Review of Finance (2005) 9: 165–200
DOI: 10.1007/s10679-005-7591-5

Optimal Liquidity Trading *

GUR HUBERMAN[1] and WERNER STANZL[2]
[1]*Columbia Business School;* [2]*Ziff Brothers Investments, L.L.C.*

Abstract. A liquidity trader wishes to trade a fixed number of shares within a certain time horizon and to minimize the mean and variance of the costs of trading. Explicit formulas for the optimal trading strategies show that risk-averse liquidity traders reduce their order sizes over time and execute a higher fraction of their total trading volume in early periods when price volatility or liquidity increases. In the presence of transaction fees, traders want to trade less often when either price volatility or liquidity goes up or when the speed of price reversion declines. In the multi-asset case, price effects across assets have a substantial impact on trading behavior.

1. Introduction

What is the optimal trading sequence of a person who wishes to buy (or sell) a certain portfolio within a certain time, and knows how trades affect prices? Institutional investors usually transact portfolios of considerable size and thus incur permanent and temporary price impacts. The temporary impact represents the transitory cost of demanding liquidity and only affects an individual trade. On the contrary, the permanent component of the price impact not only influences the price of the first trade but also the prices of all subsequent trades of an agent. Modelling this price dynamics explicitly enables us to derive cost-efficient execution strategies for multi-trade orders. We focus exclusively on linear price-impact functions because Huberman and Stanzl (2004) show that in the absence of price manipulation, price-impact functions are linear.

To minimize the price impact, an agent would choose to trade patiently and split his order into many small pieces (e.g., see Bertsimas and Lo (1998)). Such a strategy ignores the opportunity costs that arise from unfavorable price movements during the execution of an order. The longer the trade duration the higher the uncertainty of the realized prices. Hence, to balance price-impact costs against opportunity costs, we assume that a liquidity trader not only cares about the expected value but also about the variance of his execution costs. Risk aversion in this

* We are grateful to Prajit Dutta and Larry Glosten for numerous conversations and comments and to Marc Lipson for help with the Plexus data. Comments and suggestions of the referee and the editor, Josef Zechner, helped us improve the paper. We also thank the participants of the Chicago Board of Trade 13th Annual European Futures Research Symposium 2000 and the participants of the EFA Annual Meetings 2001.

context means that an investor is willing to trade lower opportunity costs for higher price-impact costs.

ITG Inc., a trade execution firm which services institutional investment funds, uses a similar classification for execution costs and groups managers according to their sensitivity to opportunity costs. On page 4 of the ITG ACE$^{\text{TM}}$ document (2004) it says:

> When executing an agency order the balance between price impact and opportunity cost is chosen on the basis of the motivation for the order, which ultimately comes from the investment manager. Passive managers are primarily concerned with price impact. Growth or momentum managers are more worried about opportunity costs.

In our parlance, growth or momentum managers are more risk-averse than passive ones.

Assuming that the trader wishes to minimize the mean and variance of the total execution costs, we prove that a time-consistent solution exists and is unique when price manipulation is ruled out. The most important features of the optimal execution strategy are that trade sizes decline over time if the price-impact function is time-independent, and that orders are independent of past random shocks such as the arrival of new information. The comparative statics show that higher aversion to risk, higher price or trading volume volatility, lower speed of price reversion, and higher liquidity of price changes to trade size all lead to more aggressive initial trading.

These theoretical results give rise to two different sets of empirical tests. First, for a given level of risk aversion, or equivalently, type of portfolio manager, how do asset-specific characteristics such as liquidity, price volatility, or pace of price reversion affect the optimal trading strategy? Second, for a fixed asset or portfolio, what is the risk aversion implied by a portfolio manager's trading sequence? Since our price model can be estimated econometrically, the formulas we obtain for the optimal execution can be used to tackle both questions.

In general, the dynamics of institutional trading has not been studied empirically. More specifically, the hypothesis that when institutions allocate their trades over a few days, they trade more aggressively initially has not been examined empirically, its intuitive appeal notwithstanding. In Section 2 we summarize evidence assembled for Huberman, Jones and Lipson (2004). They document that when institutions trade the same stock in the same direction over a few days, on average they trade 2% more during the first half of the period than during its second half. This finding is consistent with the special case of our model in which the price-impact function is constant over time.

Some institutions, like mutual funds, are prohibited from short selling. We prove that an optimal execution strategy also exists for a buyer who faces short-sale constraints. An incentive to short arises when early trading periods are relatively illiquid compared to later ones. In this case, by shorting the trader pushes the prices favorably down for future purchases. However, if short-sale constraints are in place, the trader refrains from trading in the less liquid periods and redistributes his volume across the other more liquid periods.

Chan and Lakonishok (1995) study the effects of commission costs, market capitalization, and managerial strategy on the price impact and execution costs of institutional trades. Consistent with our results, they report that execution costs are in general higher for less patient traders and low market-capitalization stocks. In addition, they report that the manager's type is the most important cost factor. Thus, as investors' risk aversion varies with type, price volatility would be expected to be a significant determinant of the execution costs. This relation has not been empirically tested yet, but our model predicts that price volatility increases costs substantially for traders with high opportunity costs.

This paper explicitly analyzes how mean-reversion of prices following trades affects the optimal trading strategy, and shows that trading costs are negatively related to the speed of price reversion. Thereby this speed can be interpreted as another dimension of liquidity. Foucault, Kandel and Kadan (2004) provide a theoretical model in which the speed at which spreads revert after a sequence of trades is endogenous and is negatively correlated with trading costs. Empirically, Biais, Hillion and Spatt (1995), Degryse et al. (2003) and Coppejans et al. (2003) have documented that spreads and prices tend to revert following trades.

In the presence of fixed per-trade transaction fees, the trade duration becomes endogenous. A higher price volatility or higher liquidity with respect to trade size decreases the optimal number of trades. When the price volatility is high, the trader lowers his risk exposure by shortening his trading horizon. When the price impact is large, the investor reduces the overall impact by submitting many small orders. Further, the trader favors a longer trade duration if prices revert more quickly to their original levels after large trades. Keim and Madhavan (1995) provide evidence that value managers exhibit the longest trade duration, followed by index and technical managers. Our model explains this pattern.

In practice, multiple assets are traded simultaneously. In this case, the traded volume of one asset presumably affects not only its own price but also the prices of other assets. To account for cross-price impacts, we extend the analysis to allow for trading a portfolio of securities and derive dynamic trading rules that describe how to optimally rebalance a portfolio. Volume is deferred to later periods when either cross-price impacts are high or asset prices are negatively correlated.

The solution to the liquidity trader's problem derived here can also be used to find the optimal "program trade" of a potential insider. A new SEC rule (FD Rule 25/1 and 25/2) on insider trading requires potential insiders to announce their trades before they actually trade and that they commit to their announced trades. Hence, potential insiders have to specify when they want to trade and how much they want to trade. But this is exactly the liquidity trader's problem.

The remainder of this paper is structured as follows. Section 2 looks at the daily trading pattern of money managers in the Plexus data set. The liquidity trader's minimization problem is introduced in Section 3. Section 4 establishes the existence and uniqueness of the sequence of optimal trades and discusses its properties. Section 5 examines portfolio trading. Section 6 studies the optimal trading frequency in the presence of fixed transaction costs. Section 7 explores the effects of autocorrelated noise trades on the optimal trading strategy. Three papers that are closely related to our work, namely, Bertsimas and Lo (1998), Almgren and Chriss (2000) and Vayanos (2001), are discussed in Section 8. Section 9 concludes. All proofs are in the Appendix.

2. Daily Trading

The analysis in this paper is normative, and considers the optimal strategy of a trader, given his objective and market circumstances. One of the analysis' key insights is that a trader who allocates his trades over time should trade larger quantities earlier on. Surprisingly, not much is known about the extent to which traders actually follow this intuitively appealing prescription. This section reports a summary statistic on the intertemporal allocation of trades by large US money managers who trade the same stock in the same direction over a few days. Taken from Huberman, Jones and Lipson (2004, in preparation), this statistic illustrates the extent to which the size of these trades declines over the execution period.

The Plexus Group provided the data underlying the statistics reported in this section. Plexus is a consulting firm that works with institutional investors to monitor and reduce their equity transaction costs. Its clients manage over \$1.5 trillion and the firm has access to transactions covering about 25% of US trading volume. The Plexus data have been studied in earlier work. (e.g., Keim and Madahvan, 1995, 1997, Jones and Lipson, 1999, 2001, Conrad et al., 2001.)

The statistics cover the behavior of 120 institutional investors over the period 1997–2001. The basic construct is a run, which is defined by its direction (buy or sell), length k, first day t, the manager involved i, and the stock involved j.

A k-day buy (sell) run on stock j by manager i that begins on day t is a series of purchases (sales) of stock j by manager i such that there is no purchase (sale) on day $t-1$, there are purchases (sales) on each of the days $t, t+1, \ldots, t+k-1$, and no purchase on day $t+k$. Fix the length of the run k for $k \geq 2$. Let the amount manager i purchased (sold) of stock j on day τ be $B(i, j, \tau)$ $(S(i, j, \tau))$ and

$$b(i, j, \tau) = \frac{B(i, j, \tau)}{\sum_{\tau=t}^{t+k-1} B(i, j, \tau)},$$

where $b(i, j, \tau)$ is the fraction of the run executed on day τ. Define similarly $s(i, j, \tau)$.

Consider the regression equation

$$b(i, j, \tau) = \alpha_b + \beta_b(\tau - t) + \epsilon(i, j, t).$$

The slope coefficient β_b depends on the run, and therefore is a function of i, j, and t. It is an estimate of the difference of the fraction of purchases executed on two

Table I. Institutional trading over time

Run length (days)	Average difference between the fraction of run traded in the first and second half of run
2	2.25%
3	1.88%
4	0.74%
5	2.38%

consecutive days. Similarly, one can estimate β_s. The average of the β_b's and β_s's and their dependence on k is the focus of this section.

Of the runs that take at least two days to complete, 95% are complete within five days. Table I summarizes the average difference between the quantity traded in the first half of the run and the second half of the run (across buys and sells). In general, if the run is of length k, the slope coefficient of the regression translates to a volume difference between the first and second half of the run of $k^2/4$ when k is even and $(k-1)(k+1)/4$ if k is odd. The table suggests that on average institutional investors trade around 2% more in the first half of the run than in the second half.

3. The Optimization Problem

Consider a market in which a *single asset* is traded over N periods. At each period, traders submit their orders simultaneously, and the price change from one period to the next depends on the aggregate excess demand. (Presumably, there is a market maker outside the model who absorbs this excess demand.) Orders are placed before the price change is known. Only market orders are considered.

From the perspective of an individual trader, the total trading volume at time n is given by $q_n + \eta_n$, where q_n denotes the trader's order size and η_n is a random variable representing the unknown volume of the others. (Negative quantities are sales.) We assume that $\{\eta_n\}_{n=1}^N$ is an $i.i.d.$ stochastic process with zero mean and finite variance σ_η^2, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The initial price of the asset at time n, $\hat{p}_n$, which is observed by each trader before choosing his quantity q_n, is the last price update computed after the trades in the previous period $n-1$. Given the initial price, an individual trader faces the transaction price $p_n = \hat{p}_n + \lambda_n(q_n + \eta_n)$, where the real number λ_n measures the liquidity with respect to trading volume. Hence, a trader expects to pay $(\hat{p}_n + \lambda_n q_n)q_n$ if he wants to buy the quantity q_n. After all trades have been executed at time n, the new price update for the next period is calculated according to $\hat{p}_{n+1} = \alpha\hat{p}_n + (1-\alpha)p_n + \varepsilon_{n+1}$, where $0 \leq \alpha \leq 1$ and ε_{n+1} incorporates news into the price.

The *updating weight* α determines the size of the price updates. It can also be interpreted as the speed of price reversion. The lower the α, the stronger is the permanent impact. If $\alpha = 0$, expected price updates and transaction prices coincide, and the price dynamics reduce to

$$p_n = p_{n-1} + \lambda_n(q_n + \eta_n) + \varepsilon_n. \tag{1}$$

Trade size has only a temporary price impact if $\alpha = 1$. The stochastic process $\{\varepsilon_n\}_{n=1}^N$, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, is *i.i.d.* with zero mean and variance σ_ε^2, and it is independent of $\{\eta_n\}_{n=1}^N$. The zero-mean assumptions are not made for convenience; if one of the two stochastic processes exhibited a nonzero mean, then price manipulation as discussed in Huberman and Stanzl (2004) would arise.

