Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Seungki Min, Costis Maglaras, Ciamac C. Moallemi

To cite this article:
Seungki Min, Costis Maglaras, Ciamac C. Moallemi (2022) Cross-Sectional Variation of Intraday Liquidity, Cross-Impact, and Their Effect on Portfolio Execution. Operations Research

Published online in Articles in Advance 12 Jan 2022

. https://doi.org/10.1287/opre.2021.2201

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–17
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Contextual Areas

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

Seungki Min,[a] Costis Maglaras,[b] Ciamac C. Moallemi[b]

[a] Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea; [b] Graduate School of Business, Columbia University, New York, New York 10027

Contact: skmin@kaist.ac.kr, https://orcid.org/0000-0002-2887-1030 (SM); c.maglaras@gsb.columbia.edu, https://orcid.org/0000-0002-4283-2177 (CM); ciamac@gsb.columbia.edu, https://orcid.org/0000-0002-4489-9260 (CCM)

Received: November 13, 2018
Revised: July 21, 2020; June 18, 2021
Accepted: July 27, 2021
Published Online in Articles in Advance: January 12, 2022

Area of Review: Financial Engineering

https://doi.org/10.1287/opre.2021.2201

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 that (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 toward 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 because of 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 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 toward 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 data set of realized portfolio transaction records by exploiting the low-rank structure of its coefficient matrix suggested by our analysis.

Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2021.2201.

Keywords: portfolio management • optimal execution • market microstructure • market impact • factor model

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-β 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 toward 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 Section 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 one to two 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 because of the participation of natural portfolio liquidity providers. We formulate and solve a multiperiod 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 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 suboptimality 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.

1.1. Contributions

In detail, the main contributions of the paper are the following.

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 (nondiagonal) low-rank matrix where the diagonal components capture the effect of single-stock liquidity providers and the nondiagonal 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 multiperiod 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, it 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 2001, Tsoukalas et al. 2019), 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 because of 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, 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.

1.2. Linear Cross-Impact Model

In the derivation of the 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. Additionally, 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 nonlinear (e.g., square root or some other rational power), then although 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. 2017).

1.3. 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 (2001) 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 Roşu (2009), Alfonsi et al. (2010), and Obizhaeva and Wang (2013), treats the market as one limit order book and uses an aggregated and stylized model of market impact to capture how the price moves as a function of trading intensity. Tsoukalas et al. (2019) 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 multiasset 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 sublinear temporary impact. Tóth et al. (2011, 2018) report that the price impact at the metaorder level is a concave function of total order size, which is known as a square root impact law; see also Bucci et al. (2019) and Capponi and Cont (2019). Building these empirical findings, a number of impact models have been proposed, such as a transient impact model (Bouchaud et al. 2009, Gatheral et al. 2012), a history-dependent permanent impact model (Bouchaud et al. 2009), 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. (2017), 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 the 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 (Lo and Wang 2000, Hasbrouck and Seppi 2001) that adopt common factor models to analyze comovement in returns/trading volumes across stocks and attribute such a commonality to the portfolio order flows. Although 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., Karoli et al. 2012, Koch et al. 2016, Ben-David et al. 2017, 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]

1.4. 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 (nondiagonal) terms in market impact. We argue in Section 2 that the correlation in volume is attributable to the portfolio order flows and in Section 3 that the cross-asset impact is partially attributable to the liquidity provision at a portfolio level, both of which are interpreted as reflections of the portfolio investors' participation. We further motivate in Section 4 a parametric intraday variation of the cross-asset impact from the observed intraday pattern of volume correlation illustrated in Section 2. A more explicit connection is made in Section 5 based on a Poisson process analogy so as to provide a numerical illustration. We also discuss this issue in Online 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 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 data set contains 240 days ($D = 240$), denoted by d, excluding days that are known to exhibit abnormal trading activity, namely the Federal Reserve 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 database and extract all trades, excluding those (a) that occur before 09:35 or after 16:00; (b) opening auction prints or closing auction prints (trading condition code contains "O," "Q," "M," or "6"); and (c) trades corrected later (correction code is not zero or trading condition code contains "G" or "Z"). We divide a day into five-minute intervals ($T = 77$, 0935–0940 hours, …, 1555–1600 hours), denoted by t. We denote by DVol_{idt} the aggregate notional (dollars) 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 0935 and 1600 hours, excluding auction and corrected prints):

$$\overline{\mathrm{DVol}}_{it} \triangleq \frac{1}{D}\sum_{d=1}^{D}\mathrm{DVol}_{idt},\quad \mathrm{VolAlloc}_{it} \triangleq \frac{\overline{\mathrm{DVol}}_{it}}{\sum_{s=1}^{T}\overline{\mathrm{DVol}}_{is}},\quad \text{and}$$

$$\mathrm{AvgVolAlloc}_t \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathrm{VolAlloc}_{it}. \tag{1}$$

