Cross-Sectional Variation of Intraday Liquidity, Cross-Impact, and their Effect on Portfolio Execution

Seungki Min
Graduate School of Business
Columbia University
smin20@gsb.columbia.edu

Costis Maglaras
Graduate School of Business
Columbia University
c.maglaras@gsb.columbia.edu

Ciamac C. Moallemi
Graduate School of Business
Columbia University
ciamac@gsb.columbia.edu

Abstract

An analysis of intraday volumes for the S&P 500 constituent stocks illustrates that (i) volume surprises, i.e., deviations from forecasted trading volumes, are correlated across stocks, and (ii) this correlation increases during the last few hours of the trading session. These observations can be attributed partly to the prevalence of portfolio trading activity that is implicit in the growth of passive (systematic) investment strategies, and partly to the increased trading intensity of such strategies towards the end of the trading session. In this paper, we investigate the consequences of such portfolio liquidity on price impact and portfolio execution. We derive a linear cross-asset market impact from a stylized model that explicitly captures the fact that a certain fraction of natural liquidity providers trade only portfolios of stocks whenever they choose to execute. We find that due to cross-impact and its intraday variation, it is optimal for a risk-neutral cost-minimizing liquidator to execute a portfolio of orders in a coupled manner, as opposed to the separable volume-weighted-average-price (VWAP) execution schedule that is often assumed. The optimal schedule couples the execution on the individual stocks so as to take advantage of increased portfolio liquidity towards the end of the day. A worst-case analysis shows that the potential cost reduction from this optimized execution schedule over the separable approach can be as high as 15% for plausible model parameters. Finally, we discuss how to estimate cross-sectional price impact if one had a dataset of realized portfolio transaction records by exploiting the low-rank structure of its coefficient matrix suggested by our analysis.

1. Introduction

Throughout the past decade or so we have experienced a so-called movement of assets under management in the equities markets from actively managed to passively and systematically managed strategies. This migration of assets has also been accompanied by the simultaneous growth of exchange-traded funds (ETFs). In very broad strokes, passive strategies tend to base investment and trade decisions on systematic portfolio-level procedures – e.g., invest in all S&P 500 constituents

proportionally to their respective market capitalization weights, invest in low-volatility stocks, high-beta stocks, high dividend stocks, etc. By contrast, active strategies tend to base investment decisions on individual firm-level procedures – e.g., invest in a particular stock selectively. In the sequel, we will refer to passive strategies as "index fund" strategies.

This gradual shift in investment styles has affected the nature of trade order flows, which motivates our subsequent analysis. We make three specific observations. First, passive and systematic strategies tend to generate portfolio trade order flows, i.e., trades that simultaneously execute orders in multiple securities in a coordinated fashion, e.g., buying a $50 million slice of the S&P 500 over the next two hours that involves the simultaneous execution of buy orders along most or all of the index constituents. Second, passive strategies tend to concentrate their trading activity towards the end of the day, partly so as to focus around times with increased market liquidity, and partly because mutual funds that implement such strategies have to settle buy and sell trade instructions from their (retail) investors at the closing market price at the end of each day; ETF products exhibit similar behavior. Third, the shift in asset ownership over time and the changes in the regulatory environment have changed the composition and strategies under which natural liquidity[1] is provided in the market; these are the counterparties that step in to either sell or buy stock against institutional investors so as to clear the market.

In §2 we will provide some empirical evidence that pairwise correlations amongst trading volumes across the S&P 500 constituents are positive throughout the trading day, and increase by about a factor of two over the last 1–2 hours of the trading day. That is, trading volumes exhibit common intraday variation away from their deterministic forecast in a way that is consistent with our earlier observations.

In this paper we study the effect of portfolio liquidity provision in the context of optimal trade execution. Specifically, we consider a stylized model of natural liquidity provision that incorporates the behavior of single-stock and portfolio participants, and that leads to a market impact model that incorporates cross-security impact terms; these arise due to the participation of natural portfolio liquidity providers. We formulate and solve a multi-period optimization problem to minimize the expected market impact cost incurred by a risk-neutral investor that seeks to liquidate a portfolio. We characterize the optimal policy, which is "coupled" – i.e., the liquidation schedules for the various orders in the portfolio are jointly determined so as to incorporate and exploit the cross-security impact phenomena. We contrast this optimal schedule with a separable execution schedule, where

[1]We use the term "natural" liquidity to indicate demand or supply from the investors who make investment decisions with their own perspective as opposed to the liquidity provided by market makers or arbitrageurs who may not affect the equilibrium market price.

the orders in the portfolio are executed independently of each other; this is commonly adopted by risk-neutral investors. Separable execution is suboptimal, in general, and we derive a bound on the sub-optimality gap of such a separable execution schedule against the optimal portfolio execution schedule, which can be interpreted as the execution cost reduction that an investor can achieve by optimizing around such cross-impact effects.

Contributions. In detail, the main contributions of the paper are the followings.

First, we propose a stylized model of cross-sectional price impact that highlights the effect of portfolio liquidity provision. Under the assumption that the magnitude of single-stock and portfolio liquidity provision is linear in the change in short-term trading prices, we show that market impact is itself linear in the trade quantity vector, and characterize the coefficient matrix that exhibits an intuitive structure: it is the inverse of a matrix that is decomposed into a diagonal matrix plus a (non-diagonal) low-rank matrix where the diagonal components capture the effect of single-stock liquidity providers and the non-diagonal terms capture the effect of portfolio liquidity providers that are assumed to trade along a set of portfolio weight vectors, such as the market and sector portfolios. Cross-impact is the result of portfolio liquidity provision.

Second, we show that optimal trade scheduling for risk-neutral minimum cost liquidation is coupled. We formulate and solve a multi-period optimization problem that selects the quantities to be traded in each security over time so as to liquidate the target portfolio over the span of a finite horizon (a day in our case) in a way that minimizes the cumulative expected market impact costs. The optimal trade schedule is coupled, and, specifically, incorporates and exploits the presence and intraday variation of cross-impact effects. Coupling is not the result of a risk penalty that captures the covariance of intraday price returns, as is typically the case (Almgren and Chriss, 2000; Tsoukalas et al., 2017), but the result of correlated liquidity. We identify the special cases where a separable execution approach would be optimal, namely, when a) there was no portfolio liquidity provision, or b) the intensity of portfolio liquidity provision varies proportionally to the intensity of single-stock liquidity provision throughout the trading day.

Third, we compare the optimal policy to a separable volume-weighted-average-price (VWAP) execution policy, and characterize the worst-case liquidation portfolios and the magnitude of the benefit that one derives from the optimized solution. A straightforward estimation of the mixture of single-stock and portfolio liquidity providers that would be consistent with the intraday volume profile and the intraday profile of pairwise volume correlations can be converted back into a numerical value for the aforementioned bound, which is around 15%. The worst-case analysis provides some intuition on the settings where this effect may be more pronounced.

Last, we propose an efficient procedure to estimate the suggested cross-asset impact model, i.e., a practical scheme for estimating the (time-varying) coefficient matrix for price impact. A direct estimation procedure for all cross-impact coefficients between each pair of stocks seems intractable due to the low signal-to-noise ratio that often characterizes market impact model estimation and the increased sparsity of trade data when we study pairs of stocks. Exploiting the low-rank structure of our stylized impact model derived above, we propose a procedure that involves the estimation of only a few parameters, e.g., one parameter per sector. We do not calibrate the cross-impact model, as this typically requires access to proprietary trading information, but we do specify a detailed procedure verified using synthetic data.

Linear cross-impact model. In the derivation of cross-impact model, we make several stylized assumptions that lead to a linear and transient impact. This is clearly a simplification that allows us to show through a tractable and insightful analysis the implication of portfolio natural liquidity provision to impact costs and execution schedules. Specifically, cross-impact terms arise from this source of liquidity, and optimal schedules are coupled across orders to account for such interaction effects: in settings where portfolio liquidity provision is attractive in terms of its supply curve's price sensitivity parameter, the optimized schedules deviate from separable VWAP-like ones to "tilt" and take advantage of portfolio liquidity provision. We believe that these insights are robust to the functional form of the impact model as long as the portfolio liquidity is present. And, specifically, a linear permanent impact cost contribution can be readily incorporated still within the tractability of a different quadratic optimization problem, and similarly for linear transient impact with decay. If the functional form of the impact cost function is non-linear, e.g., square-root or some other rational power, then while we expect that the key insights should continue to hold, the optimal execution schedule can only be found numerically and tends to be extreme (Curato et al., 2014).

Literature survey. One set of papers that is related to our work focuses on optimal trade scheduling, where the investor considers a dynamic control problem of splitting the liquidation of a large order over a predetermined time horizon so as to optimize some performance criterion. Bertsimas and Lo (1998) solve this problem in the context of minimizing the expected market impact cost, and Almgren and Chriss (2000) extend the analysis to the mean-variance criterion; see also Almgren (2003) and Huberman and Stanzl (2005). Bertsimas and Lo (1998) show that the cost-minimizing solution under a linear impact model schedules each order in proportion to the stock's forecasted volume profile. In these papers, multiple-security trading is briefly discussed as an extension of single-stock execution, and a similar setup can be found in recent studies (e.g., Brown

et al. (2010), Haugh and Wang (2014)). A separate strand of work, which includes Obizhaeva and Wang (2013), Rosu (2009), and Alfonsi et al. (2010), treat the market as one limit order book and use an aggregated and stylized model of market impact to capture how the price moves as a function of trading intensity. Tsoukalas et al. (2017) build on Obizhaeva and Wang (2013) to consider a portfolio liquidation problem incorporating risk and cross-impact effects and illustrate that the coupled execution is optimal for a risk-averse trader. Finally, closest to our paper is the recent work of Mastromatteo et al. (2017) that looks at portfolio execution with a linear cost model with cross-impact terms; their analysis predicates that the portfolio impact matrix has the same eigenvectors as the return correlation matrix, and is stationary. The problem structure allows for their model to be estimable – in a way similar to what we suggest in our paper, and the stationary model leads to a separable optimal trading schedule, which agrees with our results for that special case.

Apart from the execution scheduling problem, consistent efforts have been made to understand the nature of price impact theoretically and empirically. The seminal work of Kyle (1985) justifies a linear (permanent) price impact within a framework of rational expectations in which the market price is understood as an outcome of an equilibrium among the traders; our stylized derivation partly adopts the ideas therein. Huberman and Stanzl (2004) show using a no-arbitrage argument that the permanent price impact must be a linear function of the quantity traded (in the absence of temporary impact) and extend the argument to a multi-asset and time-dependent framework. Similarly, Schneider and Lillo (2019) show that linearity and symmetry are required for a cross-impact model to exclude arbitrage opportunities in a continuous time and transient impact setting. Identifying the functional form of price impact and its interaction with the price dynamics has been a topic of many empirical studies. Almgren et al. (2005) report an estimation result that supports a linear permanent impact and a sub-linear temporary impact. Tóth et al. (2011, 2018) report that the price impact at the meta-order level is a concave function of total order size, which is known as a square-root impact law; see also Capponi and Cont (2019) and Bucci et al. (2019). Building these empirical findings, a number of impact models have been proposed such as a transient impact model (Bouchaud et al., 2008; Gatheral et al., 2012), a history dependent permanent impact model (Bouchaud et al., 2008), and a latent order book model (Donier et al., 2015); however, nonlinear impact models are less amenable to direct analysis.

The topic of cross-impact has recently started to be explored. Specifically in Benzaquen et al. (2016), the authors postulate and estimate a linear propagator impact model based on the trade sign imbalance vector in each period, and observe that the eigenvectors of cross-impact matrix

coincide with those of return covariance matrix, which had motivated the aforementioned work of Mastromatteo et al. (2017) and a recent work of Tomas et al. (2020). A similar characterization can be found in the earlier works of the financial econometrics literature (Hasbrouck and Seppi, 2001; Lo and Wang, 2000, 2009) that adopt common factor models to analyze co-movement in returns/trading volumes across stocks, and attribute such a commonality to the portfolio order flows. While we do not estimate the cross-asset impact as this typically requires proprietary trade data as opposed to publicly available market data, our model motivation and predictions are consistent with these studies. In our paper, we further incorporate the temporal pattern of liquidity to examine its consequence in portfolio execution scheduling.

An important motivation of our work is the gradual shift of assets under management from active to passive and systematic strategies, and its implication for market behavior and the composition and timing of trading flows. In particular, focusing on the topic of liquidity, which is our main concern, this literature has found a causal relationship between ETF or mutual fund ownership and the commonality in the liquidity of the underlying constituents, e.g., Ben-David et al. (2017), Karoli et al. (2012), Koch et al. (2016), Agarwal et al. (2018); the motivation of that cross-sectional dependency is attributed to the arbitrage mechanism of ETFs or the correlated trading of mutual funds.[2]

Commonality in trading volume and portfolio liquidity provision. Throughout the paper, we connect two concepts – the correlation in trading volume across stocks and the cross-asset (non-diagonal) terms in market impact. We argue in §2 that the correlation in volume is attributable to the portfolio order flows, and in §3 that the cross-asset impact is partially attributable to the liquidity provision at a portfolio level, both of which are interpreted as reflection of the portfolio investors' participation. We further motivate in §4 a parametric intraday variation of cross-asset impact from the observed intraday pattern of volume correlation illustrated in §2. A more explicit connection is made in §5 based on a Poisson process analogy so as to provide a numerical illustration. We discuss this issue also in Appendix C, where we suggest and demonstrate an estimation procedure.

2. Preliminary Empirical Observations

To motivate our downstream analysis, we provide some empirical evidence for the cross-sectional behavior of intraday trading volume, focusing on the level and intraday variation of the pairwise

[2]The concentration of trading flows towards the end of the trading day has been a popular topic in the financial press; see, e.g., Driebusch et al. (2018).

correlations among trading volumes of the S&P 500 constituent stocks. We analyzed 482 stocks ($N = 482$), denoted by i, that were constituents of S&P 500 throughout the calendar year of 2018.[3] Our dataset contains 240 days ($D = 240$), denoted by d, excluding days that are known to exhibit abnormal trading activity, namely, the FOMC/FED announcement days on 01/31, 03/21, 05/02, 06/13, 08/01, 09/26, 11/08, and 12/19, and the half trading days on 07/03, 11/23, 12/05, and 12/24.

We use a Trade-and-Quote (TAQ) database, and extract all trades, excluding those that: a) occur before 09:35 or after 16:00; b) opening auction prints or closing auction prints (COND field contains "O," "Q," "M," or "6"); and c) trades corrected later (CORR field is not 0, or COND field contains "G" or "Z"). We divide a day into five-minute intervals ($T = 77$, 09:35–09:40, ..., 15:55–16:00), denoted by t. We denote by DVol_{idt} the aggregate notional ($) volume traded on stock i across all transactions that took place in time interval t on day d. We define $\overline{\mathrm{DVol}}_{it}$ to be the yearly average notional volume traded on stock i in time period t, and $\mathrm{AvgVolAlloc}_t$ to be the cross-sectional average percentage of daily volume traded in period t ("daily volume" in this definition accounts for all trading activity between 9:35 and 16:00, excluding auction and corrected prints):

$$\overline{\mathrm{DVol}}_{it} \triangleq \frac{1}{D}\sum_{d=1}^{D}\mathrm{DVol}_{idt}, \quad \mathrm{VolAlloc}_{it} \triangleq \frac{\overline{\mathrm{DVol}}_{it}}{\sum_{s=1}^{T}\overline{\mathrm{DVol}}_{is}} \quad \text{and} \quad \mathrm{AvgVolAlloc}_t \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathrm{VolAlloc}_{it}. \quad (1)$$

For each pair of stocks (i, j) we denote by Correl_{ijt} the pairwise correlation between the respective intraday notional traded volumes across days for each time period t. As a measure of cross-sectional dependency, we subsequently calculate the average pairwise correlation over all pairs of stocks:

$$\mathrm{Correl}_{ijt} \triangleq \frac{\sum_{d=1}^{D}(\mathrm{DVol}_{idt} - \overline{\mathrm{DVol}}_{it})(\mathrm{DVol}_{jdt} - \overline{\mathrm{DVol}}_{jt})}{\sqrt{\sum_{d=1}^{D}(\mathrm{DVol}_{idt} - \overline{\mathrm{DVol}}_{it})^2 \cdot \sum_{d=1}^{D}(\mathrm{DVol}_{jdt} - \overline{\mathrm{DVol}}_{jt})^2}}, \quad (2)$$

$$\mathrm{AvgCorrel}_t \triangleq \frac{1}{N(N-1)}\sum_{i\neq j}\mathrm{Correl}_{ijt}. \quad (3)$$

Figure 1 depicts the graphs of $\mathrm{AvgVolAlloc}_t$ and $\mathrm{AvgCorrel}_t$. $\mathrm{AvgVolAlloc}_t$ exhibits the commonly observed U-shaped behavior that shows that trading activity is concentrated in the morning and the end of the day. The graph of $\mathrm{AvgCorrel}_t$ reveals that (i) trading volumes are positively correlated throughout the day, and (ii) the cross-sectional average pairwise correlation increases

[3] See Appendix §A for additional empirical analysis on the years before 2018.

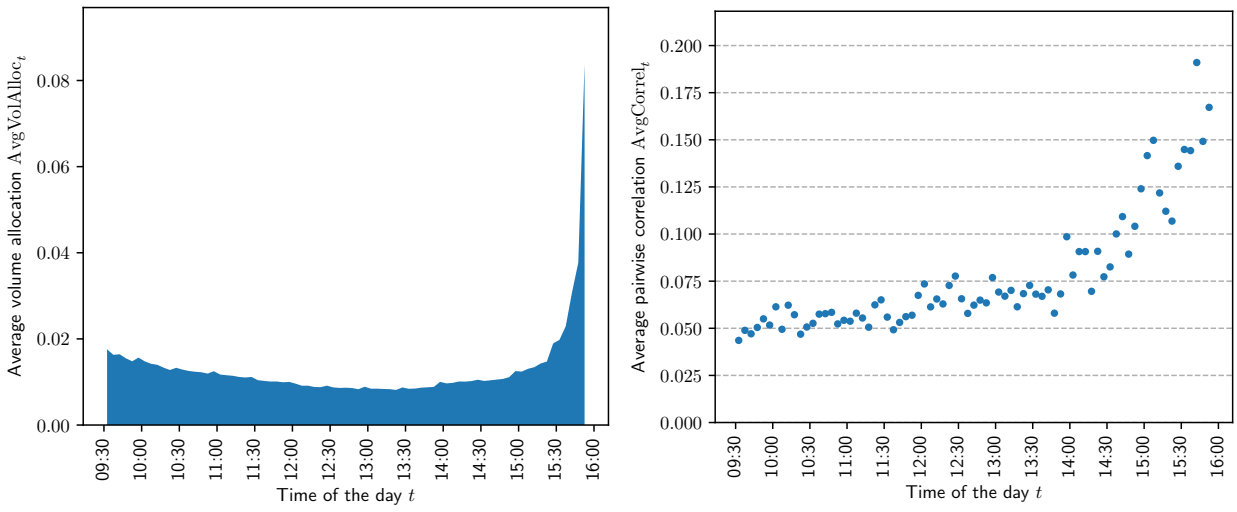

Figure 1: Cross-sectional average intraday traded volume profile (left) and cross-sectional average pairwise correlation (right): S&P 500 constituent stocks in 2018.

significantly during the last few hours of the day.[4]

One possible explanation of the observed intraday volume correlation profile could be the non-stationary participation of portfolio order flow throughout the course of the trading session. Market participants that trade portfolio order flow cause correlated stochastic volume deviations across stocks that, in turn, could contribute to the observed pairwise correlation profile. Interpreting portfolio order flows as the primary source of cross-sectional dependency in trading volume, AvgCorrel_t indirectly reflects the intensity of portfolio order flow within the total market order flow. Our empirical observation indicates that (i) portfolio order flow contributes a certain fraction of trading activity throughout the day, which (ii) is increasing towards the end of the day. In particular, with the increasing popularity of ETFs and passive funds in recent years, people now trade similar portfolios which may induce stronger cross-sectional dependency; Karoli et al. (2012), Koch et al. (2016), and Agarwal et al. (2018) provide empirical evidence that the commonality in trading volume indeed arises from the trading activity in ETFs or passive funds. Similarly, transactions to buy or sell shares of mutual funds are settled at the closing prices, and mutual fund companies tend to execute the net inflows or outflows near or at the end of the trading session.

We will return to these findings on AvgVolAlloc_t and AvgCorrel_t in §5 in order to approximate the relative magnitude of each different type of natural liquidity providers (portfolio vs. single-stock

[4] Alternative calculations of the intraday volume and correlation patterns produce similar findings. For example, one could compute stock-specific average traded volume profiles, and for each day compute the stock-specific normalized volume deviation profiles between the realized and forecasted volume profiles; these could be used for the pairwise correlation analysis. Similar findings are obtained when we study stocks clustered by their sector, e.g., among financial, energy, manufacturing, etc., stock sub-universes.

investors), and characterize its effect on the optimal execution schedule and execution costs.

3. Model

We assume that there are two types of investors – single-stock and index-fund investors – that provide natural liquidity in the market. In this section, we derive the cross-sectional market impact model from a stylized assumption on the liquidity provision mechanism of these investors. The term "single stock" here refers to discretionary or active investors that are willing to supply liquidity on individual securities.

3.1. Single-stock Investors and Index-fund Investors

Single-stock (discretionary) investors are assumed to trade and provide opportunistic liquidity on individual stocks by adjusting their holdings in response to changes in the price of the stock. A change in single-stock investor holdings in stock i is assumed to be linear in the change in the market price with a coefficient $\psi_{\text{id},i}$. Single-stock investors will sell (or buy) $\psi_{\text{id},i}$ shares of stock i when its price p_i rises (or drops) by one dollar.

A linear supply relationship between holdings and price is often assumed in the market microstructure literature (Tauchen and Pitts, 1983; Kyle, 1985). It is typically justified under the assumption that a risk-averse investor chooses his holdings to maximize his expected utility given his own belief on the future price. With a constant absolute risk aversion (CARA) utility function and normally distributed beliefs, the optimal holding position is proportional to the gap between the current price and his own reservation price, with a proportionality coefficient that incorporates his confidence in his belief and his preference on uncertainty. Our parameter $\psi_{\text{id},i}$ can be thought as a sum of the individual investors' sensitivity parameters.

