 papers, we rely on asymptotic arguments that do not rely on assumptions about investor preferences, but these results should be seen as an asymptotic approximation for a more realistic setting with a finite number of assets in which weak preference restrictions rule out Sharpe ratios far above the Sharpe ratios of diversified factor portfolios. The statistical limits to arbitrage that we highlight in this paper relax 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 another large strand of literature on the limit of arbitrage, see Gromb and Vayanos (2010) for a comprehensive review. Complementary to the existing literature, the arbitrage limit in our setting stems from statistical uncertainty, instead of being induced from risk, costs, frictions, and other constraints rational expectation investors are facing.

Kozak et al. (2018) argue that the absence of near-arbitrage opportunities enforces the expected returns to approximately line up linearly with common factor covariances, even in a world in which belief distortions affect asset prices. Our study focuses on the deviations of expected returns from this approximate linear relation and how statistical limits to arbitrage allow bigger deviations. 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. Relatedly, Uppal and Zaffaroni (2018) propose a methodology to construct robust portfolios that can be decomposed into alpha (arbitrage) portfolios and beta (factor) portfolios. Our setting is considerably different from both papers 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. On the empirical side, Guijarro-Ordonez et al. (2022) propose a deep learning approach to statistical arbitrage that achieves a sizable out-of-sample Sharpe ratio. The profits of their trading strategy stem from generalized return reversals at daily to weekly frequencies, potentially due to liquidity provision and other microstructure channels. Our empirical analysis is not targeted towards characterizing the reward-to-risk ratios for high frequency traders, nor for traders that turnover a large portion of their portfolios daily.

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 focus on testing a null that all alphas are equal to zero. This is certainly an interesting null hypothesis, but as we emphasize in this paper, the APT does allow for alphas as long as they do not induce an explosive feasible Sharpe ratio. The second literature focuses on detecting strong alphas, in which widely

used multiple testing methods, such as the BH method by Benjamini and Hochberg (1995), or its extensions can be applied to control the false discovery rate (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.[1] 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.

There has been a long-standing critique of rational expectation models in macroeconomics and finance in which economic agents are not confronted with statistical uncertainty over structure parameters, see Hansen (2007). Bayesian learning is one way to expose model agents 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 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.[2]

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 expectation to bound the set of subjective probabilities, and seek robust inference with this form of misspecification. In the spirit of Hansen (2014), we develop an optimal feasible Sharpe ratio for arbitrageurs inside the economic model, which is in contrast with the (infeasible) one from an outside econometrician's point of view. In our setting, the deviation from rational expectations stems naturally from the statistical obstacles economic agents are facing. A subtle and important point we strive to make here is that economic agents embracing machine learning methods in a high dimensional environment could achieve a distinct outcome as opposed to what rational expectation agents could asymptotically.

From a methodological perspective, the optimal portfolio weights are proportional to the posterior mean of alpha, which resembles the classical normal mean problem in empirical Bayes, dating back to Robbins (1956), where the unknown parameters, alpha, are regarded as random draws from some common distribution, and only a noisy version of alpha (in the form of ex-factor returns) is observed. Our nonparametric approach thereby shares the same spirit of nonparametric empirical

[1] 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.

[2] Our 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).

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 digs further into the Sharpe ratios above and beyond the posterior mean of alphas.

Our paper proceeds as follows. Section 2 develops our main result on 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 and 2.4 derive and illustrate the upper bound of feasible 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 technical details.

2 Statistical Limit of Arbitrage

We start by revisiting the arbitrage pricing framework developed by Ross (1976). This setting is ideal for explaining the statistical limit of arbitrage because the arbitrage pricing theory is largely developed based on a reduced-form statistical model for asset returns. This stylized model is sufficiently sophisticated to deliver theoretical insight, and is sufficiently relevant to guide empirical investment decisions.

2.1 Factor Model Setup

To be more concrete, 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, for $t = 1, 2, \ldots, T$:

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

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

Throughout we will consider asymptotic limits as N and T increase while K is fixed. To facilitate our asymptotic analysis along the cross-sectional dimension, N, we regard high dimensional objects such as α, β, and Σ_u as random variables drawn from some cross-sectional distributions, whereas γ and Σ_v are regarded as deterministic parameters, since their dimensions are fixed. We assume that

α has mean zero, and is cross-sectionally independent of β, and that β has full column rank and is pervasive. These conditions are essential for identification of γ in a model that allows for pricing errors.

We formalize the conditions below.

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

(a) $\|\beta\|_{\text{MAX}} \lesssim_{\text{P}} 1$, *and* $\lambda_{\min}(\beta^{\intercal}\beta) \gtrsim_{\text{P}} N$.[3]

(b) v_t *is i.i.d. across* t, $\mathrm{E}(v_t) = 0$, *and its covariance matrix* Σ_v *satisfies* $1 \lesssim \lambda_{\min}(\Sigma_v) \leq \lambda_{\max}(\Sigma_v) \lesssim 1$.

(c) (α_i, σ_i) *is i.i.d. across* i, *where* $\sigma_i^2 = (\Sigma_u)_{i,i}$. *Moreover,* $\mathrm{E}(\alpha_i|\sigma_i) = 0$ *and* $\mathrm{E}\big(\max_{i:i\leq N} \alpha_i^2\big) = o(1)$.

(d) $u_{i,t} = \sigma_i \varepsilon_{i,t}$, *where* $\varepsilon_{i,t}$ *is i.i.d. across* (i,t), *independent of* Σ_u, *satisfying* $\mathrm{E}(\varepsilon_{i,t}) = 0$ *and* $\mathrm{Var}(\varepsilon_{i,t}) = 1$. *In addition,* Σ_u *satisfies* $1 \lesssim_{\text{P}} \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) \lesssim_{\text{P}} 1$.

(e) The pricing errors α, *factors* v, *factor loadings* β, *and idiosyncratic errors* u *are, conditionally on* Σ_u, *mutually independent.*

Assumption 1 (a) and (b) are commonly seen in the literature of factor models. In particular, the assumption on $\lambda_{\min}(\beta^{\intercal}\beta)$ requires that all factors are pervasive.[4] Condition (c) on α implies that $\mathrm{Var}(\alpha_i) = \mathrm{E}(\alpha_i^2) = o(1)$. As will become clear (from footnote 8), a diminishing variance on α is necessary for precluding near-arbitrage opportunities in Ross' APT. (c) and (d) together suggest that the alpha signals in our model are weak, in that as N increases their magnitudes shrink towards 0, whereas volatilities are bounded from above and from below. More importantly, the assumptions imply that learning about alpha is a more arduous task than learning about volatilities.

There are at least three variations of the factor model (1), depending on what econometricians assume to be observable. The most common setup in academic finance literature imposes that factors are observable as in e.g., Fama and French (1993).[5] 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, i.e., characteristics, are assumed observable. The advantage of the last model lies in the fact that estimating a large number of (potentially)

[3] For a matrix A, we use $\|A\|$ and $\|A\|_{\text{MAX}} = \max_{i,j} |a_{ij}|$ to denote the operator norm (or $\mathbb{L}_2$ norm) and the $\mathbb{L}_\infty$ norm of A on the vector space. We use C to denote a generic constant that may change from line to line. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of A. We also use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$, $a \lesssim_{\text{P}} b$ to denote $a = O_{\text{P}}(b)$, $a \eqsim b$ if $a \lesssim b$ and $b \lesssim a$, and use $a \eqsim_{\text{P}} b$ accordingly.

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

[5] This 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 not observable.

time-varying stock-level factor exposures is statistically inefficient and computationally expensive, as opposed to directly specifying risk exposures as (linear functions of) observable characteristics.[6]

Our core theoretical results below (e.g., Theorem 1) directly apply to all three cases aforementioned. In our empirical analysis we will adopt the third framework most convenient for modeling individual stocks. This makes our analysis highly relevant for practitioners. When it comes to portfolios as test assets, we could adopt either of the first two settings, depending on which factor model is of interest (latent or say, Fama-French factor models), because there are no natural observable proxies for portfolios' betas empirically.

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 $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathrm{P})$, if it satisfies $w \in \mathcal{F}_t$, and along some diverging subsequence,[7] with probability approaching one,*

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

Intuitively, no near-arbitrage means there exists no sequence of portfolios that earn positive expected returns with vanishing risks. Under conditions similar to those in Assumption 1, Ingersoll (1984) established that a sufficient and necessary condition for the absence of near-arbitrage is that

$$S^\star = \sqrt{\alpha^\intercal \Sigma_u^{-1} \alpha} \lesssim_{\mathrm{P}} 1. \tag{3}$$

$S^\star$ is the theoretically optimal Sharpe ratio arbitrageurs can achieve in this economy using a portfolio strategy that has zero exposure to factor risks, namely, a "statistical arbitrage" strategy in the jargon of practitioners. This result suggests that moderate mispricing in the form of nonzero alphas is permitted in an economy without near-arbitrage opportunities, but there cannot be too many alphas that are too large, to the extent that $S^\star$ explodes along some diverging sequence.[8]

[6] Strictly speaking, the MSCI Barra model is cast in a conditional version of (1):

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

where β_t is a vector of observed characteristics and γ_{t-1} is a vector of time-varying risk premia. Analyzing this conditional model will not yield additional economic insight relative to the unconditional model with respect to the theoretical limit of arbitrage. Our theory remains valid with β replaced by β_{t-1} without much change. This model is overly parametrized that parameters are not identifiable without additional restrictions. Some examples of parsimonious conditional factor models include Connor et al. (2012), Gagliardini et al. (2016), and Kelly et al. (2019).

[7] We 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.

[8] Using equation (3) and the fact that $\alpha^\intercal \alpha \lesssim_{\mathrm{P}} \alpha^\intercal \lambda_{\min}(\Sigma_u^{-1}) \alpha \lesssim_{\mathrm{P}} \alpha^\intercal \Sigma_u^{-1} \alpha \lesssim_{\mathrm{P}} 1$, we have $\mathrm{E}(\alpha_i^2) = o(1)$ by the law of large numbers.

To achieve this optimal Sharpe ratio, arbitrageurs should hold a portfolio with weights given by $w^\star = \Sigma_u^{-1}\alpha$, according to Ingersoll (1984).[9] Under the rational expectation assumption, arbitrageurs (agents in this model) know the true (population) parameters: α and Σ_u. In reality, however, the true parameters are blind to arbitrageurs as they can only learn these parameters from a finite sample of data. This learning effect is sometimes harmless since it can be expected that when the sample size is large enough, the true parameters are (asymptotically) revealed, and hence the predictions under rational expectation hold approximately. Fundamentally, this phenomenon is due to the assumption 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, it becomes increasingly difficult for arbitrageurs to determine which among all assets truly have nonzero alphas for a given sample size, T. If the learning problem remains difficult as N and T increase, the learning effect persists, which could lead to distinct limiting implications as opposed to the rational expectation case. It turns out that the rational expectation limit $S^\star$ 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 $S^\star$ because of their inability to make error-free inference. Therefore, the condition (3) 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:*

$$\alpha_i \overset{i.i.d.}{\sim} \begin{cases} \mu & \text{with prob. } \rho/2 \\ -\mu & \text{with prob. } \rho/2 \\ 0 & \text{with prob. } 1-\rho \end{cases}, \quad 1 \le i \le N, \tag{4}$$

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

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 dimensions of the investment set, we can accurately characterize the difficulty of the finite sample problem arbitrageurs are facing.[10] 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.

[9] In Ingersoll (1984), α is defined to be the cross-sectional projection of the expected returns onto β in the population model such that $\alpha^\intercal \Sigma_u^{-1} \beta = 0$. In this paper, we assume instead that α is random, satisfying $\mathrm{E}(\alpha^\intercal \beta) = 0$, and hence in our setting the optimal strategy is given by $w^\star = \mathbb{M}_\beta \Sigma_u^{-1} \alpha$, where $\mathbb{M}_\beta = \mathbb{I}_N - \beta(\beta^\intercal\beta)^{-1}\beta^\intercal$ and $\mathbb{I}_N$ is the $N \times N$ identity matrix. $w^\star$ achieves the Sharpe ratio $S^\star$ (asymptotically as N increases).

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

Now suppose, more specifically, that the magnitude of (μ, ρ) satisfies

$$\mu \approx T^{-1/2} \quad \text{and} \quad \rho \approx N^{-1/2}. \tag{5}$$

This condition (5) implies that the signal strength μ vanishes as the sample size T increases and the signal percentage count ρ decays as the investment universe expands ($N \to \infty$). This setup is used to approximate a reality with only a small portion of assets having a nonzero yet small alpha. σ is assumed a fixed constant, since in reality idiosyncratic risks never vanish, whereas alphas can be small driven by competition among arbitrageurs. This model rests on an uncommon territory in the existing literature of asset pricing: weak and rare alphas. In fact, the classical no near-arbitrage condition (3) imposes, implicitly, weakness or rareness on alphas; otherwise, if alphas are strong and dense, $\alpha^\intercal\alpha$ would explode rather rapidly. Even in the current setting, in light of the fact that $\mathrm{E}(\alpha^\intercal\alpha) = \rho\mu^2 N$, we still have $\alpha^\intercal\alpha \xrightarrow{p} \infty$ as long as $N^{1/2}/T \to \infty$. In other words, a near-arbitrage opportunity arises according to (3), with a strategy $w = \sigma^{-2}\alpha$.

However, the statistical obstacle 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 $T^{-1/2}$.[11] Since by design the true alpha is of the same order of magnitude as its level of statistical uncertainty, i.e., $\mu \approx T^{-1/2}$, it is impossible for arbitrageurs to determine precisely which assets among all have nonzero alpha.

For illustration purpose, suppose that arbitrageurs adopt the strategy $\widehat{w} = \sigma^{-2}\widehat{\alpha}$, replacing α in w with $\widehat{\alpha} = \bar{r} = \alpha + \bar{u}$.[12] Out of sample, this portfolio's conditional expected return and conditional variance can be written as:

$$\begin{aligned}
\mathrm{E}\left(\sigma^{-2}\left(\alpha+\bar{u}\right)^\intercal\left(\alpha+u_t\right)|\mathcal{F}_{t-1}\right) &= \sigma^{-2}(\alpha^\intercal\alpha + \bar{u}^\intercal\alpha), \\
\mathrm{Var}\left(\sigma^{-2}\left(\alpha+\bar{u}\right)^\intercal\left(\alpha+u_t\right)|\mathcal{F}_{t-1}\right) &= \sigma^{-2}(\alpha^\intercal\alpha + 2\alpha^\intercal\bar{u} + \bar{u}^\intercal\bar{u}),
\end{aligned}$$

where u_t denotes a future return at t, that shares the same distribution as $\{u_s\}_{s\le t-1}$, but is independent of $\bar{u}$ which belongs to the information set up to $t-1$, $\mathcal{F}_{t-1}$. The resulting squared conditional Sharpe ratio is given by:

$$S^2 = \frac{\sigma^{-4}(\alpha^\intercal\alpha + \bar{u}^\intercal\alpha)^2}{\sigma^{-2}(\alpha^\intercal\alpha + 2\alpha^\intercal\bar{u} + \bar{u}^\intercal\bar{u})} \lesssim_{\mathrm{P}} T^{-1} \to 0, \tag{7}$$

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

$$\sqrt{T}(\widehat{\alpha}_i - \alpha_i) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(1 + \gamma^\intercal(\Sigma_v)^{-1}\gamma)), \tag{6}$$

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

[12] For any time series of random vector a_t, we use $\bar{a}$ to denote its sample average. As we will point out later in the paper, this strategy $\widehat{w}$, which we will denote by $\widehat{w}^{\mathrm{CSR}}$, fails to achieve the optimal Sharpe ratio in all scenarios. We will discuss the optimal strategy in Section 2.5.

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

Is it possible to find a trading strategy better than $\widehat{w}$ that achieves a non-vanishing Sharpe ratio? In fact, as we will show later, in this setting, the optimal Sharpe ratio among all *feasible* trading strategies arbitrageurs adopt, denoted by S^{OPT}, vanishes asymptotically as $N, T \to \infty$, even though the *infeasible* optimal Sharpe ratio $S^\star \to \infty$ if $N^{1/2}/T \to \infty$. The gap between S^{OPT} and $S^\star$, as shown by this example, is enormous.

We say a strategy is *feasible* if it only uses observable data, together with some necessary statistical method for inference. We formalize the definition of a feasible portfolio strategy below:

Definition 2. *A portfolio strategy $\widehat{w}$ is said to be feasible at time t, if it is a function of observables from $t - T + 1$ to t, where T is the sample size.*

In other words, a feasible strategy needs be adapted to the filtration (information set) generated by observables. The performance of a feasible portfolio depends on the difficulty of the statistical learning problem. In many cases, the statistical uncertainty vanishes as the sample size increases, so that learning makes no difference as opposed to rational expectations asymptotically. The learning problem in the above example, however, remains difficult as N and T increase, to the extent that the learning effect does not diminish in the limit and that the asymptotic limit is distinct from what rational expectation assumption implies. As we show below, the learning problem in practice is often rather difficult, hence the optimal arbitrage Sharpe ratio achievable is expected to be much smaller than $S^\star$.

2.3 Upper Bound on Feasible Sharpe Ratios

We now demonstrate the impact of the feasibility constraint on the optimal arbitrage portfolio. For any feasible strategy $\widehat{w}$, its (conditional) Sharpe ratio can be written as:

$$S(\widehat{w}) \coloneqq \mathrm{E}(\widehat{w}^\intercal r_{t+1} | \mathcal{F}_t) / \mathrm{Var}(\widehat{w}^\intercal r_{t+1} | \mathcal{F}_t)^{1/2}.$$

Arbitrageurs, in our setting, strive to find a feasible strategy that maximizes $S(\widehat{w})$.

The next theorem provides an upper bound on $S(\widehat{w})$:

Theorem 1. *Suppose that r_t follows (1) and that Assumption 1 holds. For any feasible portfolio weight $\widehat{w}$, its Sharpe ratio, $S(\widehat{w})$, satisfies, as $N \to \infty$:*

$$S(\widehat{w}) \leq \left(S(\mathcal{G})^2 + \gamma^\intercal \Sigma_v^{-1} \gamma \right)^{1/2} + o_{\mathrm{P}}(1), \quad \textit{with} \quad S(\mathcal{G})^2 \coloneqq \mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma_u^{-1} \mathrm{E}(\alpha|\mathcal{G}), \tag{8}$$

where $\mathcal{G}$ is the information set generated by $\{(r_s, \beta, v_s, \Sigma_u) : t - T + 1 \leq s \leq t\}$. If, in addition, that, $\widehat{w}$ satisfies that $\widehat{w}^\intercal \beta = 0$, that is, the portfolio is factor-neutral, then

$$S(\widehat{w}) \leq S(\mathcal{G}) + o_{\mathrm{P}}(1). \tag{9}$$

It is known that $\gamma^\intercal \Sigma_v^{-1} \gamma$ is the optimal Sharpe ratio earned from factor portfolios. Theorem 1 further points out that $S(\mathcal{G})$ is an upper bound for Sharpe ratios of all feasible portfolio strategies that have no factor exposures. It is $\mathrm{E}(\alpha|\mathcal{G})$, the posterior estimate of the pricing errors, α, that dictates the optimal feasible Sharpe ratio for arbitrageurs, rather than α themselves. In fact, it holds by the definition of $S(\mathcal{G})$ that

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

with the equality holds only when $\mathrm{E}(\alpha|\mathcal{G}) = \alpha$ almost surely, where the right-hand side corresponds to the infeasible scenario in which arbitrageurs can learn α perfectly using their information set, which echoes (3), the result given by Huberman (1982).[13]

Theorem 1 also provides one solution to a long standing problem in optimal portfolio allocation in the presence of parameter uncertainty. It has been known in the literature that the plug-in mean-variance portfolio using sample mean and sample covariance matrix performs poorly. Assuming returns are normally distributed, Kan and Zhou (2007) studied the expected performance of the plug-in mean-variance portfolio and found its Sharpe ratio is smaller than that of the infeasible Sharpe ratio. Nevertheless, they did not provide an upper bound of the feasible optimal Sharpe ratio in the presence of parameter uncertainty. Our result establishes such a bound in a general factor model setup. In what follows we will discuss different portfolio formation strategies as well as propose a new and optimal strategy that achieves this upper bound.

In light of 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 1 hold. For any given return-generating process satisfying (1), there exists no feasible strategy $\widehat{w}$ that leads to a near-arbitrage, if*

$$S(\mathcal{G}) \lesssim_{\mathrm{P}} 1, \quad \textit{as} \quad N \to \infty. \tag{10}$$

The form of $S(\mathcal{G})$ in Theorem 1 appears 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 (diagonal) covariance matrix of idiosyncratic errors, Σ_u. In fact, the upper bound in (8) 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 $\mathrm{E}(S(\mathcal{G}')^2) \leq \mathrm{E}(S(\mathcal{G})^2)$. And yet, we will show in Section 2.5 that $S(\mathcal{G})$ is in fact achievable by a feasible strategy we construct, which only assumes knowledge of β and r_t – the setting in which factor exposures are observable, implying that the no near-arbitrage bound in (10) is sufficient and necessary.

The reason that Σ_u plays no significant role is that in our model idiosyncratic variances do not vanish as N and T increase, unlike alphas. This assumption makes sense empirically, because alphas

[13] For ease of discussion, we assume alpha is random. This difference with Huberman (1982) by itself does not affect any economic or statistical conclusions we draw in this paper.

are (potentially) small and rare, driven by competition among arbitrageurs, whereas idiosyncratic risks never diminish. Consequently, detecting alphas is more challenging as opposed to estimating idiosyncratic variances, and hence the latter plays a secondary (and negligible) role as opposed to the former in the limit of arbitrage.

2.4 Explicit Formula of the Sharpe Ratio Bound

To gain insight on $S(\mathcal{G})$, we seek a more explicit expression in this section. For that purpose, we need impose an additional assumption:

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

(a) $s_i := \alpha_i/\sigma_i$ is independent of σ_i and satisfies $\mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i| \geq c_N\}}) \leq c_N N^{-1}$ for some sequence $c_N \to 0$.

(b) $\varepsilon_{i,t}$ follows a standard normal distribution.

Assumption 2(a) imposes some restriction on the dependence of α_i and σ_i. It is thereby stronger than Assumption 1(c) that imposes no such restriction. Under Assumption 2, $\widehat{\alpha}_i/\sigma_i \sim \mathcal{N}(s_i, 1)$, where s_i is i.i.d., following some prior distribution. Conditional on σ_i, the problem is translated into the classical problem of estimating a high-dimensional vector of normal means in empirical Bayes, see, e.g., Efron (2019). Together with a technical condition on the tail behavior of s_i and the normality assumption on $\varepsilon_{i,t}$, we can derive a more explicit expression of $S(\mathcal{G})$.

Proposition 1. *Suppose that r_t follows (1) and Assumptions 1 and 2 hold. We define*

$$\psi(a) = \frac{\mathrm{E}\left(s_i \phi(a - T^{1/2} s_i)\right)}{\mathrm{E}\left(\phi(a - T^{1/2} s_i)\right)},$$

where $\phi(\cdot)$ is the normal pdf function, and $\mathrm{E}(\cdot)$ is the expectation taken with respect to the cross-sectional distributions of (α_i, σ_i). Then it holds that

$$\mathrm{E}(\alpha_i | \mathcal{G}) = \sigma_i \psi\big(\widetilde{z}_i\big),$$

where $\widetilde{z}_i = T^{1/2}(\alpha_i + \bar{u}_i)/\sigma_i$, $\bar{u}$ is the average of u_t based on a sample of size T. Moreover, we have

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

where $p(a) = \mathrm{E}\left(\phi(a - T^{1/2} s_i)\right)$ is the probability distribution function of $\widetilde{z}_i$.

The first part of Proposition 1 provides a closed-form formula for $\mathrm{E}(\alpha_i|\mathcal{G})$. Under the stated conditions on the independence across i, $\widehat{\alpha}_i = \alpha_i + \bar{u}$ and σ_i, are sufficient summaries of $\mathcal{G}$ for α_i, so that $\mathrm{E}(\alpha_i|\mathcal{G}) = \mathrm{E}(\alpha_i|\widehat{\alpha}_i, \sigma_i)$. Furthermore, by exploiting Assumption 2(a), we can further write

$\mathrm{E}(\alpha_i|\widehat{\alpha}_i, \sigma_i) = \sigma_i \mathrm{E}(s_i|\widehat{\alpha}_i, \sigma_i) = \sigma_i \mathrm{E}(s_i|\widetilde{z}_i)$.[14] The latter expectation can then be evaluated with the help of the normality assumption on $u_{i,t}$. The second part of this proposition aims to simplify $S(\mathcal{G})$ on the basis of $\mathrm{E}(\alpha_i|\mathcal{G})$ according to Theorem 1, which needs this technical condition on the tail behavior of the cross-sectional distribution of s_i given by Assumption 2(a).

