Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Financial Markets 16 (2013) 1–32

Journal of FINANCIAL MARKETS

www.elsevier.com/locate/finmar

Optimal trading strategy and supply/demand dynamics ☆

Anna A. Obizhaeva[a,1], Jiang Wang[b,c,d,*]

[a] *Robert H. Smith School of Business, University of Maryland, 4428 Van Munching Hall, College Park, MD 20742, USA*

[b] *Sloan School of Management, MIT, 100 Main Street, Cambridge, MA 02142, USA*

[c] *CAFR, China*

[d] *NBER, USA*

Received 28 July 2012
Available online 13 September 2012

Abstract

In this paper, we study how the intertemporal supply/demand of a security affects trading strategy. We develop a general framework for a limit order book market to capture the dynamics of supply/demand. We show that the optimal strategy to execute an order does not depend on the static properties of supply/demand such as bid–ask spread and market depth, it depends on their dynamic properties such as resilience: the speed at which supply/demand recovers to its steady state after a trade. In general, the optimal strategy is quite complex, mixing large and small trades, and can substantially lower execution cost. Large trades remove the existing liquidity to attract new liquidity, while small trades allow the trader to further absorb any incoming liquidity flow.

JEL classification: G00; G11; G12

Keywords: Liquidity; Trading; Optimal order execution; Limit order book

☆ The authors thank Robert Fernstenberg, Thierry Foucault, William Goetzmann, Bruce Lehmann, Christine Parlour, Ioanid Roşu, Dimitri Vayanos, the referee, and participants of the 2006 AFA Meetings, the 2005 EFA Meetings and seminars at MIT, Morgan Stanley, and NYSE for helpful comments.

* Corresponding author at: Sloan School of Management, MIT, 100 Main Street, Cambridge, MA 02142, USA. Tel.: +1 617 253 2632; fax: +1 617 258 6855.

E-mail addresses: obizhaeva@rhsmith.umd.edu (A.A. Obizhaeva), wangj@mit.edu (J. Wang).

[1] Tel.: +1 301 405 7934.

1386-4181/$ - see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.finmar.2012.09.001

1. Introduction

The supply/demand of financial securities is in general not perfectly elastic.[2]

What trading strategy is optimal in a market with limited supply/demand or liquidity? How do different aspects of supply/demand affect the optimal strategy? How significant are cost savings from the optimal trading strategy? Traders face these questions each time when they trade. The answers to these questions are thus essential for our understanding of how market participants behave, how liquidity is provided and consumed, how it affects security prices, and more generally, how securities markets function.

We approach this problem by focusing on the optimal strategy of a trader who has to execute an order over a given time period.[3] This problem is also referred to as the optimal execution problem.[4] Previous work has provided valuable insights about how liquidity affects trading behavior of market participants (e.g., Bertsimas and Lo, 1998; Almgren and Chriss, 1999; Huberman and Stanzl, 2005). This literature tends to view supply/demand as a static object when analyzing their effect on optimal trading strategies. In particular, it describes the demand or supply of a security facing a large trade (depending on its sign) by specifying an instantaneous price impact function (i.e., a time-insensitive demand/supply schedule). Liquidity is, however, dynamic by its nature. Our contribution is to show that it is the dynamic properties of supply/demand such as its time evolution after trades, rather than its static properties, such as spread and depth, that are central to the cost of trading and the design of optimal strategy.

We propose a general framework to model the dynamics of supply/demand. We consider a limit order book market, in which the supply/demand of a security is represented by the limit orders posted on the ‘‘book’’ and trade occurs when buy and sell orders match. We describe the shape of the limit order book and especially how it evolves over time to capture the intertemporal nature of supply/demand that a large trader faces. We choose to focus on the limit order book market merely for convenience. Our main goal is to demonstrate the importance of supply/demand dynamics in determining the optimal trading strategy, and our main conclusions remain applicable to other market structures.

Our model explicitly incorporates three basic characteristics of liquidity documented empirically: bid–ask spread, market depth, and resilience. The first two features — bid–ask spread and market depth — capture the static aspects of liquidity. They are related to the shape of the limit order book, which determines how much the current price moves in response to a trade. Bid–ask spread and market depth therefore are key for determining the transaction cost that the trader incurs upon the execution of his trades instantaneously. The third feature — resilience — reflects the dynamic aspect of liquidity. Resilience is

[2] See, for example, Scholes (1972), Shleifer (1986), Holthausen, Leftwitch, and Mayers (1987, 1990), Kaul, Mehrotra, and Morck (2000), and more recently, Greenwood (2005) for empirical evidence on imperfect elasticity in the supply/demand of individual securities. Extensive theoretical work justifies such an inelasticity based on market frictions and asymmetric information (e.g., Kyle, 1985; Grossman and Miller, 1998; Vayanos, 1999, 2001).

[3] Ideally, we should consider both the optimal size of an order and its execution, taking into account the underlying motives to trade (e.g., return, risk, preferences, and constraints) and the costs to execute trades. The diversity in trading motives makes it difficult to tackle such a problem at a general level. Given that in practice the execution of trades is often separated from the investment decisions, we focus on the execution problem as an important and integral part of a more general problem of optimal trading behavior.

[4] The relevance of this problem for practitioners is highlighted in Chan and Lakonishok (1995, 1997), Keim and Madhavan (1995, 1997), and Obizhaeva (2008), among others.

related to how future limit-order book evolves in response to the current trade. We assume that initial price impact gradually dissipates over time as new liquidity providers step in to replenish the book. The further away the current quotes are from steady-state levels, the more aggressive liquidity providers post new orders.

We show that the optimal strategy crucially depends on the dynamic properties of the limit order book. The strategy consists of an initial large trade, followed by a sequence of small trades, and a final discrete trade to finish the order. The combination of large and small trades for the optimal execution strategy is in a sharp contrast to the simple strategies of splitting a order evenly into small trades, as suggested in previous studies (e.g., Bertsimas and Lo, 1998; Almgren and Chriss, 1999). The intuition behind the complex trading pattern is simple. The initial large trade is aimed at pushing the limit order book away from its steady state in order to attract new liquidity providers. The size of the large trade is chosen optimally to draw sufficient number of new orders while not incurring too high transaction costs. The subsequent small trades then pick off incoming orders and keep the inflow at desirable prices. A final discrete trade finishes off any remaining order at the end of the trading horizon when future demand/supply is no longer of concern.

Surprisingly, the optimal strategy and the cost saving depend primarily on the dynamic properties of supply/demand and is not very sensitive to their static properties described by instantaneous price-impact function, which has been the main focus in previous work. In particular, the speed at which the limit order book rebuilds itself after being hit by a trade, i.e., the resilience of the book or its replenish rate, plays a critical role in determining the optimal execution strategy and the cost it saves.

Moreover, we find that the cost savings from the optimal execution strategy can be substantial. As an illustration, let us consider the execution of an order of the size 20 times the market depth within a one-day horizon. Under the formulation of static supply/demand function in Bertsimas and Lo (1998) and Almgren and Chriss (1999), the proposed strategy is to spread the order evenly over time. However, when we take into account the dynamics of supply/demand, in particular the half-time for the limit-order book to recover after being hit by trades, the execution cost of the order under the optimal strategy is lower than the even strategy. For example, if the half-life for the book to recover is 0.90 minutes, which is relatively short, the cost saving is 0.33%. It becomes 1.88% when the half-life of recovery is 5.40 minutes and 7.41% when the half-life of recovery is 27.03 minutes. Clearly, cost savings increase and become substantial when the book's recovery time increases.

Many authors have studied the problem of optimal order execution. For example, Bertsimas and Lo (1998) propose a linear price impact function and solve for the optimal execution strategy to minimize the expected cost of executing a given order. Almgren and Chriss (1999, 2000) include risk considerations in a similar setting.[5] The framework used in these studies relies on static price impact functions at a set of fixed trading times. Fixing trading times is clearly undesirable because the timing of trades is an important choice variable and should be determined optimally. More importantly, the pre-specified static price impact functions fail to capture the intertemporal nature of supply/demand. They ignore how the path of trades influences the future evolution of the book. For example, Bertsimas and Lo (1998) assume a linear static price impact function. Consequently, the overall price impact of a sequence of trades depends only on their total size and is

[5]See also Grinold and Kahn (2000), Subramanian and Jarrow (2001), Dubil (2002), and Almgren (2003), among others.

independent of their distribution over time. Moreover, the execution cost becomes strategy independent when more frequent trades are allowed. Almgren and Chriss (1999, 2000) and Huberman and Stanzl (2005) introduce a temporary price impact as a modification, which depends on the pace of trades. Temporary price impact adds a dynamic element to the price impact function by penalizing speedy trades. This approach, however, restricts the execution strategy to continuous trades, which is in general sub-optimal.

What the previous analysis does not fully capture is how liquidity replenishes in the market, as well as how it interacts with trades. Our framework explicitly describes this process by directly modeling the book dynamics in a limit order book market, which, as we show, is critical in determining the optimal execution strategy.[6]

Our description of the limit order book dynamics relies on an extensive empirical literature. For example, using data from the Paris Bourse, Biais, Hillion and Spatt (1995) have shown empirically that market resilience is finite (e.g., Coppejans, Domowitz, and Madhavan, 2004, Ranaldo, 2004, Degryse, De Jong, Van Ravenswaaij, and Wuyts, 2005, Large, 2007, and Kempf, Mayston, and Yadav, 2009).

In addition to the empirical evidence, the dynamic behavior of the book we try to capture is also consistent with the equilibrium models of the limit order book markets. The idea of liquidity being consumed by a trade and then replenished as additional liquidity providers attempt to benefit is behind most of these models. For example, Foucault (1999), Foucault, Kadan, and Kandel (2005), and Goettler, Parlour, and Rajan (2005) build theoretical models of limit-order book markets, which exhibit different but finite levels of resilience in equilibrium, depending on the characteristics of market participants.[7] The level of resilience reflects the amount of hidden liquidity in the market. Our framework allows us to capture this dynamic aspect of the supply/demand in a flexible way and to examine the optimal execution strategy under more realistic market conditions.

Our analysis is partial equilibrium in nature, taking the dynamics of the limit order book as given. Although we do not attempt to provide an equilibrium justification for the specific limit order book dynamics used in the paper, our framework allows more general dynamics. In follow-up research, several authors have used this framework to incorporate richer book behavior. For example, Alfonsi et al., (2010) consider general, but continuous shapes of the limit order book and Predoiu, Shaikhet, and Shreve (2010) allow discrete orders and more general dynamics. Endogenizing the limit order book dynamics in a full equilibrium setting is certainly desirable, but challenging. Existing equilibrium models, such as those mentioned above, have to severely limit the set of admissible order-placement strategies. For example, Foucault, Kadan, and Kandel (2005), Roşu (2008, 2009) only allow orders of a fixed size and Goettler, Parlour, and Rajan (2005) focus on one-shot strategies. These simplifications are helpful in obtaining certain simple properties of the book, but they are quite restrictive when analyzing the optimal trading strategy. A more general and realistic equilibrium model must allow general strategies. From this

[6]In concurrent work, Esser and Monch (2005) also consider the effect of finite market resilience. But instead of considering the optimal strategy in the general strategy space, they only consider iceberg strategies.

[7]Back and Baruch (2007) consider a full equilibrium model of a limit order market in which an insider trades strategically with liquidity traders who choose between block orders or working orders to save cost. See, also, Goettler, Parlour, and Rajan (2009). What we focus on, as well as the papers mentioned above, is the interaction among the liquidity traders. In particular, we look at the interaction between a large strategic liquidity trader and a set of small non-strategic liquidity providers, whose behavior is described by the dynamics of the book in a reduced form.

perspective, our analysis, namely solving the optimal execution strategy under general supply/demand dynamics, is a key step in this direction.

The rest of the paper is organized as follows. Section 2 states the optimal execution problem. Section 3 introduces a limit order book framework. Section 4 shows that the conventional setting in previous work can be viewed as a special case of our framework, involving unrealistic assumptions and undesirable properties. Section 5 provides the solution for a problem in the discrete time. Section 6 provides the solution for a problem in the continuous time. Section 7 analyzes the properties and cost savings of optimal strategies. Section 8 discusses extensions. Section 9 concludes. All proofs are given in the Appendix.