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:

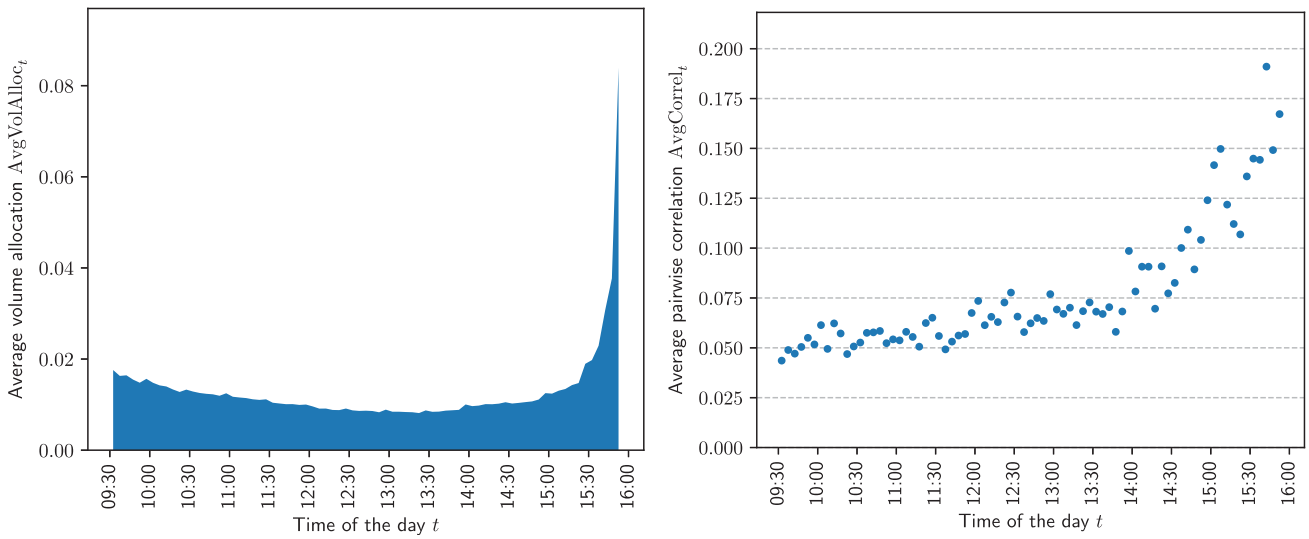
$$\mathrm{Correl}_{ijt} \triangleq \frac{\sum_{d=1}^{D}(\mathrm{DVol}_{idt} - \overline{\mathrm{DVol}}_{it})(\mathrm{DVol}_{jdt} - \overline{\mathrm{DVol}}_{jt})}{\sqrt{\sum_{d=1}^{D}(\mathrm{DVol}_{idt} - \overline{\mathrm{DVol}}_{it})^2 \cdot \sum_{d=1}^{D}(\mathrm{DVol}_{jdt} - \overline{\mathrm{DVol}}_{jt})^2}},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

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

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 that (ii) the cross-sectional average pairwise correlation increases significantly during the last few hours of the day.[4]

One possible explanation of the observed intraday volume correlation profile could be the nonstationary 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, $\mathrm{AvgCorrel}_t$ 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 toward 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

Figure 1. (Color online) Cross-Sectional Average Intraday Traded Volume Profile (Left Panel) and Cross-Sectional Average Pairwise Correlation (Right Panel): S&P 500 Constituent Stocks in 2018



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 to or at the end of the trading session.