We consider a universe of N stocks, denoted by $i = 1, \ldots, N$. Suppose that the change in the N-dimensional price vector is $\mathbf{\Delta p} \in \mathbb{R}^N$. Let $\mathbf{e}_i$ be the i^{th} standard basis vector. Single-stock investors on stock i will experience the price change $\mathbf{e}_i^\top \mathbf{\Delta p}$ and adjust their holding position by $-\psi_{\text{id},i} \cdot \mathbf{e}_i^\top \mathbf{\Delta p}$. In vector representation, the change in the holding vector of single-stock investors $\mathbf{\Delta h}_{\text{id}} \in \mathbb{R}^N$ can be written as

$$\mathbf{\Delta h}_{\text{id}}\left(\mathbf{\Delta p}\right) = -\sum_{i=1}^{N} \mathbf{e}_i \cdot \psi_{\text{id},i} \cdot \mathbf{e}_i^\top \mathbf{\Delta p} = -\mathbf{\Psi}_{\text{id}} \mathbf{\Delta p} \quad \in \mathbb{R}^N, \tag{4}$$

where $\mathbf{\Psi}_{\text{id}} \triangleq \text{diag}\left(\psi_{\text{id},1}, \ldots, \psi_{\text{id},N}\right) \in \mathbb{R}^{N \times N}$. The quantity $\mathbf{\Delta h}_{\text{id}}\left(\mathbf{\Delta p}\right)$ can be thought as "signed"-volume; i.e., it is positive when orders to buy are submitted in the market when the prices drop,

and negative when orders to sell are submitted in the market when prices rise.

In turn, index-fund investors trade "portfolios" based on some view on the entire market, a sector, or a particular group of securities such as high-beta stocks. This investor type includes many institutional investors, but the individual investors who hold ETFs or join index funds also belong to this group. We assume that there are K such funds, denoted by $k = 1, \ldots, K$. Let $\mathbf{w}_k = (w_{k1}, \ldots, w_{kN})^\top \in \mathbb{R}^N$ be the weight vector of index fund k, expressed in the number of shares: one unit of index fund k contains w_{k1} shares of stock 1, w_{k2} shares of stock 2, and so on. Given a price change $\Delta\mathbf{p} \in \mathbb{R}^N$, investors in index fund k will experience the price change $\mathbf{w}_k^\top \Delta\mathbf{p}$. Analogous to single-stock investors, index-fund investors adjust their holding position on index fund k linearly to its price change $\mathbf{w}_k^\top \Delta\mathbf{p}$ with a coefficient $\psi_{\mathrm{f},k}$. Since trading one unit of index fund k is equivalent to trading a basket of individual stocks with weight vector $\mathbf{w}_k$, we can state the change in the index-fund investors' holding position vector $\Delta\mathbf{h}_\mathrm{f} \in \mathbb{R}^N$ as a vector of changes in the constituents of that fund:

$$\Delta\mathbf{h}_\mathrm{f}(\Delta\mathbf{p}) = -\sum_{k=1}^{K} \mathbf{w}_k \cdot \psi_{\mathrm{f},k} \cdot \mathbf{w}_k^\top \Delta\mathbf{p} = -\mathbf{W}\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top \Delta\mathbf{p} \quad \in \mathbb{R}^N, \tag{5}$$

where

$$\mathbf{W} \triangleq \begin{bmatrix} | & & | \\ \mathbf{w}_1 & \ldots & \mathbf{w}_K \\ | & & | \end{bmatrix} \in \mathbb{R}^{N\times K}, \quad \boldsymbol{\Psi}_\mathrm{f} \triangleq \mathrm{diag}\left(\psi_{\mathrm{f},1}, \ldots, \psi_{\mathrm{f},K}\right) \in \mathbb{R}^{K\times K}. \tag{6}$$

Throughout the paper, we assume that all $\psi_{\mathrm{id},i}$'s and $\psi_{\mathrm{f},k}$'s are strictly positive, and that $\mathbf{w}_k$'s are linearly independent.

To better illustrate, we provide a limit order book interpretation of the model. We first consider order books for single stocks and those for index funds in "isolation". A single-stock order book for stock i consists of the limit orders submitted by the single-stock investors, where the limit orders are distributed with a constant density $\psi_{\mathrm{id},i}$ (i.e., $\psi_{\mathrm{id},i}$ shares of stock i per one dollar interval) and are symmetric on buy and sell sides with a mid-price p_i and no bid-ask spread. For each index-fund k, the index-fund investors place the limit orders in a separate order book with a constant density $\psi_{\mathrm{f},k}$ (i.e., $\psi_{\mathrm{f},k}$ shares of index fund k per one dollar interval) and its mid-price is given by $\mathbf{w}_k^\top \mathbf{p}$. Let us now focus on the amount of limit orders available within a certain range of price in two types of order books in "aggregation". Within the price deviations $\Delta\mathbf{p} \in \mathbb{R}^N$ across single stocks (equivalently, $\mathbf{W}^\top \Delta\mathbf{p} \in \mathbb{R}^K$ across index funds), there will be $\boldsymbol{\Psi}_\mathrm{id}\Delta\mathbf{p}$ shares of single stocks available in the single-stock order books, and $\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top\Delta\mathbf{p}$ shares of index funds available in the index-fund order books (that are equivalent to $\mathbf{W}\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top\Delta\mathbf{p}$ shares of single stocks), and therefore

$\left(\mathbf{\Psi}_{\text{id}} + \mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top\right)\Delta\mathbf{p}$ shares of single stocks in total. In that sense, the linear impact model corresponds to limit order book that has the same thickness across price levels.

3.2. Cross-sectional Price Impact

We wish to execute $\mathbf{v} \in \mathbb{R}^N$ shares during a given time period. Depending on whether we want to buy or sell, each component can be positive or negative. Our orders (eventually) transact against natural liquidity provided by single-stock and index-fund investors; market makers and high-frequency traders intermediate the market but tend to maintain negligible inventories at the end of the day. A price change of $\Delta\mathbf{p} \in \mathbb{R}^N$ will affect an inventory change of $\mathbf{v}$ shares if the following market-clearing condition is satisfied:

$$\mathbf{v} + \Delta\mathbf{h}_{\text{id}}\left(\Delta\mathbf{p}\right) + \Delta\mathbf{h}_{\text{f}}\left(\Delta\mathbf{p}\right) = \mathbf{0}. \tag{7}$$

By equations (4) and (5),

$$\mathbf{v} = \left(\sum_{i=1}^{N} \mathbf{e}_i \cdot \psi_{\text{id},i} \cdot \mathbf{e}_i^\top + \sum_{k=1}^{K} \mathbf{w}_k \cdot \psi_{\text{f},k} \cdot \mathbf{w}_k^\top\right)\Delta\mathbf{p} = \left(\mathbf{\Psi}_{\text{id}} + \mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top\right)\Delta\mathbf{p}. \tag{8}$$

In other words, out of $\mathbf{v}$ shares, $\mathbf{\Psi}_{\text{id}}\Delta\mathbf{p} \in \mathbb{R}^N$ shares are obtained from single-stock investors and $\mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top\Delta\mathbf{p} \in \mathbb{R}^N$ shares from index-fund investors. This linear relationship between $\mathbf{v}$ and $\Delta\mathbf{p}$ can be translated into the price impact summarized in the next proposition.

Proposition 1 (Cross-sectional price impact)**.** *When a liquidator executes $\mathbf{v} \in \mathbb{R}^N$ shares, the market-clearing price change vector $\Delta\mathbf{p} \in \mathbb{R}^N$ is such that*

$$\Delta\mathbf{p} = \mathbf{G}\mathbf{v} \quad \textit{and} \quad \mathbf{G} \triangleq \left(\mathbf{\Psi}_{\text{id}} + \mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top\right)^{-1}. \tag{9}$$

Note that the coefficient matrix $\mathbf{G}$ is an inverse of $\mathbf{\Psi}_{\text{id}} + \mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top$, which is composed of two symmetric and strictly positive-definite matrices. Therefore, $\mathbf{G}$ is itself a well-defined symmetric positive-definite matrix, with the following structure: a diagonal matrix plus a non-diagonal low-rank matrix. The following matrix expansion derived from an application of the Woodbury matrix identity will prove useful:

$$\mathbf{G} = \left(\underbrace{\mathbf{\Psi}_{\text{id}}}_{\text{diagonal}} + \underbrace{\mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top}_{\text{rank } K}\right)^{-1} = \underbrace{\mathbf{\Psi}_{\text{id}}^{-1}}_{\text{diagonal}} - \underbrace{\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\mathbf{\Psi}_{\text{f}}^{-1} + \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}}_{\text{rank } K}. \tag{10}$$

Proposition 1 characterizes the structure of the cross-price impact model. The cross-impact is captured by the non-diagonal entries in $\mathbf{W}\mathbf{\Psi}_{\mathrm{f}}\mathbf{W}^{\top}$ that result from the natural liquidity provision attributed to index-fund (portfolio) investors.

We interpret the terms $\mathbf{\Psi}_{\mathrm{id}} = \mathrm{diag}_{i=1}^{N}(\psi_{\mathrm{id},i})$ and $\mathbf{\Psi}_{\mathrm{f}} = \mathrm{diag}_{k=1}^{K}(\psi_{\mathrm{f},k})$ as *"liquidity"*. The component $\psi_{\mathrm{id},i}$ represents the amount of liquidity provided by single-stock investors in stock i and $\psi_{\mathrm{f},k}$ represents the amount of liquidity supplied by index-fund investors in index fund k. The sum $\mathbf{\Psi}_{\mathrm{id}} + \mathbf{W}\mathbf{\Psi}_{\mathrm{f}}\mathbf{W}^{\top}$ indicates the total market liquidity. As shown in (9), price impact is inversely proportional to liquidity, which agrees with the conventional definition of liquidity as a measure of ease of trading. When $\psi_{\mathrm{id},i}$ or $\psi_{\mathrm{f},k}$ is large, equivalently when liquidity is abundant, price impact is low. Since $\psi_{\mathrm{id},i}$ and $\psi_{\mathrm{f},k}$ are defined as the sensitivity of investors' holdings to market price movements, these terms are a measure of price impact that capture how many shares we can obtain from these two types of investors when the price moves by a certain amount.

3.3. One-period Transaction Cost

Consider a liquidator that wishes to execute $\mathbf{v} \in \mathbb{R}^N$ shares in a short period of time, say over 5 to 15 minutes. Let $\mathbf{p}_0 \in \mathbb{R}^N$ be the price at the beginning of this execution period. Assuming that $\mathbf{v}$ is traded continuously and at a constant rate over the duration of that time period, the liquidator will realize an average transaction price given by

$$\bar{\mathbf{p}}^{\mathrm{tr}} = \mathbf{p}_0 + \frac{1}{2}\mathbf{G}\mathbf{v} + \bar{\boldsymbol{\epsilon}}^{\mathrm{tr}}, \tag{11}$$

where $\bar{\boldsymbol{\epsilon}}^{\mathrm{tr}} \in \mathbb{R}^N$ represents a random error term that captures unpredictable market price fluctuations or the effect of trades executed in that period by other investors. The equation (11) suggests that costs accumulate linearly over the duration of the period, and that the average price change is half the end-to-end impact plus a random contribution due to fluctuations in the price due to exogenous factors. (We will return to this assumption later on.) We will assume that the error is independent of our execution $\mathbf{v}$ and zero mean, i.e., $\mathsf{E}\left[\bar{\boldsymbol{\epsilon}}^{\mathrm{tr}} | \mathbf{v}\right] = \mathbf{0}$. The single-period expected implementation shortfall incurred by the liquidator is given by

$$\bar{\mathcal{C}}\left(\mathbf{v}\right) \triangleq \mathsf{E}\left[\mathbf{v}^{\top}\left(\bar{\mathbf{p}}^{\mathrm{tr}} - \mathbf{p}_0\right)\right] = \frac{1}{2}\mathbf{v}^{\top}\mathbf{G}\mathbf{v}. \tag{12}$$

Linear price impact induces quadratic implementation shortfall costs; note that the resulting cost is always positive since $\mathbf{G}$ is positive-definite. The following proposition briefly explores how the mixture of natural liquidity providers affects the expected execution cost.

Proposition 2 (Extreme cases). *Consider a parametric scaling of the single-stock and index-fund natural liquidity, $\mathbf{\Psi}_{\text{id}}$ and $\mathbf{\Psi}_{\text{f}}$, respectively, given by*

$$\mathbf{G} = \left(\alpha \cdot \mathbf{\Psi}_{\text{id}} + \beta \cdot \mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^{\top}\right)^{-1}, \tag{13}$$

for some scalars $\alpha \in (0, 1]$ and $\beta \in (0, 1]$.

(i) If there are no index-fund investors ($\alpha \to 1$ and $\beta \to 0$), the expected execution cost becomes separable across individual assets:

$$\lim_{\alpha \to 1, \beta \to 0} \bar{\mathcal{C}}\left(\mathbf{v}\right) = \frac{1}{2}\mathbf{v}^{\top}\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{v} = \frac{1}{2}\sum_{i=1}^{N}\frac{v_i^2}{\psi_{\text{id},i}}. \tag{14}$$

(ii) If there are no single-stock investors ($\alpha \to 0$ and $\beta \to 1$), the liquidator can execute portfolio orders with finite expected execution cost only when the orders can be expressed as a linear combination of the index-fund weight vectors. Specifically,

$$\lim_{\alpha \to 0, \beta \to 1} \bar{\mathcal{C}}\left(\mathbf{v}\right) = \begin{cases} \infty & \text{if } \mathbf{v} \notin \operatorname{span}\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right), \\ \frac{1}{2}\mathbf{u}^{\top}\mathbf{\Psi}_{\text{f}}^{-1}\mathbf{u} & \text{if } \mathbf{v} = \mathbf{W}\mathbf{u}. \end{cases} \tag{15}$$

(The proof is provided in Appendix D.1.) Therefore, separable (security-by-security) market impact cost models, often assumed in practice, essentially predicate, as per our analysis, that all natural liquidity in the market is provided by opportunistic single-stock investors. And, in that case, (14) recovers the commonly used "diagonal" market impact cost model. The other extreme scenario assumes that all liquidity is provided along the weight vectors of the index-fund investors, and the resulting cost then depends on how the target execution vector $\mathbf{v}$ can be expressed as a linear combination of $(\mathbf{w}_1, \ldots, \mathbf{w}_K)$. In practice, the latter case suggests that execution costs may increase in periods with a relatively higher intensity of portfolio liquidity provision when the target portfolio that is being liquidated is not well aligned with the directions in which portfolio liquidity is supplied.

3.4. Time-varying Liquidity and Multi-period Transaction Costs

The stylized observations of Proposition 2 suggest that intraday trading costs may be affected by intraday variations in the mixture of natural liquidity providers, and, in particular, if the relative contribution of index-fund investors increases significantly over time.

We will consider the transaction cost of an intraday execution schedule $\mathbf{v}_1, \ldots, \mathbf{v}_T$ over T periods, in which $\mathbf{v}_t \in \mathbb{R}^N$ shares are executed during the time interval t. We will make the following

assumptions on the intraday behavior of price impact, price dynamics, and realized execution costs.

a) We allow the mixture of liquidity provision to fluctuate over the course of the day. We denote the time-varying liquidity by $\psi_{\mathrm{id},it}$ and $\psi_{\mathrm{f},kt}$ with an additional subscript t. We assume that the portfolio weight vectors $\mathbf{w}_k$ of index liquidity providers are fixed during a given day. Under this setting, the coefficient matrix of price impact can be represented as follows:

$$\mathbf{G}_t = \left(\mathbf{\Psi}_{\mathrm{id},t} + \mathbf{W}\mathbf{\Psi}_{\mathrm{f},t}\mathbf{W}^\top\right)^{-1}.$$

b) Let $\mathbf{p}_t$ be the fundamental price at the end of period t. The "fundamental" price denotes the price on which the market agrees as a best guess of the future price excluding the temporary deviation of the realized transaction price due to market impact. The fundamental price process $(\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_T)$ is assumed to be a martingale independent of the execution schedule:

$$\mathbf{p}_t = \mathbf{p}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{for all } t = 1, \ldots, T,$$

where the innovation term $\boldsymbol{\epsilon}_t$ satisfies $\mathsf{E}\left[\boldsymbol{\epsilon}_t | \mathcal{F}_{t-1}\right] = \mathbf{0}$ and $\mathcal{F}_{t-1}$ denotes all past information. The term $\boldsymbol{\epsilon}_t$ is commonly understood as the change in a market participant's belief perhaps due to the information revealed during period t. We are implicitly assuming that our execution conveys no information about the future price.

c) The realized "transaction" price in each period can deviate from the fundamental price temporarily, e.g., due to a short-term imbalance between buying order flow and selling order flow. In executing $\mathbf{v}_t$ shares, the liquidator is contributing to such an imbalance, which causes the temporary price impact according to the mechanism described above. We assume that this impact is temporary, and we particularly assume that the transaction price begins at the fundamental price in each period regardless of the liquidator's trading activity in prior periods.[5] Given the coefficient matrix $\mathbf{G}_t$, when $\mathbf{v}_t$ is executed smoothly, the average transaction price is

$$\bar{\mathbf{p}}_t^{\mathrm{tr}} = \mathbf{p}_{t-1} + \frac{1}{2}\mathbf{G}_t\mathbf{v}_t + \bar{\boldsymbol{\epsilon}}_t^{\mathrm{tr}},$$

where the error term $\bar{\boldsymbol{\epsilon}}_t^{\mathrm{tr}}$ satisfies $\mathsf{E}\left[\bar{\boldsymbol{\epsilon}}_t^{\mathrm{tr}} | \mathbf{v}_t\right] = \mathbf{0}$ as before.

Under these assumptions, the expected transaction cost of executing a series of portfolio trans-

[5] Even in the presence of permanent impact, if it is linear, symmetric, and time-invariant, it does not affect the optimal trading schedule. See Appendix A of Almgren and Chriss (2000).

actions $\mathbf{v}_1, \ldots, \mathbf{v}_T$ is separable over time and can be expressed as follows:

$$\bar{\mathcal{C}}\left(\mathbf{v}_1, \ldots, \mathbf{v}_T\right) \triangleq \mathsf{E}\left[\sum_{t=1}^{T} \mathbf{v}_t^\top \left(\bar{\mathbf{p}}_t^{\mathrm{tr}} - \mathbf{p}_0\right)\right] = \sum_{t=1}^{T} \frac{1}{2} \mathbf{v}_t^\top \mathbf{G}_t \mathbf{v}_t.$$

This formulation implicitly assumes that the intraday liquidity captured through $\psi_{\mathrm{id},it}$'s and $\psi_{\mathrm{f},kt}$'s is deterministic and known in advance. Although intraday liquidity evolves stochastically over the course of the day, its expected profile exhibits a fairly pronounced shape that serves as a forecast that can be used as a basis for analysis (as is done in practice). In later sections, we introduce more detailed parameterizations that utilize the intraday trading volume as an observable proxy for the intraday variation of the liquidity; see §4.2, §5.1 and Appendix C.

3.5. Discussion on Model

The cross-impact model derived in this paper can be characterized as a special case of the multi-asset version of Almgren-Chriss model (Almgren and Chriss, 2000, Appendix A), where we represent the temporary impact with symmetric positive definite matrices whose non-diagonal entries reflect the effect of portfolio liquidity provision. This would be the simplest form of cross-impact model that achieves tractability and economic soundness at the same time. One may consider a more generalized form that possibly involves non-linear, asymmetric, or transient/permanent components: for example, one could hypothesize a square root-like impact model by appropriately changing the underlying assumptions for the two types of liquidity providers. However, such a generalized model may no longer be analytically tractable and even may open the possibility of arbitrage or price manipulation. Indeed, Schneider and Lillo (2019) show that linearity and symmetry are required for a cross-impact model to exclude arbitrage opportunities although their setup is slightly different from ours,[6] and a similar conclusion can also be found in Huberman and Stanzl (2004). See also the discussion in §1 on the robustness of linearity assumption.

4. Optimal Portfolio Execution

We will formulate and solve the multi-period optimal portfolio execution problem in §4.1, and then explore the properties of the optimal solution as a function of intraday variations of the two sources of natural liquidity providers in §4.2.

[6] More specifically, Schneider and Lillo (2019) consider transient impact models in a continuous-time setting and show that if the impact is non-linear (Lemma 3.5) or asymmetric (Lemma 3.9) then there exists a round-trip trade schedule that yields a positive profit in expectation.

4.1. Optimal Trade Schedule

Consider a risk-neutral liquidator interested in executing $\mathbf{x}_0 \in \mathbb{R}^N$ shares over an execution horizon T (e.g., a day). We formulate a discrete-time optimization problem to find an optimal schedule $\mathbf{v}_1, \ldots, \mathbf{v}_T$ that minimizes the expected total transaction cost:

$$\text{minimize} \qquad \bar{\mathcal{C}}\left(\mathbf{v}_1, \ldots, \mathbf{v}_T\right) = \sum_{t=1}^{T} \frac{1}{2} \mathbf{v}_t^\top \mathbf{G}_t \mathbf{v}_t \tag{16}$$

$$\text{subject to} \qquad \sum_{t=1}^{T} \mathbf{v}_t = \mathbf{x}_0. \tag{17}$$

Proposition 3 ("Coupled" execution)**.** *The risk-neutral cost minimization problem (16–17) has a unique optimal solution given by*

$$\mathbf{v}_t^* \;=\; \mathbf{G}_t^{-1} \left(\sum_{s=1}^{T} \mathbf{G}_s^{-1} \right)^{-1} \mathbf{x}_0 \;=\; \left(\mathbf{\Psi}_{\mathrm{id},t} + \mathbf{W} \mathbf{\Psi}_{\mathrm{f},t} \mathbf{W}^\top \right) \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \mathbf{W} \bar{\mathbf{\Psi}}_{\mathrm{f}} \mathbf{W}^\top \right)^{-1} \mathbf{x}_0, \tag{18}$$

where the total daily liquidity $\bar{\mathbf{\Psi}}_{\mathrm{id}}$ *and* $\bar{\mathbf{\Psi}}_{\mathrm{f}}$ *are defined as follows:*

$$\bar{\mathbf{\Psi}}_{\mathrm{id}} \triangleq \sum_{t=1}^{T} \mathbf{\Psi}_{\mathrm{id},t}, \quad \bar{\mathbf{\Psi}}_{\mathrm{f}} \triangleq \sum_{t=1}^{T} \mathbf{\Psi}_{\mathrm{f},t}. \tag{19}$$

We make the following observations. First, the optimal solution is "coupled" across securities. Specifically, as long as the market impact is cross-sectional, the cost-minimizing solution needs to consider all orders simultaneously in optimally scheduling how to liquidate the constituent orders, as opposed to scheduling each order separately and attempting to minimize costs as if market impact were separable; such a separable execution approach is often used in practice (effectively assuming that there are no cross-impact effects). The coupled execution recognizes that the blend of natural liquidity changes intraday, and attempts to change the composition of the residual liquidation portfolio so as to take advantage of portfolio liquidity that may become available, say towards the end of the day, for example. We will explore this point further in the remainder of this section. Second, it is typical to derive coupled optimal portfolio trade schedules for risk-averse investors that consider the variance of the execution costs in the objective function or as a constraint; in that case, the covariance structure of the portfolio over its liquidation horizon intuitively leads to a coupled execution solution (Almgren and Chriss, 2000; Tsoukalas et al., 2017). In our problem formulation, the coupling of the execution path is driven by the cross-sectional dependency of natural (portfolio) liquidity provided by index-fund investors which leads to cross-impact, as opposed to the cross-

sectional dependency of intraday returns. Third, we note that in the above formulation we have not imposed side constraints that would enforce that the liquidation path is monotone; we will return to this point later on.