2. Statement of the problem

The problem we are interested in is how a trader optimally executes a given order. We assume that the trader has to buy X_0 units of a security over a fixed time period $[0,T]$. Suppose that the trader completes the order in $N+1$ trades at times $t_0, t_1, \ldots, t_N$, where $t_0 = 0$ and $t_N = T$. Let x_{t_n} denote the trade size for the trade at t_n. We then have

$$\sum_{n=0}^{N} x_{t_n} = X_0. \tag{1}$$

A strategy to execute the order is given by the number of trades, $N+1$, the set of times to trade, $\{0 \le t_0, t_1, \ldots, t_{N-1}, t_N \le T\}$ and trade sizes $\{x_{t_0}, x_{t_1}, \ldots, x_{t_N} : x_{t_n} \ge 0\ \forall n \text{ and } (1)\}$. Let Θ_D denote the set of these strategies:

$$\Theta_D = \left\{ \{x_{t_0}, x_{t_1}, \ldots, x_{t_N}\} : 0 \le t_0, t_1, \ldots, t_N \le T; x_{t_n} \ge 0 \forall n; \sum_{n=0}^{N} x_{t_n} = X_0 \right\}. \tag{2}$$

Here, we have assumed that the strategy set consists of execution strategies with a finite number of trades at discrete times. This is done merely for easy comparison with previous work. Later we will expand the strategy set to allow an uncountable number of trades over time as well (Section 6).

Let $\overline{P}_n$ denote the average execution price for trade x_{t_n}. The trader chooses his execution strategy over a given trading horizon T to minimize the expected total cost of his purchase:

$$\min_{x \in \Theta_D} \mathrm{E}_0 \left[\sum_{n=0}^{N} \overline{P}_n x_n \right]. \tag{3}$$

This objective function implies that the risk-neutral trader cares only about the expected value but not the uncertainty of the total cost. Later, we will further incorporate risk considerations (in Section 8).

It is important to recognize that the execution price $\overline{P}_n$ for trade x_n in general will depend not only on x_n, the current trade size, but also all past trades. Such a dependence reflects two dimensions of the price impact of trading. First, it changes the security's *current* supply/demand. For example, after a purchase of x units of the security at the current price of $\overline{P}$, the remaining supply of the security at $\overline{P}$ usually decreases. Second, a change in current supply/demand can affect *future* supply/demand and therefore the costs for future trades. In other words, the price impact is determined by the full dynamics of supply/demand in response to a trade. In order to fully specify and solve the optimal execution problem, we thus need to properly model the supply/demand dynamics.

3. Limit order book and supply/demand dynamics

The actual supply/demand of a security and its dynamics depend on the actual trading process. From different markets, the trading process varies significantly, ranging from a specialist market, a dealer market to a centralized electronic market with a limit order book. In this paper, we consider the limit order book market. However, our analysis is of a general nature, and we expect our results to be relevant for other market structures as well.

3.1. Limit order book

A limit order is an order to trade a certain number of shares of a security at a given price. In a market operated through a limit order book (LOB), traders post their supply/demand in the form of limit orders to an electronic trading system.[8] A trade occurs when an order, say a buy order, enters the system at the price of an opposite order on the book, in this case a sell order. The collection of all limit orders posted can be viewed as the total demand and supply in the market.

Let $q_A(P)$ be the density of limit orders to sell at price P, and let $q_B(P)$ be the density of limit orders to buy at price P. The number of sell orders in a small price interval $[P, P+dP)$ is $q_A(P)\,dP$. Typically, we have

$$q_A(P) = \begin{cases} +, & P \geq A \\ 0, & P < A \end{cases} \quad \text{and} \quad q_B(P) = \begin{cases} 0, & P > B, \\ +, & P \leq B, \end{cases}$$

where $A \geq B$ are the best ask and bid prices, respectively. We define

$$V = (A+B)/2, \quad s = A - B, \tag{4}$$

where V is the mid-quote price and s is the bid–ask spread. Then, $A = V + s/2$ and $B = V - s/2$. Because we are considering the execution of a large buy order, we focus on the upper half of the LOB and simply drop the subscript A.

In order to model the execution cost for a large order, we need to specify the initial LOB and how it evolves after been hit by a series of buy trades. Let the LOB (the upper half of it) at time t be $q(P; F_t; Z_t; t)$, where F_t denotes the fundamental value of the security and Z_t represents the set of state variables that may affect the LOB such as past trades. We consider here a simple model for the LOB that captures its dynamic nature. This model allows us to illustrate the importance of supply/demand dynamics for analyzing the optimal execution problem. We discuss below how to extend this model to better fit the empirical LOB dynamics (Section 8).

[8] The number of exchanges adopting electronic trading platforms has been increasing. Examples for the stock market include NYSE's OpenBook program, NASDAQ's SuperMontage, Toronto Stock Exchange, Vancouver Stock Exchange, Euronext (Paris, Amsterdam, Brussels), London Stock Exchange, Copenhagen Stock Exchange, Deutsche Borse, and Electronic Communication Networks. Examples for the fixed income market include eSpeed, Euro MTS, BondLink, and BondNet. Examples for the derivatives market include Eurex, Globex, ISE, and Matif.

The fundamental value F_t follows a Brownian motion reflecting the fact that, in the absence of any trades, the mid-quote price may change due to news about the fundamental value. Thus, $V_t = F_t$ in the absence of any trades, and the LOB maintains the same shape except that the mid-point, V_t, is changing with F_t. For simplicity, we assume that the only set of relevant state variables Z_t is the history of past trades, denoted by $x_{[0,t]}$.

At time 0, the mid-quote is $V_0 = F_0$ and the LOB has a simple block-shape density,

$$q_0(P) \equiv q(P; F_0; 0; 0) = q1_{\{P \geq A_0\}},$$

where $A_0 = F_0 + s/2$ is the initial ask price and $1_{\{z \geq a\}}$ is an indicator function:

$$1_{\{z \geq a\}} = \begin{cases} 1, & z \geq a, \\ 0, & z < a. \end{cases}$$

In other words, q_0 is a step function of P with a jump from zero to q at the ask price $A_0 = V_0 + s/2 = F_0 + s/2$. Panel A of Fig. 1 shows the shape of the book at time 0.

Now we consider a buy trade of size x_0 shares at $t=0$. The trade will "eat off" all the sell orders with prices from $F_0 + s/2$ up to A_{0_+}, where A_{0_+} is given by

$$\int_{F_0+s/2}^{A_{0_+}} q\, dP = x_0.$$

From this formula, we find that the new ask price is $A_{0_+} = F_0 + s/2 + x_0/q$. The average execution price for trade x_0 is linear in the size of trade and is equal to $\overline{P} = F_0 + s/2 + x_0/(2q)$. Thus, the block shape of the LOB is consistent with the linear price impact function assumed in previous work. This is also the main reason we adopted this specification here. Right after the trade, the limit order book is described as

$$q_{0_+}(P) \equiv q(P; F_0; Z_{0_+}; 0_+) = q1_{\{P \geq A_{0_+}\}},$$

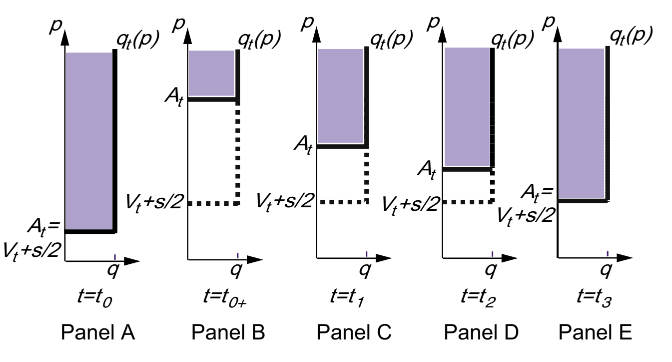

Fig. 1. The limit order book and its dynamics. This figure illustrates how the sell side of the limit-order book evolves over time in response to a buy trade. Before the trade at time $t_0 = 0$, the limit-order book is full at the ask price $A_0 = V_0 + s/2$, which is shown in the first panel from the left. The trade of size x_0 at $t=0$ "eats off" the orders on the book with the lowest prices and pushes the ask price up to $A_{0+} = (F_0 + s/2) + x_0/q$, as shown in the second panel. During the following periods, new orders will arrive at the ask price A_t. These orders fill up the book and lower the ask price until this price converges to its new steady state $A_t = F_t + \lambda x_0 + s/2$, as shown in the last panel on the right. For clarity, we assume that there are no fundamental shocks during this period.

where $A_{0_+} = F_0 + s/2 + x_0/q$ is the new ask price. Orders at prices below A_{0_+} have all been executed. The book is left with sell limit orders at prices above (including) A_{0_+}. Panel B of Fig. 1 plots the limit order book right after the trade.

3.2. Limit order book dynamics

We next specify how the LOB evolves over time after being hit by a trade. This amounts to describing how new sell limit orders arrive to fill the book. First, we need to specify the impact of the trade on the mid-quote price. Usually, the mid-quote price will be shifted up by the trade. We assume that the shift in the mid-quote price is linear in the size of the total trade. That is,

$$V_{0^+} = F_0 + \lambda x_0,$$

where $0 \le \lambda \le 1/q$ and λx_0 corresponds to the permanent price impact of trade x_0. If initial trade x_0 at $t=0$ is not followed by other trades and if there are no shocks to the fundamentals, then as time t goes to infinity, the limit order book eventually converges to its new steady state:

$$q_\infty(P) = q1_{\{P \ge A_\infty\}},$$

where the new mid-quote $V_\infty = F_0 + \lambda x_0$ and ask price $A_\infty = V_t + s/2$. Next we need to specify how the limit-order book converges to its steady state. Note that right after the trade, the ask price is $A_{0+} = F_0 + s/2 + x_0/q$, while in the steady state it is $A_\infty = F_0 + s/2 + \lambda x_0$. The difference between the two is $A_{0+} - A_\infty = x_0(1/q - \lambda)$. We assume that the limit-order book converges to its steady state exponentially,

$$q_t(P) = q1_{\{P \ge A_t\}}, \tag{5}$$

where

$$A_t = V_t + s/2 + x_0 \kappa e^{-\rho t}, \quad \kappa = 1/q - \lambda, \tag{6}$$

$V_t = V_{0_+}$ in the absence of new trades and changes in fundamental F_t, and parameter $\rho \ge 0$ corresponds to the convergence speed, which measures the "resilience" of the LOB.

If we define D_t being the deviation of current ask price A_t from its steady state level $V_t + s/2$,

$$D_t = A_t - V_t - s/2, \tag{7}$$

then Eqs. (5) and (6) imply that after a buy trade x_0, the new sell limit orders will start coming into the book at the new ask price A_t at the rate of $\rho q D_t$. Thus, the further the current ask price is from its steady state, the more aggressively liquidity providers step in and post new orders to offer replenished liquidity. Panel C to Panel E in Fig. 1 illustrate the time evolution of the LOB after a buy trade.

We can easily extend the LOB dynamics to allow multiple trades and shocks to the fundamental value. Let $n(t)$ denote the number of trades during interval $[0,t)$. Define a trading sequence with $n(t)$ trades at times $t_1, \ldots, t_{n(t)}$ of size x_{t_i}. Let X_t be the remaining order to be executed at time t, before trading at time t occurs. We have $X_{T_+} = 0$ and

$$X_t = X_0 - \sum_{t_n < t} x_{t_n}. \tag{8}$$

If $X_0 - X_t$ is the total amount of purchase during $[0, t)$, then the mid-quote V_t at any time t is

$$V_t = F_t + \lambda(X_0 - X_t) = F_t + \lambda \sum_{i=0}^{n(t)} x_{t_i}. \tag{9}$$

The ask price at any time t is

$$A_t = V_t + s/2 + \sum_{i=0}^{n(t)} x_{t_i} \kappa e^{-\rho(t-t_i)}, \tag{10}$$

and the limit order book is given by (5). The above description can be extended to include sell orders, which may occur in the meantime shifting the mid-quote V_t. But if they are not predictable, we can simply omit them, as they will not affect our analysis.

Before we go ahead with the LOB dynamics and examine its implications for execution strategy, several comments are in order. First, the key feature of the LOB is its finite resilience, which is captured by ρ, the refresh rate of the book. This is motivated by a range of empirical evidence such as those documented in Biais, Hillion, and Spatt (1995), Hamao and Hasbrouck (1995), and Coppejans, Domowitz, and Madhavan (2004), among others. Second, although the LOB dynamics specified here is taken as given, without additional equilibrium justification, its qualitative behavior, namely, the finite resilience, is consistent with those obtained in simple equilibrium models of LOB markets considered by Foucault, Kadan, and Kandel (2005) and Goettler, Parlou, and Rajan (2005).[9] Third, existing equilibrium models are inadequate for analyzing the problem of execution as they limit the admissible strategies severely by restricting trade size and frequency. Thus, in order to develop a full equilibrium model for the execution problem, we first need to know its solution under general demand/supply dynamics and then arrive at equilibrium dynamics. From this point of view, this paper focuses on the first part of this undertaking. Fourth, our setting is very flexible in allowing an arbitrary shape of the book and rich dynamics for its time evolution in response to an arbitrary set of trades. Since the main goal of this paper is to demonstrate the importance of supply/demand dynamics in determining the optimal trading behavior rather than obtaining a general solution to the problem, we narrow down our analysis to a specific case of the general setting. The qualitative conclusions we obtain from the simple case remain robust when more general forms of the book and its dynamics are allowed, as follow-up research has shown (e.g., Alfonsi et al., 2010).