We will return to these findings on $\mathrm{AvgVolAlloc}_t$ and $\mathrm{AvgCorrel}_t$ in Section 5 in order to approximate the relative magnitude of each different type of natural liquidity providers (portfolio versus single-stock 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 $\psi_{\mathrm{id},i}$. Single-stock investors will sell (or buy) $\psi_{\mathrm{id},i}$ 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 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 $\psi_{\mathrm{id},i}$ 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 $\Delta\mathbf{p} \in \mathbb{R}^N$. Let $\mathbf{e}_i$ be the ith standard basis vector. Single-stock investors on stock i will experience the price change $\mathbf{e}_i^\top \Delta\mathbf{p}$ and adjust their holding position by $-\psi_{\mathrm{id},i} \cdot \mathbf{e}_i^\top \Delta\mathbf{p}$. In vector representation, the change in the holding vector of single-stock investors $\Delta\mathbf{h}_{\mathrm{id}} \in \mathbb{R}^N$ can be written as

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

where $\mathbf{\Psi}_{\mathrm{id}} \triangleq \mathrm{diag}(\psi_{\mathrm{id},1}, \ldots, \psi_{\mathrm{id},N}) \in \mathbb{R}^{N\times N}$. The quantity $\Delta\mathbf{h}_{\mathrm{id}}(\Delta\mathbf{p})$ 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-β 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 $\psi_{\mathrm{f},k}$. Because 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 midprice 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 midprice 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, $\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top \Delta\mathbf{p}$ 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, $(\boldsymbol{\Psi}_\mathrm{id} + \mathbf{W}\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top)\Delta\mathbf{p}$ 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:

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

By Equations (4) and (5),

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

In other words, of $\mathbf{v}$ shares, $\boldsymbol{\Psi}_\mathrm{id}\Delta\mathbf{p} \in \mathbb{R}^N$ shares are obtained from single-stock investors and $\mathbf{W}\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top \Delta\mathbf{p} \in \mathbb{R}^N$ 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*

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

Note that the coefficient matrix $\mathbf{G}$ is an inverse of $\boldsymbol{\Psi}_\mathrm{id} + \mathbf{W}\boldsymbol{\Psi}_\mathrm{f}\mathbf{W}^\top$, 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 nondiagonal low-rank matrix. The following matrix expansion derived from an application of the Woodbury matrix identity will prove useful:

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

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

We interpret the terms $\boldsymbol{\Psi}_{\mathrm{id}} = \mathrm{diag}_{i=1}^{N}(\psi_{\mathrm{id},i})$ and $\boldsymbol{\Psi}_{\mathrm{f}} = \mathrm{diag}_{k=1}^{K}(\psi_{\mathrm{f},k})$ 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 $\boldsymbol{\Psi}_{\mathrm{id}} + \mathbf{W}\boldsymbol{\Psi}_{\mathrm{f}}\mathbf{W}^{\top}$ 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. Because $\psi_{\mathrm{id},i}$ 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 captures 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 5–15 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

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

where $\bar{\epsilon}^{\mathrm{tr}} \in \mathbb{R}^N$ represents a random error term that captures unpredictable market price fluctuations or the effect of trades executed in that period by other investors. 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 because of fluctuations in the price because of 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., $\mathsf{E}[\bar{\epsilon}^{\mathrm{tr}}|\mathbf{v}] = \mathbf{0}$). The single-period expected implementation shortfall incurred by the liquidator is given by

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

Linear price impact induces quadratic implementation shortfall costs; note that the resulting cost is always positive because $\mathbf{G}$ 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, $\boldsymbol{\Psi}_{\mathrm{id}}$ and $\boldsymbol{\Psi}_{\mathrm{f}}$, respectively, given by*

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

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*:

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

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,*

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

(The proof is provided in Online 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. Additionally, 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 Multiperiod 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 because of 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:

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

where the innovation term ϵ_t satisfies $\mathsf{E}[\epsilon_t | \mathcal{F}_{t-1}] = \mathbf{0}$ and $\mathcal{F}_{t-1}$ denotes all past information. The term ϵ_t is commonly understood as the change in a market participant's belief perhaps because of 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., because of 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. 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

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

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

Under these assumptions, the expected transaction cost of executing a series of portfolio transactions $\mathbf{v}_1, \ldots, \mathbf{v}_T$ is separable over time and can be expressed as follows:

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

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 Sections 4.2 and 5.1 and Online Appendix C.

3.5. Discussion on Model

The cross-impact model derived in this paper can be characterized as a special case of the multiasset version of the Almgren–Chriss model (Almgren and Chriss 2001, appendix A), where we represent the temporary impact with symmetric positive-definite matrices whose nondiagonal entries reflect the effect of portfolio liquidity provision. This would be the simplest form of the cross-impact model that achieves tractability and economic soundness at the same time. One may consider a more generalized form that possibly involves nonlinear, 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 Section 1 on the robustness of linearity assumption.