To define the information set of an individual trader we introduce the vector $H_n \equiv (\{\hat{p}_j\}_{j=1}^n, \{p_j\}_{j=1}^{n-1}, \{q_j\}_{j=1}^n, \{\eta_j\}_{j=1}^{n-1}, \{\varepsilon_j\}_{j=1}^n)$ containing the variables known to the trader before he submits his order in period n, and the sigma-algebra $\sigma(H_n)$ that it generates. Then, the set $M(H_n)$ of all $\sigma(H_n)$-measurable functions comprises all information available to the trader before his trade at time n. Unlike η_n, the trader does know the news ε_n. Furthermore, the trader can only choose a trading strategy q_n that is an element of $M(H_n)$. This setup should best capture real trading activity where the latest public news is known before submitting an order, while others' trades are not.

To make later references easier, the price dynamics are summarized by

$$\begin{aligned} \hat{p}_n &= \alpha\hat{p}_{n-1} + (1-\alpha)p_{n-1} + \varepsilon_n \\ p_n &= \hat{p}_n + \lambda_n(q_n + \eta_n), \end{aligned} \tag{2}$$

for $n \geq 1$ and initial price $p_0 = \hat{p}_0 > 0$, with the special case (1) when $\alpha = 0$.

The liquidity trader's optimization problem can be formulated as

$$\begin{gathered} L(Q, N) \equiv \inf_{\{q_n \in \mathbf{M}(H_n)\}_{n=1}^N} \mathbb{E}\left[\sum_{n=1}^N p_n q_n\right] + \frac{R}{2}\mathrm{Var}\left[\sum_{n=1}^N p_n q_n\right] \\ \text{subject to } \sum_{n=1}^N q_n = Q \text{ and (2)}, \end{gathered} \tag{3}$$

where $Q > 0$ ($Q < 0$) denotes the number of shares he wants to buy (sell) and $R \geq 0$ is the risk-aversion coefficient. Expectation and variance are evaluated at time zero before any of the random price elements are realized. The liquidity trader, aware of the price impact of his trades summarized in (2), minimizes the mean and variance of the total execution costs that must be incurred to enlarge (reduce) his portfolio by Q shares. Note that the first line in (3) reads

$$-\sup_{\{q_n \in \mathbf{M}(H_n)\}_{n=1}^N} \mathbb{E}\left[\sum_{n=1}^N p_n(-q_n)\right] - \frac{R}{2}\mathrm{Var}\left[\sum_{n=1}^N p_n(-q_n)\right] \tag{4}$$

for the seller, i.e., he sells $\{-q_n\}_{n=1}^N$ shares in order to maximize revenues minus its variance. Thus, if $\{q_n\}_{n=1}^N$ denote the optimal quantities traded by a buyer of Q shares, the optimal quantities traded by a seller of $-Q$ shares are $\{-q_n\}_{n=1}^N$. We will refer to both $L(Q, N)$ and $\mathbb{E}[\sum_{n=1}^N p_n q_n] + \frac{R}{2}\mathrm{Var}[\sum_{n=1}^N p_n q_n]$ as *cost function*, with (Q, N) and $(q_1, \dots, q_N)$ being the arguments, respectively.

REMARK 1. An efficient mean-variance frontier for the execution costs can be constructed from (3). To see this, define the efficient frontier by

$$\mathfrak{F}(V) \equiv \left\{ \inf_{\{q_n \in \mathbf{M}(H_n)\}_{n=1}^N} \mathbb{E}\left[\sum_{n=1}^N p_n q_n\right] : \mathrm{Var}\left[\sum_{n=1}^N p_n q_n\right] \leq V \right\}. \tag{5}$$

By solving (3) and varying the risk-aversion coefficient R one can trace out $\mathfrak{F}$ in the (mean,variance)-space. If the solutions to (3) are unique, then $\mathfrak{F}$ is strictly decreasing in the interval $(0, \bar{V}]$, where $\bar{V}$ is the variance produced by the expected-cost-minimizing strategy ($R = 0$). The trader's optimal execution strategy is implied by the tangency point between his utility function and the frontier $\mathfrak{F}$.

REMARK 2. The objective function in (3) can also be interpreted as minimizing the mean and variance of the value-weighted average price, $\sum_{n=1}^N p_n q_n / Q$, paid for execution (divide $L(Q, N)$ by Q and reset the risk-aversion parameter to R/Q).

REMARK 3. Problem (3) can also be employed to study insider trading as in the Kyle (1985) framework provided the total trading volume is fixed or announced before trading (e.g., the FD Rule 25/1 and 25/2 on insider trading by the SEC stipulates that potential insiders have to announce their trades before they actually trade and that they are obliged to commit to their announced trades). An insider knows that eventually the security will trade at a price v, and tries to profit from the discrepancy between that eventual price and the prices at which he can trade over the next N periods. Assume no updates on v are made during trading. Buying q_n at price p_n hence yields a profit of $\mathbb{E}[(v - p_n)q_n]$ for a risk-neutral insider. The objective is therefore

$$\begin{aligned} \pi(Q, N) &\equiv \sup_{\{q_n \in \mathbf{M}(H_n)\}_{n=1}^N} \mathbb{E}\left[\sum_{n=1}^N (v - p_n) q_n\right] \\ &\text{subject to } \sum_{n=1}^N q_n = Q \text{ and (2).} \end{aligned} \tag{6}$$

As a consequence, $\pi(Q, N) = \sup \mathbb{E}[\sum_{n=1}^N (v - p_n) q_n]$ subject to $\sum_{n=1}^N q_n = Q$ is equivalent to $L(Q, N) = \inf \mathbb{E}[\sum_{n=1}^N p_n q_n]$ subject to $\sum_{n=1}^N q_n = Q$, and $\pi(Q, N) = vQ - L(Q, N)$. But this is nothing but the liquidity trader's minimization problem in (3) if $R = 0$. As a result, risk-neutral insiders who fix the number of shares they trade are liquidity traders.

4. Cost-efficient Trade Execution

This section formulates a recursive version of the problem in (3), provides sufficient conditions for the existence and uniqueness of the solution, and presents explicit formulas for the optimal trading policy. The optimal trading path is independent of the resolution of uncertainty, and the traded amounts decline with time. The focus is only on the buyer's problem, because the seller's problem is similar. Upfront we state the most general result (the proof is in the Appendix). The following subsections discuss special cases of Theorem 1. In particular, Proposition 2 (Section 4.3) assumes that the permanent price impact is constant through time, whereas Proposition 3 (Section 4.3) allows λ_n to change over time. Both propositions restrict α to be zero (no temporary price impact). Section 4.4 permits $\alpha \neq 0$ but considers only constant λ_n's.

THEOREM 1. Consider the backward difference equation

$$\mu_n = \alpha\lambda_{n-1} + \lambda_n\left(1 + \frac{R}{2}\lambda_n\sigma_\eta^2\right) - \frac{\lambda_n^2(1 + \alpha + R\alpha\lambda_n\sigma_\eta^2)^2}{2[2\mu_{n+1} + R(\alpha^2\lambda_n^2\sigma_\eta^2 + \sigma_\varepsilon^2)]}, \tag{7}$$

for $1 \leq n \leq N-1$, with $\mu_N = \alpha\lambda_{N-1} + \lambda_N(1 + \frac{R}{2}\lambda_N\sigma_\eta^2)$. If $2\mu_n + R(\alpha^2\lambda_{n-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2) > 0$ for $2 \leq n \leq N$, then the liquidity trader's problem (3) exhibits a unique, time-consistent solution that can be obtained by solving the dynamic program in (14). It is given by

$$q_n = \theta_n Q_n, \tag{8}$$

for $1 \leq n \leq N$, where

$$\theta_n \equiv 1 - \frac{\lambda_n(1 + \alpha + R\alpha\lambda_n\sigma_\eta^2)}{2\mu_{n+1} + R(\alpha^2\lambda_n^2\sigma_\eta^2 + \sigma_\varepsilon^2)} \tag{9}$$

$1 \leq n \leq N-1$, and $\theta_N = 1$. The minimal costs equal

$$L_n(\tilde{p}_{n-1}, Q_{n-1}, Q_n) = [\tilde{p}_{n-1} - \alpha\lambda_{n-1}(Q_{n-1} + \eta_{n-1})]Q_n + \mu_n Q_n^2, \tag{10}$$

$1 \leq n \leq N$, where $Q_0 = \lambda_0 = \eta_0 \equiv 0$ and $\tilde{p}_{n-1} \equiv p_{n-1} + \varepsilon_n$.

Theorem 1 includes a sufficient condition for the existence of a solution to problem (3). Proposition 1 (Section 4.2) offers two more sufficient conditions that have a more appealing economic interpretation. Namely, if price-manipulation opportunities as defined in Huberman and Stanzl (2004) are absent, then a time-consistent solution to (3) exists.

To provide the basic intuition, we begin with a stripped-down version of (3), with $\alpha = 0$ and two periods.

4.1. THE TWO-PERIOD PROBLEM

With the price process $p_n = p_{n-1} + \lambda_n(q_n + \eta_n) + \varepsilon_n$, $n = 1, 2$, the costs of trading q_1 in period 1 and $q_2 = Q - q_1$ in period 2 amount to

$$\begin{aligned} C_2 \;=\; & [p_0 + \lambda_1(q_1 + \eta_1) + \varepsilon_1]q_1 \\ & +[p_0 + \lambda_1(q_1 + \eta_1) + \lambda_2(Q - q_1 + \eta_2) + \varepsilon_1 + \varepsilon_2](Q - q_1). \end{aligned} \tag{11}$$

The pair (q_1, q_2) that minimizes $\mathbb{E}[C_2] + \frac{R}{2}\mathrm{Var}[C_2]$ is

$$q_1 = \frac{2\lambda_2 - \lambda_1 + R(\lambda_2^2\sigma_\eta^2 + \sigma_\varepsilon^2)}{2\lambda_2 + R(\lambda_2^2\sigma_\eta^2 + \sigma_\varepsilon^2)}Q \quad \text{and} \quad q_2 = \frac{\lambda_1}{2\lambda_2 + R(\lambda_2^2\sigma_\eta^2 + \sigma_\varepsilon^2)}Q. \tag{12}$$

The stochastic term ε_1 is a sunk cost by the time the first (and only) decision is made, and therefore does not affect the optimal trades in (12).

To understand how optimal trading is affected by risk aversion, set $\lambda_1 = \lambda_2 = \lambda$ for the moment. Then, $q_1 = q_2 = Q/2$ if $R = 0$, i.e., the risk-neutral trader splits his total quantity evenly across the two periods. These amounts are not optimal for a risk-averse trader, because the marginal cost at these quantities is $-R(\lambda^2\sigma_\eta^2 + \sigma_\varepsilon^2)Q/2 < 0$. In other words, the risk-averse trader is willing to incur higher expected expenses in return for a lower variance. In fact, he always wants to equate the marginal change in the expected value of the execution costs, $\lambda(2q_1 - Q)$, to the marginal change in its variance weighted with the coefficient R, $R(\lambda^2\sigma_\eta^2 + \sigma_\varepsilon^2)(Q - q_1)$. Thus, optimality requires that $q_1 > q_2$ and that a buyer purchases shares in each period. (Note that the trader chooses $q_1 = Q$ if he aims at minimizing the variance only.)

The ratio $q_1/q_2 = 2\lambda_2/\lambda_1 - 1 + R(\lambda_2^2\sigma_\eta^2 + \sigma_\varepsilon^2)/\lambda_1$ enables us to perform comparative statics. The optimal trade size at time 1 increases when λ_1 decreases, or when λ_2, R, σ_η^2, or σ_ε^2 rises. Hence, the trader purchases less in price-sensitive periods and shifts his trading volume in the first period when price volatility or the level of risk aversion goes up.

If λ_1 is sufficiently large ($\lambda_1 > 2\lambda_2 + R(\lambda_2^2\sigma_\eta^2 + \sigma_\varepsilon^2)$), it is optimal to short-sell in the first period, implying that the benefit from reducing the price overcompensates the cost of trading more than Q shares in the second period. A short-sale constraint would prohibit such a trading policy. Indeed, if $q_1 \geq 0$ were imposed, then $q_1 = 0$ and $q_2 = Q$, i.e., the trader buys all shares in the second period.

4.2. EXISTENCE AND UNIQUENESS OF OPTIMAL STRATEGY

To solve problem (3) we define a recursive version and apply dynamic programming arguments to find a time-consistent solution. For simplicity, we consider here only $\alpha = 0$.

The state at time n consists of the price $\tilde{p}_{n-1}$, which is to be paid for zero quantity, and Q_n, the number of shares that remain to be bought. The control variable at time n is q_n, the number of shares purchased in period n. Randomness is represented by the ε_k's ($n + 1 \leq k \leq N$) and the η_k's ($n \leq k \leq N$). The objective is the weighted sum of the expectation and variance of the execution costs, and the law of motion is governed by (1) and the following state equations which describe the dynamics of the remaining number of shares to be traded:

$$Q_1 = Q,\ Q_{n+1} = Q_n - q_n,\ 1 \leq n \leq N, \text{ and } Q_{N+1} = 0. \tag{13}$$

The equations $Q_1 = Q$ and $Q_{N+1} = 0$ represent the restriction that Q shares must be traded within the next N periods.