3.3. Execution cost

Given the LOB dynamics, we can describe the total cost of an execution strategy for a given order X_0. Let x_{t_n} denote the trade at time t_n, and A_{t_n} denote the ask price at time t_n prior to this trade. Since the evolution of the ask price A_t in (10) is not continuous, we denote by A_t the left limit of A_t, $A_t = \lim_{s \to t^-} A_s$, i.e., the ask price before the trade at time t.

[9] The literature on the strategic behavior of traders submitting limit orders include Glosten (1994), Rock (1996), Seppi (1997), Harris (1998), Parlour (1998), Foucault (1999), Sandas (2001), Parlour and Seppi (2003), Roşu (2008, 2009). Cao, Hansch, and Wang (2003), Bloomfield, O'Hara, and Saar (2004), Ranaldo (2004), and Harris and Panchapagesan (2005), also show that traders indeed use the rich information revealed by the book when deciding on their order submissions.

The same convention is followed for V_t as well. The cost for a single trade x_{t_n} is then given by

$$c(x_{t_n}) = \int_0^{x_{t_n}} P_{t_n}(x)\,dx, \tag{11}$$

where $P_t(x)$ is defined by equation:

$$x = \int_{A_t}^{P_t(x)} q_t(P)\,dP. \tag{12}$$

For the block-shaped LOB given in (5), we have $P_t(x) = A_t + x/q$ and

$$c(x_{t_n}) = [A_{t_n} + x_{t_n}/(2q)]x_{t_n}. \tag{13}$$

The total cost of $N+1$ trades of size x_{t_n}, $n = 0,1,\ldots,N$, is $\sum_{n=0}^{N} c(x_{t_n})$. Thus, the optimal execution problem (3) is reduced to

$$\min_{x \in \Theta_D} \mathrm{E}_0 \left[\sum_{n=0}^{N} [A_{t_n} + x_{t_n}/(2q)]x_{t_n} \right], \tag{14}$$

under the LOB dynamics given in (9) and (10).

4. Conventional models as a special case

Previous work on the optimal execution strategy usually uses a discrete-time setting with fixed time intervals (e.g., Bertsimas and Lo, 1998; Almgren and Chriss, 1999, 2000). Such a setting, however, avoids the question of how to determine the optimal trading times. In this section, we show that it represents a special case of our framework with specific restrictions on the LOB dynamics, which lead to crucial limitations.

4.1. Conventional setup

We first consider a simple discrete-time setting proposed by Bertsimas and Lo (1998), which captures the basic features of the models used in earlier work.

In such a setting, the trader trades at fixed equally spaced time intervals, $n\tau$, where $\tau = T/N$ and $n = 0,1,\ldots,N$, while trading horizon T and the number of trades N are given. Each trade has an impact on the price, which will affect the total cost of the trade and all future trades. Most models assume a linear price-impact function of the following form:

$$\overline{P}_n = \overline{P}_{n-1} + \lambda x_n + u_n = (F_n + s/2) + \lambda \sum_{i=0}^{n} x_i, \tag{15}$$

where the subscript n denotes the n-th trade at $t_n = n\tau$, $\overline{P}_n$ is the average price at which trade x_n is executed with $\overline{P}_{0\_} = F_0 + s/2$, λ is the price impact coefficient, and u_n is an i.i.d. random variable with a mean of zero and a variance of $\sigma^2\tau$. These assumptions are reasonable given the conclusion of Huberman and Stanzl (2004) that in the absence of quasi-arbitrage, permanent price-impact functions must be linear. In the second equation, we have set $F_n = F_0 + \sum_{i=0}^{n} u_i$. Parameter λ captures the permanent price impact of a trade.

The trader wishing to execute an order of size X_0 solves the following problem:

$$\min_{\{x_0,x_1,\ldots,x_N\}} \mathrm{E}_0\left[\sum_{n=0}^{N} \overline{P}_n x_n\right] = (F_0 + s/2)X_0 + \lambda \sum_{n=0}^{N} X_n(X_{n+1} - X_n), \tag{16}$$

where $\overline{P}_n$ is defined in (15) and X_n is a number of shares left to be acquired at time t_n (before trade x_{t_n}) with $X_{N+1} = 0$.

As shown in Bertsimas and Lo (1998), given that the objective function is quadratic in x_n, it is optimal for the trader to split his order into small trades of equal sizes and execute them at regular intervals over the fixed period of time:

$$x_n = \frac{X_0}{N+1}, \tag{17}$$

where $n = 0, 1, \ldots, N$.

4.2. The continuous-time limit

Although the discrete-time setting with a linear price impact function gives a simple and intuitive solution, it leaves a key question unanswered, namely, what determines the time-interval between trades. An intuitive way to address this question is to take the continuous-time limit of the discrete-time solution (i.e., to let N go to infinity). However, as Huberman and Stanzl (2005) point out, the solution to the discrete-time model (16) does not have a well-defined continuous-time limit. In fact, as $N \to \infty$, the cost of the trades as given in (16) approaches the following limit of:

$$(F_0 + s/2)X_0 + (\lambda/2)X_0^2.$$

This limit depends only on the total trade size X_0 and not on the actual trading strategy itself. Thus, for a risk-neutral trader, the execution cost with continuous trading is a fixed number and any continuous strategy is as good as another. Consequently, the discrete-time model does not have a well-behaved continuous-time limit.[10] The intuition is that a trader can simply walk up the supply curve, and the speed of his trading is irrelevant. Without increasing the cost, the trader can choose to trade intensely at the very beginning and complete the whole order in an arbitrarily small period. For example, if the trader becomes slightly risk averse, he will choose to finish all the trades right at the beginning, irrespective of their price impact.[11] Such a situation is clearly undesirable and economically unreasonable.

[10] In taking the continuous-time limit, we have held λ constant. This is, of course, unrealistic. For different τ, λ can well be different. But the problem remains as long as λ has a finite limit when $\tau \to 0$.

[11] As $N \to \infty$, the objective function to be minimized for a risk-averse trader with a mean–variance preference approaches the following limit:

$$C(x_{[0,T]}) = \mathrm{E}\left[\int_0^T P_t\, dX_t\right] + \frac{1}{2}a\,\mathrm{Var}\left[\int_0^T P_t\, dX_t\right] = (F_0 + s/2)X_0 + (\lambda/2)X_0^2 + \frac{1}{2}a\sigma^2 \int_0^T X_t^2\, dt,$$

where $a > 0$ is the risk-aversion coefficient and σ is the price volatility. The trader cares not only about expected execution costs but also its variance, which is given by the last term. Only variance of the execution cost depends on the strategy. The optimal strategy is to choose an L-shaped profile for the trades, i.e., to trade with infinite speed at the beginning, thus making the variance term zero.

4.3. A special case of our framework

We can see the limitations of the conventional model by considering it as a special case of our framework. Indeed, we can specify the parameters in the LOB framework so that it will be equivalent to the conventional setting. First, we set the trading times at fixed intervals: $t_n = n\tau$, $n = 0, 1, \ldots, N$. Next, we make the following assumptions about the LOB dynamics as described in (5) and (9):

$$q = 1/(2\lambda), \quad \lambda = \lambda, \quad \rho = \infty, \tag{18}$$

where the second equation simply states that the price impact coefficient in the LOB framework is set to be equal to its counterpart in the conventional setting. These restrictions imply the following dynamics for the LOB. As it follows from (10), after the trade x_n at t_n ($t_n = n\tau$), the ask price A_{t_n} jumps from level $V_{t_n} + s/2$ to level $V_{t_n} + s/2 + 2\lambda x_n$. Since resilience is infinite, over the next period, ask price comes all the way down to the new steady state level of $V_{t_n} + s/2 + \lambda x_n$ (assuming no fundamental shocks from t_n to t_{n+1}). Thus, the dynamics of ask price A_{t_n} is equivalent to the dynamics of $\overline{P}_{t_n}$ in (15).

For the parameters in (18), the cost for trade x_{t_n} is given in (13), which becomes

$$c(x_{t_n}) = [F_{t_n} + s/2 + \lambda(X_0 - X_{t_n}) + \lambda x_{t_n}]x_{t_n},$$

which is the same as the trading cost in the conventional model (16). Thus, the conventional model is a special case of the LOB framework with the parameters in (18).

The main restrictive assumption we have to make to obtain the conventional setup is $\rho = \infty$. This assumption means that the LOB always converges to its steady state before the next trading time. This is not crucial if the time between trades is held fixed. If the time between trades is allowed to shrink, this assumption becomes unrealistic. It takes time for the new limit orders to come in to fill up the book again. In reality, the shape of the limit order book after a trade depends on the flow of new orders as well as the time elapsed. As the time between trades shrinks to zero, the assumption of infinite recovery speed becomes less reasonable and gives rise to the problems in the continuous-time limit of the conventional model.

4.4. Temporary price impact

This problem has led several authors to modify the conventional setting. He and Mamaysky (2005), for example, directly formulate the problem in continuous-time and impose fixed transaction costs to rule out any continuous trading strategies. Similar to the more general price impact function considered by Almgren and Chriss (1999, 2000) and Huberman and Stanzl (2005) proposes a temporary price impact of a particular form to penalize high-intensity continuous trading. Both of these modifications limit us to a subset of feasible strategies, which is in general sub-optimal. Given its closeness to our paper, we now briefly discuss the modification with a temporary price impact.

Almgren and Chriss (1999, 2000) include a temporary component in the price impact function, which can depend on the trading interval τ. The temporary price impact gives additional flexibility in dealing with the continuous-time limit of the problem. In particular, they specify the following dynamics for the execution prices of trades:

$$\hat{P}_n = \overline{P}_n + G(x_n/\tau), \tag{19}$$

where $\overline{P}_n$ is the same as given in (15), $\tau = T/N$ is the time between trades, and $G(\cdot)$ describes a temporary price impact and reflects temporary price deviations from "equilibrium" caused by trading. With $G(0)=0$ and $G'(\cdot)>0$, the temporary price impact penalizes high trading volume per unit of time, x_n/τ. Using a linear form for $G(\cdot)$, $G(z)=\theta z$, it is easy to show that as N goes to infinity, the expected execution cost approaches to

$$(F_0 + s/2)X_0 + (\lambda/2)X_0^2 + \theta \int_0^T \left(\frac{dX_t}{dt}\right)^2 dt$$

(e.g., Grinold and Kahn, 2000; Huberman and Stanzl, 2005). Clearly, with the temporary price impact, the optimal execution strategy has a continuous-time limit. In fact, it is very similar to its discrete-time counterpart: This strategy is deterministic and the trading intensity, defined by the limit of x_n/τ, is constant over time.[12]

The temporary price impact reflects an important aspect of the market, namely, the difference between short-term and long-term supply/demand. If a trader speeds up his buy trades, as he can do in the continuous-time limit, he will deplete the short-term supply and increase the immediate cost for additional trades. As more time is allowed between trades, supply will gradually recover. However, as a heuristic modification, the temporary price impact does not provide an accurate and complete description of the supply/demand dynamics. This leads to several drawbacks. For example, the temporary price impact function in the form considered by Almgren and Chriss (2000) and Huberman and Stanzl (2005) rules out the possibility of discrete trades. This is not only artificial but also undesirable. As we show later, the optimal execution strategy generally involves both discrete and continuous trades. Moreover, introducing the temporary price impact does not capture the full dynamics of supply/demand. For example, two sets of trades close to each other in time versus far apart will generate different supply/demand dynamics, while in Huberman and Stanzl (2005) they lead to the same dynamics. Finally, simply specifying a particular form for the temporary price impact function says little about the underlying economic factors that determine it.

5. Discrete-time solution

We now return to our general framework and solve for the optimal execution strategy. Suppose that trading times are fixed at $t_n = n\tau$, where $\tau = T/N$ and $n = 0,1,\ldots,N$. We consider the corresponding strategies $x_{[0,T]} = \{x_0, x_1, \ldots, x_n\}$ within the strategy set Θ_D defined in Section 2. Using (3), (9), (10) and (14), the optimal execution problem is reduced to

$$J_0 = \min_{\{x_0,\ldots,x_N\}} \quad \mathrm{E}_0\left[\sum_{n=0}^{N}[A_{t_n} + x_n/(2q)]x_n\right]$$

$$\text{s.t.} \quad A_{t_n} = F_{t_n} + \lambda(X_0 - X_{t_n}) + s/2 + \sum_{i=0}^{n-1} x_i \kappa e^{-\rho\tau(n-i)}, \tag{20}$$

where F_t follows a random walk. This problem can be solved using dynamic programming.

