IMES DISCUSSION PAPER SERIES

Optimal Execution of Multiasset Block Orders under Stochastic Liquidity

Naoki Makimoto and Yoshihiko Sugihara

Discussion Paper No. 2010-E-25

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

2-1-1 NIHONBASHI-HONGOKUCHO
CHUO-KU, TOKYO 103-8660
JAPAN

IMES Discussion Paper Series 2010-E-25
November 2010

Optimal Execution of Multiasset Block Orders under Stochastic Liquidity

Naoki Makimoto* and Yoshihiko Sugihara**

Abstract

In this paper, we develop a multiasset model of market liquidity and derive the optimal strategy for block order execution under both liquidity and volatility risk. The market liquidity flowing into and out of an order book is modeled as a mean-reverting stochastic process. Given the shape of the order book for each asset, we express the market impact of an execution as a recursive impact that recovers gradually with associated uncertainty. We then derive the optimal execution strategy as a closed-form solution to the mean–variance problem that optimizes the trade-off between the market impact and the volatility/liquidity risk given investor risk aversion. Using our model, we analyze some implications of the optimal execution strategy with comparative statics and simulations. We also discuss whether we avoid price manipulation with our optimal execution strategy.

Keywords: optimal execution strategy; market impact; transaction cost; stochastic liquidity; limit order book; price manipulation; mean-variance optimization
JEL classification: C61, G11, G12

***** Professor, Graduate School of Business Sciences, University of Tsukuba
(E-mail: makimoto@gssm.gsbs.tsukuba.ac.jp)
****** Deputy Director and Economist, Institute for Monetary and Economic Studies, Bank of Japan (E-mail: yoshihiko.sugihara@boj.or.jp)

This paper was prepared in part while Naoki Makimoto was a visiting scholar at the Institute for Monetary and Economic Studies, Bank of Japan. He would like to acknowledge the generous hospitality and research support extended. The authors would also like to thank Masamitsu Ohnishi at Osaka University, participants in financial modeling and analysis seminar at Kyoto University, and research staff at the Bank of Japan for helpful discussion and comments. Views expressed in this paper are those of the authors and do not necessarily reflect the official views of the Bank of Japan.

Contents

1 Introduction

Executing a large volume of block securities under fluctuating market liquidity may cause a significant market impact resulting in large transaction costs. The importance of transaction costs has been recognized since at least the 1980s, when Perold [1988] introduced the concept of implementation shortfall. Financial crises have also drawn much attention to the dramatic increase in transaction costs under declining market liquidity. As a result, since the 1990s, various models of asset price processes with market impacts have been developed to control for execution costs, including Bertsimas and Lo [1998], Almgren and Chriss [2001], Konishi and Makimoto [2001], Subramanian and Jarrow [2001], Kissell, Glantz, and Malamut [2004], Obizhaeva and Wang [2005], and Engle and Ferstenberg [2007], Alfonsi, Fruth, and Schied [2010], from which optimal execution strategies have been constructed.

Financial institutions, even in ordinary market periods, have applied some of these models to minimize execution costs. Several empirical studies, such as Chen *et al.* [2004] and Xuemin [2008], have shown that the size of mutual funds erodes the performance of these models, which implies that execution costs are considered to account for a large part of the total cost of portfolio management. This is problematic in that these funds necessarily face the need for block trades for rebalancing purposes. Furthermore, the minimization of execution costs is one of the tasks currently undertaken by financial institutions, with practitioners attempting to apply optimal execution strategies as an algorithmic trading schedule. These have rapidly developed given the availability of high-frequency market data and improvements in information technology, combined with technologies that search for off-exchange liquidity. Johnson [2010] surveys recent developments in these practices and their related technologies.

We propose an approach that minimizes execution costs involved in multiasset trades while controlling for the risk of liquidity (or the fluctuation in prices). To start with, we develop a model of market liquidity employing a mean-reverting stochastic process. We interpret this liquidity model as the fluctuating number of orders on a limit order book, or the queue waiting for an execution. Next, we solve a static problem that optimizes the trade-off between the market impact and volatility and liquidity risk to obtain an optimal execution strategy. The model assumes a multiasset environment, and we intend that it will be used by a portfolio manager handling a basket of trades. Finally, we analyze the performance of this optimal strategy by comparing it with strategies derived from earlier studies omitting uncertainty in liquidity. We also examine the behavior of the strategy in relation to the price or liquidity correlations among assets.

Several existing studies possess a similar motivation to our analysis. For instance, Almgren and Chriss [2001] obtained an optimal execution strategy that optimizes the trade-off between the market impact and price volatility costs by assuming both a temporal market impact that recovers promptly after executions and a permanent impact that does not. Similarly, Konishi and Makimoto [2001] extended their model and obtained optimal slices of a block trade by solving a mean–standard deviation problem. Conversely, Kissell, Glantz, and Malamut [2004], proposed a model that divides and allocates the market impact into permanent and temporary components in a top-down manner. Obizhaeva and Wang [2005] also considered an optimal execution strategy by assuming recursive liquidity in a block-shaped limit order book, where the price diffuses and recovers gradually after execution. Finally, Alfonsi, Fruth, and Schied [2010] extended Obizhaeva and Wang [2005] to derive an optimal strategy where the limit order book has a general but continuous and differentiable shape.

We provide an analytical solution to the problem of execution cost minimization that incorporates uncertainty in market liquidity. Our approach therefore extends the model in Obizhaeva and Wang [2005] by adding uncertainty in liquidity, while our minimization problem is almost the same as Almgren and Chriss [2001]. However, unlike Obizhaeva and Wang [2005], queuing theory provides the inspiration for our model construction. Our approach is therefore an extended version of their model but with an important key difference. Namely, the construction process suggests that uncertainty be added to Obizhaeva and Wang [2005] for liquidity recovery after each execution; we assume the liquidity after the execution asymptotically declines at random to the liquidity level in place before the execution. We provide an analytical solution to the problem that can be used in the pretrade analysis of transactions to see how fluctuating liquidity or volatility affects the optimal execution strategies.

One of the other defining characteristics of our analysis is that our strategy incorporates the correlations between prices and the liquidity volatilities of assets. These correlations play an important role in the cost of multiasset executions. Suppose, for instance, that a portfolio manager trades a basket of assets with a time horizon for portfolio rebalancing. The manager should then incorporate the impact of trades in one asset on the prices and liquidity of other assets. Borkovec and Heidle [2010] pointed out this issue, indicating that most execution cost models are not perfectly designed to handle multiasset trades. We partly overcome this problem.

By the numerical analyses of optimal strategies, we obtain various implications. For the single-asset case, we confirm that execution velocity should increase as liquidity increases, as the investor becomes more risk averse, or as the volatility/resilience of liquidity rises. For the multiasset case, we find that the investor should sell/buy a liquid asset more quickly as the correlation of liquidity or the fundamental price increases among assets while keeping the execution velocity of an illiquid asset unchanged. Some of these features are not evident in earlier studies, and this leads to useful insight into multiasset trading. We investigate these results from the perspective of an efficient frontier of trading strategies.

From a theoretical perspective, we should construct these execution strategies to avoid the possibility of price manipulation. Huberman and Stanzl [2004] define this mathematically as round-trip trade with a negative cost—which is conceptually different from the market manipulation restricted by law—with Gatheral [2008] and Alfonsi and Schied [2010] pointing out that the model in Obizhaeva and Wang [2005] may violate the no-price-manipulation condition if the liquidity decay is a nonconvex function of time. In this regard, we confirm that our optimal strategy does not violate the no-price-manipulation condition in the single-asset case, but it may contain round-trip trades in the multiasset case.

The remainder of this paper is structured as follows. Section 2 defines the model of market prices and the liquidity recovery after execution and discusses its relation with earlier works. Section 3 derives the optimal execution strategy based on our model. Section 4 analyzes the optimal strategy using comparative statics and efficient frontiers of execution costs. Section 5 provides a summary of the paper.

2 Model and optimal execution problem

2.1 The model

We consider a discrete execution schedule at a regular time interval τ. Suppose an investor (or portfolio manager) has N assets and $K+1$ of time slots for execution starting from the current time 0 to a finite time horizon T. Let $t_k = k\tau$ $(k = 0, \ldots, K)$ denote the time just before the k-th execution where $t_0 = 0$ and $t_K = T$. Given a block of buy orders with a total size ω_i for an asset i $(i = 1, \ldots, N)$,[1] we will determine the volume of slice orders $\xi_0^i, \ldots, \xi_K^i$, where ξ_k^i denotes the volume of asset i to be executed as a market order in the k-th time slot, and $\sum_{k=0}^{K} \xi_k^i = \omega_i$. We consider the optimal execution in a limit order book market given the total purchasing volume of ω_i.[2]

First, we build the following model, which describes the dynamic behavior of asset prices and limit order books. Let P_t^i and A_t^i be the fundamental price and the best ask price of asset i at time t, respectively. The fundamental price is given by a random walk:

$$P_{t_k}^i = P_0^i + \sum_{s=1}^{k} \Delta_s^i, \quad (k = 1, \ldots, K,\ i = 1, \ldots, N), \tag{2.1}$$

where P_0^i is the current fundamental price.[3] The increment $\{\Delta_s^i\}$ obeys a stationary process with a finite variance, which satisfies $\mathrm{E}\left[\Delta_s^i\right] = 0$ for any s and i. The fundamen-

[1] We use super/subscript i to denote the variable for asset i.

[2] The optimal execution of selling ω_i of a block order can be analyzed in a similar manner.

[3] Here, we assume the investor does not possess superior information about the fundamental value of the asset.

tal price indicates the intrinsic value of the asset irrelevant to its liquidity. We assume that the covariance matrix of $\Delta_s^1, \ldots, \Delta_s^N$, denoted as $\mathbf{\Sigma}_\Delta = [\sigma_\Delta^{ij}]$, is independent of s, and that Δ_s^i and Δ_l^j $(l \neq s)$ are uncorrelated.

Second, we consider the volume of sell limit orders on the limit order book of an asset, interpreted as the current liquidity level of the asset. Let M_t^i denote the volume of sell limit orders placed above A_t^i in the order book of asset i. Let $f_i(p)$ denote an order volume density of existing or potential limit orders at price level $P_t^i + p$ for asset i. Here, $f_i(p)dp$ indicates the volume of limit orders placed on the price level between $P_t^i + p$ and $P_t^i + p + dp$. A new order is supposed to be placed in such a way that the order fills up the density down from A_t^i marginally. In other words, if a new sell limit order is placed just below A_t^i, A_t^i consequently moves down. The total volume of sell limit orders between P_t^i and $P_t^i + p$ is given by $F_i(p) = \int_0^p f_i(q)dq$. The volume of sell limit orders, M_t^i, can be written in terms of $F_i(p)$ as $M_t^i = F_i(\infty) - F_i(A_t^i - P_t^i)$. Here, we assume that the total volume of limit orders on the book is finite; i.e., $F_i(\infty) < \infty$ for any i. We also define a volume V_t^i by $V_t^i = F_i(A_t^i - P_t^i)$. V_t^i is interpreted as the volume of potential sell limit orders that are not placed in the order book at time t but will be placed from A_t^i down to P_t^i after time t.

Third, we model the time development of market liquidity M_t^i. We simply assume that M_t^i follows a mean-reverting stochastic process with a fixed variance as:

$$dM_t^i = (\nu_i - \mu M_t^i)dt + d\tilde{Z}_t^i, \quad (i = 1, \ldots, N), \tag{2.2}$$

where $d\tilde{Z}_t^i$ represents stationary and independent stochastic increments with a finite variance, and ν_i, μ are parameters; ν_i is assumed to be different from asset to asset while μ is common for all assets.[4] The mean-reversion property of M_t^i in Eq.(2.2) represents the effect that the gap between the best ask price A_t^i and the fundamental price P_t^i, which is generated by the preceding trades, is gradually resolved. Further assuming that $F_i(\infty)$ equals the stationary level of liquidity, i.e., $F_i(\infty) = \nu^i/\mu$, indicating that the mean-reversion level of the best ask price corresponds to the fundamental price, Eq.(2.2) is rewritten in terms of V_t^i as:

$$dV_t^i = -\mu V_t^i dt - d\tilde{Z}_t^i, \quad (i = 1, \ldots, N). \tag{2.3}$$

The model is interpreted as a queue waiting for an execution. Suppose a new sell limit order arriving in an order book obeys a Poisson process at a rate $\nu(M_t^i)$, while existing limit orders on the book are supposed to be canceled or executed by other traders at a rate $\mu(M_t^i)$. Both the order arriving rate ν and the cancel/execution rate

[4] Under the above conditions, A_t^i can fall below P_t^i. Such a case is the situation where the ex post market impacts are extreme, and as a result, the price stays below the fundamental price. However, even if this situation occurs, it does not affect the following calculations of the execution costs and optimal execution strategies.

μ are a function of the number of existing orders M_t^i. The configuration of the model is illustrated in Figure 1 (a). Given that the change in the process of M_t^i during the small time interval dt between execution time slots is written as:

$$\begin{cases} \Pr\left(M_{t+dt}^i - M_t^i = +1 \mid M_t^i\right) = \nu(M_t^i)dt + o(dt), \\ \Pr\left(M_{t+dt}^i - M_t^i = -1 \mid M_t^i\right) = \mu(M_t^i)dt + o(dt), \end{cases}$$

the mean and the variance of the change in the order volume are simply computed as:

$$\begin{cases} \mathrm{E}\left[M_{t+dt}^i - M_t^i \mid M_t^i\right] = \{\nu(M_t^i) - \mu(M_t^i)\}dt + o(dt), \\ \mathrm{V}\left[M_{t+dt}^i - M_t^i \mid M_t^i\right] = \{\nu(M_t^i) + \mu(M_t^i)\}dt + o(dt). \end{cases}$$

The simplest case of rate functions is when $\nu(M_t^i) = \nu_i$ and $\mu(M_t^i) = \mu M_t^i$ that organizes the $M/M/\infty$ queue. Here, the arriving rate ν_i is assumed to be different from asset to asset, while the cancel/execution rate μ is common for all assets. In this case, the process M_t^i is mean reverting:

$$\mathrm{E}\left[M_{t+dt}^i - M_t^i \mid M_t^i\right] = (\nu_i - \mu M_t^i)dt + o(dt),$$

which constitutes the same mean as the model in Eq.(2.2). The variance of the process in this case differs a little from Eq.(2.2) as the variance also depends on M_t^i.