4. Optimal Portfolio Execution

We will formulate and solve the multiperiod optimal portfolio execution problem in Section 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 Section 4.2.

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:

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

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

Proposition 3 ("Coupled" Execution). *The risk-neutral cost minimization Problems* (16) *and* (17) *have a unique optimal solution given by*

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

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 was 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 toward 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 2001, Tsoukalas et al. 2019). 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 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, $\mathbf{\Psi}_{\mathrm{f},t} = \mathbf{0}$ 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 because $\mathbf{G}_t$ 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 conditions of the convex minimization problem in (16) and (17) require that there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^N$ such that

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

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 $\mathbf{v}_t^{*} = \mathbf{G}_t^{-1} \boldsymbol{\lambda} = \mathbf{G}_t^{-1} \left(\sum_{s=1}^{T} \mathbf{G}_s^{-1} \right)^{-1} \mathbf{x}_0$. Because 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, and 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: that is,

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

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, that all index funds share the profile β_t. The empirical findings of Section 2 indicate that pairwise correlations of trading volumes increase toward 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 vary differently from each other so as to generate the time-varying pairwise correlation relationship; this is supported by the behavior of market participants toward 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 toward 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):

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

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 $\widehat{\mathbf{W}} \triangleq \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1} \mathbf{W} (\bar{\mathbf{\Psi}}_{\mathrm{f}}^{-1} + \mathbf{W}^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1} \mathbf{W})^{-1}$ *and* $\widehat{\mathbf{w}}_k$ *denotes the kth 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 Online 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, this 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 VWAP algorithm. As we show in Section 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 Section 4.2 simplified to the case of a single index fund (e.g., the market portfolio), we establish a worst case bound on the suboptimality gap of such a separable execution schedule against the optimized portfolio execution schedule derived.

Specifically, restricting attention to the parameterization introduced in Section 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 $\widehat{\mathbf{w}}_1 \triangleq (\bar{\psi}_{\mathrm{f},1}^{-1} + \mathbf{w}_1^\top \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1} \mathbf{w}_1)^{-1} \bar{\mathbf{\Psi}}_{\mathrm{id}}^{-1} \mathbf{w}_1$. 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).

Section 5.1 (and in more detail, Online 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 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 AvgVolAlloc_t and AvgCorrel_t, discussed in Section 2. Section 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 Section 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 Online 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 Online Appendix D.3), the intraday volume and pairwise correlation profiles are given by

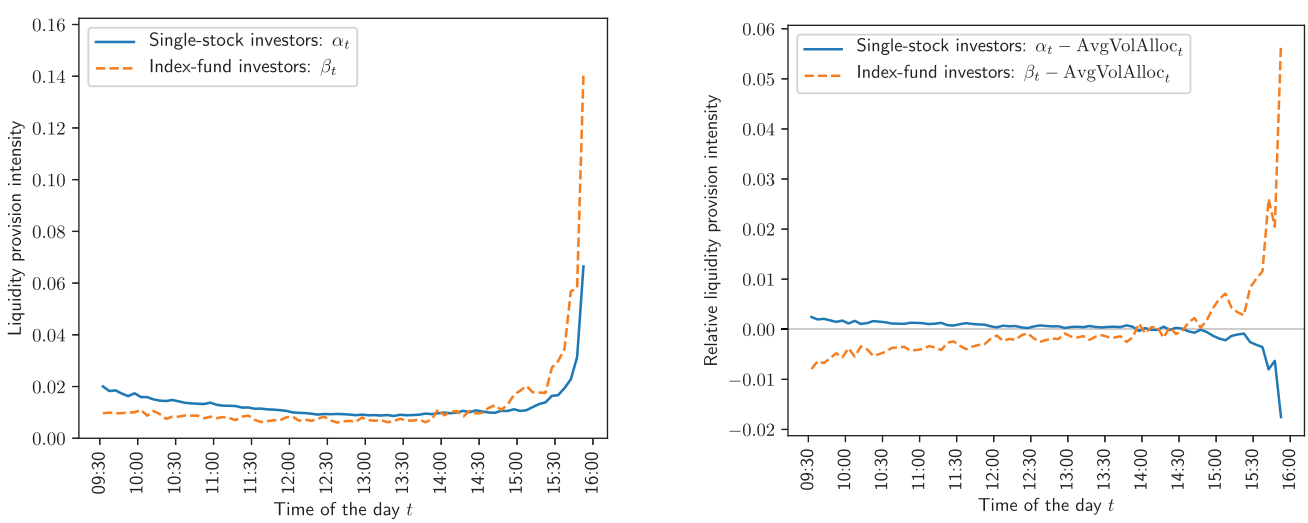
$$\text{AvgVolAlloc}_t = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathsf{E}[\text{DVol}_{idt}]}{\sum_{s=1}^{T} \mathsf{E}[\text{DVol}_{ids}]} = \alpha_t \cdot (1-\theta) + \beta_t \cdot \theta, \tag{28}$$