Since the objective function in (3) is not additive-separable, it is not obvious whether there exists an equivalent dynamic program for (3). The Appendix shows that such a dynamic program indeed exists and is given by

$$\begin{aligned} L_n(\tilde{p}_{n-1}, Q_n) \;=\; & \inf_{q_n \in \mathbf{M}(H_n)} \mathbb{E}_n[p_n q_n + L_{n+1}(\tilde{p}_n, Q_{n+1})] \\ & + \frac{R}{2} \mathrm{Var}_n[p_n q_n + L_{n+1}(\tilde{p}_n, Q_{n+1})] \end{aligned} \tag{14}$$

subject to (1) and (13). $\mathbb{E}_n$ and Var_n denote the conditional expectation and variance in period n. Beginning at the end in period N and applying the recursive equation above together with (1) and (13), the functional equation can be solved backwards as a function of the state variables. The procedure ends when we reach the first period in which we know the whole optimal trading sequence and the total costs.

The question of the existence of a solution to (3) and (14) is closely related to whether price manipulation as studied in Huberman and Stanzl (2004) is possible. In an environment like here, where prices are unknown when trades are submitted, Huberman and Stanzl require expected expenditures, $\mathbb{E}[\sum_{n=1}^{N} p_n q_n]$, to be nonnegative when $\sum_{n=1}^{N} q_n = 0$ to rule out (expected) profits from price manipulation. They show that the price process (1) cannot be manipulated if and only if the symmetric matrix defined by

$$[\Lambda_N]_{m,n} \equiv \begin{cases} 2\lambda_{n+1} & \text{if } n = m \\ \lambda_{m+1} & \text{if } n > m \end{cases},\ 1 \leq m, n \leq N - 1, \tag{15}$$

is positive semidefinite.

Now, the existence of a solution to (3) and (14) is guaranteed by (1) being manipulation-free whenever traders are risk-averse. In the case of risk neutrality, a slightly stronger condition is needed to ensure the existence of a solution. The proposition below documents these facts. (See Theorem 2 in the Appendix for a proof.)

PROPOSITION 1. Suppose one of the following conditions is met:

(i) $R > 0$ (trader is risk-averse) and the price process (1) is manipulation-free, or

(ii) $R = 0$ (trader is risk-neutral) and the matrix Λ_N is positive definite.
Then, the liquidity trader's problem (3) has a unique, time-consistent solution that can be derived by solving the dynamic programming problem in (14).

Obviously, if the price-impact slopes are constant and positive, then the matrix Λ_N is positive definite. Thus, according to Proposition 1, a solution to the liquidity trader's problem exists, regardless of the trader's type. If the λ_n's change over time, then Λ_N has to be evaluated numerically to apply Proposition 1. Huberman and Stanzl (2004) derive a recursive formula for the determinant of Λ_N that can be used for this computational evaluation. Empirical studies imply varying price-impact slopes (for example, see Chordia et al. (2001)).

4.3. OPTIMAL TRADING BEHAVIOR

Having established the equivalence between (3) and the corresponding dynamic program, we turn now to the optimal solution itself. If $\alpha = 0$ and the price-impact function is time-independent, i.e., the λ_n's are constant, then the solution to (3) is summarized as follows:

PROPOSITION 2. If the price-impact sequence is constant and positive, then a solution to (3) exists and is unique. The optimal trading quantities of a risk-averse trader are given by

$$\begin{aligned} q_n &= D\left(A_+ r_+^{N-2-n} - A_- r_-^{N-2-n}\right) Q \quad \text{for } 1 \le n \le N-1, \\ &\text{and} \quad q_N = D\lambda^2 (r_+ - r_-)\, Q, \end{aligned} \tag{16}$$

where

$$r_\pm \equiv 1 + \frac{\sigma}{2\lambda}\left[R\sigma \pm \sqrt{R(4\lambda + R\sigma^2)}\right],$$

$$A_\pm \equiv [\lambda^2 + 3\lambda R\sigma^2 + R^2\sigma^4] r_\pm - \lambda(\lambda + R\sigma^2), \tag{17}$$

$$D^{-1} \equiv \frac{r_+^{N-3} - 1}{1 - r_-} A_+ - \frac{r_-^{N-3} - 1}{1 - r_+} A_- + \left(3\lambda + R\sigma^2\right)\left(\lambda + R\sigma^2\right)(r_+ - r_-) > 0,$$

and $\sigma^2 \equiv \lambda^2\sigma_\eta^2 + \sigma_\varepsilon^2$. All quantities are positive and the sequence of trades is strictly decreasing. If $R = 0$, then $q_n = \frac{Q}{N}$ for $1 \le n \le N$.

Therefore, it is optimal for a buyer to purchase shares in each period. The intuition behind declining trade size is as in the two-period example. Due to the price dynamics (1), the variance of the execution costs at time n depends only on the remaining shares to be traded, $Q_n - q_n$, and is increasing in $Q_n - q_n$. Given the risk-averse utility, the variance that is produced by distributing trades evenly across time is too high. Thus, the risk-averse trader wants to reduce the variance by submitting a larger order in period n than in $n + 1$.

Furthermore, the optimal strategy does not explicitly depend on the random shocks induced by the η_n's and ε_n's. History only enters through Q_n, the remaining shares to be traded. This fact can be interpreted as follows. The total execution cost of trading the remaining shares in periods n and $n + 1$ can be written as the difference between $p_{n+1}Q_n$, the execution costs of buying the remaining shares in period $n+1$, and $(p_{n+1} - p_n)q_n$, the "cost savings" of trading at time n ($p_{n+1} \geq p_n$ holds in expectation). From (1), however, it follows that the conditional expectation and variance of the price differential, $p_{n+1} - p_n$, are not a function of ε_n and η_n, and that $Q_n\varepsilon_n$ is the only term contained in $\mathbb{E}_n[p_{n+1}Q_n]$ that includes ε_n. Therefore, the shocks ε_n and η_n have no impact on the optimal strategy.

To illustrate the shape of the solution to (3) given in Proposition 2, we conduct some numerical analysis. Figures 1–3 show the form and the basic comparative-static properties of the formulas. In the simulations we divide a trading day into 30-minute intervals to get 13 trading periods (the NYSE is open from 9:30am to 4:00pm). The amount to be traded is 100,000 shares and the initial price of the financial asset is \$20. A reasonable value for λ is 10^{-5} (see Hausman et al. (1992) or Kempf and Korn (1999)); if 1,000 shares are traded, then the price moves by one cent (provided that the price is measured in dollars). The range for the risk-aversion coefficient, R, is assumed to be between zero and 10^{-4}. Randomness is quantified by the magnitudes of the variances of the news revelation and the residual trades, which are set at 0.02 dollar^2 and 1,000 shares^2, respectively. The graphs below show that the trajectory of the optimal trades is typically a geometrically decreasing function of time.

Figure 1a computes the sequence of optimal trades for different values of the risk-aversion factor. The horizontal line (at $7692 \approx 100{,}000/13$) shows the strategy of a risk-neutral trader. As can be seen from this figure, higher risk aversion causes the trades to be shifted to early periods. For example, if $R = 2.5 * 10^{-4}$, then approximately 75% of the 100,000 shares are bought in the first two periods. The smaller R becomes, the more closely it approaches the risk-neutral horizontal. The sensitivity of trades to R is also reflected in the price changes. Figure 1b illustrates that expected prices change significantly in the early periods for high levels of R.

Figures 2a and 2b look at the reaction of optimal trades to various levels of the variance σ_ε^2. Like larger values of R, a higher σ_ε^2 causes traders to redistribute their trades from later to earlier periods (Figure 2a). Hence, the expected prices in the first periods are very sensitive to the level of σ_ε^2 (Figure 2b). Besides the direct effect of the variance on the price dynamics, it also alters the trading behavior. This further increases the change in price. Different levels of σ_η^2 show the same effect; a numerical illustration is therefore omitted here.

Figures 3a and 3b consider different values of λ. The higher the λ, the smaller are the orders in the first periods (Figure 3a). When λ is big, large trades in the beginning of trading would drive up prices too much, so that the succeeding pur-

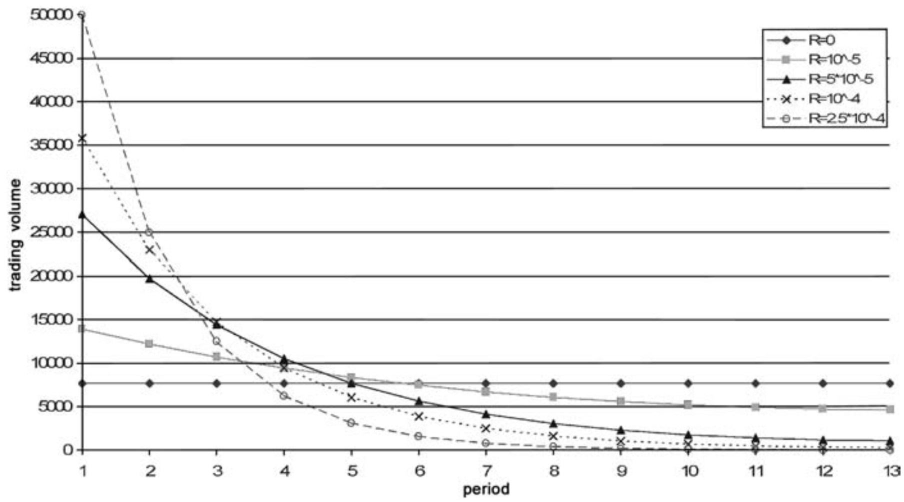

Figure 1a. Optimal trading volume and risk aversion. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $\lambda = 10^{-5}$, $\sigma_\varepsilon^2 = 0.02$, and $\sigma_\eta^2 = 1000$.

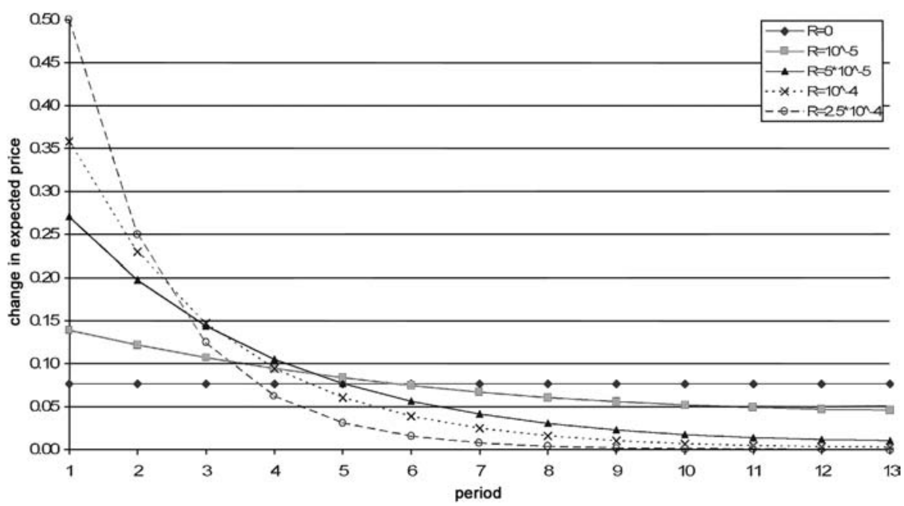

Figure 1b. Change in expected price and risk aversion. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $\lambda = 10^{-5}$, $\sigma_\varepsilon^2 = 0.02$, and $\sigma_\eta^2 = 1000$.

chases would take place at too high a price. Figure 3b demonstrates the sensitivity of expected price changes to different levels of λ.

If the price slopes, λ_n, are time-dependent, then one cannot derive a closed-form solution to (3), but at least a recursive solution can be obtained by solving the dynamic program in (14) (see Theorem 1).

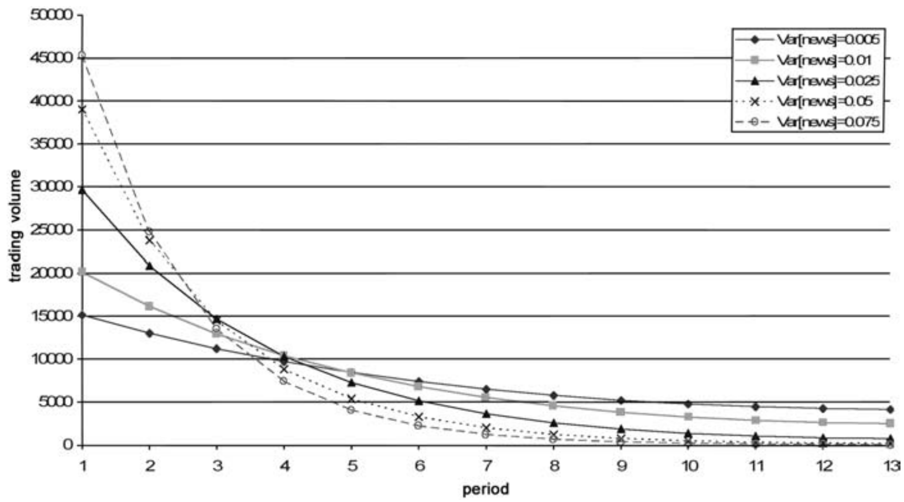

Figure 2a. Optimal trading volume and price volatility. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $\lambda = 10^{-5}$, $R = 5 * 10^{-5}$, and $\sigma_\eta^2 = 1000$.