The other interpretation is when $\nu(M_t^i) = (\kappa - M_t^i)\tilde{\nu}$ and $\mu(M_t^i) = \tilde{\mu} M_t^i$, which configures the finite source queue, where $\tilde{\nu}$, $\tilde{\mu}$, and κ are considered to be the rate of order submission per market participant, the diminishing rate of existing orders, and the number of potential participants, respectively. In this case, M_t^i is considered to be a source of an order submission, or the number of market participants. When we assume that $\tilde{\nu} = \tilde{\mu}$, the finite source queue generates a mean-reverting process with a fixed variance such that:

$$\begin{cases} \mathrm{E}\left[M_{t+dt}^i - M_t^i \mid M_t^i\right] = (\kappa - 2M_t^i)\tilde{\nu}dt + o(dt), \\ \mathrm{V}\left[M_{t+dt}^i - M_t^i \mid M_t^i\right] = \tilde{\nu}\kappa dt + o(dt). \end{cases}$$

This reduces to Eq.(2.2) by letting $\tilde{\nu} = \mu/2$ and $\kappa = 2\nu/\mu$.

Let $a = e^{-\mu\tau}$ and $Z_{t_k}^i = -\int_{t_{k-1}+}^{t_k} e^{-\mu(t_k - t)} d\tilde{Z}_t^i$, where t_k+ denotes the time just after the k-th execution. Integrating Eq.(2.3) from time t_k+ to t_{k+1} yields $V_{t_{k+1}}^i = aV_{t_k+}^i + Z_{t_{k+1}}^i$ within any execution time slot k. Note that $\mathrm{E}\left[Z_t^i\right] = 0$ and $\mathrm{V}\left[Z_t^i\right] < \infty$. We know that $V_{t_k+}^i = V_{t_k}^i + \xi_k^i$ as our market order in time slot t_k ($k = 0, \ldots, K$) consumes existing limit orders M_t^i. In what follows, we rewrite for simplicity $V_{t_k}^i$, $Z_{t_k}^i$, $A_{t_k}^i$, and $P_{t_k}^i$ as V_k^i, Z_k^i, A_k^i, and P_k^i, respectively. Consequently, the time development of V_k is expressed as:

$$V_{k+1}^i = a(V_k^i + \xi_k^i) + Z_{k+1}^i, \quad (i = 1, \ldots, N,\ k = 0, \ldots, K-1). \tag{2.4}$$

We also assume that the initial state equals the stationary state of the dynamics; hence, $V_0^i = 0$ for any asset i. The variance–covariance matrix of $Z_k^1, \ldots, Z_k^N$ is denoted as $\mathrm{Cov}\left(Z_k^i, Z_k^j\right) = [\sigma_Z^{ij}] = \mathbf{\Sigma}_Z$ independently from k. We further assume that Z_k^i and Z_l^j are uncorrelated for any $k \neq l$, and $\{Z_k^i;\ k = 1, \ldots, K\}$ and $\{\Delta_k^i;\ k = 1, \ldots, K\}$ are uncorrelated for any i. The dynamics of our model are described in Figure 1 (b).

2.2 Execution costs

This section considers execution costs. Because a buy execution consumes existing orders, purchasing ξ_k^i of asset i by a market order makes the price rise up to the level of A_{k+}^i after the purchase. Because $A_k^i = F_i^{-1}(V_k^i) + P_k^i$ from the definition of F_i, the cost of purchasing ξ_k^i of asset i, denoted as C_k^i, is given by:

$$C_k^i = \int_{A_k^i}^{A_{k+}^i} p\tilde{f}_i(p,k)dp = \int_{F_i^{-1}(V_k^i)+P_k^i}^{F_i^{-1}(V_k^i+\xi_k^i)+P_k^i} p\tilde{f}_i(p,k)dp,$$

where $\tilde{f}_i(p,k) = f_i(p - P_k^i)$ denotes the order density at the actual price level. This formulation of execution costs also applies in Alfonsi, Fruth, and Schied [2010]. This is further transformed into:

$$C_k^i = P_k^i\xi_k^i + \int_{F_i^{-1}(V_k^i)}^{F_i^{-1}(V_k^i+\xi_k^i)} pf_i(p)dp = P_k^i\xi_k^i + G_i(V_k^i + \xi_k^i) - G_i(V_k^i), \tag{2.5}$$

where:

$$G_i(v) = \int_0^{F_i^{-1}(v)} pf_i(p)dp,$$

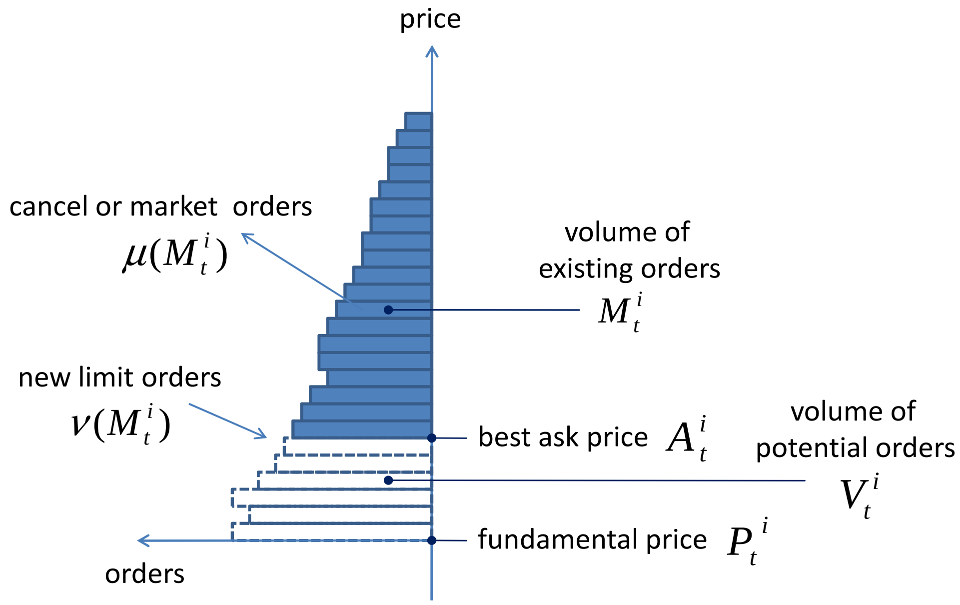
denotes the average price of purchasing asset i above the fundamental price P_k^i.

According to Perold [1988] and Kissell [2006], the implementation shortfall of the execution is defined as the difference between the actual portfolio returns and the portfolio returns based on a portfolio manager's decision price. Suppose the portfolio manager decides execution based on the fundamental price just before the liquidation begins. We define the implementation shortfall of the i-th asset, IS_i, as:

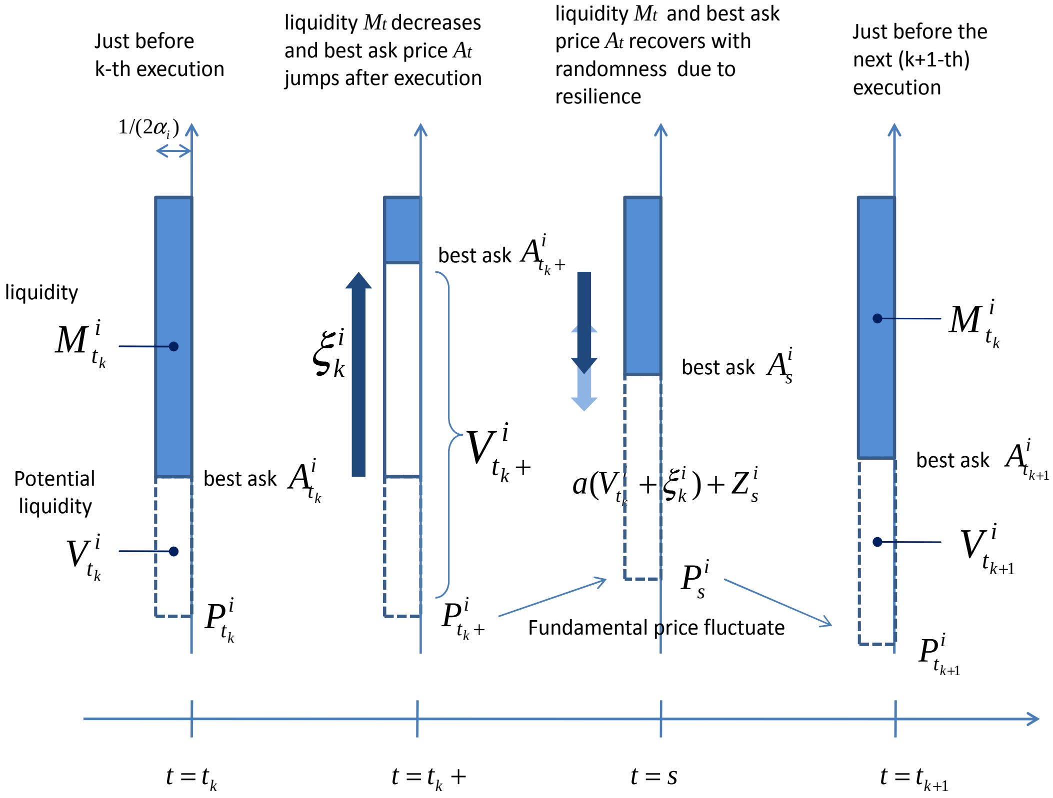
$$IS_i = \sum_{k=0}^{K} C_k^i - P_0^i w_i. \tag{2.6}$$

This paper focuses only on the implicit costs of trading, comprised of market impact costs, missed trading opportunity costs, and delay costs (slippage). Our analysis therefore omits other direct transaction costs, such as broker commissions, taxes, fees paid to the exchange, and bid–ask spreads. We also assume a block-shaped limit order book as in Obizhaeva and Wang [2005]. We define $f_i(p) = (2\alpha_i)^{-1}$, where α_i refers to the liquidity parameter.[5] Note that a smaller α_i implies the higher liquidity of asset

[5] The domain of f_i, i.e., the price range where limit orders exist, is considered to be finite. It is acceptable that the limit order book $f_i(p)$ is defined up to the price level p_+ where both the



(a) Limit order book model



(b) Dynamics of liquidity recovery after executions

Figure 1: Dynamics of the model

i. Because $G_i(v) = \alpha_i v^2$ ($v \geq 0$) for $f_i(p)$ under consideration,[6] the cost of purchasing ξ_k^i units of asset i at time t_k is given by:

$$C_k^i = P_k^i \xi_k^i + 2\alpha_i V_k^i \xi_k^i + \alpha_i (\xi_k^i)^2, \tag{2.7}$$

using Eq.(2.5).

From Eq.(2.7), the execution cost C_k^i is decomposed into three separate marginal costs. The first term in Eq.(2.7), $P_k^i \xi_k^i$, is the fundamental cost paid for the intrinsic value of an asset. The second term, $2\alpha_i V_k^i \xi_k^i$, is the liquidity cost incurred for the fluctuation in liquidity. This cost is recursive because it is contingent on past executions. Finally, the third term, $\alpha_i(\xi_k^i)^2$, is considered a temporal cost proportional to the difference between the cost of purchasing ξ_k^i of asset i at the current ask price A_k^i without any market impact, $A_k^i \xi_k^i$, and that with market impact, $(A_k^i + A_{k+}^i)\xi_k^i/2 = (2A_k^i + \alpha_i \xi_k^i)\xi_k^i/2$. This cost is "temporal" because it does not affect executions in other time slots.

2.3 The optimal execution problem

In what follows, we define $\boldsymbol{\xi}_k = (\xi_k^1, \ldots, \xi_k^N)^\top$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)^\top$, where $\top$ denotes the transpose, and the entire execution strategy is represented as $\boldsymbol{\xi} = \{\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_K\}$. The total execution cost $C(\boldsymbol{\xi})$ for strategy $\boldsymbol{\xi}$ is then given by:

$$C(\boldsymbol{\xi}) = \sum_{i=1}^{N}\sum_{k=0}^{K} C_k^i = \underbrace{\sum_{i=1}^{N}\sum_{k=0}^{K} P_k^i \xi_k^i}_{=I_1(\boldsymbol{\xi})} + \underbrace{\sum_{i=1}^{N}\alpha_i \sum_{k=0}^{K}(2V_k^i + \xi_k^i)\xi_k^i}_{=I_2(\boldsymbol{\xi})}, \tag{2.8}$$

where $I_1(\boldsymbol{\xi})$ and $I_2(\boldsymbol{\xi})$ are defined as the first and second term in Eq.(2.8), respectively. Given we restrict ourselves to considering static optimization, our objective is to determine the execution schedule $\boldsymbol{\xi}$ at time $t = 0$ so as to minimize $\mathrm{E}\left[C(\boldsymbol{\xi})\right] + \lambda \mathrm{V}\left[C(\boldsymbol{\xi})\right]$, where λ denotes the investor's coefficient of risk aversion ($\lambda \geq 0$). The mean and the variance of $I_1(\boldsymbol{\xi})$ are calculated as:

$$\mathrm{E}\left[I_1(\boldsymbol{\xi})\right] = \sum_{i=1}^{N} P_0^i \omega_i, \tag{2.9}$$

$$\mathrm{V}\left[I_1(\boldsymbol{\xi})\right] = \sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_\Delta^{ij}\sum_{k=1}^{K}\left(\sum_{s=k}^{K}\xi_s^i\right)\left(\sum_{s=k}^{K}\xi_s^j\right). \tag{2.10}$$

---

fluctuation in liquidity, Z_t^i, and every execution ξ_k^i are absorbed (hence, $f_i(p) = (2\alpha_i)^{-1}$ for $|p| \leq p_+$, and $f_i(p) = 0$ for $|p| > p_+$). Mathematically more rigorously, the upper limit of price is $p_+ = 2\alpha_i\{(K+1)\delta + \omega_i\}$, which is obtained by assuming the δ-truncated distribution for Z_t^i (e.g., the normal distribution truncated at $\pm\delta$). However, the finiteness of f_i does not affect the following derivation of the optimal execution strategy in Section 3 and related discussion.