Applying this result, we compare the optimal Sharpe ratio S^{OPT} with $S^\star$ of Huberman (1982) in Example 1.

Corollary 2. *Suppose that the same assumptions as in Proposition 1 hold. In addition, we assume alpha follows (4) as in Example 1. Then we have* $S^\star = \sigma^{-1}\mu(\rho N)^{1/2} + o_{\mathrm{P}}(1)$. *Further, assuming that* $\sigma^{-1}\mu(\rho N)^{1/2}$ *does not vanish, then it holds that* $S^{\mathrm{OPT}} \leq (1-\epsilon)\sigma^{-1}\mu(\rho N)^{1/2}$ *for some* $\epsilon > 0$, *if and only if*

$$T^{1/2}\mu/\sigma - \sqrt{-2\log\rho} \lesssim 1. \tag{11}$$

Corollary 2 suggests that when $T^{1/2}\mu/\sigma$ is large that the constraint (11) is violated, $S^\star \eqsim_{\mathrm{P}} S^{\mathrm{OPT}}$, that is, in the limit, the learning effect does not play any role, so that arbitrageurs in this scenario achieve the same optimal Sharpe ratio as in Huberman (1982). 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 to the extent 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 hence S^{OPT} is dominated by $S^\star$ in the limit, resulting in a strictly smaller Sharpe ratio than the infeasible Sharpe ratio in the classical case.

To give a concrete example of Corollary 2, consider an alternative DGP assumption as opposed to (5):[15]

$$\mu \eqsim N^{-\lambda} \quad \text{and} \quad \rho > 0 \quad \text{is fixed.} \tag{12}$$

In this scenario, $(S^\star)^2 \eqsim_{\mathrm{P}} N^{1-2\lambda}$, which explodes unless $\lambda > 1/2$. If further assuming that $N/T \to \psi > 0$, then the left-hand-side of condition (11) is of order $N^{1/2-\lambda} \vee 1$, so that (11) holds if and only if $\lambda \geq 1/2$. Therefore, $\lambda < 1/2$ is not consistent with absence of (feasible) near arbitrage in that the infeasible Sharpe ratio explodes, while in the mean time the feasible Sharpe ratio approximately equals the infeasible Sharpe ratio (by Corollary 2) and hence explodes. If $\lambda > 1/2$, the infeasible Sharpe ratio (and hence the feasible one) vanishes, which does not seem like an economically plausible case. If we think that arbitrageur activity is required to prevent substantial mispricing, then a setting where mispricing disappears asymptotically even if the frictions faced by arbitrageurs are very large is not plausible. This suggests that under this DGP (12), the only economically plausible case with absence of near-arbitrage is $\lambda = 1/2$. That is, λ can be thought as determined in equilibrium, in which there are substantial asset demand distortions such that mispricing in the absence of

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

[15] It is easy to show that the setup (12) satisfies all assumptions of Proposition 1 for all fixed $\lambda > 0$.

arbitrageur action would be non-negligible asymptotically, and arbitrageurs are aggressive enough so that near-arbitrage opportunities do not exist asymptotically.

We now illustrate the behavior of S^{OPT} numerically and verify the theoretical predictions of Corollary 2 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 (4), a ρ percentage 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 landscape of Sharpe ratios on signal weakness and rareness, despite that some of the resulting portfolio Sharpe ratios (the top left conner 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 the easiest when signal is strong and count is large (top left conner), and the most challenging towards the right bottom corner, where the optimal Sharpe ratios drop to near 0.

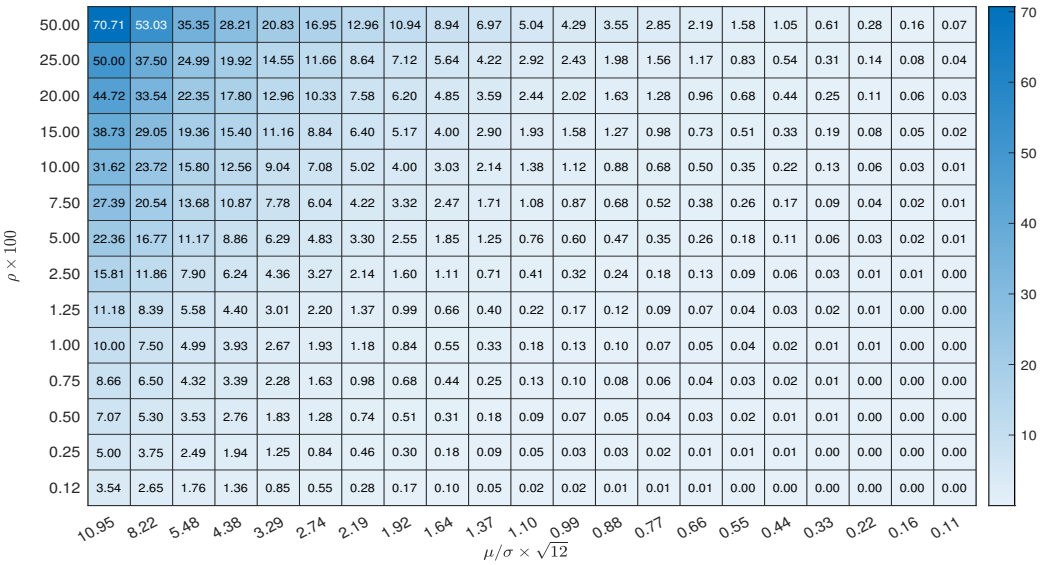

Figure 1: Optimal Sharpe Ratios (S^{OPT}) of Feasible Arbitrage Portfolios

Note: The figure reports optimal Sharpe ratios of feasible arbitrage portfolios in model (4), 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 on Figure 1 are only a fraction of the corresponding (infeasible) Sharpe ratios, $S^\star = \sqrt{\alpha^\intercal(\Sigma_u)^{-1}\alpha} = \mu/\sigma\sqrt{\rho N}$, as shown by Figure 2. The pattern we see from Figure 2 agrees with theoretical predictions of Corollary 2. When the annualized Sharpe ratio

$\mu/\sigma \times \sqrt{12}$ is larger than 2.74, regardless of the values of ρ, the signal-to-noise ratio of the learning problem is sufficiently strong that the statistical limit to arbitrage does not matter much, and hence $S^{\text{OPT}}/S^\star$ is close to 1. Nonetheless, this regime is irrelevant in practice, since it is mostly associated with unrealistically high Sharpe ratios (see Figure 1). In contrast, as μ/σ diminishes, the gap between $S^\star$ and S^{OPT} widens. In almost all empirically relevant scenarios, $S^\star$ is largely exaggerated.

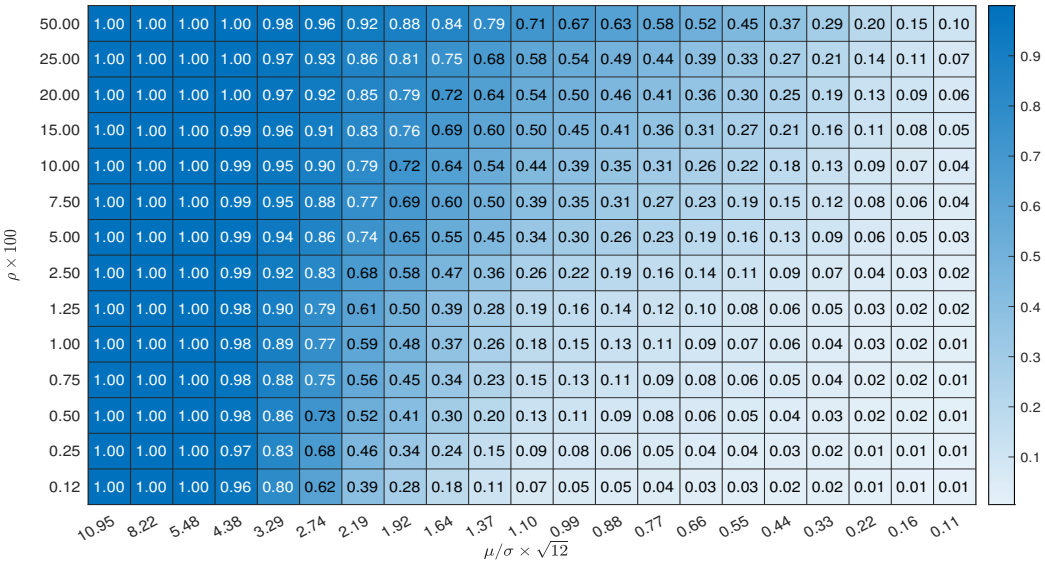

Figure 2: Ratios between S^{OPT} and $S^\star$

Note: The figure reports the ratios of optimal Sharpe ratios between feasible and infeasible arbitrage portfolios. The simulation setting is based on model (4), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

2.5 Constructing the Optimal Arbitrage Portfolio

In our previous discussion, we have shown in Theorem 1 that the optimal Sharpe ratio for any feasible strategy is bounded by $S(\mathcal{G})$. In Proposition 1, we have shown that $S(\mathcal{G}) \approx S^{\text{OPT}}$ under additional assumptions. Corollary 2 further demonstrates that the optimal Sharpe ratio can vary with sequences of DGPs. In light of this, the optimal strategy should depend on the unobserved DGP as well, which poses a serious challenge to arbitrageurs.

It turns out, nevertheless, that arbitrageurs can construct a uniformly optimal strategy, which achieves S^{OPT} over a large class of data generating precesses, without perfect knowledge of the true DGP. We describe this portfolio strategy as "all weather" in that it can be applied in all scenarios of DGPs under consideration. Moreover, the fact that this strategy achieves S^{OPT} implies that the Sharpe ratio upper bound we derive is sharp.

We build this strategy in the setting where factors are latent but factor exposures are observable,

since this is the case we analyze empirically.

Algorithm 1 (Constructing the Optimal Arbitrage Portfolio)**.**
Inputs: r_t, $t \in \mathcal{T} = \{t - T + 1, \ldots, t\}$ *and* β.

S1. We we construct cross-sectional regression estimates of alpha, idiosyncratic volatilities, and the induced t-statistics, for each $i = 1, 2, \ldots, N$*:*

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

S2 We non-parametrically estimate the marginal density of t-statistics using Gaussian kernel function $\phi(x)$ *and bandwidth* $k_N \sim (\log N)^{-1}$*:*

$$\widehat{p}(a) = \frac{1}{N k_N} \sum_i \phi\Big(\frac{\widehat{z}_i - a}{k_N} \Big),$$

S3. We construct an estimate of $\psi(a)$ *by plugging* $\widehat{p}(a)$ *into the Tweedie's formula (Robbins (1956)):*

$$\widehat{\psi}(a) = \frac{1}{\sqrt{T}} a + \frac{1 + k_N^2}{\sqrt{T}} \frac{d}{da} \log \widehat{p}(a).$$

S4. We choose the arbitrage portfolio weights as $\widehat{w}^{\mathrm{OPT}} = \mathbb{M}_\beta \breve{w}$*, with* $\breve{w}_i = \widehat{\psi}(\widehat{z}_i)/\widehat{\sigma}_i$.

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

As we have discussed in footnote 9, the optimal strategy in the case that arbitrageurs know the true DGP is given by

$$w^\star = \mathbb{M}_\beta \Sigma_u^{-1} \alpha, \tag{13}$$

where $\mathbb{M}_\beta = \mathbb{I}_N - \beta(\beta^\intercal \beta)^{-1} \beta^\intercal$. Intuitively, part of the construction in (13), $\Sigma_u^{-1}\alpha$, 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. Multiplying by $\mathbb{M}_\beta$ in (13) simply eliminates factor exposures in r_t, because $\mathbb{M}_\beta r_t \approx \alpha + u_t$. In light of this and Theorem 1,

$$w^{\mathrm{OPT}} = \mathbb{M}_\beta \Sigma_u^{-1} \mathrm{E}(\alpha | \mathcal{G}) \tag{14}$$

appears a reasonable target for portfolio weights when arbitragers do not observe true alphas in the DGP.

Our objective is to construct portfolio weights that approximate the optimal feasible weight given by (14). As explained by the discussion following Proposition 1, the key result is $\mathrm{E}(\alpha_i | \mathcal{G}) = \sigma_i \mathrm{E}(s_i | \widehat{\alpha}_i / \sigma_i) = \sigma_i \mathrm{E}(s_i | \widetilde{z}_i) = \sigma_i \psi(\widetilde{z}_i)$, thanks to the independence assumption between s_i and σ_i. We thereby need an estimator of the conditional expectation function $\psi(a)$. Because $\widetilde{z}_i \sim \mathcal{N}(\sqrt{T} s_i, 1)$, the Tweedie's formula allows us to connect conditional expectation $\psi(a)$ to $p(a)$, the marginal density of $\widetilde{z}_i$, as follows

$$\psi(a) = \frac{1}{\sqrt{T}} a + \frac{1}{\sqrt{T}} \frac{d}{da} \log p(a).$$

In light of this connection, we rely on the empirical Bayes method, following Brown and Greenshtein (2009). Concretely, we conduct in Step S2 kernel density estimation of $p(a)$, which, combined with Tweedie's formula, leads to estimate of the conditional expectation function in Step S3. The additional factor $(1+k_N^2)$ is to correct the bias arising from the estimation error embedded in $\widehat{p}(a)$. With the estimate of $\psi(a)$, the optimal weights on ex-factor returns are constructed as $\breve{w}$. This, in turn, leads to the optimal weight estimates, $\widehat{w}^{\text{OPT}}$, on original input asset returns.

An essential step towards uniform optimality is that we consolidate information of assets with similar $\widehat{z}_i$, as in Step S2, to obtain an estimate of the conditional expectation of their signal strength, using which we obtain their optimal portfolio weights. This strategy outperforms the alternatives, some of which directly use estimated alphas as if these estimates are not susceptible to 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 $\widehat{w}^{\text{OPT}}$:

Theorem 2. *Let $\mathbb{P}$ denote the collection of all data-generating processes under which r_t follows (1), and Assumptions 1 and 2 hold. 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 generated by the portfolio strategy $\widehat{w}^{\text{OPT}}$ as $\widehat{S}^{\text{OPT}} := \text{E}(r_{t+1}^\intercal \widehat{w}^{\text{OPT}} | \mathcal{F}_t)/\text{Var}(r_{t+1}^\intercal \widehat{w}^{\text{OPT}} | \mathcal{F}_t)^{1/2}$. Then it holds that $\widehat{w}^{\text{OPT}}$ achieves, asymptotically, the upper bound S^{OPT} uniformly over all sequences of data-generating processes. That is, for any $\epsilon > 0$,*

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

Theorem 2 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.

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 $\widehat{S}^{\text{OPT}}$ and S^{OPT}. Otherwise, if $S^{\text{OPT}} \to \infty$, the estimation error is described in percentage terms. This is necessary because we simultaneously consider a large class of models.

With Theorem 2, we establish the necessity for the no near-arbitrage condition given by (10).

Corollary 3. *Suppose the same assumptions as in Theorem 2 hold. The portfolio weights by $\widehat{w}^{\text{OPT}}$ yields a near-arbitrage strategy under any sequences of data-generating processes for which condition (10) does not hold.*

We have shown that arbitrageurs can construct an optimal strategy that realizes S^{OPT}. Now suppose that the equilibrium "cost" of implementing an arbitrage is C in an economy with statistical limit of arbitrage. In equilibrium, $S^{\text{OPT}} = C$, otherwise arbitrageurs can trade until it is no longer

profitable to do so. We can thereby interpret $\widehat{S}^{\text{OPT}}$ as an empirical estimate of the arbitrage cost, which we will estimate empirically.

2.6 Estimating Optimal Infeasible Sharpe Ratio

We are also interested in estimating the optimal infeasible Sharpe ratio, $S^\star$, which can be perceived as the optimal Sharpe ratio from an outside econometrician's point of view, and yet cannot be realized by a feasible portfolio. Existing literature on testing APT often construct test statistics in the spirit of Gibbons et al. (1989), which are effectively based on $S^\star$, see, e.g., Pesaran and Yamagata (2017) and Fan et al. (2015). While such tests are powerful and may lead to discoveries of alpha signals, they are not relevant for arbitrageurs in that arbitrageurs may not construct a feasible portfolio to profit from these statistical discoveries.

To construct an estimator for $S^\star$, we consider the following choice motivated from its sample analog:

$$\widetilde{S}^\star = \left(\bar{r}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{r}\right)^{1/2}, \tag{15}$$

where $\bar{r} = T^{-1}\sum_{t\in\mathcal{T}} r_t$, $\widehat{\sigma}_i^2 = T^{-1}\sum_{t\in\mathcal{T}}(r_{i,t}-\bar{r}_i)^2$, and $\widehat{\Sigma}_u = \text{diag}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_N^2)$.

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

$$\widehat{S}^\star = \left(\bar{r}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{r} - N/T\right)^{1/2}. \tag{16}$$

The second estimator again takes the form of a summation over individual squared Sharpe ratios, but it eliminates the term that will be dominated by the estimation bias under some data generating processes. The next proposition summarizes the asymptotic properties of both estimators.

Proposition 2. *Suppose that r_t follows (1) and that Assumption 1 holds. Assume that $\mathrm{E}(\alpha_i^2 \mathbb{1}_{\{|\alpha_i|\geq c_N\}}) \leq c_N N^{-1}$, $T \lesssim N$, $T^{-1}N^{1/2}\log N \leq c_N$, for some sequence $c_N \to 0$, and that $\varepsilon_{i,t}$ has finite eighth moment. Then we have*

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

Similar to Theorem 2, the estimation error is relative when $S^\star$ dominates one asymptotically, and in absolute terms if $S^\star$ is dominated by one.[16] This accommodates a large class of models, some of which have an exploding or a shrinking $S^\star$. While it is possible to estimate $S^\star$, it is not possible to build a portfolio that realize it, unless the signal-to-noise ratio is sufficiently large such that $S^\star = S^{\text{OPT}}$. Empirically, the difference between $\widehat{S}^\star$ and $\widehat{S}^{\text{OPT}}$ thereby tells us the signal strength in the data.

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

2.7 Alternative Strategies for Arbitrage Portfolios

Algorithm 1 suggests a relatively sophisticated procedure that distinguishes weaker and strong signals using t-statistics before constructing, separately, the optimal weights for these signals. In this section, we study several alternative methods, neither of which can achieve optimality uniformly across all DGPs we consider, but they are simpler and somewhat prevalent in practice. The contrast among these strategies helps illustrate their pros and cons in different scenarios.

2.7.1 Cross-Sectional Regression

The conventional approach to estimating alphas is through the cross-sectional regression:

$$\widehat{\alpha} = (\beta^{\intercal}\beta)^{-1}\beta^{\intercal}\bar{r},$$

with which the arbitrage portfolio weights can be constructed directly as:

$$\widehat{w}^{\mathrm{CSR}} = \mathbb{M}_{\beta}\widehat{\Sigma}_u^{-1}\widehat{\alpha}. \tag{17}$$

This choice of portfolio weight is the sample analog of the optimal weight given by (13).

We now exploit Example 1 to illustrate the pros and cons of the CSR strategy. We will point out that it is not optimal in all DGPs. For this purpose, we only need focus on the case in which returns are driven by idiosyncratic errors and alpha. For convenience, we adopt a simplified volatility estimator: $\widehat{\Sigma}_u = \widehat{\sigma}^2\mathbb{I}_N$, where $\widehat{\sigma}^2$ is averaged over all volatility estimates, because in this example, all assets share the same volatility. This further simplifies the analysis because the scaling factor, $\widehat{\sigma}^2$, is cancelled out, and hence $\widehat{\sigma}^2$ does not play any role in the portfolio's Sharpe ratio, $\widehat{S}^{\mathrm{CSR}}$.

Proposition 3. *Suppose that r_t follows (1) with $\beta = 0$, $u_{i,t} \sim \mathcal{N}(0, \sigma^2)$, and α following (4) as in Example 1. We also assume $\mu \lesssim 1$. The Sharpe ratio of the arbitrage portfolio, whose weights are given by $\widehat{w}^{\mathrm{CSR}} = \widehat{\sigma}^{-2}\widehat{\alpha}$, satisfies $\widehat{S}^{\mathrm{CSR}} - S^{\mathrm{CSR}} = o_{\mathrm{P}}(1)$, where*

$$\widehat{S}^{\mathrm{CSR}} = \mathrm{E}(r_{t+1}^{\intercal}\widehat{w}^{\mathrm{CSR}}|\mathcal{F}_t)/\mathrm{Var}(r_{t+1}^{\intercal}\widehat{w}^{\mathrm{CSR}}|\mathcal{F}_t)^{1/2}, \qquad S^{\mathrm{CSR}} = \frac{N^{1/2}\rho\mu^2\sigma^{-2}}{(T^{-1}+\rho\mu^2\sigma^{-2})^{1/2}}.$$

Figure 3 plots the ratio of S^{CSR} against S^{OPT} for a range of parameters. Evidently, this former is dominated by the latter when alpha signals are both sparse and strong. This dominance regime is highlighted in black numbers on the heatmap from Figure 3. As $\mu/\sigma \times \sqrt{12}$ approaches 1.0 (a vertical line) from the right or the upper left corner, the gap between the two Sharpe ratios shrinks.

Intuitively, this approach takes all signals directly without distinguishing the insignificant ones from the significant ones. Consequently, even fake signals (pure noise) are assigned non-zero weights, which, in turn, 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 with respect to parameters for which the cross-sectional regression approach is strongly dominated by our optimal strategy are associated with realistic Sharpe ratios.

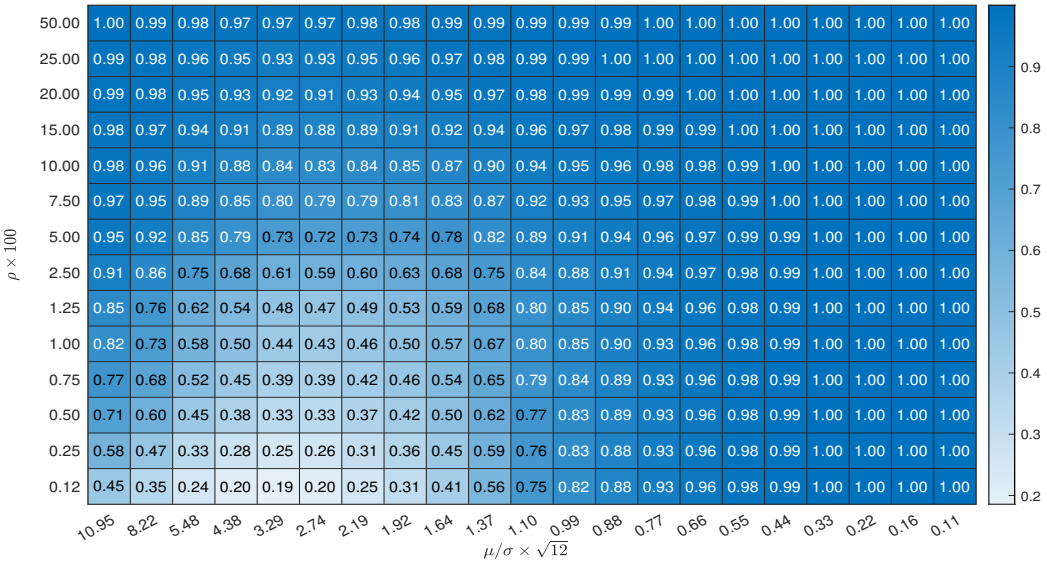

Figure 3: Ratios between S^{CSR} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the OLS based portfolio and the feasible optimal arbitrage portfolio. The simulation setting is based on model (4), 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

From the statistical point of view, we can formalize the search for alpha as a multiple testing problem. Say, there are N assets potentially with nonzero α, and for each i, we can define a null hypothesis: $\mathbb{H}_0^i : \alpha_i = 0$, so that rejecting this null leads to a discovery of α_i. With multiple testing comes the concern of data snooping, meaning that a large fraction of tests that appear positive are in fact due to chance. One sensible approach is to control the false discovery rate (FDR), instead of the size of individual tests, a proposal advocated by Barras et al. (2010), Bajgrowicz and Scaillet (2012), and Harvey et al. (2016) in different asset pricing contexts.

The B-H procedure is often adopted to control FDR in multiple testing problems. Giglio et al. (2021) have proved its validity in a general factor model setting for alpha detection. Below we

describe the algorithm for constructing alpha estimates, which will be used as inputs to the construction of an arbitrage portfolio.

Algorithm 2 (The B-H based Alpha Selection)**.** *Let $\{p_i : i = 1, \ldots, N\}$ be the p-values of the t-test statistics corresponding to the cross-sectional regression estimates of alpha.*

S1. Sort in ascending order the collection of p-values, with the sorted p-values given by $p_{(1)} \leq \ldots \leq p_{(N)}$.

S2. For $i = 1, \ldots, N$, reject $\mathbb{H}_0^i : \alpha_i = 0$, if $p_i \leq p_{(\widehat{k})}$, where $\widehat{k} = \max\{i \leq N : p_{(i)} \leq \tau i/N\}$, for any pre-determined level τ, say, 5%.