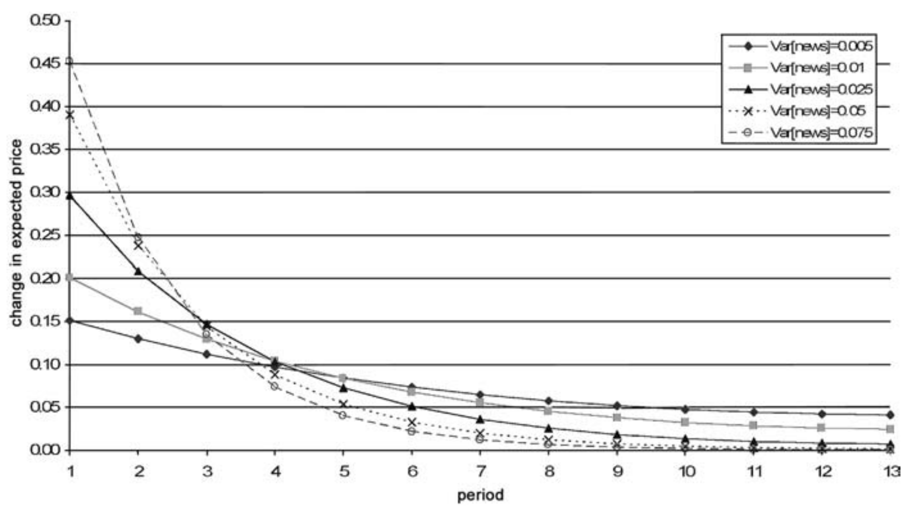

Figure 2b. Change in expected price and price volatility. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $\lambda = 10^{-5}$, $R = 5 * 10^{-5}$, and $\sigma_\eta^2 = 1000$.

PROPOSITION 3. Let Λ_N in (15) be positive semidefinite if $R > 0$ and positive definite if $R = 0$. The optimal trading sequence of (3) is given by

$$q_n = \left[1 - \frac{\lambda_n}{2\mu_{n+1} + R\sigma_\varepsilon^2}\right] Q_n \quad \text{for } 1 \le n \le N-1, \quad \text{and} \quad q_N = Q_N, \quad (18)$$

where

$$\mu_n = \lambda_n \left[1 + \frac{R}{2}\lambda_n \sigma_\eta^2 - \frac{\lambda_n}{2(2\mu_{n+1} + R\sigma_\varepsilon^2)}\right] \quad (19)$$

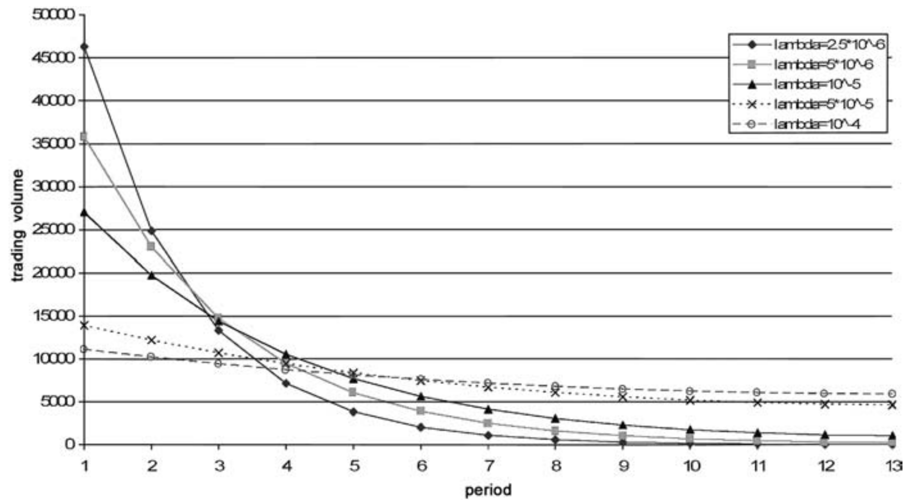

Figure 3a. Optimal trading volume and permanent price impact. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $R = 5 * 10^{-5}$, $\sigma_\varepsilon^2 = 0.02$, and $\sigma_\eta^2 = 1000$.

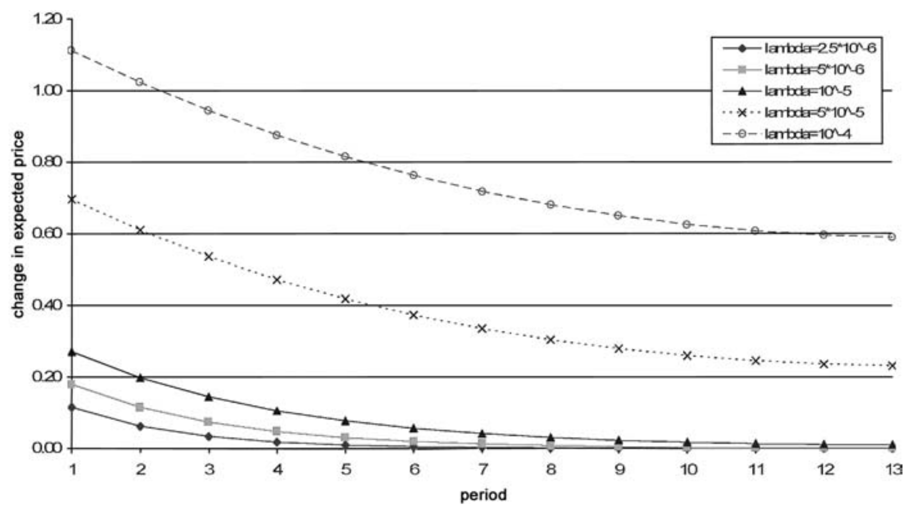

Figure 3b. Change in expected price and permanent price impact. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $R = 5 * 10^{-5}$, $\sigma_\varepsilon^2 = 0.02$, and $\sigma_\eta^2 = 1000$.

with boundary condition $\mu_N = \lambda_N(1 + \frac{R}{2}\lambda_N\sigma_\eta^2)$, and the Q_n's satisfy (13). The minimal costs evolve according to

$$L_n(\tilde{p}_{n-1}, Q_n) = \tilde{p}_{n-1}Q_n + \mu_n Q_n^2 \tag{20}$$

for $1 \leq n \leq N$.

Previous papers, e.g., Chan, Chung and Johnson (1995), find that the spread of NYSE stocks follows a U-shape pattern, with the spread widest after the open and

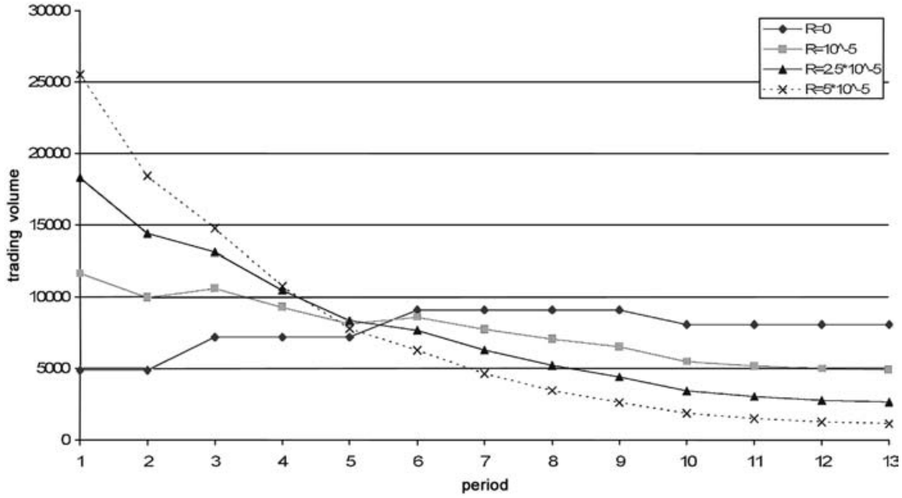

Figure 4. Optimal trading volume and time-dependent λ_n's. $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $\lambda_1 = \lambda_2 = 1.05 * 10^{-5}$, $\lambda_n = 10^{-5}$ if $n = 6, 7, 8, 9$, $= 1.025 * 10^{-5}$ if $n = 3, 4, 5, 10, 11, 12, 13$, $\sigma_\varepsilon^2 = 0.02$, and $\sigma_\eta^2 = 1000$.

prior to the close. Figure 4 assumes such a scenario by having the price-impact slopes of the early and last periods higher than those of the middle periods. The figure demonstrates that the optimal trades need no longer decline like in Proposition 2. Only if the trader is sufficiently risk averse, his trade pattern exhibits a geometrically decreasing shape. The risk-neutral trader always opts to shift more volume to the more liquid periods.

After the asset-specific parameters in (1) have been estimated, our formulas are useful to conduct two types of empirical tests. First, for a fixed level of risk aversion, one can check whether observed trading patterns of portfolio managers match our formulas. In particular, do liquidity, price or trade volume volatility affect the observed execution strategies as we predict? Second, if we assume that an agent follows Proposition 3, then his implied risk aversion can be derived from his actual trading behavior. This would allow to estimate and compare the level of risk aversion across different types of portfolio managers. Are growth or momentum managers more concerned about opportunity costs than other managers?

4.4. PRICE RULE WITH CONVEX UPDATING

The focus of this subsection is on (3) with a positive updating weight α in (2). In this case, the transaction price is higher than the price update after a trade, i.e., trades also have a temporary price impact. Moreover, the larger α the faster the price reverts to its original level after a trade.

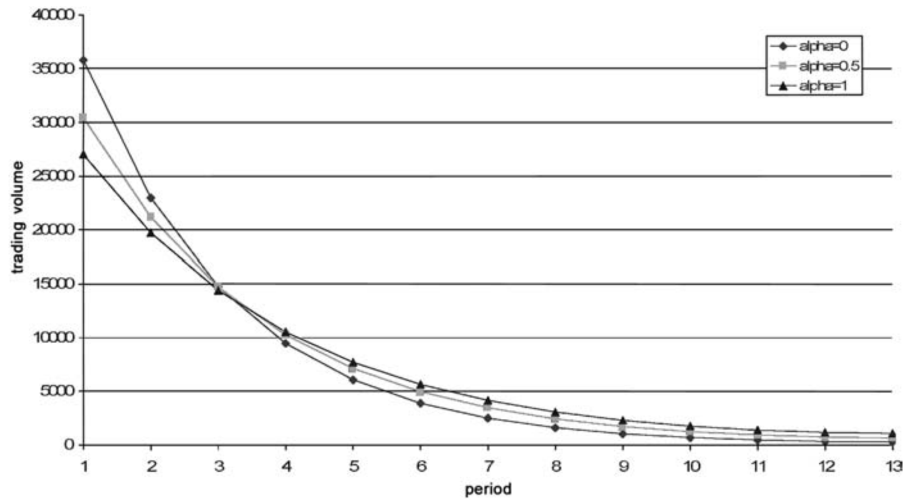

Figure 5. Optimal trading volume and different price-updating weights. $N = 13$, $p_0 = 20$, $Q = 10^5$, $R = 10^{-4}$, $\lambda = 10^{-5}$, $\sigma_\varepsilon^2 = 0.02$, and $\sigma_\eta^2 = 1000$.

Consider the two-period problem with $\lambda_1 = \lambda_2 = \lambda$. Execution costs are given by

$$\begin{aligned} C_2 = {} & [p_0 + \lambda(q_1 + \eta_1) + \varepsilon_1]q_1 \\ & + [p_1 + \lambda(Q - q_1 + \eta_2) - \alpha\lambda(q_1 + \eta_1) + \varepsilon_2](Q - q_1). \end{aligned}$$

Minimizing the cost function yields the ratio of the unique optimal solution

$$\frac{q_1}{q_2} = \frac{(1+\alpha)\lambda + R(\lambda^2\sigma_\eta^2 + \sigma_\varepsilon^2 - \alpha(1-\alpha)\lambda^2)}{\lambda(1 + \alpha + R\alpha\lambda)}. \tag{21}$$

From (21) follows that the effect of α on the optimal trades is ambiguous. However, for small values of λ and R, trading is postponed to the second period when α increases, with the right-hand side of (21) decreasing in α. This is because the effect of the trade q_1 on the variance of the execution costs becomes more pronounced when α rises. In addition, $q_1 \geq q_2$ or $q_1 \leq q_2$ may hold.