[6] This is because $F_i^{-1}(p) = 2\alpha_i p$, $G(v)$ is computed as $G(v) = \int_0^{2\alpha_i v} p(2\alpha_i)^{-1}dp = \alpha_i v^2$.

As we obtain

$$V_k^i = \sum_{s=0}^{k-1} a^{k-s}\xi_s^i + \sum_{s=1}^{k} a^{k-s} Z_s^i, \tag{2.11}$$

by solving Eq.(2.4) recursively, the mean and the variance of $I_2(\boldsymbol{\xi})$ are calculated, after a permutation of summations, as:

$$\begin{aligned} \mathrm{E}\left[I_2(\boldsymbol{\xi})\right] &= \sum_{i=1}^{N} 2\alpha_i \sum_{k=0}^{K-1} \left(\sum_{s=k+1}^{K} a^{s-k}\xi_s^i \right) \xi_k^i + \sum_{i=1}^{N} \alpha_i \sum_{k=0}^{K} (\xi_k^i)^2, & (2.12) \\ \mathrm{V}\left[I_2(\boldsymbol{\xi})\right] &= \sum_{i=1}^{N}\sum_{j=1}^{N} (2\alpha_i)(2\alpha_j)\sigma_Z^{ij} \sum_{k=1}^{K} \left(\sum_{s=k}^{K} a^{s-k}\xi_s^i \right) \left(\sum_{s=k}^{K} a^{s-k}\xi_s^j \right). & (2.13) \end{aligned}$$

Noting that Eq.(2.9) is a constant independent of $\boldsymbol{\xi}$, we formulate the optimization problem as:

$$\min_{\boldsymbol{\xi}} \ \mathrm{E}\left[I_2(\boldsymbol{\xi})\right] + \lambda \left\{ \mathrm{V}\left[I_1(\boldsymbol{\xi})\right] + \mathrm{V}\left[I_2(\boldsymbol{\xi})\right] \right\} \qquad \text{s.t. } \boldsymbol{\omega} = \sum_{k=0}^{K} \boldsymbol{\xi}_k. \tag{2.14}$$

In Eq.(2.14), we do not require ξ_i to be positive, a priori; i.e., we also allow for intermediate sell orders. However, the positivity can be proved for the single-asset case, as will be shown in Corollary 2.

This formulation corresponds to the mean–variance minimization of the implementation shortfall $IS(\boldsymbol{\xi}) = \sum_{i=1}^{N} IS_i$. This is because the mean and variance for the total execution are computed from Eq.(2.6) as:

$$\begin{aligned} \mathrm{E}\left[IS(\boldsymbol{\xi})\right] &= \sum_{i=1}^{N} \mathrm{E}\left[IS_i\right] = \mathrm{E}\left[I_2(\boldsymbol{\xi})\right], \\ \mathrm{V}\left[IS(\boldsymbol{\xi})\right] &= \sum_{i=1}^{N} \mathrm{V}\left[C(\boldsymbol{\xi})\right] = \mathrm{V}\left[I_1(\boldsymbol{\xi})\right] + \mathrm{V}\left[I_2(\boldsymbol{\xi})\right]. \end{aligned}$$

Thus, Eq.(2.14) is rewritten as:

$$\min_{\boldsymbol{\xi}} \ \mathrm{E}\left[IS(\boldsymbol{\xi})\right] + \lambda \mathrm{V}\left[IS(\boldsymbol{\xi})\right] \qquad \text{s.t. } \boldsymbol{\omega} = \sum_{k=0}^{K} \boldsymbol{\xi}_k. \tag{2.15}$$

2.4 Discussion of market impact

The market impact is the response (rise or decline) of market prices associated with an execution. We employ the well-known framework in Holthausen, Leftwich, and Mayers [1987], Grinold and Kahn [1999], or Almgren and Chriss [2001], where market impacts are exogenously decomposed into two parts: a temporary impact assumed to recover by the next execution time slot, and a permanent impact assumed to remain at least to the execution time horizon. We refer to this type of execution cost model as an exogenous market impact model. The following discusses the execution costs of our limit order book type model, comprising fundamental, recursive, and temporal costs by comparing these costs with the framework of the exogenous market impact models.

In our model, Eq.(2.11) implies how an execution affects potential liquidity. This indicates that the impact of an execution on liquidity is inherited by subsequent executions, which decay geometrically at a rate $a \in (0,1)$, accompanied by a random fluctuation. We refer to this type of impact as the recursive market impact. When a is close to 1, the impact decays slowly and persists for a considerable length of time. Thus, most of the impact of $a \simeq 1$ is considered to be a permanent impact, which remains at least until the end of the total execution T. In contrast, when a is close to 0, the impact diminishes quickly, and for that reason, most of the impact is considered to be temporary.

We now reconsider these points more specifically. Our model is interpreted as an exogenous market impact model where the temporary and permanent impacts are increasing linear functions of an execution volume with additional uncertainty. Given the best ask price just before the k-th execution, $A^i_k = 2\alpha_i V^i_k + P^i_k$, which increases to $A^i_{k+} = A^i_k + 2\alpha_i \xi^i_k$ because of the execution and then declines to $A^i_{k+1} = 2\alpha_i[a(V^i_k + \xi^i_k) + Z^i_{k+1}] + P^i_{k+1}$ just before the next execution, we identify the temporary impact of the k-th execution on the best ask price, which recovers by the next time slot, and the permanent impact, which does not, as:

$$\begin{cases} \text{temporary impact} = A^i_{k+} - A^i_{k+1} = 2\alpha_i(1-a)(\xi^i_k + V^i_k) - \Delta^i_{k+1} - 2\alpha Z^i_{k+1}, \\ \text{permanent impact} = A^i_{k+1} - A^i_k = 2\alpha_i a \xi^i_k - 2\alpha_i(1-a)V^i_k + \Delta^i_{k+1} + 2\alpha_i Z^i_{k+1}. \end{cases} \tag{2.16}$$

This shows that the recursive impact can be forcibly decomposed into temporary and permanent impacts, both of which are increasing linear functions of ξ^i_k with an additional liquidity factor $2\alpha_i(1-a)V^i_k$ and stochastic factors $\Delta^i_k + 2\alpha Z^i_k$. However, because the temporal and permanent impacts are mutually complementary with respect to these factors, the permanent impact can be negative and may be absorbed by the temporal impact in the current or following executions; therefore, the permanent impact in Eq.(2.16) does not always actually leave a "permanent" effect on subsequent prices in our model. This feature is missing from earlier studies and is unique to our recursive market impact. This feature also leads to the differences in optimal execution strategies from those provided in the exogenous market impact models, which are discussed later.

Almgren and Chriss [2001] provided an analytical execution strategy, which minimizes the mean and variance of the execution cost, the same problem setting as ours, where both impacts are assumed to be a linear function of the execution volume. In such a case, the total execution cost becomes a quadratic function of each execution volume, which is analytically solvable. This is true to our model; the cost function in

Eq.(2.7) is a quadratic function of ξ.

The shape of an order book plays an important role in determining the functional form of the temporary and permanent impacts. We can easily see from the discussion in Section 2.1 that our assumption of a block-shaped order book makes both of the impacts linear functions of the execution volume. When an order book on the ask side is a declining function of price, which is a more realistic assumption, both the temporary and permanent impacts are nonlinear increasing functions of execution volume. Taking an example, consider when the shape of the ask side of an order book is a reciprocal function of price, $f_i(p) = 1/\{\beta_i(p+\bar{\alpha}_i)\}$, where $\beta_i > 0$ and $0 < \bar{\alpha}_i < A_0^i$ for any asset i. The temporary and permanent impact functions are represented in the exponential functions of ξ_k^i as[7] :

$$\begin{cases} \text{temporary impact} = \bar{\alpha}_i(e^{\beta_i(V_k^i+\xi_k^i)} - e^{\beta_i a(V_k^i+\xi_k^i)+\beta_i Z_k^i}) - \Delta_{k+1}^i, \\ \text{permanent impact} = \bar{\alpha}_i(e^{\beta_i a(V_k^i+\xi_k^i)+\beta_i Z_k^i} - e^{\beta_i V_k^i}) + \Delta_{k+1}^i. \end{cases}$$

While these functions are still an increasing function of ξ, the magnitude of the impacts differs from the case of the block-shaped order book.

This paper, however, focuses on optimal execution in the block-shaped limit order book as it is analytically tractable.[8] It is possible to understand the general features of an optimal execution schedule based on the block-shaped order book case because the execution cost under the block-shaped order book is a first-order approximation of that under a general-shaped order book. Because $G_i'(x) = F_i^{-1}(x)$, Eq.(2.5) is expanded as:

$$C_k^i = P_k^i\xi_k^i + F_i^{-1}(V_k^i)\xi_k^i + \frac{(\xi_k^i)^2}{2}(F_i^{-1})'(V_k^i) + \cdots .$$

Under the block-shaped order book assumption, the third term is constant and the higher-order terms vanish, which is not the case for a general-shaped order book. The effects of the third- and higher-order terms are large when the order density decays faster as price increases. Consequently, the optimal execution velocity is faster when the (ask-side) limit order book is a decreasing function of price compared with that of the block-shaped order book.

3 Optimal execution strategies

In this section, we solve Eq.(2.15) to obtain an optimal execution strategy.

[7] In this case, $F_i(p) = \dfrac{1}{\beta_i}\ln\dfrac{p+\bar{\alpha}_i}{\bar{\alpha}_i}$, and $F_i^{-1}(p) = \bar{\alpha}_i(e^{\beta_i p} - 1)$; hence, $A_k^i = \bar{\alpha}_i(e^{\beta_i V_k^i} - 1) + P_k^i$.

[8] Though Alfonsi, Fruth, and Schied [2010] consider a mean minimization problem of execution costs under a general-shaped limit order book, we consider the mean–variance minimization problem, which makes it difficult to optimize under a general-shaped order book.

3.1 General case

The first-order optimality condition of the minimization problem in Eq.(2.14) with respect to ξ_k^i is given by:

$$\frac{\partial \mathrm{E}\left[I_2(\boldsymbol{\xi})\right]}{\partial \xi_k^i} + \lambda \frac{\partial \mathrm{V}\left[I_1(\boldsymbol{\xi})\right]}{\partial \xi_k^i} + \lambda \frac{\partial \mathrm{V}\left[I_2(\boldsymbol{\xi})\right]}{\partial \xi_k^i} = \beta_i, \quad (i = 1, \ldots, N,\ k = 0, \ldots, K), \tag{3.1}$$

where β_i is a Lagrange multiplier for the constraint of the total volume to execute for asset i. Here and in what follows, we define $\sum_{k=a}^{b} x_k = 0$ if $a > b$ for any $\{x_k\}$. Eq.(3.1) with the constraint $\boldsymbol{\omega} = \sum_{k=0}^{K} \boldsymbol{\xi}_k$ forms a system of $N(K+2)$ linear equations with the same number of unknown parameters ξ_k^i $(i = 1, \ldots, N, k = 0, \ldots, K)$ and β_i $(i = 1, \ldots, N)$. Therefore, our objective reduces to finding the explicit solution to Eq.(3.1) and the constraint $\boldsymbol{\omega} = \sum_{k=0}^{K} \boldsymbol{\xi}_k$.