[12]If the trader is risk-averse with a mean–variance preference, the optimal execution strategy has a decreasing trading intensity over time. See Almgren and Chriss (2000) and Huberman and Stanzl (2005).

Proposition 1. *The solution to the optimal execution problem (20) is*

$$x_n = -\frac{1}{2}\delta_{n+1}[D_{t_n}(1-\beta_{n+1}e^{-\rho\tau}+2\kappa\gamma_{n+1}e^{-2\rho\tau})-X_{t_n}(\lambda+2\alpha_{n+1}-\beta_{n+1}\kappa e^{-\rho\tau})], \tag{21}$$

with $x_N = X_N$ and $D_t = A_t - V_t - s/2$. The expected cost for future trades under the optimal strategy is determined according to

$$J_{t_n} = (F_{t_n}+s/2)X_{t_n} + \lambda X_0 X_{t_n} + \alpha_n X_{t_n}^2 + \beta_n D_{t_n} X_{t_n} + \gamma_n D_{t_n}^2, \tag{22}$$

where the coefficients α_{n+1}, β_{n+1}, γ_{n+1}, and δ_{n+1} are determined recursively as follows:

$$\alpha_n = \alpha_{n+1} - \frac{1}{4}\delta_{n+1}(\lambda+2\alpha_{n+1}-\beta_{n+1}\kappa e^{-\rho\tau})^2, \tag{23a}$$

$$\beta_n = \beta_{n+1}e^{-\rho\tau} + \frac{1}{2}\delta_{n+1}(1-\beta_{n+1}e^{-\rho\tau}+2\kappa\gamma_{n+1}e^{-2\rho\tau})(\lambda+2\alpha_{n+1}-\beta_{n+1}\kappa e^{-\rho\tau}), \tag{23b}$$

$$\gamma_n = \gamma_{n+1}e^{-2\rho\tau} - \frac{1}{4}\delta_{n+1}(1-\beta_{n+1}e^{-\rho\tau}+2\gamma_{n+1}\kappa e^{-2\rho\tau})^2, \tag{23c}$$

with $\delta_{n+1} = [1/(2q)+\alpha_{n+1}-\beta_{n+1}\kappa e^{-\rho\tau}+\gamma_{n+1}\kappa^2 e^{-2\rho\tau}]^{-1}$ and terminal conditions

$$\alpha_N = 1/(2q)-\lambda, \quad \beta_N = 1 \quad \text{and} \quad \gamma_N = 0. \tag{24}$$

Proposition 1 describes the optimal execution strategy when we fix the trading times at a certain interval τ. This strategy is optimal only among strategies with the same fixed trading interval. In principle, we want to choose the trading interval to minimize the execution costs. One way to allow different trading intervals is to take the limit $\tau \to 0$, i.e., $N \to \infty$, in the problem (20). Fig. 2 plots optimal execution strategies $\{x_n, n=0,1,\ldots,N\}$ for different values of N: N=10, 25, 100, respectively. As N becomes large, the strategy splits into two parts, large trades at both ends of the trading horizon (at the beginning and

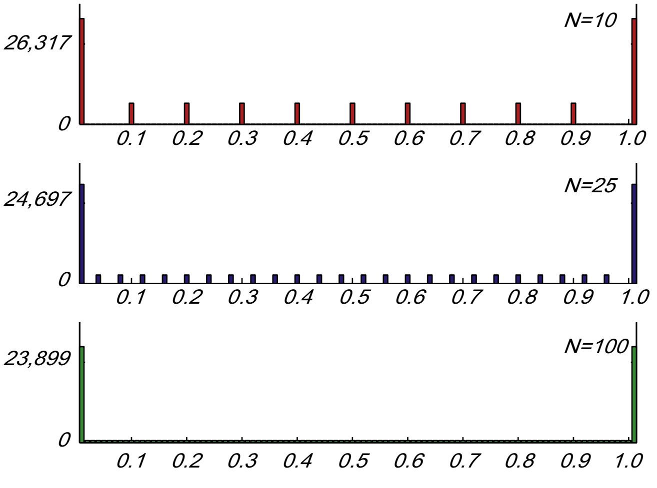

Fig. 2. Optimal strategy with N fixed discrete trading intervals. This figure plots the optimal trades for N fixed intervals, where N is 10, 25, and 100 for respectively the top, middle and bottom panels. The initial order to trade is set at $X_0 = 100{,}000$ units, the time horizon is set at T=1 day, the market depth is set at q=5,000 units, the price-impact coefficient is set at $\lambda = 1/(2q) = 10^{-4}$, and the resilience coefficient is set at $\rho = 2.231$.

at the end) and small trades in between. Clearly, these strategies are very different from the conventional strategy (17) obtained previously when the dynamics of demand/supply is ignored.

Proposition 2 describes the continuous-time limit of the optimal execution strategy and the expected cost.

Proposition 2. *As $N \to \infty$, the optimal execution strategy becomes*

$$\lim_{N\to\infty} x_0 = x_{t=0} = \frac{X_0}{\rho T + 2}, \tag{25a}$$

$$\lim_{N\to\infty} x_n/(T/N) = \dot{X}_t = \frac{\rho X_0}{\rho T + 2}, \quad t \in (0,T), \tag{25b}$$

$$\lim_{N\to\infty} x_N = x_{t=T} = \frac{X_0}{\rho T + 2}, \tag{25c}$$

where x_0 is the trade at the beginning of trading period, x_N is the trade at the end of trading period, and $\dot{X}_t$ is the speed of trading in between these trades. The expected cost is determined according to

$$J_t = (F_0 + s/2)X_t + \lambda X_0 X_t + \alpha_t X_t^2 + \beta_t X_t D_t + \gamma_t D_t^2,$$

where coefficients α_t, β_t, and γ_t are given by

$$\alpha_t = \frac{\kappa}{\rho(T-t)+2} - \frac{\lambda}{2}, \quad \beta_t = \frac{2}{\rho(T-t)+2}, \quad \gamma_t = -\frac{\rho(T-t)}{2\kappa[\rho(T-t)+2]}. \tag{26}$$

What is the intuition underlying this complex trading pattern? The initial discrete trade x_0 is aimed at pushing the limit-order book away from its steady state. This deviation makes liquidity providers to step in and place new orders onto the book. The size of discrete trade x_0 is chosen optimally to draw a sufficient number of new orders while not incurring too high transaction costs. The subsequent continuous trades then pick off incoming orders and keep the inflow coming at desirable prices. A final discrete trade x_N finishes off any remaining order at the end of trading horizon when future demand/supply is no longer of concern. In Section 7, we examine in more detail the properties of the optimal execution strategy and their dependence on the LOB dynamics.

6. Continuous-time solution

The continuous-time limit of the discrete-time solution suggests that limiting ourselves to discrete strategies can be suboptimal. Instead, we should formulate the problem in a continuous-time setting and allow for both continuous and discrete trading strategies. We show next how to derive the optimal strategy in the continuous-time version of the LOB framework.

Let the fundamental value be $F_t = F_0 + \sigma Z_t$, where Z_t is a standard Brownian motion defined on $[0,T]$. Variable F_t fully captures the uncertainty in the model. Let $\mathcal{F}_t$ denote the filtration generated by Z_t. A general execution strategy can consist of two components: a set of discrete trades at certain times and a flow of continuous trades. A set of discrete trades is also called an "impulse" trading policy.

Definition 1. Let $N_+ = \{1,2,\ldots\}$. An impulse trading policy $(\tau_k, x_k) : k \in N_+$ is a sequence of trading times τ_k and trade amounts x_k such that: (1) $0 \le \tau_k \le \tau_{k+1}$ for $k \in N_+$, (2) τ_k is a stopping time with respect to $\mathcal{F}_t$, and (3) x_k is measurable with respect to $\mathcal{F}_{t_k}$.

The continuous trades can be defined by a continuous trading policy described by the trading intensity $\mu_{[0,t]}$, where μ_t is measurable with respect to $\mathcal{F}_t$ and $\mu_t\, dt$ represents the trades during time interval $[t, t+dt)$. Let $\hat{T}$ denote the set of impulse trading times. Then, the set of admissible execution strategies for a buy order is

$$\Theta_C = \left\{ \mu_{[0,T]}, x_{\{t \in \hat{T}\}} : \mu_t,\ x_t \ge 0,\ \int_0^T \mu_t\, dt + \sum_{t \in \hat{T}} x_t = X_0 \right\}, \tag{27}$$

where μ_t is the rate of continuous buy trades at time t and x_t is the size of the discrete buy trade for $t \in \hat{T}$. The dynamics of X_t, the number of shares yet to acquire at time t, is then given by the following equation:

$$X_t = X_0 - \int_0^t \mu_s\, ds - \sum_{s \in \hat{T}, s < t} x_s.$$

Now let us specify the dynamics of the ask price A_t. Similar to the discrete-time setting, we have $A_0 = F_0 + s/2$ and

$$A_t = A_0 + \int_0^t [dV_s - \rho D_s\, ds - \kappa\, dX_s], \tag{28}$$

where mid-quote is $V_t = F_t + \lambda(X_0 - X_t)$ as in (9) and deviation is $D_t = A_t - V_t - s/2$ as in (7). The dynamics of ask price A_t captures the evolution of the limit-order book, in particular the changes in mid-quote V_t, the inflow of new orders, and the continuous execution of trades.

Next, we compute the execution cost consisting of two parts: the costs from continuous trades and discrete trades, respectively. The execution cost from t to T is

$$C_t = \int_t^T A_s \mu_s\, ds + \sum_{s \in \hat{T}, t \le s \le T} [A_s + x_s/(2q)] x_s. \tag{29}$$

Given the dynamics of the state variables in (9), (28), and cost function in (29), the optimal execution problem now becomes

$$J_t \equiv J(X_t, A_t, V_t, t) = \min_{\{\mu_{[0,T]}, \{x_{t \in \hat{T}}\}\} \in \Theta_C} \mathrm{E}_t[C_t], \tag{30}$$

where J_t is the value function at time t equal to the expected cost for future trades under the optimal execution strategy. At time T, the trader is forced to buy all of the remaining order X_T, which leads to the following boundary condition:

$$J_T = [A_T + 1/(2q) X_T] X_T.$$

Proposition 3 gives the solution to the problem.

Proposition 3. *The value function for the optimization problem* (30) *is*

$$J_t = (F_t + s/2) X_t + \lambda X_0 X_t + \alpha_t X_t^2 + \beta_t D_t + \gamma_t D_t^2,$$

where $D_t = A_t - V_t - s/2$. The optimal execution strategy is

$$x_0 = x_T = \frac{X_0}{\rho T + 2}, \quad \mu_t = \frac{\rho X_0}{\rho T + 2} \ \forall t \in (0, T), \tag{31}$$

where the coefficients α_t, β_t, and γ_t are the same as in Proposition 2.

The optimal strategy consists of an initial discrete trade, followed by a sequence of continuous trades, and finished with a final discrete trade. Obviously, the solution in the continuous-time setting from Proposition 3 is identical to the continuous-time limit of the solution in the discrete-time setting from Proposition 2. The optimal execution strategy is, however, different from strategies obtained in the conventional setting (17). Since the strategy involves both discrete and continuous trades, this clearly indicates that the timing of trades is a critical part of the optimal strategy. This also shows that ruling out discrete or continuous trades ex ante is in general suboptimal. Our solution demonstrates that both static and dynamic properties of supply/demand, which are captured by the LOB dynamics in our framework, are important in analyzing the optimal execution strategy and its cost.

7. Optimal strategy and cost savings

In contrast with previous work, the optimal execution strategy includes discrete and continuous trading. We now analyze the properties of the optimal execution strategy in more detail and quantify the cost reduction it accomplishes.

7.1. Properties of optimal execution strategy

The first thing to notice is that the execution strategy given in (31) does not depend on the price impact λ and market depth q. Coefficient λ captures the permanent price impact of a trade and, given a linear form of the price impact function, fully describes the instantaneous (static) supply/demand. Independence of optimal strategy on λ is a rather striking result given that most of the previous work focuses on λ as the key parameter determining the execution strategy and cost. As we show earlier, λ affects the execution strategy when the times to trade are exogenously set at fixed intervals. When the times to trade are determined optimally, the impact of λ on execution strategy disappears.

Coefficient q captures the depth of the LOB market. In the simple model we consider, it is assumed to be constant at all price levels above the ask price. In this case, the actual value of market depth does not affect the optimal execution strategy. For more general (and possibly more realistic) shapes of the limit order book, the optimal execution strategy may well depend on the static characteristics of the book. Our analysis clearly shows that the static aspects of the supply/demand does not fully capture the factors that determine the optimal execution strategy.