Similar to the case of $\widehat{S}^{\text{OPT}}$, we adopt a sample-splitting method. We divide the entire sample $\mathcal{T}$ into two subsamples S and S'. We apply B-H to select signals, whose p-values are based on t-statistics $\breve{z}$ using S, and then construct portfolio weights with alpha estimates $\breve{\alpha}'$ using S':

$$\widehat{\alpha}_i^{\text{BH}}(\tau) = \breve{\alpha}_i' \mathbb{1}_{\{p_i \leq p_{(\widehat{k})}\}}. \tag{18}$$

The B-H procedure guarantees (in expectation) that at least a fraction $(1 - \tau)$ of selected assets have nonzero alphas, regardless of the actual percentage of alphas in the data generating process. Effectively, it imposes a hard-thresholding procedure on the alpha estimates, replacing less significant alphas by zero. Similar to (17), the optimal portfolio weights are thus given by:

$$\widehat{w}^{\text{BH}} = \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \widehat{\alpha}^{\text{BH}}(\tau). \tag{19}$$

Controlling the false discovery rate on top of the CSR estimates is intuitively appealing, but doing so incurs a potential loss of power, leading to less investment opportunities. Our focus is on optimal portfolio construction instead of false discovery control. The next proposition shows that in the context of Example 1, arbitrageurs who adopt the B-H based alpha estimator cannot achieve optimal portfolio for a large class of DGP sequences.

Proposition 4. *Suppose that r_t follows (1) with $\beta = 0$, $u_{i,t} \sim \mathcal{N}(0, \sigma^2)$, and α following (4) as in Example 1. We assume $\mu, \rho \lesssim N^d$ with fixed $d < 0$, and $|S| \asymp |S'| \asymp T$. The Sharpe ratio of the arbitrage portfolio with weights given by $\widehat{w}^{\text{BH}} = \widehat{\sigma}^{-2} \widehat{\alpha}^{\text{BH}}(\tau)$ satisfies $\widehat{S}^{\text{BH}} = S^{\text{BH}} + o_{\text{P}}(1 + S^{\text{BH}})$, where*[17]

$$\widehat{S}^{\text{BH}} = \text{E}(r_{t+1}^\intercal \widehat{w}^{\text{BH}} | \mathcal{F}_t) / \text{Var}(r_{t+1}^\intercal \widehat{w}^{\text{BH}} | \mathcal{F}_t)^{1/2}$$

is the Sharpe ratio, and

$$S^{\text{BH}} = \mu \sigma^{-1} \sqrt{\rho N \Phi(|S|^{1/2} \mu / \sigma - z^*)},$$

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

where $\Phi(\cdot)$ is the normal cumulative distribution function, and z^* is the positive solution of the equation

$$2(1-\tau(1-\rho))\Phi(-z) = \tau\rho\Phi(T^{1/2}\mu/\sigma - z). \tag{20}$$

We note that S^{BH} is upper bounded by $\sqrt{1-\tau}S^{\text{OPT}}$, where τ is the pre-determined level that controls the false discovery rate. Intuitively, as τ increases, the B-H procedure tends to fail in guarding against fake signals, so that the performance of the B-H portfolio would deteriorate.

Similar to CSR, the B-H procedure cannot achieve the optimal Sharpe ratio, as shown by Figure 4. The scenarios that B-H achieves optimality correspond to the white values on Figure 4, where the border of the dominant region is located near the vertical line at $\mu/\sigma\sqrt{12} = 2.19$. Intuitively, the B-H is effective in singling out strong signals, so it leads to almost optimal portfolios as long as all signals are strong. However, when signals are weak, the B-H procedure, which amounts to hard-thresholding, performs worse than the cross-sectional regression. As shown by Figure 1, even if alphas are individually weak, their empirical relevance should not be ignored because their collective contribution to the portfolio's Sharpe ratio can be highly non-trivial. The B-H approach is overly conservative compared to alternatives in this parameter regime, even though B-H remains a preferable approach to selecting truly significant alphas and controlling false discoveries. In contrast, the optimal arbitrage portfolio exploits information embedded in all alpha estimates, including false positives, beyond the set of significant ones selected via B-H procedure. This result also demonstrates a clear distinction between two objectives: alpha testing and portfolio construction, the objectives of which 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

The analysis above suggests that we can construct the optimal portfolio out of the ex-factor returns, while imposing regularization on portfolio weights, before rewriting the regularized portfolio weights in terms of raw returns (i.e., multiplying the weights by $\mathbb{M}_\beta$). Regularizing portfolio weights amounts to imposing priors directly on the alpha estimates. To see this, we adopt a shrinkage approach, when constructing arbitrage portfolios on residual returns:

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

where $p_\lambda(w) = \lambda\|w\|_1$ or $\lambda\|w\|_2^2$, for some $\lambda > 0$. Since $\widehat{\Sigma}_u$ is diagonal, the closed-form solution is $\psi_q(\widehat{\alpha}, \widehat{\Sigma}_u, \lambda)$, where $q = 1$ corresponds to the LASSO penalty and $q = 2$ the ridge, and for $i = 1, 2, \ldots, N$,

$$\Big(\psi_1(\widehat{\alpha}, \widehat{\Sigma}_u, \lambda)\Big)_i = (\widehat{\sigma}_i)^{-2} sgn(\widehat{\alpha}_i)(|\widehat{\alpha}_i| - \lambda)_+, \quad \Big(\psi_2(\widehat{\alpha}, \widehat{\Sigma}_u, \lambda)\Big)_i = \big((\widehat{\sigma}_i)^2 + \lambda\big)^{-1}\widehat{\alpha}_i.$$

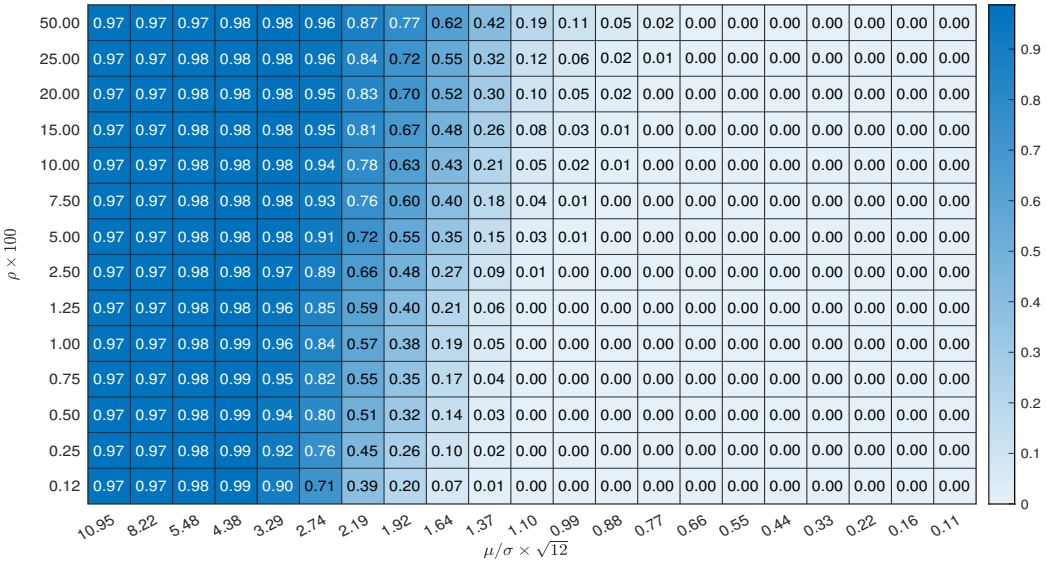

Figure 4: Ratios between S^{BH} and S^{OPT}

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 simulation setting is based on model (4), in which a $100 \times \rho\%$ of assets have alphas that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$.

This leads to the optimal portfolio weight on r_t:[18]

$$\widehat{w}^q = \mathbb{M}_\beta \psi_q(\widehat{\alpha}, \widehat{\Sigma}_u, \lambda), \quad q = 1, 2.$$

Depending on the magnitude of λ, the LASSO approach replaces all smaller signals by zero and shrinks the larger signals 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 with a shrinkage factor depending on $\widehat{\sigma}_i^2$. Like the above analysis, when specialized to example (1), we can adopt $\widehat{\Sigma}_u = \widehat{\sigma}^2 \mathbb{I}_N$, in which case ridge becomes equivalent to CSR! This "embedded" shrinkage effect of CSR explains why it performs well in the case of small signals.

Proposition 5. *Suppose that r_t follows (1) with $\beta = 0$, $u_{i,t} \sim \mathcal{N}(0, \sigma^2)$, and α following (4) as in Example 1. We assume $\mu \leq c_N$. The Sharpe ratio of the arbitrage portfolio with weights given by $\widehat{w}^q$, denoted as $\widehat{S}^q$ for $q = 1, 2$, satisfies $\widehat{S}^1 - S^{\text{LASSO}} = o_{\text{P}}(1)$ and $\widehat{S}^2 - S^{\text{CSR}} = o_{\text{P}}(1)$, where*

$$S^{\text{LASSO}} = \rho\mu\sigma^{-1}N^{1/2} \frac{\int_{-\infty}^{\infty} sgn(x)(T^{-1/2}\sigma|x| - \lambda)_+ \phi(T^{1/2}\sigma^{-1}\mu - x)dx}{\sqrt{\int_{-\infty}^{\infty} \big((T^{-1/2}\sigma|x| - \lambda)_+\big)^2 \big((1-\rho)\phi(x) + \rho\phi(T^{1/2}\sigma^{-1}\mu - x)\big)\big)dx}},$$

and S^{CSR} is defined in Proposition 3.

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

Proposition 5, along with Proposition 3, provides explicit formula of S^{LASSO}. Figure 5 compares S^{LASSO} with S^{OPT}. The LASSO approach involves a tuning parameter, which calls for a cross-validation procedure. We adopt an infeasible and theoretically optimal tuning parameter, λ, that maximizes S^{LASSO}, making this approach a stronger competitor. Even though Proposition 5 suggests that LASSO is not uniformly optimal, it performs quite 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 does not play too much a role. When signals are rather weak, LASSO behaves like Ridge (and hence CSR), because shrinking these signals does not change the fact that they are almost indistinguishable from noise.

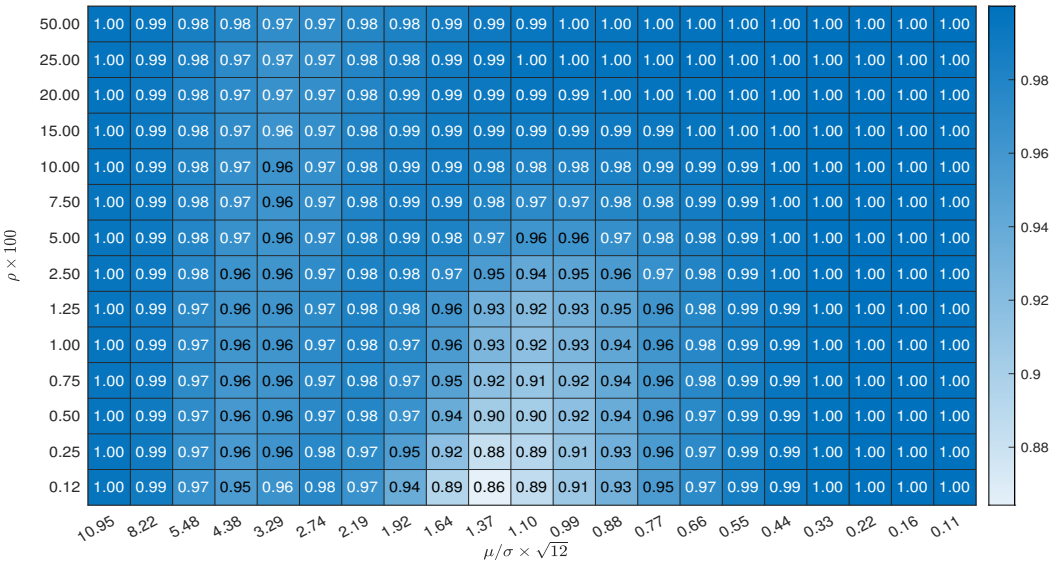

Figure 5: Ratios between S^{LASSO} and S^{OPT}

Note: The figure reports the ratios between the Sharpe ratios of the LASSO based portfolio and the feasible optimal arbitrage portfolio. The simulation setting is based on model (4), 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^{LASSO}.

3 Simulation Evidence

This section demonstrates the empirical relevance of our theory via simulations and examines the finite sample performance of the proposed portfolio strategies.

3.1 Comparison of Portfolio Strategies in Finite Sample

For simplicity and clarity, we simulate a one-factor (CAPM) model of returns given by (1). We choose the factor risk premium as 5% per year and set the annualized volatility at 25%. We model the cross-section of betas using a normal distribution with mean 1 and variance 1. Since we focus

on the arbitrage portfolio, the parameters about the factor component (including the number of factors) are inconsequential, because factors, if any, are eliminated by $\mathbb{M}_\beta$ in the first step when constructing these trading strategies. In addition, we adopt model (4) in Example 1 for the cross-sectional distribution of alpha, and fix the idiosyncratic volatilities of all assets at σ, since it is α/σ that determines the signal strength and that there is no need of varying both α and σ in the cross section.

We now compare the finite sample performance of our portfolio estimators over different DGPs. For any given parameter value $(\mu/\sigma, \rho)$ in a DGP, we estimate the portfolio weights, $\widehat{w}^{\text{OPT}}$, using our Algorithm 1, and calculate the resulting (theoretical) Sharpe ratio: $\widehat{w}^{\text{OPT}\intercal}\mu/\sqrt{\widehat{w}^{\text{OPT}\intercal}\Sigma_u^{-1}\widehat{w}^{\text{OPT}}}$. We then calculate the average Sharpe ratio over all Monte Carlo repetitions. Our approach requires a tuning parameter k_n. For robustness, we report results based on three parameter values $(0.5k_n, k_n, 2k_n)$ with $k_n = 0.25$. We repeat this exercise for the CSR, B-H, and LASSO methods for comparison.

In light of Theorem 2, a sensible choice of the estimation error can be written as:

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

where A denotes OPT, CSR, BH, or LASSO, and the dependence of $\widehat{S}^{\text{A}}$ and S^{OPT} on μ/σ and ρ is omitted. When S^{OPT} is large (i.e., $>> 1$), this error is in percentages relative to S^{OPT}; when S^{OPT} is small (i.e., $o_{\text{P}}(1)$), the error is measured in terms of the absolute difference. The error is defined this way because S^{OPT} itself can diverge or diminish depending on different parameters in the simulated DGPs.

Table 1 reports the maximal error over all values of μ/σ and ρ. The results show that OPT has a smaller error in almost all cases for all tuning parameters than CSR, BH, or LASSO. As T increases from 10 years to 40 years, the maximum error drops from 0.377 to 0.263 in the case of $N = 1,000$ for $k_n = 0.25$, whereas CSR, BH and LASSO stay above 0.44. The maximal error for CSR is achieved at the lower left conner of Figure 1, where signals are strong but rare; for BH, the worst performance occurs around the upper right corner, where many weak signals exist; for LASSO, the worse is near the bottom but in the middle, where signals are neither too strong nor too weak.

3.2 Finite Sample Performance of the Infeasible Sharpe Ratio Estimator

Finally, Figure 6 reports the estimation error $\left|\widehat{S}^\star - S^\star\right|/\left(1+S^\star\right)$ in simulations. The result confirms the consistency result given by Proposition 2. The error is relative when $S^\star$ is large or moderate ($>> 1$). We find the relative error is around 1% towards the left top corner. For DGPs near the bottom right corner of Figure 6, $S^\star$ vanishes as shown by Figures 1 and 2, the error becomes absolute ($S^\star << 1$) and is moderately small given the sample size and the cross-sectional dimension.

	$N = 1,000$, Monthly			$N = 3,000$, Monthly			$N = 1,000$, Daily		
	$T = 10$	$T = 20$	$T = 40$	$T = 10$	$T = 20$	$T = 40$	$T = 10$	$T = 20$	$T = 40$
	0.385	0.332	0.289	0.442	0.367	0.320	0.449	0.440	0.408
OPT	0.377	0.309	0.263	0.437	0.333	0.282	0.411	0.382	0.356
	0.381	0.282	0.233	0.446	0.318	0.247	0.370	0.334	0.303
CSR	0.540	0.489	0.441	0.618	0.570	0.515	0.537	0.485	0.427
BH	0.742	0.703	0.651	0.814	0.789	0.748	0.760	0.715	0.657
LASSO	0.537	0.488	0.440	0.615	0.568	0.512	0.536	0.483	0.426

Table 1: Sharpe Ratio Comparison in Simulations

Note: This table reports the maximum error, defined by $\sup_{\mu/\sigma,\rho} \mathrm{Err}^{\mathrm{A}}(\mu/\sigma, \rho)$, where A denotes either OPT, or CSR, or BH, over all values of μ/σ and ρ in Figure 1, for several choices of N, T (in years), and data frequencies. The first three rows correspond to the OPT approach with three different values of tuning parameters, $0.5k_n$, k_n, and $2k_n$, respectively, where $k_n = 0.25$. The BH approach controls false discovery rate at a level 5%. The LASSO approach uses the optimal (infeasible) tuning parameter that optimizes S^{LASSO}.

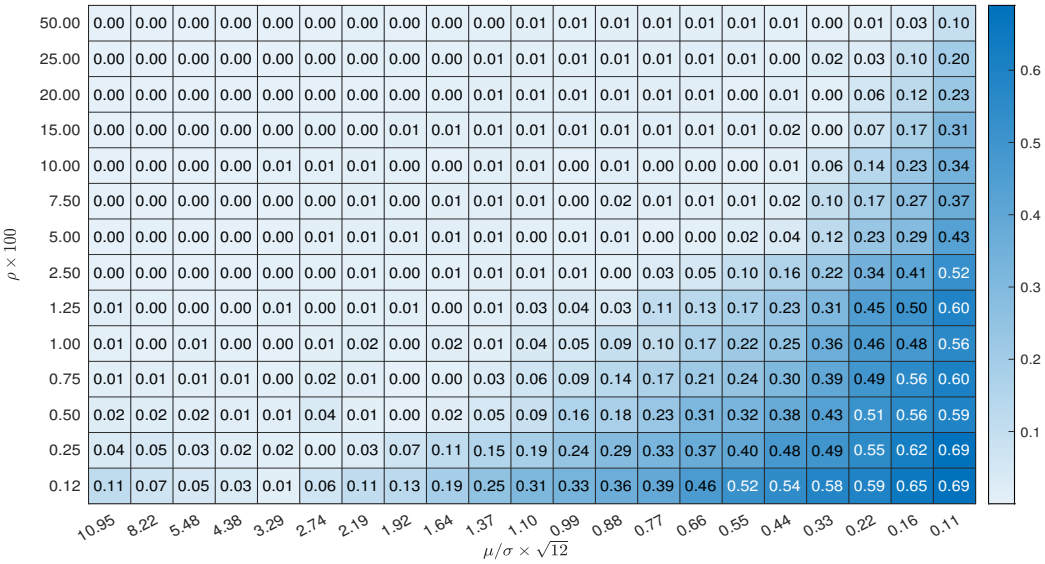

Figure 6: Comparison between $\widehat{S}^\star$ and $S^\star$

Note: The figure reports the error between $\widehat{S}^\star$ and $S^\star$ defined as $\left|\widehat{S}^\star - S^\star\right| / \left(1 + S^\star\right)$. The simulation setting is based on model (4), in which a $100 \times \rho\%$ of assets have αs that correspond to an annualized Sharpe ratio $\mu/\sigma \times \sqrt{12}$. In this experiment, $N = 1,000$ and $T = 20$ years.

4 Empirical Analysis of US Equities

To demonstrate the empirical relevance of the statistical limit of arbitrage, we study US monthly equity returns from January 1965 to December 2020.

4.1 Data Preprocessing

We adopt a multi-factor model with 16 characteristics and 11 GICS sectors, which are selected to incorporate empirical insight 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, see Chen and Zimmermann (2020) for construction details.

We download the monthly return data for individual equities from CRSP. We take a number of steps to preprocess the data. First, we single out delisted stocks, and attach delisting returns as their last returns (on the delisting months). Next, we merge the returns data with the aforementioned characteristics database using permnos. The total number of unique permnos on average per month is 6,536. We then apply the usual filters (share codes 10 and 11 and exchange codes 1, 2, and 3) to the database, to eliminate (part of) the sampling periods for stocks that fail to meet these criteria. The remaining average number of stocks per month is 4,756. For stocks whose returns are missing for more than 3 months, we eliminate the missing periods, otherwise we fill the missing returns by zeros.

We now deal with missing characteristics. We start by removing all characteristics data for any stocks since their delisting months. We then fill missing GICS codes with the corresponding stocks' most recent records prior to their missing dates. Stocks without any GICS codes over the entire sample period are eliminated. If the GICS codes become available later in the sample for some stocks, their sample prior to the first dates when GICS become available are eliminated, which mainly occurs prior to 1990. With GICS information, we adopt a two-step procedure to fill in other missing characteristics. For any missing value in a stock's characteristic, we fill it with the sector-wise median of this characteristic each month. If a characteristic's values are not available for an entire sector in a certain month, we fill them with this characteristic's cross-sectional median over all stocks in this month. After data preprocessing, the final average number of stocks per month is reduced to 4,067.

The resulting panel is not balanced, because we do not fill in missing data before a stock's IPO or after its delisting. Our approach to filling missing data thereby avoids forward-looking bias.

4.2 Model Performance

At the end of each month, we run cross-sectional regressions of next month returns onto the 27 cross-sectional predictors (including the intercept). We do so using all stocks in the current month's cross sections. Following Gu et al. (2020), the 16 characteristics are rank-normalized within each cross-section, alleviating the impact of extreme outliers in characteristics, though this barely changes any follow-up results.

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

on the decline since the beginning of the sample till 1990s. This coincides with the period when the number of stocks in the US equity markets increases. The R^2s are moderately low, with an average of 8.25%. The low R^2s suggest that a substantial portion of cross-sectional variation of individual equity returns is idiosyncratic noise. Therefore, learning alphas from residuals of the factor model is an incredibly difficult statistical task.

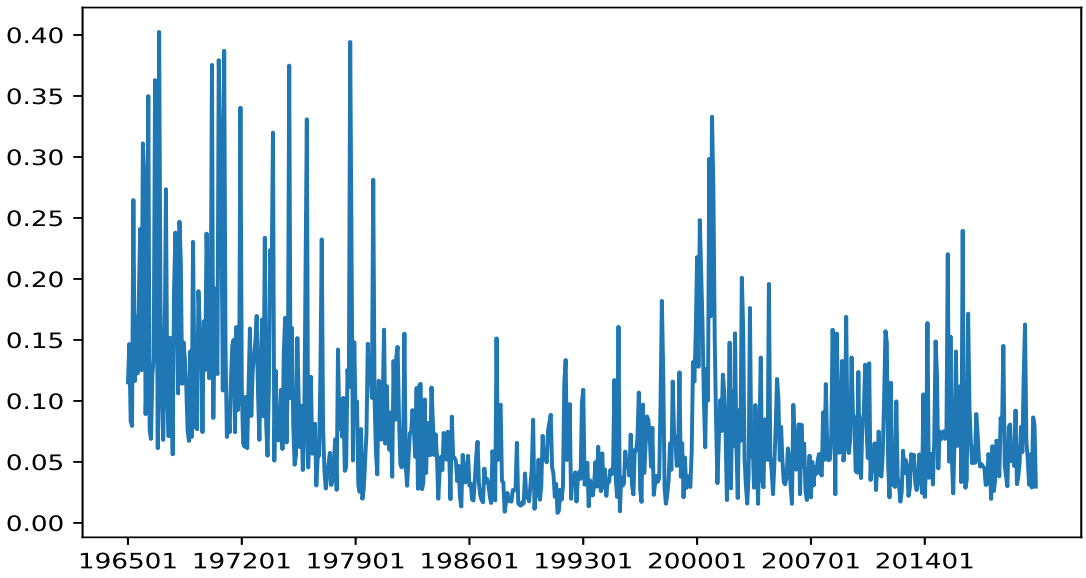

Figure 7: Time-series of the Cross-sectional R^2s

4.3 Rare and Weak Alphas

We now study the statistical properties of 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 impose that all residuals have at least 60 observations. This ensures enough sample size for inference on alpha, although the distribution of alphas' t-statistics turns out not sensitive to this requirement. Figure 8 provides histograms of the t-statistics and Sharpe ratios for alphas of all 12,415 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 6.35% of the t-statistics exceed 2.0 in magnitude, and more than 0.63% exceed 3.0. This suggests that truly significant alphas are extremely rare. Moreover, the largest Sharpe ratio of all individual stocks' alphas is rather modest, about 1.699. Only 0.505% 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.