The previous analysis follows through. In particular, analogues of Propositions 1-3 can be derived, although with more complicated expressions (see Theorem 1 above and Theorem 2 in the Appendix). As in the case $\alpha = 0$, the optimal trades depend on the history only through the state variables Q_n, and past random shocks do not enter the formulas.

Numerical simulations show that including the updating weight α does not change the qualitative properties of the sequence of optimal trades. When $\alpha > 0$ the liquidity trader typically postpones trades to later periods, as can be seen from Figure 5. The impact of a trade on the cost's volatility increases in α, which induces trades to be less aggressive in the early periods. Thus, in real markets we expect more uniform trading volume across time for assets with higher estimates of α.

4.5. SHORT-SALE CONSTRAINTS

Some financial institutions are prohibited from short selling, for instance, mutual funds. It is hence of interest to re-examine the optimization problem in (3) for a liquidity trader who faces a short-sale constraint.

Formally, a trader never shorts if and only if $q_n \geq Q_n - Q$, or equivalently, $Q_n \leq Q$, $1 \leq n \leq N$. Our approach can accommodate such a restriction. In fact, the absence of price manipulation again implies the existence of a deterministic, time-consistent solution to (3) (see Lemma 3 in the Appendix). Formulas for the optimal trading sequence cannot be derived explicitly. Yet, numerical optimization can be applied efficiently for finding a minimum, as the solution is deterministic.

Evaluations of the formulas presented in Proposition 3 or Theorem 1 reveal that the short-sale constraint becomes binding whenever the price-impact slopes in the early periods are large relative to those in later periods. In such a situation, it would pay to short in the early periods in order to push down the prices of future purchases. Figure 6a illustrates how trading volume would be transferred to later periods when the trader wants to but is not allowed to short.

If the temporary price impact becomes larger (α increases), then the trader is less inclined to short: a higher α means that he can drive future prices and thus costs down by less (not shown in Figure 6a).

It can be optimal for an investor to sell in some periods without violating short sale constraints. As Figure 6b demonstrates, this occurs when a few periods are more illiquid half-way through the execution of an order. Imposing $q_n \geq 0$, $1 \leq n \leq N$, rules out such trading patterns. A sale constraint is more restrictive than a short-sale constraint, nonetheless, it may be reasonable to forbid sales of investors who ultimately must acquire a certain amount of shares. As Lemma 3 also holds for sale constraints, it is easy to find numerically the optimal execution strategy for a trader who is not allowed to sell. The effect of a sale constraint is that an investor abstains from trading in the less liquid periods and redistributes his individual orders across the other more liquid periods (see Figure 6b).

5. Portfolio Trading

In most applications, portfolio managers trade whole portfolios. In fact, in many cases they merely rebalance a portfolio and the aggregate value of their purchases is approximately equal to the aggregate value of their sales. To examine optimal portfolio trading we extend the setup in a straightforward manner. Individuals are now allowed to trade a portfolio of at most $K \geq 1$ securities. Prices, trades, and the stochastic variables in (1)-(3) then become K-dimensional vectors. Although the η_n's are *i.i.d.*, the components of η_n can be intratemporally correlated; the same

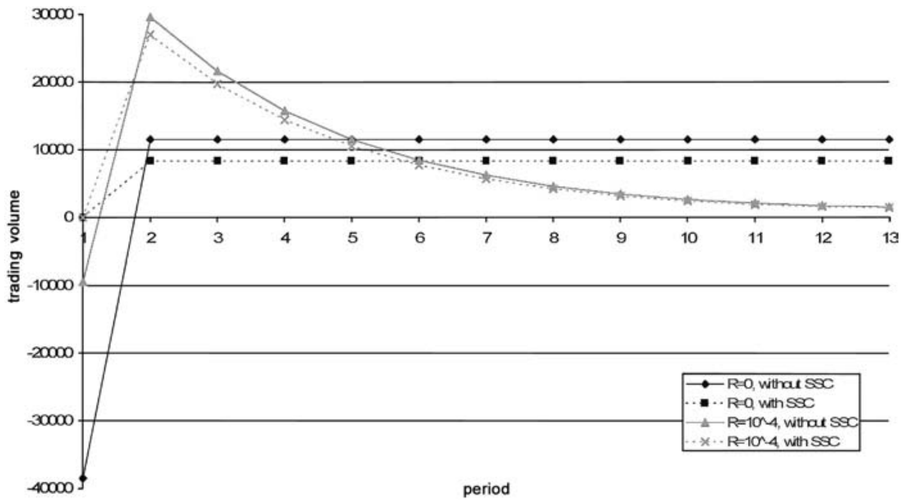

Figure 6a. Trading with a short-sale constraint (SSC). $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0$, $\lambda_1 = 1.5 * 10^{-5}$, $\lambda_n = 10^{-5}$ if $n = 2, \ldots, 13$, $\sigma_\varepsilon^2 = 0.01$, and $\sigma_\eta^2 = 1000$.

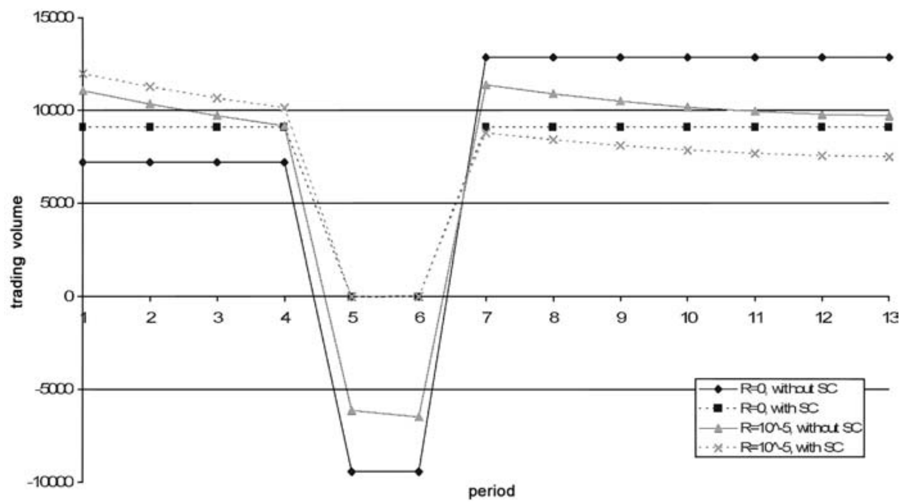

Figure 6b. Trading with a sale constraint (SC). $N = 13$, $p_0 = 20$, $Q = 10^5$, $\alpha = 0.25$, $\lambda_5 = \lambda_6 = 1.5 * 10^{-5}$, $\lambda_n = 10^{-5}$ if $n \neq 5, 6$, $\sigma_\varepsilon^2 = 0.01$, and $\sigma_\eta^2 = 1000$.

applies to ε_n. We consider here only $\alpha = 0$, i.e., the price process in (1), where trades have only a permanent price impact. The price-impact slopes λ_n become $K \times K$ positive definite matrices that incorporate not only all assets' individual price impacts but also *cross-price impacts*. The vector $Q \in \mathbf{R}^K$ in (3) summarizes the number of shares to be traded for each asset. It can include both purchases and sales. The covariance matrices of ε_n and η_n are Σ_ε and Σ_η, respectively, and $I_{K \times K}$ is the $K \times K$ identity matrix.

Similar to the single-asset case, sufficient conditions for the existence of a time-consistent solution can be found; if these conditions hold, the solution is unique and can be obtained by solving the dynamic program in (14). For the risk-averse case, the absence of price manipulation guarantees the existence of a solution, while a more technical condition is required for the risk-neutral case, whose details are not presented here.

PROPOSITION 4. The optimal trading sequence for the multi-asset version of (3) (if a solution exists) is

$$q_n = \left[I_{K\times K} - (2\Psi_{n+1} + R\Sigma_\varepsilon)^{-1}\lambda_n\right] Q_n \quad \text{for } 1 \le n \le N-1, \quad \text{and} \quad q_N = Q_N, \tag{22}$$

where

$$\Psi_n = \lambda_n \left[I_{K\times K} + \frac{R}{2}\Sigma_\eta\lambda_n - \frac{1}{2}(2\Psi_{n+1} + R\Sigma_\varepsilon)^{-1}\lambda_n\right], \tag{23}$$

$1 \le n \le N-1$, with boundary condition $\Psi_N = \lambda_N(I_{K\times K} + \frac{R}{2}\Sigma_\eta\lambda_N)$, and the Q_n's satisfy (13). The minimal costs can be obtained from

$$L_n(\tilde{p}_{n-1}, Q_n) = \tilde{p}_{n-1}^T Q_n + Q_n^T \Psi_n Q_n \tag{24}$$

for $1 \le n \le N$.

Again, as for the single-asset case, the optimal trades are deterministic functions of the history. Evidently, the optimal trading strategy for one security depends on the parameters and state variables of all the other securities, unless the λ_n's and the covariance matrices are all diagonal. Diagonal λ_n's mean that trading one asset has no impact on the prices of the other assets, and diagonal covariance matrices imply that the stochastic terms are uncorrelated.

The formulas (22)–(24) can be conveniently used for comparative statics to assess how price uncertainty and cross-price impacts affect the optimal portfolios. In particular, Figure 7 demonstrates the effect of price volatility on trading behavior when two assets are held in the portfolio. As benchmark serves the case where the variances are the same across assets, resulting in the same trading quantities for both assets (shown as Case I in Figure 7). If, ceteris paribus, the volatility of asset B increases, then more shares of B and fewer of A are traded than before in early periods (Case II). Consequently, volume in the beginning of trading is biased toward the more volatile asset, thanks to risk aversion. Case III shows that a negative price correlation between the assets causes not as much of aggressive trading early on. In this situation, each asset provides a hedge for the other and therefore a risk-averse trader needs less to hurry up with his orders.

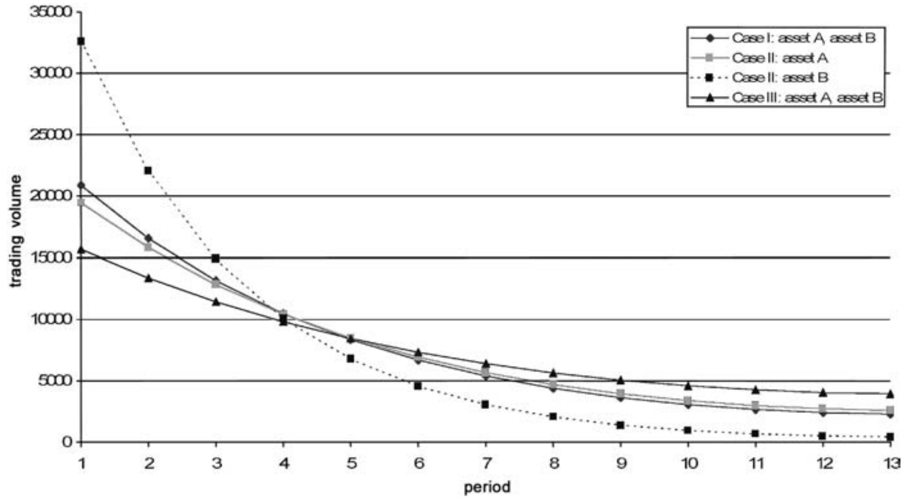

Figure 7. Trading a portfolio of two assets – the effect of price volatility on volume. Case I: $N = 13$, $p_0 = (20, 20)$, $Q = (10^5, 10^5)$, $R = 10^{-5}$, $\alpha = 0$, $\lambda_{AA} = \lambda_{BB} = 10^{-5}$, $\lambda_{AB} = \lambda_{BA} = 10^{-6}$, $\sigma_\eta^2(A, A) = \sigma_\eta^2(B, B) = 1000$, $\sigma_\eta^2(A, B) = \sigma_\eta^2(B, A) = 10$, $\sigma_\varepsilon^2(A, A) = \sigma_\varepsilon^2(B, B) = 0.05$, $\sigma_\varepsilon^2(A, B) = \sigma_\varepsilon^2(B, A) = 0.01$; Case II: (only change) $\sigma_\varepsilon^2(B, B) = 0.15$; Case III: (only change) $\sigma_\varepsilon^2(A, B) = \sigma_\varepsilon^2(B, A) = -0.02$.

The optimal trading quantities are sensitive to the levels of the cross-price impacts. Trades are deferred to later periods when cross-price impacts rise. Moreover, trading volume in early periods is always concentrated on those assets with the lowest individual price-impacts in the portfolio.

6. Analysis of the Cost Function

If the price-impact sequence, λ_n, is constant, then a closed-form expression for the cost function in (3) can be derived. To account for commission fees, we incorporate fixed transaction costs, denoted by $k(N)$. For simplicity, we calculate the cost function only for a single asset.