The structure of the optimal schedule in (18) takes an intuitive form: the proportion of the trade that is liquidated in period t is proportional to the available liquidity in that period, as captured by the time-dependent numerator matrix $\mathbf{\Psi}_{\mathrm{id},t} + \mathbf{W}\mathbf{\Psi}_{\mathrm{f},t}\mathbf{W}^\top$, normalized by the total liquidity made available throughout the day, as captured by the time-independent denominator matrix $\bar{\mathbf{\Psi}}_{\mathrm{id}} + \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top$. An alternative interpretation also given by (18) is that the optimal schedule splits the order in a way that is inversely proportional to a normalized time-dependent market impact matrix.

Corollary 1 (No index-fund investors, $\mathbf{\Psi}_{\mathrm{f},t} = 0$ for $t = 1, \ldots, T$)**.** *When there are no index-fund investors: i.e., $\bar{\boldsymbol{\psi}}_{\mathrm{f}} = \mathbf{0}$, a separable VWAP-like trade schedule is optimal:*

$$v^*_{it} = \frac{\psi_{\mathrm{id},it}}{\sum_{s=1}^{T} \psi_{\mathrm{id},is}} \cdot x_{i0}, \quad \textit{for } i = 1, \ldots, N. \tag{20}$$

Proof of Proposition 3. Note that since $\mathbf{G}_t$ is symmetric, $\frac{\partial}{\partial \mathbf{v}_t} \frac{1}{2} \mathbf{v}_t^\top \mathbf{G}_t \mathbf{v}_t = \mathbf{G}_t \mathbf{v}_t$. The Karush-Kuhn-Tucker (KKT) conditions of the convex minimization problem in (16)–(17) require that there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^N$ such that

$$\boldsymbol{\lambda} = \frac{\partial}{\partial \mathbf{v}_t} \frac{1}{2} \mathbf{v}_t^\top \mathbf{G}_t \mathbf{v}_t \Big|_{\mathbf{v}_t = \mathbf{v}_t^*} = \mathbf{G}_t \mathbf{v}_t^*, \quad \text{for all } t = 1, \ldots, T,$$

which together with the inventory constraint in (17) implies that

$$\mathbf{x}_0 = \sum_{t=1}^{T} \mathbf{v}_t^* = \sum_{t=1}^{T} \mathbf{G}_t^{-1} \boldsymbol{\lambda}.$$

It follows that $\mathbf{v}_t^* = \mathbf{G}_t^{-1}\boldsymbol{\lambda} = \mathbf{G}_t^{-1} \left(\sum_{s=1}^{T} \mathbf{G}_s^{-1} \right)^{-1} \mathbf{x}_0$. Since all $\mathbf{G}_t$'s are invertible, the optimal solution exists and is unique. ■

In a market where all natural liquidity is provided by single-stock, opportunistic investors, there are no cross-security impact effects, market impact is separable, and the minimum cost schedule for a risk-neutral liquidator is also separable across securities – the optimal solution simply needs to minimize expected impact costs separately for each order in the portfolio. Each individual order can be scheduled independently of the others, and the resulting schedule is VWAP-like in that the execution quantity v_{it} is proportional to the available liquidity $\psi_{\mathrm{id},it}$ at that moment.

Indeed, the overall market trading volume profile is treated as the observable proxy for the natural liquidity profile, the solution spreads each order separately and in a way that is proportional to the percentage of the market volume that is forecasted for each time period; this is what a typical VWAP execution algorithm does.

Conversely, if some of the natural liquidity is provided by index-fund investors that wish to trade portfolios, e.g., liquidate some amount of an energy-tracking portfolio if the energy sector has had a significant positive return intraday, the separable VWAP schedule would not minimize expected market impact costs, and would not be optimal for the motivating trade scheduling problem.

4.2. Optimal Trade Schedule under a Parametric Liquidity Profile

To gain some insight into the structure of the optimal policy, we explore a setting where the intensity of single-stock and index-fund investors' liquidity provision varies parametrically as follows: single-stock investors' liquidity $\psi_{\mathrm{id},it}$ varies over time $t = 1, 2, \ldots, T$ according to a profile α_t, and index-fund investors' liquidity $\psi_{\mathrm{f},kt}$ varies according to another profile β_t, i.e.,

$$\boldsymbol{\Psi}_{\mathrm{id},t} = \alpha_t \cdot \bar{\boldsymbol{\Psi}}_{\mathrm{id}}, \quad \boldsymbol{\Psi}_{\mathrm{f},t} = \beta_t \cdot \bar{\boldsymbol{\Psi}}_{\mathrm{f}}, \quad \text{for } t = 1, \ldots, T, \tag{21}$$

where $\sum_{t=1}^{T} \alpha_t = \sum_{t=1}^{T} \beta_t = 1$.

We will assume that all single stocks share the same time-varying profile α_t, and likewise all index funds share the profile β_t. The empirical findings of §2 indicate that pairwise correlations of trading volumes increase towards the end of the day. If a primary source of stochastic fluctuations in intraday trading volumes is the stochastic arrivals of single stock and portfolio trades, then one would expect that the profiles α_t, β_t vary intraday so as to generate the well-known U-shaped volume profile, and to vary differently from each other so as as to generate the time-varying pairwise correlation relationship; this is supported by the behavior of market participants towards the end of the day, as discussed earlier. Indeed, if the two sources of natural liquidity had the same trading activity profiles, i.e., $\alpha_t = \beta_t$, then the average correlation in intraday trading volume would not vary intraday. We expect that towards the end of the day, the intensity of index-fund liquidity provision (β_t) increases relatively faster than the intensity of single-stock liquidity provision (α_t).

Proposition 4 (Optimal execution under structured variation). *Under the parameterization of* (21)*, the schedule* $\mathbf{v}_t^*$ *is optimal for risk-neutral cost minimization* (16)*:*

$$\mathbf{v}_t^* = \alpha_t \cdot \mathbf{x}_0 + (\beta_t - \alpha_t) \cdot \mathbf{W} \left(\bar{\boldsymbol{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top \bar{\boldsymbol{\Psi}}_{\mathrm{id}}^{-1} \mathbf{W} \right)^{-1} \mathbf{W}^\top \bar{\boldsymbol{\Psi}}_{\mathrm{id}}^{-1} \mathbf{x}_0, \tag{22}$$

or, equivalently,

$$\mathbf{v}_t^* = \alpha_t \cdot \mathbf{x}_0 + (\beta_t - \alpha_t) \cdot \sum_{k=1}^{K} (\widehat{\mathbf{w}}_k^\top \mathbf{x}_0) \cdot \mathbf{w}_k, \tag{23}$$

where $\widehat{\mathbf{W}} \triangleq \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}$*, and* $\widehat{\mathbf{w}}_k$ *denotes the* k^{th} *column of* $\widehat{\mathbf{W}}$*.*

Before offering an interpretation for (23) we state the following corollary.

Corollary 2 (Optimal execution under common variation)**.** *If* $\alpha_t = \beta_t$ *for all* $t = 1, \ldots, T$*, a separable, VWAP-like strategy is again optimal:*

$$\mathbf{v}_t^* = \alpha_t \cdot \mathbf{x}_0. \tag{24}$$

The proof of Proposition 4 is given in Appendix D.2. Corollary 2 states that when the intensity of natural liquidity provision is the same for single-stock and index-fund investors, i.e., $\alpha_t = \beta_t$, the optimal schedule $\mathbf{v}_t^*$ is again aligned with $\mathbf{x}_0$ scaled by α_t. As α_t $(= \beta_t)$ represents the market activity at time t, the above policy can be interpreted as a VWAP-like execution that spreads each individual order proportionally to the total volume available at each point in time; this is separable across orders. As noted earlier, the setting where $\alpha_t = \beta_t$ is inconsistent with the empirical findings on the intraday behavior of pairwise correlations of trading volumes.

In contrast, (23) highlights that when the mixture of natural liquidity varies intraday (through the difference between α_t and β_t), the optimal schedule tilts away from the VWAP-like execution encountered in (24) so as to take advantage of an increase in available index-fund liquidity, e.g., offered along the direction of sector portfolios.

5. Illustration of Optimal Execution and Performance Bounds

In this section we provide a brief illustration of the optimized execution path that incorporates the effect of index-fund (portfolio) liquidity. Risk-neutral investors often adopt a separable execution style, i.e., trade each asset separately, most often using a volume-weighted average-price (VWAP) algorithm. As we show in §4, this separable strategy, under some assumptions, can be shown to minimize expected impact costs per order, but disregards the effect of portfolio liquidity and cross-impact costs when multiple orders are traded side by side. For a stylized model of natural liquidity of the form introduced in §4.2 simplified to the case of a single index fund (e.g., the market portfolio), we establish a worst-case bound on the sub-optimality gap of such a separable execution schedule against the optimized portfolio execution schedule derived above.

Specifically, restricting attention to the parameterization introduced in §4.2 in a setting with a single index fund ($K = 1$), we have $\mathbf{\Psi}_{\mathrm{id},t} = \alpha_t \cdot \bar{\mathbf{\Psi}}_{\mathrm{id}}$ and $\mathbf{\Psi}_{\mathrm{f},t} = \beta_t \cdot \bar{\mathbf{\Psi}}_{\mathrm{f}}$ with $\sum_{t=1}^{T} \alpha_t = \sum_{t=1}^{T} \beta_t = 1$. Proposition 4 states that the optimal execution $\mathbf{v}_t^*$ is

$$\mathbf{v}_t^* = \alpha_t \cdot \mathbf{x}_0 + (\beta_t - \alpha_t) \cdot \left(\widehat{\mathbf{w}}_1^\top \mathbf{x}_0\right) \cdot \mathbf{w}_1, \quad \text{for } t = 1, \ldots, T, \tag{25}$$

where $\mathbf{w}_1 \in \mathbb{R}^N$ is the weight vector of the index fund (e.g., the market portfolio), expressed in number of shares, and $\widehat{\mathbf{w}}_1 \triangleq \left(\bar{\psi}_{\mathrm{f},1}^{-1} + \mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1} \mathbf{w}_1\right)^{-1} \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1} \mathbf{w}_1$. By contrast, the separable execution $\mathbf{v}_t^{\mathrm{sep}}$ liquidates each order in the portfolio independently, allocating quantities to be traded in each period in a way that is proportional to the total traded volume that is forecasted to be executed in that period:

$$v_{it}^{\mathrm{sep}} = \mathrm{VolAlloc}_{it} \cdot x_{0i}, \quad \text{for } t = 1, \ldots, T, \quad \text{for each } i = 1, \ldots, N, \tag{26}$$

where $\mathrm{VolAlloc}_{it}$ is the percentage of the daily volume in security i that trades in period t, defined in (1).

§5.1 (and, in more detail, Appendix §D.3) posits a stylized stochastic-process generative model for single-stock and index-fund (portfolio) investor order flow that results in a simple parametric structure in the total traded volume profile $\mathrm{VolAlloc}_{it}$ and the resulting pairwise correlation profile (among traded volumes) Correl_{ijt}. The model's primitive parameters can be estimated so as to be consistent with $\mathrm{AvgVolAlloc}_t$ and $\mathrm{AvgCorrel}_t$, discussed in §2. §5.2 provides analytic results on the optimality gap between the separable and the optimal execution schedules, in (26) and (25), respectively, which for the parameters estimated in §5.1 can be as high as 15%.

5.1. A Useful Parameterization of Intraday Liquidity

We will posit a simple generative model of single-stock and index-fund (portfolio) order flow (driven by two underlying Poisson processes). This mixture of order flows comprises the total volume for the day, and also generates a certain correlation structure in the traded volumes per period across securities. (We will offer a brief overview in this section, and defer to Appendix §D.3 for additional details on this model.) Let θ_i denote the fraction of traded volume in a day for stock i that is generated by the order flow submitted by index-fund investors. Formally,

$$\theta_i \triangleq \frac{|\tilde{w}_{1i}| \cdot \bar{q}_{\mathrm{f}}}{\bar{q}_{\mathrm{id},i} + |\tilde{w}_{1i}| \cdot \bar{q}_{\mathrm{f}}}, \tag{27}$$

where $\bar{q}_{\mathrm{f}}$ is the notional traded by index-fund investors, $\tilde{w}_{1i}$ is the weight of security i in this index fund (notionally weighted), and $\bar{q}_{\mathrm{id},i}$ is the notional traded by single-stock investors in security i. For simplicity, further assume that $\theta_1 = \theta_2 = \ldots = \theta_N = \theta$; i.e., all securities have the same composition of order flow as contributed by single-stock and index-fund investors.

In such a model (as explained in §D.3), the intraday volume and pairwise correlation profiles are given by

$$\mathrm{AvgVolAlloc}_t = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathsf{E}\left[\mathrm{DVol}_{idt}\right]}{\sum_{s=1}^{T}\mathsf{E}\left[\mathrm{DVol}_{ids}\right]} = \alpha_t \cdot (1-\theta) + \beta_t \cdot \theta, \tag{28}$$

$$\mathrm{AvgCorrel}_t = \frac{1}{N(N-1)}\sum_{i\neq j}\mathrm{Correl}_{ijt} = \frac{\beta_t \cdot \theta^2}{\alpha_t \cdot (1-\theta)^2 + \beta_t \cdot \theta^2}. \tag{29}$$

We note that the assumption that $\theta_i = \theta$ for all securities i leads to the conclusion that all securities have the same intraday volume profile, and, perhaps more importantly, that the intraday volume correlation profile Correl_{ijt} is the same across all pairs of stocks. The latter is arguably a fairly strong restriction, and it is only imposed so as to allow for a more tractable closed-form performance analysis.

Given the empirically observed profiles for $\mathrm{AvgVolAlloc}_t$ and $\mathrm{AvgCorrel}_t$, illustrated in Figure 1, we can solve a set of coupled equations defined by (28)–(29), where the respective left hand sides are given by the empirically estimated values, so as to identify the values of θ, $\alpha_1, \ldots, \alpha_T$ and $\beta_1, \ldots, \beta_T$. The results are summarized in Figures 2 and 3. We can observe that θ is estimated to be .24, implying that 24% of total traded volume originates from the index fund. We also observe that, at the beginning of the day, the trading activity of index-fund investors β_t is smaller than that of single-stock investors α_t, but β_t far exceeds α_t in the last hour of the day, as expected. Such an intraday variation in the composition of order flow is consistent with the increasing pairwise correlation in volumes towards the end of the trading day.

Finally, Figure 4 provides a graphical illustration of the optimal execution schedule in (25) with respect to these estimated parameters. The example depicts an investor that wants to liquidate a portfolio $\mathbf{x}_0$ with two orders, where the weights of the liquidation portfolio deviate significantly from the weights of the index portfolio $\mathbf{w}_1$ (as captured by the angle between $\mathbf{x}_0$ and $\mathbf{w}_1$). To exploit the increased end-of-day liquidity along the direction of the index portfolio, $\mathbf{w}_1$, the optimal schedule trades stock 2 more aggressively in the morning session, as shown by $\mathbf{v}_t^*$, thus tilting away from a separable VWAP-like execution that would be aligned with $\mathbf{x}_0$. As a consequence, the residual portfolio executed towards the end of the day is better aligned with the index portfolio (in the

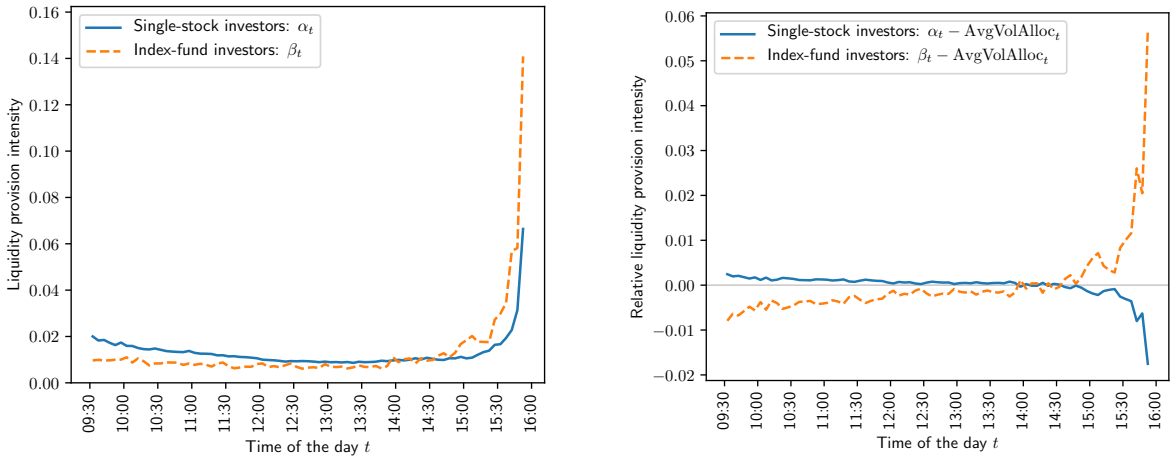

Figure 2: Intensity of single-stock investors, α_t, and index-fund (portfolio) investors, β_t, calibrated to best match empirical profiles of traded volume, $\mathrm{AvgVolAlloc}_t$, and pairwise volume correlations, $\mathrm{AvgCorrel}_t$ (left) and deviation of α_t, β_t from the market profile, $\mathrm{AvgVolAlloc}_t$ (right). The data of the year 2018 is used as in §2.

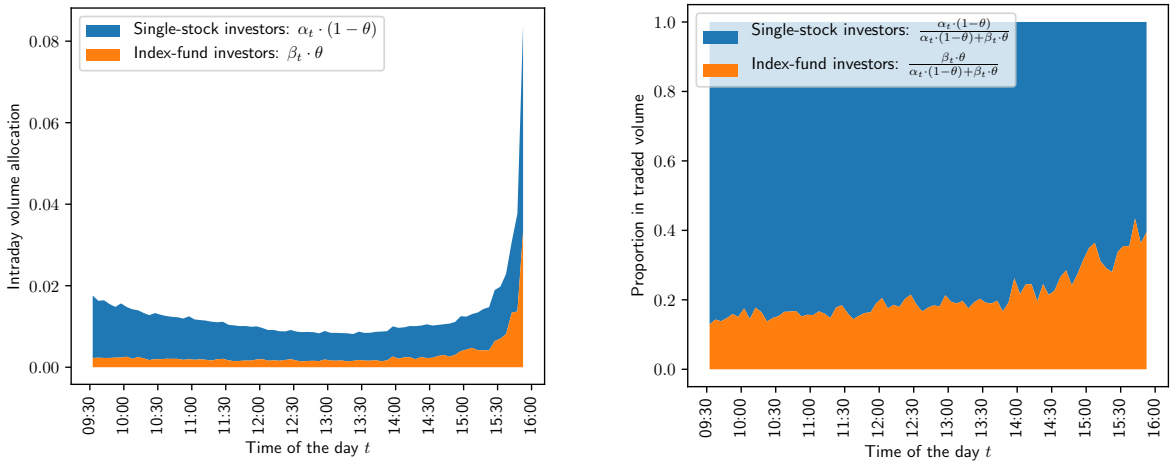

Figure 3: Decomposition of average volume allocation, $\mathrm{AvgVolAlloc}_t$, into single-stock investors' contribution, $\alpha_t \cdot (1-\theta)$, and index-fund investors' contribution, $\beta_t \cdot \theta$ (left) and their proportions (right). In this generative model, every security is assumed to have the same intraday profiles.

afternoon, $\mathbf{v}_t^*$ is closer to the index portfolio $\mathbf{w}_1$).