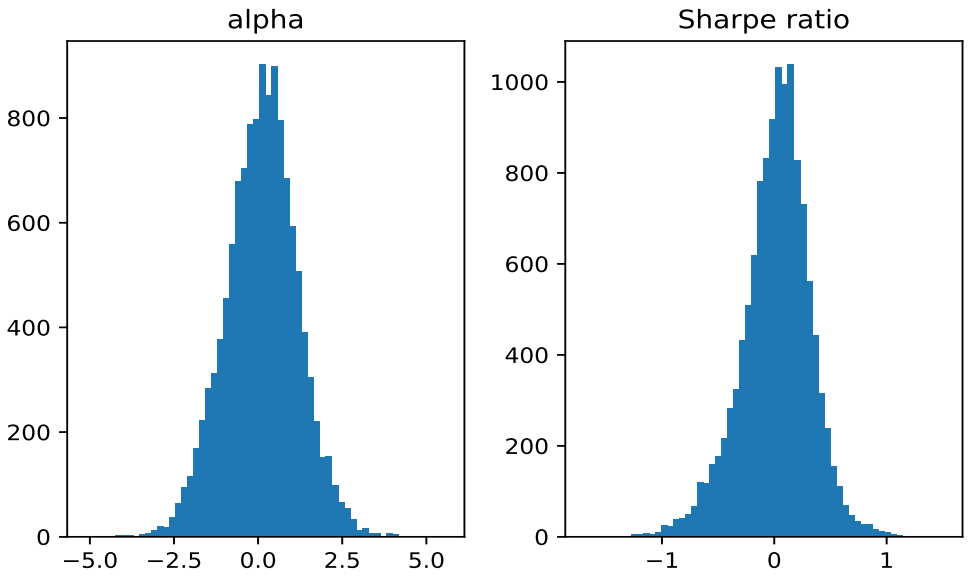

Figure 8: Histograms of the t-Statistics and Sharpe Ratios of Estimated Alphas

Note: The figure provides the histograms of the 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 12,482.

4.4 Performance of Arbitrage Portfolios

Throughout we assume alphas do not vary over time. If alphas are driven by some observable characteristics, then it is possible to construct a factor using these characteristics via cross-sectional regressions, which turns "alpha" into risk premia. In this regard, alphas are meaningless without reference to a specific factor model. Extracting more "factors" out of alphas would lead to even smaller arbitrage profits.

We now compare arbitrage portfolios based on various strategies, including the optimal strategy, the cross-sectional regression (CSR) approach, the multiple-testing based procedure (BH), and LASSO approach. The ridge approach is omitted, since it is equivalent to the CSR.

Specifically, at the end of each month, we build optimal portfolio weights using these strategies. We only invest in stocks with a continuous record for at least 96 months. We rebalance these portfolios at the end of each month, with weights recalculated using a 120-month rolling window. Both Lasso and the optimal strategy require a tuning parameter. Out of the 10-year rolling window, we leave the last 2 years as the validation sample for tuning parameter selection. As expected, optimal tuning parameter is difficult to select, which undermines the performance of both strategies.

All these strategies yield similar Sharpe ratios. BH and OPT tie for the top of the chart, yielding 0.497 and 0.496, respectively, followed by CSR that scores 0.450. The LASSO approach only obtains 0.384. The Sharpe ratios of different strategies are not influenced by risk aversion, though the cumulative returns are. 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 9.

Closely examining these strategies reveals more insight. BH is highly conservative. Out of 46 years of out-of sample trading months (1975/01 - 2020/12), 289 months have no trading activities. The largest number of stocks selected for trading in a month is 10, and the average over all non-zero periods is 2.43. In contrast, CSR trade all stocks that meet our trading criteria, with an average of 2,366 stocks per month. OPT almost does so, with an average of 2,359. The number of stocks traded by LASSO is rather volatile, varying between none and all stocks from month to month, with an average of 757.6 per month. This is likely caused by the noise in the tuning process.

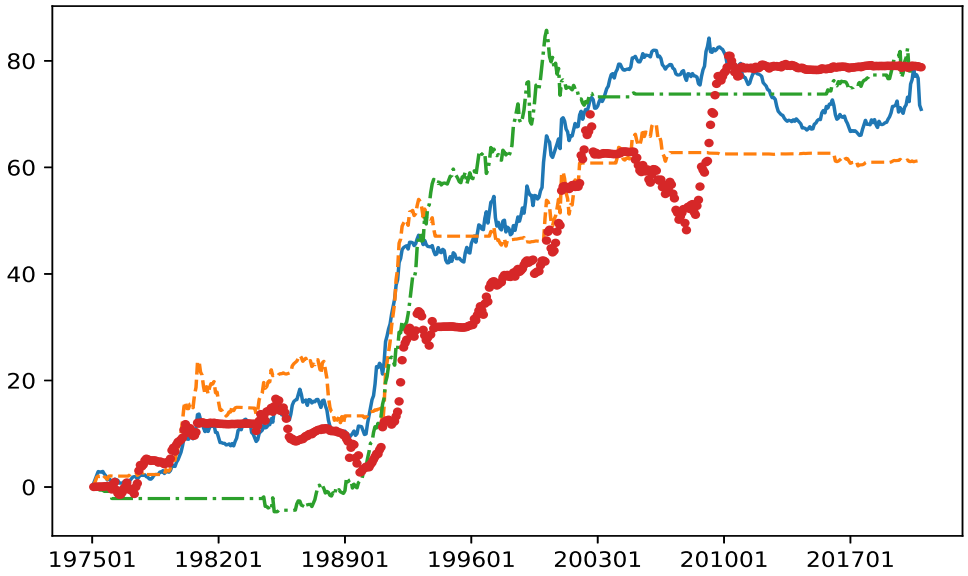

Figure 9: Normalized Cumulative Returns of Arbitrage Portfolios

Note: This figure compares the cumulative returns of OPT (red dotted), CSR (blue solid), BH (green dot-dashed), and LASSO (orange 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.

We also calculate the perceived Sharpe ratios using (15), and provide a time-series plot of $\widehat{S}^\star$ in Figure 10. We also compare it with the biased estimates $\widetilde{S}^\star$ using (16). We observe a huge gap between the estimated perceived Sharpe ratios using these formulae. As predicted by Proposition 2, $\widetilde{S}^\star$ overestimates $S^\star$, though it guarantees positive values. Our estimate $\widetilde{S}^\star$ is averaged around 2.55 (we truncate negative estimates by 0), but can sometimes exceed 7.5. These estimates are far greater than the feasible Sharpe ratios we obtain for any of these strategies. That said, even the infeasible Sharpe ratios can be as low as 0 for certain periods of the sample. The feasible portfolio returns seem in agreement with the prediction. For instance, OPT, LASSO, and BH's cumulative returns are almost flat post 2010, whereas CSR has negative returns.

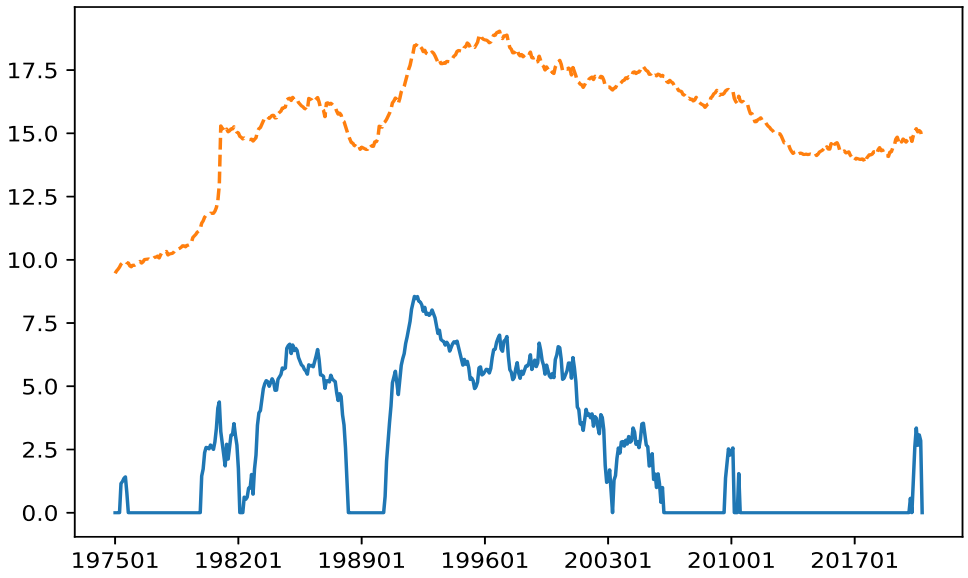

Figure 10: Time Series of Sharpe Ratios

Note: The figure compares naive estimates (orange dashed) and their bias-corrected estimates (blue solid) of the infeasible Sharpe ratios based on a rolling window of 120 months.

5 Conclusion

Taking stock, our paper provides a new theoretical framework to understanding the implications of statistical learning in asset pricing. In the age of big data, rational expectations assumption often fails to retain its relevance in practice, and hence understanding its limitation and the role of statistical learning is vitally important. We introduce new econometric tools in the spirit of nonparametric empirical Bayes, which could be adopted in other contexts.

The empirical message should be confined within the context of monthly rebalancing strategies via linear factor models. The gap between feasible and infeasible Sharpe ratios will further increase if arbitrageurs face additional statistical challenges, e.g., model misspecification, omitted factors, weak factors, large non-sparse idiosyncratic covariance matrix, etc. Consequently, the empirical gap should remain for any arbitrageurs, including those who engage in higher frequency trading or use more complex nonlinear models.

More broadly, existing literature have documented impressive Sharpe ratios on various machine learning based trading strategies. 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 in investment. Our paper provides a theoretical framework to understand the "upper bound" on the performance of any strategy in the specific context of arbitrage pricing theory, tying together this statistical limit with economic rationale.

On a side note, our theoretical and empirical analyses also have implications on the econometric

analysis in asset pricing. Examining the economic performance of asset pricing models is as important as and complementary to statistical tests. The criteria of a good statistical test are primarily statistical in nature, such as Type I and Type II errors, false discovery rate, etc; whereas in practice, it is the economic performance that agents in the economy fundamentally care about. There is often a wedge between these two objectives. For instance, a statistical procedure that guards against false discovery rate may be overly conservative for investment purpose; rejection by a powerful test statistic may not necessarily lead to the practical irrelevance of an economic theory. While the asset pricing literature has seen an explosion of statistical machine learning tools imported from other areas, we caution against their use without guidance of economics.

References

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets 5*(1), 31–56.

Andrews, D. W. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*(5), 2153–2211.

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.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *Journal of Finance 61*(1), 259–299.

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.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *Journal of Finance 43*(2), 507–528.

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.

Cooper, M. J., H. Gulen, and M. J. Schill (2008). Asset growth and the cross-section of stock returns. *Journal of Finance 63*(4), 1609–1651.

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

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.

Fama, E. F. and K. R. French (2006). Profitability, investment and average returns. *Journal of Financial Economics 82*(3), 491–518.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*, 607–636.

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

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., 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.

Gu, S., B. Kellly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Gu, S., B. T. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics 222*, 429–450.

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

Hahn, J. and H. Lee (2009). Financial constraints, debt capacity, and the cross-section of stock returns. *Journal of Finance 64*(2), 891–921.

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.

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.

Heston, S. L. and R. Sadka (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics 87*(2), 418–445.

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.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *Journal of Finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

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

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., S. Nagel, and S. Santosh (2018). Interpreting factor models. *Journal of Finance 73*(3), 1183–1223.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation, and risk. *Journal of Finance 49*(5), 1541–1578.

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

Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review 4*(1), 1–44.

Litzenberger, R. H. and K. Ramaswamy (1979). The effects of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics 7*, 163–195.

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.

Martin, I. and S. Nagel (2021). Market efficiency in the age of big data. *Journal of Financial Economics, forthcoming*.

Moskowitz, T. J. and M. Grinblatt (1999). Do industries explain momentum? *Journal of Finance 54*(4), 1249–1290.

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.

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

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.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review 71*(3), 289–315.

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

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

Uppal, R. and P. Zaffaroni (2018). Robust portfolio choice. Technical report, EDHEC Business School and Imperial College London.

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

Appendix A Mathematical Proofs

A.1 Proof of Theorem 1 and Proposition 1, and Corollary 2

Proof of Theorem 1. To simplify the notation, we omit the dependence of β, Σ on N, and $\widehat{w}$ on N and T. All limits are taken as $N \to \infty$. The derivation applies to either fixed T or $T \to \infty$ together with N.

We first note that, given (1), conditioning on $\mathcal{G}$ is equivalent to conditioning on the information set generated by

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

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

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

$$S(\widehat{w}) = \mathrm{E}(\widehat{w}^\intercal r_{t+1}|\mathcal{G})/\mathrm{Var}(\widehat{w}^\intercal r_{t+1}|\mathcal{F}_t)^{1/2} + \mathcal{E}/\mathrm{Var}(\widehat{w}^\intercal r_{t+1}|\mathcal{F}_t)^{1/2}. \tag{A.1}$$

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

$$\mathrm{P}\left(|\mathcal{E}|/\|\widehat{w}\| \geq \epsilon\right) \leq \mathrm{E}(\mathcal{E}^2/\|\widehat{w}\|^2)/\epsilon^2 = \mathrm{E}(\widehat{w}^\intercal \mathrm{Var}(\alpha|\mathcal{G})\widehat{w}/\|\widehat{w}\|^2)/\epsilon^2. \tag{A.2}$$

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

$$\mathrm{E}(\widehat{w}^\intercal \mathrm{Var}(\alpha|\mathcal{G})\widehat{w}/\|\widehat{w}\|^2) \leq \mathrm{E}\left(\max_{i\leq N} \mathrm{Var}(\alpha_i|\mathcal{G})\right) \leq \mathrm{E}\left(\max_{i\leq N} \alpha_i^2\right) = o(1), \tag{A.3}$$

where the last step comes from condition (c) of Assumption 1. Combining (A.75) and (A.76), and using $\mathrm{Var}(\widehat{w}^\intercal r_{t+1}|\mathcal{F}_t) = \widehat{w}^\intercal \Sigma \widehat{w} \geq \lambda_{\min}(\Sigma_u)\|\widehat{w}\|^2 \gtrsim_{\mathrm{P}} \|\widehat{w}\|^2$, we obtain

$$|\mathcal{E}|/\mathrm{Var}(\widehat{w}^\intercal r_{t+1}|\mathcal{F}_t)^{1/2} \lesssim_{\mathrm{P}} |\mathcal{E}|/\|\widehat{w}\| = o_{\mathrm{P}}(1). \tag{A.4}$$

(A.4) and (A.1) lead to

$$S(\widehat{w}) = \widehat{w}^\intercal \mathrm{E}(r_{t+1}|\mathcal{G})(\widehat{w}^\intercal \Sigma \widehat{w})^{-1/2} + o_{\mathrm{P}}(1). \tag{A.5}$$

Furthermore, applying Cauchy-Schwarz inequality, we obtain

$$|\widehat{w}^\intercal \mathrm{E}(r_{t+1}|\mathcal{G})|^2 (\widehat{w}^\intercal \Sigma \widehat{w})^{-1} \leq \mathrm{E}(r_{t+1}|\mathcal{G})^\intercal \Sigma^{-1} \mathrm{E}(r_{t+1}|\mathcal{G}). \tag{A.6}$$

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

$$\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\beta(\Sigma_v^{-1} + \beta^\intercal\Sigma_u^{-1}\beta)^{-1}\beta^\intercal\Sigma_u^{-1}. \tag{A.7}$$

By direct calculations, we have

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

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

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

Therefore, using the fact that $\lambda_{\min}(\beta^\intercal\beta) \gtrsim_{\mathrm{P}} N$ and that $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$ in light of condition (a) and (d) of Assumption 1, we have

$$\lambda_{\max}((\beta^\intercal\Sigma_u^{-1}\beta)^{-1}) = \lambda_{\min}^{-1}(\beta^\intercal\Sigma_u^{-1}\beta) \leq \lambda_{\min}^{-1}(\beta^\intercal\beta)\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} N^{-1}. \tag{A.8}$$

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

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

we have

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

which in turn leads to

$$\gamma^\intercal \beta^\intercal \Sigma^{-1}\beta\gamma = \gamma^\intercal \Sigma_v^{-1}\gamma + o_{\mathrm{P}}(1). \tag{A.9}$$

Next, we show

$$\mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma^{-1}\beta\gamma = o_{\mathrm{P}}(1). \tag{A.10}$$

Notice that $\mathrm{E}\left(\mathrm{E}(\alpha|\mathcal{G})|\Sigma,\beta\right) = \mathrm{E}\left(\alpha|\Sigma,\beta\right) = \mathrm{E}\left(\alpha|\Sigma\right) = 0$ (by conditions (c) and (e) of Assumption 1), and that, conditionally on (Σ,β), $\mathrm{E}(\alpha_i|\mathcal{G})$ is independent across i. Therefore,

$$\mathrm{E}\left(\left(\mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma^{-1}\beta\gamma\right)^2 |\Sigma,\beta\right) \leq \sum_{i\leq N} \mathrm{E}\left(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma,\beta\right) \max_{j\leq N}(\gamma^\intercal \beta^\intercal \Sigma^{-1})_j^2. \tag{A.11}$$

On the other hand, from (A.7), we obtain

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

Because of $\lambda_{\min}(\Sigma_v) \gtrsim 1$, $\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} 1$, $\|\Sigma_u\|_{\mathrm{MAX}} \leq \|\Sigma_u\| \lesssim_{\mathrm{P}} 1$, $\lambda_{\min}(\Sigma_u) \gtrsim_{\mathrm{P}} 1$, Σ_u is diagonal, and (A.8), we have

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

Hence, we have, for all positive fixed ϵ,

$$\mathrm{P}\left(|\mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma^{-1}\beta\gamma| \geq \epsilon|\Sigma,\beta\right) \leq \mathrm{E}\left(\left(\mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma^{-1}\beta\gamma\right)^2 |\Sigma,\beta\right)/\epsilon^2 = o_{\mathrm{P}}(1), \tag{A.12}$$

where the last equality comes from (A.11) and that $\mathrm{E}\left(\sum_{i\leq N}\mathrm{E}\left(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma,\beta\right)\right) \leq \sum_{i\leq N}\mathrm{E}(\alpha_i^2) = o(N)$ by condition (c) of Assumption 1. Since $\mathrm{P}\left(|\mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma^{-1}\beta\gamma| \geq \epsilon|\Sigma,\beta\right) \leq 1$ are uniformly bounded for all N (by definition), we obtain by taking expectations on both sides of (A.12) that, for all positive fixed ϵ,

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

which is equivalent to (A.10).

Finally, we derive

$$\mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma^{-1}\mathrm{E}(\alpha|\mathcal{G}) = \mathrm{E}(\alpha|\mathcal{G})^\intercal \Sigma_u^{-1}\mathrm{E}(\alpha|\mathcal{G}) + o_{\mathrm{P}}(1). \tag{A.13}$$

Following the same derivation for (A.11), we obtain

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

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

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

Then given the above result that $\mathrm{E}\left(\sum_{i\leq N}\mathrm{E}\left(\mathrm{E}(\alpha_i|\mathcal{G})^2|\Sigma,\beta\right)\right)=o(N)$, we obtain that $\mathrm{E}\left(\left\|\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\beta\right\|_{\mathrm{F}}^2|\Sigma,\beta\right)=o_{\mathrm{P}}(N)$. Therefore, similar to the derivation of (A.10), we obtain

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

On the other hand, using (A.8), we obtain

$$\|(\Sigma_v^{-1}+\beta^\intercal\Sigma_u^{-1}\beta)^{-1}\|=\lambda_{\min}^{-1}(\Sigma_v^{-1}+\beta^\intercal\Sigma_u^{-1}\beta)\leq\lambda_{\max}((\beta^\intercal\Sigma_u^{-1}\beta)^{-1})\lesssim_{\mathrm{P}}N^{-1}. \tag{A.14}$$

Then, using (A.14), we have

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

and hence, in light of (A.7), we obtain (A.13).

Given that $\mathrm{E}(r_{t+1}|\mathcal{G})=\mathrm{E}(\alpha|\mathcal{G})+\beta\gamma$, it follows from (A.9), (A.10), and (A.13) that

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

In light of (A.6), we conclude the proof of the first statement.

Furthermore, if $\widehat{w}^\intercal\beta=0$, then it follows that $\widehat{w}^\intercal r_t=\widehat{w}^\intercal(\alpha+u_t)$ and $\widehat{w}^\intercal\Sigma\widehat{w}=\widehat{w}^\intercal\Sigma_u\widehat{w}$. Equation (A.5) then becomes

$$S(\widehat{w})=\widehat{w}^\intercal\mathrm{E}(\alpha|\mathcal{G})(\widehat{w}^\intercal\Sigma_u\widehat{w})^{-1/2}+o_{\mathrm{P}}(1). \tag{A.15}$$

Similar to (A.6), a direct application of Cauchy-Schwarz inequality

$$|\widehat{w}^\intercal\mathrm{E}(\alpha|\mathcal{G})|^2(\widehat{w}^\intercal\Sigma_u\widehat{w})^{-1}\leq\mathrm{E}(\alpha|\mathcal{G})^\intercal\Sigma_u^{-1}\mathrm{E}(\alpha|\mathcal{G})=S(\mathcal{G})^2, \tag{A.16}$$

which concludes the proof. ∎

Proof of Proposition 1. Step 1. We have established in the beginning of the proof of Theorem 1 that the $\mathcal{G}$-conditional distribution of α_i is the same as the $\mathcal{G}_i$-conditional distribution of α_i, where $\mathcal{G}_i$ is

the information set generated by $\{(\alpha_i + u_{i,s}) : t - T + 1 \leq s \leq t\}$ and σ_i. Note that $u_{i,s}$ is centered normal, we have that the conditional probability density of $\{r^*_{i,s} \coloneqq \alpha_i + u_{i,s}, t - T + 1 \leq s \leq t\}$ given α_i and σ_i, denoted by $p(r^*_i|\alpha_i, \sigma_i)$, is

$$p(r^*_i|\alpha_i, \sigma_i) = \prod_{t-T+1\leq s\leq t} \sigma_i^{-1}\phi\left(\frac{r^*_{i,s} - \alpha_i}{\sigma_i}\right) = \phi(T^{1/2}\sigma_i^{-1}(\bar{r}^*_i - \alpha_i))f(r^*_i).$$

Here $\bar{r}^*_i = T^{-1}\sum_{t-T+1\leq s\leq t} r^*_{i,s}$ and $f(r^*_i)$ is a function of r^*_i that does not depend on α_i. Hence, applying Bayes' theorem, we have

$$\begin{aligned}
\mathrm{E}(\alpha_i|\mathcal{G}) = \mathrm{E}(\alpha_i|\mathcal{G}_i) = \sigma_i\mathrm{E}(s_i|\mathcal{G}_i) &= \sigma_i \int x p(s_i = x|r^*_i, \sigma_i)dx \\
&= \sigma_i \int x \frac{p(r^*_i|s_i = x, \sigma_i)p(s_i = x|\sigma_i)}{\int p(r^*_i|s_i = x', \sigma_i)p(s_i = x'|\sigma_i)dx'}dx \\
&= \sigma_i \int x \frac{p(r^*_i|s_i = x, \sigma_i)p_s(x)}{\int p(r^*_i|s_i = x', \sigma_i)p_s(x')dx'}dx \\
&= \sigma_i \int x \frac{\phi(\widehat{z}_i - T^{1/2}x)p_s(x)}{\int \phi(\widehat{z}_i - T^{1/2}x')p_s(x')dx'}dx = \sigma_i\psi(\widehat{z}_i),
\end{aligned}$$

where $\widehat{z}_i = T^{1/2}\sigma_i^{-1}\bar{r}^*_i$, $p_s(\cdot)$ 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. This concludes the proof of the first statement.