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

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.

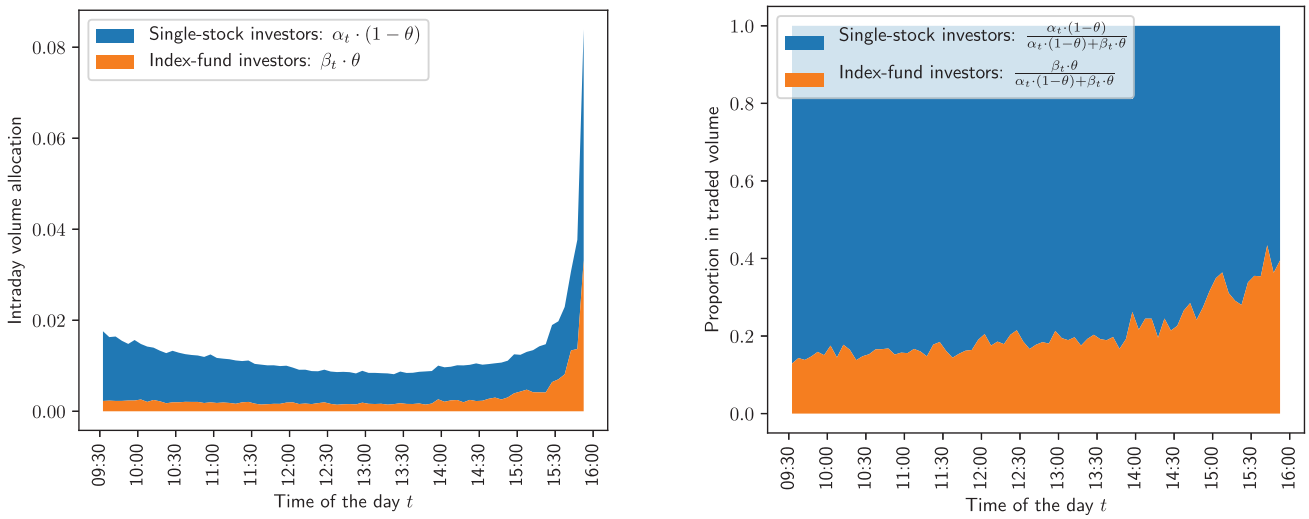
Given the empirically observed profiles for AvgVolAlloc_t and AvgCorrel_t, illustrated in Figure 1, we can solve a set of coupled equations defined by (28) and (29), where the respective left-hand sides are given by the empirically estimated values so as to identify the values of θ, $\alpha_1, \ldots, \alpha_T$, and $\beta_1, \ldots, \beta_T$. The results are summarized in Figures 2 and 3. We can observe that θ is estimated to be 0.24, 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,

Figure 2. (Color online) Intensity of Single-Stock Investors, α_t, and Index-Fund (Portfolio) Investors, β_t, Calibrated to Best Match Empirical Profiles of Traded Volume, AvgVolAlloc_t, and Pairwise Volume Correlations, AvgCorrel_t (Left Panel) and Deviation of α_t, β_t from the Market Profile, AvgVolAlloc_t (Right Panel)



Note. The data of the year 2018 are used as in Section 2.