5.2. Implementation Shortfall Comparison: Optimal vs. Separable Execution Schedules

From (26) and (28) we get that the separable schedule $\mathbf{v}_t^{\mathrm{sep}}$ is given by

$$\mathbf{v}_t^{\mathrm{sep}} = (\alpha_t \cdot (1-\theta) + \beta_t \cdot \theta) \cdot \mathbf{x}_0, \tag{30}$$

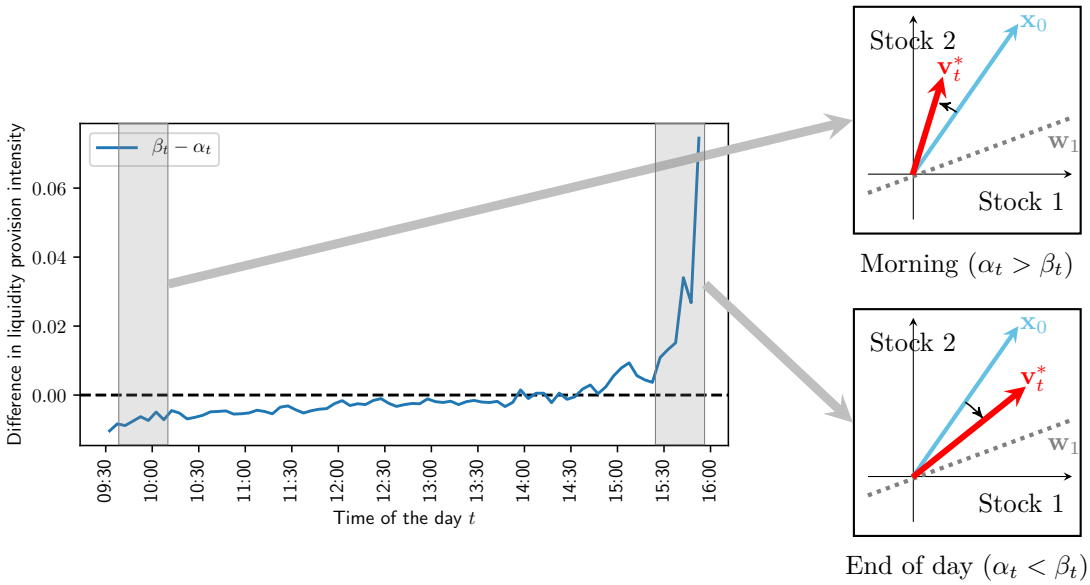

Figure 4: Illustration of the optimized schedule $\mathbf{v}_t^*$, which is shown to tilt away from or toward the direction of index fund $\mathbf{w}_1$ depending on the difference between single-stock and index-fund liquidity, $\beta_t - \alpha_t$.

and for $\mathbf{v}_t^*$ and $\mathbf{v}_t^{\text{sep}}$, the expected implementation shortfall can be written as follows:

$$\bar{\mathcal{C}}(\mathbf{v}_t^*) = \frac{1}{2}\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\text{id}} + \mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0, \tag{31}$$

$$\bar{\mathcal{C}}(\mathbf{v}_t^{\text{sep}}) = \frac{1}{2}\sum_{t=1}^{T}(\alpha_t\cdot(1-\theta)+\beta_t\cdot\theta)^2\cdot\mathbf{x}_0^\top\left(\alpha_t\bar{\mathbf{\Psi}}_{\text{id}} + \beta_t\mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0. \tag{32}$$

Note that under the assumption that there is only one index fund, $\mathbf{w}_1$, we can simplify the above expressions and reduce $\mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\mathbf{W}^\top$ to $\mathbf{w}_1\bar{\psi}_{\text{f},1}\mathbf{w}_1^\top$. The expression for $\bar{\mathcal{C}}(\mathbf{v}_t^{\text{sep}})$ is obtained by substituting (30) into (16). We define as a relative performance measure the ratio between the expected transaction costs incurred by the two execution schedules:

$$\Upsilon(\mathbf{x}_0) \triangleq \frac{\bar{\mathcal{C}}(\mathbf{v}_t^{\text{sep}})}{\bar{\mathcal{C}}(\mathbf{v}_t^*)}; \tag{33}$$

this ratio is clearly greater than or equal to 1, and captures the additional cost incurred by the separable VWAP-like schedule over the optimal (coupled) execution schedule.

Proposition 5 (Exact cost ratio). *For any* $\mathbf{x}_0 \in \mathbb{R}^N$,

$$\Upsilon(\mathbf{x}_0) = 1 + \theta^2\cdot\left(\sum_{t=1}^{T}\frac{\beta_t^2}{\alpha_t} - 1\right) + \Delta\cdot\left(\frac{\mathbf{x}_0^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{x}_0}{\left(\mathbf{w}_1^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{x}_0\right)^2}\cdot\frac{1+\eta_1}{\bar{\psi}_{\text{f},1}} - 1\right)^{-1}, \tag{34}$$

where

$$\gamma_t \triangleq \frac{\beta_t}{\alpha_t}, \quad \eta_1 \triangleq \bar{\psi}_{\text{f},1}\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{w}_1, \quad \textit{and} \quad \Delta \triangleq \sum_{t=1}^{T} \frac{\alpha_t \cdot (1-\theta\cdot(1-\gamma_t))^2\,(1-\gamma_t)}{1+\eta_1\cdot\gamma_t}. \tag{35}$$

(The proof is given in Appendix D.4.1.) The parameter η_1 is the ratio between index-fund liquidity ($1/\bar{\psi}_{\text{f},1}$) and single-stock liquidity along the index-fund weights ($\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{w}_1$). Equivalently, it is the ratio between the price change of trading along the index-fund direction $\mathbf{w}_1$ against only the fraction of single-stock investors in the market,[7] and the price change of trading along $\mathbf{w}_1$ against only the fraction of index-fund investors in the market,[8] which is $1/\bar{\psi}_{\text{f},1}$.

The last expression in the performance metric is a product two terms: the first is associated with the intraday variation of liquidity and trading volume (Δ), and the second is associated with the degree of alignment between the execution portfolio $\mathbf{x}_0$ and the index fund weights $\mathbf{w}_1$.

Worst case liquidation portfolios. First, we explore the structure of the portfolios that would exhibit the largest optimality gap under a separable execution.

Remark 1 (Maximum/minimum cost ratio). *Let* Υ_{market} *and* Υ_{orth} *be the cost ratio when* $\mathbf{x}_0 = \mathbf{w}_1$ *and* $\mathbf{x}_0 = \mathbf{w}_1^\perp$*, respectively, where* $\mathbf{w}_1^\perp$ *is an arbitrary portfolio such that* $\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{w}_1^\perp = 0$ *with* $\mathbf{w}_1^\top \neq \mathbf{0}$*:*

$$\Upsilon_{\text{market}} \triangleq \Upsilon(\mathbf{x}_0 = \mathbf{w}_1) = 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1\right) + \eta_1 \cdot \Delta, \tag{36}$$

$$\Upsilon_{\text{orth}} \triangleq \Upsilon(\mathbf{x}_0 = \mathbf{w}_1^\perp) = 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1\right). \tag{37}$$

Then, the largest and smallest cost ratios are obtained at either $\mathbf{x}_0 = \mathbf{w}_1$ *or* $\mathbf{x}_0 = \mathbf{w}_1^\perp$ *depending on the sign of* Δ*:*

$$\max_{\mathbf{x}_0\in\mathbb{R}^N}\{\Upsilon(\mathbf{x}_0)\} = \begin{cases} \Upsilon_{\text{market}} & \textit{if } \Delta \geq 0 \\ \Upsilon_{\text{orth}} & \textit{if } \Delta \leq 0 \end{cases}, \quad \min_{\mathbf{x}_0\in\mathbb{R}^N}\{\Upsilon(\mathbf{x}_0)\} = \begin{cases} \Upsilon_{\text{orth}} & \textit{if } \Delta \geq 0 \\ \Upsilon_{\text{market}} & \textit{if } \Delta \leq 0 \end{cases}. \tag{38}$$

[7] If we trade $\mathbf{w}_1$ against single-stock investors we cause a change in prices given by $\Delta\mathbf{p} = \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{w}_1$, which implies a change in the price of the market portfolio equal to $\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{w}_1$.

[8] To gain some intuition of the magnitude of that parameter, imagine wanting to buy a \$100 million slice of the S&P 500, where in one case it is acquired from distinct liquidity providers, each trading only one of the constituent orders, while in the other case it is acquired from the same (portfolio) liquidity provider. The mere difference in the aggregate volatility held by the distinct liquidity providers in the first scenario versus the unique market portfolio liquidity provider in the second scenario would suggest a potentially significant difference in trading costs, and therefore a high ($\gg 1$) value for η_1.

In particular, for fixed $(\eta_1, \alpha_1, \ldots, \alpha_T, \beta_1, \ldots, \beta_T)$*, there exists* $\theta^* \in [0,1]$ *such that*

$$\Delta \geq 0 \quad \textit{if } \theta \leq \theta^* \quad \textit{and} \quad \Delta \leq 0 \quad \textit{if } \theta \geq \theta^*. \tag{39}$$

Similarly, for fixed $(\theta, \alpha_1, \ldots, \alpha_T, \beta_1, \ldots, \beta_T)$*, there exists* $\eta_1^* \in [\frac{\theta}{1-\theta}, \infty]$ *such that*

$$\Delta \leq 0 \quad \textit{if } \eta_1 \leq \eta_1^* \quad \textit{and} \quad \Delta \geq 0 \quad \textit{if } \eta_1 \geq \eta_1^*. \tag{40}$$

This remark identifies which portfolios give rise to the largest and smallest cost ratios, respectively. It is straightforward that the cost ratio has extreme values at $\mathbf{x}_0 = \mathbf{w}_1$ and $\mathbf{x}_0 = \mathbf{w}_1^\perp$, i.e., when $\mathbf{x}_0$ is most and least aligned with the market portfolio $\mathbf{w}_1$. From (30) and (23) we get that in these two extreme cases the separable and optimal schedules are given by

$$\mathbf{v}_t^{\text{sep}} = (\alpha_t \cdot (1-\theta) + \beta_t \cdot \theta) \cdot \mathbf{x}_0, \quad \text{and} \quad \mathbf{v}_t^* = \begin{cases} \left(\alpha_t \cdot \left(1 - \frac{\eta_1}{1+\eta_1}\right) + \beta_t \cdot \frac{\eta_1}{1+\eta_1}\right) \cdot \mathbf{x}_0 & \text{if } \mathbf{x}_0 = \mathbf{w}_1, \\ \alpha_t \cdot \mathbf{x}_0 & \text{if } \mathbf{x}_0 = \mathbf{w}_1^\perp. \end{cases}$$

When $\mathbf{x}_0 = \mathbf{w}_1$, the sensitivity of the optimized execution schedule to the intensity of index-fund liquidity provision, β_t, is $\frac{\eta_1}{1+\eta_1}$, whereas the sensitivity of the separable execution is θ. If $\theta > \frac{\eta_1}{1+\eta_1}$, the separable execution schedule will trade above the optimal level in the morning, and trade below the optimal level towards the end of the day; the opposite happens if $\theta < \frac{\eta_1}{1+\eta_1}$. We can expect that the suboptimality of separable execution roughly scales with $\left(\theta - \frac{\eta_1}{1+\eta_1}\right)^2$. A similar argument suggests that when $\mathbf{x}_0 = \mathbf{w}_1^\perp$, the suboptimality of separable execution roughly scales with $(\theta - 0)^2$. Comparing $\left(\theta - \frac{\eta_1}{1+\eta_1}\right)^2$ and θ^2 as proxies for Υ_{market} and Υ_{orth}, respectively, the findings of Remark 1 follow.

Performance implications when trading the market portfolio. Next we characterize Υ_{market} as a function of the parameter η_1.

Remark 2 (Characterization of Υ_{market})**.** *For fixed* $(\theta, \alpha_1, \ldots, \alpha_T, \beta_1, \ldots, \beta_T)$*, as a function of* η_1*,*

$$\Upsilon_{\text{market}}(\eta_1) \textit{ decreases if } \eta_1 \leq \frac{\theta}{1-\theta}, \quad \textit{and} \quad \Upsilon_{\text{market}}(\eta_1) \textit{ increases if } \eta_1 \geq \frac{\theta}{1-\theta}. \tag{41}$$

For particular values of η_1,

$$\Upsilon_{\text{market}}(\eta_1 = 0) = 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1 \right), \tag{42}$$

$$\Upsilon_{\text{market}}\left(\eta_1 = \frac{\theta}{1-\theta}\right) = 1, \tag{43}$$

$$\lim_{\eta_1 \to \infty} \Upsilon_{\text{market}}(\eta_1) = 1 + (1-\theta)^2 \cdot \left(\sum_{t=1}^{T} \frac{\alpha_t^2}{\beta_t} - 1 \right). \tag{44}$$

This implies that Υ_{market} first decreases and then increases as η_1 varies. This can be similarly understood as Remark 1: separable execution correctly reacts to the liquidity provided by index-fund investors only when $\eta_1 = \frac{\theta}{1-\theta}$, and overreacts or underreacts when η_1 deviates from $\frac{\theta}{1-\theta}$.

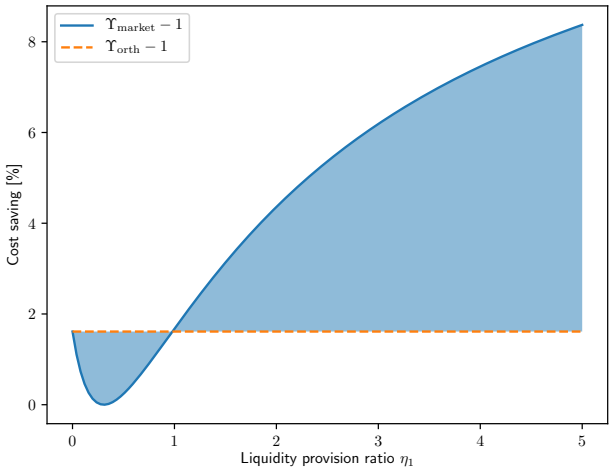

Figure 5: Possible range of cost ratio Υ with respect to η_1 given the values of $\theta, \alpha_1, \ldots, \alpha_T, \beta_1, \ldots, \beta_T$ obtained in §5.1. Compared to the separable execution, the coupled execution can save up to 14.0 % when trading the market portfolio.

Our estimate of the fraction of index-fund liquidity, $\theta = .24$, suggests a threshold value of $\theta/(1-\theta) \approx .31$. Even though the value of η_1 is unidentifiable in our context, one would expect the value of η_1 to be moderately large (see Footnote 8), and that the realized benefits from using optimal vs. separable execution schedule to approach the upper bound in (44). That upper bound is equal to 14.0% for the parameters $\theta, \alpha_1, \ldots, \alpha_T, \beta_1, \ldots, \beta_T$ estimated in §5.1, and Figure 5 graphs Υ_{market} and Υ_{orth} as functions of η_1. That is, under the assumptions of our stylized generative model of order flow, one can reduce execution costs by as much as 14.0% by optimally coupling the execution schedules of the various orders that are being liquidated so as to exploit the effects of cross-impact induced due to portfolio liquidity provision.

Liquidating single orders. Finally, we apply our results to the special case where the target

portfolio to be liquidated is an order on a single security.

Remark 3 (Individual orders). *When trading a single stock, the cost ratio is given by*

$$\Upsilon(\mathbf{x}_0 = \mathbf{e}_i) = 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1 \right) + \frac{\eta_{1,i}}{1 + \eta_1 - \eta_{1,i}} \cdot \Delta, \tag{45}$$

where $\eta_{1,i} \triangleq \frac{w_{1i}^2 \cdot \bar{\psi}_{\mathrm{f},1}}{\psi_{\mathrm{id},i}}$. *We can further identify the stock that induces the largest cost ratio:*

$$\underset{i=1,\ldots,N}{\operatorname{argmax}} \{\Upsilon(\mathbf{x}_0 = \mathbf{e}_i)\} = \begin{cases} \operatorname{argmax}_{i=1,\ldots,N} \left\{ \frac{w_{1i}^2}{\bar{\psi}_{\mathrm{id},i}} \right\} & \text{if } \Delta \geq 0 \\ \operatorname{argmin}_{i=1,\ldots,N} \left\{ \frac{w_{1i}^2}{\bar{\psi}_{\mathrm{id},i}} \right\} & \text{if } \Delta \leq 0 \end{cases}. \tag{46}$$

Here we are comparing the performance implications of liquidating a single order using a separable execution schedule vs. the optimal execution schedule that may add positions early in the day, so as to unwind the residual portfolio later in the day in a way that benefits from the liquidity provided by index-fund investors. The fraction $w_{1i}^2/\bar{\psi}_{\mathrm{id},i}$ determines which security is most costly to trade, and depends both on the market weight of the security in the index-fund portfolio, and the liquidity provided by its own single-stock investors. Assuming that, for our estimated value for θ, η_1 is sufficiently large ($\Delta \geq 0$), equation (46) suggests that the optimized execution schedule may be most beneficial when trading in securities with large market weights.

6. Extensions

Estimation of cross-asset market impact. Estimating a cross-security impact model that explicitly measures the impact coefficient between any pair of securities i, j is hard due to the high dimensionality of the unknown coefficient matrix (an $N \times N$ matrix), and because the underlying data tends to be very noisy. We propose an efficient procedure to estimate an impact model by exploiting the low-rank structure of the cost model postulated in §3, which would take as input a large set of proprietary portfolio transactions.

The derivation in §3 predicts a linear relationship between the portfolio transactions $\tilde{\mathbf{v}}_{dt} \in \mathbb{R}^N$ (measured in dollar amount) and the realized implementation shortfalls $\tilde{\mathbf{r}}_{dt}^{\mathrm{tr}} \in \mathbb{R}^N$ (measured in return) of the form

$$\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} = \frac{1}{2} \left(\tilde{\boldsymbol{\Psi}}_{\mathrm{id},dt} + \tilde{\mathbf{W}}_d \tilde{\boldsymbol{\Psi}}_{\mathrm{f},dt} \tilde{\mathbf{W}}_d^\top \right)^{-1} \tilde{\mathbf{v}}_{dt} + \bar{\boldsymbol{\epsilon}}_{dt}^{\mathrm{tr}}, \tag{47}$$

where a diagonal matrix $\tilde{\boldsymbol{\Psi}}_{\mathrm{id},dt} \in \mathbb{R}^{N \times N}$ describes the liquidity provided by single-stock investors, a diagonal matrix $\tilde{\boldsymbol{\Psi}}_{\mathrm{f},dt} \in \mathbb{R}^{K \times K}$ describes the liquidity provided by index-fund investors, and the

noise term $\bar{\boldsymbol{\epsilon}}^{\mathrm{tr}}_{dt} \in \mathbb{R}^N$ describes the random fluctuation of price. As analogous to a common practice to estimate the price impact for individual stocks,[9] we further parameterize the diagonal entries of the liquidity matrices as follows:

$$\tilde{\psi}_{\mathrm{id},idt} = \gamma_{\mathrm{id}} \times \frac{\widehat{\mathrm{DVol}}_{\mathrm{id},idt}}{\widehat{\sigma}_{\mathrm{id},idt}}, \quad \tilde{\psi}_{\mathrm{f},kdt} = \gamma_{\mathrm{f},k} \times \frac{\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}}{\widehat{\sigma}_{\mathrm{f},kdt}}, \tag{48}$$

where $\widehat{\mathrm{DVol}}_{\mathrm{id},idt}$ and $\widehat{\sigma}_{\mathrm{id},idt}$ (resp., $\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}$ and $\widehat{\sigma}_{\mathrm{f},kdt}$) are some forecasted trading volume and volatility of the stock i (resp., the index fund k) generated by the single-stock investors (resp., the index-fund investors), and γ_{id} and $\gamma_{\mathrm{f},k}$'s are unknown parameters that describe time-invariant characteristics of liquidity providers. As a result, we can substantially reduce the number of unknowns to estimate: assuming that such forecasts are available, it suffices to estimate $K+1$ parameters.

In Appendix C, we motivate the above parameterization in detail, propose an effective regression scheme including the case where the forecasts are not available, and also verify the procedure based on a carefully synthesized dataset.

Trading constraints. Trade execution algorithms used to liquidate portfolios may impose additional constraints, starting with side constraints that force the liquidation schedule to only trade in the securities that are included in the target liquidation portfolio, and to only trade in the direction of the parent orders themselves – i.e., only sell stock in securities that were submitted as "sell" orders, and vice versa for "buy" orders.

§4–§5 do not impose these side trading constraints, and the derived optimal schedules may violate these restrictions, e.g., by choosing to trade in securities that are not included in the target liquidation portfolio, $\mathbf{x}_0$, so that the residual liquidation portfolio can take advantage of (cheaper) natural portfolio liquidity towards the end of the day.[10] Similarly, the optimal schedule may choose to increase the size of an existing order (as opposed to start liquidating it) early in the day, if that would be beneficial when liquidating the residual portfolio towards the end of the day.

The constrained portfolio liquidation problem is similar in nature to the one studied in the previous section, and the (numerically) optimized schedule will continue to incorporate and exploit the effects of cross-impact and natural portfolio liquidity provision. One exception is when a single parent order is liquidated, in which case these cross-impact and portfolio liquidity factors are not

[9] The parameterization (48) is consistent with most of the literature in estimating market-impact models; e.g., assuming that there are only single stock natural liquidity providers, one would recover a commonly encountered cost model of the form $\gamma_{\mathrm{id}}^{-1} \times \widehat{\sigma}_{idt}(\tilde{v}_{idt}/\widehat{\mathrm{DVol}}_{idt})$; see Almgren et al. (2005).

[10] In a market where all liquidity is provided by single-stock investors, the optimal schedule would never choose to trade outside the universe of securities that are in the liquidation portfolio or to trade against the direction of the respective parent orders.

relevant to such a constrained formulation.[11]

Mean-variance optimization. This paper primarily focuses on the risk-neutral liquidation problem for which we characterize the optimal schedule that minimizes the expected execution cost. One possible extension would be to incorporate the variance of execution cost into the objective, as done in Almgren and Chriss (2000, Appendix A), so as to formulate a risk-averse liquidation problem into a mean-variance optimization, and the optimal schedule can be readily found via a quadratic programming (with $N \times T$ decision variables). We anticipate that it will be an interesting research topic to characterize the optimal schedule that exploits the cross-sectional properties in liquidity provision and random fluctuation of the prices.

References

V. Agarwal, P. Hanouna, R. Moussawi, and C. W. Stahel. Do ETFs increase the commonality in liquidity of underlying stocks? *28th Annual Conference on Financial Economics and Accounting*, 2018.

A. Alfonsi, A. Fruth, and A. Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10:143–157, 2010.

R. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10:1–18, 2003.

R. Almgren and N. Chriss. Optimal control of portfolio transactions. *Journal of Risk*, 3:5–39, 2000.

R. Almgren, C. Thum, E. Hauptmann, and H. Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005.

I. Ben-David, F. A. Franzoni, and R. Moussawi. Exchange-traded funds. *Annual Review of Financial Economics*, 9:169–189, 2017.

M. Benzaquen, I. Mastromatteo, Z. Eisler, and J.-P. Bouchaud. Dissecting cross-impact on stock markets: An empirical analysis. Working paper, 2016.

D. Bertsimas and A. W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1: 1–50, 1998.

J.-P. Bouchaud, J. D. Farmer, and F. Lillo. How markets slowly digest changes in supply and demand. In *Handbook of Financial Markets: Dynamics and Evolution*, pages 57–156. Elsevier: Academic Press, 2008.

D. B. Brown, B. Carlin, and M. S. Lobo. Optimal portfolio liquidation with distress risk. *Management Science*, 56(11):1997–2014, 2010.

F. Bucci, I. Mastromatt, M. Benzaquen, and J.-P. Bouchaud. Impact is not just volatility. 2019.

F. Capponi and R. Cont. Trade duration, volatility and market impact. *Working paper*, 2019.

G. Curato, J. Gatheral, and F. Lillo. Optimal execution with nonlinear transient market impact. 2014.

[11] In practice, the liquidation problem may impose additional trading constraints, e.g., upper bounds on the speed of execution, (linear) exposure constraints, etc. In addition, the liquidator may either incorporate a risk term into her objective function, or add a risk budget constraint. The resulting problem continues to be a convex quadratic program with similar structural properties.