Step 2. Apparently, $\widehat{z}_i = T^{1/2}\sigma_i^{-1}(\alpha_i + \bar{u}_i)$ is i.i.d. across i, whose conditional distribution given (α_i, σ_i) is normal, it follows that its unconditional density function $p(a) = \mathrm{E}(\phi(a - T^{1/2}s_i))$. By direction calculation and the definition of S^{OPT} in the statement of the proposition, we have

$$\mathrm{E}(\mathrm{E}(\alpha_i|\mathcal{G})^\intercal\Sigma_u^{-1}\mathrm{E}(\alpha_i|\mathcal{G})) = \sum_i \mathrm{E}\left(\mathrm{E}(s_i|\mathcal{G})^2\right) = N\int \psi(a)^2p(a)da = \left(S^{\mathrm{OPT}}\right)^2. \tag{A.17}$$

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

$$\begin{aligned}
& \mathrm{E}\left(\left|\mathrm{E}(s_i\mathbb{1}_{\{|s_i|\leq c_N\}}|\mathcal{G})^2 - \mathrm{E}(s_i|\mathcal{G})^2\right|\right) \\
\leq\ & \mathrm{E}\left(\mathrm{E}(s_i\mathbb{1}_{\{|s_i|>c_N\}}|\mathcal{G})^2\right) + 2\mathrm{E}\left(\left|\mathrm{E}(s_i|\mathcal{G})\mathrm{E}(s_i\mathbb{1}_{\{|s_i|>c_N\}}|\mathcal{G})\right|\right) \\
\leq\ & \mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|>c_N\}}) + 2\sqrt{\mathrm{E}\left(\mathrm{E}(s_i|\mathcal{G})^2\right)\mathrm{E}\left(\mathrm{E}(s_i\mathbb{1}_{\{|s_i|>c_N\}}|\mathcal{G})^2\right)} \\
\leq\ & \mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|>c_N\}}) + 2\sqrt{\mathrm{E}\left(\mathrm{E}(s_i|\mathcal{G})^2\right)\mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|>c_N\}})} \\
\leq\ & c_N N^{-1} + \sqrt{\mathrm{E}\left(\mathrm{E}(s_i|\mathcal{G})^2\right)c_N N^{-1}},
\end{aligned} \tag{A.18}$$

where the last step comes from condition (a) of Assumption 2. Then we have

$$\begin{aligned} & \mathrm{E}\left(\left|\sum_i \mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2 - \sum_i \mathrm{E}(s_i|\mathcal{G})^2\right|\right) \\ \le\ & \sum_i \mathrm{E}\left(\left|\mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2 - \mathrm{E}(s_i|\mathcal{G})^2\right|\right) \\ \le\ & c_N + \sqrt{\sum_i \mathrm{E}\left(\mathrm{E}(s_i|\mathcal{G})^2\right) c_N} = o\left(1 + S^{\mathrm{OPT}}\right), \end{aligned} \tag{A.19}$$

where the second inequality is a direct result of (A.18), and the last estimate is given by (A.17). From (A.19) and (A.17), it follows, respectively, using Markov's inequality and triangle inequality that

$$\sum_i \mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2 = \sum_i \mathrm{E}(s_i|\mathcal{G})^2 + o_{\mathrm{P}}\left(1 + S^{\mathrm{OPT}}\right), \tag{A.20}$$

$$\mathrm{E}\left(\sum_i \mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2\right) = \left(S^{\mathrm{OPT}}\right)^2 + o\left(1 + S^{\mathrm{OPT}}\right). \tag{A.21}$$

Further, we have

$$\begin{aligned} \mathrm{Var}\left(\sum_i \mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2\right) &= \sum_i \mathrm{Var}\left(\mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2\right) \\ &\le c_N^2 \sum_i \mathrm{E}\left(\mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})^2\right) = o\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \end{aligned} \tag{A.22}$$

For the first line, we use that $\mathrm{E}(s_i \mathbb{1}_{\{|s_i|\le c_N\}}|\mathcal{G})$ is independent across i. The second line is obvious as $|s_i|\mathbb{1}_{\{|s_i|\le c_N\}} \le c_N$. The last line comes from (A.21). Combining (A.21) and (A.22), we obtain

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

Along with (A.20), we obtain

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

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

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

we conclude the proof. ∎

Proof of Corollary 2. Because of the tail condition $\mathrm{E}(\alpha_i^2 \mathbb{1}_{\{|\alpha_i|\ge c_N\}}) \le c_N N^{-1}$ for some sequence

$c_N \to 0$, we have

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

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

$$\alpha^\intercal\alpha = \sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i|<c_N\}} + o_{\mathrm{P}}(1), \quad \mathrm{E}\left(\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i|<c_N\}}\right) = \mu^2\rho N. \tag{A.23}$$

On the other hand, it holds that

$$\mathrm{Var}\left(\sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i|<c_N\}}\right) \leq \sum_i \mathrm{E}(\alpha_i^4 \mathbb{1}_{\{|\alpha_i|<c_N\}}) \leq c_N^2 \sum_i \mathrm{E}(\alpha_i^2) = c_N^2\mu^2\rho N. \tag{A.24}$$

Combining (A.23) and (A.24), we obtain

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

As a result, it holds that

$$S^\star = \sigma^{-1}\sqrt{\alpha^\intercal\alpha} = \sigma^{-1}\mu(\rho N)^{1/2} + o_{\mathrm{P}}(1). \tag{A.25}$$

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

$$\begin{aligned}
\psi(a) &= \frac{\mu\rho\phi(a - T^{1/2}\mu/\sigma) - \mu\rho\phi(a + T^{1/2}\mu/\sigma)}{(2-2\rho)\phi(a) + \rho\phi(a - T^{1/2}\mu/\sigma) + \rho\phi(a + T^{1/2}\mu/\sigma)}, &\text{(A.26)}\\
\left(S^{\mathrm{OPT}}\right)^2 &= \frac{\mu\rho N}{2\sigma^2}\int \psi(a)(\phi(a - T^{1/2}\mu/\sigma) - \phi(a + T^{1/2}\mu/\sigma))da. &\text{(A.27)}
\end{aligned}$$

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

$$\begin{aligned}
\sup_{a\geq C} \frac{\rho\phi(a)}{\phi(a - T^{1/2}\mu/\sigma)} &= \exp\left(\log\rho + T^{1/2}\mu\sigma^{-1}\left(\frac{1}{2}T^{1/2}\mu\sigma^{-1} - C\right)\right)\\
&\leq \exp\left(\log\rho + \frac{1}{2}\left(\sqrt{-2\log\rho} + C\right)\left(\sqrt{-2\log\rho} - C\right)\right) \leq 1. \quad \text{(A.28)}
\end{aligned}$$

On the other hand, in light of (A.26) and (A.27), we have

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

We hence obtain from (A.28) that, for N sufficiently large,

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

This proves the "if" part, given (A.25) and that $\mu^2\rho N/\sigma^2$ does not vanish. Now suppose $T^{1/2}\mu\sigma^{-1}-\sqrt{-2\log\rho}\to\infty$. Then, for all fixed $x>0$, we have, for sufficiently large N,

$$\begin{aligned}
\sup_{a:|a|\leq x}\frac{\phi(a+T^{1/2}\mu/\sigma)}{\rho\phi(a)} &= \exp\left(-\log\rho-T^{1/2}\mu\sigma^{-1}\left(\frac{1}{2}T^{1/2}\mu\sigma^{-1}-x\right)\right)\\
&\leq \exp\left(-\log\rho-\frac{1}{2}\left(\sqrt{-2\log\rho}+c_N^{-1}\right)\left(\sqrt{-2\log\rho}+c_N^{-1}\right)\right)\\
&\leq \exp\left(-c_N^{-2}/2\right)\to 0, \qquad \text{(A.29)}\\
\sup_{a:|a|\leq x}\frac{\phi(a+2T^{1/2}\mu/\sigma)}{\phi(a)} &= \exp\left(-2T^{1/2}\mu\sigma^{-1}(T^{1/2}\mu\sigma^{-1}-x)\right)\to 0. \qquad \text{(A.30)}
\end{aligned}$$

Given (A.26), it holds that

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

Substituting (A.30) into the numerator, and (A.29) and (A.30) into the denominator, we obtain that, for all fixed $x>0$,

$$\sup_{a:|a|\leq x}\left|\mu^{-1}\psi\left(a+T^{1/2}\mu/\sigma\right)-1\right|\to 0. \qquad \text{(A.31)}$$

Since the integrand of (A.27) is always positive and even in a, it holds that, for all fixed $x>0$,

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

Here the second inequality comes from (A.30), the third inequality is a result of (A.31), and the last inequality is obvious. Because this result holds for all fixed $x>0$, the "only if" part is proved. ∎

A.2 Proof of Theorem 2

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 ψ, which we recall stands for $\Sigma_u^{-1/2}\mathrm{E}(\alpha|\mathcal{G})$, and the estimate $\widehat{\psi} := (\widehat{\psi}(\widehat{z}_1), \ldots, \widehat{\psi}(\widehat{z}_N))^\intercal$, measured by L2 norm, is small compared to S^{OPT}. This leads to that the Sharpe ratio generated by $\widehat{w}^{\mathrm{OPT}} = \mathbb{M}_\beta \Sigma_u^{-1/2}\widehat{\psi}$ converges to S^{OPT}, proved in the last step.

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

Step 1. Throughout the proof, we use the following notation, introduced in the statement of Proposition 1,

$$\widetilde{z}_i = T^{-1/2}\sum_{s\in\mathcal{T}}(s_i + \varepsilon_{i,s}), \quad p(a) = \mathrm{E}(\phi(a - T^{1/2}s_i)), \quad \psi(a) = \frac{\int x\phi(a - T^{1/2}x)p_s(x)dx}{p(a)}. \tag{A.32}$$

As in that statement, $p(a)$ is the density of $\widetilde{z}_i$, and $\psi(a)$ is the expectation of s_i, conditional on $\widetilde{z}_i = a$. We also write for convenience $\widetilde{s}_i := T^{-1/2}\widetilde{z}_i$.

Intuitively, for assets with large $\widetilde{z}_i$, $\widetilde{s}_i$ is a relatively precisely estimate the true s_i. In contrast, for assets with small $\widetilde{z}_i$, more likely $\widetilde{z}_i$ is driven by noise. As a result, we introduce $B = \{i \le N : |\widetilde{z}_i| \le \widetilde{k}_N\}$ to separate the two cases, where $\widetilde{k}_N = k_N^{-2}$. Moreover, we set $\widehat{\psi}$ and ψ as the N-dimensional vectors with entries $\widehat{\psi}_i := \widehat{\psi}(\widehat{z}_i)$ and $\psi_i := \psi(\widetilde{z}_i)$. It holds that

$$\|\widehat{\psi} - \psi\|^2 \le \sum_{i\in B}(\widehat{\psi}_i - \psi_i)^2 + \sum_{i\in B^c}(\widehat{\psi}_i - \psi_i)^2. \tag{A.33}$$

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

$$\|\widehat{\psi} - \psi\|^2 = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \tag{A.34}$$

The last step proves optimality of our portfolio strategy based on the above result. We end this step by noting that (the last part of) Proposition 1 states

$$\|\psi\| = S(\mathcal{G}) = S^{\mathrm{OPT}} + o_{\mathrm{P}}(1). \tag{A.35}$$

Step 2. This step control the magnitude of $\sum_{i\in B}(\widehat{\psi}_i - \psi_i)^2$ of (A.33). It does so by showing

$$\sum_{i\in B^c}(\psi_i - \widetilde{s}_i)^2 = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right) \quad \text{and} \quad \sum_{i\in B^c}(\widehat{\psi}_i - \widetilde{s}_i)^2 = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \tag{A.36}$$

Since $\psi_i := \psi(\widetilde{z}_i)$, to bound $\sum_{i\in B^c}(\psi_i - \widetilde{s}_i)^2$, we show that $|\psi(a) - T^{-1/2}a|$ is small. On the other hand, Tweedie's formula reads

$$\psi(a) - T^{-1/2}a = T^{-1/2}\frac{p'(a)}{p(a)}. \tag{A.37}$$

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

$$\begin{aligned}|p'(a)| &\leq \int |T^{1/2}x-a|\phi(T^{1/2}x-a)p_s(x)dx \\ &\leq b_N\int_{|T^{1/2}x-a|\leq b_N}\phi(T^{1/2}x-a)p_s(x)dx+\sup_{x:|T^{1/2}x-a|>b_N}|T^{1/2}x-a|\phi(T^{1/2}x-a) \\ &\leq b_N p(a)+\sup_{y:|y|>b_N}|y|\exp(-y^2/2). \end{aligned} \tag{A.38}$$

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

$$|p'(a)| \leq c_N \widetilde{k}_N p(a) + c_N N^{-2}. \tag{A.39}$$

It hence holds that

$$\max_i \frac{|p'(\widetilde{z}_i)|}{p(\widetilde{z}_i)} \lesssim_{\mathrm{P}} \sup_a \frac{|p'(a)|}{p(a)} 1_{\{p(a)\geq N^{-3/2}\}} \leq c_N \widetilde{k}_N. \tag{A.40}$$

The first inequality comes from(B.154) of Lemma B3. The second directly follows from (A.39) (note that $T = o(N)$ by assumption). Combining (A.40) and (A.37), we obtain

$$\mathrm{P}((\widetilde{s}_i - \psi(\widetilde{z}_i))^2 \leq c_N T^{-1}\widetilde{k}_N^2, \forall i \leq N) \geq 1 - c_N. \tag{A.41}$$

As a result,

$$\sum_{i\in B^c}(\widetilde{s}_i-\psi(\widetilde{z}_i))^2 \lesssim_{\mathrm{P}} c_N T^{-1}\widetilde{k}_N^2|B^c| \leq c_N \sum_{i\in B^c}\widetilde{s}_i^2 \lesssim_{\mathrm{P}} c_N \sum_{i\in B^c}\psi(\widetilde{z}_i)^2. \tag{A.42}$$

Here the first inequality is simply (A.41), the second holds since $\widetilde{s}_i^2 \geq T^{-1}\widetilde{k}_N^2$ for all $i \in B^c$ by definition, and the last inequality is a direct implication of the first two. Given (A.42), we obtain the first part of (A.36) by noting $\sum_{i\in B^c}\psi(\widetilde{z}_i)^2 \lesssim_{\mathrm{P}} \left(S^{\mathrm{OPT}}\right)^2 + 1$ due to (A.35).

Now we establish the second part of (A.36). By construction we have

$$\widehat{\psi}(a) - T^{-1/2}a = T^{-1/2}\frac{\widehat{p}'(a)}{\widehat{p}(a)}, \quad \text{with} \quad \widehat{p}(a) = \frac{1}{Nk_N}\sum_i \phi\left(\frac{\widehat{z}_i - a}{k_N}\right). \tag{A.43}$$

Similar to (A.38), we have, for all positive sequence b_N and all a,

$$\begin{aligned}|\widehat{p}'(a)| &\leq \frac{1}{Nk_N^2}\sum_i \frac{|\widehat{z}_i-a|}{k_N}\phi\left(\frac{\widehat{z}_i-a}{k_N}\right) \\ &\leq \frac{1}{Nk_N^2}\sum_{i:|\widehat{z}_i-a|/k_N\leq b_N}\frac{|\widehat{z}_i-a|}{k_N}\phi\left(\frac{\widehat{z}_i-a}{k_N}\right)+\frac{1}{k_N^2}\sup_{i:|\widehat{z}_i-a|/k_N>b_N}\frac{|\widehat{z}_i-a|}{k_N}\phi\left(\frac{\widehat{z}_i-a}{k_N}\right) \\ &\leq \frac{b_N}{k_N}\widehat{p}(a)+\frac{1}{k_N^2}\sup_{y:|y|>b_N}|y|\exp(-y^2/2). \end{aligned}$$

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

we obtain, for all a,

$$|\widehat{p}'(a)| \leq c_N \widetilde{k}_N \widehat{p}(a) + c_N N^{-2}. \tag{A.44}$$

Therefore, it holds that

$$\max_i \frac{|\widehat{p}'(\widehat{z}_i)|}{\widehat{p}(\widehat{z}_i)} \leq c_N \widetilde{k}_N, \tag{A.45}$$

which comes from (A.44) and that $\widehat{p}(\widehat{z}_i) \geq \frac{1}{Nk_N}$ for all i. As a result, we obtain the second part of (A.36):

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

Here the first inequality is simply substituting (A.45) into (A.43), the second inequality comes from $\max_{i \leq N} |\widehat{z}_i - \widetilde{z}_i| \leq c_N \widetilde{k}_N$ by Lemma B2, the last inequality holds by (A.35) and (the last two inequalities of) (A.42).

Step 3. To analyze $\sum_{i \in B} (\widehat{\psi}_i - \psi_i)^2$ of (A.33), we introduce an auxiliary function:

$$\bar{\psi}(a) = \frac{\int x \phi((a - T^{1/2}x)/v) p_s(x) dx}{\int \phi((a - T^{1/2}x)/v) p_s(x) dx}, \quad \text{with} \quad v := \sqrt{1 + k_N^2}. \tag{A.46}$$

$\bar{\psi}(a)$ is essentially the expectation of s_i, conditional on $\check{z}_i = a$, where $\check{z}_i \sim \mathcal{N}(T^{1/2} s_i, v^2)$, i.e., $\check{z}_i$ has slightly more noisy than $\widetilde{z}_i$. The goal is to establish

$$\sum_{i \in B} (\psi_i - \bar{\psi}(\widetilde{z}_i))^2 = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right) \quad \text{and} \quad \sum_{i \in B} (\widehat{\psi}_i - \bar{\psi}(\widetilde{z}_i))^2 = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \tag{A.47}$$

Then the triangle inequality would give us the desired bound on $\sum_{i \in B} (\widehat{\psi}_i - \psi_i)^2$. 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 (A.46), and use $\pi(a)$ to denote the numerator of $\psi(a)$ as in (A.32). 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 $\phi(y/v)$ and $\phi(y)$ are close in that, for all y,

$$\begin{aligned} |\phi(y/v) - \phi(y)| &\leq \sup_{y: |y| \leq k_N^{-1}} |\phi(y/v) - \phi(y)| + \sup_{y: |y| > k_N^{-1}} |\phi(y/v) + \phi(y)| \\ &\leq c_N k_N^{-1} \phi(y) \sup_{y: |y| \leq k_N^{-1}} |y/v - y| + c_N N^{-2} \leq c_N \phi(y) + c_N N^{-2}. \end{aligned} \tag{A.48}$$

Here we use Lemma B4 (choose $j = 0$) and that $|v^{-1} - 1| \sim k_N^2$. Using (A.48), we directly obtain that, for all a,

$$|\bar{p}(a) - p(a)| \leq \int |\phi((a - T^{1/2}x)/v) - \phi(a - T^{1/2}x)| p_s(x) dx$$

$$\leq \quad c_N \int \phi(a - T^{1/2}x) p_s(x) dx + c_N N^{-2} = c_N p(a) + c_N N^{-2}. \tag{A.49}$$

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

$$\pi(a) = \int_0^\infty x\bar{\phi}(|a|, x) p_s(x) dx, \quad \text{and} \quad \bar{\pi}(a) = \int_0^\infty x\bar{\phi}(a/v, x/v) p_s(x) dx, \tag{A.50}$$

where

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

Since $|(1-e^{-y}) - (1-e^{-y/v})| \leq c_N(1-e^{y})$ for all $y \geq 0$, it follows from (A.48) and direct calculations that, for all $a \geq 0$ and $x \geq 0$,

$$|\bar{\phi}(a/y, x/y) - \bar{\phi}(a, x)| \leq c_N \bar{\phi}(a, x) + c_N N^{-2}. \tag{A.51}$$

Substituting (A.51) into (A.50), we obtain, for all $a \geq 0$,

$$|\bar{\pi}(a) - \pi(a)| \leq c_N |\pi(a)| + c_N N^{-2} \int_0^\infty x p_s(s) dx \leq |\pi(a)| + c_N N^{-2}. \tag{A.52}$$

Here the last inequality holds by $\mathrm{E}(|s|) \leq \sqrt{\mathrm{E}(s^2)} \leq c_N$ due to condition (a) of Assumption 2. Because $\pi(a)$ and $\bar{\pi}(a)$ are both odd functions in a due to that $p_s(x)$ is a even function of x, (A.52) apparently holds for all a.

To establish from (A.49) and (A.52) that $\bar{\psi}(a)$ and $\psi(a)$ are close, we set $A \coloneqq \{a : |a| \leq \widetilde{k}_N, p(a) \geq N^{-2}\}$. Then we obtain that, for all $a \in A$,

$$\begin{aligned} |\bar{\psi}(a) - \psi(a)| \quad &= \quad \left| \frac{\bar{\pi}(a)}{\bar{p}(a)} - \frac{\pi(a)}{\bar{p}(a)} \right| + \left| \frac{\pi(a)}{\bar{p}(a)} - \frac{\pi(a)}{p(a)} \right| \\ &\leq \quad (1 + c_N) \frac{|\bar{\pi}(a) - \pi(a)|}{p(a)} + c_N \frac{|\pi(a)|}{p(a)} \leq c_N \frac{N^{-2}}{p(a)} + c_N \psi(a). \end{aligned} \tag{A.53}$$

Here the first equality is obvious, the first inequality comes from the lower bound of $p(a)$ (by the definition of A) and (A.49), the second inequality is a result of (A.52). From (A.53), it follows that, for all a satisfying $a \in A$,

$$|\bar{\psi}(a) - \psi(a)|^2 \leq c_N \frac{N^{-2}}{p(a)} + c_N \psi(a)^2. \tag{A.54}$$

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

$$N \int_A |\bar{\psi}(a) - \psi(a)|^2 p(a) da \leq c_N + c_N N \int_{-\infty}^\infty \psi(a)^2 p(a) da \leq c_N + c_N \left(S^{\mathrm{OPT}}\right)^2, \tag{A.55}$$

which comes from (A.54) and that $\int_A da \leq 2\widetilde{k}_N$. Therefore, using Chebyshev's inequality and

comparing the definitions of sets A and B, we obtain

$$\begin{aligned}\sum_{i\in B}(\psi_i-\bar\psi(\widetilde z_i))^2\mathbb{1}_{\{p(\widetilde z_i)\geq N^{-2}\}} &\lesssim_{\mathrm{P}} \mathrm{E}\sum_{i\in B}(\psi_i-\bar\psi(\widetilde z_i))^2\mathbb{1}_{\{p(\widetilde z_i)\geq N^{-2}\}}\\ &= N\int_A|\bar\psi(a)-\psi(a)|^2p(a)da\leq c_N+c_N\left(S^{\mathrm{OPT}}\right)^2,\end{aligned}$$

where the last inequality holds by (A.55). Given (B.154) of Lemma B3, we obtain the first part of (A.47).

Step 4. This step proves the second part of (A.47), i.e., we bound $\sum_{i\in B}(\widehat\psi_i-\bar\psi(\widetilde z_i))^2$. We introduce $\widetilde p(z)$ and $\widetilde\psi(z)$ that mimick $\widehat p(z)$ and $\widehat\psi(z)$ by replacing the data input $\widetilde z_i$ with $\widehat z_i$:

$$\widetilde p(z)=\frac{1}{Nk_N}\sum_i\phi\left(\frac{\widetilde z_i-z}{k_N}\right),\quad\text{and}\quad\widetilde\psi(z)=\frac{1}{\sqrt{T}}z+\frac{v^2}{\sqrt{T}}\frac{\widetilde p'(z)}{\widetilde p(z)}.\tag{A.56}$$

Then we can decompose the quantity of interest:

$$\sum_{i\in B}(\widehat\psi_i-\bar\psi(\widetilde z_i))^2\leq\sum_{i\in B}(\widetilde\psi(\widetilde z_i)-\bar\psi(\widetilde z_i))^2+\sum_{i\in B}(\widehat\psi(\widehat z_i)-\widetilde\psi(\widetilde z_i))^2.\tag{A.57}$$

We first show that $\sum_{i\in B}(\widetilde\psi(\widetilde z_i)-\bar\psi(\widetilde z_i))^2$ is small. Since we have $\widetilde p(\widetilde z_i)\geq\frac{1}{Nk_N}$ for all i, symmetric to the derivation of (A.45), we have

$$\max_i\frac{\widetilde p'(\widetilde z_i)}{\widetilde p(\widetilde z_i)}\leq c_N\widetilde k_N.\tag{A.58}$$

On the other hand, symmetric to the derivation of (A.40), we obtain

$$\max_i\frac{\bar p'(\widetilde z_i)}{\bar p(\widetilde z_i)}\lesssim_{\mathrm{P}}\max_i\frac{\bar p'(a)}{\bar p(a)}\mathbb{1}_{\{p(a)\geq N^{-3/2}\}}\lesssim c_N\widetilde k_N.\tag{A.59}$$

where for the second inequality we note $\bar p(a)\gtrsim p(a)$ for all a due to $v\geq1$. Substituting (A.58) and (A.59) into the definitions of $\widetilde\psi(z)$ and $\bar\psi(z)$ ((A.56) and (A.46)), we obtain

$$\max_i|\widetilde\psi(\widetilde z_i)-\bar\psi(\widetilde z_i)|\lesssim_{\mathrm{P}}c_N\widetilde k_NT^{-1/2}.\tag{A.60}$$

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

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

A scrutiny of their proof of the lemma reveals that this additional condition is only indispensable (a) to derive three equalities: (48), (59), and (62) (the way it is used is similar across the three), and

(b) to guarantee that $\max_{i\le N}\sqrt{T}\bar{\psi}(\widetilde{z}_i) = o(N^d)$ for every $d > 0$. In the absence of this additional condition, a weaker result holds: for every $d > d' > 0$,

$$\mathrm{E}\Bigg(\sum_i \min\{T(\widetilde{\psi}(\widetilde{z}_i) - \bar{\psi}(\widetilde{z}_i))^2, N^{d'}\}\mathbb{1}_{\{|\widetilde{z}_i|\le N^{d'}, p(\widetilde{z}_i)\ge N^{d'-1}\}}\Bigg) \lesssim N^d. \tag{A.61}$$

(A.61) turns out sufficient for establishing a desired bound on $\sum_{i\in B}(\widetilde{\psi}(\widetilde{z}_i) - \bar{\psi}(\widetilde{z}_i))^2$, which we demonstrate now. Then we have, for every $d > d' > 0$,

$$\begin{aligned}\sum_{i\in B} T(\widetilde{\psi}(\widetilde{z}_i) - \bar{\psi}(\widetilde{z}_i))^2 &\lesssim_{\mathrm{P}} \sum_{i\in B}\min\{T(\widetilde{\psi}(\widetilde{z}_i) - \bar{\psi}(\widetilde{z}_i))^2, N^{d'}\} \\ &\lesssim_{\mathrm{P}} N^d + \sum_{i\in B}\min\{T(\widetilde{\psi}(\widetilde{z}_i) - \bar{\psi}(\widetilde{z}_i))^2, N^{d'}\}\mathbb{1}_{\{p(\widetilde{z}_i)\ge N^{d'-1}\}} \lesssim N^d. \end{aligned}\tag{A.62}$$

Here the first inequality comes from (A.60), the second comes from $\mathrm{E}\left(\sum_{i\in B}\mathbb{1}_{\{p(\widetilde{z}_i)<N^{d'-1}\}}\right) \lesssim \widetilde{k}_N N^{d'}$, and the last is simply (A.61).