From Eqs.(2.10)$\sim$(2.13), Eq.(3.1) can be explicitly expressed in matrix form as:

$$\begin{aligned} 2\Bigg\{ \boldsymbol{\alpha} \left(\sum_{s=0}^{k-1} a^{k-s} \boldsymbol{\xi}_s + \sum_{s=k}^{K} a^{s-k} \boldsymbol{\xi}_s \right) &+ \lambda \boldsymbol{\Sigma}_\Delta \sum_{s=1}^{k} \sum_{u=s}^{K} \boldsymbol{\xi}_u \\ &+ 4\lambda \boldsymbol{\alpha} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \sum_{s=1}^{k} a^{k-s} \sum_{u=s}^{K} a^{u-s} \boldsymbol{\xi}_u \Bigg\} = \boldsymbol{\beta}, \quad (k = 0, \ldots, K), \end{aligned} \tag{3.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^\top$ and $\boldsymbol{\alpha} = \mathrm{Diag}(\alpha_i)$ is a diagonal matrix with α_i's. To represent the optimal solution, we define:

$$\boldsymbol{A} = (1 - a^2)\boldsymbol{\alpha} + a\lambda \boldsymbol{\Sigma}_\Delta + 4\lambda \boldsymbol{\alpha} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha}, \qquad \boldsymbol{B} = (1 - a)^2 \lambda \boldsymbol{A}^{-1} \boldsymbol{\Sigma}_\Delta. \tag{3.3}$$

As $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}_\Delta$, and $\boldsymbol{\Sigma}_Z$ are positive definite, so is $\boldsymbol{A}$, which ensures the existence of $\boldsymbol{A}^{-1}$. Let $\boldsymbol{C}$ be the Cholesky factorization of $\boldsymbol{\Sigma}_\Delta$; i.e., $\boldsymbol{\Sigma}_\Delta = \boldsymbol{C}^\top \boldsymbol{C}$. From Eq.(3.3), we obtain:

$$\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{-1} = (1 - a)^2 \lambda \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}^\top. \tag{3.4}$$

The right-hand side of Eq.(3.4) is positive definite and thus diagonalizable as:

$$\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{-1} = (1 - a)^2 \lambda \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}^\top = \boldsymbol{D}^{-1} \boldsymbol{\Gamma} \boldsymbol{D}, \tag{3.5}$$

where $\boldsymbol{\Gamma} = \mathrm{Diag}(\gamma_i)$ with γ_i $(i = 1, \ldots, N)$ being the eigenvalues of both sides of Eq.(3.4) and $\boldsymbol{D}$ being the $N \times N$ matrix composed of the associated left eigenvectors. It is noted here that from the positive definiteness, $\gamma_i > 0$ for all i. Letting $\boldsymbol{R} = (\boldsymbol{D}\boldsymbol{C})^{-1}$, we see from Eq.(3.5) that:

$$\boldsymbol{B} = \boldsymbol{R}\boldsymbol{\Gamma}\boldsymbol{R}^{-1}, \qquad \boldsymbol{B}\boldsymbol{R} = \boldsymbol{R}\boldsymbol{\Gamma}, \qquad \boldsymbol{R}^{-1}\boldsymbol{B} = \boldsymbol{\Gamma}\boldsymbol{R}^{-1}. \tag{3.6}$$

Using the above setting, we obtain the closed-form solution to the problem in Eq.(2.14) of the optimal execution strategy $\boldsymbol{\xi}^*$.

Theorem 1 *Let:*

$$\theta_i = \frac{\gamma_i + 2 + \sqrt{\gamma_i^2 + 4\gamma_i}}{2}, \quad (i = 1, \ldots, N), \tag{3.7}$$

and $\boldsymbol{\Theta} = Diag(\theta_i)$. *The optimal execution strategy in time slot* K *is given in terms of* $\boldsymbol{\Theta}$ *by:*

$$\begin{aligned} \boldsymbol{\xi}_K^* &= (\boldsymbol{I} - \boldsymbol{\Theta}^2) \Big[(\boldsymbol{I} + \frac{\lambda}{1-a} \boldsymbol{\alpha}^{-1} \boldsymbol{\Sigma}_\Delta) \left\{ (\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+1} - (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^K \right\} \\ &\quad + \left(\boldsymbol{I} + \frac{4\lambda}{1-a} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \right) (\boldsymbol{I} - \boldsymbol{\Theta})(\boldsymbol{\Theta}^K + \boldsymbol{\Theta}^{-K+1}) \Big]^{-1} \boldsymbol{\omega}. \end{aligned} \tag{3.8}$$

In time slots $k = 0, \ldots, K-1$, *the optimal execution strategy is given in terms of* $\boldsymbol{\Theta}$ *and* $\boldsymbol{\xi}_K^*$ *in Eq.(3.8) by:*

$$\begin{aligned} \boldsymbol{\xi}_0^* &= \Big[\frac{\lambda}{1-a} \boldsymbol{\alpha}^{-1} \boldsymbol{\Sigma}_\Delta \left\{ (\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+1} - (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^K \right\} \\ &\quad + \left(\boldsymbol{I} + \frac{4\lambda}{1-a} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \right) (\boldsymbol{I} - \boldsymbol{\Theta}) \left(\boldsymbol{\Theta}^K + \boldsymbol{\Theta}^{-K+1} \right) \Big] (\boldsymbol{I} - \boldsymbol{\Theta}^2)^{-1} \boldsymbol{\xi}_K^*, \end{aligned} \tag{3.9}$$

$$\boldsymbol{\xi}_k^* = (\boldsymbol{I} + \boldsymbol{\Theta})^{-1} \left\{ (\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+k} + (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^{K-k} \right\} \boldsymbol{\xi}_K^*, \quad (k = 1, \ldots, K-1). \tag{3.10}$$

Proof. See the appendix.

So far, we have assumed $V_0^i = 0$ for all i, meaning that no potential sell orders are executed at time 0. However, if the effect of past large buy orders has not yet vanished, some of the potential sell orders may have been executed at time 0; i.e., $V_0^i > 0$. In this case, the optimal execution strategy in Theorem 1 is modified as follows.

Corollary 1 *When* $V_0^i > 0$ $(i = 1, \ldots, N)$, *the optimal execution strategy in time slot* K *is given by:*

$$\begin{aligned} \boldsymbol{\xi}_K^* &= (\boldsymbol{I} - \boldsymbol{\Theta}^2) \Big[(\boldsymbol{I} + \frac{\lambda}{1-a} \boldsymbol{\alpha}^{-1} \boldsymbol{\Sigma}_\Delta) \left\{ (\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+1} - (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^K \right\} \\ &\quad + \left(\boldsymbol{I} + \frac{4\lambda}{1-a} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \right) (\boldsymbol{I} - \boldsymbol{\Theta})(\boldsymbol{\Theta}^K + \boldsymbol{\Theta}^{-K+1}) \Big]^{-1} (\boldsymbol{V}_0 + \boldsymbol{\omega}). \end{aligned} \tag{3.11}$$

where $\boldsymbol{V}_0 = (V_0^1, \ldots, V_0^N)^\top$. *In time slots* $k = 0, \ldots, K-1$, *the optimal execution strategy is given in terms of* $\boldsymbol{\xi}_K^*$ *in Eq.(3.11) by:*

$$\begin{aligned} \boldsymbol{\xi}_0^* &= \Big[\frac{\lambda}{1-a} \boldsymbol{\alpha}^{-1} \boldsymbol{\Sigma}_\Delta \left\{ (\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+1} - (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^K \right\} \\ &\quad + \left(\boldsymbol{I} + \frac{4\lambda}{1-a} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \right) (\boldsymbol{I} - \boldsymbol{\Theta}) \left(\boldsymbol{\Theta}^K + \boldsymbol{\Theta}^{-K+1} \right) \Big] (\boldsymbol{I} - \boldsymbol{\Theta}^2)^{-1} \boldsymbol{\xi}_K^*, \end{aligned} \tag{3.12}$$

$$\boldsymbol{\xi}_k^* = (\boldsymbol{I} + \boldsymbol{\Theta})^{-1} \left\{ (\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+k} + (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^{K-k} \right\} \boldsymbol{\xi}_K^* - \boldsymbol{V}_0, \quad (k = 1, \ldots, K-1). \tag{3.13}$$

The optimal strategy in Corollary 1 is obtained by replacing the total execution volume ω_i by $V_0^i + \omega_i$ in the strategy of $V_0^i = 0$ in Eq.(3.8). On the other hand, $\boldsymbol{\xi}_k^i$ in Eq.(3.13) equals that in Eq.(3.10) minus V_0^i, compensating for the difference in total execution volumes. We omit the proof as it is similar to that of Theorem 1 in the appendix.

3.2 Special cases

In this section, we consider two specific cases. The first is when the fundamental price is deterministic, and the second is when there is only a single asset.

First, in order to extract the effect of the market impact more explicitly, we delete the effect caused by the fundamental price movement by setting $\boldsymbol{\Sigma}_\Delta = \boldsymbol{O}$.[9] It is easy to check that the optimal execution strategy in Theorem 1 is reduced to the following simplified form.

Theorem 2 *When $\boldsymbol{\Sigma}_\Delta = \boldsymbol{O}$, the optimal execution strategy is given by:*

$$
\begin{aligned}
\boldsymbol{\xi}_K^* &= \left[\{(K-1)(1-a)+2\} \boldsymbol{I} + \frac{4\lambda}{1-a} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \right]^{-1} \boldsymbol{\omega}, \\
\boldsymbol{\xi}_0^* &= \left(\boldsymbol{I} + \frac{4\lambda}{1-a} \boldsymbol{\Sigma}_Z \boldsymbol{\alpha} \right) \boldsymbol{\xi}_K^*, \\
\boldsymbol{\xi}_k^* &= (1-a)\boldsymbol{\xi}_K^*, \quad (k = 1, \ldots, K-1).
\end{aligned}
$$

Proof. See the appendix.

A remarkable feature in this case is that the optimal strategy is the same for each k except $k = 0$ and $k = K$. Optimal strategies of the same type have been found in Alfonsi, Fruth, and Schied [2010] for a single-asset case with a more general market impact function. See also Obizhaeva and Wang [2005].

Second, when there is only a single asset, the optimal execution strategy can be simplified. Let:

$$
\lambda_\Delta = \frac{\lambda \Sigma_\Delta}{\alpha(1-a)}, \quad \lambda_Z = \frac{4\lambda\alpha\Sigma_Z}{1-a}, \quad \gamma = \frac{(1-a)^2\lambda_\Delta}{a\lambda_\Delta + \lambda_Z + a + 1}, \quad \theta = \frac{\gamma + 2 + \sqrt{\gamma^2 + 4\gamma}}{2},
$$

then, we have the following.

Theorem 3 *For a single-asset case, the optimal execution strategy is given as follows.*
(1) When $\lambda\Sigma_\Delta \neq 0$:

$$
\begin{aligned}
\xi_0^* &= \{\lambda_\Delta \phi + (\lambda_Z + 1)\psi\}\varphi, \\
\xi_k^* &= \left\{ (\theta - a)\theta^{K-k} + (1 - a\theta)\theta^{-K+k} \right\} \varphi, \quad (k = 1, \ldots, K-1), \\
\xi_K^* &= (\theta + 1)\varphi,
\end{aligned}
$$

[9] In the above discussions for the general case, we assume the positive definiteness of $\boldsymbol{\Sigma}_\Delta$. This does not hold for the case where the fundamental price is deterministic.

where $\phi = \dfrac{(\theta-a)\theta^K - (1-a\theta)\theta^{-K+1}}{\theta-1}$, $\varphi = \dfrac{\omega}{(\lambda_\Delta+1)\phi + (\lambda_Z+1)\psi}$,
and $\psi = \theta^K + \theta^{-K+1}$.

(2) When $\lambda\Sigma_\Delta = 0$*:*

$$\begin{aligned}
\xi_0^* &= \frac{\lambda_Z+1}{\lambda_Z+2+(K-1)(1-a)}\omega,\\
\xi_k^* &= \frac{1-a}{\lambda_Z+2+(K-1)(1-a)}\omega, \quad (k=1,\dots,K-1),\\
\xi_K^* &= \frac{1}{\lambda_Z+2+(K-1)(1-a)}\omega.
\end{aligned}$$

Corollary 2 $\xi_k^* > 0$ *($k = 0,\dots,K$) for the optimal strategies in Theorem 3.*

Proof. Because $\theta > 1$,

$$\phi > \frac{(\theta-a)-(1-a\theta)}{\theta-1} = 1+a > 0,$$

hence, $\xi_0^* > 0$ is proved.[10] $\xi_k^* > 0$ can be proved in a similar manner. □

Corollary 2 confirms the optimal strategies are composed only of buy orders for the single-asset case.

4 Properties of optimal execution strategies

We analyze the properties of the optimal execution strategy in Theorem 1, 2, and 3 in this section. We first consider the properties of a single-asset execution and then summarize those in a multiasset environment.

4.1 Properties of single-asset execution

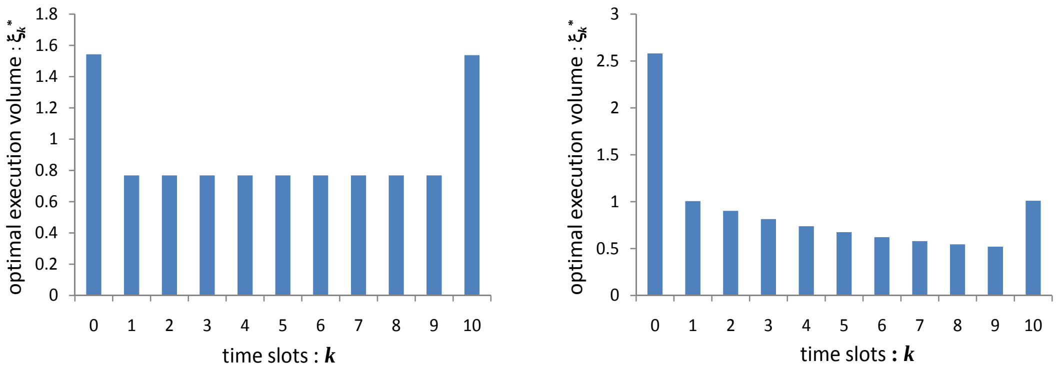
4.1.1 The case of $\lambda\Sigma_\Delta = 0$

First, we investigate the determinants of the optimal execution schedule in the case of $\lambda\Sigma_\Delta = 0$, shown in Theorem 3 (2). Figure 2 (a) displays the typical optimal execution schedule when $K = 10$ and $\omega = 10$. The optimal volumes at the first and last time slots are larger than those in the remaining time slots, while the volumes

[10] Because $\theta > 1$ from Eq.(3.7), $\theta - a > 0$ is proved. $1 - a\theta > 0$ is proved as follows. Let $g(x) = x^2 - (\gamma+2)x + 1$. By direct substitution, we obtain:

$$\begin{aligned}
g(1) &= -\gamma < 0,\\
g(1/a) &= \left(1-\frac{1}{a}\right)^2 \frac{\lambda_Z+a+1}{a\lambda_\Delta+\lambda_Z+a+1} > 0.
\end{aligned}$$

Given that θ is one of the solutions of the quadratic equation $g(x) = 0$, $1 - a\theta > 0$ holds.