J. Donier, J. Bonart, I. Mastromatteo, and J.-P. Bouchaud. A fully consistent, minimal model for non-linear market impact. *Quantitative Finance 15(7)*, 2015.

C. Driebusch, A. Osipovich, and G. Zuckerman. What's the biggest trade on the New York stock exchange? The last one, March 2018. URL `https://www.wsj.com/articles/at-closing-time-the-stock-market-heats-up-like-a-bar-at-last-call-1521038300`.

J. Gatheral, A. Schied, and A. Slynko. Transient linear price impact and fredholm integral equations. *Mathematical Finance*, 2012.

J. Hasbrouck and D. J. Seppi. Common factors in prices, order flows, and liquidity. *Journal of Financial Economics*, 2001.

M. Haugh and C. Wang. Dynamic portfolio execution and information relaxations. *SIAM Journal of Financial Mathematics*, 5(1):316–359, 2014.

G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 74(4):1247–1276, 2004.

G. Huberman and W. Stanzl. Optimal liquidity trading. *Review of Finance*, 9:165–200, 2005.

L. S. Junior and I. P. Franca. Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 391(1-2), 2011.

G. A. Karoli, K.-H. Lee, and M. A. van Dijk. Understanding commonality in liquidity around the world. *Journal of Financial Economics*, 105(1):82–112, 2012.

A. Koch, S. Ruenzi, and L. Starks. Commonality in liquidity: A demand-side explanation. *The Review of Financial Studies*, 29(8):1943–1974, 2016.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

A. W. Lo and J. Wang. Trading volume: Definitions, data analysis, and implications of portfolio theory. *The Review of Financial Studies*, 2000, 2009.

I. Mastromatteo, M. Benzaquen, Z. Eisler, and J.-P. Bouchaud. Trading lightly: Cross-impact and optimal portfolio execution. Working paper, 2017.

A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013.

I. Rosu. A dynamic model of the limit order book. *Review of Financial Studies*, 22:4601–4641, 2009.

M. Schneider and F. Lillo. Cross-impact and no-dynamic-arbitrage. *Quantitative Finance*, 19(1): 137–154, 2019.

G. E. Tauchen and M. Pitts. The price variability-volume relationship on speculative markets. *Econometrica*, 51(2):485–505, 1983.

M. Tomas, I. Mastromatteo, and M. Benzaquen. How to build a cross-impact model from first principles: Theoretical requirements and empirical results. 2020.

B. Tóth, Y. Lempérière, C. Deremble, J. De Lataillade, J. Kockelkoren, and J.-P. Bouchaud. Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006, 2011.

B. Tóth, E. Zoltán, and J.-P. Bouchaud. The short-term price impact of trades is universal. 2018.

G. Tsoukalas, J. Wang, and K. Giesecke. Dynamic portfolio execution. *Management Science*, 2017.

A. Additional Empirical Analysis

We provide the empirical analysis illustrated in §2 for the longer timespan from 2007 to 2018. The same procedure was conducted for each year and for the stocks that had been the constituents of S&P 500 throughout that year. Figure 6 shows the intraday pattern of correlation in liquidity for each year with various ways of visualization. First of all, we observe consistently across all years that the correlation increases over the course of the day. While the overall level of correlation fluctuates over years[12] (bottom left figure), we observe that the end-of-day increase had become significant (bottom right figure), which can be attributable to the increasing popularity of index-fund investing.

Figure 7 shows that the intraday pattern exists among both large-cap stocks and small-cap stocks, and we can observe that the large-cap stocks are more correlated than the small-cap stocks.

Recall that in §5.2 we have argued that the benefit from incorporate cross-asset impact into execution scheduling depends not only on the relative magnitude of index-fund liquidity provision versus single-stock liquidity provision (that is captured by the proportion of index-fund liquidity θ), but also on the intraday variation of their composition (that is captured by changes in the ratio $\frac{\alpha_t}{\beta_t}$). Table 1 shows that the maximum cost savings according to the analysis of §5 have been increasing in recent years: although the overall proportion of index-fund liquidity was relatively higher during the financial crisis period '07–'11, the composition of liquidity was stable throughout the day in this period, and as a result a separable VWAP execution would work fine compared to the optimized coupled execution.

Year	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18
Prop. of fund liquidity (%)	26.5	29.7	24.3	26.7	27.4	20.3	19.1	21.7	20.9	22.9	17.9	23.5
Maximum cost saving (%)	8.3	3.1	7.8	7.2	6.7	11.2	16.4	8.7	12.0	6.4	7.9	14.0

Table 1: The illustrative statistics introduced in §5: the proportion of index-fund liquidity, θ, and the maximum cost saving, $\Upsilon_{\max} - 1$, estimated on years 2007–2018.

B. Change of Units

In §3 and §4, the price impact was the equilibrium expected price change $\Delta\mathbf{p}$, expressed in dollars, required for the market to clear when executing a vector $\mathbf{v}$, expressed in number of shares for each security in the executed portfolio. We can restate the market impact in terms of the return $\mathbf{r} \in \mathbb{R}^N$

[12] We observe that the correlation is relatively higher during the period 2007–2011 that coincides with the time of financial crisis. This will be consistent with a common belief that high volatility of markets is directly linked with strong correlations between stocks (Junior and Franca, 2011).

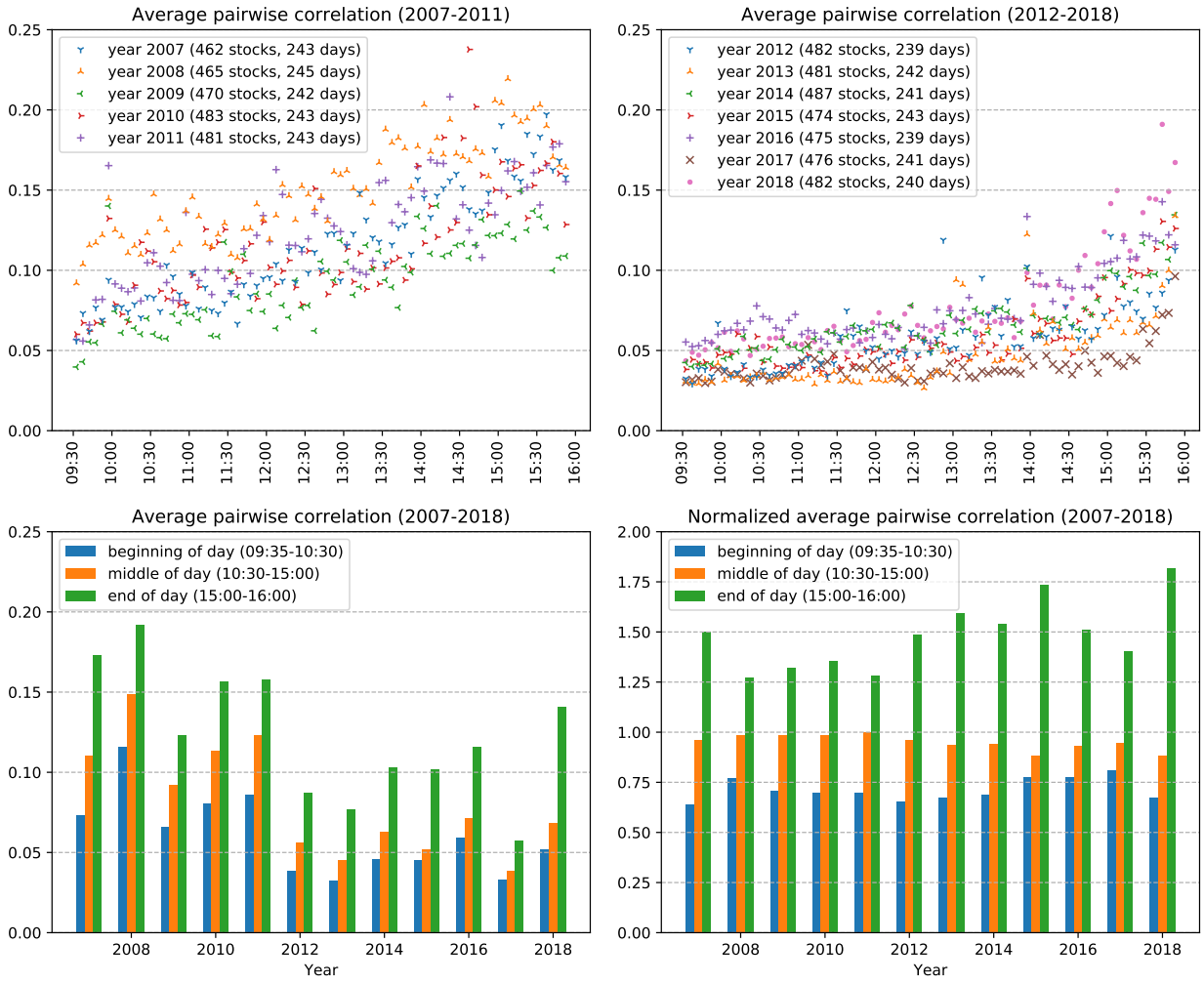

Figure 6: Intraday variations of average pairwise correlation in liquidity for years 2007–2011 (top left) and for years 2012–2018 (top right), and their alternative visualization with ones averaged for each part of day (bottom left) and ones further normalized by the average correlation level over the entire day (bottom right).

as a function of the vector of notional execution quantities $\tilde{\mathbf{v}} \in \mathbb{R}^N$. We let p denote the (arrival) equilibrium price vector $\mathbf{p} \in \mathbb{R}^N$, snapped at the beginning of the execution period, and define the diagonal matrix $\mathbf{P} \triangleq \mathrm{diag}(\mathbf{p}) \in \mathbb{R}^{N \times N}$. Then,

$$\mathbf{r} \triangleq \mathbf{P}^{-1}\Delta\mathbf{p} \quad \text{and} \quad \tilde{\mathbf{v}} \triangleq \mathbf{P}\mathbf{v}. \tag{49}$$

We redefine the liquidity variable $\psi_{\mathrm{id},i}$, $\psi_{\mathrm{f},k}$ and weight vectors $\mathbf{w}_k$ accordingly:

$$\tilde{\psi}_{\mathrm{id},i} \triangleq p_i^2 \cdot \psi_{\mathrm{id},i}, \quad \tilde{\psi}_{\mathrm{f},k} \triangleq (\mathbf{w}_k^\top \mathbf{p})^2 \cdot \psi_{\mathrm{f},k} \quad \text{and} \quad \tilde{\mathbf{w}}_k \triangleq \frac{\mathbf{P}\mathbf{w}_k}{\mathbf{p}^\top \mathbf{w}_k}. \tag{50}$$

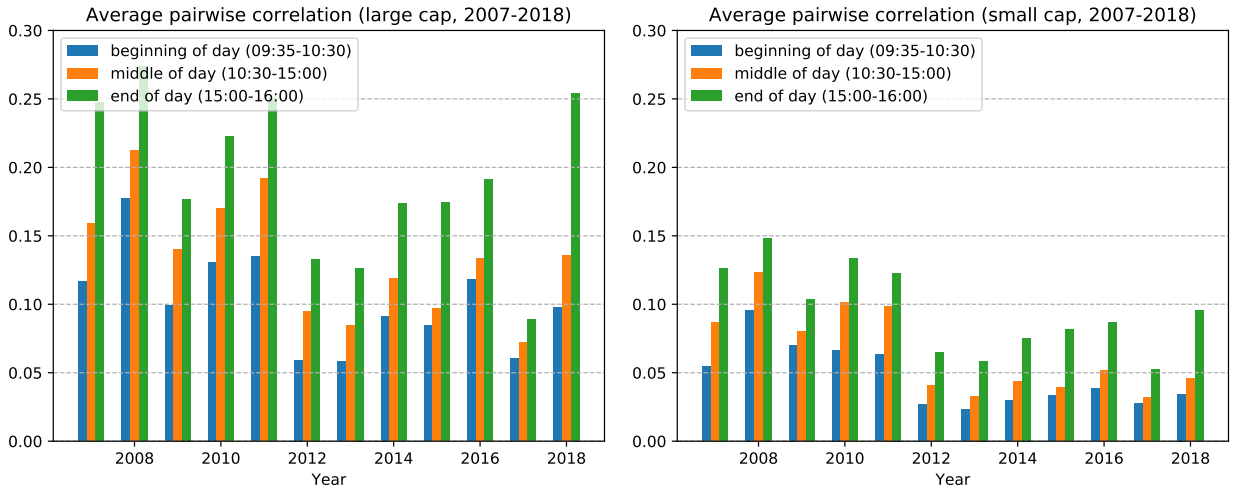

Figure 7: The average pairwise correlation for large-cap stocks (top-100 stocks in S&P 500, left) and for small-cap stocks (bottom-100 stocks in S&P 500, right).

The redefined liquidity variable $\psi_{\mathrm{id},i}$ now has the following interpretation: single-stock investors will sell (or buy) $1\% \cdot \tilde{\psi}_{\mathrm{id},it}$ dollar amount of stock i, when its price rises (or drops) by one percent. The rescaled weight vector $\tilde{\mathbf{w}}_k$ represents the normalized dollar-weighted portfolio. Putting it all together, we get

$$\begin{aligned} \mathbf{r} &= \mathbf{P}^{-1}\mathbf{G}\mathbf{v} = \mathbf{P}^{-1}\left(\mathbf{\Psi}_{\mathrm{id}} + \mathbf{W}\mathbf{\Psi}_{\mathrm{f}}\mathbf{W}^{\top}\right)^{-1}\mathbf{P}^{-1}\cdot\mathbf{P}\mathbf{v} \\ &= \left(\mathbf{P}\mathbf{\Psi}_{\mathrm{id}}\mathbf{P} + \mathbf{P}\mathbf{W}\mathbf{\Psi}_{\mathrm{f}}\mathbf{W}^{\top}\mathbf{P}\right)^{-1}\tilde{\mathbf{v}} = \left(\tilde{\mathbf{\Psi}}_{\mathrm{id}} + \tilde{\mathbf{W}}\tilde{\mathbf{\Psi}}_{\mathrm{f}}\tilde{\mathbf{W}}^{\top}\right)^{-1}\tilde{\mathbf{v}}. \end{aligned}$$

The resulting expected implementation shortfall cost is unchanged:

$$\bar{\mathcal{C}}(\mathbf{v}) \triangleq \frac{1}{2}\mathbf{v}^{\top}\Delta\mathbf{p} = \frac{1}{2}\tilde{\mathbf{v}}^{\top}\mathbf{r} = \frac{1}{2}\tilde{\mathbf{v}}^{\top}\left(\tilde{\mathbf{\Psi}}_{\mathrm{id}} + \tilde{\mathbf{W}}\tilde{\mathbf{\Psi}}_{\mathrm{f}}\tilde{\mathbf{W}}^{\top}\right)^{-1}\tilde{\mathbf{v}}.$$

C. Estimation of Cross-asset Market Impact

In this section, we provide a detailed description of the estimation procedure sketched in §6, and verify the procedure based on a carefully synthesized dataset.

C.1. Estimation Scheme

Required data. We assume that we have access to realized portfolio executions, their realized shortfalls, and reference information about the prevailing weight vectors of popular index funds (such as the market and sector portfolios). More specifically, we assume that the given data

contains the following information. First, the portfolio transactions $\tilde{\mathbf{v}}_{dt} \in \mathbb{R}^N$ that is a portfolio vector executed during time interval t on day d, expressed in (signed) notional dollar amounts. Second, the realized implementation shortfalls (return) $\bar{\mathbf{r}}^{\text{tr}}_{dt} \in \mathbb{R}^N$ incurred in the execution of portfolio $\tilde{\mathbf{v}}_{dt}$ relative to the arrival price vector at the beginning of time interval t on day d.[13] Third, some reference information that are publicly available: (i) the realized end-to-end returns $\mathbf{r}_{dt} \in \mathbb{R}^N$ during time interval t, (ii) the intraday trading volume DVol_{idt} that is the total market volume of stock i during time interval t on day d, expressed in (unsigned) notional dollar amounts, and (iii) the daily allocation of index funds $\tilde{\mathbf{W}}_d = [\tilde{\mathbf{w}}_{1d}, \ldots, \tilde{\mathbf{w}}_{Kd}] \in \mathbb{R}^{N\times K}$ where $\tilde{\mathbf{w}}_{kd}$ is the dollar-weighted vector of index fund k on day d, normalized (i.e., $\mathbf{1}^\top \tilde{\mathbf{w}}_{kd} = 1$, for all d and k). Optionally, a proxy for the amount of index-fund order flows, i.e., a quantity that reflects the trade volume generated by index-fund investors. Depending on the availability of such a proxy, we may adopt different parameterizations of the cross-impact model, (M1) or (M2), which will be introduced below.

Cross-asset impact model. The derivation in §3 predicts the following relationship between the executed quantity $\tilde{\mathbf{v}}_{dt}$ and the realized shortfall[14] $\bar{\mathbf{r}}^{\text{tr}}_{dt}$: analogous to (11), we derive

$$\bar{\mathbf{r}}^{\text{tr}}_{dt} = \frac{1}{2}\tilde{\mathbf{G}}_{dt}\tilde{\mathbf{v}}_{dt} + \bar{\boldsymbol{\epsilon}}^{\text{tr}}_{dt}, \quad \tilde{\mathbf{G}}_{dt} = \left(\tilde{\boldsymbol{\Psi}}_{\text{id},dt} + \tilde{\mathbf{W}}_d\tilde{\boldsymbol{\Psi}}_{\text{f},dt}\tilde{\mathbf{W}}_d^\top\right)^{-1}, \tag{51}$$

where the rescaled liquidity matrices are given by $\tilde{\boldsymbol{\Psi}}_{\text{id},dt} = \text{diag}_{i=1}^N(\tilde{\psi}_{\text{id},idt}) \in \mathbb{R}^{N\times N}$ and $\tilde{\boldsymbol{\Psi}}_{\text{f},dt} = \text{diag}_{k=1}^K(\tilde{\psi}_{\text{f},kdt}) \in \mathbb{R}^{K\times K}$, and the noise term $\bar{\boldsymbol{\epsilon}}^{\text{tr}}_{dt} \in \mathbb{R}^N$ describes the random fluctuation of price. Next we introduce a further parameterization of $\tilde{\boldsymbol{\Psi}}_{\text{id},dt}$ and $\tilde{\boldsymbol{\Psi}}_{\text{f},dt}$, in which we reduce the number of free parameters for the idiosyncratic components (as is typically done), and further simplify how we capture the non-stationary behavior of the various terms so as to be able to rely on market observable quantities as proxies.

A parameterization with "idiosyncratic" and "factor" trading volume. As discussed in §3, the liquidity variable $\tilde{\psi}_{\text{id},idt}$ (resp., $\tilde{\psi}_{\text{f},kdt}$) represents the notional amount of stock i (resp., index fund k) that will be supplied by single-stock investors (resp., index-fund investors) in response to a movement in the price of the stock (resp., index). We interpret that the variable $\tilde{\psi}$ captures (i) the number of investors, or participation intensity, present in each period, and (ii) the sensitivity of

[13] We require the return dataset $\bar{\mathbf{r}}^{\text{tr}}_{dt}$ to have no missing entries. For an entry (i, d, t) such that no execution was made at all (i.e., $\tilde{v}_{idt} = 0$), we recommend to set $\bar{r}^{\text{tr}}_{idt}$ to be the return measured with the market volume-weighted-average-price relative to the arrival price at the beginning of interval, or simply a half of the end-to-end market return $\frac{1}{2}r_{idt}$.

[14] In this section, we are using the realized shortfall (return) $\bar{\mathbf{r}}^{\text{tr}}_{dt}$ instead of the absolute price change $\Delta\mathbf{p}_{dt}$, and notional traded vectors $\tilde{\mathbf{v}}_{dt}$ instead of number of shares $\mathbf{v}_{dt}$. Similarly, we use dollar-weighted vectors $\tilde{\mathbf{w}}_k$ instead of share-weighted vectors $\mathbf{w}_k$. With the rescaled liquidity parameters $\tilde{\boldsymbol{\Psi}}_{\text{id}}$ and $\tilde{\boldsymbol{\Psi}}_{\text{f}}$, the structure of the price-impact model remains the same. See Appendix B.

these investors to price movements. The first factor roughly scales in proportion to trading volume, while the second factor varies in a way that depends on the volatility of the underlying security or index, and, specifically, it is plausible to imagine that it scales in a way that it is inversely proportional to the volatility itself, i.e., $\tilde{\psi} \propto \frac{\mathrm{DVol}}{\sigma}$. Based on this interpretation, we consider the parameterizations of $\tilde{\psi}_{\mathrm{id},idt}$ and $\tilde{\psi}_{\mathrm{f},kdt}$ with the following reduced form:

$$\tilde{\psi}_{\mathrm{id},idt} = \gamma_{\mathrm{id}} \times \frac{\widehat{\mathrm{DVol}}_{\mathrm{id},idt}}{\widehat{\sigma}_{\mathrm{id},idt}}, \quad \tilde{\psi}_{\mathrm{f},kdt} = \gamma_{\mathrm{f},k} \times \frac{\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}}{\widehat{\sigma}_{\mathrm{f},kdt}}, \tag{M1}$$

where (i) $\widehat{\mathrm{DVol}}_{\mathrm{id},idt}$ and $\widehat{\sigma}_{\mathrm{id},idt}$ denote (forecasted) "idiosyncratic" trading volume and volatility of stock i that describe the trading activity of single-stock investors, (ii) $\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}$ and $\widehat{\sigma}_{\mathrm{f},kdt}$ denote (forecasted) "factor" trading volume and volatility of index fund k that describe the trading activity of index-fund investors, and (iii) $\gamma_{\mathrm{id}} \in \mathbb{R}$ and $\boldsymbol{\gamma}_{\mathrm{f}} \triangleq (\gamma_{\mathrm{f},1}, \ldots, \gamma_{\mathrm{f},K})^\top \in \mathbb{R}^K$ are unknown time-invariant leading coefficients. We have selected a simple parameterization where all single-stock terms $\tilde{\psi}_{\mathrm{id},idt}$ share the same coefficient γ_{id} that is believed to reflect some invariant characteristic of all single-stock investors.