The difference equation in (7) is a Riccati equation when all λ_n's are constant. By solving (7), we obtain

$$L(Q, N) = p_0 Q + \left\{ \lambda + \frac{R}{2}(\lambda^2 \sigma_\eta^2 + \sigma_\varepsilon^2) - \frac{\sqrt{b}}{2} \frac{r^{2N-1} - r}{r^{2N} - 1} \right\} Q^2 + k(N), \qquad (25)$$

where

$$
\begin{aligned}
r &= \left(a + \frac{c}{2}\right)/\sqrt{b} + \sqrt{\left(a + \frac{c}{2}\right)^2/b - 1},\\
a &= \lambda\left(1 + \alpha + \frac{R}{2}\lambda\sigma_\eta^2\right),\\
b &= \lambda^2(1 + \alpha + R\alpha\lambda\sigma_\eta^2)^2,\\
c &= R(\alpha^2\lambda^2\sigma_\eta^2 + \sigma_\varepsilon^2).
\end{aligned} \tag{26}
$$

Observe that $L(Q, N)$ converges to the cost function for a risk-neutral trader when $R \downarrow 0$, namely,

$$
L(Q, N) = p_0 Q + \frac{\lambda}{2}\left[1 - \alpha + \frac{1+\alpha}{N}\right] Q^2 + k(N). \tag{27}
$$

The fees $k(N)$ are assumed to be increasing and convex in the number of trades.

As can be seen from the equations in (25)–(27), the cost function is increasing in the total number of shares traded, the level of risk aversion, the price and volume volatilities σ_ε^2 and σ_η^2, and the magnitude of the permanent price impact, and decreasing in the price-updating weight α. These relationships hold for the whole parameter space.

It would be interesting to gauge the empirical relation between these parameters and the execution costs and see whether our predictions apply. The function $L(Q, N)$ is difficult to measure empirically because it depends on the level of risk aversion which cannot be observed directly. Therefore, it is more natural to employ a measure like $TC(Q, N) \equiv \mathbb{E}[\sum_{n=1}^{N}(p_n - p_0)q_n]$ to estimate the execution costs. Since the optimal trading strategy is a function of the model parameters, so is $TC(Q, N)$.

The level of risk aversion is positively related with $TC(Q, N)$. If R rises, more volume is traded in the early periods which causes the expected price-impact costs to go up. The permanent price-impact parameter has two opposite effects on $TC(Q, N)$. A higher λ, on the one hand, increases $TC(Q, N)$ directly by amplifying the price impact of each trade. On the other hand, it decreases $TC(Q, N)$ as it induces a more uniform trading pattern across time (recall Figures 3a and 3b). The first effect dominates and hence $\partial TC/\partial\lambda > 0$, as long as R is not too big. Chan and Lakonishok (1995) analyze the effects of commission costs, market capitalization (as a proxy for market liquidity), and managerial strategy on the price impact and execution costs of institutional trades. Consistent with our model, they report that execution costs are higher for low market-capitalization stocks. Perhaps surprisingly, they report that the type of the manager is the most important cost factor. Since less patient managers incur a higher price impact and execution costs than patient ones, this could be interpreted as $\partial TC/\partial R$ being relatively large.

To our knowledge, there exist no empirical studies on how price volatility, trade volume volatility, or the speed of mean reversion influence execution costs. As long as a portfolio manager is risk averse, price (or trade volume) volatility should be a significant determinant of the execution costs ($\partial TC/\partial\sigma_\varepsilon^2 > 0$ and $\partial TC/\partial\sigma_\eta^2 > 0$ because of large early trades). In particular, one could test the impact of price volatility across different types of managers and asset classes. As regards to mean reversion, one could estimate how the price-updating weight affects the execution costs across stocks. We predict that costs are lower for stocks with high price-updating weights, because a higher α causes future prices to be less sensitive to trades.

The cost function also determines the optimal number of trades, since with fixed, positive transaction costs, the number of trades emerges endogenously. The following subsections discuss two cases regarding the timing of trades. First, we set the time between trades equal to a fixed number τ, which will enable us to deduce and examine the optimal trade duration τN. Then, we assume that all trading occurs within the time interval $[0, \tau]$ and investigate the optimal trade frequency.

6.1. CONSTANT TIME BETWEEN TRADES

Given a constant time between trades, τ, the liquidity trader aims at minimizing the cost function with respect to N in order to find the optimal number of trades, N^*. Since L is strictly convex in N, the cost function always exhibits a global minimum. The optimal trade duration, τN^*, is finite when k is an increasing and convex function. For example, if $k(N) = kN$ and $R = 0$, then $N^* = Q\sqrt{(1+\alpha)\lambda/(2k)}$, provided we treat N as a continuous variable. As a consequence, the optimal number of trades increases in Q, α, and λ, and decreases in k. The more sensitive the price reaction to trades (the higher the λ), or the larger the temporary price impact, the more often the trader chooses to trade.

More generally, we obtain the following comparative statics (for the proof see the Appendix).

PROPOSITION 5. Regardless of the trader's type we have $\partial N^*/\partial Q > 0$ and $\partial N^*/\partial\alpha > 0$. The sign of $\partial N^*/\partial\lambda$ is positive when either (i) $R = 0$ or (ii) $R > 0$ and $\lambda^2\sigma_\eta^2$ is sufficiently small relative to σ_ε^2. If the trader is risk-averse, then N^* is always decreasing in σ_ε^2, while $\partial N^*/\partial\sigma_\eta^2 < 0$ and $\partial N^*/\partial R < 0$ only if α is sufficiently small. In case $k(N) = kN$, $\partial N^*/\partial k < 0$ for all $R \geq 0$.

Thus, it is optimal for a risk-averse trader to submit fewer orders when the price volatility goes up. Intuitively, the trader compensates for the higher price volatility by reducing the trade duration. Figure 8a illustrates the monotone relation between N^* and σ_ε^2 for various levels of R. If the time between trades is 30 minutes, then the range of the trade duration would be between 1.5 hours and 3 days, given the parameter values in the figure.

Also the volatility of other traders' volume, σ_η^2, has a negative effect on the optimal number of trades, unless α is too large. If the trading activity of other mar-

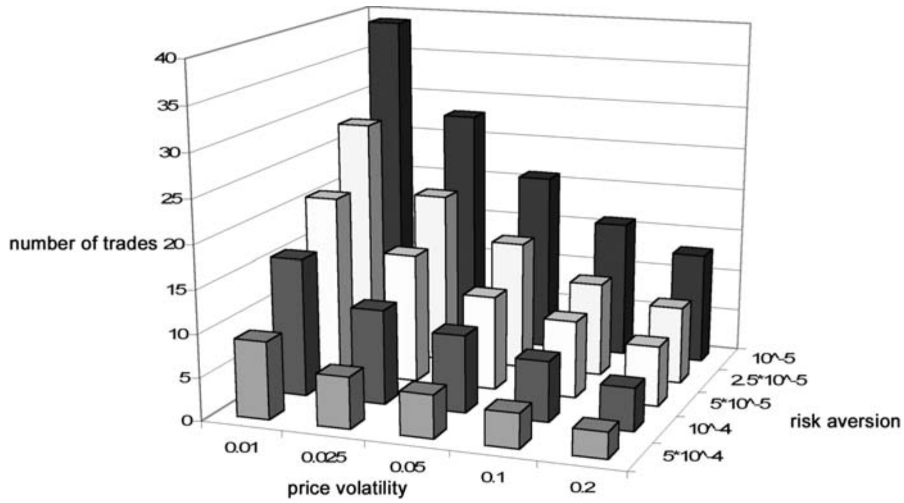

Figure 8a. Optimal number of trades for different levels of price volatility and risk aversion. $p_0 = 20$, $Q = 10^5$, $\lambda = 10^{-5}$, $\alpha = 0.25$, $\sigma_\eta^2 = 1000$, and $k(N) = 2N$.

ket participants makes prices more uncertain, then the trader wishes to complete his trades earlier. The situation is different when α is close to one, where the temporary price impact is high and the permanent one is low. In this case, it pays to trade more often, as the impact of others' trades decays very fast.

Combining the effects of σ_ε^2 and σ_η^2, the number of trades diminishes with the level of risk aversion. Only if α is large enough, $\partial N^*/\partial R$ may flip sign and become positive.

Both the permanent and the temporary price-impact parameters are positively related with N^*. If λ increases, the trader attempts to reduce the total price impact by submitting more but smaller orders. If α rises, the trader prefers a longer trade duration in order to benefit from the improved price reversion. While $\partial N^*/\partial\alpha > 0$ throughout the parameter space, the sign of $\partial N^*/\partial\lambda$ can become negative if $\lambda^2\sigma_\eta^2 \gg \sigma_\varepsilon^2$. A negative $\partial N^*/\partial\lambda$ arises when the marginal benefit of decreasing the volatility of the execution costs by trading fewer times overcompensates the marginal benefit of lowering the mean of the execution costs by trading more often. The inequality $\lambda^2\sigma_\eta^2 \gg \sigma_\varepsilon^2$, though, is unlikely to hold in practice, as λ is typically of order 10^{-5}. Figure 8b shows the relation between N^* and λ for a subset of the parameter space.

Keim and Madhavan (1995) consider three different investment styles of institutional portfolio managers and present evidence that value managers exhibit the longest trade duration, followed by index and technical managers (Table 2, p. 379). An interesting exercise would be to estimate our model parameters conditional on the investment style. Given the trade duration and the estimates of the stock-specific parameters, one could deduce the level of risk aversion implied by the trade duration for each style. Most likely, the variables Q and R are directly determined

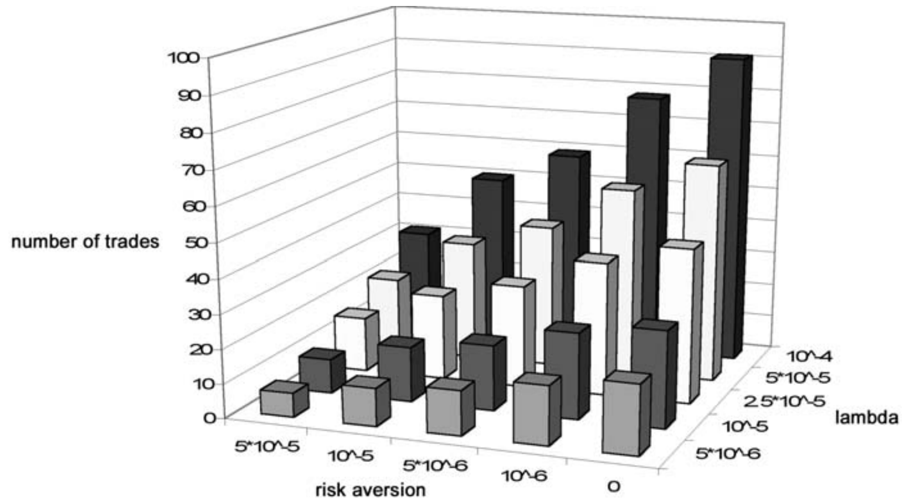

Figure 8b. Optimal number of trades for different levels of permanent price impact and risk aversion. $p_0 = 20$, $Q = 10^5$, $\alpha = 0.25$, $\sigma_\eta^2 = 1000$, $\sigma_\varepsilon^2 = 0.02$, and $k(N) = 75N$.

by the investment style. But also stock-specific parameters may be affected by the investment style, because some styles may be biased toward a particular subset of the stock universe. For example, growth-oriented managers may tend to invest in high-risk and low-liquidity stocks. In any case, if value managers are less risk-averse than index managers and index managers less risk-averse than technical managers, then Keim and Madhavan's finding would be confirmed.

Keim and Madhavan (1995) also find that, contrary to theory, trade duration increases with market capitalization even when order quantity is controlled for (Table 4, p. 384, and Figures 1 and 2, pp. 386–387). A possible explanation is that submission costs decline with market liquidity, making a trade break-up that reduces the overall price impact more attractive. To investigate this hypothesis, we could model the transaction-cost function k to depend on λ, and then ask what parameter values would cause $\partial N^*/\partial\lambda$ to be negative.

6.2. VARIABLE TIME BETWEEN TRADES

Suppose that calendar time is a fixed interval $[0, \tau]$ and that trades are equally spaced in time. Hence, if the trader chooses to transact N times, the time between trades is τ/N. The arrival of news and other people's trades are both still $i.i.d.$ processes. Their per-unit-of-time variances over the whole trading interval are $\sigma_{\varepsilon\tau}^2$ and $\sigma_{\eta\tau}^2$, respectively. Therefore, the per-interval variances, $\sigma_\eta^2(N)$ and $\sigma_\varepsilon^2(N)$, satisfy $\sigma_\eta^2(N) = \tau\sigma_{\eta\tau}^2/N$ and $\sigma_\varepsilon^2(N) = \tau\sigma_{\varepsilon\tau}^2/N$.

The risk-aversion coefficient, however, is not assumed to change with the trading frequency. The trader's dislike of price volatility is independent of how short the time between trades is. This is a reasonable premise, since the declining variance per period already decreases the risk part of the trader's utility function, taking into account the desired effect that volatility matters less to the trader if the time between trading becomes smaller.