(a) $\lambda\Sigma_\Delta = 0$ (b) $\lambda\Sigma_\Delta \neq 0$

Figure 2: A typical optimal execution schedule for a single asset

in the intermediate slots $t_1, \ldots, t_{K-1}$ are equivalent. This property is governed by the following three factors. The first factor is the temporal cost $\alpha\xi_k^2/2$, from which the uniform distribution of orders becomes optimal. The second factor is the recursive cost, from which a relatively slower execution becomes optimal. This is because an execution in earlier time slots generates some permanent impact, which increases the cost of following executions, as discussed in Section 2.1. The third factor is the assumption that the order book is initially in the stationary state, from which the investor has an incentive to increase the purchase in the first time slot because execution under the stationary state entails a smaller cost than the following time slots. These factors, taken as a whole, form a U-shaped optimal execution schedule; that is, the equally divided schedule is optimal in the intermediate time slots $t = t_1, \ldots, t_{K-1}$, while execution volumes in the first and last time slots, $t = 0$ and $t = T$, are higher than those in the intermediate slots.

Next, we consider the comparative statics of the optimal execution schedule. Table 1 summarizes the signs of the derivatives of ξ_k^* with respect to each parameter a, α, λ, and Σ_Z, showing whether the optimal volume increases or decreases in each time slot when these parameters increase. As shown from the table, faster execution becomes optimal as parameters α, λ, or Σ_Z increase. This is mainly because the first and the third factors explained above (the temporal cost and stationary initial state) dominate cost when liquidity decreases or variance rises. However, the direction of the last execution volume with varying a is indefinite depending on the balance of dominance between the second and third factors.

4.1.2 The case of $\lambda\Sigma_\Delta \neq 0$

First, we consider the determinants of the optimal execution schedule in the case of $\lambda\Sigma_\Delta \neq 0$, shown in Theorem 3 (1). Figure 2 (b) displays a typical optimal schedule in this case. Compared with the previous $\lambda\Sigma_\Delta = 0$ case, the optimal volumes in the

Table 1: Comparative statics when $\lambda\Sigma_\Delta = 0$

$\partial\xi_k^*/\partial x$	$t=0$	$t=t_1,\ldots,t_{K-1}$	$t=T$
$x:\ a$	$+$	$-$	indefinite
$x:\ \alpha,\lambda,\Sigma_Z$	$+$	$-$	$-$

Note: "$\pm$" indicates the sign of $\partial\xi_k^*/\partial x$.

Table 2: Comparative statics when $\lambda\Sigma_\Delta \neq 0$

$\partial(\xi_k^*/\xi_K^*)/\partial x$	$t=t_1,\ldots,t_{K-1}$
$x:\ a,\alpha,\Sigma_Z$	$-$
$x:\ \lambda,\Sigma_\Delta$	$+$

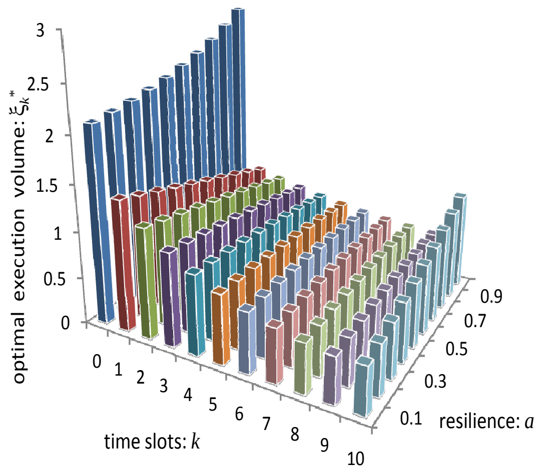
Note: "$\pm$" indicates the sign of $\partial(\xi_k^*/\xi_K^*)/\partial x$.

intermediate time slots are not flat; hence, the optimal execution velocity increases because of uncertainty in the fundamental price. We consider this to be the fourth factor concerning liquidity fluctuation. We numerically investigate this case as the direction of ξ_k^* is not as simple as that in the previous $\lambda\Sigma_\Delta = 0$ case. We detect that, while ξ_k^* $(k=0,\ldots,K-1)$ moves down when some parameters are very small, the overall feature is almost the same as the $\lambda\Sigma_\Delta = 0$ case; the larger the parameters a, α, λ, σ_Z, or Σ_Δ, the faster the execution should be. In other words, the investor should generally execute faster as resilience increases, as liquidity drops, as variance increases, or as the investor becomes more risk averse. This is consistent with our intuition of the behavior shown by investors.

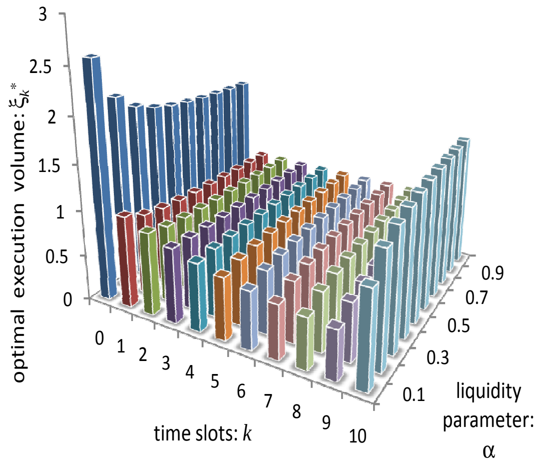
Next, we consider the comparative statics. Unlike the case of $\lambda\Sigma_\Delta = 0$, the derivatives of an execution schedule with respect to the parameters in this case have a much more complex form. Instead, we compute the derivatives of ξ_k^*/ξ_K^* $(k=1,\ldots,K-1)$, which gives us partial information on the structure of the optimal execution schedule. Table 2 summarizes the signs of the derivatives of ξ_k^*/ξ_K^* with respect to each parameter, a, α, λ, Σ_Z and Σ_Δ. We observe that the ratio of ξ_k^* to ξ_K^* for $k=1,\ldots,K-1$ increases when a, α and Σ_Z increase, while it decreases when λ and Σ_Δ increase. Roughly speaking, an increase in liquidity-related risk (i.e., a,α and Σ_Z) enhances early execution, while an increase in risk aversion and the fundamental price variance delays execution, when compared with the execution in the final period.

To investigate how each parameter affects the optimal execution schedule more specifically, we analyze the simulated schedule with varying parameters when $\lambda\Sigma_\Delta \neq 0$. Figure 3 displays the optimal strategy with varying parameters: a, α, Σ_Z, and λ.

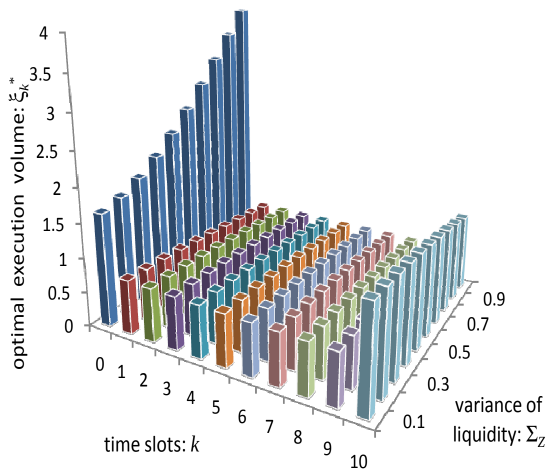
Panel (a) displays how the optimal execution schedule changes with varying liquid-



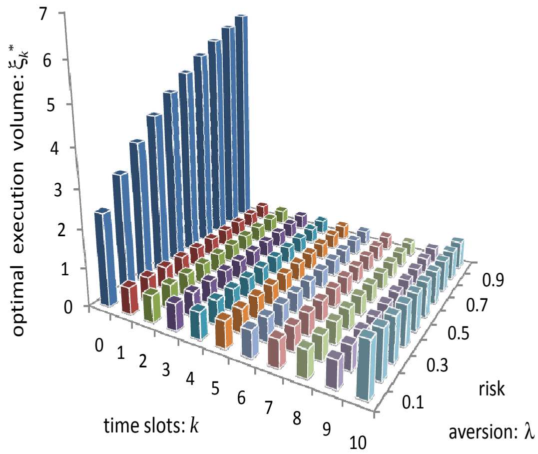
(a) Resilience : a



(b) Liquidity parameter: α



(c) Variance of liquidity : Σ_Z



(d) Coefficient of risk aversion : λ

Figure 3: Change in optimal execution schedules by varying parameters for a single-asset execution

Note: Parameters other than those varied are set to be $K = 10$, $\alpha = 0.1$, $a = 0.5$, $\lambda = 0.3$, $w = 10$, and $\Sigma_\Delta = \Sigma_Z = 0.1$.

ity resilience a. As shown, faster execution becomes optimal as the market resilience of liquidity goes up. In other words, the quicker and the greater the recovery in liquidity after execution, up to the next time slot, the slower the optimal execution. Underlying this is the fact that traders should wait for liquidity recovery by withholding buying volume if liquidity recovers quickly. This results in a decrease in the market impact of each trade.

Panel (b) displays the change in the optimal execution schedule by varying liquidity α. As α increases, hence as market liquidity declines, the optimal execution becomes slower and closer to the equally divided schedule. This implies that traders should avoid the larger temporal market impact in a smaller liquidity market, which results in expanding the effect of the equally divided schedule.

Panel (c) displays the change in the optimal execution schedule by varying the liquidity variance parameter Σ_Z. Similar to the resilience case in Panel (a), the larger the liquidity variance becomes, the faster the optimal execution. This is consistent with our intuition that an increase in the uncertainty of liquidity leads to faster execution for risk-averse investors.

Panel (d) displays the change in the optimal execution schedule with varying risk aversion coefficient λ. As traders become risk averse, the faster the optimal execution. Given that risk-averse traders avoid exposing themselves to volatility, faster execution becomes optimal. When $\lambda = 0$, the execution volumes in the intermediate time slots are equivalent, and these become a decreasing function of time when $\lambda \neq 0$.

4.2 Multiple assets

We focus on the joint property of the optimal execution strategy in Theorem 1 in this section. When considering a two-asset case, the joint property is determined by the four parameters in our model: asset correlation ρ_Δ implicit in $\mathbf{\Sigma}_\Delta$, liquidity correlation ρ_Z implicit in $\mathbf{\Sigma}_Z$, the difference in total volume ω_i, and the difference in liquidity α_i. As the correlation parameters mostly reflect the joint property, we analyze numerically the effect of the optimal strategy by varying the correlations. We only analyze a two-asset case here, with one liquid asset where $\alpha = 0.1$ and one illiquid asset, which has relatively low liquidity with $\alpha = 10$. However, this analysis can be generalized to cases of three or more assets.

Figure 4 displays the optimal remaining volume up to time slot k, $\mathbf{\Xi}_k = \boldsymbol{\omega} - \sum_{j=0}^{k-1} \boldsymbol{\xi}_j^*$, of the liquid and the illiquid assets with varying correlations.[11] Panel (a) shows the optimal remaining volume with varying asset correlation, and Panel (b) shows that with varying liquidity correlation. The other parameters are set to $T = 10$, $\lambda =$

[11] The remaining volume in each time slot is evaluated at the time just before the execution of the time slot. We denote the remaining volume evaluated at the time just after the time slot k as $\mathbf{\Xi}_{k+}$.

0.7, $\boldsymbol{\omega} = (10, 10)^\top$, and $a = 0.8$. We also set the marginal standard deviation of both the fundamental price and the liquidity to be 0.1 for both assets (hence, $\sigma_\Delta^{11} = \sigma_\Delta^{22} = \sigma_Z^{11} = \sigma_Z^{22} = 0.01$).

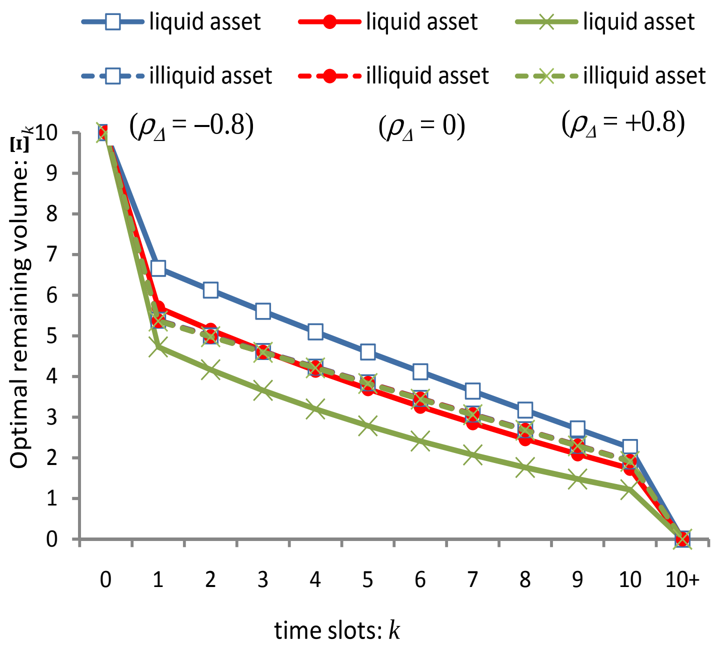
As shown from Figure 4, it is optimal to buy the liquid asset faster as the correlation coefficient increases, while the optimal execution schedule of the illiquid asset remains almost unchanged. This result indicates that execution costs should be controlled by the liquid asset only. We intuitively interpret this as meaning that investors can reduce volatility and liquidity risk by buying or selling a liquid asset, which produces a smaller market impact faster than an illiquid asset, when they are aware of a positive correlation among asset movements in advance. However, this is not the case when assets are negatively correlated, because buying one asset faster negatively affects the price of the other asset as it inherits higher risk afterwards.