We do not further formulate "idiosyncratic" and "factor" trading volumes in this paper: they will be latent variables, i.e., they are not immediately quantifiable from the market data, since the actual trading volume that we observe from the market is the mixture of these two components. Someone can estimate the intraday pattern of the proportion of factor trading volume explicitly as suggested in §5.1, or can use some additional market information such as trading volume grouped by investor type if available. Given such proxies, $K+1$ unknown parameters (γ_{id} and $\boldsymbol{\gamma}_{\mathrm{f}}$) can be estimated via a procedure described later.

A parameterization with observables. As an alternative of the parameterization (M1), we propose a more specific parameterization that relies on directly observable quantities:

$$\tilde{\psi}_{\mathrm{id},idt} = \nu_{\mathrm{id},t} \times \frac{\mathrm{MADVol}_{idt}}{\bar{\sigma}_{idt}}, \quad \tilde{\psi}_{\mathrm{f},kdt} = \nu_{\mathrm{f},kt} \times \frac{\sum_{i \in \mathcal{S}_k} \mathrm{MADVol}_{idt}}{\bar{\sigma}_{\mathrm{f},kdt}}. \tag{M2}$$

The coefficients $\nu_{\mathrm{id},t}$ and $\boldsymbol{\nu}_{\mathrm{f},t}$ are the unknowns here, and the others variables are the moving-average measures defined as

$$\mathrm{MADVol}_{idt} \triangleq \frac{1}{\tau}\sum_{s=0}^{\tau-1} \mathrm{DVol}_{i,d-s,t}, \quad \bar{\sigma}_{idt} \triangleq \sqrt{\frac{1}{\tau}\sum_{s=0}^{\tau-1} r_{i,d-s,t}^2}, \quad \bar{\sigma}_{\mathrm{f},kdt} \triangleq \sqrt{\frac{1}{\tau}\sum_{s=0}^{\tau-1} \left(\tilde{\mathbf{w}}_{k,d-s}^\top \mathbf{r}_{d-s,t}\right)^2}, \tag{52}$$

where τ is the length of sliding window and $\mathcal{S}_k$ is the set of stocks that belong to the index fund k. One can adopt different averaging scheme as long as it provides reasonable forecasts for trading

volume and volatility.

Compared to the previous parameterization (M1), the leading coefficients in (M2), $\nu_{\mathrm{id},t}$ and $\boldsymbol{\nu}_{\mathrm{f},t}$, have the subscript t (as opposed to γ_{id} and $\boldsymbol{\gamma}_{\mathrm{f}}$) so as to reflect the intraday variation in composition of the two types of liquidity provision (i.e., $\tilde{\psi}_{\mathrm{id},idt}$ vs. $\tilde{\psi}_{\mathrm{f},kdt}$) correctly. Such a variation is not well captured in the moving-averaged measures introduced above (i.e., MADVol_{idt} vs. $\sum_{i\in\mathcal{S}_k}\mathrm{MADVol}_{idt}$), since the observable trading volume DVol_{idt} is a simple reflection of sum of two types of liquidity.[15] By allowing the unknown coefficients dependent on the time of the day, we let the estimation procedure to find the right values of $\nu_{\mathrm{id},t}$ and $\boldsymbol{\nu}_{\mathrm{f},t}$ that fairly describe the expected intraday profile of liquidity composition.

There are $T\times(K+1)$ values to estimate, too many considering the noise level in the dataset. We further reduce the number of unknowns by imposing a simple intraday variation pattern: we divide a day into three segments and assume that the coefficients are constant within each segment. More specifically, let $\mathcal{T}_{\mathrm{beg}}$, $\mathcal{T}_{\mathrm{end}}$, and $\mathcal{T}_{\mathrm{mid}}$ be the first one hour (09:30–10:30), the last one hour (15:00–16:00), and the remaining trading session (10:30–15:00), respectively, and assume that

$$\nu_{\mathrm{id},t}=\begin{cases}\nu_{\mathrm{id}}^{\mathrm{beg}} & \text{if } t\in\mathcal{T}_{\mathrm{beg}}\\ \nu_{\mathrm{id}}^{\mathrm{mid}} & \text{if } t\in\mathcal{T}_{\mathrm{mid}}\\ \nu_{\mathrm{id}}^{\mathrm{end}} & \text{if } t\in\mathcal{T}_{\mathrm{end}}\end{cases},\quad \nu_{\mathrm{f},kt}=\begin{cases}\nu_{\mathrm{f},k}^{\mathrm{beg}} & \text{if } t\in\mathcal{T}_{\mathrm{beg}}\\ \nu_{\mathrm{f},k}^{\mathrm{mid}} & \text{if } t\in\mathcal{T}_{\mathrm{mid}}\\ \nu_{\mathrm{f},k}^{\mathrm{end}} & \text{if } t\in\mathcal{T}_{\mathrm{end}}\end{cases}. \tag{M2-seg}$$

With this segmentation, we have $3\times(K+1)$ unknowns in total, and the intraday variation in liquidity composition can be represented with the change in their relative magnitude across the segments, i.e., $\nu_{\mathrm{id}}^{\mathrm{beg}}$ vs. $\boldsymbol{\nu}_{\mathrm{f}}^{\mathrm{beg}}$, $\nu_{\mathrm{id}}^{\mathrm{mid}}$ vs. $\boldsymbol{\nu}_{\mathrm{f}}^{\mathrm{mid}}$, and $\nu_{\mathrm{id}}^{\mathrm{end}}$ vs. $\boldsymbol{\nu}_{\mathrm{f}}^{\mathrm{end}}$.

Estimation procedure. We illustrate a simple procedure that estimates the unknown coefficients in the parameterization (M2) for intraday segment $\mathcal{T}_{\mathrm{end}}$. The same procedure can apply for the other intraday segments as well as the parameterization (M1).

We aim to find the values of $\nu_{\mathrm{id}}^{\mathrm{end}}\in\mathbb{R}$ and $\boldsymbol{\nu}_{\mathrm{f}}^{\mathrm{end}}\in\mathbb{R}^K$ such that their corresponding cross-impact model fits the actual price changes realized during the time periods $t\in\mathcal{T}_{\mathrm{end}}$. Let us denote the cross-impact matrix parameterized with ν_{id} and $\boldsymbol{\nu}_{\mathrm{f}}$ by $\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}},\boldsymbol{\nu}_{\mathrm{f}})$: More specifically,

$$\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}},\boldsymbol{\nu}_{\mathrm{f}})\triangleq\left(\mathrm{diag}_{i=1}^{N}\left(\nu_{\mathrm{id}}\cdot\frac{\mathrm{MADVol}_{idt}}{\bar{\sigma}_{idt}}\right)+\tilde{\mathbf{W}}_d\mathrm{diag}_{k=1}^{K}\left(\nu_{\mathrm{f},k}\cdot\frac{\sum_{i\in\mathcal{S}_k}\mathrm{MADVol}_{idt}}{\bar{\sigma}_{\mathrm{f},kdt}}\right)\tilde{\mathbf{W}}_d^{\top}\right)^{-1}.$$

[15] Suppose that the intensity of liquidity provision by index-fund investors stays constant over time, i.e., $\tilde{\psi}_{\mathrm{f},kdt}$ does not vary over the course of the day but does $\tilde{\psi}_{\mathrm{id},idt}$ only. The market-wide volume $\sum_{i\in\mathcal{S}_k}\mathrm{MADVol}_{idt}$ will still fluctuate according to the variation of single-stock investors' liquidity provision, and therefore, the time-variation of $\sum_{i\in\mathcal{S}_k}\mathrm{MADVol}_{idt}$ does not correctly reflect the time-variation of index-fund investors' liquidity provision.

We first introduce the empirical loss $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}$ with respect to the realized single-stock shortfalls $\bar{\mathbf{r}}_{dt}^{\mathrm{tr}}$: as we expect that $\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} \approx \frac{1}{2}\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}})\tilde{\mathbf{v}}_{dt}$,

$$\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}}) \triangleq \frac{1}{N}\sum_{d=1}^{D}\sum_{t\in\mathcal{T}_{\mathrm{end}}}\left(\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} - \frac{1}{2}\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}})\tilde{\mathbf{v}}_{dt}\right)^{\top}\bar{\boldsymbol{\Sigma}}_{dt}^{-1}\left(\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} - \frac{1}{2}\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}})\tilde{\mathbf{v}}_{dt}\right),$$

where $\bar{\boldsymbol{\Sigma}}_{dt} \triangleq \mathrm{diag}_{i=1}^{N}(\bar{\sigma}_{idt}^2)$ is the empirical diagonal covariance matrix. Analogously, the empirical loss $\mathcal{L}_{\mathrm{f}}^{\mathrm{end}}$ with respect to the realized index-fund shortfalls $\tilde{\mathbf{W}}_d^{\top}\bar{\mathbf{r}}_{dt}^{\mathrm{tr}}$ can be defined as follows:

$$\mathcal{L}_{\mathrm{f}}^{\mathrm{end}}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}}) \triangleq \frac{1}{K}\sum_{d=1}^{D}\sum_{t\in\mathcal{T}_{\mathrm{end}}}\left(\tilde{\mathbf{W}}_d^{\top}\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} - \frac{1}{2}\tilde{\mathbf{W}}_d^{\top}\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}})\tilde{\mathbf{v}}_{dt}\right)^{\top}\bar{\boldsymbol{\Sigma}}_{\mathrm{f},dt}^{-1}\left(\tilde{\mathbf{W}}_d^{\top}\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} - \frac{1}{2}\tilde{\mathbf{W}}_d^{\top}\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}})\tilde{\mathbf{v}}_{dt}\right),$$

where $\bar{\boldsymbol{\Sigma}}_{\mathrm{f},dt} \triangleq \mathrm{diag}_{k=1}^{K}(\bar{\sigma}_{kdt}^2)$ is the empirical diagonal covariance matrix of index-fund returns.

Observe that minimizing $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}$ or $\mathcal{L}_{\mathrm{f}}^{\mathrm{end}}$ is identical to performing a least squares estimation with the heteroscedastic residual terms. In particular when the index-fund investors' contribution is absent (i.e., when we are restricted to have $\boldsymbol{\nu}_{\mathrm{f}} = \mathbf{0}$), minimizing $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}$ is equivalent to the estimation procedure proposed in Almgren et al. (2005). Similarly, when the single-stock investors' contribution is absent (i.e., when we are restricted to have $\nu_{\mathrm{id}} = 0$) and the index funds are orthogonal, minimizing $\mathcal{L}_{\mathrm{f}}^{\mathrm{end}}$ is equivalent to fitting a separable linear model under which each index fund is treated in isolation. In other words, the loss $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}$ focuses more on the diagonal entries of cross-impact matrix $\tilde{\mathbf{G}}_{dt}$ and the loss $\mathcal{L}_{\mathrm{f}}$ rather focuses on the non-diagonal entries of $\tilde{\mathbf{G}}_{dt}$.

Based on those loss measures, we suggest a four-step procedure that yields the estimates $\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}$ and $\widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}$:

1. (Initial guess) Find a single scalar value $\widehat{\nu}$ via an ordinary least-squares regression based on the following linear model:

$$\bar{r}_{idt}^{\mathrm{tr}} = \frac{1}{2} \times \nu^{-1} \times \frac{\bar{\sigma}_{idt}}{\mathrm{MADVol}_{idt}} \times \tilde{v}_{idt} + e_{idt},$$

   where e_{idt}'s are i.i.d. Initialize the estimates with $\widehat{\nu}$: i.e., $\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}} \leftarrow \widehat{\nu}$ and $\widehat{\nu}_{\mathrm{f},k}^{\mathrm{end}} \leftarrow \widehat{\nu}$ for all $k = 1, \ldots, K$.

2. (Estimation of diagonal entries) Fix $\widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}$ and find $\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}$ that best explains the realized single-stock shortfalls by minimizing loss $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}$:

$$\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}} \leftarrow \underset{\nu_{\mathrm{id}}\in\mathbb{R}_+}{\mathrm{argmin}}\, \mathcal{L}_{\mathrm{id}}^{\mathrm{end}}(\nu_{\mathrm{id}}, \widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}).$$

3. (Estimation of non-diagonal entries) Fix $\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}$ and find $\widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}$ that best explains the realized index-fund shortfalls by minimizing loss $\mathcal{L}_{\mathrm{f}}^{\mathrm{end}}$:

$$\widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}} \leftarrow \operatorname*{argmin}_{\boldsymbol{\nu}_{\mathrm{f}} \in \mathbb{R}_{+}^{K}} \mathcal{L}_{\mathrm{f}}^{\mathrm{end}}(\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}, \boldsymbol{\nu}_{\mathrm{f}}).$$

4. (Fine tuning) Finally adjust the estimates by minimizing two loss measures simultaneously:

$$(\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}, \widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}) \leftarrow \operatorname*{argmin}_{(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}}) \in \mathbb{R}_{+} \times \mathbb{R}_{+}^{K}} \left\{ \mathcal{L}_{\mathrm{id}}^{\mathrm{end}}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}}) + \mathcal{L}_{\mathrm{f}}^{\mathrm{end}}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}}) \right\}$$

Step 1 replicates the procedure to estimate a separable (idiosyncratic) impact model that is commonly adopted in the literature, and the estimate $\widehat{\nu}$ found in step 1 is utilized as a baseline value for $\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}$ and $\widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}$ in the next steps. In steps 2 and 3, it performs further estimation of $\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}$ and $\widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}}$ by minimizing the losses $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}}$ and $\mathcal{L}_{\mathrm{f}}^{\mathrm{end}}$ individually, and in step 4 it performs a fine tuning by minimizing $\mathcal{L}_{\mathrm{id}}^{\mathrm{end}} + \mathcal{L}_{\mathrm{f}}^{\mathrm{end}}$ together. In the implementation, we suggest to use a simple gradient descent method in each step: the loss minimizer needs not to be the global optimum since the results from the previous steps will provide reasonable initial solutions to the next steps.

This four-step procedure aims to estimate the coefficients $\nu_{\mathrm{id}}^{\mathrm{end}}$ and $\boldsymbol{\nu}_{\mathrm{f}}^{\mathrm{end}}$ in a robust and efficient way. By sequentially improving the estimates starting from the parameter value estimated from a simple separable impact model, it prevents the final outcomes from taking extreme values and accelerates the optimization procedure. One may performs step 4 only, but we anticipate that the outcome will be very sensitive to initialization values because the objective is non-convex and possibly multimodal. The matrix $\tilde{\mathbf{G}}_{dt}(\nu_{\mathrm{id}}, \boldsymbol{\nu}_{\mathrm{f}})$ can be computed efficiently in practice by using the Woodbury matrix identity.

Comparison with Schneider and Lillo (2019). Schneider and Lillo (2019) perform a direct non-parametric estimation of cross-impact among Italian and European bonds ($N = 33$) using high-frequency market data in an effort to validate their theoretical findings on no-arbitrage conditions. Compared to their estimation procedure, our estimation heavily relies on the model: we postulate a parsimonious parametric representation of cross-impact and exploit its structure to alleviate the difficulty of direct estimation. Since our model is based on a stylized characterization of index-fund investors participating the stock markets, our suggested estimation scheme may not be appropriate to estimate cross-impact among fixed-income securities that may share a different kind of commonality following from risk and term structure. On the other hand, their non-parametric approach may not be appropriate to handle non-stationary cross-impact as opposed to our para-

metric approach that uses trading volume as a proxy for time-varying liquidity. Although it would be an interesting research topic to adopt direct estimation methods and compare the results, we believe that it would be beyond the scope of this paper.

C.2. Illustration with Synthetic Data

We illustrate and verify the suggested estimation procedure by using a (partly) synthetic dataset. We construct test portfolios, one for each day d and period t. We will then simulate the execution of these portfolios by adding some market impact on top of the actual market return in (d, t). We pick randomly a set of market impact coefficients for the model in (M1), and simulate execution costs for the test portfolios by adding the expected impact cost contribution to the realized market return. We then forget the impact cost coefficients and the detailed specification of (M1), and given the set of test portfolios and realized execution costs we estimate an impact model given the (observable) parameterization (M2). Even though we use slightly different market impact models for the dataset generation and for the parameter estimation, we will illustrate that it can still predict the transaction costs well enough in spite of the model misspecification.

Original dataset. We consider three sectors ($K = 3$; energy, finance, and technology sectors) and S&P 500 stocks that belong to these sectors ($N = 153$) throughout the year 2018. As in §2, we consider five-minute intervals ($T = 78$) per each day, and exclude half-trading days and the days on which FED announcements were made. In the calculation of moving-averaged measures (52), we use the time window of $\tau = 60$ days, and therefore the estimation is performed after excluding the first 60 days ($D = 240 - 60 = 180$).

Ground truth market impact model. We assume that the ground truth model is given by the parameterization (M1) where the idiosyncratic volume forecast $\widehat{\mathrm{DVol}}_{\mathrm{id},idt}$ and the factor volume forecast $\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}$ are assumed to have the following form:

$$\widehat{\mathrm{DVol}}_{\mathrm{f},kdt} = \theta_{kt} \times \sum_{i \in \mathcal{S}_k} \mathrm{MADVol}_{idt}, \quad \widehat{\mathrm{DVol}}_{\mathrm{id},idt} = \mathrm{MADVol}_{idt} - w_{kdi} \times \widehat{\mathrm{DVol}}_{\mathrm{f},kdt}. \tag{53}$$

The variable θ_{kt} represents the proportion of the factor trading volume out of total market trading volume across all stocks in sector k on the intraday time interval t.[16] This formulation follows from the assumption that individual stock's trading volume is decomposed into an idiosyncratic component and a factor component, i.e., $\mathrm{MADVol}_{idt} = \widehat{\mathrm{DVol}}_{\mathrm{id},idt} + w_{kdi} \times \widehat{\mathrm{DVol}}_{\mathrm{f},kdt}$, where the sector-wide contribution of factor volume, $\frac{\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}}{\sum_{i \in \mathcal{S}_k} \mathrm{MADVol}_{idt}}$, is assumed to be constant across days.

[16] Compared to the parameterization introduced in §5, the variable θ_{kt} would correspond to the proportion of index-fund order flows, $\frac{\beta_t \cdot \theta}{\alpha_t \cdot (1-\theta) + \beta_t \cdot \theta}$, which is plotted in Figure 3.

Reflecting the empirical observations in §2, we make up the values of θ_{kt}'s as plotted in Figure 8.

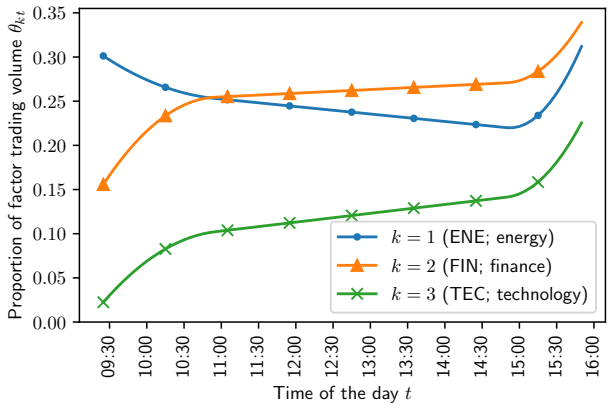

Figure 8: Hypothetical intraday profiles of the proportion of factor trading volume, $(\theta_{kt})_{t=1}^{T}$, that are plugged in (53) for synthetic dataset generation. For example, the second profile curve implies that the index-fund investors in finance sector ($k = 2$) account for 15% of the total sector-wide traded volume at the beginning of the day, and 35% of it at the end of the day.

The values of unknown coefficients γ's are chosen as $\gamma_{\mathrm{id}} = 0.04$ and $\boldsymbol{\gamma}_{\mathrm{f}} = (0.30, 0.10, 0.20)^{\top}$. To gain some intuition of the magnitude of these parameters, when $\gamma = 0.04$ and $\sigma = 0.1\% \approx \frac{1\%}{\sqrt{78}}$ (five-minute volatility), the given parameterization predicts that the expected cost of executing a trade with 2% participation rate will be $\frac{1}{2} \cdot \frac{\sigma}{\gamma} \cdot 2\% \approx 2.5$ basis points.

Given the hypothetical values of θ's and γ's, we assume that the true coefficient matrix of market impact is given by

$$\tilde{\mathbf{G}}_{dt}^{\mathrm{true}} = \left(\mathrm{diag}_{i=1}^{N} \left(\gamma_{\mathrm{id}} \cdot \frac{\widehat{\mathrm{DVol}}_{\mathrm{id},idt}}{\bar{\sigma}_{idt}} \right) + \tilde{\mathbf{W}}_d \mathrm{diag}_{k=1}^{K} \left(\gamma_{\mathrm{f},k} \cdot \frac{\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}}{\bar{\sigma}_{\mathrm{f},kdt}} \right) \tilde{\mathbf{W}}_d^{\top} \right)^{-1}. \tag{54}$$

In what follows, we show that our proposed estimation procedure finds some approximation of $\tilde{\mathbf{G}}_{dt}^{\mathrm{true}}$ with a different parameterization rather than directly estimating the values of θ's and γ's.

Hypothetical portfolio transactions. We imagine a situation that investing decisions are made on a daily basis and the associated portfolio transactions are being executed over the course of the day. More specifically, the hypothetical portfolio transactions $\tilde{\mathbf{v}}_{dt}$'s are generated according to the following procedure: on each day $d = 1, \ldots, D$, independently,

1. we randomly select the single stocks and the sectors to trade: A single stock is selected with probability 5%, and a sector is selected with probability 25%.
2. For each selected single stock (or a selected sector), the trading direction (i.e., buy or sell) is determined randomly, and the participation rate is drawn from the log-normal distribution

with mean 1.5% and standard deviation 0.5% for single stocks, and with mean 0.5% and standard deviation 0.1% for sectors.

3. Given the trading direction and the participation rate, we imagine a VWAP trading schedule over the course of the day: The notional amount of transactions at time t on a selected asset (a stock i or a sector k) is given by

$$\begin{aligned} q_{\mathrm{id},idt} &= (\text{trading direction})_{id} \times (\text{participation rate})_{id} \times \mathrm{MADVol}_{idt}, \\ q_{\mathrm{f},kdt} &= (\text{trading direction})_{kd} \times (\text{participation rate})_{kd} \times \sum_{i \in \mathcal{S}_k} \mathrm{MADVol}_{idt}. \end{aligned}$$

Accordingly, the portfolio transactions on day d are simply given by

$$\tilde{\mathbf{v}}_{dt} = \mathbf{q}_{\mathrm{id},dt} + \tilde{\mathbf{W}}_d \mathbf{q}_{\mathrm{f},dt}, \quad \forall t = 1, \ldots, T,$$

where $\mathbf{q}_{\mathrm{id},dt} \triangleq (q_{\mathrm{id},1dt}, \ldots, q_{\mathrm{id},Ndt})^\top \in \mathbb{R}^N$ and $\mathbf{q}_{\mathrm{f},dt} \triangleq (q_{\mathrm{f},1dt}, \ldots, q_{\mathrm{f},Kdt})^\top \in \mathbb{R}^K$.