Next, we show that $\sum_{i\in B}(\widehat{\psi}(\widehat{z}_i) - \widetilde{\psi}(\widetilde{z}_i))^2$ is small. Lemma B2 states that

$$\max_{i\in B}|\widehat{z}_i - \widetilde{z}_i| \lesssim_{\mathrm{P}} \chi_N \coloneqq \sqrt{T/N}\big(\epsilon_N + \mathrm{E}(s_j^2)^{1/2}\big), \quad \text{with} \quad \epsilon_N \coloneqq k_N^5. \tag{A.63}$$

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

$$|\widehat{p}(\widehat{z}_i) - \widetilde{p}(\widetilde{z}_i)| \le \frac{1}{Nk_N}\sum_j\left|\phi\left(\frac{\widehat{z}_j - \widehat{z}_i}{k_N}\right) - \phi\left(\frac{\widetilde{z}_j - \widetilde{z}_i}{k_N}\right)\right| \lesssim_{\mathrm{P}} \chi_N k_N^{-2}\widetilde{p}(\widetilde{z}_i) + N^{-2}k_N^{-1}, \tag{A.64}$$

$$\begin{aligned}|\widehat{p}'(\widehat{z}_i) - \widetilde{p}'(\widetilde{z}_i)| &\le \frac{1}{Nk_N^3}\sum_j\left|(\widehat{z}_j - \widehat{z}_i)\phi\left(\frac{\widehat{z}_j - \widehat{z}_i}{k_N}\right) - (\widetilde{z}_j - \widetilde{z}_i)\phi\left(\frac{\widetilde{z}_j - \widetilde{z}_i}{k_N}\right)\right| \\ &\lesssim_{\mathrm{P}} \chi_N k_N^{-4}\widetilde{p}(\widetilde{z}_i) + N^{-2}k_N^{-1}. \end{aligned}\tag{A.65}$$

Here the first inequalities for both lines hold by definition (note $\phi'(a) = -a\phi(a)$). The second inequalities for both lines comes from substituting (A.63) into (B.158) of Lemma B4 (note $\widetilde{k}_N \lesssim k_N^{-1}$). Since $\widetilde{p}(\widetilde{z}_i) \ge \frac{1}{Nk_N}$ by definition, we obtain from (A.64) and (A.65) that

$$\max_{i\in B}\frac{|\widehat{p}(\widehat{z}_i) - \widetilde{p}(\widetilde{z}_i)|}{\widetilde{p}(\widetilde{z}_i)} \lesssim_{\mathrm{P}} \chi_N k_N^{-2} + N^{-1} \lesssim \chi_N k_N^{-2}, \tag{A.66}$$

$$\max_{i\in B}\frac{|\widehat{p}'(\widehat{z}_i) - \widetilde{p}'(\widetilde{z}_i)|}{\widetilde{p}(\widetilde{z}_i)} \lesssim_{\mathrm{P}} \chi_N k_N^{-4} + N^{-1} \lesssim \chi_N k_N^{-4}. \tag{A.67}$$

Then we have

$$\max_{i\in B}\left|\frac{\widehat{p}'(\widehat{z}_i)}{\widehat{p}(\widehat{z}_i)}-\frac{\widetilde{p}'(\widetilde{z}_i)}{\widetilde{p}(\widetilde{z}_i)}\right|\le\max_{i\in B}\frac{\widetilde{p}(\widetilde{z}_i)}{\widehat{p}(\widehat{z}_i)}\frac{|\widehat{p}'(\widehat{z}_i)-\widetilde{p}'(\widetilde{z}_i)|}{\widetilde{p}(\widetilde{z}_i)}+\max_{i\in B}\frac{\widetilde{p}(\widetilde{z}_i)}{\widehat{p}(\widehat{z}_i)}\frac{\widetilde{p}'(\widetilde{z}_i)}{\widetilde{p}(\widetilde{z}_i)}\frac{|\widehat{p}(\widehat{z}_i)-\widetilde{p}(\widetilde{z}_i)|}{\widetilde{p}(\widetilde{z}_i)}\lesssim_{\mathrm{P}}\chi_N k_N^{-4}. \tag{A.68}$$

The first line is direct algebra. Substituting (A.58), (A.66), and (A.67) into the right-hand-side of the first line, we obtain the second line. Combining (A.63) and (A.68) with the definitions of $\widehat{\psi}$ and $\widetilde{\psi}$ ((A.43) and (A.56)), we obtain

$$\sum_{i\in B}(\widehat{\psi}(\widehat{z}_i)-\widetilde{\psi}(\widetilde{z}_i))^2\le\frac{N}{T}\max_{i\in B}|\widehat{z}_i-\widetilde{z}_i|^2+\frac{N}{T}\max_{i\in B}\left|\frac{\widehat{p}'(\widehat{z}_i)}{\widehat{p}(\widehat{z}_i)}-\frac{\widetilde{p}'(\widetilde{z}_i)}{\widetilde{p}(\widetilde{z}_i)}\right|^2\lesssim_{\mathrm{P}}k_N^{-8}(\epsilon_N^2+\mathrm{E}(s_j^2)). \tag{A.69}$$

The goal is to show $\sum_{i\in B}(\widehat{\psi}(\widehat{z}_i)-\widetilde{\psi}(\widetilde{z}_i))^2=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$, which is apparently true from (A.69) if $\mathrm{E}(s_j^2)\le\epsilon_N^2$. For the case $\mathrm{E}(s_j^2)>\epsilon_N^2$, we observe

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

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

$$N\epsilon_N^4\lesssim_{\mathrm{P}}\sum_i\epsilon_N^2\mathbb{1}_{\{|s_i|\ge\epsilon_N/2\}}\lesssim_{\mathrm{P}}\sum_i\widetilde{s}_i^2\mathbb{1}_{\{|\widetilde{s}_i|\ge\epsilon_N/4\}}\lesssim\sum_{i\in B^c}\widetilde{s}_i^2\lesssim_{\mathrm{P}}1+(S^{\mathrm{OPT}})^2. \tag{A.70}$$

Here the second inequality comes from $\widetilde{s}_i-s_i=\bar{\varepsilon}_i$ and $\max_i|\bar{\varepsilon}_i|\lesssim\sqrt{(\log N)/T}$ by the uniform bound on i.i.d normal variables. The thrid inequality holds by the definition of B, and the last inequality can be established from holds by (A.35) and (the last two inequalities of) (A.42). Since $\mathrm{E}(s_j^2)\le 1+\mathrm{E}(s_j^2\mathbb{1}_{\{|s_i|>1\}})\lesssim 1$ by condition (a) of Assumption 2, it follows from (A.70) that $k_N^{-8}\mathrm{E}(s_j^2)=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$. Given (A.69), we prove $\sum_{i\in B}(\widehat{\psi}(\widehat{z}_i)-\widetilde{\psi}(\widetilde{z}_i))^2=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$. Substituting this result and (A.62) into (A.57), we obtain $\sum_{i\in B}(\widehat{\psi}_i-\bar{\psi}(\widetilde{z}_i))^2=o_{\mathrm{P}}(1+(S^{\mathrm{OPT}})^2)$, i.e., the second part of (A.47). Substituting (A.36) and (A.47) into (A.33), we finally establish (A.34).

Step 5. This step combines (A.34) with (A.35) to prove the theorem, i.e., that the Sharpe ratio of the strategy $\widehat{w}^{\mathrm{OPT}}$ we construct achieves $\left(S^{\mathrm{OPT}}\right)^2$ asymptotically (recall $\widehat{w}^{\mathrm{OPT}}:=\mathbb{M}_\beta\breve{w}$ and $\breve{w}_i=\widehat{\psi}(\widehat{z}_i)/\widehat{\sigma}_i$).

Using condition (d) of Assumption 1 and (B.143) of Lemma B1, we have $\max_{i\le N}|\widehat{\sigma}_i/\sigma_i-1|\lesssim_{\mathrm{P}}c_N$. As a result, we obtain $\|\Sigma^{1/2}\breve{w}-\widehat{\psi}\|\lesssim_{\mathrm{P}}c_N\|\widehat{\psi}\|$, where vector $\widehat{\psi}$ has components $\widehat{\psi}_i:=\widehat{\psi}(\widehat{z}_i)$. Then, it follows from (A.34) and (A.35) that

$$\|\Sigma_u^{1/2}\breve{w}-\psi\|\le\|\Sigma^{1/2}\breve{w}-\widehat{\psi}\|+\|\widehat{\psi}-\psi\|\lesssim_{\mathrm{P}}c_N\|\psi\|+\|\widehat{\psi}-\psi\|=o_{\mathrm{P}}\left(1+S^{\mathrm{OPT}}\right). \tag{A.71}$$

Hence we have

$$|(\breve{w}^\intercal \Sigma_u^{1/2} - \psi^\intercal)\psi| \leq \|\Sigma_u^{1/2}\breve{w} - \psi\|\|\psi\| = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right), \tag{A.72}$$

$$|\breve{w}^\intercal \Sigma_u \breve{w} - \psi^\intercal\psi| \leq \|\Sigma_u^{1/2}\breve{w} - \psi\|^2 + 2\|\Sigma_u^{1/2}\breve{w} - \psi\|\|\psi\| = o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \tag{A.73}$$

Here for both (A.72) and (A.73), the first inequalities come from Cauchy-Schwarz, whereas the last equalities come from (A.71) and (A.35). Further, substituting (A.35) into (A.72) and (A.73), we obtain

$$\breve{w}^\intercal \Sigma_u^{1/2}\psi = \left(S^{\mathrm{OPT}}\right)^2 + o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right), \quad \breve{w}^\intercal \Sigma_u \breve{w} = \left(S^{\mathrm{OPT}}\right)^2 + o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \tag{A.74}$$

On the other hand, we define $\mathcal{E} = \breve{w}^\intercal\alpha - \breve{w}^\intercal\Sigma_u^{1/2}\psi$. Since $\breve{w}$, Σ_u, and ψ are all $\mathcal{G}$-measurable and, according to Proposition 1, $\mathrm{E}(\alpha|\mathcal{G}) = \Sigma_u^{1/2}\psi$, it follows that $\mathrm{E}(\mathcal{E}^2|\mathcal{G}) = \mathrm{Var}(\mathcal{E}|\mathcal{G}) = \breve{w}^\intercal\Sigma_u^{1/2}\mathrm{Var}(\alpha|\mathcal{G})\Sigma_u^{1/2}\breve{w}$. Then, using Chebyshev's inequality, we have, for all positive fixed ϵ,

$$\mathrm{P}(|\mathcal{E}|/\breve{w}^\intercal\Sigma_u\breve{w} \geq \epsilon) \leq \mathrm{E}(\mathcal{E}^2/\breve{w}^\intercal\Sigma_u\breve{w})/\epsilon^2 = \mathrm{E}(\breve{w}^\intercal\Sigma_u^{1/2}\mathrm{Var}(\alpha|\mathcal{G})\Sigma_u^{1/2}\breve{w}/\breve{w}^\intercal\Sigma_u\breve{w})/\epsilon^2. \tag{A.75}$$

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

$$\mathrm{E}(\breve{w}^\intercal\Sigma_u^{1/2}\mathrm{Var}(\alpha|\mathcal{G})\Sigma_u^{1/2}\breve{w}/\breve{w}^\intercal\Sigma_u\breve{w}) \leq \mathrm{E}\left(\max_{i\leq N}\mathrm{Var}(\alpha_i|\mathcal{G})\right) \leq \mathrm{E}(\max_{i\leq N}\alpha_i^2) = o(1), \tag{A.76}$$

where the last step comes from condition (c) of Assumption 1. Combining (A.75) and (A.76), we obtain

$$\breve{w}^\intercal\alpha = \breve{w}^\intercal\Sigma_u^{1/2}\psi + o_{\mathrm{P}}(\breve{w}^\intercal\Sigma_u\breve{w}). \tag{A.77}$$

Next, we note that

$$\mathrm{E}(\psi_i) = \mathrm{E}(s_i) = 0, \quad \mathrm{E}(\psi_i^2) \leq \mathrm{E}(\mathrm{E}(s_i^2|\mathcal{G})) = \mathrm{E}(s_i^2) \leq c_N.$$

Here $\mathrm{E}(s_i) = 0$ comes from condition (c) of Assumption 1, and $\mathrm{E}(s_i^2) \leq c_N$ holds by condition (a) of Assumption 2. Because ψ_i is i.i.d. across i and independent of (β, Σ_u), it follows

$$\mathrm{E}(\|\beta^\intercal\Sigma_u^{-1/2}\psi\|^2|\beta, \Sigma_u) \leq c_N N\|\beta\|_{\mathrm{MAX}}^2\lambda_{\max}(\Sigma_u^{-1}) \lesssim_{\mathrm{P}} c_N N. \tag{A.78}$$

Then we have

$$\|\beta^\intercal\breve{w}\| \leq \|\beta^\intercal\Sigma_u^{-1/2}\psi\| + \|\beta^\intercal(\breve{w} - \Sigma_u^{-1/2}\psi)\| \lesssim_{\mathrm{P}} c_N N^{1/2}\left(1 + S^{\mathrm{OPT}}\right), \tag{A.79}$$

where the last inequality comes from (A.78) and (A.71). As a result, we obtain

$$\breve{w}^\intercal \mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta \breve{w} = \breve{w}^\intercal \Sigma_u \breve{w} + \breve{w}^\intercal \mathbb{P}_\beta \Sigma_u \mathbb{P}_\beta \breve{w} - 2\breve{w}^\intercal \Sigma_u \mathbb{P}_\beta \breve{w} = \breve{w}^\intercal \Sigma_u \breve{w} + o_{\mathrm{P}}\left(1 + \left(S^{\mathrm{OPT}}\right)^2\right). \tag{A.80}$$

For the last equality, we use (A.79), the first part of (A.74), and $\lambda_{\min}(\beta^\intercal \beta) \gtrsim_{\mathrm{P}} N$. Similarly, using $\|\beta^\intercal \alpha\| \lesssim_{\mathrm{P}} N^{1/2} \mathrm{E}(\alpha_i^2)^{1/2} \lesssim_{\mathrm{P}} c_N N^{1/2}$, we write

$$\breve{w}^\intercal \mathbb{M}_\beta \alpha = \breve{w}^\intercal \alpha + \breve{w}^\intercal \mathbb{P}_\beta \alpha = \breve{w}^\intercal \alpha + o_{\mathrm{P}}\left(1 + S^{\mathrm{OPT}}\right). \tag{A.81}$$

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

$$\begin{aligned} \widehat{S}^{\mathrm{OPT}} &= \frac{\left(\widehat{w}^{\mathrm{OPT}}\right)^\intercal \alpha}{\sqrt{\left(\widehat{w}^{\mathrm{OPT}}\right)^\intercal \Sigma_u \widehat{w}^{\mathrm{OPT}}}} = \frac{\breve{w}^\intercal \mathbb{M}_\beta \alpha}{\sqrt{\breve{w}^\intercal \mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta \breve{w}}} \\ &= \frac{\breve{w}^\intercal \alpha}{\sqrt{\breve{w}^\intercal \Sigma_u \breve{w}}} + o_{\mathrm{P}}\left(1 + S^{\mathrm{OPT}}\right) = S^{\mathrm{OPT}} + o_{\mathrm{P}}\left(1 + S^{\mathrm{OPT}}\right). \end{aligned} \tag{A.82}$$

The first two equalities hold by definition. The third one comes from (A.80), (A.81), and the second part of (A.74). The last equality comes from (A.77) and (A.74). Because $\widehat{w}^{\mathrm{OPT}}$ is $\mathcal{G}$-measurable and $\beta^\intercal \widehat{w}^{\mathrm{OPT}} = 0$, the second part of Theorem 1 and Proposition 1 apply. We hence have $\widehat{S}^{\mathrm{OPT}} \leq S(\mathcal{G}) + o_{\mathrm{P}}(1) = S^{\mathrm{OPT}} + o_{\mathrm{P}}(1)$. Because $-\widehat{S}^{\mathrm{OPT}}$ is the Sharpe ratio generated by $-\widehat{w}^{\mathrm{OPT}}$, we also have $-\widehat{S}^{\mathrm{OPT}} \leq S^{\mathrm{OPT}} + o_{\mathrm{P}}(1)$. As a result, when S^{OPT} does vanish, we have $\widehat{S}^{\mathrm{OPT}} = o_{\mathrm{P}}(1)$. Therefore, given (A.82) and using the subsequence argument (see, e.g., Andrews and Cheng (2012)), we have

$$\widehat{S}^{\mathrm{OPT}} = S^{\mathrm{OPT}} + o_{\mathrm{P}}\left(1 + S^{\mathrm{OPT}}\right).$$

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

$$\lim_{N,T\to\infty} \mathrm{P}\left(\left|\widehat{S}^{\mathrm{OPT}} - S^{\mathrm{OPT}}\right| \geq \epsilon S^{\mathrm{OPT}} + \epsilon\right) = 0. \tag{A.83}$$

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

$$\limsup_{N,T\to\infty} \lim_{k\to\infty} \mathrm{P}_k\left(\left|\widehat{S}^{\mathrm{OPT}} - S^{\mathrm{OPT}}\right| \geq \epsilon S^{\mathrm{OPT}} + \epsilon\right) > 0.$$

This contradicts (A.83), and the theorem is proved.

A.3 Proof of Proposition 2

Proof of Proposition 2. By definition we have

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

We start with the analysis of $\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha$. From (B.143) of Lemma B1, it follows

$$\|\widehat{\Sigma}_u - \Sigma_u\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} \sqrt{T^{-1} \log N}. \tag{A.84}$$

As a result, noting $\mathrm{P}(0 \leq \mathbb{M}_\beta \leq \mathbb{I}_N) \to 1$ and $\mathrm{P}(\Sigma_u \sim \mathbb{I}_N) \to 1$ by condition (a) of Assumption 1, and recalling $(S^\star)^2 = \alpha^\intercal \Sigma_u^{-1} \alpha$ we have

$$|\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha - \alpha^\intercal \mathbb{M}_\beta \Sigma_u^{-1} \mathbb{M}_\beta \alpha| \lesssim_{\mathrm{P}} \sqrt{T^{-1} \log N} (S^\star)^2. \tag{A.85}$$

On the other hand, it holds that

$$\begin{aligned}
|\alpha^\intercal \mathbb{M}_\beta \Sigma_u^{-1} \mathbb{M}_\beta \alpha - \alpha^\intercal \Sigma_u^{-1} \alpha| &\leq \alpha^\intercal \mathbb{P}_\beta \Sigma_u^{-1} \mathbb{P}_\beta \alpha + 2\sqrt{(\alpha^\intercal \Sigma_u^{-1} \alpha)(\alpha^\intercal \mathbb{P}_\beta \Sigma_u^{-1} \mathbb{P}_\beta \alpha)} \\
&\lesssim_{\mathrm{P}} \alpha^\intercal \mathbb{P}_\beta \alpha + \sqrt{(\alpha^\intercal \Sigma_u^{-1} \alpha)(\alpha^\intercal \mathbb{P}_\beta \alpha)} \\
&\lesssim_{\mathrm{P}} N^{-1} \|\alpha^\intercal \beta\|^2 + \sqrt{N^{-1} (\alpha^\intercal \Sigma_u^{-1} \alpha) \|\alpha^\intercal \beta\|^2} \\
&\lesssim_{\mathrm{P}} \mathrm{E}(\alpha_i^2) + S^\star \mathrm{E}(\alpha_i^2)^{1/2}.
\end{aligned} \tag{A.86}$$

Here the first inequality comes from Cauchy-Schwarz inequality. The second comes from $\mathrm{P}(\Sigma_u \sim \mathbb{I}_N) \to 1$ and $\mathbb{P}_\beta^2 = \mathbb{P}_\beta$. We obtain the third line by using $\lambda_{\min}(\beta^\intercal \beta) \gtrsim N$. The last line holds because $\mathrm{E}(\|\alpha^\intercal \beta\|^2 | \beta) \lesssim N \|\beta\|_{\mathrm{MAX}}^2 \mathrm{E}(\alpha_i^2) \lesssim_{\mathrm{P}} N \mathrm{E}(\alpha_i^2)$ by condition (a) of Assumption 1. On the other hand, because of the condition $\mathrm{E}(\alpha_i^2 \mathbb{1}_{\{|\alpha_i| \geq c_N\}}) \leq c_N N^{-1}$, we have

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

which, by Markov's inequality, leads to

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

Moreover, it holds that

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

Using Chebyshev's inequality, we obtain

$$\alpha^\intercal \alpha \geq \sum_i \alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}} \gtrsim_{\mathrm{P}} \sum_i \mathrm{E}(\alpha_i^2 \mathbb{1}_{\{|\alpha_i| < c_N\}}) \geq N \mathrm{E}(\alpha_i^2) + o(1).$$

Since $(S^\star)^2 \gtrsim_{\mathrm{P}} \alpha^\intercal \alpha$ due to condition (d) of Assumption 1, we have $N \mathrm{E}(\alpha_i^2) \lesssim_{\mathrm{P}} (S^\star)^2 + 1$. Hence, we can

$$\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha = (S^\star)^2 + o_{\mathrm{P}}\left(\sqrt{T^{-1} \log N} ((S^\star)^2 + 1) \right). \tag{A.87}$$

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

$$\alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \lesssim \alpha^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \alpha \sqrt{\bar{u}^\intercal \bar{u}} \lesssim \alpha^\intercal \alpha \sqrt{\bar{u}^\intercal \bar{u}} = O_{\mathrm{P}}(((S^\star)^2 + 1) T^{-1/2}). \tag{A.88}$$

The first inequality comes from Cauchy-Schwarz. The second inequality holds because $\mathrm{P}(\mathbb{M}_\beta \sim \mathbb{I}_N) \to 1$, $\mathbb{M}_\beta^2 = \mathbb{M}_\beta$, and $\mathrm{P}(\widehat{\Sigma}_u \sim \mathbb{I}_N) \to 1$ due to $\mathrm{P}(\Sigma_u \sim \mathbb{I}_N) \to 1$ and (A.84). The third inequality holds by $\mathrm{P}(\widehat{\Sigma}_u \sim \mathbb{I}_N) \to 1$ as well.