Figure 3. (Color online) Decomposition of Average Volume Allocation, AvgVolAlloc_t, into Single-Stock Investors' Contribution, $\alpha_t \cdot (1-\theta)$, and Index-Fund Investors' Contribution, $\beta_t \cdot \theta$ (Left Panel) and Their Proportions (Right Panel)

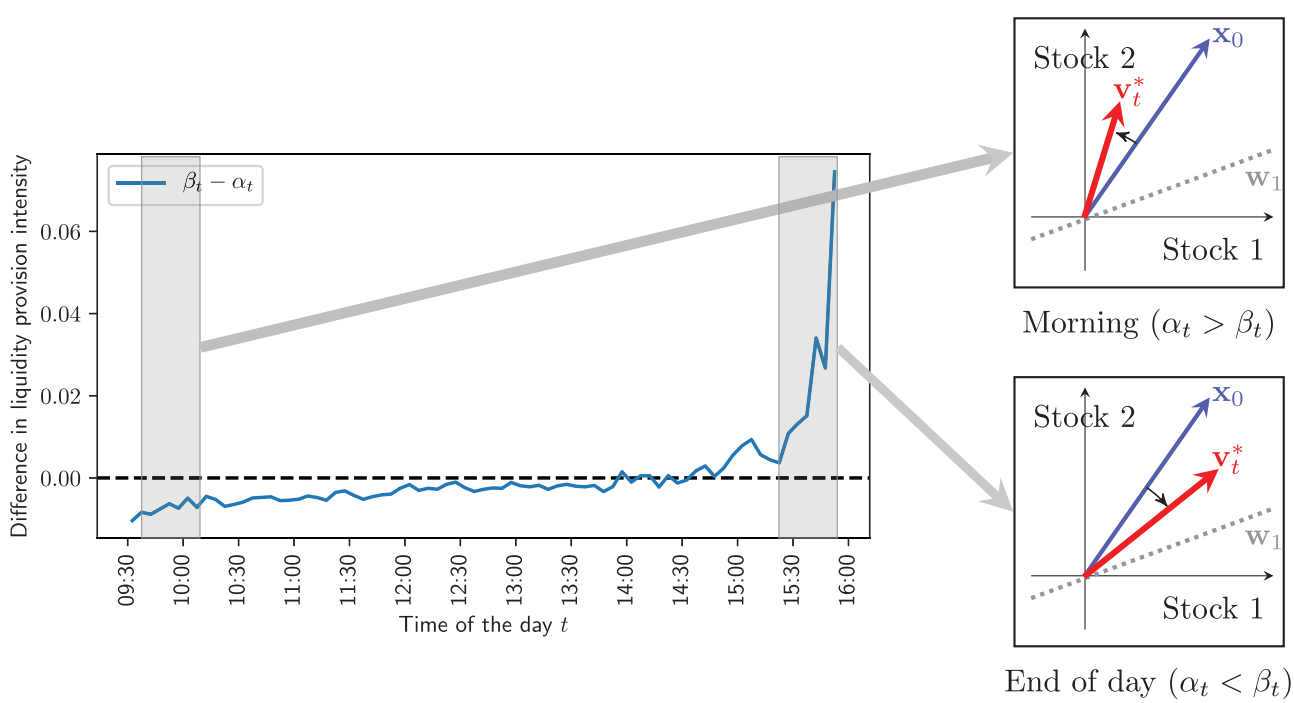


Note. In this generative model, every security is assumed to have the same intraday profiles.

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 toward 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

Figure 4. (Color online) 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, $\beta_t - \alpha_t$



portfolio executed toward the end of the day is better aligned with the index portfolio (in the 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 $\mathbf{v}_t^{\text{sep}}$ is given by

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

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

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

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

Note that under the assumption that there is only one index fund, $\mathbf{w}_1$, we can simplify these 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 one 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* $\mathbf{x}_0 \in \mathbb{R}^N$,

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

where

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

(The proof is given in Online 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 of 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 $\mathbf{w}_1$.

5.2.1. 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:

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

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

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

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

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

This remark identifies that 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

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

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 toward 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.

5.2.2. 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,

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

For particular values of η_1,

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

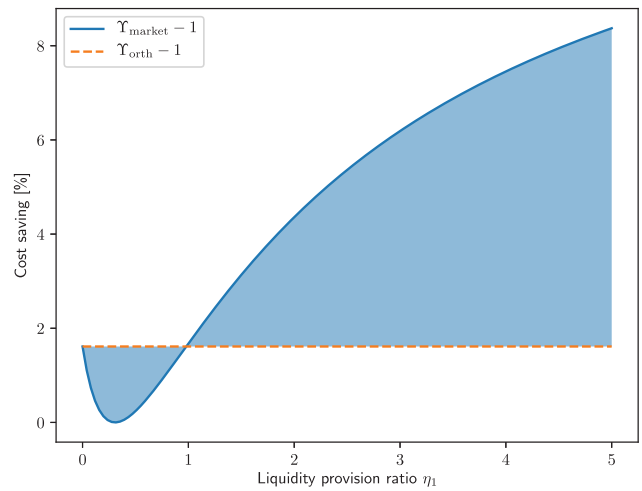
$$\Upsilon_{\text{market}}\left(\eta_1 = \frac{\theta}{1-\theta}\right) = 1, \tag{43}$$

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

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 $\frac{\theta}{1-\theta}$.