The optimal execution strategy depends on two parameters, the LOB resilience ρ and the execution horizon T. We consider these dependencies separately.

Panel A of Fig. 3 plots the optimal execution strategy, namely, the time path X_t of the remaining order. Clearly, the nature of the optimal strategy is different from strategies proposed in the literature and involving a smooth flow of small trades. When the timing of trades is determined optimally, the optimal execution strategy consists of both large

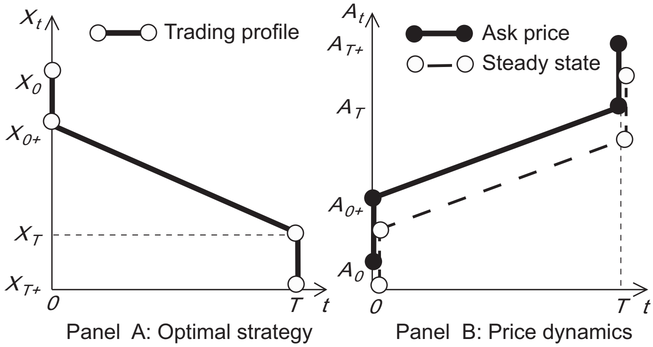

Fig. 3. Profiles of the optimal execution strategy and ask price. Panel A plots the profile of optimal execution policy as described by X_t, the number of units left to be traded at time t. Panel B plots the profile of realized ask price A_t. A discrete trade occurs at time 0 and moves the ask price up. After the initial trade, continuous trades are executed as a constant fraction of newly incoming sell orders to keep the deviation of the ask price A_t from its steady state $V_t + s/2$, shown with dashed line in Panel B, at a constant level. A discrete trade occurs at the last moment T to complete the order.

discrete trades and continuous trades. In particular, under the LOB dynamics we consider here, the optimal execution involves a discrete trade at the beginning, followed by a flow of small trades and then a discrete terminal trade. Such a strategy seems intuitive given the dynamics of the limit order book. The initial discrete trade pushes the limit order book away from its stationary state so that new orders are lured in. The subsequent flow of small trades will "eat up" these incoming orders thus keeping them coming. At the end, a discrete trade finishes the remaining part of the order.

The size of the initial trade is chosen optimally to draw sufficient number of new orders while not incurring too high transaction costs. If the initial trade is too large, then it will raise the average prices of the new orders. If the initial trade too small, then it will not lure in enough orders before the terminal time. The trade off between these two factors largely determines the size of the initial trade.

The subsequent continuous trades are intended to maintain the flow of new limit orders at desirable prices. To see how this works, let us consider the path of the ask price A_t under the optimal execution strategy. It is plotted in Panel B of Fig. 3. The initial discrete trade consumes the liquidity by "eating up" bottom limit orders and pushes up the ask price from its initial level of $A_0 = V_0 + s/2$ to its new level of $A_{0+} = V_0 + s/2 + X_0/(\rho T + 2)/q$. Afterwards, the optimal execution strategy keeps $D_t = A_t - V_t - s/2$, the deviation of the current ask price A_t from its steady state level $V_t + s/2$, at a constant level of $\kappa X_0/(\rho T + 2)$. Consequently, the rate of new sell order flow, which is given by $\rho \times D_t$, is also maintained at a constant level. The ask price A_t goes up together with $V_t + s/2$, the steady-state "value" of the security, which is shown with the dashed line in Fig. 3(b). Plugging $dA_t = dV_t$ for $0 < t < T$ into (28), we find that $\rho D_t = \kappa \mu_t$. In other words, $\mu_t = (1/\kappa)\rho D_t$, implying that under the optimal execution strategy, a constant fraction of $1/\kappa$ of the new sell orders is executed to maintain a constant order flow.

The final discrete trade is determined by two factors. First, the order has to be completed within the given horizon. Second, the evolution of supply/demand afterwards no longer matters. In practice, both of these factors can take different forms. For example, the trading horizon T can be endogenously determined rather than exogenously given. We consider this extension by allowing risk considerations in Section 8.

Our discussion above shows that the LOB dynamics captured by the resilience parameter ρ is the key factor in determining optimal execution strategy. In order to better understand this link, let us consider two extreme cases, when $\rho = 0$ and $\rho = \infty$. When $\rho = 0$, we have no recovery of the limit order book after a trade. The execution costs will be strategy independent; it does not matter when and at what speed the trader eats up the limit-order book. This result is also true in a discrete setting for any value of N, as well as in its continuous-time limit. When $\rho = \infty$, the limit order book rebuilds itself immediately after a trade. As we discussed in Section 4, this case corresponds to the conventional setting. Again, the execution cost becomes strategy independent. It should be pointed out that even though in the limit of $\rho \to 0$ or $\rho \to \infty$, the optimal execution strategy given in Proposition 3 converges to a pure discrete strategy or a pure continuous strategy, many other strategies are equally good given the degeneracy in these two cases.

When the resilience of the limit order book is finite, $0 < \rho < \infty$, the optimal strategy is a mixture of discrete and continuous trades. The fraction of the total order executed through continuous trades is $\int_0^T \mu_t \, dt / X_0 = \rho T / (\rho T + 2)$. This fraction increases with ρ. It is more efficient to use small trades when the limit order book is more resilient. The intuition is that, given a larger resilience, smaller discrete trades are required to lure the same amount of new order flows, against which one can take full advantage by trading continuously.

Another important parameter in determining the optimal execution strategy is the time-horizon T to complete the order. From Proposition 3, we see that as T increases, the size of the two discrete trades decreases. This result is intuitive. The more time we have to execute the order, the more we can spread continuous trades to benefit from the inflow of new orders mitigating the total cost.

7.2. Cost savings

So far, we have focused on the optimal execution strategy. We now turn to the savings the optimal strategy can yield. For this purpose, we use the strategy obtained in the conventional setting and its cost as the benchmark. As shown in Section 4, the conventional strategy is a constant flow of trades with intensity $\mu_\infty = X_0/T, \ t \in [0, T]$. The subscript ∞ denotes here the infinite resilience implicitly assumed in this setting. Under this simple strategy, we have the mid-quote $V_t = F_t + \lambda(t/T)X_0$, the deviation $D_t = [\kappa X_0/(\rho T)](1 - e^{-\rho t})$, and the ask price $A_t = V_t + D_t + s/2$. The expected net execution cost for the strategy with a constant rate of execution μ_∞ is given by

$$\tilde{J}_0^{\mathrm{CM}} = \mathrm{E}_0\left[\int_0^T (A_t - F_t - s/2)(X_0/T)\, dt\right] = (\lambda/2)X_0^2 + \kappa \frac{\rho T - (1 - e^{-\rho T})}{(\rho T)^2} X_0^2,$$

where the superscript CM in $\tilde{J}_0^{\mathrm{CM}}$ stands for the "Conventional Model." The total expected execution cost of a buy order of size X_0 is equal to its fundamental value $(F_0 + s/2)\, X_0$ plus the extra cost from the price impact of trading. Since the first term is unrelated to the execution strategy, we consider only the net cost, not including the expenses related to the fundamental value.

From Proposition 3, the net expected cost under the optimal execution strategy is

$$\tilde{J}_0 = J_0 - (F_0 + s/2)X_0 = (\lambda/2)X_0^2 + \frac{\kappa}{\rho T + 2} X_0^2.$$

The improvement in expected execution cost can be calculated as $J_0^{\mathrm{CM}} - J_0$, given by

$$\tilde{J}_0^{\mathrm{CM}} - \tilde{J}_0 = \kappa \frac{2\rho T - (\rho T + 2)(1 - e^{-\rho T})}{(\rho T + 2)(\rho T)^2} X_0^2.$$

It can be shown that this improvement is always non-negative. The relative gain in execution quality can be defined as $\Delta = (\tilde{J}^{\mathrm{CM}} - \tilde{J}_0)/\tilde{J}^{\mathrm{CM}}$.

In order to calibrate the magnitude of costs reduction by the optimal execution strategy, we consider several numerical examples. Let the size of the order to be executed be $X_0 = 100{,}000$ shares and the initial security price be $A_0 = F_0 + s/2 = \$100$. We choose the width of the limit order book, which gives the depth of the market, to be $q = 5{,}000$. This implies that if the order is executed at once, the ask price will move up by 20%. Without loss of generality, we consider the execution horizon to be one day, $T = 1$. The other parameters, especially ρ, may well depend on the security under consideration. We will analyze how the optimal strategy and its cost savings depend on a range of values for resilience ρ and price impact λ.

Table 1 reports the numerical values of the optimal execution strategy for different values of ρ. As discussed above, for small values of ρ, most of the order is executed through two discrete trades, while for large values of ρ, most of the order is executed through a flow of continuous trades as in the conventional models. For intermediate ranges of ρ, a mixture of discrete and continuous trades is used.

Table 2 reports the relative improvement in the expected net execution cost by the optimal execution strategy over the simple strategy of the conventional setting. Let us first consider the extreme case in which the resilience of the LOB is very small, e.g., $\rho = 0.001$ and the half-life for the LOB to rebuild itself after being hit by a trade is 693.15 days. In this case, even though the optimal execution strategy looks very different from the simple execution strategy, as shown in Fig. 4, the improvement in execution cost is minuscule. This is not surprising as we

Table 1
Profiles of the optimal execution strategy for different levels of LOB resilience ρ.

ρ	Half-life ($\log 2/\rho$)	Trade x_0	Trade over $(0, T)$	Trade x_N
0.001	693.15 day	49,975	50	49,975
0.01	69.31 day	49,751	498	49,751
0.5	1.39 day	40,000	20,000	40,000
1	270.33 minutes	33,333	33,334	33,333
2	135.16 minutes	25,000	50,000	25,000
4	67.58 minutes	16,667	66,666	16,667
5	54.07 minutes	14,286	71,428	14,286
10	27.03 minutes	8,333	83,334	8,333
20	13.52 minutes	4,545	90,910	4,545
50	5.40 minutes	1,921	96,153	1,921
300	0.90 minutes	331	99,338	331
1000	0.20 minutes	100	99,800	100
10000	0.03 minutes	10	99,980	10

The table reports values of optimal discrete trades x_0 and x_T at the beginning and the end of the trading horizon and the intensity of continuous trades in between for an order of $X_0 = 100{,}000$ for different values of the LOB resilience parameter ρ or the half-life of an LOB disturbance $\tau_{1/2}$, which is defined as $\exp\{-\rho\tau_{1/2}\} = 1/2$. The initial ask price is \$100, the market depth is set at $q = 5{,}000$ units, the (permanent) price-impact coefficient is set at $\lambda = 1/(2q) = 10^{-4}$, and the trading horizon is set at $T = 1$ day, which is 6.5 hours (390 minutes).

Table 2
Cost savings by the optimal execution strategy from the simple trading strategy.

ρ	Half-life	λ				
		$\frac{1}{2q}$	$\frac{1}{10q}$	$\frac{1}{50q}$	$\frac{1}{100q}$	0
0.001	693.15 day	0.00	0.01	0.02	0.02	0.02
0.01	69.31 day	0.08	0.15	0.16	0.16	0.17
0.5	1.39 day	2.82	5.42	5.99	6.06	6.13
1	270.33 minutes	3.98	8.16	9.14	9.26	9.39
2	135.16 minutes	4.32	9.97	11.51	11.71	11.92
4	67.58 minutes	3.19	9.00	11.05	11.35	11.65
5	54.07 minutes	2.64	8.07	10.21	10.53	10.86
10	27.03 minutes	1.13	4.58	6.65	7.01	7.41
20	13.52 minutes	0.37	1.98	3.54	3.89	4.31
50	5.40 minutes	0.07	0.49	1.24	1.50	1.88
300	0.90 minutes	0.00	0.02	0.08	0.13	0.33
1000	0.20 minutes	0.00	0.00	0.01	0.02	0.10
10000	0.03 minutes	0.00	0.00	0.00	0.00	0.09

Relative improvement in expected net execution cost $\Delta = (\tilde{J}^{CM} - \tilde{J}_0)/\tilde{J}^{CM}$ is reported for different values of LOB resilience coefficient ρ and the permanent price-impact coefficient λ. The order size is set at 100,000, the market depth is set at q=5,000, and the horizon for execution is set at T=1 day (equivalent of 390 minutes).