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

$$\begin{aligned} N = \mathrm{tr}(\widehat{\Sigma}_u^{-1} \widehat{\Sigma}_u) &= \sum_{i \leq N} (\widehat{\Sigma}_u^{-1})_{i,i} \left(T^{-1} \sum_{s \in T} (\mathbb{M}_\beta u_s)_i^2 - (\mathbb{M}_\beta \bar{u})_i^2 \right) \\ &= T^{-1} \sum_{s \in T} u_s^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_s - \bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} \\ &= T^{-1} \sum_{s \in T} u_s^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_s + O_{\mathrm{P}}(N/T). \end{aligned} \tag{A.89}$$

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

$$\begin{aligned} \bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - \bar{u}^\intercal \widehat{\Sigma}_u^{-1} \bar{u} &\leq 2|\bar{u}^\intercal \widehat{\Sigma}_u^{-1} \mathbb{P}_\beta \bar{u}| + \bar{u}^\intercal \mathbb{P}_\beta \widehat{\Sigma}_u^{-1} \mathbb{P}_\beta \bar{u} \\ &\lesssim_{\mathrm{P}} \sqrt{\bar{u}^\intercal \bar{u}} \sqrt{\bar{u}^\intercal \mathbb{P}_\beta \bar{u}} + \bar{u}^\intercal \mathbb{P}_\beta \bar{u} \lesssim_{\mathrm{P}} N^{1/2}/T. \end{aligned} \tag{A.90}$$

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

$$\begin{aligned} T^{-2} \sum_{t \in T} (u_t^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta u_t - u_t^\intercal \widehat{\Sigma}_u^{-1} u_t) &= T^{-2} \sum_{t \in T} \left(2 \sqrt{u_t^\intercal u_t} \sqrt{u_t^\intercal \mathbb{P}_\beta \widehat{\Sigma}_u^{-1} \mathbb{P}_\beta u_t} + u_t^\intercal \mathbb{P}_\beta \widehat{\Sigma}_u^{-1} \mathbb{P}_\beta u_t \right) \\ &\lesssim_{\mathrm{P}} T^{-2} \sum_{t \in T} \left(\sqrt{u_t^\intercal u_t} \sqrt{u_t^\intercal \mathbb{P}_\beta u_t} + u_t^\intercal \mathbb{P}_\beta u_t \right) \lesssim_{\mathrm{P}} \frac{N^{1/2}}{T}. \end{aligned} \tag{A.91}$$

From (A.89), (A.90), and (A.91), it directly follows

$$\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1} N = \bar{u}^\intercal \widehat{\Sigma}_u^{-1} \bar{u} - T^{-2} \sum_{s \in T} u_s^\intercal \widehat{\Sigma}_u^{-1} u_s + O_{\mathrm{P}}(N^{1/2}/T + N/T^2). \tag{A.92}$$

On the other hand, we have

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

It then follows from (A.84) and $\mathrm{P}(\Sigma_u \sim \mathbb{I}_N) \to 1$ that

$$\begin{aligned}\bar{u}^\intercal \widehat{\Sigma}_u^{-1} \bar{u} &= -\bar{u}^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)\bar{u} + O_{\mathrm{P}}(T^{-1}(\log N)\bar{u}^\intercal \bar{u}) \\ &= -\bar{u}^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)\bar{u} + O_{\mathrm{P}}(T^{-2}N \log N). \end{aligned} \tag{A.93}$$

Similarly, we have

$$T^{-2}\sum_{t\in T} u_t^\intercal \widehat{\Sigma}_u^{-1} u_t = -T^{-2}\sum_{t\in T} u_t^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)u_t + O_{\mathrm{P}}(T^{-2}N\log N). \tag{A.94}$$

Substituting (A.93) and (A.94) into (A.92), we have

$$\begin{aligned}\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1}\mathbb{M}_\beta \bar{u} - T^{-1}N &= -\bar{u}^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)\bar{u} + T^{-2}\sum_{t\in T} u_t^\intercal \Sigma_u^{-2}(\widehat{\Sigma}_u - 2\Sigma_u)u_t \\ &\quad + O_{\mathrm{P}}(T^{-1}N^{1/2} + T^{-2}N\log N). \end{aligned} \tag{A.95}$$

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

$$\begin{aligned}(\widehat{\Sigma}_u)_{i,i} &= (T^{-1}uu^\intercal)_{i,i} + (\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta)_{i,i} \\ &\quad + (\mathbb{P}_\beta T^{-1}uu^\intercal)_{i,i} + (T^{-1}uu^\intercal \mathbb{P}_\beta)_{i,i} + (\mathbb{P}_\beta T^{-1}uu^\intercal \mathbb{P}_\beta)_{i,i}. \end{aligned} \tag{A.96}$$

From the uniform bound on i.i.d. random variables and $\|\mathbb{P}_\beta\|_{\mathrm{MAX}} \lesssim N^{-1}$ by condition (a) of Assumption 1, it follows

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

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

$$\sum_{i\le N} |\bar{u}_i^2 (\Sigma_u^{-2})_{i,i}(\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta)_{i,i}| \lesssim_{\mathrm{P}} T^{-1}(\log N)\sum_{i\le N}\bar{u}_i^2 \lesssim_{\mathrm{P}} T^{-2}N\log N, \tag{A.97}$$

and

$$T^{-2}\sum_{t\in T}\sum_{i\le N} |u_{i,t}^2 (\Sigma_u^{-2})_{i,i}(\mathbb{M}_\beta \bar{u}\bar{u}^\intercal \mathbb{M}_\beta)_{i,i}| \lesssim_{\mathrm{P}} T^{-3}(\log N)\sum_{t\in T}\sum_{i\le N} u_{i,t}^2 \lesssim_{\mathrm{P}} T^{-2}N\log N. \tag{A.98}$$

Further, we obtain

$$\begin{aligned}&\sum_{i\le N, j\le K} \mathrm{E}(|(T^{-1}uu^\intercal \beta)_{i,j}||\beta, \Sigma_u) \\ \le &\sum_{i\le N, j\le K} \sqrt{\mathrm{E}((T^{-1}uu^\intercal)_{i,i}|\Sigma_u)\mathrm{E}((T^{-1}\beta^\intercal uu^\intercal \beta)_{j,j}|\beta,\Sigma_u)}\end{aligned}$$

$$= \sum_{i\leq N, j\leq K} \sqrt{(\Sigma_u)_{i,i}(\beta^\intercal \Sigma_u \beta)_{j,j}} \leq N^{3/2} K \|\beta\|_{\text{MAX}} \lambda_{\max}(\Sigma_u) \lesssim_{\text{P}} N^{3/2}. \tag{A.99}$$

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

$$\begin{aligned} & \sum_{j\leq K, k\leq K} \text{E}(|(T^{-1}\beta uu^\intercal \beta)_{j,k}||\beta, \Sigma_u) \\ \leq & \sum_{j\leq K, k\leq K} \sqrt{\text{E}((T^{-1}\beta^\intercal uu^\intercal \beta)_{j,j}|\beta, \Sigma_u)\text{E}((T^{-1}\beta^\intercal uu^\intercal \beta)_{j,j}|\beta, \Sigma_u)} \\ = & \sum_{j\leq K, k\leq K} \sqrt{(\beta^\intercal \Sigma_u \beta)_{j,j}(\beta^\intercal \Sigma_u \beta)_{k,k}} \leq KN \|\beta\|_{\text{MAX}} \lambda_{\max}(\Sigma_u) \lesssim_{\text{P}} N. \end{aligned} \tag{A.100}$$

From (A.99) and (A.100), it directly follows

$$\sum_{i\leq N, j\leq K} |(T^{-1}uu^\intercal \beta)_{i,j}| \lesssim_{\text{P}} N^{3/2}, \quad \sum_{j\leq K, k\leq K} |(T^{-1}\beta uu^\intercal \beta)_{j,k}| \lesssim_{\text{P}} N. \tag{A.101}$$

Using (A.101), and noting $\max_{i\leq N} |\bar{u}_i| \lesssim_{\text{P}} \sqrt{T^{-1}\log N}$ from the uniform bound on i.i.d. random variables and $\|(\beta^\intercal\beta)^{-1}\beta\|_{\text{MAX}} \lesssim N^{-1}$ by condition (a) of Assumption 1, we obtain

$$\begin{aligned} \sum_{i\leq N} |\bar{u}_i^2 (T^{-1}uu^\intercal \mathbb{P}_\beta)_{i,i}| \quad \leq & \quad \sum_{i\leq N, j\leq K} |\bar{u}_i^2 (T^{-1}uu^\intercal \beta)_{i,j}| \|(\beta^\intercal\beta)^{-1}\beta\|_{\text{MAX}}. \\ \lesssim_{\text{P}} & \quad N^{-1}T^{-1}\log N \sum_{i\leq N, j\leq K} |(T^{-1}uu^\intercal \beta)_{i,j}| \lesssim_{\text{P}} N^{1/2}T^{-1}\log N, \end{aligned} \tag{A.102}$$

and

$$\begin{aligned} \sum_{i\leq N} |\bar{u}_i^2 (\mathbb{P}_\beta (T^{-1}uu^\intercal)\mathbb{P}_\beta)_{i,i}| \quad \leq & \quad \sum_{i\leq N, j\leq K, k\leq K} \bar{u}_i^2 |(T^{-1}\beta uu^\intercal \beta)_{j,k}| \|(\beta^\intercal\beta)^{-1}\beta\|_{\text{MAX}}^2 \\ \lesssim_{\text{P}} & \quad N^{-1}T^{-1}\log N \sum_{j\leq K, k\leq K} |(T^{-1}\beta uu^\intercal \beta)_{j,k}| \lesssim_{\text{P}} T^{-1}\log N. \end{aligned} \tag{A.103}$$

Symmetric reasoning leads to

$$\frac{1}{T^2}\sum_{s\in T}\sum_{i\leq N} |u_{i,s}^2 (T^{-1}uu^\intercal \mathbb{P}_\beta)_{i,i}| \lesssim_{\text{P}} N^{1/2}T^{-1}, \tag{A.104}$$

$$\frac{1}{T^2}\sum_{s\in T}\sum_{i\leq N} |u_{i,s}^2 (\mathbb{P}_\beta (T^{-1}uu^\intercal)\mathbb{P}_\beta)_{i,i}| \lesssim_{\text{P}} T^{-1}. \tag{A.105}$$

Substituting (A.97), (A.98), (A.102), (A.104), (A.103), and (A.105) into (A.95) and (A.96), we

obtain

$$\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1}N = -T^{-2} \sum_{i:i\leq N} A_i + O_{\mathrm{P}}(T^{-1}N^{1/2}\log N + T^{-2}N\log N). \tag{A.106}$$

Here and only here we use short-hand notation

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

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

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

It hence follows that

$$T^{-2}\sum_{i:i\leq N} A_i \lesssim_{\mathrm{P}} T^{-1}N^{1/2}\mathrm{E}(u_{i,t}^8(\Sigma_u^{-4})_{i,i}) \lesssim T^{-1}N^{1/2}, \tag{A.107}$$

where the last inequality comes from that $\varepsilon_{i,t}$ has finite eighth moment by assumption. Subsituting (A.107) into (A.106), we obtain

$$\bar{u}^\intercal \mathbb{M}_\beta \widehat{\Sigma}_u^{-1} \mathbb{M}_\beta \bar{u} - T^{-1}N = O_{\mathrm{P}}(T^{-1}N^{1/2}\log N + T^{-2}N\log N). \tag{A.108}$$

Combining (A.87), (A.88), and (A.108), and noting $N^{1/2}T \leq c_N$ and $T \lesssim N$ by assumption, we obtain

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

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

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

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

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

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

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

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

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

The proof concludes. ∎

A.4 Proof of Proposition 3

Proof. From Assumption 1, it holds that

$$\alpha^\intercal\alpha = \mu^2\rho N + O_{\mathrm{P}}(\mu^2(\rho N)^{1/2}), \quad \alpha^\intercal\bar{u} = O_{\mathrm{P}}(\mu(\rho N)^{1/2}T^{-1/2}), \tag{A.109}$$

$$\bar{u}^\intercal\bar{u} = T^{-1}N\sigma^2 + O_{\mathrm{P}}(T^{-1}N^{1/2}). \tag{A.110}$$

Here the first result of (A.109) comes from $\mathrm{E}(\alpha^\intercal\alpha) = \mu^2\rho N$ and $\mathrm{Var}(\alpha^\intercal\alpha) \leq \mu^4\rho N$; the last result of (A.109) relies on the independence of $u_{i,t}$ across t and with α; the result of (A.110) comes from the independence of $u_{i,t}$ across (i,t). Since by definition $\widehat{\alpha}_i = \alpha_i + \bar{u}_i$, we obtain, using (A.109) and (A.110), and $\mu \leq c_N$,

$$\alpha^\intercal\widehat{\alpha} = \mu^2\rho N + o_{\mathrm{P}}(\mu(\rho N)^{1/2}), \quad \widehat{\alpha}^\intercal\widehat{\alpha} = T^{-1}N\sigma^2 + \mu^2\rho N + o_{\mathrm{P}}(T^{-1/2}N^{1/2} + \mu(\rho N)^{1/2}). \tag{A.111}$$

Because $\mathrm{E}\left(r_{t+1}^\intercal\widehat{w}^{\mathrm{CSR}}|\mathcal{F}_t\right) = \widehat{\sigma}^{-2}\alpha^\intercal\widehat{\alpha}$ and $\mathrm{Var}\left(r_{t+1}^\intercal\widehat{w}^{\mathrm{CSR}}|\mathcal{F}_t\right) = \widehat{\sigma}^{-4}\sigma^2\widehat{\alpha}^\intercal\widehat{\alpha}$, we prove $\widehat{S}^{\mathrm{CSR}} = S^{\mathrm{CSR}} + o_{\mathrm{P}}(1)$ directly from (A.111). ∎

A.5 Proof of Proposition 4

Proof of Proposition 4. We let $\widehat{z} = -\Phi^{-1}(p_{\widehat{(k)}}/2)$ and $\widehat{z}' = -\Phi^{-1}(p_{\widehat{(k+1)}}/2)$, where we recall Φ is the standard normal cdf. In other words, $\widehat{z}$ and $\widehat{z}'$ are the t-statistics whose p-values, calculated based on standard normal distribution, are $p_{\widehat{(k)}}$ and $p_{\widehat{(k+1)}}$. We also set $c_\lambda = \sqrt{(2-\lambda)\log N}$. We note, for all sequences $a_N \to \infty$,

$$a_N\Phi(-a_N)/\phi(a_N) \to 1. \tag{A.112}$$

As a result, it holds, for all fixed $0 < \lambda \leq 2$,

$$N\Phi(-c_\lambda) \to \infty. \tag{A.113}$$

We further define

$$H_0 = \{i \leq N : \alpha_i = 0\}, \quad H_+ = \{i \leq N : \alpha_i = \mu\}, \quad H_- = \{i \leq N : \alpha_i = -\mu\},$$

$$m_0(a) = \sum_{i \in H_0} \mathbb{1}_{\{|\breve{z}_i| \geq a\}}, \quad m_+(a) = \sum_{i = H_+} \mathbb{1}_{\{|\breve{z}_i| \geq a\}}, \quad m_-(a) = \sum_{i = H_-} \mathbb{1}_{\{|\breve{z}_i| \geq a\}},$$

where $\breve{z}_i = |S|^{1/2}\breve{\alpha}_i/\breve{\sigma}_i$ is the t-statistic of stock i calculated from subsample S. From the definitions of $\widehat{z}$ and $\widehat{z}'$, we obtain

$$\frac{2N\Phi(-\widehat{z})}{m_0(\widehat{z}) + m_+(\widehat{z}) + m_-(\widehat{z})} \leq \tau, \quad \frac{2N\Phi(-\widehat{z}')}{m_0(\widehat{z}') + m_+(\widehat{z}') + m_-(\widehat{z}')} > \tau, \tag{A.114}$$

$$m_0(\widehat{z}') + m_+(\widehat{z}') + m_-(\widehat{z}') = m_0(\widehat{z}) + m_+(\widehat{z}) + m_-(\widehat{z}) + 1. \tag{A.115}$$

Noting $\breve{z}_i$ is i.i.d. across i, we have, for all deterministic positive sequences (a_N, a'_N) satisfying $N\Phi(-a_N) \to \infty$ and $\rho N\Phi(|S|^{1/2}\mu/\sigma - a'_N) \to \infty$,

$$m_0(a_N) = 2N\Phi(-a_N)\left(1 + o_{\mathrm{P}}(1)\right), \tag{A.116}$$

$$m_\pm(a'_N) = \frac{\rho}{2} N(\Phi(|S|^{1/2}\mu/\sigma - a'_N) + \Phi(-|S|^{1/2}\mu/\sigma - a'_N))\left(1 + o_{\mathrm{P}}(1)\right). \tag{A.117}$$

Now suppose $z^* \geq c_\lambda$ for all fixed $0 < \lambda \leq 2$. Then by direct calculations it follows from (20) that, for all fixed $0 < \lambda \leq 2$,

$$\frac{\rho\Phi(|S|^{1/2}\mu/\sigma - c_\lambda)}{\Phi(-c_\lambda)} \to 0. \tag{A.118}$$

Using (A.113), (A.118), (A.116), (A.117) and the monotonicity of Φ, we obtain, for all fixed $0 < \lambda \leq 2$,

$$m_0(c_\lambda) = 2N\Phi(-c_\lambda)\left(1 + o_{\mathrm{P}}(1)\right), \quad m_\pm(c_\lambda) = o_{\mathrm{P}}(N\Phi(-c_\lambda)).$$

Hence, from (A.114) and (A.115), we conclude, for all fixed $0 < \lambda \leq 2$,

$$\mathrm{P}(\widehat{z} \geq c_\lambda) \to 1. \tag{A.119}$$

Combining (A.119), (A.118), and (A.117), we have, for all fixed $\lambda > 0$,

$$\mathrm{P}(m_0(\widehat{z}) + m_+(\widehat{z}) + m_-(\widehat{z}) \leq CN^\lambda) \to 1.$$

Moreover, we observe

$$\left|\widehat{S}^{\mathrm{BH}}\right| \leq \frac{\mu}{\sigma} \frac{\sum_{i \leq N} \left|\widehat{w}_i^{\mathrm{BH}}\right|}{\sqrt{\sum_{i \leq N} \left(\widehat{w}_i^{\mathrm{BH}}\right)^2}} \leq \frac{\mu}{\sigma} \sqrt{\sum_{i \leq N} \mathbb{1}_{\{\widehat{w}_i^{\mathrm{BH}} \neq 0\}}} \leq \frac{\mu}{\sigma} \sqrt{m_0(\widehat{z}) + m_+(\widehat{z}) + m_-(\widehat{z})}.$$

Then from that $\mu \leq CN^\lambda$ for some fixed $\lambda < 0$, it follows $\widehat{S}^{\mathrm{BH}} = o_{\mathrm{P}}(1)$. On the other hand, from (A.118) it follows $\rho N\Phi(T^{1/2}\mu/\varsigma - z^*) \leq CN^\lambda$ for all fixed $0 < \lambda \leq 2$. Hence $S^{\mathrm{BH}} = o(1)$ and we prove $\widehat{S}^{\mathrm{BH}} - S^{\mathrm{BH}} = o_{\mathrm{P}}\left(S^{\mathrm{BH}} + 1\right)$ under $z^* \geq c_\lambda$.

Next, we suppose $z^* \leq c_\lambda$ for some fixed $0 < \lambda \leq 2$. Then, using (20) and (A.112), it holds that

for some fixed $0 < \lambda \leq 2$,

$$\frac{\rho\Phi(|S|^{1/2}\mu/\sigma - c_\lambda)}{\Phi(-c_\lambda)} \to \infty. \tag{A.120}$$

We combine (A.113), (A.120), (A.116), and (A.117) to conclude that, for some fixed $0 < \lambda \leq 2$ and in probability,

$$\frac{m_\pm(c_\lambda)}{m_0(c_\lambda)} \to \infty.$$

It then follows from (A.114) and (A.115) that $\mathrm{P}(\widehat{z} \leq c_\lambda) \to 1$ for some fixed $0 < \lambda \leq 2$. Given (A.113) and (A.120), we have, in probability

$$N\Phi(-\widehat{z}) \to \infty, \quad \rho N\Phi(|S|^{1/2}\mu/\sigma - \widehat{z}) \to \infty. \tag{A.121}$$

Applying equation (13) of Liu and Shao (2014) to (A.121), we obtain

$$m_0(\widehat{z}) = 2N\Phi(-\widehat{z})\,(1 + o_{\mathrm{P}}(1))\,, \quad m_\pm(\widehat{z}) = \frac{\rho}{2}N(\Phi(|S|^{1/2}\mu/\sigma - \widehat{z}) + \Phi(-|S|^{1/2}\mu/\sigma - \widehat{z}))\,(1 + o_{\mathrm{P}}(1))\,. \tag{A.122}$$

Since $\widehat{z}' \geq \widehat{z}$, (A.122) would still hold if all $\widehat{z}$ are replaced by $\widehat{z}'$. Hence, substituting (A.122) back into (A.114) and (A.115), and noting $\Phi(-|S|^{1/2}\mu/\sigma - \widehat{z}) \leq \Phi(-\widehat{z})$, we have

$$\frac{2(1-\tau)\Phi(-\widehat{z})}{\tau\rho\Phi(|S|^{1/2}\mu/\sigma - \widehat{z})} = 1 + o_{\mathrm{P}}(1). \tag{A.123}$$

Next, using (20) and (A.112), we note that $z^* \leq c_\lambda$ for some fixed $0 < \lambda \leq 2$ leads to that $|S|^{1/2}\mu \geq c_{\lambda'}$ for some fixed $\lambda' < 2$. As a result, using (A.112), and comparing (20) and (A.123), we have

$$\Phi(|S|^{1/2}\mu/\sigma - \widehat{z}) = \Phi(|S|^{1/2}\mu/\sigma - z^*)\,(1 + o_{\mathrm{P}}(1))\,. \tag{A.124}$$

In light of (A.122) - (A.124), we have

$$m_+(\widehat{z}) + m_-(\widehat{z}) = \rho N\Phi(|S|^{1/2}\mu/\sigma - z^*)\,(1 + o_{\mathrm{P}}(1))\,. \tag{A.125}$$

Moreover, from (A.112) and (A.120), and that $\rho \lesssim N^d$ with $d < 0$, it follows $\mu \gtrsim \sqrt{(\log N)/T}$.

Now we analyze the Sharpe ratio, we have

$$\begin{aligned}
\widehat{\sigma}^2\alpha^\intercal\widehat{w}^{\mathrm{BH}} &= \sum_{i=H_+}\mu\breve{\alpha}_i'\mathbb{1}_{\{|\widehat{z}_i|\geq\widehat{z}\}} - \sum_{i=H_-}\mu\breve{\alpha}_i'\mathbb{1}_{\{|\widehat{z}_i|\geq\widehat{z}\}} \\
&= \mu^2(m_+(\widehat{z}) + m_-(\widehat{z})) + \sum_{i=H_+}\mu\bar{u}_i'\mathbb{1}_{\{|\widehat{z}_i|\geq\widehat{z}\}} - \sum_{i=H_-}\mu\bar{u}_i'\mathbb{1}_{\{|\widehat{z}_i|\geq\widehat{z}\}} \\
&= \mu^2(m_+(\widehat{z}) + m_-(\widehat{z})) + O_{\mathrm{P}}\left(\mu\sqrt{m_+(\widehat{z}) + m_-(\widehat{z})}|S'|^{-1/2}\right).
\end{aligned}$$

Here $\bar{u}_i' = |S'|^{-1}\sum_{s\in S'}u_{i,s}$ and in the last step we utilize the independence of of $\varepsilon_{i,s}$ across split

samples. Similarly, we have

$$\begin{aligned}\widehat{\sigma}^4 \left\|\widehat{w}^{\mathrm{BH}}\right\|^2 &= \sum_{i=H_+} (\breve{\alpha}_i')^2 \mathbb{1}_{\{|\widehat{z}_i|\geq \widehat{z}\}} + \sum_{i=H_-} (\breve{\alpha}_i')^2 \mathbb{1}_{\{|\widehat{z}_i|\geq \widehat{z}\}} + \sum_{i=H_0} (\breve{\alpha}_i')^2 \mathbb{1}_{\{|\widehat{z}_i|\geq \widehat{z}\}} \\ &= \mu^2(m_+(\widehat{z}) + m_-(\widehat{z})) + 2\mu \sum_{i\in H_+} \bar{u}_i' \mathbb{1}_{\{|\widehat{z}_i|\geq \widehat{z}\}} - 2\mu \sum_{i\in H_-} \bar{u}_i' \mathbb{1}_{\{|\widehat{z}_i|\geq \widehat{z}\}} + \sum_{i\leq N} (\bar{u}_i')^2 \mathbb{1}_{\{|\widehat{z}_i|\geq \widehat{z}\}} \\ &= \mu^2(m_+(\widehat{z}) + m_-(\widehat{z})) + O_{\mathrm{P}}\left(\mu\sqrt{m_+(\widehat{z}) + m_-(\widehat{z})}|S'|^{-1/2}\right) \\ &\quad + O_{\mathrm{P}}((m_+(\widehat{z}) + m_-(\widehat{z}) + m_0(\widehat{z}))|S'|^{-1}).\end{aligned}$$

Given (A.122) and (A.121), we have $m_+(\widehat{z})+m_-(\widehat{z}) \to \infty$ in probability, $m_0(\widehat{z})/(m_+(\widehat{z})+m_-(\widehat{z})) \lesssim_{\mathrm{P}} 1$. Thus, noting $|S'|^{-1/2} \leq c_N\mu$ as $\mu \gtrsim \sqrt{(\log N)/T}$, we obtain

$$\begin{aligned}\alpha^\intercal \widehat{w}^{\mathrm{BH}} &= (1 + o_{\mathrm{P}}(1))\, \widehat{\sigma}^{-2}\mu^2(m_+(\widehat{z}) + m_-(\widehat{z})), \\ \left\|\widehat{w}^{\mathrm{BH}}\right\|^2 &= (1 + o_{\mathrm{P}}(1))\, \widehat{\sigma}^{-4}\mu^2(m_+(\widehat{z}) + m_-(\widehat{z})).\end{aligned}$$