Our estimate of the fraction of index-fund liquidity, $\theta = 0.24$, suggests a threshold value of $\theta/(1-\theta) \approx 0.31$. 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 the realized benefits from using optimal versus 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 Section 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 because of portfolio liquidity provision.

Figure 5. (Color online) 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 Section 5.1



Note. Compared with the separable execution, the coupled execution can save up to 14.0% when trading the market portfolio.

5.2.3. 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 $\eta_{1,i} \triangleq \frac{w_{1i}^2 \cdot \bar{\psi}_{\ell,1}}{\bar{\psi}_{\text{id},i}}$. We can further identify the stock that induces the largest cost ratio:

$$\underset{i=1,\ldots,N}{\arg\max}\{\Upsilon(\mathbf{x}_0 = \mathbf{e}_i)\} = \begin{cases} \arg\max_{i=1,\ldots,N}\left\{\dfrac{w_{1i}^2}{\bar{\psi}_{\text{id},i}}\right\} & \text{if } \Delta \ge 0 \\ \arg\min_{i=1,\ldots,N}\left\{\dfrac{w_{1i}^2}{\bar{\psi}_{\text{id},i}}\right\} & \text{if } \Delta \le 0. \end{cases} \tag{46}$$

Here, we are comparing the performance implications of liquidating a single order using a separable execution schedule versus 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

6.1. 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 because of the high dimensionality of the unknown coefficient matrix (an $N \times N$ matrix) and because the underlying data tend 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 Section 3, which would take as input a large set of proprietary portfolio transactions.

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

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

where a diagonal matrix $\tilde{\mathbf{\Psi}}_{\mathrm{id},dt} \in \mathbb{R}^{N\times N}$ describes the liquidity provided by single-stock investors, a diagonal matrix $\tilde{\mathbf{\Psi}}_{\mathrm{f},dt} \in \mathbb{R}^{K\times K}$ describes the liquidity provided by index-fund investors, and the noise term $\bar{\epsilon}_{dt}^{\mathrm{tr}} \in \mathbb{R}^N$ 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 $\widehat{\sigma}_{\mathrm{id},idt}$ ($\widehat{\mathrm{DVol}}_{\mathrm{f},kdt}$ and $\widehat{\sigma}_{\mathrm{f},kdt}$, respectively) are some forecasted trading volume and volatility of the stock i (the index fund k, respectively) generated by the single-stock investors (the index-fund investors, respectively), and γ_{id}'s 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 Online Appendix C, we motivate this 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 data set.

6.2. 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).

Sections 4 and 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 toward 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 toward 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 relevant to such a constrained formulation.[11]

6.3. 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 (2001, 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.

Endnotes

[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.

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

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

[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 subuniverses).

[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 (2001).

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

[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, whereas 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.

[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 $\gamma_{\text{id}}^{-1} \times \widehat{\sigma}_{idt}(\tilde{v}_{idt}/\widehat{\text{DVol}}_{idt})$); 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.

[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.

References

Agarwal V, Hanouna P, Moussawi R, Stahel CW (2018) Do ETFs increase the commonality in liquidity of underlying stocks? Preprint, submitted July 13, 2017; revised December 11, 2018, https://dx.doi.org/10.2139/ssrn.3001524.

Alfonsi A, Fruth A, Schied A (2010) Optimal execution strategies in limit order books with general shape functions. *Quant. Finance* 10(2):143–157.

Almgren R (2003) Optimal execution with nonlinear impact functions and trading-enhanced risk. *Appl. Math. Finance* 10(1):1–18.

Almgren R, Chriss N (2001) Optimal execution of portfolio transactions. *J. Risk* 3:5–40.

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

Ben-David I, Franzoni FA, Moussawi R (2017) Exchange-traded funds. *Annual Rev. Financial Econom.* 9:169–189.