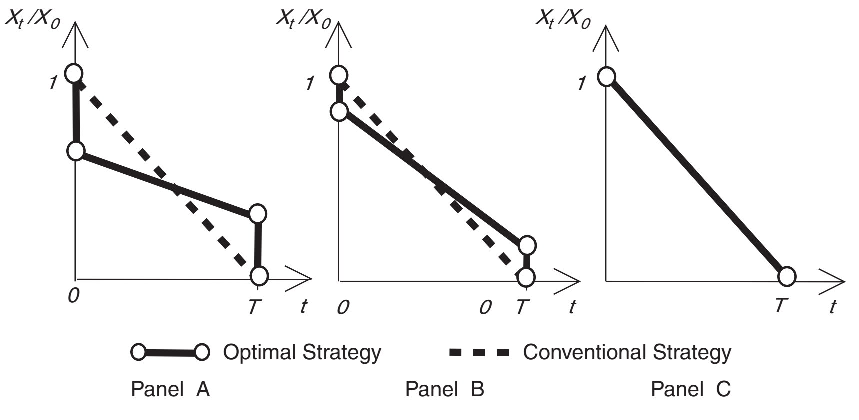

Fig. 4. Optimal strategy versus simple strategy from the conventional models. The figure plots the time paths of remaining order to be executed for the optimal strategy (solid line) and the simple strategy obtained from the conventional models (dashed line), respectively. The order size is set at $X_0 = 100{,}000$, the initial ask price is set at \$100, the market depth is set at q=5,000 units, the (permanent) price-impact coefficient is set at $\lambda = 1/(2q) = 10^{-4}$, and the trading horizon is set at T=1 day, which is assumed to be 6.5 hours (390 minutes). Panels A, B, and C plot the strategies for $\rho = 0.001$,2 and 1,000, respectively.

know the execution cost becomes strategy independent when $\rho = 0$. For a modest value of ρ, e.g., $\rho = 2$ with a half life of 135 minutes (2 hours and 15 minutes), the improvement in execution cost ranges from 4.32% for $\lambda = 1/(2q)$ to 11.92% for $\lambda = 0$. When ρ becomes large and the LOB becomes very resilient, e.g., $\rho = 300$ and the half-life of LOB deviation is 0.90 minute, the improvement in execution cost becomes small again, with a maximum of 0.33%

when $\lambda = 0$. This is again expected as we know that the simple strategy is close to the optimal strategy when $\rho \to \infty$ (as in this limit, the cost becomes strategy independent).

Table 2 also reveals an interesting result. The relative savings in execution cost by the optimal execution strategy is the highest when $\lambda = 0$, i.e., when the permanent price impact is zero. Of course, the magnitude of net execution cost becomes very small as λ goes to zero.[13]

In order to see the difference between the optimal strategy and the simple strategy obtained in conventional settings, we compare their profiles X_t in Fig. 4. The solid line shows the optimal execution strategy of the LOB framework and the dashed line shows the execution strategy of the conventional setting. Obviously, the difference between the two strategies are more significant for smaller values of ρ.

8. Extensions

We have used a parsimonious LOB model to analyze the impact of supply/dynamics on optimal execution strategy. Obviously, the simple characteristics of the model does not reflect the richness in the LOB dynamics observed in the market. The framework we developed, however, is quite flexible to allow for extensions in various directions. In this section, we briefly discuss some of them.

8.1. Time varying LOB resilience

Our model can easily incorporate time variation in LOB resilience. It has been documented that trading volume, order flows, and transaction costs all exhibit U-shaped intraday patterns. These variables are high at the opening of the trading day, then fall to lower levels during the day and finally rise again towards the close of a trading day. This suggests that the liquidity in the market may well vary over a trading day. Monch (2004) has attempted to incorporate such a time-variation in the conventional models.

We can easily allow for deterministic time variation in LOB dynamics. In particular, we can allow the resilience coefficient to be time dependent, $\rho = \rho_t$ for $t \in [0, T]$. The results in Propositions 1–3 still hold if we replace ρ by ρ_t, ρT by $\int_0^T \rho_t \, dt$, and $\rho(T-t)$ by $\int_t^T \rho_t \, dt$.

8.2. Different shapes for LOB

We have considered a simple shape for the LOB described by a step function with the constant density of limit orders placed at various price levels. As shown in Section 3, this form of the LOB is consistent with the static linear price-impact function widely used in the literature. Although Huberman and Stanzl (2004) have provided theoretical arguments in support of the linear price impact functions, the empirical literature has suggested that the shape of the LOB can be more complex (e.g., Hopman, 2007). Addressing this issue, we can allow more general shapes of the LOB in our framework. This will also make the LOB dynamics more convoluted. As a trade eats away the tip of the LOB, we have to specify how the LOB converges to its steady state. With a complicated shape for the LOB, this convergence process can take many forms. Modeling more complex shapes of the LOB

[13] When X_0 is big, the execution costs are largely determined by substantial costs related to the permanent price impact when it is present. Any cost reductions due to optimal dealing with the temporary price impact will seem small as a percentage of total costs.

involves assumptions about the flow of new orders at a range of prices. Recently, Alfonsi, Schied, and Schulz (2009) extended our analysis to LOB with a general density of placed limit orders. Remarkably, the authors find a close-form solution for a broad class of limit-order books and show that the suggested optimal strategies are qualitatively similar to those derived for a block-shaped LOB. Their findings thus confirm the robustness of our results.

8.3. Risk aversion

We have considered the optimal execution problem for a risk-neutral trader. We can extend our framework to consider the optimal execution problem for a risk-averse trader as well. For tractability, we assume that this trade has a mean-variance objective function with a risk-aversion coefficient of a. The optimization problem (30) now becomes

$$J_t \equiv J(X_t, A_t, V_t, t) = \min_{\{\mu_{[0,T]}, \{x_{t\in\hat{T}}\}\}\in\Theta_C} \mathrm{E}_t[C_t] + \frac{1}{2} a\, \mathrm{Var}_t[C_t], \tag{32}$$

with (9), (28), and (29). At time T, the trader is forced to buy all of the remaining order X_T. This leads to the following boundary condition:

$$J_T = [A_T + 1/(2q) X_T] X_T.$$

Since the only source of uncertainty in (32) is F_t and only the trades executed in interval $[t, t+dt)$ will be subject to uncertainty in F_t, we can rewrite this formula in a more convenient form:

$$J_t = \min_{\{\mu_{[0,T]}, \{x_{t\in\hat{T}}\}\}\in\Theta_C} \mathrm{E}_t[C_t] + \frac{1}{2} a \int_t^T \sigma^2 X_s^2\, ds. \tag{33}$$

Proposition 4 gives the solution to the problem for a risk-averse trader:

Proposition 4. *The optimal execution strategy for the optimization problem (33) is*

$$x_0 = X_0 \frac{\kappa f'(0) + a\sigma^2}{\kappa\rho f(0) + a\sigma^2},$$

$$\mu_t = \kappa x_0 \frac{\rho g(t) - g'(t)}{1 + \kappa g(t)} e^{-\int_0^t ((\kappa g'(s) + \rho)/(1+\kappa g(s)))\, ds}, \quad \forall t \in (0, T),$$

$$x_T = X_0 - x_0 - \int_0^T \mu_s\, ds.$$

The value function is determined by

$$J_t = (F_t + s/2) X_t + \lambda X_0 X_t + \alpha_t X_t^2 + \beta_t D_t + \gamma_t D_t^2,$$

where $D_t = A_t - V_t - s/2$. *The coefficients* $\alpha_t, \beta_t, \gamma_t$ *are given by*

$$\alpha_t = \frac{\kappa f(t) - \lambda}{2}, \quad \beta_t = f(t), \quad \gamma_t = \frac{f(t) - 1}{2\kappa},$$

where functions $f(t)$ *and* $g(t)$ *are defined as*

$$f(t) = (v - a\sigma^2)/(\kappa\rho) + \left[-\frac{\kappa\rho}{2v} + e^{(2\rho v/(2\kappa\rho + a\sigma^2))(T-t)} \left(\frac{\kappa\rho}{2v} - \frac{\kappa\rho}{v - a\sigma^2 - \kappa\rho} \right) \right]^{-1},$$

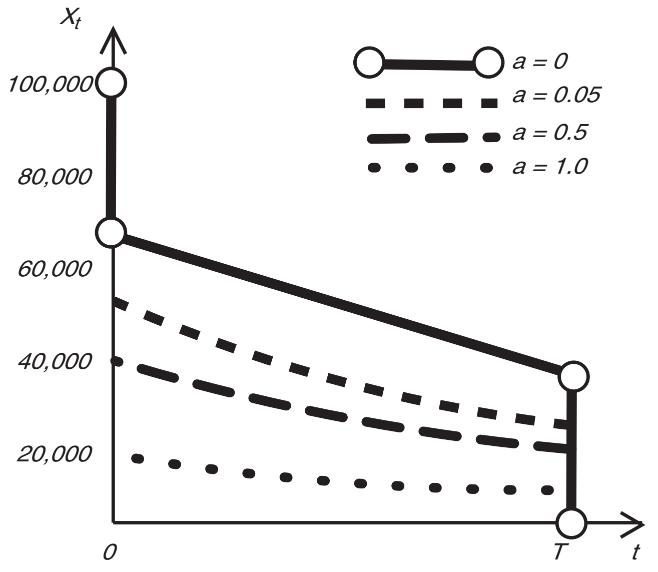

Fig. 5. Profiles of optimal strategies for different coefficients of risk aversion a. This figure shows the profiles of optimal execution policies X_t for the traders with different coefficients of risk aversion $a=0$ (solid line), $a=0.05$ (dashed line), and $a=0.5$ (dashed-dotted line) and $a=1$ (dotted line), respectively. The variable X_t indicates how much shares still has to be executed before trading at time t. The order size is set at $X_0=100{,}000$, the market depth is set at $q=5{,}000$ units, the permanent price-impact coefficient is set at $\lambda=0$, the trading horizon is set at $T=1$, and the resilience coefficient is set at $\rho=1$.

$$g(t)=-\frac{f'(t)-\rho f(t)}{kf'(t)+a\sigma^2}$$

with $v=\sqrt{a^2\sigma^4+2a\sigma^2\kappa\rho}$.

It can be shown that as the risk aversion coefficient a goes to 0, the coefficients α_t, β_t, and γ_t converge to those in Proposition 2 that were obtained for a risk-neutral trader. The nature of the execution strategy that is optimal for a risk-averse trader remains qualitatively similar to the strategy that is optimal for a risk-neutral trader. A risk-averse trader will place discrete trades at the beginning and at the end of trading period and trade continuously in between. The initial and final discrete trades are, however, of different magnitude. The more risk averse the trader is, the faster he wants to execute his order to avoid future uncertainty and the more aggressive orders he submits in the beginning. The effect of trader's risk aversion a on the optimal trading profile is shown in Fig. 5.

9. Conclusion

In this paper, we examine how the limited elasticity of the supply/demand of a security affects trading behavior of market participants. Our main goal is to demonstrate the importance of supply/demand dynamics in determining optimal trading strategies.

The execution of orders is usually not costless. The execution prices are different from pre-trade benchmarks, since implemented transactions consume liquidity and change the remaining supply/demand. The supply/demand schedule right after a transaction will be determined by its static properties. Furthermore, trades often trigger a complex evolution of supply/demand. Rather then being permanent, its initial changes may partially dissipate over time as liquidity providers step in and replenish liquidity. Thus, supply/demand represents a complex object in the marketplace that changes in response to executed trades. While designing trading strategies traders have to take into account a full dynamics of supply/demand since their transactions are often spread over time.

In this paper, we focus on the optimal execution problem faced by a trader who wishes to execute a large order over a given period of time. We explicitly model supply/demand as a limit order book market. The shape of a limit-order book determines static properties of supply/demand such as bid–ask spread and price impact. The dynamics of a limit order book in response to trades determines its dynamic properties such as resilience. We are interested in how various aspects of liquidity influence trading strategies. We show that when trading times are chosen optimally, the resilience is the key factor in determining the optimal execution strategy. The strategy involves discrete trades as well as continuous trades, instead of merely continuous trades as in previous work that focuses only on price impact and spread. The intuition is that traders can use discrete orders to aggressively consume available liquidity and induce liquidity providers to step in and place new orders into the trading system, thus making the execution of future trades cheaper. The developed framework for supply/demand is based on the limit order book market for convenience. Our main conclusions remain applicable to any other market structures. The framework is fairly general to accommodate rich forms of supply/demand dynamics. It represents a convenient tool for those who wish to fine-tune their trading strategies to realistic dynamics of supply/demand in the marketplace.

Appendix A

A.1. Proof of Proposition 1

From (7), we have

$$D_{t_n} = A_{t_n} - V_{t_n} - s/2 = \sum_{i=0}^{n-1} x_{t_i} \kappa e^{-\rho\tau(n-i)}. \tag{A.1}$$

From (A.1), the dynamics of D_t between trades will be

$$D_{t_{n+1}} = (D_{t_n} + x_{t_n}\kappa)e^{-\rho\tau}, \tag{A.2}$$

with $D_0 = 0$. We can then express the optimal execution problem (20) in terms of X_t and D_t:

$$\min_{x \in \Theta_D} \mathrm{E}_0 \sum_{n=0}^{N} [(F_{t_n} + s/2) + \lambda(X_0 - X_{t_n}) + D_{t_n} + x_{t_n}/(2q)] x_{t_n}, \tag{A.3}$$

under the dynamics of D_t given by (A.2).