This procedure is selected intentionally to match up to calibrating the model on a set of full day VWAP-like executions.

Hypothetical dataset. Given the ground truth impact model $\tilde{\mathbf{G}}_{dt}^{\mathrm{true}}$ and the simulated portfolio transactions $\tilde{\mathbf{v}}_{dt}$, we make perturbation on the original dataset: The realized implementation shortfalls $\bar{\mathbf{r}}_{dt}^{\mathrm{tr}}$, the realized end-to-end returns $\mathbf{r}_{dt}$ and the market trading volume DVol_{idt} are overwritten as

$$\bar{\mathbf{r}}_{dt}^{\mathrm{tr}} \leftarrow \bar{\mathbf{r}}_{dt}^{\mathrm{tr}} + \frac{1}{2} \tilde{\mathbf{G}}_{dt}^{\mathrm{true}} \tilde{\mathbf{v}}_{dt}, \quad \mathbf{r}_{dt} \leftarrow \mathbf{r}_{dt} + \tilde{\mathbf{G}}_{dt}^{\mathrm{true}} \tilde{\mathbf{v}}_{dt}, \quad \mathrm{DVol}_{idt} \leftarrow \mathrm{DVol}_{idt} + |\tilde{v}_{idt}|, \tag{55}$$

and we obtain a hypothetical dataset in which the effects of the transactions $\tilde{\mathbf{v}}_{dt}$ are reflected. In this hypothetical dataset, 59% of our trades are due to index-fund investing, and we pay 2.4 basis points for the transaction cost in total.

Estimation result. We perform the estimation based on the parameterization (M2) with the segmentation scheme (M2-seg). We estimate $3 \times (K+1)$ unknown coefficients $(\widehat{\nu}_{\mathrm{id}}^{\mathrm{beg}}, \widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{beg}})$, $(\widehat{\nu}_{\mathrm{id}}^{\mathrm{mid}}, \widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{mid}})$, and $(\widehat{\nu}_{\mathrm{id}}^{\mathrm{end}}, \widehat{\boldsymbol{\nu}}_{\mathrm{f}}^{\mathrm{end}})$ by applying the four-step procedure described in §C.1 for each of intraday segments $\mathcal{T}_{\mathrm{beg}}$, $\mathcal{T}_{\mathrm{mid}}$, and $\mathcal{T}_{\mathrm{end}}$ separately.

Since the estimation model is different from the model used for data generation, there are no true parameter values that exactly correspond to the estimated parameters. Instead, we introduce

the "effective" true coefficients for (M2) that are expressed in terms of θ and γ used in (M1):

$$\bar{\nu}_{\mathrm{id},t} \triangleq \gamma_{\mathrm{id}} \times \left(1 - \frac{1}{K}\sum_{k=1}^{K} \theta_{kt}\right), \quad \bar{\nu}_{\mathrm{f},kt} \triangleq \gamma_{\mathrm{f},k} \times \theta_{kt}.$$

Figure 9 (left) shows the comparison between the estimated coefficients $(\widehat{\nu}_{\mathrm{id},t}, \widehat{\boldsymbol{\nu}}_{\mathrm{f},t})_{t=1}^{T}$ and their effective true values $(\bar{\nu}_{\mathrm{id},t}, \bar{\boldsymbol{\nu}}_{\mathrm{f},t})_{t=1}^{T}$. We observe that the estimated model approximates the intraday variation of the ground truth market impact with piecewise constant profiles.

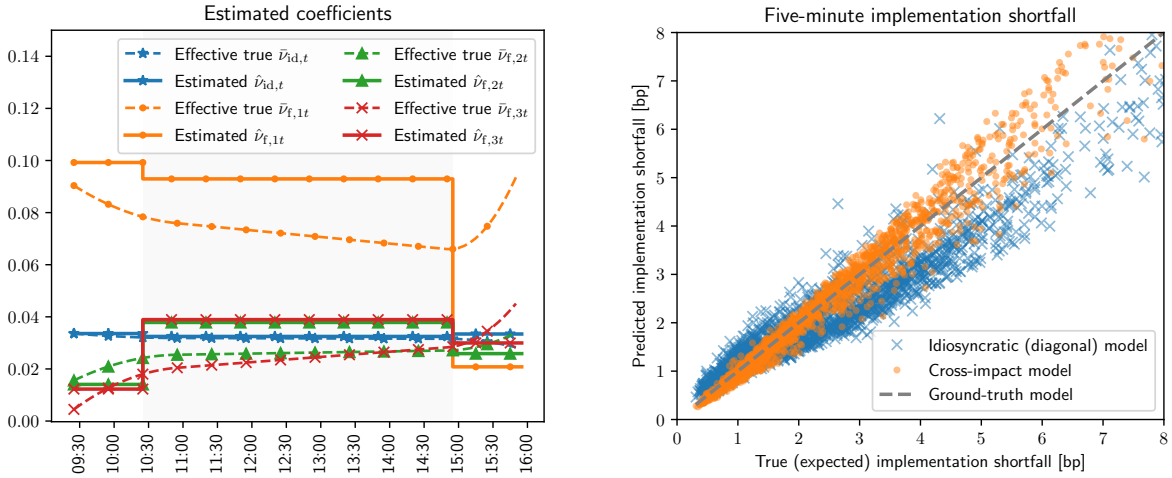

Figure 9: Estimated coefficients $\widehat{\boldsymbol{\nu}}$ vs. effective true coefficients $\bar{\boldsymbol{\nu}}$ (left). The estimation results approximate the intraday variation of the cross-impact matrix. The estimated five-minute implementation shortfall $\hat{y}_{dt}^{\mathrm{cross}}$ and $\hat{y}_{dt}^{\mathrm{diag}}$ vs. the true expected five-minute implementation shortfall $\hat{y}_{dt}^{\mathrm{true}}$ (right, the data points are randomly selected for visualization).

We further investigate the performance of estimated model relative to the idiosyncratic (diagonal) market impact model[17] that may be adopted by someone who ignores the cross-asset impact. More specifically, we compute the realized five-minute implementation shortfall y_{dt}^{real} (expressed in percentage) incurred by the hypothetical portfolio transaction $\tilde{\mathbf{v}}_{dt}$ and the model predictions:

$$y_{dt}^{\mathrm{real}} \triangleq \frac{\bar{\mathbf{r}}_{dt}^{\top}\tilde{\mathbf{v}}_{dt}}{\|\tilde{\mathbf{v}}_{dt}\|_1}, \; \hat{y}_{dt}^{\mathrm{true}} \triangleq \frac{\frac{1}{2}\tilde{\mathbf{v}}_{dt}^{\top}\tilde{\mathbf{G}}_{dt}^{\mathrm{true}}\tilde{\mathbf{v}}_{dt}}{\|\tilde{\mathbf{v}}_{dt}\|_1}, \; \hat{y}_{dt}^{\mathrm{cross}} \triangleq \frac{\frac{1}{2}\tilde{\mathbf{v}}_{dt}^{\top}\tilde{\mathbf{G}}_{dt}(\widehat{\nu}_{\mathrm{id},t}, \widehat{\boldsymbol{\nu}}_{\mathrm{f},t})\tilde{\mathbf{v}}_{dt}}{\|\tilde{\mathbf{v}}_{dt}\|_1}, \; \hat{y}_{dt}^{\mathrm{diag}} \triangleq \frac{\frac{1}{2}\tilde{\mathbf{v}}_{dt}^{\top}\tilde{\mathbf{G}}_{dt}(\widehat{\nu}_{t}^{\mathrm{diag}}, \mathbf{0})\tilde{\mathbf{v}}_{dt}}{\|\tilde{\mathbf{v}}_{dt}\|_1},$$

where $\hat{y}_{dt}^{\mathrm{true}}$, $\hat{y}_{dt}^{\mathrm{cross}}$, and $\hat{y}_{dt}^{\mathrm{diag}}$ are the expected costs predicted with, respectively, the ground truth model, the cross-asset (non-diagonal) model estimated above, and the idiosyncratic (diagonal) model. We also compute associated R^2 values as a performance measure of each model: with

[17] The idiosyncratic impact model can be seen as a special case of our estimation model where index-fund investors do not exist (i.e., $\tilde{\mathbf{G}}_{dt}(\nu_t^{\mathrm{diag}}, \mathbf{0})$). For the estimation of the diagonal model, we find the best coefficient $\widehat{\nu}_t^{\mathrm{diag}}$ that minimizes the loss $\mathcal{L}_{\mathrm{id}}$ for each intraday segment (only steps 1 & 2 are performed).

$\bar{y} \triangleq \frac{1}{DT} \sum_d \sum_t y_{dt}^{\text{real}}$,

$$R^2_{\text{true}} \triangleq 1 - \frac{\sum_d \sum_t (y_{dt}^{\text{real}} - \hat{y}_{dt}^{\text{true}})^2}{\sum_d \sum_t (y_{dt}^{\text{real}} - \bar{y})^2},\ R^2_{\text{cross}} \triangleq 1 - \frac{\sum_d \sum_t (y_{dt}^{\text{real}} - \hat{y}_{dt}^{\text{cross}})^2}{\sum_d \sum_t (y_{dt}^{\text{real}} - \bar{y})^2},\ R^2_{\text{diag}} \triangleq 1 - \frac{\sum_d \sum_t (y_{dt}^{\text{real}} - \hat{y}_{dt}^{\text{diag}})^2}{\sum_d \sum_t (y_{dt}^{\text{real}} - \bar{y})^2}.$$

From Figure 9 (right), we can visually verify that, in this simulated setup, the estimated model improves the accuracy of prediction compared to the prevalent idiosyncratic (diagonal) impact model. Table 2 shows R^2 values of each model for each intraday segment: while the overall R^2 values are small (even for the ground truth model) due to the high noise level in return realizations, the estimated model works better than the idiosyncratic model.

Intraday segment	R^2_{true}	R^2_{cross}	R^2_{diag}
Beginning of the day (09:30–10:30)	0.1123	0.1184	0.1043
Middle of the day (10:30–15:00)	0.0718	0.0673	0.0611
End of the day (15:00–16:00)	0.0729	0.0683	0.0639
All day (09:30–16:00)	0.1019	0.1023	0.0927

Table 2: R^2 values of the ground truth model, our suggested model, and a separable diagonal impact model on a synthetic dataset.

We expect that the estimation may not work well in some situations. For example, if the realized transactions account for a negligible fraction of the total market volume, their market impact will be insignificant and hardly distinguishable from the noise, and this may result in inaccurate estimates. If the realized transactions are mainly driven by the single-stock level investments, their aggregate impact on the index-fund prices will be relatively smaller than their impact on the individual stock prices, and hence the estimates related to index-fund liquidity (i.e., $\widehat{\nu}_{\text{f},kt}$) will involve larger estimation errors than the ones related to single-stock liquidity (i.e., $\widehat{\nu}_{\text{id},t}$). A similar issue may arise when the market liquidity provision is mainly driven by the index-fund investors since the impact on the index fund prices will be relatively small. We observe relatively large estimation errors for the last intraday segment in our numerical demonstration (Figure 9), and this would be partly attributable to the above concern because near the end of the day a larger proportion of liquidity is provided along the index funds.

D. Proofs

D.1. Proof of Proposition 2

We first focus on the case where $\mathbf{v} = \mathbf{W}\mathbf{u}$ in (15). When $\alpha = \beta = 1$, by Woodbury's identity we get

$$\mathbf{G} = \left(\mathbf{\Psi}_{\text{id}} + \mathbf{W}\mathbf{\Psi}_{\text{f}}\mathbf{W}^\top\right)^{-1} = \mathbf{\Psi}_{\text{id}}^{-1} - \mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\mathbf{\Psi}_{\text{f}}^{-1} + \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}.$$

Consequently,

$$\begin{aligned}\mathbf{W}^\top\mathbf{G}\mathbf{W} &= \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W} - \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\mathbf{\Psi}_{\text{f}}^{-1} + \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W} \\ &= \left(\left(\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1} + \mathbf{\Psi}_{\text{f}}\right)^{-1}.\end{aligned}$$

Next, we incorporate the effect of α and β as follows:

$$\mathbf{W}^\top\mathbf{G}\mathbf{W} = \left(\alpha\cdot\left(\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1} + \beta\cdot\mathbf{\Psi}_{\text{f}}\right)^{-1} \longrightarrow \mathbf{\Psi}_{\text{f}}^{-1} \quad \text{as } \alpha\to 0 \text{ and } \beta\to 1.$$

Therefore, for any $\mathbf{u}\in\mathbb{R}^K$,

$$\lim_{\alpha\to 0,\beta\to 1}\bar{\mathcal{C}}\left(\mathbf{v}=\mathbf{W}\mathbf{u}\right) = \lim_{\alpha\to 0,\beta\to 1}\frac{1}{2}\mathbf{u}^\top\mathbf{W}^\top\mathbf{G}\mathbf{W}\mathbf{u} = \frac{1}{2}\mathbf{u}^\top\mathbf{\Psi}_{\text{f}}^{-1}\mathbf{u}.$$

Next we consider the case where $\mathbf{v}\notin\text{span}(\mathbf{w}_1,\cdots,\mathbf{w}_K)$. Let $\mathbf{v} = \mathbf{W}\mathbf{u}+\mathbf{e}$ for some $\mathbf{e}\in\mathbb{R}^N$ such that $\mathbf{W}^\top\mathbf{e} = \mathbf{0}$ and $\mathbf{e}\neq\mathbf{0}$. By the Woodbury matrix identity, we get

$$\mathbf{G} = \alpha^{-1}\cdot\mathbf{\Psi}_{\text{id}}^{-1} - \alpha^{-1}\cdot\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\frac{\alpha}{\beta}\mathbf{\Psi}_{\text{f}}^{-1} + \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}.$$

Therefore,

$$\lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{G}\right\} = \mathbf{\Psi}_{\text{id}}^{-1} - \mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}.$$

With $\mathbf{r}\triangleq\mathbf{\Psi}_{\text{id}}^{-1/2}\mathbf{e}$ and $\mathbf{A}\triangleq\mathbf{\Psi}_{\text{id}}^{-1/2}\mathbf{W}$,

$$\mathbf{e}^\top\left(\mathbf{\Psi}_{\text{id}}^{-1} - \mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\right)\mathbf{e} = \mathbf{r}^\top\mathbf{r} - \mathbf{r}^\top\mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{r}.$$

Note that $\mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{r}$ is a projection of $\mathbf{r}$ onto the space spanned by $\mathbf{A}$ (denoted by $\text{span}(\mathbf{A})$). Therefore,

$$\lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{e}^\top\mathbf{G}\mathbf{e}\right\} = 0 \quad \text{if and only if} \quad \mathbf{r}\in\text{span}(\mathbf{A}).$$

If $\mathbf{r} \in \text{span}(\mathbf{A})$, i.e., $\mathbf{r} = \mathbf{A}\mathbf{s}$ for some $\mathbf{s} \in \mathbb{R}^K$, then $\mathbf{e} = \mathbf{\Psi}_{\text{id}}^{1/2}\mathbf{r} = \mathbf{\Psi}_{\text{id}}^{1/2}\mathbf{\Psi}_{\text{id}}^{-1/2}\mathbf{W}\mathbf{s} = \mathbf{W}\mathbf{s}$, and hence $\mathbf{v} \in \text{span}(\mathbf{W})$. Since we are assuming $\mathbf{v} \notin \text{span}(\mathbf{W})$, we have $\mathbf{r} \notin \text{span}(\mathbf{A})$, and hence

$$\lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{e}^\top\mathbf{G}\mathbf{e}\right\} > 0.$$

Furthermore,

$$\begin{aligned}\mathbf{G}\mathbf{W} &= \alpha^{-1}\cdot\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W} - \alpha^{-1}\cdot\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\left(\frac{\alpha}{\beta}\mathbf{\Psi}_{\text{f}}^{-1} + \mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W} \\ &= \alpha^{-1}\cdot\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\underbrace{\left(\mathbf{I}_K - \left[\frac{\alpha}{\beta}\left(\mathbf{W}^\top\mathbf{\Psi}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{\Psi}_{\text{f}}^{-1} + \mathbf{I}_K\right]^{-1}\right)}_{\longrightarrow\mathbf{O}\text{ as }\alpha\to 0}.\end{aligned}$$

Therefore,

$$\lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{e}^\top\mathbf{G}\mathbf{W}\mathbf{u}\right\} = 0.$$

To summarize, since $\lim_{\alpha\to 0,\beta\to 1}\left\{\mathbf{u}\mathbf{W}^\top\mathbf{G}\mathbf{W}\mathbf{u}\right\} = \mathbf{u}^\top\mathbf{\Psi}_{\text{f}}^{-1}\mathbf{u}$, it follows that

$$\begin{aligned}&\lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot(\mathbf{W}\mathbf{u}+\mathbf{e})^\top\mathbf{G}(\mathbf{W}\mathbf{u}+\mathbf{e})\right\} \\ =\ &\lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{u}\mathbf{W}^\top\mathbf{G}\mathbf{W}\mathbf{u}\right\} + \lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot 2\mathbf{e}^\top\mathbf{G}\mathbf{W}\mathbf{u}\right\} + \lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{e}^\top\mathbf{G}\mathbf{e}\right\} \\ =\ &0 + 0 + \lim_{\alpha\to 0,\beta\to 1}\left\{\alpha\cdot\mathbf{e}^\top\mathbf{G}\mathbf{e}\right\} > 0. \qquad (56)\end{aligned}$$

It then follows that $\lim_{\alpha\to 0,\beta\to 1}\left\{(\mathbf{W}\mathbf{u}+\mathbf{e})^\top\mathbf{G}(\mathbf{W}\mathbf{u}+\mathbf{e})\right\} = \infty$. ■

D.2. Proof of Proposition 4

Note that

$$\begin{aligned}&\mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1} - \mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1} \\ =\ &\mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\cdot\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)\cdot\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1} \\ &-\mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\cdot\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\cdot\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1} \\ =\ &\mathbf{W}\bar{\mathbf{\Psi}}_{\text{f}}\cdot\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W} - \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)\cdot\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1} \\ =\ &\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\text{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}.\end{aligned}$$

Again, using the Woodbury matrix identity, we get

$$\begin{aligned}
\mathbf{G}_t^{-1}\left(\sum_{s=1}^{T}\mathbf{G}_s^{-1}\right)^{-1} &= \left(\alpha_t\bar{\mathbf{\Psi}}_{\mathrm{id}}+\beta_t\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)\left(\bar{\mathbf{\Psi}}_{\mathrm{id}}+\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\\
&= \left(\alpha_t\bar{\mathbf{\Psi}}_{\mathrm{id}}+\beta_t\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)\left(\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}-\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1}+\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\right)\\
&= \alpha_t\mathbf{I}_N-\alpha_t\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1}+\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\\
&\quad+\beta_t\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}-\beta_t\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}+\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\\
&= \alpha_t\mathbf{I}_N+(\beta_t-\alpha_t)\mathbf{W}\underbrace{\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1}+\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}}_{\triangleq\widehat{\mathbf{W}}^\top}.
\end{aligned}$$

■

D.3. Simple Generative Order Flow Model used in §5

We first establish an explicit relationship between intraday variation of natural liquidity and intraday variation of the resulting traded volume, by introducing a stochastic-process generative model for trading volume. The underlying motivation is simple yet intuitive: single-stock and index-fund investors create (stochastic) order flows onto the securities they wish to trade. The arrival intensity of these order flows per type of investor in each time period is proportional to the corresponding trading activity or liquidity provided by this investor type in this time period. This is captured by the profiles α_t and β_t, respectively.

Specifically, we assume that the notional trade volume of stock i in time interval t on day d, DVol_{idt}, is composed of order flows made by single-stock investors $Q_{\mathrm{id},idt}$ and a $|\tilde{w}_{1i}|$ proportion of order flows made by index-fund investors $Q_{\mathrm{f},dt}$. We let $|\tilde{w}_{1i}|$ be dollar-weighted ownership of stock i in the index fund so that trading one dollar amount of an index fund accumulates $|\tilde{w}_{1i}|$ dollar amount of notional trade volume onto stock i. Each order flow can naturally be decomposed into small transactions:

$$\mathrm{DVol}_{idt}=Q_{\mathrm{id},idt}+|\tilde{w}_{1i}|\cdot Q_{\mathrm{f},dt}=\sum_{j=1}^{N_{\mathrm{id},idt}}q_{\mathrm{id},idt}(j)+|\tilde{w}_{1i}|\cdot\sum_{j=1}^{N_{\mathrm{f},dt}}q_{\mathrm{f},dt}(j),\tag{57}$$

where $N_{\mathrm{id},idt}$ and $q_{\mathrm{id},idt}(j)$ represent the number of transactions and the absolute size of the j^{th} transaction made by single-stock investors in time interval t on day d. For the transactions made by index-fund investors, $N_{\mathrm{f},dt}$ and $q_{\mathrm{f},dt}(j)$ are defined analogously. We treat $N_{\mathrm{id},idt}$, $N_{\mathrm{f},dt}$, $q_{\mathrm{id},idt}(j)$ and $q_{\mathrm{f},dt}(j)$ as random variables that follow particular distribution assumptions.

The order arrival processes for the two investor types are assumed to be Poisson with time-varying rates that are proportional to α_t and β_t:

$$N_{\mathrm{id},idt} \sim \mathrm{Poisson}(\alpha_t \cdot \Lambda) \quad \text{and} \quad N_{\mathrm{f},dt} \sim \mathrm{Poisson}(\beta_t \cdot \Lambda). \tag{58}$$

We further assume that the individual order quantities $q_{\mathrm{id},idt}(j)$'s (and $q_{\mathrm{f},dt}(j)$'s) are all independent and identically distributed with the following moment conditions:

$$\mathsf{E}\left[q_{\mathrm{id},idt}(j)\right] = \bar{q}_{\mathrm{id},i}, \quad \mathrm{Var}\left[q_{\mathrm{id},idt}(j)\right] = c_v^2 \cdot \bar{q}_{\mathrm{id},i}^2, \quad \mathsf{E}\left[q_{\mathrm{f},dt}(j)\right] = \bar{q}_{\mathrm{f}}, \quad \mathrm{Var}\left[q_{\mathrm{id},idt}(j)\right] = c_v^2 \cdot \bar{q}_{\mathrm{f}}^2, \tag{59}$$

where c_v represents a coefficient of variation.