We have one other type of solution that generates a round-trip trade of the liquid asset, while the optimal trade does not include the round-trip trade in the single-asset case, as proved in Section 4.1. This arises because the joint parameters (the correlations of the fundamental price or liquidity) are constants in our model. In other words, the investor knows the correlation in advance. Suppose the investor knows that the liquidity of assets i and j are positively correlated. The investor realizes that if he/she sells asset i to increase M_t^i, M_t^j is likely to move up; hence, the price of asset j moves down. This effect results in a reduction in the cost of buying asset j in later time slots, and the sold volume of asset i is bought back afterwards. This mechanism causes the round-trip trade and may violate the no-price-manipulation condition discussed in Gatheral [2008].[12]

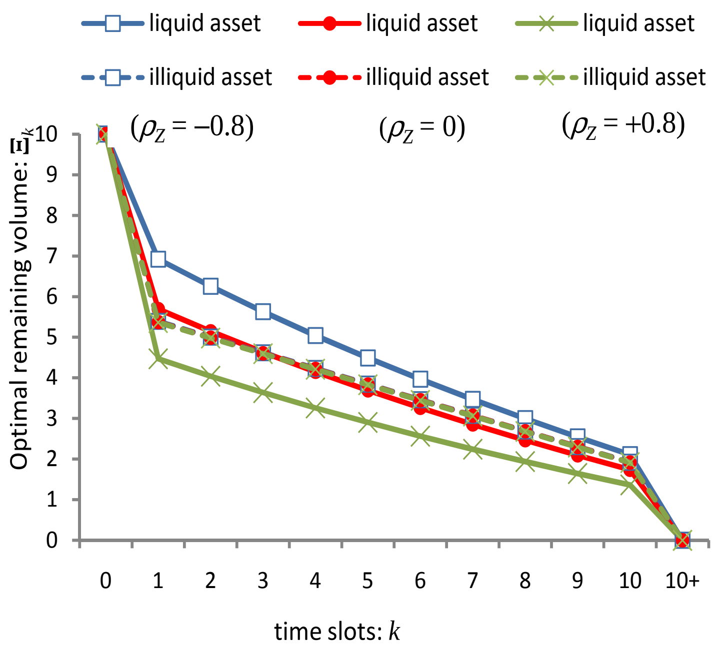
4.3 Efficient frontier of trading strategies

This section evaluates the efficient frontier of execution costs introduced by Almgren and Chriss [2001] in order to visualize the trade-off between the mean and variance of expected execution costs. The efficient frontier indicates the execution strategy with minimum variance for a given level of execution cost, which is constructed by computing the means and variances of the optimal execution strategies with varying risk aversion λ. Therefore, optimal execution strategies lie on the frontier, while the other feasible strategies lie above the frontier. Figure 5 illustrates the efficient frontier of two assets with various execution strategies when the parameters are $N = 2$, $K =$

[12] Note that the price-manipulation condition is a purely mathematical concept that differs from the form of market manipulation restricted by law. It is natural for a trader to generate profit from round-trip trades in the actual market. A market maker generally makes profit from round-trip trade mainly because of accompanying transaction costs, such as the bid-ask spread. It is also natural that the basket or portfolio trader makes profit from executions, which is similar to statistical arbitrage trading among assets.



(a) Optimal remaining volume with different asset correlations ρ_Δ



(b) Optimal remaining volume with different liquidity correlations ρ_Z

Figure 4: Optimal remaining volume with varying correlation parameters

Note: Because the optimal remaining volumes of the illiquid asset with different correlation parameters are almost unchanged and overlap considerably, they appear as a consolidated (dotted) line in the charts.

10, $a = 0.5$, $\boldsymbol{\alpha} = (0.5, 0.5)^\top$, $\boldsymbol{w} = (10, 10)^\top$, $(\sigma_\Delta^{11}, \sigma_\Delta^{22}) = (0.09, 0.09)^\top$, $(\sigma_Z^{11}, \sigma_Z^{22}) = (0.64, 0.64)^\top$, $\rho_\Delta = 0.5$, and $\rho_Z = 0.8$. The solid line indicates the frontier, and the points designate the following five types of strategies for both assets.

- Instant strategy: executes the entire volume at the current time 0; i.e., $\xi_0 = \omega$, $\xi_k = 0$ $(k > 0)$.
- Uniform strategy: equally allocates the volume to every execution slot; i.e., $\xi_k = \omega/(K+1)$ $(\forall k)$.
- First-and-last strategy: executes half of the total volume instantly, with the remaining half in the last time slot; i.e., $\xi_0 = \xi_K = \omega/2$, $\xi_k = 0$ $(0 < k < K)$.
- First-and-second strategy: executes half of the total volume in both the first and second time slots; i.e., $\xi_0 = \xi_1 = \omega/2$, $\xi_k = 0$ $(1 < k \leq K)$.
- Exponential-decay strategy: decreases the execution volume exponentially as time passes while executing the remaining volumes in the last time slot. Here, we suppose $\xi_k = \omega/2^{k+1}$ $(0 \leq k < K)$, $\xi_K = \omega(1 - \sum_{k=0}^{K-1} 2^{-k-1}) = \omega/2^K$.

The shape of the efficient frontier is quite similar to those of other models used in practice, such as that displayed in Exhibit 10 in Borkovec and Heidle [2010]. This suggests that our model of fluctuating liquidity on an order book is likely to work well in practice. The instant and uniform strategies are the two extreme strategies; the former minimizes the variance while entirely ignoring the level of expected cost, while the latter minimizes the expected cost while entirely ignoring the variance. These two extreme strategies anchor the efficient frontier at both endpoints.[13] The other strategies lie above the frontier, though each strategy has different mean and variance features, indicating that these strategies are not efficient in the model. The first-and-last strategy and the first-and-second strategy are both equally divided two-times schedules, though these means and variances are located at quite different levels. The first-and-last strategy waits for the recovery of liquidity until the end of the execution horizon while taking the risk of execution cost moves. Therefore, the strategy locates in the lower-right region of the chart, where expected costs are relatively low and variance is relatively high. The first-and-second strategy, on the contrary, executes just after the first execution, which results in higher cost with lower variance. Hence, the strategy locates in the upper-left region. The exponential-decay strategy is located closer to the efficient frontier, though reducing the expected costs or the variances could further improve the strategy.

[13] In Figure 5, the rightmost extreme of the efficient frontier does not reach the uniform strategy as the execution volumes in the first and last time slots do not always equal to those in the intermediate time slots because of the factors explained in Section 4.1.1.

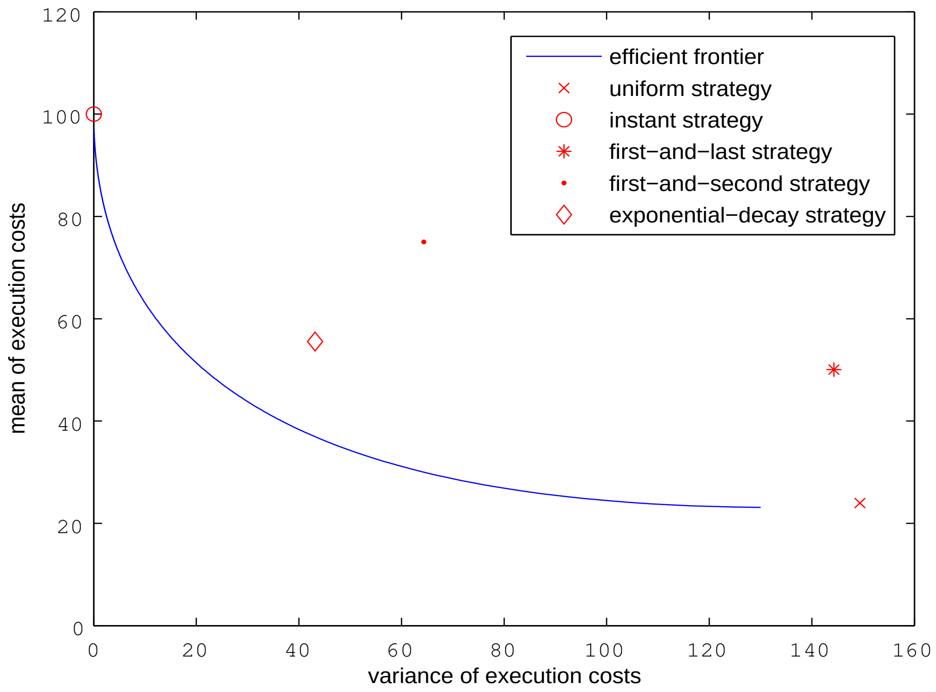


Figure 5: Efficient frontier and strategies

Note: The parameters are $N = 2$, $K = 10$, $a = 0.5$, $\boldsymbol{\alpha} = (0.5, 0.5)^\top$, $\boldsymbol{w} = (10, 10)^\top$, $(\sigma_\Delta^{11}, \sigma_\Delta^{22}) = (0.09, 0.09)^\top$, $(\sigma_Z^{11}, \sigma_Z^{22}) = (0.64, 0.64)^\top$, $\rho_\Delta = 0.5$, and $\rho_Z = 0.8$.

Figure 6 indicates how the efficient frontier moves by varying the parameters. While Sections 4.1 and 4.2 analyze the optimal execution strategy from the perspective of the optimal volume allocation, this section evaluates it from the perspective of the efficient frontier. The upper panel in Figure 6 illustrates the changes in the efficient frontier by varying parameter a. The other parameters are the same as in Figure 5. As a increases, the efficient frontier moves upward and is stretched to the right. This indicates that parameter a adjusts the trade-off between expected costs and variances, and moreover determines the balance between the permanent impact, which cannot be reduced by the execution horizon, and the temporal impact, which can be reduced in exchange for an increase in variance.

The lower panel in Figure 6 illustrates the changes in the efficient frontier by varying the liquidity correlation ρ_Z for the two assets. The other parameters are the same as in Figure 5. In addition to these, the efficient frontier when the variance of liquidity is close to zero ($\boldsymbol{\Sigma}_Z \approx \boldsymbol{O}$) is plotted. From the panel, we can see that the frontier moves upward when liquidity is positively correlated and downward when liquidity

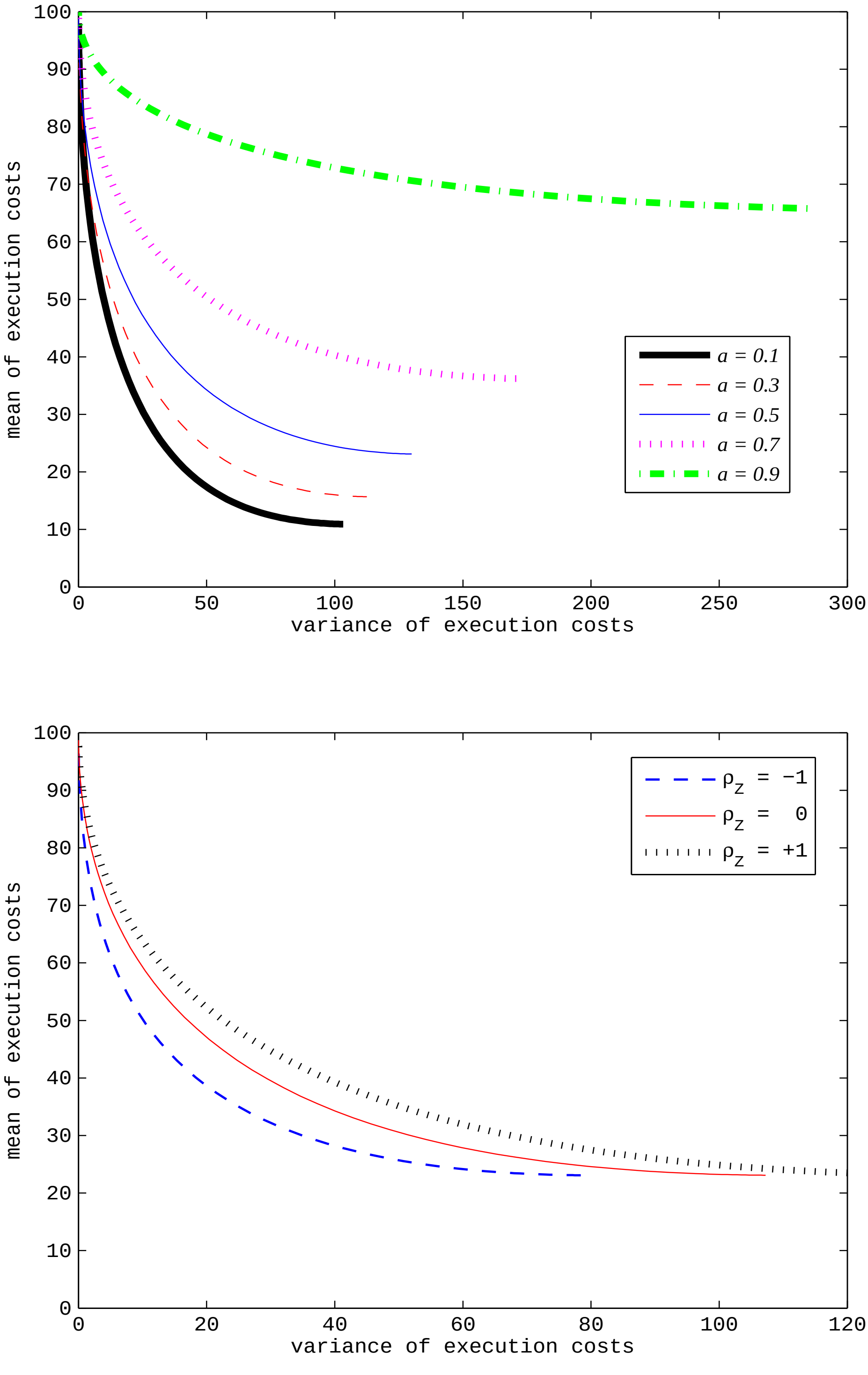


Figure 6: Efficient frontiers by parameters

Note: The upper and lower panel dipict the efficient frontiers versus a and ρ_Z, respectively. The other parameters are the same as in Figure 5.

is negatively correlated, while the minimum of execution costs or variance remains unchanged. This indicates that a negative correlation between the liquidity of the two assets makes it possible to reduce execution costs, while the level of execution risk remains unchanged. Conversely, positive liquidity correlation pushes execution costs up. Notably, the line for $\rho_Z = -1$ almost coincides with the line for $\boldsymbol{\Sigma}_Z \approx \boldsymbol{O}$. This implies that a portfolio manager can perfectly hedge execution costs by using other assets that are negatively correlated with the original asset.