First, by induction we prove that the value function for (A.3) is quadratic in X_t and D_t. It has a form implied by (22):

$$J(X_{t_n},D_{t_n},F_{t_n},t_n)=(F_{t_n}+s/2)X_{t_n}+\lambda X_0X_{t_n}+\alpha_nX_{t_n}^2+\beta_nX_{t_n}D_{t_n}+\gamma_nD_{t_n}^2. \tag{A.4}$$

At time $t=t_N=T$, the trader has to finish the order and the cost is

$$J(X_T,D_T,F_T,T)=(F_T+s/2)X_T+[\lambda(X_0-X_T)+D_T+X_T/(2q)]X_T.$$

Hence, $\alpha_N=1/(2q)-\lambda$, $\beta_N=1$, $\gamma_N=0$. Recursively, the Bellman equation yields

$$\begin{aligned}J_{t_{n-1}}=\min_{x_{n-1}}\{&[(F_{t_{n-1}}+s/2)+\lambda(X_0-X_{t_{n-1}})+D_{t_{n-1}}+x_{n-1}/(2q)]x_{n-1}\\&+\mathrm{E}_{t_{n-1}}J[X_{t_{n-1}}-x_{n-1},(D_{t_{n-1}}+\kappa x_{n-1})e^{-\rho\tau},F_{t_n},t_n]\}.\end{aligned}$$

Since F_{t_n} follows a Brownian motion and the value function is linear in F_{t_n}, it follows that the optimal trade size x_{n-1} is a linear function of $X_{t_{n-1}}$ and $D_{t_{n-1}}$ and the value function is a quadratic function of $X_{t_{n-1}}$ and $D_{t_{n-1}}$, satisfying (A.4), which leads to the recursive equation (23) for the coefficients.

A.2. Proof of Proposition 2

First, we prove the convergence of the value function. As $\tau=T/N\to 0$, the first order approximation of the system (23) in τ leads to the following restrictions on the coefficients:

$$\lambda+2\alpha_t-\beta_t\kappa=0,\quad 1-\beta_t+2\kappa\gamma_t=0 \tag{A.5}$$

and

$$\dot{\alpha}_t=\frac{1}{4}\kappa\rho\beta_t^2,\quad \dot{\beta}_t=\rho\beta_t-\frac{1}{2}\rho\beta_t(\beta_t-4\kappa\gamma_t),\quad \dot{\gamma}_t=2\rho\gamma_t+\frac{1}{4\kappa}\rho(\beta_t-4\kappa\gamma_t)^2. \tag{A.6}$$

It is easy to verify that α_t, β_t, and γ_t given in (26) provide the solution of (A.6), satisfying (A.5) and (24). Thus, as $\tau\to 0$, the coefficients of the value function (23) converges to (26).

Second, we prove the convergence result for the optimal execution policy $\{x_t\}$. Substituting α_t, β_t, and γ_t into (21), we can show that as $\tau\to 0$, the execution policy converges to

$$x_t=\left\{X_t\frac{1}{\rho(T-t)+2}-D_t\frac{1+\rho(T-t)}{\kappa[\rho(T-t)+2]}\right\}\left[1-\frac{1}{2}\rho^2(T-t)\tau\right]+\frac{1}{2}(\rho/\kappa)D_t\tau+\mathrm{o}(\tau), \tag{A.7}$$

where $\mathrm{o}(\tau)$ denotes the terms of the higher orders of τ. At $t=0$, $D_0=0$ and we have $\lim_{\tau\to 0}x_0=X_0/(\rho T+2)$. Moreover, after the initial discrete trade x_0 at time 0, all trades will be small (except possibly for the trade at time T) and equal to

$$x_t=\frac{1}{\kappa}\rho D_t\tau+\mathrm{o}(\tau),\quad t=n\tau,\ n=1,\ldots,N-1. \tag{A.8}$$

We prove this by induction. First, using (A.7), where $X_\tau=X_0-x_0$ and $D_\tau=kx_0(1-\rho\tau)$, we check that (A.8) holds for x_τ. Second, it we assume that (A.8) holds for some x_t, where $t=n\tau$, then we can show that $x_{t+\tau}$ will satisfy this condition as well. In fact, the dynamics of X_t and D_t is defined by

$$X_{t+\tau}=X_t-x_t,\quad D_{t+\tau}=(D_t+kx_t)(1-\rho\tau),\ \ t=n\tau,\ n=0,\ldots,N-1. \tag{A.9}$$

Substituting these equations into (A.7) and using the induction assumption, we get

$$x_{t+\tau} = (\rho/\kappa)D_{t+\tau}\tau + \mathrm{o}(\tau).$$

After the discrete trade x_0 at time $t=0$, the consequent trades will be continuous. Moreover, (A.8) implies the following form of X_t and D_t dynamics:

$$X_{t+\tau} = X_t - \frac{1}{\kappa}\rho D_t\tau + \mathrm{o}(\tau), \quad D_{t+\tau} = D_t + \mathrm{o}(\tau). \tag{A.10}$$

Taking into account the initial condition right after the trade at time 0, we find that

$$D_t = D_\tau = \frac{kX_0}{\rho T + 2} + \mathrm{o}(\tau).$$

From (A.8) as $\tau \to 0$ for any $t \in (0, T)$, the trade size x_t converges to $(\rho X_0/(\rho T+2))\tau$. Since all shares X_0 should be acquired by time T, it is obvious that $\lim_{\tau\to 0} x_T = X_0/(\rho T+2)$.

A.3. Proof of Propositions 3 and 4

We give the proof of Proposition 4 along with the proof of Proposition 3 as a special case. Let us first formulate the problem (33) in terms of variables X_t and $D_t = A_t - V_t - s/2$ whose dynamics, similar to (A.2), is

$$dD_t = -\rho D_t\, dt - \kappa\, dX_t, \tag{A.11}$$

with $D_0 = 0$. If we write down the cost of continuous and discrete trading as following:

$$dC_t^c = (F_t + s/2)\mu_t\, dt + \lambda(X_0 - X_t)\mu_t\, dt + D_t\mu_t\, dt, \tag{A.12}$$

$$\Delta C_t^d = 1_{\{t\in\hat{T}\}}[(F_t + s/2)x_t + \lambda(X_0 - X_t)x_t + D_t x_t + x_t^2/(2q)], \tag{A.13}$$

then (33) is equivalent to

$$\min_{\{\mu_{[0,T]},\{x_{t\in\hat{T}}\}\}\in\Theta_C} \mathrm{E}_t\left[\int_0^t dC_t^c + \sum_{t\in\hat{T}}\Delta C_t^d\right] + (a/2)\int_t^T \sigma^2 X_s^2\, ds, \tag{A.14}$$

with (A.11)–(A.13).

This is the optimal control problem with a single control variable X_t. We can apply standard methods to find its solution. In particular, the solution will be characterized by three regions, where it will be optimal to trade discretely, continuously, and do not trade at all. We can specify the necessary conditions for each region that any value function should satisfy. Under some regularity conditions on the value function, we can use Ito's lemma together with the dynamic programming principle to derive the Bellman equation associated with (A.14). For this problem, the Bellman equation is a variational inequality involving the first-order partial differential equation with the gradient constraints. Moreover, the value function should also satisfy the boundary conditions. Below, we heuristically derive the variational inequalities and show the candidate function that satisfies them. To prove that this function is the solution, we have to check the sufficient conditions for optimality using the verification principle.

We proceed with the proof of Proposition 4 in three steps. First, we define the variational inequalities (VIs) and the boundary conditions for the optimization problem (A.14). Second, we show that the solution to the VIs exists and implies the candidate value function and the candidate optimal strategy. Third, we verify that our candidate value function and optimal strategy are indeed solutions to the optimization problem. Finally, we discuss the properties of the optimal strategies.

A.4. Variational inequalities

Let $J(X_t,D_t,F_t,t)$ be the value function for our problem. Under some regularity conditions, it has to satisfy the necessary conditions for the optimality or the Bellman equation associated with (A.14). For this problem, the Bellman equation is the variational inequality involving the first-order partial differential equation with the gradient constraints, i.e.,

$$\min\{J_t-\rho D_t J_D+\frac{1}{2}\sigma^2 J_{FF}+a\sigma^2 X_t^2,(F_t+s/2)+\lambda(X_0-X_t)+D_t-J_X+\kappa J_D\}=0.$$

Thus, the space can be divided into three regions. In the discrete trade (DT) region, the value function J has to satisfy

$$J_t-\rho D_t J_D+\frac{1}{2}\sigma^2 J_{FF}+a\sigma^2 X_t^2>0,(F_t+s/2)+\lambda(X_0-X_t)+D_t-J_X+\kappa J_D=0. \tag{A.15}$$

In the no-trade (NT) region, the value function J satisfies

$$J_t-\rho D_t J_D+\frac{1}{2}\sigma^2 J_{FF}+a\sigma^2 X_t^2=0,(F_t+s/2)+\lambda(X_0-X_t)+D_t-J_X+\kappa J_D>0. \tag{A.16}$$

In the continuous-trade (CT) region, the value function J has to satisfy

$$J_t-\rho D_t J_D+\frac{1}{2}\sigma^2 J_{FF}+a\sigma^2 X_t^2=0,(F_t+s/2)+\lambda(X_0-X_t)+D_t-J_X+\kappa J_D=0. \tag{A.17}$$

In addition, we have the boundary condition at the terminal point:

$$J(X_T,D_T,F_T,T)=(F_T+s/2)X_T+\lambda(X_0-X_T)X_T+D_T X_T+X_T^2/(2q). \tag{A.18}$$

The inequalities (A.15)–(A.18) are the variational inequalities (VIs), which are the necessary conditions for any solutions to the problem (A.14).

A.5. Candidate value function

Using the intuition from the discrete-time case, we can derive the candidate value function that will satisfy the variational inequalities (A.15)–(A.18). We will be searching for the solution in a class of functions quadratic in X_t and D_t. Note that it is always optimal to trade at time 0. Moreover, the nature of the problem implies that there is no NT region. In fact, if we assume that there exists a strategy with no trading at period (t_1,t_2), then it will be always suboptimal relative to a similar strategy except for the trade at time t_1 being reduced by a sufficiently small amount ϵ that is instead continuously executed over the period (t_1,t_2). Thus, the candidate value function has to satisfy (A.17) in the CT region and (A.15) in any other region. Since there is no NT region, $(F_t+s/2)+\lambda(X_0-X_t)+D_t-J_X+\kappa J_D=0$ holds for any point (X_t,D_t,F_t,t). This implies a particular form for the quadratic candidate value function:

$$J(X_t,D_t,F_t,t)=(F_t+s/2)X_t+\lambda X_0X_t+[\kappa f(t)-\lambda)]X_t^2/2+f(t)X_tD_t+[f(t)-1]D_t^2/(2\kappa), \tag{A.19}$$

where $f(t)$ is a function that depends only on t. Substituting (A.19) into $J_t-\rho D_tJ_D+\frac{1}{2}\sigma^2J_{FF}+a\sigma^2X_t^2\geq 0$, we have

$$(\kappa f'+a\sigma^2)X_t^2/2+(f'-\rho f)X_tD_t+(f'+2\rho-2f)D_t^2/(2\kappa)\geq 0, \tag{A.20}$$

which holds with an equality for any point of the CT region.

Minimizing with respect to X_t, we show that the CT region is specified by

$$X_t=-\frac{f'-\rho f}{\kappa f'+a\sigma^2}D_t. \tag{A.21}$$

For (X_t,D_t) in the CT region, (A.20) holds with the equality. The function $f(t)$ can be found from the Riccati equation:

$$f'(t)(2\rho\kappa+a\sigma^2)-\kappa\rho^2f^2(t)-2a\sigma^2\rho f(t)+2a\sigma^2\rho=0. \tag{A.22}$$

This equation guarantees that $J_t-\rho D_tJ_D+\frac{1}{2}\sigma^2J_{FF}+a\sigma^2X_t^2$ is equal to zero for each point in the CT region and greater than zero for any other points. Taking into account the terminal condition $f(T)=1$, we can solve for $f(t)$. As a result, if the trader is risk neutral and $a=0$, then

$$f(t)=\frac{2}{\rho(T-t)+2}.$$

Substituting the expression for $f(t)$ into (A.19), we get the candidate value function of Proposition 3. If the trader is risk averse and $a\neq 0$, then

$$f(t)=\frac{1}{\kappa\rho}(v-a\sigma^2)-\left[\frac{\kappa\rho}{2v}+\left(\frac{\kappa\rho}{v-a\sigma^2-\kappa}-\frac{\kappa\rho}{2v}\right)e^{(2\rho v/(2\rho\kappa+a\sigma^2))(T-t)}\right]^{-1},$$

where v is the constant defined in Proposition 4. From (A.19), this results in the candidate value function specified in Proposition 4.