Unfortunately, the literature provides little guidance on the possible relationship between the permanent and temporary price impacts and the trading rate. A model in which α, λ, and N^* are determined simultaneously in equilibrium would be required to address this question. As this is outside of the scope of this paper, we assume that the permanent price-impact is independent of the speed at which the investor trades, while the price-updating weight is nonincreasing in the trading frequency. The slope λ reflects the degree of asymmetric information and does not change when the liquidity trader transacts more frequently, because his trades are per definition not motivated by private information about the asset's value. In spite of this, by raising the submission frequency the trader initially not only clears the order book at the prevailing quotes, but also intensifies other market participants, suspicion that an insider is trading. Both effects cause the bid-ask spread to widen. Once the market reassesses the orders to be liquidity-initiated, the likelihood of insider trading is updated downwards, which leads to a partial price reversion. Typically, the more concentrated the liquidity trading, the longer it takes for the price reversion to get underway, that is, α goes down with the number of trades. Biais et al. (1995) provide empirical evidence of this phenomenon.

Albeit the cost function may no longer be convex when σ_η^2, σ_ε^2, and α are functions of N, a global minimum still exists in the half-open interval $[1, \infty)$. This is because the second term in (25) converges to a constant, whereas k is increasing in N. As in the previous subsection, we are able to determine how the model parameters affect the optimal number of trades. Here are the main findings. (Since the proof follows the same steps as those presented in the proof of Proposition 5 and is rather lengthy, we omit it.)

PROPOSITION 6. Let $\sigma_\eta^2(N) = \tau\sigma_{\eta\tau}^2/N$, $\sigma_\varepsilon^2(N) = \tau\sigma_{\varepsilon\tau}^2/N$, $\alpha(N) = \alpha_\tau/N$, and $N^* > 2$. Then, for all sufficiently small R, $\partial N^*/\partial Q$ and $\partial N^*/\partial\lambda$ are both positive, while $\partial N^*/\partial\alpha_\tau$ is negative. Furthermore, for small λ and α_τ, the signs of $\partial N^*/\partial R$, $\partial N^*/\partial\sigma_{\varepsilon\tau}^2$, and $\partial N^*/\partial\sigma_{\eta\tau}^2$ are all positive when $R > 0$.

The trader's response to higher price or volume volatility or higher risk aversion is to shorten the time between his trades. Smaller orders move the price less and a higher trade frequency reduces the price volatility between trades. The optimal trade frequency goes up when an asset is more illiquid, because splitting orders into smaller pieces, which brings down the overall price impact, pays particularly when liquidity is low. If α_τ rises, liquidity traders decelerate the speed at which they submit orders to benefit from the higher level of price reversion.

Using tick data, the parameters $\sigma_{\varepsilon\tau}^2$, λ, and α_τ could be estimated for each stock an institutional portfolio manager is trading. Then, after subtracting the orders of the manager, one could infer $\sigma_{\eta\tau}^2$. Equipped with these estimates, one could test whether the intraday trading pattern of the portfolio manager confirms our predictions.

7. Autocorrelated Noise Trades

So far we assumed that noise trades, η_n, are independent across time. In contrast, Choi, Salandro and Shastri (1988) provide empirical evidence that order flow is positively serially correlated. To incorporate autocorrelation, we extend the price dynamics in (2) by modelling the noise trades as an AR(1) process, i.e.,

$$\eta_n = \rho\eta_{n-1} + \nu_n, \tag{28}$$

where $\rho \in \mathbf{R}$ and $\nu_n \sim i.i.d(0, \sigma_\nu^2)$. The stochastic innovation ν_n is known to the liquidity trader only after trades took place at time n.

The liquidity trader uses Equation (28) to forecast future noise trades and adjusts his optimal trading strategy accordingly. Unfortunately, the Bellman equation in (14) cannot be employed to derive the optimal strategy for a risk-averse trader. If $R > 0$, the equivalence between (14) and the original problem in (3) breaks down, since the optimal trading volume typically will be a nonlinear function of the state variables. Hence, we only discuss here the case $R = 0$.

For a risk-neutral trader, the optimal policy is described by

$$\begin{aligned} q_n &= \left[1 - \frac{(1+\alpha)\lambda_n}{2\mu_{n+1}}\right] Q_n + \frac{\rho\beta_{n+1}}{2\mu_{n+1}}\eta_{n-1}, \quad 1 \le n \le N-1, \\ q_N &= Q_N, \\ L_n(\tilde{p}_{n-1}, Q_{n-1}, Q_n, \eta_{n-1}) &= (\tilde{p}_{n-1} - \alpha\lambda_{n-1}Q_{n-1})Q_n \\ &\quad + \beta_n Q_n \eta_{n-1} + \gamma_n \eta_{n-1}^2 + \mu_n Q_n^2 + \delta_n, \end{aligned} \tag{29}$$

where

$$\begin{aligned} \beta_n &= \frac{\rho(1+\alpha)\lambda_n}{2\mu_{n+1}}\beta_{n+1} + \rho\lambda_n - \alpha\lambda_{n-1}, \\ \gamma_n &= \rho^2\gamma_{n+1} - \frac{\rho^2\beta_{n+1}^2}{4\mu_{n+1}}, \\ \mu_n &= \alpha\lambda_{n-1} + \lambda_n - \frac{(1+\alpha)^2\lambda_n^2}{4\mu_{n+1}}, \\ \delta_n &= \delta_{n+1} + \gamma_{n+1}\sigma_\nu^2, \end{aligned} \tag{30}$$

for $1 \le n \le N-1$, $\beta_N = \rho\lambda_N - \alpha\lambda_{N-1}$, $\gamma_N = \delta_N = 0$, and $\mu_N = \alpha\lambda_{N-1} + \lambda_N$. The existence of a solution is guaranteed by $\mu_n > 0$, $2 \le n \le N$.

By (29), the optimal trade size is no longer deterministic when noise trades are autocorrelated but depends on the level of the noise trades. The relation between the parameters α and ρ determines how η_{n-1} affects the optimal trading volume.

To simplify the analysis of the equations in (29) and (30) let λ_n be constant. If $\rho > \alpha$, then β_n is positive and thus q_n is increasing in η_{n-1}, implying that positive realizations of η_{n-1} cause the liquidity trader to buy more shares in period n, even though the current execution price, p_n, goes up when $\eta_{n-1} > 0$. Since future prices are expected to rise due to the positive autocorrelation of the noise trades, the higher price-impact costs incurred at time n are more than offset by the cost savings of trading less in the future.

Alternatively, if $\rho < \alpha$, a positive η_{n-1} decreases the optimal trade size in period n. When the transitory price impact exceeds the effect stemming from the serial correlation of the noise trades, it is beneficial to defer trading volume to later periods, because prices are expected to revert. Finally, if $\rho = \alpha$, then $\beta_n = \gamma_n = \delta_n = 0$ and η_{n-1} has no influence on the optimal trading sequence. In this situation, any price movements induced by noise trades are expected to be counterbalanced by the temporary price impact.

8. Extant Literature

Bertsimas and Lo (1998), using a simpler form of the price process (1), show that to minimize expected execution costs of trading a fixed number of shares, a trader should split his orders evenly over time. However, institutions typically trade more up-front as we observed in Section 2. To accommodate this empirical regularity, Bertsimas and Lo add an AR(1) news process (with a positive coefficient) to the price equation (1) and demonstrate that the even-split trading strategy is no longer optimal in this case.

If the trader is a buyer and the news is good, he will accelerate his trades in anticipation of more good news which will drive the price higher. This modeling approach raises a few issues such as, who is on the other side of the accelerated trades? Shouldn't the prices anticipate the future news? If the goal is to explain institutional trading volume profiles over short time intervals, Bertsimas and Lo presumably predict positively (negatively) sloped profiles for purchases (sales) following bad news and negatively (positively) sloped profiles for purchases (sales) following good news. In contrast to Bertsimas and Lo, we follow a classical paradigm in finance which says that news have no persistence and occur surprisingly. This prevents the sequence of optimal trades to be driven by the explicit shape of the news process.

By assuming either risk-averse preferences or prices with time-dependent price impact our model mimics actual trading strategies of large traders. This is the main contribution of this paper.

Almgren and Chriss (2000) independently from Bertsimas and Lo's and our work extend the Bertsimas and Lo framework to allow risk aversion. In particular, they construct an efficient frontier of optimal execution similar to that in Remark 1 of Section 3. The focus and analysis of Almgren and Chriss differ considerably from ours. While they limit their attention to the shape of the efficient frontier, we study the underlying optimization problem and the properties of the optimal policy. More important, their framework is more restrictive than ours.

First, Almgren and Chriss assume as Bertsimas and Lo (1998) that the price impact is time-independent. Aware of this limitation, they mention in their conclusion the importance of extending their setting to permit time-dependent price impacts. Intraday data show that the price impact of trade size varies during the day (see Madhavan et al. (1997)). Hence, the optimal trading strategies should be time-dependent. If the price impact is time-dependent, the existence of a solution to (3) requires separate analysis. Another contribution of this paper is the finding that the absence of price manipulation guarantees the existence and uniqueness of the solution to (3). Second, both Almgren and Chriss and Bertsimas and Lo assume away fixed transactions costs and short-sale constraints. Third, the existence of an optimum in Almgren and Chriss's model requires that trades have a temporary price impact. Finally, they include a drift term to their price process. But as Huberman and Stanzl (2004) prove, this assumption is problematic, because price manipulation and quasi-arbitrage opportunities could arise.

Vayanos (2001) studies a general equilibrium model of a large trader who receives random shocks to his holding of a risky asset and interacts with competitive market makers to share risk; it is another important antecedent to our paper. Vayanos shows that the large trader who has received an unexpectedly large endowment of the risky asset reduces his risk exposure by selling it at a decreasing rate or by first selling it and then buying back some of the units sold. An increase in the large trader's aversion to risk is associated with more aggressive trading initially, and average execution time increases with the level of noise trades.

Some of our model's predictions overlap those of Vayanos (2001) although the models themselves are very different. The overlap suggests robustness of the results. Moreover, our model considers cases and makes predictions that Vayanos' work does not. These include: (1) the effect of the speed of price mean reversion on the optimal trading strategy and execution costs; (2) the optimal strategy of executing a portfolio of assets that impounds cross-price impacts; (3) the impact of short-sales constraints on the optimal execution (many large money managers cannot sell short!); (4) the derivation of optimal trading frequency and its dependence on asset- and trader-specific parameters; (5) the possibility of autocorrelated noise trades; (6) non-normal random shocks. A result found in Vayanos, but which is outside the scope of our paper is that the large trader's holdings increase over time when his aversion to risk is large relative to that of the market makers.

9. Concluding Remarks

This paper studies the optimal behavior of a trader who wishes to buy (or sell) a given quantity of a security within a certain number of trading rounds. He is constrained to submit only market orders and his trades affect current and future prices of the security. He therefore breaks up his trades into a sequence of smaller orders. Risk neutrality implies that these smaller orders are equal. If the trader is risk averse, though, the magnitude of his trades declines over time. We also examine the optimal trade duration and how short-sale constraints affect the trading behavior.

In case noise trades are autocorrelated, order flow can be decreasing over time even when the trader is risk-neutral. A trader facing positively autocorrelated noise trades expects a sequence of purchases by noise traders in the future after observing them buying today. To mitigate the future price impacts of his trades, he lowers his future trading volume by submitting a higher order today.

Our results are useful for conducting empirical tests. First, having estimated the asset-specific model parameters, one can compare actual trading patterns with those our formulas predict. In particular, how do price (or volume) volatility, liquidity, or speed of price reversion affect the trading profiles of money managers compared to our findings? On the other hand, unobserved manager characteristics, such as the level of risk aversion, can be inferred from observed trading patterns by applying our formulas. Second, this paper implies that price volatility and the speed of price reversion are significant determinants of the execution costs, a hypothesis that has not been tested yet empirically. Third, how do cross-price impacts and price correlations influence portfolio rebalancing in practice?

Finally, our model makes explicit predictions about the optimal trade duration and the optimal trade frequency. If the time between trades is constant, investors wish to trade more often when risk aversion, price volatility, or liquidity are low, or speed of price reversion is high. If the time horizon of trading is fixed, higher risk aversion or price volatility cause a higher trading frequency, as do lower liquidity or speed of price reversion. Note that agents always trade more patiently when price reversion is faster.

Several extensions to our analysis are desirable. On the theoretical level, this paper does not study the optimal policy of a trader who can submit limit orders in addition to market orders, nor does it study the circumstances under which the intertrading intervals are chosen in a continuous-time setup.

Moreover, liquidity, which is measured by the price-impact slopes, may be stochastic. One way to take into account the uncertainty of the market's liquidity is to model the price-impact slopes as a stochastic process and study the optimal trading behavior in the same spirit as in this paper. Presumably, liquidity risk induces investors to trade fewer times and submit larger quantities in early periods. Nevertheless, if price-impact slopes are positively correlated, it may be advantageous to postpone orders to see whether liquidity drifts upwards. One could model the price-impact slopes as an autoregressive process and try to find an optimal execution strategy.