Generally, when entire markets are very risk averse, market liquidity is unlikely to recover well, and liquidity is likely to be positively correlated among assets; hence, a is large and ρ_Z is positive. In such market conditions, our results indicate that one should execute trades immediately or avoid trading altogether. This is because the efficient frontier moves upward, particularly in the middle to high variance region, which results in both higher costs and risk required.

4.4 Impact on market volatility

The execution of a large block of securities impacts price volatility as well as the execution cost. We now analyze how the optimal slice of a block trade affects the volatility of the market price in the single-asset case. In order to do this, we compute the quadratic variation of the best ask price during the execution time period $t \in [0, T]$, denoted as $\Sigma_A(\boldsymbol{\xi})$. Because $A_t = P_t + F^{-1}(V_t) = P_t + 2\alpha V_t$ where P_t and V_t are independent processes, the quadratic variation can be easily computed as:

$$\Sigma_A(\boldsymbol{\xi}) = \int_0^T d[A, A]_s = K\Sigma_\Delta + 4\alpha^2 \left(K\Sigma_Z + \sum_{k=0}^{K} (\xi_k)^2 \right). \tag{4.1}$$

We can easily see that only the last term $\sum_{k=0}^{K} (\xi_k)^2$ includes the execution schedule; hence, we hereafter focus only on this term. We compare the effect of the optimal execution schedule $\boldsymbol{\xi}^*$ on volatility with two types of execution schedules: instant execution where $\xi_0 = \omega$, $\xi_k = 0$ $(k > 0)$, and uniform schedule where $\xi_k = \omega/(K + 1)$ $(\forall k)$. Given $\xi > 0$:

$$\sum_{k=0}^{K} \left(\frac{\omega}{K+1} \right)^2 \leq \sum_{k=0}^{K} (\xi_k^*)^2 \leq \left(\sum_{k=0}^{K} \xi_k \right)^2 = \omega^2. \tag{4.2}$$

Obviously, the quadratic variation is largest for the instant execution and smallest for the uniform schedule, while that of the optimal execution schedule falls between.

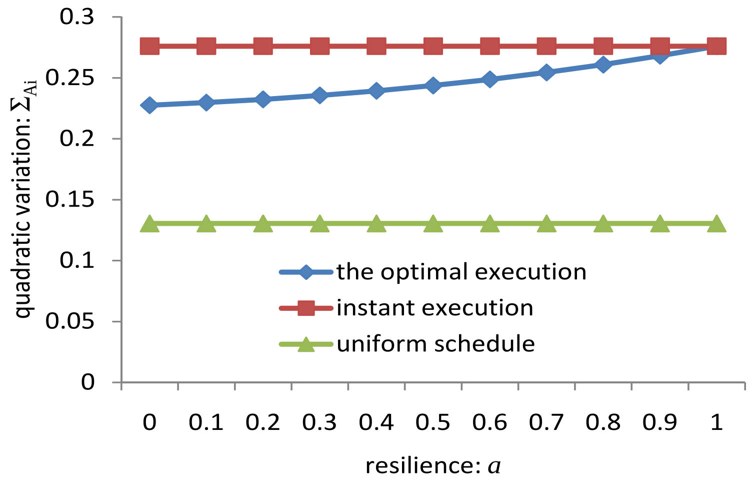
Figure 7 plots the quadratic variation in the single-asset case with varying a, α and Σ_Z. We confirm that the quadratic variation caused by the optimal execution is higher than that of the uniform schedule and lower than that of the instant execution. Also, the quadratic variation caused by optimal execution is closer to the instant execution than the uniform schedule. In Particular, when α is large, i.e., when liquidity is low, the difference in the impact of the optimal and uniform schedules on the quadratic variation increases.

5 Summary

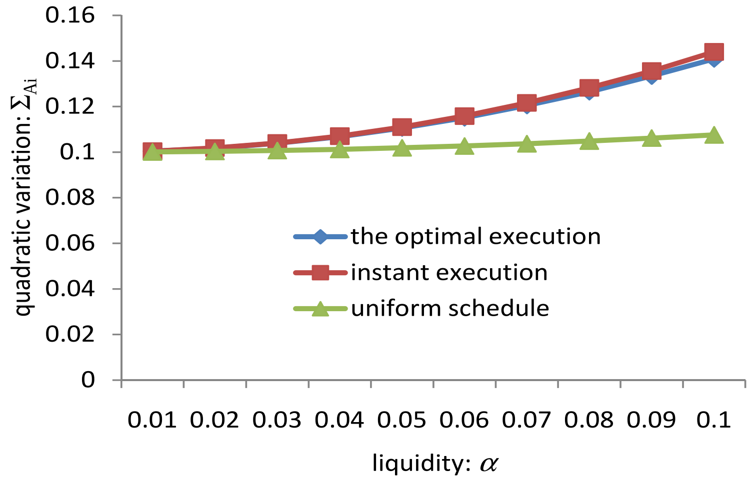
We have developed a multiasset model of market liquidity and derived the closed-form solution of optimal execution strategies under both liquidity and volatility risk. Market liquidity was modeled as a mean-reverting process, which was interpreted as a queue in an order book waiting for an execution. We then solved the mean–variance problem by optimizing the trade-off between market impacts and volatility/liquidity risk, and obtained an optimal execution strategy in an analytical form.

Our model and the optimal execution strategy allow us to understand the properties of the execution schedule of a risk-averse investor under fluctuating liquidity. In this, the investor attempts to execute faster as the market becomes less liquid, as the volatility of liquidity or price increases, or as the investor becomes more risk averse. For the multiasset case, we detected that the investor should sell/buy a liquid asset more quickly as the correlation of liquidity or the fundamental price increases among assets while keeping the execution velocity of an illiquid asset unchanged. By analysis of the efficient trading frontier of the model, we have also confirmed that both trading costs and risk increase dramatically in a distressed environment where liquidity recovery is slow and the correlation of liquidity is high. These findings remain consistent with the intuitive behavior of investors and with typical algorithmic trading strategies such as the implementation shortfall strategy.

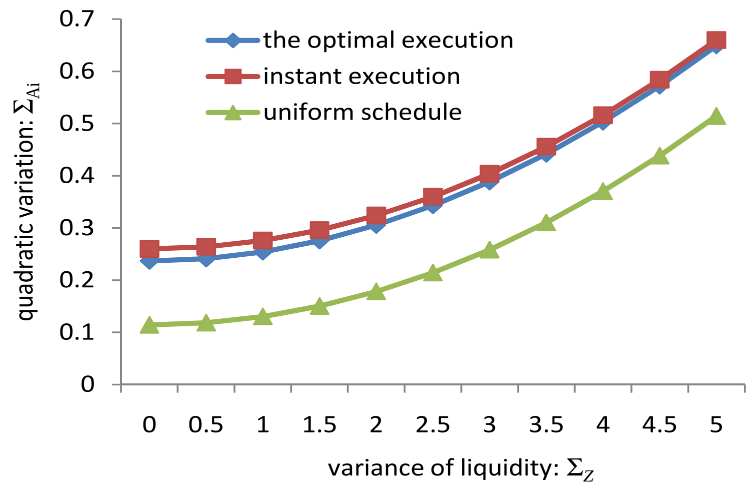
With respect to our future work, we propose to pursue detailed analyses of the optimal execution strategy in the multiasset case. In particular, we intend to examine more closely the relationship of optimal strategies to price manipulation. We should also contemplate the generalization of the shape of the order book. Obtaining a market-adaptive execution schedule by solving the problem dynamically is yet another challenge.



(a) The level of quadratic variation by resilience a



(b) The level of quadratic variation by resilience α



(c) The level of quadratic variation by variance of liquidity Σ_Z

Figure 7: The level of quadratic variation by parameters for the single asset case

Note: The other parameters are $T = 10,\ a = 0.9,\ \lambda = 0.7,\ \alpha = 0.2,\ \omega = 10,\ \rho_\Delta = 1,\ \rho_Z = 1,\ \Sigma_\Delta = 1$, and $\Sigma_Z = 1$.

Appendix Proofs

Proof of Theorem 1

We first show that there exists a minimum of the objective function:

$$H(\boldsymbol{\xi}) = \mathrm{E}\left[I_2(\boldsymbol{\xi})\right] + \lambda[\mathrm{V}\left[I_1(\boldsymbol{\xi})\right] + \mathrm{V}\left[I_2(\boldsymbol{\xi})\right]] \tag{A.1}$$

in Eq.(2.14). By rearranging the terms, we can rewrite Eq.(2.12) as:

$$\begin{aligned}
\mathrm{E}\left[I_2(\boldsymbol{\xi})\right] &= \sum_{i=1}^{N} \alpha_i \sum_{k=0}^{K} \left(2 \sum_{s=0}^{k-1} a^{k-s} \xi_s^i + \xi_k^i \right) \xi_k^i \\
&= \sum_{i=1}^{N} \alpha_i \left\{ (1-a^2) \sum_{m=0}^{K-1} \left(\sum_{l=0}^{m} a^{m-l} \xi_l^i \right)^2 + \left(\sum_{l=0}^{K} a^{K-l} \xi_l^i \right)^2 \right\}.
\end{aligned} \tag{A.2}$$

When $\|\boldsymbol{\xi}\| \to \infty$, where $\|\boldsymbol{\xi}\|$ denotes the quadratic norm of $\boldsymbol{\xi}$, there exists at least one ξ_k^i such that $|\xi_k^i| \to \infty$. It then follows from Eqs.(A.1) and (A.2) that $\lim_{\|\boldsymbol{\xi}\| \to \infty} H(\boldsymbol{\xi}) = \infty$. Because $H(\boldsymbol{\xi})$ is continuous and nonnegative, there exists a minimum of $H(\boldsymbol{\xi})$.

We recall that the first-order optimality condition given in Eq.(3.2) is a system of linear equations. This means that the optimal solution is unique provided that the coefficient matrix is invertible. If this does not hold, all we have to do is to change one of the coefficients slightly so as to satisfy invertibility. Therefore, the optimization problem is reduced to finding the explicit solution to Eq.(3.2).

For this purpose, we define an operator $\mathcal{D}$ by:

$$\mathcal{D}x_k = ax_{k-1} - (1+a)^2 x_k + 2(1+a+a^2)x_{k+1} - (1+a)^2 x_{k+2} + ax_{k+3}.$$

Using some algebra, $\mathcal{D}$ is shown to satisfy:

$$\mathcal{D}\left\{ \sum_{s=0}^{k-1} a^{k-s} \boldsymbol{\xi}_s + \sum_{s=k}^{K} a^{s-k} \boldsymbol{\xi}_s \right\} = (1-a^2)(-\boldsymbol{\xi}_k + 2\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_{k+2}), \tag{A.3}$$

$$\mathcal{D}\left\{ \sum_{s=1}^{k} \sum_{u=s}^{K} \boldsymbol{\xi}_u \right\} = -a\boldsymbol{\xi}_k + (1+a^2)\boldsymbol{\xi}_{k+1} - a\boldsymbol{\xi}_{k+2}, \tag{A.4}$$

$$\mathcal{D}\left\{ \sum_{s=1}^{k} a^{k-s} \sum_{u=s}^{K} a^{u-s} \boldsymbol{\xi}_u \right\} = -\boldsymbol{\xi}_k + 2\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_{k+2}. \tag{A.5}$$

Applying Eqs.(A.3)$\sim$(A.5) to Eq.(3.2), we obtain:

$$\boldsymbol{A}\boldsymbol{\xi}_{k+2} - \{2\boldsymbol{A} + (1-a)^2 \lambda \boldsymbol{\Sigma}_\Delta\}\boldsymbol{\xi}_{k+1} + \boldsymbol{A}\boldsymbol{\xi}_k = \mathbf{0}, \quad (k = 1, \ldots, K-3), \tag{A.6}$$

where $\mathbf{0}$ denotes a zero vector. By premultiplying $\boldsymbol{R}^{-1}\boldsymbol{A}^{-1}$ to Eq.(A.6), and using Eq.(3.6), we obtain:

$$\overline{\boldsymbol{\xi}}_{k+2} - (2\boldsymbol{I} + \boldsymbol{\Gamma})\overline{\boldsymbol{\xi}}_{k+1} + \overline{\boldsymbol{\xi}}_k = \mathbf{0}, \quad (k = 1, \ldots, K-3), \tag{A.7}$$

where $\overline{\boldsymbol{\xi}}_k = (\overline{\xi}_k^1, \ldots, \overline{\xi}_K^N)^\top = \boldsymbol{R}^{-1}\boldsymbol{\xi}_k$. Given that $\boldsymbol{\Gamma}$ is diagonal, Eq.(A.7) can be written elementwise as:

$$\overline{\xi}_{k+2}^i - (\gamma_i + 2)\overline{\xi}_{k+1}^i + \overline{\xi}_k^i = 0, \quad (i = 1, \ldots, N,\ k = 1, \ldots, K-3). \tag{A.8}$$

A general solution to Eq.(A.8) is: $\overline{\xi}_k^i = c_i\theta_i^k + d_i\theta_i^{-k}$ with θ_i defined in Eq.(3.7), which is expressed in a matrix form as:

$$\overline{\boldsymbol{\xi}}_k = \boldsymbol{\Theta}^k\boldsymbol{c} + \boldsymbol{\Theta}^{-k}\boldsymbol{d}, \quad (k = 1, \ldots, K-1). \tag{A.9}$$

Here, $\boldsymbol{c} = (c_1, \ldots, c_N)^\top$ and $\boldsymbol{d} = (d_1, \ldots, d_N)^\top$ are unknown coefficients determined by the boundary conditions.