A.6. Verification principle

Now we verify that the candidate value function $J(X_0,D_0,F_0,0)$ obtained above is greater or equal to the value achieved by any other trading policy. Let $X_{[0,T]}$ be an arbitrary feasible policy from Θ_C and let $V(X_t,D_t,F_t,t)$ be the corresponding value function. We have

$$X(t)=X(0)-\int_0^t\mu_t\,dt-\sum_{s\in\hat{T},s<t}x_s,$$

where $\mu_t\geq 0$ and $x_t\geq 0$ for $t\in\hat{T}$. For any τ and X_0, we consider a hybrid policy that follows the policy X_t on the interval $[0,\tau]$ and the candidate optimal policy on the interval $[\tau,T]$. The value function for this modified policy is

$$V_\tau(X_0,D_0,F_0,0)=\mathrm{E}_0\left[\int_0^\tau[(F_t+s/2)+\lambda(X_0-X_t)+D_t]\mu_t\,dt\right.$$

$$+\sum_{t_i<\tau,t_i\in\hat{T}}[(F_{t_i}+s/2)x_{t_i}+\lambda(X_0-X_{t_i})x_{t_i}$$
$$+D_{t_i}x_{t_i}+x_{t_i}^2/(2q)]+J(X_\tau,D_\tau,F_\tau,\tau)\Big]. \tag{A.23}$$

For any function, e.g., $J(X_t,D_t,F_t,t)$ and any (X_t,D_t,F_t,t), we have

$$J(X_t,D_t,F_t,t)=J(X_0,D_0,F_0,0)+\int_0^t J_s\,ds+\int_0^t J_X\,dX+\int_0^t J_D\,dD$$
$$+\int_0^t J_F\,dF+\int_0^t\frac{1}{2}J_{FF}(dF)^2+a\sigma^2\int_0^t X_s^2\,ds+\sum_{t_i<t,t_i\in\hat{T}}\Delta J. \tag{A.24}$$

Using $dD_t=-\rho D_t\,dt-\kappa\,dX_t$ and substituting (A.24) for $J(X_\tau,D_\tau,F_\tau,\tau)$ into (A.23), we have

$$V_\tau(X_0,D_0,F_0,0)=J(X_0,D_0,F_0,0)+\mathrm{E}_0\int_0^\tau\left[F_t+\frac{s}{2}+\lambda(X_0-X_t)+D_t-J_X+\kappa J_D\right]\mu_t\,dt$$
$$+\mathrm{E}_0\int_0^\tau\left(J_t-\rho D_tJ_D+\frac{1}{2}\sigma^2J_{FF}+a\sigma^2X_t^2\right)dt$$
$$+\mathrm{E}_0\sum_{t_i<t,t_i\in\hat{T}}\left[\Delta J+(F_t+\frac{s}{2}+\lambda(X_0-X_t)+D_t+x_{t_i}/(2q))x_{t_i}\right]$$
$$=J(X_0,D_0,F_0,0)+I_1+I_2+I_3. \tag{A.25}$$

We can show that for any arbitrary strategy X_t and for any moment τ,

$$V_\tau(X_0,D_0,F_0,0)\geq J(X_0,D_0,F_0,0). \tag{A.26}$$

It is clear that the VIs (A.15)–(A.17) imply the non-negativity of I_1 and I_2 in (A.25). Moreover, if we write $\Delta J(X_{t_i},D_{t_i},F_{t_i},t_i)$ as $J(X_{t_i}-x_{t_i},D_{t_i}+\kappa x_{t_i},F_{t_i}+\sigma Z_{t_i},t_i)-J(X_{t_i},D_{t_i},F_{t_i},t_i)$, we can show that $I_3\geq 0$. This completes the proof of (A.26).

Applying it for $\tau=0$, we see that $J(X_0,D_0,F_0,0)\leq V(X_0,D_0,F_0,0)$. Moreover, there is a strict equality for our candidate optimal strategy. This completes the proof of Proposition 3.

A.7. Properties of the optimal execution policy

We now analyze the properties of optimal execution strategies. First, let us consider the risk-neutral trader with $a=0$. Substituting the expression for $f(t)$ into (A.21), we find that the CT region is given by

$$X_t=\frac{\rho(T-t)+1}{\kappa}D_t.$$

This implies that after the initial trade $x_0=X_0/(\rho T+2)$, which pushes the system from its initial state X_0 and $D_0=0$ into the CT region, the trader trades continuously at the rate $\mu_t=\rho X_0/(\rho T+2)$ staying in the CT region, and executes the rest $x_T=X_0/(\rho T+2)$ at the end of the trading horizon. In fact, this is the same solution as we had for the continuous-time limit of problem (20).

If the trader is risk averse, then the CT region is given by

$$X_t=g(t)D_t\quad\text{where } g(t)=-\frac{f'(t)-\rho f(t)}{f'(t)\kappa+a\sigma^2}.$$

This implies that after discrete trade $x_0 = X_0(\kappa f'(0) + a\sigma^2)/(\rho\kappa f(0) + a\sigma^2)$ at the beginning, which pushes the system from its initial state into the CT region, the trader will trade continuously at the rate,

$$\mu_t = \kappa x_0 \frac{\rho g(t) - g'(t)}{1 + \kappa g(t)} e^{-\int_0^t ((\kappa g\prime(s) + \rho)/(1 + \kappa g(s)))\, ds}.$$

This can be shown taking into account the dynamics of D_t in (A.2) and the specification of the CT region. At the end, the trader finishes off the order.

References

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

Alfonsi, A., Schied, A., Schulz, A., 2009. Optimal execution strategies in limit order books with general shape functions. Quantitative Finance 10, 143–157.

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

Almgren, R., Chriss, N., 1999. Value under liquidation. Risk 12 (12).

Almgren, R., Chriss, N., 2000. Optimal execution of portfolio transactions. University of Toronto Working Paper.

Back, K., Baruch, S., 2007. Working orders in limit order markets and floor exchanges. Journal of Finance 62, 1589–1621.

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

Biais, B., Hillion, P., Spatt, C., 1995. An empirical analysis of the limit order book and order flow in Paris Bourse. Journal of Finance 50, 1655–1689.

Bloomfield, R., O'Hara, M., Saar, G., 2004. The "Make or Take" decision in an electronic market: evidence on the evolution of liquidity. Journal of Financial Economics 75, 165–199.

Cao, C., Hansch, O., Wang, X., 2003. The informational content of an open limit order book. In: EFA 2004 Maastricht Meetings Paper No. 4311.

Chan, L., Lakonishok, J., 1995. The behavior of stock prices around institutional trades. Journal of Finance 50, 1147–1174.

Chan, L., Lakonishok, J., 1997. Institutional equity trading costs: NYSE versus Nasdaq. Journal of Finance 52, 713–735.

Coppejans, M., Domowitz, I., Madhavan, A., 2004. Resilience in an automated auction. Working Paper.

Degryse, H., De Jong, F., Van Ravenswaaij, M., Wuyts, G., 2005. Aggressive orders and the resilience of a limit order market. Review of Finance 9, 201–242.

Dubil, R., 2002. Optimal liquidation of large security holdings in thin markets. In: Shachmurove Yochanan (Ed.), Research in International Entrepreneurial Finance and Business Ventures at the Turn of the Third Millennium. Academy of Entrepreneurial Finance.

Esser, A., Monch, B., 2005. The navigation of an iceberg: the optimal use of hidden orders. Working Paper.

Foucault, T., 1999. Order flow composition and trading costs in a dynamic limit order market. Journal of Financial Markets 2, 99–134.

Foucault, T., Kadan, O., Kandel, E., 2005. Limit order book as a market for liquidity. Review of Financial Studies 18, 1171–1217.

Glosten, L., 1994. Is the electronic open limit order book inevitable? Journal of Finance 49, 1127–1161.

Goettler, R., Parlour, C., Rajan, U., 2005. Equilibrium in a dynamic limit order market. Journal of Finance 60, 2149–2192.

Goettler, R., Parlour, C., Rajan, U., 2009. Informed traders and limit order markets. Journal of Financial Markets 93, 67–87.

Greenwood, R., 2005. Short- and long-term demand curves for stock: theory and evidence on the dynamics of arbitrage. Journal of Financial Economics 75, 607–649.

Grinold, R., Kahn, R., 2000. Active Portfolio Management. McGraw-Hill, New York.

Grossman, S., Miller, M., 1998. Liquidity and market structure. Journal of Finance 43, 617–637.

Hamao, Y., Hasbrouck, J., 1995. Securities trading in the absence of dealers: trades and quotes on the Tokyo Stock Exchange. Review of Financial Studies 8, 849–878.
Harris, L., 1998. Optimal dynamic order submission strategies in some stylized trading problems. Financial Markets, Institutions and Instruments 7, 1–76.
Harris, L., Panchapagesan, V., 2005. The information-content of the limit order book: evidence from NYSE specialist trading decisions. Journal of Financial Markets 8, 25–67.
He, H., Mamaysky, H., 2005. Dynamic trading policies with price impact. Journal of Economic Dynamics and Control 29, 891–930.
Holthausen, R., Leftwich, R., Mayers, D., 1987. The effect of large block transactions on security prices: a cross-sectional analysis. Journal of Financial Economics 19, 237–267.
Holthausen, R., Leftwitch, R., Mayers, D., 1990. Large block transactions, the speed of response of temporary and permanent stock-price effect. Review of Financial Studies 9, 71–95.
Hopman, C., 2007. Do supply and demand drive stock prices? Quantitative Finance 7, 37–53.
Huberman, G., Stanzl, W., 2004. Arbitrage-free price update and price-impact functions. Econometrica 72, 1247–1275.
Huberman, G., Stanzl, W., 2005. Optimal liquidity trading. Review of Finance 9, 165–200.
Kaul, A., Mehrotra, V., Morck, R., 2000. Demand curves for stocks do slope down: new evidence from an index weights adjustment. Journal of Finance 55, 893–912.
Keim, D., Madhavan, A., 1995. Anatomy of the trading process: empirical evidence on the behavior of institutional traders. Journal of Financial Economics 37, 371–398.
Keim, D., Madhavan, A., 1997. Transactions costs and investment style: an inter-exchange analysis of institutional equity trades. Journal of Financial Economics 46, 265–292.
Kempf, A., Mayston, D., Yadav, P., 2009. Resilience in limit order book markets: a dynamic view of liquidity. Working Paper.
Kyle, A., 1985. Continuous auctions and insider trading. Econometrica 53 (6), 1315–1335.
Large, J., 2007. Measuring the resiliency of an electronic limit order book. Journal of Financial Markets 10, 1–25.
Monch, B., 2004. Optimal liquidation strategies. Goethe University Working Paper.
Obizhaeva, A., 2008. Portfolio transitions and stock price dynamics. Working Paper, University of Maryland.
Parlour, C., 1998. Price dynamics in limit order markets. Review of Financial Studies 11, 789–816.
Parlour, C., Seppi, D., 2003. Liquidity-based competition for order flow. Review of Financial Studies 16, 301–303.
Predoui, S., Shaikhet, G., Shreve, S., 2010. Optimal execution strategy in a general one-sided limit-order-book. Working Paper.
Ranaldo, A., 2004. Order aggressiveness in limit order book markets. Journal of Financial Markets 7, 53–74.
Rock, K., 1996. The specialist order book and price anomalies. Working Paper.
Roşu, I., 2008. Liquidity and information in order driven markets. NBER Working Paper.
Roşu, I., 2009. A dynamic model of the limit order book. Review of Financial Studies 22, 4601–4641.
Sandas, P., 2001. Adverse selection and competitive market making: evidence from a pure limit order book. Review of Financial Studies 14, 705–734.
Scholes, M., 1972. The market for securities: substitution versus price pressure and the effects of information on share price. Journal of Business 45, 179–211.
Seppi, D., 1997. Liquidity provision with limit orders and a strategic specialist. Review of Financial Studies 10, 103–150.
Shleifer, A., 1986. Do demand curves for stocks slope down? Journal of Finance 41, 579–590.
Subramanian, A., Jarrow, R., 2001. The liquidity discount. Mathematical Finance 11, 447–474.
Vayanos, D., 1999. Strategic trading and welfare in a dynamic market. Review of Economic Studies 66, 219–254.
Vayanos, D., 2001. Strategic trading in a dynamic noisy market. Journal of Finance 56, 131–171.