We prove the proposition, given (A.125). ∎

A.6 Proof of Proposition 5

Proof of Proposition 5. We start by defining

$$H_0 = \{i \leq N : \alpha_i = 0\}, \quad H_+ = \{i \leq N : \alpha_i = \mu\}, \quad H_- = \{i \leq N : \alpha_i = -\mu\}, \quad \widetilde{w}^1 = \widetilde{w}^1\widehat{\sigma}^2.$$

Then from the definition of $\widetilde{w}^1$, it follows

$$\mathrm{E}(r_{t+1}^\intercal \widetilde{w}^1|\mathcal{F}_t) = \alpha^\intercal\widetilde{w}^1 = \mu \sum_{i\in H_+} \mathrm{sgn}(\widehat{\alpha}_i)(|\widehat{\alpha}_i| - \lambda)_+ - \mu \sum_{i\in H_-} \mathrm{sgn}(\widehat{\alpha}_i)(|\widehat{\alpha}_i| - \lambda)_+, \tag{A.126}$$

$$\mathrm{Var}(r_{t+1}^\intercal \widetilde{w}^1|\mathcal{F}_t) = \sigma^2\|\widetilde{w}^1\|^2 = \sigma^2 \sum_{i\leq N}((|\widehat{\alpha}_i| - \lambda)_+)^2. \tag{A.127}$$

Because, conditional on α_i, the distribution of $\widehat{\alpha}_i$ is $\mathcal{N}(\alpha_i, \sigma^2/T)$, the statistical moments of $\widetilde{w}^1$ satisfy, for $j \in \{0, 1, 2, 4\}$,

$$\begin{aligned}\mathrm{E}((\widetilde{w}_i^1)^j|\alpha_i = \mu) &= T^{1/2}\sigma^{-1} \int_{-\infty}^{\infty} (\mathrm{sgn}(x)(|x| - \lambda)_+)^j\, \phi(T^{1/2}\sigma^{-1}(x - \mu))dx \\ &= (T^{-1/2}\sigma)^j \int_{-\infty}^{\infty} (\mathrm{sgn}(x)(|x| - \lambda^*)_+)^j\, \phi(x - \mu^*)dx.\end{aligned} \tag{A.128}$$

where we use the short-hand notation $\lambda^* = T^{1/2}\sigma^{-1}\lambda$ and $\mu^* = T^{1/2}\sigma^{-1}\mu$, and here and below we set $z^0 = \mathbb{1}_{\{z\neq 0\}}$ by convention. Moreover, by direct calculations, we have, for $j \in \{0, 2, 4\}$,

$$\int_{-\infty}^{\infty} \mathrm{sgn}(x)(|x| - \lambda^*)_+\phi(x - \mu^*)dx = \int_{\lambda^*}^{\infty} (x - \lambda^*)(\phi(x - \mu^*) - \phi(x + \mu^*))dx$$

$$\begin{aligned} &\gtrsim (1 \wedge \mu^*) \int_{\lambda^*}^{\infty} (x - \lambda^*)\phi(x - \mu^*)dx \\ &\sim \frac{(1 \wedge \mu^*)(1 \vee (\mu^* - \lambda^*))}{1 \vee (\lambda^* - \mu^*)} \Phi(\mu^* - \lambda^*), \end{aligned} \tag{A.129}$$

$$\begin{aligned} \int_{-\infty}^{\infty} ((|x| - \lambda^*)_+)^j \phi(x - \mu^*)dx &= \int_{\lambda^*}^{\infty} (x - \lambda^*)^j (\phi(x - \mu^*) + \phi(x + \mu^*))dx \\ &\sim \int_{\lambda^*}^{\infty} (x - \lambda^*)^j \phi(x - \mu^*)dx \\ &\sim \frac{(1 \vee (\mu^* - \lambda^*))^j}{(1 \vee (\lambda^* - \mu^*))^j} \Phi(\mu^* - \lambda^*). \end{aligned} \tag{A.130}$$

By symmetry, we have, for all integer $j \geq 0$,

$$\mathrm{E}((\widetilde{w}_i^1)^j | \alpha_i = \mu) = (-1)^j \mathrm{E}((\widetilde{w}_i^1)^j | \alpha_i = -\mu). \tag{A.131}$$

Similarly, we have, for $j \in \{2, 4\}$,

$$\mathrm{E}((\widetilde{w}_i^1)^j | \alpha_i = 0) = (T^{-1/2}\sigma)^j \int_{-\infty}^{\infty} ((|x| - \lambda^*)_+)^j \phi(x)dx, \tag{A.132}$$

$$\int_{-\infty}^{\infty} ((|x| - \lambda^*)_+)^j \phi(x)dx = 2\int_{\lambda^*}^{\infty} (x - \lambda^*)^j \phi(x)dx \sim \frac{1}{(1 \vee \lambda^*)^j} \Phi(-\lambda^*). \tag{A.133}$$

Using (A.128), we observe that S^{LASSO} defined in the statement of the proposition satisfies

$$S^{\mathrm{LASSO}} = \rho\mu\sigma^{-1}N^{1/2} \frac{\mathrm{E}(\widetilde{w}_i^1 | \alpha_i = \mu)}{\sqrt{(1 - \rho)\mathrm{E}((\widetilde{w}_i^1)^2 | \alpha_i = 0) + \rho\mathrm{E}((\widetilde{w}_i^1)^2 | \alpha_i = \mu)}}. \tag{A.134}$$

We now prove $\widehat{S}^1 - S^{\mathrm{LASSO}} = o_{\mathrm{P}}(1)$.

We first consider the case where $\rho N\mu^2\Phi(\mu^* - \lambda^*) \leq c_N$. Then, using (A.128) and (A.130) (setting $j = 0$), and (A.131), we obtain

$$\mathrm{E}\left(\sum_{i \in H_+ \cup H_-} \mathbb{1}_{\{(|\widehat{\alpha}_i| - \lambda)_+ \neq 0\}} \right) = \rho N \mathrm{E}((\widetilde{w}_i^1)^0 | \alpha_i = \mu) \sim \rho N \Phi(\mu^* - \lambda^*) \leq c_N \mu^{-2}. \tag{A.135}$$

As a result, we have

$$\begin{aligned} |\widehat{S}^1| = \frac{|\mathrm{E}(r_{t+1}^{\intercal} \widetilde{w}^1 | \mathcal{F}_t)|}{\mathrm{Var}(r_{t+1}^{\intercal} \widetilde{w}^1 | \mathcal{F}_t)^{1/2}} &\leq \frac{\mu}{\sigma} \frac{\sum_{i \in H_+} |(|\widehat{\alpha}_i| - \lambda)_+|}{\left(\sum_{i \in H_+} |(|\widehat{\alpha}_i| - \lambda)_+|^2 \right)^{1/2}} + \frac{\mu}{\sigma} \frac{\sum_{i \in H_-} |(|\widehat{\alpha}_i| - \lambda)_+|}{\left(\sum_{i \in H_-} |(|\widehat{\alpha}_i| - \lambda)_+|^2 \right)^{1/2}} \\ &\leq \frac{\mu}{\sigma} \sqrt{\sum_{i \in H_+ \cup H_-} \mathbb{1}_{\{(|\widehat{\alpha}_i| - \lambda)_+ \neq 0\}}} \lesssim_{\mathrm{P}} c_N. \end{aligned} \tag{A.136}$$

Here the second inequality holds by Cauchy-Schwarz and the last holds by Markov's inequality and

(A.135). On the other hand, it follows from (A.134) that

$$\begin{aligned} \left|S^{\text{LASSO}}\right| &\leq \sqrt{\rho N}\mu\sigma^{-1}\frac{\left|\mathrm{E}(\widetilde{w}_i^1|\alpha_i=\mu)\right|}{\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=\mu)^{1/2}} = \sqrt{\rho N}\mu\sigma^{-1}\frac{\left|\mathrm{E}(\widetilde{w}_i^1(\widetilde{w}_i^1)^0|\alpha_i=\mu)\right|}{\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=\mu)^{1/2}} \\ &\leq \sqrt{\rho N}\mu\sigma^{-1}\mathrm{E}((\widetilde{w}_i^1)^0|\alpha_i=\mu)^{1/2} \leq c_N, \end{aligned} \tag{A.137}$$

where the second inequality comes from Cauchy-Schwarz and the last holds by (A.135). Combining (A.136) and (A.137), we obtain $\widehat{S}^1 - S^{\text{LASSO}} = o_{\mathrm{P}}(1)$, under $\rho N\mu^2\Phi(\mu^* - \lambda^*) \leq c_N$.

Next, we consider the case where $\rho N\mu^2\Phi(\mu^* - \lambda^*) \gtrsim 1$. From (A.128), (A.129), and (A.130), it follows that, for $j \in \{2, 4\}$,

$$\mathrm{E}(\widetilde{w}_i^1|\alpha_i=\mu) \gtrsim T^{-1/2}\sigma\frac{(1\wedge\mu^*)(1\vee(\mu^*-\lambda^*))}{1\vee(\lambda^*-\mu^*)}\Phi(\mu^*-\lambda^*), \tag{A.138}$$

$$\mathrm{E}((\widetilde{w}_i^1)^j|\alpha_i=\mu) \sim (T^{-1/2}\sigma)^j\frac{(1\vee(\mu^*-\lambda^*))^j}{(1\vee(\lambda^*-\mu^*))^j}\Phi(\mu^*-\lambda^*), \tag{A.139}$$

$$\mathrm{E}((\widetilde{w}_i^1)^j|\alpha_i=0) \sim (T^{-1/2}\sigma)^j\frac{1}{(1\vee\lambda^*)^j}\Phi(-\lambda^*). \tag{A.140}$$

Using (A.138) - (A.140), $\mu \leq c_N\mu^*$ (by $T \to \infty$), and $\mu \leq c_N$ (by assumption), we have

$$\frac{\rho N\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=\mu)}{(\rho N)^2\mathrm{E}(\widetilde{w}_i^1|\alpha_i=\mu)^2} \lesssim \frac{1}{\rho N(1\wedge\mu^*)^2\Phi(\mu^*-\lambda^*)} \lesssim c_N\frac{1}{\rho N\mu^2\Phi(\mu^*-\lambda^*)} \leq c_N, \tag{A.141}$$

$$\widetilde{w}\frac{\rho N\mathrm{E}((\widetilde{w}_i^1)^4|\alpha_i=\mu)+N\mathrm{E}((\widetilde{w}_i^1)^4|\alpha_i=0)}{(\rho N)^2\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=\mu)^2+N^2\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=0)^2} \leq c_N. \tag{A.142}$$

On the other hand, because $(\alpha_i, \widetilde{w}_i^1)$ is i.i.d. across i, we obtain from (A.126) and (A.127) that

$$\begin{aligned} \mathrm{E}(\alpha^\intercal\widetilde{w}^1) &= \rho N\mu\mathrm{E}(\widetilde{w}_i^1|\alpha_i=\mu), \\ \mathrm{Var}(\alpha^\intercal\widetilde{w}^1) &\lesssim \rho N\mu^2\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=\mu), \\ \mathrm{E}(\|\widetilde{w}^1\|^2) &= \rho N\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=\mu)+(1-\rho)N\mathrm{E}((\widetilde{w}_i^1)^2|\alpha_i=0), \\ \mathrm{Var}(\|\widetilde{w}^1\|^2) &\lesssim \rho N\mathrm{E}((\widetilde{w}_i^1)^4|\alpha_i=\mu)+(1-\rho)N\mathrm{E}((\widetilde{w}_i^1)^4|\alpha_i=0). \end{aligned}$$

Combining these four results with (A.141) and (A.142), and using Chebyshev's inequality, we obtain

$$\frac{\alpha^\intercal\widetilde{w}^1}{\mathrm{E}(\alpha^\intercal\widetilde{w}^1)} = 1 + o_{\mathrm{P}}(1), \quad \frac{\|\widetilde{w}^1\|^2}{\mathrm{E}(\|\widetilde{w}^1\|^2)} = 1 + o_{\mathrm{P}}(1).$$

Given (A.126), (A.127), and (A.134), and noting that $\widetilde{w}^1$ and $\widehat{w}^1$ are collinear, we conclude that $\widehat{S}^1 - S^{\text{LASSO}} = o_{\mathrm{P}}(1)$, under $\rho N\mu^2\Phi(\mu^* - \lambda^*) \gtrsim 1$. The proof ends. ∎

Appendix B Proofs of Technical Lemmas

Lemma B1. *We define* $\bar{u}_i = T^{-1}\sum_{s\in\mathcal{T}} u_{i,s}$. *Suppose Assumptions 1 and 2 holds. Also suppose* $T \lesssim N^d$ *with fixed* $d < 1$. *Then it holds that, as* $N, T \to \infty$,

$$\max_{1\le i\le N} |\widehat{\sigma}_i^2 - \sigma_i^2| = O_{\mathrm{P}}\left(\sqrt{(\log N)/T}\right), \tag{B.143}$$

$$\max_{1\le i\le N} |(\mathbb{P}_\beta \bar{u})_i| = O_{\mathrm{P}}\left(1/\sqrt{TN}\right), \tag{B.144}$$

$$\max_{1\le i\le N} |(\mathbb{P}_\beta \alpha)_i| = O_{\mathrm{P}}\left(N^{-1/2}\mathrm{E}(s_i^2)^{1/2}\right). \tag{B.145}$$

Proof. We start with (B.143). First of all, we write

$$\begin{aligned}\max_{1\le i\le N} |\widehat{\sigma}_i^2 - \sigma_i^2| &\le \|T^{-1}\mathbb{M}_\beta uu^\intercal \mathbb{M}_\beta - \Sigma_u\|_{\mathrm{MAX}} \\ &\le \|\mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta - \Sigma_u\|_{\mathrm{MAX}} + \|\mathbb{M}_\beta(T^{-1}uu^\intercal - \Sigma_u)\mathbb{M}_\beta\|_{\mathrm{MAX}}. \end{aligned}\tag{B.146}$$

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

$$\begin{aligned}\|\mathbb{M}_\beta \Sigma_u \mathbb{M}_\beta - \Sigma_u\|_{\mathrm{MAX}} &\le \|\mathbb{P}_\beta \Sigma_u \mathbb{P}_\beta\|_{\mathrm{MAX}} + 2\|\mathbb{P}_\beta \Sigma_u\|_{\mathrm{MAX}} \\ &\le (N\|\mathbb{P}_\beta\|_{\mathrm{MAX}} + 2)\|\mathbb{P}_\beta\|_{\mathrm{MAX}}\|\Sigma_u\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} N^{-1}. \end{aligned}\tag{B.147}$$

The last inequality comes from $\|\mathbb{P}_\beta\|_{\mathrm{MAX}} \le C\|\beta\|_{\mathrm{MAX}}^2\|(\beta^\intercal\beta)^{-1}\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} N^{-1}$, which is true because of condition (a) of Assumption 1. On the other hand, we have

$$\|\mathbb{M}_\beta(T^{-1}uu^\intercal - \Sigma_u)\mathbb{M}_\beta\|_{\mathrm{MAX}} \le \|T^{-1}uu^\intercal - \Sigma_u\|_{\mathrm{MAX}}(1 + 2N\|\mathbb{P}_\beta\|_{\mathrm{MAX}} + N^2\|\mathbb{P}_\beta\|_{\mathrm{MAX}}) \lesssim_{\mathrm{P}} \sqrt{(\log N)/T}, \tag{B.148}$$

where the last inequality comes from the uniform bound on i.i.d. normal variables and that $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$ by condition (d) of Assumption 1. Substituting (B.147) and (B.148) into (B.146), and noting $N^{-1} \le C\sqrt{(\log N)/T}$ by assumption, we obtain (B.143).

We obtain (B.144) by writing

$$\max_{1\le i\le N} |(\mathbb{P}_\beta \bar{u})_k| \le C\|\beta\|_{\mathrm{MAX}}\|(\beta^\intercal\beta)^{-1}\|_{\mathrm{MAX}} \max_{1\le k\le K} |(\beta^\intercal \bar{u})_k| \lesssim_{\mathrm{P}} \max_{1\le k\le K} N^{-1}|(\beta^\intercal \bar{u})_k| \lesssim_{\mathrm{P}} 1/\sqrt{TN}.$$

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

Finally, we write

$$\max_{1\le i\le N} |(\mathbb{P}_\beta \alpha)_i| \le C\|\beta\|_{\mathrm{MAX}}\|(\beta^\intercal\beta)^{-1}\|_{\mathrm{MAX}} \max_{1\le k\le K} |(\beta^\intercal \alpha)_k| \le CN^{-1} \max_{1\le k\le K} |(\beta^\intercal \alpha)_k|. \tag{B.149}$$

On the other hand, from condition (c) and (e) of Assumption 1 and condition (a) of Assumption 2,

we have $\mathrm{E}(\alpha_i\alpha_j|\beta,\Sigma_u) = \delta_{i,j}\sigma_i^2\mathrm{E}(s_i^2)$. Therefore, as K is fixed and $\lambda_{\max}(\Sigma_u) \lesssim_{\mathrm{P}} 1$, we have

$$\max_{1\le k\le K} |(\beta^\intercal\alpha)_k| \lesssim_{\mathrm{P}} c_N N^{1/2}\|\beta\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} N^{1/2}\mathrm{E}(s_i^2)^{1/2}. \tag{B.150}$$

Substituting (B.150) into (B.149), we obtain (B.145). ∎

Lemma B2. *Suppose Assumptions 1 and 2 hold. Also assume $N^d \lesssim T \lesssim N^{d'}$ with fixed $d > 1/2$ and $d < 1$. Then it holds that, as $N, T \to \infty$,*

$$\max_{1\le i\le N} |\widehat{z}_i - \widetilde{z}_i| \le c_N\widetilde{k}_N, \qquad \max_{i\in B} |\widehat{z}_i - \widetilde{z}_i| \lesssim_{\mathrm{P}} \chi_N, \tag{B.151}$$

where $\widetilde{z}_i := T^{1/2}(s_i + \bar{\varepsilon}_i)$, $\chi_N := \sqrt{T/N}(k_N^5 + \mathrm{E}(s_j^2)^{1/2})$, and set B is $B := \{i \in N : |\widetilde{z}_i| \le \widetilde{k}_N\}$, with $\widetilde{k}_N := k_N^{-2}$.

Proof. By definition we have

$$\widehat{z}_i - \widetilde{z}_i = -T^{1/2}\frac{(\mathbb{P}_\beta\alpha)_i}{\widehat{\sigma}_i} - T^{1/2}\frac{(\mathbb{P}_\beta\bar{u})_i}{\widehat{\sigma}_i} + \left(\frac{\sigma_i}{\widehat{\sigma}_i} - 1\right)\widetilde{z}_i. \tag{B.152}$$

Since $T \gtrsim N^d$ with $d > 1/2$ by assumption, (B.143) of Lemma B1 leads to $\max_{1\le i\le N}|\widehat{\sigma}_i^2 - \sigma_i^2| = o_{\mathrm{P}}(1)$. Then, noting $\min_i \sigma_i \gtrsim_{\mathrm{P}} 1$ by condition (d) of Assumption 1, we obtain $\min_i \widehat{\sigma}_i \gtrsim_{\mathrm{P}} 1$. Applying (B.143) again, we have $\max_i \left|\frac{\sigma_i}{\widehat{\sigma}_i} - 1\right| \lesssim_{\mathrm{P}} \sqrt{(\log N)/T}$. Using these two results, and substituting (B.144) and (B.145) of Lemma B1 into (B.152), we obtain

$$\max_{1\le i\le N} |\widehat{z}_i - \widetilde{z}_i| \lesssim_{\mathrm{P}} \chi_N + \sqrt{(\log N)/T}\max_{1\le i\le N}|\widetilde{z}_i|, \qquad \max_{i\in B} |\widehat{z}_i - \widetilde{z}_i| \lesssim_{\mathrm{P}} \chi_N + \sqrt{(\log N)/T}\max_{i\in B}|\widetilde{z}_i|. \tag{B.153}$$

The definition of set B leads to $\max_{i\in B}|\widetilde{z}_i| \le \widetilde{k}_N$. Then, noting $T \gtrsim N^d$ with $d > 1/2$ by assumption, we have $\sqrt{(\log N)/T}\max_{i\in B}|\widetilde{z}_i| \lesssim \sqrt{T/N}k_N^5 \le \chi_N$. Given the second part of (B.153), we obtain the second part of (B.151).

On the other hand, since $\mathrm{P}(|s_i| \ge 1) \le \mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|\ge 1\}}) \le c_N N^{-1}$ by condition (a) of Assumption 2, we have $\mathrm{P}(\max_i |s_i| \ge 1) \le c_N$. Combining this result with $\max_i |\bar{\varepsilon}_i| \lesssim_{\mathrm{P}} \sqrt{(\log N)/T}$ by the uniform bound on i.i.d. normal variables, we obtain $\max_{1\le i\le N}|\widetilde{z}_i| \lesssim_{\mathrm{P}} c_N T^{1/2}$ (again noting $T \gtrsim N^d$ with $d > 1/2$ by assumption). Then we have $\sqrt{(\log N)/T}\max_{1\le i\le N}|\widetilde{z}_i| \le c_N\sqrt{\log N}$. Also, we have $\chi_N \le c_N$ since $T = o(N)$ by assumption and $\mathrm{E}(s_j^2) \lesssim 1 + \mathrm{E}(s_i^2\mathbb{1}_{\{|s_i|\ge 1\}}) \lesssim 1$ by condition (a) of Assumption 2. Substituting the two bounds into the first part of (B.153), we achieve the first part of (B.151). ∎

Lemma B3. *Suppose Assumptions 1 and 2 hold. $N^d \lesssim T \lesssim N^{d'}$ with fixed $d > 1/2$ and $d < 1$. Then it holds that, as $N, T \to \infty$,*

$$\mathrm{P}(p(\widetilde{z}_i) \ge N^{-3/2}, \forall i \le N) \ge 1 - c_N. \tag{B.154}$$

Proof. Note that when $|x| < 1$, we can find $C > 1$ such that $a \geq C\sqrt{T}$ implies $|a - \sqrt{T}x| \geq (C-1)\sqrt{T}$. Therefore, for $|x| < 1$, we have

$$\int_{|a|\geq CT^{1/2}} \phi(T^{1/2}x - a)da \leq \int_{|a|\geq (C-1)T^{1/2}} \exp(-a^2/2)da \lesssim T^{-1/2}\exp(-T) \leq c_N N^{-1}. \quad \text{(B.155)}$$

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

$$\begin{aligned} \int_{|a|\geq CT^{1/2}} p(a)da &\leq \int_{|x|\geq 1} p_s(x)dx + \int_{|x|<1}\int_{|a|\geq CT^{1/2}} \phi(T^{1/2}x - a)da p_s(x)dx \\ &\leq \int_{|x|\geq 1} p_s(x)dx + \sup_{x:|x|<1}\int_{|a|\geq CT^{1/2}} \phi(T^{1/2}x - a)da \leq c_N N^{-1}. \quad \text{(B.156)} \end{aligned}$$

Here the last inequality comes from (B.155) and $\int_{|x|\geq 1} p_s(x)dx \leq \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|\geq 1\}}) \leq \mathrm{E}(s_i^2 \mathbb{1}_{\{|s_i|\geq c_N\}}) \leq c_N N^{-1}$ by condition (a) of Assumption 2. It follows from (B.156) that

$$\begin{aligned} \mathrm{P}(p(\widetilde{z}_i) < N^{-3/2}) &= \int \mathbb{1}_{\{p(a)<N^{-3/2}\}} p(a)da \\ &\leq c_N N^{-1} + \int_{|a|<CT^{1/2}} \mathbb{1}_{\{p(a)<N^{-3/2}\}} p(a)da \leq c_N N^{-1}. \quad \text{(B.157)} \end{aligned}$$

The last inequality also uses $T = o(N)$ by assumption. (B.157) proves the lemma by Bonferroni inequalities. ∎

Lemma B4. *It holds that, as $N \to \infty$, for $j \in \{0, 1\}$ and for all $(a, \bar{a})$ satisfying $|\bar{a} - a| \lesssim k_N$,*

$$|a^j\phi(a) - \bar{a}^j\phi(\bar{a})| \lesssim c_N k_N^{-j-1}|\bar{a} - a|\phi(a) + c_N N^{-2}. \quad \text{(B.158)}$$

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 diverging sequence b_N, 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 $j \in \{0, 1\}$,

$$|a^j\phi(a) - \bar{a}^j\phi(\bar{a})| \lesssim b_N^{j+1}|\bar{a} - a|\phi(a). \quad \text{(B.159)}$$

Moreover, $\sup_{a:|a|\geq b_N} |a^j\phi(a)| \leq c_N N^{-2}$ for $j \in \{0, 1\}$ and for all b_N that satisfies $b_N \gtrsim (\log N)^d$ with $d > 1/2$. Then, choosing b_N that satisfies $b_N \gtrsim (\log N)^d$ with $d > 1/2$ and $b_N \lesssim c_N k_N^{-1}$, we obtain (B.158). ∎