To obtain $\boldsymbol{c}$ and $\boldsymbol{d}$, we define another difference operator $\widetilde{\mathcal{D}}x_k = -ax_{k-1} + (a+1)x_k - x_{k+1}$. Applying $\widetilde{\mathcal{D}}$ to Eq.(3.2) at $k = K-1$ yields:

$$\boldsymbol{A}\boldsymbol{\xi}_{K-1} = (1-a)\{\boldsymbol{A} + (1-a)\lambda\boldsymbol{\Sigma}_\Delta\}\boldsymbol{\xi}_K. \tag{A.10}$$

Thus, premultiplying $\boldsymbol{R}^{-1}\boldsymbol{A}^{-1}$ to Eq.(A.10) and rearranging terms, we get the boundary condition:

$$\overline{\boldsymbol{\xi}}_{K-1} = \{(1-a)\boldsymbol{I} + \boldsymbol{\Gamma}\}\overline{\boldsymbol{\xi}}_K. \tag{A.11}$$

Similarly, applying $\widetilde{\mathcal{D}}$ to Eq.(3.2) at $k = K-2$ and premultiplying by: $\boldsymbol{R}^{-1}\boldsymbol{A}^{-1}$ together with Eq.(A.11) yields:

$$\overline{\boldsymbol{\xi}}_{K-2} = (\boldsymbol{I} + \boldsymbol{\Gamma})\overline{\boldsymbol{\xi}}_{K-1} + \boldsymbol{\Gamma}\overline{\boldsymbol{\xi}}_K. \tag{A.12}$$

Substituting Eq.(A.9) into Eq.(A.12) and using Eq.(A.11), we obtain:

$$\begin{bmatrix} \boldsymbol{\Theta}^{-1} & \boldsymbol{\Theta} \\ \boldsymbol{\Theta}^{-2} & \boldsymbol{\Theta}^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Theta}^K\boldsymbol{c} \\ \boldsymbol{\Theta}^{-K}\boldsymbol{d} \end{bmatrix} = \begin{bmatrix} (1-a)\boldsymbol{I} + \boldsymbol{\Gamma} \\ (\boldsymbol{I} + \boldsymbol{\Gamma})\{(1-a)\boldsymbol{I} + \boldsymbol{\Gamma}\} + \boldsymbol{\Gamma} \end{bmatrix} \overline{\boldsymbol{\xi}}_K.$$

After some manipulation, we obtain:

$$\begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Theta}^{-K}(\boldsymbol{I} - a\boldsymbol{\Theta}) \\ \boldsymbol{\Theta}^K(\boldsymbol{\Theta} - a\boldsymbol{I}) \end{bmatrix} (\boldsymbol{I} + \boldsymbol{\Theta})^{-1}\boldsymbol{R}^{-1}\boldsymbol{\xi}_K, \tag{A.13}$$

where we use the relation:

$$\begin{bmatrix} \boldsymbol{\Theta}^2 & -\boldsymbol{\Theta} \\ -\boldsymbol{\Theta}^{-2} & \boldsymbol{\Theta}^{-1} \end{bmatrix} \begin{bmatrix} (1-a)\boldsymbol{I} + \boldsymbol{\Gamma} \\ (\boldsymbol{I} + \boldsymbol{\Gamma})\{(1-a)\boldsymbol{I} + \boldsymbol{\Gamma}\} + \boldsymbol{\Gamma} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{I} - \boldsymbol{\Theta}^{-1})(\boldsymbol{I} - a\boldsymbol{\Theta}) \\ (\boldsymbol{I} - \boldsymbol{\Theta}^{-1})(\boldsymbol{\Theta} - a\boldsymbol{I}) \end{bmatrix},$$

because $\boldsymbol{\Gamma} = \boldsymbol{\Theta} - 2\boldsymbol{I} + \boldsymbol{\Theta}^{-1}$, and because $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$ are commutative. Eqs.(A.9) and (A.13) prove:

$$\boldsymbol{\xi}_k = (\boldsymbol{I}+\boldsymbol{\Theta})^{-1}\left\{(\boldsymbol{I} - a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+k} + (\boldsymbol{\Theta} - a\boldsymbol{I})\boldsymbol{\Theta}^{K-k}\right\}\boldsymbol{\xi}_K, \quad (k = 1, \ldots, K-1). \tag{A.14}$$

We subtract Eq.(3.2) at $k=1$ from Eq.(3.2) at $k=0$ to obtain:

$$(1-a)\boldsymbol{\alpha}\boldsymbol{\xi}_0 = (1-a)\boldsymbol{\alpha}\sum_{k=1}^{K} a^{k-1}\boldsymbol{\xi}_k + \lambda\boldsymbol{\Sigma}_\Delta\sum_{k=1}^{K}\boldsymbol{\xi}_k + 4\lambda\boldsymbol{\alpha}\boldsymbol{\Sigma}_Z\boldsymbol{\alpha}\sum_{k=1}^{K} a^{k-1}\boldsymbol{\xi}_k. \tag{A.15}$$

Premultiplying Eq.(A.15) by $\{(1-a)\boldsymbol{\alpha}\}^{-1}$ and adding $\sum_{k=1}^{K}\boldsymbol{\xi}_k$ to both sides, we obtain:

$$\boldsymbol{\omega} = \left\{\boldsymbol{I} + \frac{4\lambda}{1-a}\boldsymbol{\Sigma}_Z\boldsymbol{\alpha}\right\}\sum_{k=1}^{K} a^{k-1}\boldsymbol{\xi}_k + \left\{\boldsymbol{I} + \frac{\lambda}{1-a}\boldsymbol{\alpha}^{-1}\boldsymbol{\Sigma}_\Delta\right\}\sum_{k=1}^{K}\boldsymbol{\xi}_k. \tag{A.16}$$

From Eq.(A.14), we obtain:

$$\sum_{k=1}^{K} a^{k-1}\boldsymbol{\xi}_k = (\boldsymbol{I}+\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^K + \boldsymbol{\Theta}^{-K+1})\boldsymbol{\xi}_K, \tag{A.17}$$

$$\sum_{k=1}^{K}\boldsymbol{\xi}_k = \left\{(\boldsymbol{I}-a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+1} - (\boldsymbol{\Theta}-a\boldsymbol{I})\boldsymbol{\Theta}^K\right\}(\boldsymbol{I}-\boldsymbol{\Theta}^2)^{-1}\boldsymbol{\xi}_K. \tag{A.18}$$

Substituting Eqs.(A.17) and (A.18) into Eq.(A.16), we finally obtain:

$$\begin{aligned}\boldsymbol{\omega} = & \left[\left(\boldsymbol{I} + \frac{\lambda}{1-a}\boldsymbol{\alpha}^{-1}\boldsymbol{\Sigma}_\Delta\right)\left\{(\boldsymbol{I}-a\boldsymbol{\Theta})\boldsymbol{\Theta}^{-K+1} - (\boldsymbol{\Theta}-a\boldsymbol{I})\boldsymbol{\Theta}^K\right\}\right.\\ & \left. + \left(\boldsymbol{I} + \frac{4\lambda}{1-a}\boldsymbol{\Sigma}_Z\boldsymbol{\alpha}\right)(\boldsymbol{I}-\boldsymbol{\Theta})\left(\boldsymbol{\Theta}^K + \boldsymbol{\Theta}^{-K+1}\right)\right](\boldsymbol{I}-\boldsymbol{\Theta}^2)^{-1}\boldsymbol{\xi}_K,\end{aligned}$$

which proves Eq.(3.8) for $\boldsymbol{\xi}_K^*$. Eq.(3.10) for $\boldsymbol{\xi}_k^*$ $(k=1,\ldots,K-1)$ is then easily obtained from Eq.(A.14). Finally, we substitute Eqs.(3.8) and (3.10) into Eq.(A.15), then Eq.(3.9) for $\boldsymbol{\xi}_0^*$ is obtained after some algebra. This completes the proof of Theorem 1. □

Proof of Theorem 2

When $\boldsymbol{\Sigma}_\Delta = \boldsymbol{O}$, Eqs.(A.6), (A.10) and (A.15) are reduced to:

$$\boldsymbol{A}\boldsymbol{\xi}_{k+2} - 2\boldsymbol{A}\boldsymbol{\xi}_{k+1} + \boldsymbol{A}\boldsymbol{\xi}_k = \boldsymbol{0}, \quad (k=1,\ldots,K-3), \tag{A.19}$$

$$\boldsymbol{A}\boldsymbol{\xi}_{K-1} = (1-a)\boldsymbol{A}\boldsymbol{\xi}_K, \tag{A.20}$$

$$(1-a)\boldsymbol{\alpha}\boldsymbol{\xi}_0 = (1-a)\boldsymbol{\alpha}\sum_{k=1}^{K} a^{k-1}\boldsymbol{\xi}_k + 4\lambda\boldsymbol{\alpha}\boldsymbol{\Sigma}_Z\boldsymbol{\alpha}\sum_{k=1}^{K} a^{k-1}\boldsymbol{\xi}_k, \tag{A.21}$$

respectively, where $\boldsymbol{A} = (1-a^2)\boldsymbol{\alpha} + 4\lambda\boldsymbol{\alpha}\boldsymbol{\Sigma}_Z\boldsymbol{\alpha}$. Because $\boldsymbol{A}$ is invertible, Eqs.(A.19) and (A.20) imply:

$$\boldsymbol{\xi}_{k+2} - \boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k, \quad (k=1,\ldots,K-3) \tag{A.22}$$

$$\boldsymbol{\xi}_{K-1} = (1-a)\boldsymbol{\xi}_K. \tag{A.23}$$

Moreover, applying $\widetilde{\mathcal{D}}$ to Eq.(3.2) at $k = K - 2$ yields:

$$\boldsymbol{\xi}_{K-2} = \boldsymbol{\xi}_{K-1},$$

which together with (A.22) and (A.23) proves:

$$\boldsymbol{\xi}_k = (1-a)\boldsymbol{\xi}_K, \quad (k = 1, \ldots, K-1). \tag{A.24}$$

Solving Eqs.(A.21), (A.24) and $\sum_{k=0}^{K} \boldsymbol{\xi}_k = \boldsymbol{\omega}$, we obtain the desired result. □

References

Alfonsi, A. and A. Schied, "Optimal execution and absence of price manipulations in limit order book models," *SIAM (Society for Industrial and Applied Mathematics) Journal on Financial Mathematics*, 1, 2010, pp. 490–522.

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

Almgren, R. and N. Chriss, "Optimal execution of portfolio transactions," *Journal of Risk*, 3(2), 2001, pp. 5–39.

Bertsimas, D. and A. W. Lo, "Optimal control of execution costs," *Journal of Financial Markets*, 1, 1998, pp. 237–268.

Borkovec, M. and H. G. Heidle, "Building and evaluating a transaction cost model: A primer," *Journal of Trading*, 5(2), 2010, pp. 57–77.

Chen, J., H. Hong, M. Huang, and J. D. Kubik, "Does fund size erode mutual fund performance? The role of liquidity and organization," *The American Economic Review*, 94(4), 2004, pp. 1276–1302.

Engle, R. and R. Ferstenberg, "Execution risk: It's the same as investment risk," *Journal of Portfolio Management*, 33(2), 2007, pp. 34–44.

Gatheral, J., "No-dynamic-arbitrage and market impact," Working Paper 2008-6, Courant Institute of Mathematical Sciences, New York University, 2008.

Grinold, R. C. and R. L. Kahn, *Active Portfolio Management (2nd edition)*, McGraw-Hill, 1999.

Holthausen, R. W., R. W. Leftwich, and D. Mayers, "The effect of large block transaction on security prices: A cross-sectional analysis," *Journal of Financial Economics*, 19, 1987, pp. 237–267.

Huberman, G. and W. Stanzl, "Price manipulation and quasi-arbitrage," *Econometrica*, 72, 2004, pp. 1247–1275.

Johnson, B., *Algorithmic Trading and DMA: An Introduction to Direct Access Trading Strategies*, 4Myeloma Press, 2010.

Kissell, R., "The expanded implementation shortfall: Understanding transaction cost components," *The Journal of Trading*, 1(3), 2006, pp. 6–16.

Kissell, R., M. Glantz, and R. Malamut, "A practical framework for estimating transaction costs and developing optimal trading strategies to achieve best execution," *Financial Research Letters*, 1, 2004, pp. 35–46.

Konishi, H. and N. Makimoto, "Optimal slice of a block trade," *Journal of Risk*, 3(4), 2001, pp. 33–51.

Obizhaeva, A. and J. Wang, "Optimal trading strategy and supply/demand dynamics," Discussion paper, MIT Sloan School of Management, 2005.

Perold, A. F., "The implementation shortfall: Paper versus reality," *Journal of Portfolio Management*, 14(3), 1988, pp. 4–9.

Subramanian, A. and R. Jarrow, "The liquidity discount," *Mathematical Finance*, 11(4), 2001, pp. 447–474.

Xuemin, Y., "Liquidity, investment style, and the relation between fund size and fund performance," *Journal of Financial and Quantitative Analysis*, 43, 2008, pp. 741–767.