Benzaquen M, Mastromatteo I, Eisler Z, Bouchaud J-P (2017) Dissecting cross-impact on stock markets: An empirical analysis. *J. Statist. Mech.* 2017(2):023406.

Bertsimas D, Lo AW (1998) Optimal control of execution costs. *J. Financial Marketing* 1(1):1–50.

Bouchaud J-P, Farmer JD, Lillo F (2009) How markets slowly digest changes in supply and demand. Hens T, Schenk-Hoppé KR, eds. *Handbook of Financial Markets: Dynamics and Evolution, Handbooks in Finance* (Elsevier, Amsterdam), 57–156.

Brown DB, Carlin B, Lobo MS (2010) Optimal portfolio liquidation with distress risk. *Management Sci.* 56(11):1997–2014.

Bucci F, Mastromatt I, Benzaquen M, Bouchaud J-P (2019) Impact is not just volatility. *Quant. Finance* 19(11):1763–1766.

Capponi F, Cont R (2019) Trade duration, volatility and market impact. Preprint, submitted April 8, https://dx.doi.org/10.2139/ssrn.3351736.

Curato G, Gatheral J, Lillo F (2017) Optimal execution with non-linear transient market impact. *Quant. Finance* 17(1):41–54.

Donier J, Bonart J, Mastromatteo I, Bouchaud J-P (2015) A fully consistent, minimal model for non-linear market impact. *Quant. Finance* 15(7):1109–1121.

Driebusch C, Osipovich A, Zuckerman G (2018) What's the biggest trade on the New York stock exchange? The last one. *Wall Street Journal* (March 14), https://www.wsj.com/articles/at-closing-time-the-stock-market-heats-up-like-a-bar-at-last-call-1521038300.

Gatheral J, Schied A, Slynko A (2012) Transient linear price impact and Fredholm integral equations. *Math. Finance* 22(3):445–474.

Hasbrouck J, Seppi DJ (2001) Common factors in prices, order flows, and liquidity. *J. Financial Econom.* 59(3):382–411.

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

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

Huberman G, Stanzl W (2005) Optimal liquidity trading. *Rev. Finance* 9(2):165–200.

Karoli GA, Lee K-H, Van Dijk MA (2012) Understanding commonality in liquidity around the world. *J. Financial Econom.* 105(1):82–112.

Koch A, Ruenzi S, Starks L (2016) Commonality in liquidity: A demand-side explanation. *Rev. Financial Stud.* 29(8):1943–1974.

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

Lo AW, Wang J (2000) Trading volume: Definitions, data analysis, and implications of portfolio theory. *Rev. Financial Stud.* 13(2):257–300.

Mastromatteo I, Benzaquen M, Eisler Z, Bouchaud J-P (2017) Trading lightly: Cross-impact and optimal portfolio execution. Preprint, submitted April 10, https://dx.doi.org/10.2139/ssrn.2949748.

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

Roşu I (2009) A dynamic model of the limit order book. *Rev. Financial Stud.* 22(11):4601–4641.

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

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

Tomas M, Mastromatteo I, Benzaquen M (2020) How to build a cross-impact model from first principles: Theoretical requirements and empirical results. Preprint, submitted April 29, https://dx.doi.org/10.2139/ssrn.3567815.

Tóth B, Eisler Z, Bouchaud J-P (2018) The short-term price impact of trades is universal. *Market Microstructure Liquidity* 3(2):1850002.

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

Tsoukalas G, Wang J, Giesecke K (2019) Dynamic portfolio execution. *Management Sci.* 65(5):2015–2040.

Seungki Min is an assistant professor in the Department of Industrial and Systems Engineering at Korea Advanced Institute of Science and Technology. His research interests are focused on sequential decision-making problems arising in modern business applications such as algorithmic trading and online advertisement.

Costis Maglaras is Dean and the David and Lyn Silfen Professor of Business at Columbia Business School. His current research centers on stochastic modeling and data science, with an emphasis on stochastic networks, financial engineering, and quantitative pricing and revenue management.

Ciamac C. Moallemi is the William von Mueffling Professor of Business in the Decision, Risk, and Operations Division of the Graduate School of Business at Columbia University. His research interests are in the development of mathematical and computational tools for optimal decision making under uncertainty, with a focus on applications areas including market microstructure, quantitative and algorithmic trading, and blockchain technology.