Under the above assumptions, the single-stock investors' order flow $Q_{\mathrm{id},idt}$ is a compound Poisson process with the following mean and variance:

$$\begin{aligned}
\mathsf{E}\left[Q_{\mathrm{id},idt}\right] &= \mathsf{E}\left[N_{\mathrm{id},idt}\right] \cdot \mathsf{E}\left[q_{\mathrm{id},idt}(j)\right] = \alpha_t \cdot \Lambda \cdot \bar{q}_{\mathrm{id},i}, && (60)\\
\mathrm{Var}\left[Q_{\mathrm{id},idt}\right] &= \mathsf{E}\left[\mathrm{Var}\left(Q_{\mathrm{id},idt}|N_{\mathrm{id},idt}\right)\right] + \mathrm{Var}\left[\mathsf{E}\left(Q_{\mathrm{id},idt}|N_{\mathrm{id},idt}\right)\right] && (61)\\
&= \mathsf{E}\left[N_{\mathrm{id},idt} \cdot c_v^2 \cdot \bar{q}_{\mathrm{id},i}^2\right] + \mathrm{Var}\left[N_{\mathrm{id},idt} \cdot \bar{q}_{\mathrm{id},i}\right] && (62)\\
&= \alpha_t \cdot \Lambda \cdot (c_v^2 + 1) \cdot \bar{q}_{\mathrm{id},i}^2. && (63)
\end{aligned}$$

The mean and variance of $Q_{\mathrm{f},dt}$ can be expressed in a similar manner. Summing these flows for each security we get that

$$\begin{aligned}
\mathsf{E}\left[\mathrm{DVol}_{idt}\right] &= \alpha_t \cdot \Lambda \cdot \bar{q}_{\mathrm{id},i} + \beta_t \cdot \Lambda \cdot |\tilde{w}_{1i}| \cdot \bar{q}_{\mathrm{f}}, && (64)\\
\mathrm{Var}\left[\mathrm{DVol}_{idt}\right] &= \alpha_t \cdot \Lambda \cdot (1 + c_v^2) \cdot \bar{q}_{\mathrm{id},i}^2 + \beta_t \cdot \Lambda \cdot |\tilde{w}_{1i}|^2 \cdot (1 + c_v^2) \cdot \bar{q}_{\mathrm{f}}^2, && (65)\\
\mathrm{Cov}\left[\mathrm{DVol}_{idt}, \mathrm{DVol}_{jdt}\right] &= \beta_t \cdot \Lambda \cdot |\tilde{w}_{1i}| \cdot |\tilde{w}_{1j}| \cdot (1 + c_v^2) \cdot \bar{q}_{\mathrm{f}}^2. && (66)
\end{aligned}$$

The common order flow $Q_{\mathrm{f},dt}$ made by index-fund investors results in a positive correlation between stocks represented in the index.

Define θ_i to be the proportion of daily traded volume generated by index-fund investors out of the total daily traded volume of stock i:

$$\theta_i \triangleq \frac{\sum_{t=1}^{T} \mathsf{E}\left[|\tilde{w}_{1i}| \cdot Q_{\mathrm{f},dt}\right]}{\sum_{t=1}^{T} \mathsf{E}\left[\mathrm{DVol}_{idt}\right]} = \frac{|\tilde{w}_{1i}| \cdot \bar{q}_{\mathrm{f}}}{\bar{q}_{\mathrm{id},i} + |\tilde{w}_{1i}| \cdot \bar{q}_{\mathrm{f}}}. \tag{67}$$

The intraday traded volume profile $\mathrm{VolAlloc}_{it}$ and the pairwise correlation Correl_{ijt}, defined in (1)

and (2), can be simply expressed with θ_i and θ_j:

$$\begin{aligned}
\text{VolAlloc}_{it} &\equiv \frac{\mathsf{E}\left[\text{DVol}_{idt}\right]}{\sum_{s=1}^{T}\mathsf{E}\left[\text{DVol}_{ids}\right]} = \alpha_t \cdot (1-\theta_i) + \beta_t \cdot \theta_i, && (68)\\
\text{Correl}_{ijt} &\equiv \frac{\text{Cov}\left[\text{DVol}_{idt}, \text{DVol}_{jdt}\right]}{\sqrt{\text{Var}\left[\text{DVol}_{idt}\right]} \cdot \sqrt{\text{Var}\left[\text{DVol}_{jdt}\right]}} && (69)\\
&= \frac{\beta_t \cdot \theta_i \cdot \theta_j}{\sqrt{\alpha_t \cdot (1-\theta_i)^2 + \beta_t \cdot \theta_i^2} \cdot \sqrt{\alpha_t \cdot (1-\theta_j)^2 + \beta_t \cdot \theta_j^2}}. && (70)
\end{aligned}$$

If we further assume that the proportions θ_i are the same across all securities,

$$\theta \equiv \theta_1 = \theta_2 = \cdots = \theta_N, \tag{71}$$

then VolAlloc_{it} is the same for all stocks i and Correl_{ijt} is identical across all pairs of stocks, i, j, as given in (28)–(29).

D.4. Proofs for §5.2

D.4.1. Proof of Proposition 5

Note that

$$\begin{aligned}
\Upsilon(\mathbf{x}_0) &= \frac{\sum_{t=1}^{T}(\alpha_t \cdot (1-\theta) + \beta_t \cdot \theta)^2 \cdot \mathbf{x}_0^\top \left(\alpha_t \bar{\mathbf{\Psi}}_{\text{id}} + \beta_t \mathbf{W} \bar{\mathbf{\Psi}}_{\text{f}} \mathbf{W}^\top\right)^{-1} \mathbf{x}_0}{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\text{id}} + \mathbf{W} \bar{\mathbf{\Psi}}_{\text{f}} \mathbf{W}^\top\right)^{-1} \mathbf{x}_0}\\
&= \sum_{t=1}^{T} \alpha_t \cdot (1 + \theta \cdot (\gamma_t - 1))^2 \cdot \frac{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\text{id}} + \gamma_t \mathbf{W} \bar{\mathbf{\Psi}}_{\text{f}} \mathbf{W}^\top\right)^{-1} \mathbf{x}_0}{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\text{id}} + \mathbf{W} \bar{\mathbf{\Psi}}_{\text{f}} \mathbf{W}^\top\right)^{-1} \mathbf{x}_0}.
\end{aligned}$$

By Woodbury's matrix identity,

$$\begin{aligned}
&\frac{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \gamma_t \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0} \\
= \;& \frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0 - \mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\gamma_t^{-1}\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0 - \mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0} \\
= \;& 1 + \frac{\left(\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0\right) - \left(\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\gamma_t^{-1}\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0\right)}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0 - \mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0} \\
= \;& 1 + \frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1} - \left(\gamma_t^{-1}\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\right)\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0 - \mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0} \\
= \;& 1 + \frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\cdot\left(\gamma_t^{-1} - 1\right)\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1}\cdot\left(\gamma_t^{-1}\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0} \\
= \;& 1 + (1-\gamma_t)\cdot\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\left(\mathbf{I}_K + \gamma_t\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0}.
\end{aligned}$$

When $K = 1$, we get

$$\left(\mathbf{I}_K + \gamma_t\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\right)^{-1} = \left(1 + \gamma_t\mathbf{w}_1^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{w}_1\bar{\psi}_{\mathrm{f},1}\right)^{-1} = \frac{1}{1+\gamma_t\cdot\eta_1}.$$

Consequently,

$$\begin{aligned}
&\frac{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \gamma_t \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \left(\bar{\mathbf{\Psi}}_{\mathrm{id}} + \mathbf{W}\bar{\mathbf{\Psi}}_{\mathrm{f}}\mathbf{W}^\top\right)^{-1}\mathbf{x}_0} \\
= \;& 1 + \frac{1-\gamma_t}{1+\eta_1\cdot\gamma_t}\cdot\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0 - \mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0} \\
= \;& 1 + \frac{1-\gamma_t}{1+\eta_1\cdot\gamma_t}\cdot\left(\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\left(\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{W}\right)^{-1}\mathbf{W}^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0} - 1\right)^{-1} \\
= \;& 1 + \frac{1-\gamma_t}{1+\eta_1\cdot\gamma_t}\cdot\left(\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{w}_1\cdot\frac{\bar{\psi}_{\mathrm{f},1}}{1+\bar{\psi}_{\mathrm{f},1}\mathbf{w}_1^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{w}_1}\cdot\mathbf{w}_1^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0} - 1\right)^{-1} \\
= \;& 1 + \frac{1-\gamma_t}{1+\eta_1\cdot\gamma_t}\cdot\left(\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0}{\left(\mathbf{w}_1^\top\bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1}\mathbf{x}_0\right)^2}\cdot\frac{1+\eta_1}{\bar{\psi}_{\mathrm{f},1}} - 1\right)^{-1}.
\end{aligned}$$

To simplify notation, define

$$f(x) \triangleq \left(\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0}{\left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0 \right)^2} \cdot \frac{1+\eta_1}{\bar{\psi}_{\text{f},1}} - 1 \right)^{-1}.$$

Then,

$$\Upsilon(\mathbf{x}_0) \;=\; \sum_{t=1}^{T} \alpha_t \cdot (1+\theta\cdot(\gamma_t-1))^2 \cdot \left(1 + \frac{1-\gamma_t}{1+\eta_1\cdot\gamma_t} \cdot f(\mathbf{x}_0)\right).$$

Note that

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t \cdot (1+\theta\cdot(\gamma_t-1))^2 &= \sum_{t=1}^{T} \alpha_t \cdot (1 + 2\theta\cdot(\gamma_t-1) + \theta^2\cdot(\gamma_t^2 - 2\gamma_t + 1)) \\
&= \sum_{t=1}^{T} \alpha_t + 2\theta\cdot(\beta_t - \alpha_t) + \theta^2 \cdot \left(\frac{\beta_t^2}{\alpha_t} - 2\beta_t + \alpha_t \right) \\
&= 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1 \right).
\end{aligned}$$

As a result,

$$\Upsilon(\mathbf{x}_0) = 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1 \right) + \underbrace{\left(\sum_{t=1}^{T} \frac{\alpha_t \cdot (1-\theta\cdot(1-\gamma_t))^2 (1-\gamma_t)}{1+\eta_1\cdot\gamma_t} \right)}_{\triangleq \Delta} \times f(\mathbf{x}_0).$$

■

D.5. Proof of Remarks 1 – 3

Maximum/minimum cost ratio. Note that $f(\mathbf{x}_0)$ is a decreasing function of $\frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0}{\left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0 \right)^2}$, and

$$\begin{aligned}
\min_{\mathbf{x}_0 \in \mathbb{R}^N} \frac{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0}{\left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0 \right)^2} &= \left(\max_{\mathbf{x}_0 \in \mathbb{R}^N} \frac{\left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0 \right)^2}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{x}_0} \right)^{-1} = \left(\max_{\mathbf{y} \in \mathbb{R}^N} \frac{\left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1/2} \mathbf{y} \right)^2}{\mathbf{y}^\top \mathbf{y}} \right)^{-1} \\
&= \left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1} \mathbf{w}_1 \right)^{-1} = \frac{\bar{\psi}_{\text{f},1}}{\eta_1}.
\end{aligned}$$

The above value is obtained at $\mathbf{x}_0 = \mathbf{w}_1$. Therefore,

$$\max_{\mathbf{x}_0 \in \mathbb{R}^N} f(\mathbf{x}_0) = f(\mathbf{x}_0 = \mathbf{w}_1) = \left(\frac{\bar{\psi}_{\text{f},1}}{\eta_1} \cdot \frac{1+\eta_1}{\bar{\psi}_{\text{f},1}} - 1 \right)^{-1} = \eta_1.$$

On the other hand, since $\min_{\mathbf{x}_0\in\mathbb{R}^N} \frac{\left(\mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{x}_0\right)^2}{\mathbf{x}_0^\top \bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{x}_0} = 0$ at $\mathbf{x}_0 = \mathbf{w}_1^\perp$, it follows that

$$\min_{\mathbf{x}_0\in\mathbb{R}^N} f(\mathbf{x}_0) = f(\mathbf{x}_0 = \mathbf{w}_1^\perp) = 0.$$

Combining these two results, we have that

$$\begin{aligned}
\max_{\mathbf{x}_0\in\mathbb{R}^N} \Upsilon(\mathbf{x}_0) &= 1+\theta^2\cdot\left(\sum_{t=1}^T \frac{\beta_t^2}{\alpha_t} - 1\right) + \max_{\mathbf{x}_0\in\mathbb{R}^N}\{\Delta\cdot f(\mathbf{x}_0)\} = \max\{\Upsilon_{\text{market}}, \Upsilon_{\text{orth}}\} \\
\min_{\mathbf{x}_0\in\mathbb{R}^N} \Upsilon(\mathbf{x}_0) &= 1+\theta^2\cdot\left(\sum_{t=1}^T \frac{\beta_t^2}{\alpha_t} - 1\right) + \min_{\mathbf{x}_0\in\mathbb{R}^N}\{\Delta\cdot f(\mathbf{x}_0)\} = \min\{\Upsilon_{\text{market}}, \Upsilon_{\text{orth}}\},
\end{aligned}$$

and, trivially, $\Upsilon_{\text{market}} \geq \Upsilon_{\text{orth}}$ if and only if $\Delta \geq 0$.

Sign of Δ with respect to θ. Note that $\Delta(\theta)$ is a quadratic function of θ. It suffices to show that $\Delta(\theta=0)\geq 0$ and $\Delta(\theta=1)\leq 0$. Note that for an arbitrary function $h(\cdot)$,

$$\begin{cases} \text{if } h(\cdot) \text{ is non-decreasing,} & h(\gamma)\cdot(1-\gamma) \leq h(1)\cdot(1-\gamma), \quad \forall\gamma \\ \text{if } h(\cdot) \text{ is non-increasing,} & h(\gamma)\cdot(1-\gamma) \geq h(1)\cdot(1-\gamma), \quad \forall\gamma \end{cases}. \tag{72}$$

In the case of $\theta=0$, by setting $h(\gamma_t) \triangleq \frac{1}{1+\eta_1\cdot\gamma_t}$ which is a non-increasing function, we get

$$\Delta(\theta=0) = \sum_{t=1}^T \frac{\alpha_t\cdot(1-\gamma_t)}{1+\eta_1\cdot\gamma_t} \overset{(72)}{\geq} \sum_{t=1}^T \frac{\alpha_t\cdot(1-\gamma_t)}{1+\eta_1} = (1+\eta_1)^{-1}\sum_{t=1}^T(\alpha_t-\beta_t) = 0.$$

In the case of $\theta=1$, by setting $h(\gamma_t) \triangleq \frac{\gamma_t^2}{1+\eta_1\cdot\gamma_t}$, which is a non-decreasing function, we get

$$\Delta(\theta=1) = \sum_{t=1}^T \frac{\alpha_t\cdot\gamma_t^2(1-\gamma_t)}{1+\eta_1\cdot\gamma_t} \overset{(72)}{\leq} \sum_{t=1}^T \frac{\alpha_t\cdot(1-\gamma_t)}{1+\eta_1} = 0.$$

Change of Υ_{market} with respect to η_1. Note that

$$\begin{aligned}
\frac{\partial\Upsilon_{\text{market}}}{\partial\eta_1} &= \frac{\partial}{\partial\eta_1}\left(\eta_1\cdot\Delta(\eta_1)\right) \\
&= \frac{\partial}{\partial\eta_1}\left(\sum_{t=1}^T \frac{\alpha_t\cdot\left(1-\theta\cdot(1-\gamma_t)\right)^2(1-\gamma_t)}{\eta_1^{-1}+\gamma_t}\right) \\
&= \sum_{t=1}^T \frac{\alpha_t\cdot\left(1-\theta\cdot(1-\gamma_t)\right)^2(1-\gamma_t)}{\eta_1^2\cdot(\eta_1^{-1}+\gamma_t)^2} \\
&= \frac{\theta^2}{\eta_1^2}\cdot\sum_{t=1}^T \alpha_t\cdot\left(1+\frac{\theta^{-1}-1-\eta_1^{-1}}{\eta_1^{-1}+\gamma_t}\right)^2(1-\gamma_t).
\end{aligned}$$

Set $h(\gamma_t) \triangleq \left(1 + \frac{\theta^{-1}-1-\eta_1^{-1}}{\eta_1^{-1}+\gamma_t}\right)^2$. If $\eta_1 \leq \frac{\theta}{1-\theta}$, then $\theta^{-1} - 1 - \eta_1^{-1} \leq 0$, and hence $h(\cdot)$ is non-decreasing. Therefore,

$$\begin{aligned}
\frac{\partial \Upsilon_{\text{market}}}{\partial \eta_1} &= \frac{\theta^2}{\eta_1^2} \cdot \sum_{t=1}^{T} \alpha_t \cdot \left(1 + \frac{\theta^{-1} - 1 - \eta_1^{-1}}{\eta_1^{-1} + \gamma_t}\right)^2 (1 - \gamma_t) \\
&\overset{(72)}{\leq} \frac{\theta^2}{\eta_1^2} \cdot \sum_{t=1}^{T} \alpha_t \cdot \left(1 + \frac{\theta^{-1} - 1 - \eta_1^{-1}}{\eta_1^{-1} + 1}\right)^2 (1 - \gamma_t) \\
&= \frac{\theta^2}{\eta_1^2} \cdot \left(1 + \frac{\theta^{-1} - 1 - \eta_1^{-1}}{\eta_1^{-1} + 1}\right)^2 \cdot \sum_{t=1}^{T} \alpha_t (1 - \gamma_t) \\
&= \frac{\theta^2}{\eta_1^2} \cdot \left(1 + \frac{\theta^{-1} - 1 - \eta_1^{-1}}{\eta_1^{-1} + 1}\right)^2 \cdot \sum_{t=1}^{T} (\alpha_t - \beta_t) \\
&= 0.
\end{aligned}$$

If $\eta_1 \geq \frac{\theta}{1-\theta}$, then $h(\cdot)$ is non-increasing, and hence the sign of the inequality reverses. Therefore,

$$\frac{\partial \Upsilon_{\text{market}}}{\partial \eta_1} \leq 0 \quad \text{if } \eta_1 \leq \frac{\theta}{1-\theta}, \quad \text{and} \quad \frac{\partial \Upsilon_{\text{market}}}{\partial \eta_1} \geq 0 \quad \text{if } \eta_1 \geq \frac{\theta}{1-\theta}.$$

■

When $\eta_1 = 0$,

$$\Upsilon_{\text{market}}(\eta_1 = 0) = 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1\right).$$

Note that $\Upsilon_{\text{market}}(\eta_1 = 0) = \Upsilon_{\text{orth}}$. Since $\Upsilon_{\text{market}}(\eta_1)$ is decreasing in $[0, \frac{\theta}{1-\theta}]$, this completes proof of (40). When $\eta_1 = \frac{\theta}{1-\theta}$, since $\frac{(1-\theta+\theta\cdot\gamma_t)^2}{1+\eta_1\cdot\gamma_t} = (1-\theta)\cdot(1-\theta+\theta\cdot\gamma_t)$, it follows that

$$\begin{aligned}
\Upsilon_{\text{market}}(\eta_1 = \frac{\theta}{1-\theta}) &= 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1\right) + \frac{\theta}{1-\theta} \cdot (1-\theta) \cdot \sum_{t=1}^{T} \alpha_t \cdot (1 - \theta \cdot (1 - \gamma_t))\,(1 - \gamma_t) \\
&= 1 + \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1\right) - \theta^2 \cdot \left(\sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_t} - 1\right) \\
&= 1.
\end{aligned}$$

As $\eta_1 \to \infty$,

$$\begin{aligned}
\lim_{\eta_1\to\infty} \Upsilon_{\text{market}} &= 1+\theta^2\cdot\left(\sum_{t=1}^{T}\frac{\beta_t^2}{\alpha_t}-1\right)+\lim_{\eta_1\to\infty}\left(\eta_1\cdot\Delta(\eta_1)\right)\\
&= 1+\theta^2\cdot\left(\sum_{t=1}^{T}\frac{\beta_t^2}{\alpha_t}-1\right)+\sum_{t=1}^{T}\frac{\alpha_t\cdot\left(1-\theta\cdot(1-\gamma_t)\right)^2(1-\gamma_t)}{\gamma_t}\\
&= 1+\theta^2\cdot\left(\sum_{t=1}^{T}\frac{\beta_t^2}{\alpha_t}-1\right)+(1-\theta)^2\cdot\left(\sum_{t=1}^{T}\frac{\alpha_t^2}{\beta_t}\right)-1+2\theta-\theta^2\cdot\left(\sum_{t=1}^{T}\frac{\beta_t^2}{\alpha_t}\right)\\
&= 1+(1-\theta)^2\cdot\left(\sum_{t=1}^{T}\frac{\alpha_t^2}{\beta_t}-1\right).
\end{aligned}$$

■

Cost ratio of single-stock trading. Note that

$$\begin{aligned}
f(\mathbf{e}_i) &= \left(\frac{\mathbf{e}_i^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{e}_i}{\left(\mathbf{w}_1^\top\bar{\mathbf{\Psi}}_{\text{id}}^{-1}\mathbf{e}_i\right)^2}\cdot\frac{1+\eta_1}{\bar{\psi}_{\text{f},1}}-1\right)^{-1}=\left(\frac{\bar{\psi}_{\text{id},i}^{-1}}{\left(w_{1i}\cdot\bar{\psi}_{\text{id},i}^{-1}\right)^2}\cdot\frac{1+\eta_1}{\bar{\psi}_{\text{f},1}}-1\right)^{-1}\\
&= \left(\frac{1+\eta_1}{\eta_{1,i}}-1\right)^{-1}=\frac{\eta_{1,i}}{1+\eta_1-\eta_{1,i}}.
\end{aligned}$$

Also note that

$$\frac{\eta_{1,i}}{1+\eta_1-\eta_{1,i}}\geq\frac{\eta_{1,j}}{1+\eta_1-\eta_{1,j}}\quad\text{if and only if}\quad\frac{w_{1i}^2}{\bar{\psi}_{\text{id},i}}\geq\frac{w_{1j}^2}{\bar{\psi}_{\text{id},j}}.$$

The results immediately follow from (34). ■