Appendix

Proof of Theorem 1. As first step, we derive the solution to (14) by backward induction. In period N, the optimal cost function, as a function of the state variables $\tilde{p}_{N-1}$, Q_{N-1}, and Q_N, is given by

$$\begin{aligned} L_N(\tilde{p}_{N-1}, Q_{N-1}, Q_N) &= \mathbb{E}_N[p_N q_N] + \frac{R}{2}\mathrm{Var}_N[p_N q_N] \\ &= [\tilde{p}_{N-1} - \alpha\lambda_{N-1}(Q_{N-1} + \eta_{N-1})]Q_N \\ &\quad + \left[\alpha\lambda_{N-1} + \lambda_N\left(1 + \frac{R}{2}\lambda_N\sigma_\eta^2\right)\right]Q_N^2, \end{aligned} \tag{31}$$

since in the last period $q_N = Q_N$ must be traded. Let us define $\mu_N \equiv \alpha\lambda_{N-1} + \lambda_N(1 + \frac{R}{2}\lambda_N\sigma_\eta^2)$.

We will show that if the formulas in (7)–(10) hold for $n+1$, they are also true for n, completing the induction argument. Employing the price dynamics in (2) and the induction hypothesis, we have

$$\begin{aligned} L_n(\tilde{p}_{n-1}, Q_{n-1}, Q_n) &= \inf_{q_n}[\tilde{p}_{n-1} - \alpha\lambda_{n-1}(Q_{n-1} + \eta_{n-1})]Q_n \\ &\quad + \alpha(\lambda_{n-1} - \lambda_n)Q_n^2 + (1+\alpha)\lambda_n Q_n q_n \\ &\quad + \mu_{n+1}(Q_n - q_n)^2 + \frac{R}{2} \\ &\quad \times\{\lambda_n^2[\alpha q_n + (1-\alpha)Q_n]^2\sigma_\eta^2 + (Q_n - q_n)^2\sigma_\varepsilon^2\}. \end{aligned} \tag{32}$$

Taking the first-order condition for this expression yields (8) to be the unique minimum, provided that the second-order condition $2\mu_{n+1} + R(\alpha^2\lambda_n^2\sigma_\eta^2 + \sigma_\varepsilon^2) > 0$ is met. The minimal cost, $L_n(\tilde{p}_{n-1}, Q_{n-1}, Q_n)$, then can be computed to be

$$\begin{aligned} &[\tilde{p}_{n-1} - \alpha\lambda_{n-1}(Q_{n-1} + \eta_{n-1})]Q_n \\ &\quad + \left[\alpha\lambda_{n-1} + \lambda_n(1 + \frac{R}{2}\lambda_n\sigma_\eta^2) - \frac{\lambda_n^2(1 + \alpha + R\alpha\lambda_n\sigma_\eta^2)^2}{2[2\mu_{n+1} + R(\alpha^2\lambda_n^2\sigma_\eta^2 + \sigma_\varepsilon^2)]}\right]Q_n^2. \end{aligned}$$

Consequently, all formulas in (7)–(10) hold for n, too.

As last step, we need to demonstrate that the solution given in (7)–(10) is the only time-consistent solution. To this end, let $\{q_n, L_n\}_{n=1}^N$ be the unique solution to (14) obtained above. Evidently, in the last period, $q_N = Q_N$ for all possible values of Q_N, and therefore $\{q_N, L_N\}$ is the only solution to (3) for period N. Next, assume as induction hypothesis that $\{q_k, L_k\}_{k=n+1}^N$ is the unique, time-consistent solution to (3) for the periods $n+1 \leq k \leq N$. In this case, simple calculations reveal that

$$\mathbb{E}_n\left[\sum_{j=n}^{N} p_j q_j\right] + \frac{R}{2}\mathrm{Var}_n\left[\sum_{j=n}^{N} p_j q_j\right]$$

$$= \mathbb{E}_n[p_n q_n + L_{n+1}] + \frac{R}{2}\mathrm{Var}_n[p_n q_n + L_{n+1}], \tag{33}$$

which implies that $\{q_k, L_k\}_{k=n}^{N}$, in turn, is the unique, time-consistent solution for the periods $n \leq k \leq N$, completing the proof. □

LEMMA 1. Define symmetric matrices $\Lambda_{1,N}, \ldots, \Lambda_{N-1,N}$ by

$$[\Lambda_{k,N}]_{m,n} \equiv \begin{cases} 2(\alpha\lambda_k + \lambda_{n+k}) & \text{if } n = m \\ 2\alpha\lambda_k + (1-\alpha)\lambda_{m+k} & \text{if } n > m \end{cases}, \tag{34}$$

$1 \leq m, n \leq N-k$. If the matrix $\Lambda_{1,N}$ is positive semidefinite (positive definite), then so are the matrices $\Lambda_{k,N}$, $2 \leq k \leq N-1$.

Proof. Follows at once from the fact that $x^T \Lambda_{k,N} x = y^T \Lambda_{k+1,N} y$, where $\sum_{j=1}^{N-k} x_j = 0$ and $y = (x_2, x_3, \ldots, x_{N-k})$. □

LEMMA 2. The price process (2) is manipulation-free if and only if $\Lambda_{1,N}$ is positive semidefinite.

Proof. If $\sum_{j=n}^{N} q_j = 0$, $q_j \in \mathbf{M}(H_j)$, then $\mathbb{E}_n[\sum_{j=n}^{N} p_j q_j] = \mathbb{E}[q^T \Lambda_{n,N} q]/2$, where $q = (q_{n+1}, q_{n+2}, \ldots, q_N)$. Applying Lemma 1 proves this Lemma. □

THEOREM 2. Suppose one of the following conditions holds:
(i) $R > 0$ (trader is risk-averse) and the price process (2) is manipulation-free, or
(ii) $R = 0$ (trader is risk-neutral) and the matrix $\Lambda_{1,N}$ is positive definite.
Then, the liquidity trader's problem (3) has a unique, time-consistent solution given by (7)–(10).

Proof. Each condition (i) or (ii) implies the second-order condition, $2\mu_n + R(\alpha^2\lambda_{n-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2) > 0$ for $2 \leq n \leq N$, as stated in Theorem 1. To see this, let us again use backward induction. In period $N-1$, the cost function, $\mathbb{E}_{N-1}[p_{N-1}q_{N-1} + L_N] + \frac{R}{2}\mathrm{Var}_{N-1}[p_{N-1}q_{N-1} + L_N]$, equals

$$\left\{\tilde{p}_{N-2} + \left[\lambda_{N-1} + \frac{R}{2}(\lambda_{N-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2)\right] Q_{N-1}\right\} Q_{N-1}$$

$$+\lambda_{N-1}(1 + \alpha + R\alpha\lambda_{N-1}\sigma_\eta^2) Q_{N-1}(q_{N-1} - Q_{N-1})$$

$$+\frac{1}{2}(\Lambda_{N-1,N} + R\Upsilon_{N-1,N})(q_{N-1} - Q_{N-1})^2,$$

where $\Upsilon_{N-1,N} = (\alpha^2\lambda_{N-1}^2 + \lambda_N^2)\sigma_\eta^2 + \sigma_\varepsilon^2 > 0$. This expression is strictly convex in q_{N-1} if either condition (i) or (ii) hold (by Lemmas 1–2), in which case $2\mu_N + R(\alpha^2\lambda_{N-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2) > 0$.

As induction hypothesis, suppose that $2\mu_k + R(\alpha^2\lambda_{k-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2) > 0$ for $n+1 \leq k \leq N$. It can be shown that the cost function at time $n-1$, $\mathbb{E}_{n-1}[p_{n-1}q_{n-1} + L_n] + \frac{R}{2}\mathrm{Var}_{n-1}[p_{n-1}q_{n-1} + L_n]$, amounts to

$$\begin{aligned}&\left\{\tilde{p}_{n-2} + \left[\lambda_{n-1} + \frac{R}{2}(\lambda_{n-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2)\right] Q_{n-1}\right\} Q_{n-1}\\ &-\lambda_{n-1}(1 + \alpha + R\alpha\lambda_{n-1}\sigma_\eta^2)Q_{n-1}1_{N-n+1}^T x_{n-1}(q_{n-1}) \\ &+\frac{1}{2}x_{n-1}(q_{n-1})^T(\Lambda_{n-1,N} + R\Upsilon_{n-1,N})x_{n-1}(q_{n-1}),\end{aligned} \tag{35}$$

where $x_{n-1}(q_{n-1}) = (q_n(q_{n-1}), q_{n+1}(q_{n-1}), \ldots, q_N(q_{n-1}))$, each component being an affine function of q_{n-1} and $\sum_{j=n-1}^N q_j = Q_{n-1}$, $\Upsilon_{n-1,N}$ is a positive definite matrix, and 1_k is the k-dimensional vector containing only ones. If either condition (i) or (ii) holds, then the function in (35) is strictly convex in q_{n-1}, thanks to Lemmas 1–2. Hence $2\mu_n + R(\alpha^2\lambda_{n-1}^2\sigma_\eta^2 + \sigma_\varepsilon^2) > 0$ and we are done. □

Proof of Proposition 2. Since $\Lambda_{1,N}$ is positive definite, a time-consistent solution exists and is unique by Theorem 2. Writing out the first-order conditions of (3) for the price process (1) yields

$$\lambda q_{n+2} - (2\lambda + R\sigma^2)q_{n+1} + \lambda q_n = 0 \quad \text{for } 1 \leq n \leq N-3 \tag{36}$$

$$(\lambda^2 + 3\lambda R\sigma^2 + R^2\sigma^4)q_{N-1} - \lambda(\lambda + R\sigma^2)q_{N-2} = 0 \tag{37}$$

$$(3\lambda + R\sigma^2)q_{N-1} + \lambda\sum_{n=1}^{N-3} q_n = \lambda Q, \tag{38}$$

where $\sigma^2 \equiv \lambda^2\sigma_\eta^2 + \sigma_\varepsilon^2$.

Solving the difference Equation (36) subject to the boundary conditions (37) and (38) gives (16) and (17).

The proof that the optimal trades are positive and strictly decreasing if $R > 0$ can be easily verified by looking directly at the formulas in (16) and (17). □

LEMMA 3. If either condition (i) or (ii) in Theorem 2 is satisfied, then the liquidity trader's problem (3) has a deterministic, time-consistent solution even when a short-sale constraint is imposed.

Proof. In the last period, the trader respects the short-sale constraint by default because $q_N = Q_N$. Then, suppose that $\{q_k = g_k(Q_k)\}_{k=n}^N$ represents a time-consistent solution to the subproblem regarding the periods n to N, where the g_k's are deterministic functions. At time $n-1$, we need to minimize the cost function $L_{n-1} = \mathbb{E}_{n-1}[\sum_{k=n-1}^{N} p_k q_k] + \frac{R}{2}\mathrm{Var}_{n-1}[\sum_{k=n-1}^{N} p_k q_k]$ subject to the constraints $\sum_{k=n-1}^{N} q_k = Q_{n-1}$ and $\{\sum_{m=k}^{N} q_m \leq Q\}_{k=n}^{N}$ (the inequalities imply that the trader never shorts). A solution to this optimization problem exists, because (i) the cost function, being equal to (35), is convex, quadratic, and bounded below, and (ii) the constraints describe a polyhedral convex set. Furthermore, since L_{n-1} can also be written as

$$\mathbb{E}_{n-1}\left[p_{n-1}q_{n-1} + \mathbb{E}_n\left[\sum_{k=n}^{N} p_k q_k\right] + \frac{R}{2}\mathrm{Var}_n\left[\sum_{k=n}^{N} p_k q_k\right]\right]$$
$$+\frac{R}{2}\{\lambda_{n-1}^2\sigma_\eta^2[\alpha q_{n-1} + (1-\alpha)Q_{n-1}]^2 + \sigma_\varepsilon^2(Q_{n-1} - q_{n-1})^2\},$$

we infer that $\{q_k = g_k(Q_k)\}_{k=n-1}^{N}$ is a time-consistent solution to the problem for the periods $n-1$ to N. Continuing in this fashion we arrive at the first period, where we conclude, due to $q_1 = g_1(Q)$, that a deterministic, time-consistent solution exists for the whole problem. □

Proof of Proposition 5. We can apply the Implicit Function Theorem to the first-order condition $\frac{\partial}{\partial N}L(Q, N^*, x) = 0$ to analyze the dependence of N^* on the underlying parameters Q and $x = (\lambda, \alpha, R, \sigma_\eta^2, \sigma_\varepsilon^2)$. As $\partial N^*/\partial x_j = -\frac{\partial^2 L}{\partial N \partial x_j} / \frac{\partial^2 L}{\partial N^2}$ and L is strictly convex in N, the sign of $\partial N^*/\partial x_j$ will be equal to the opposite sign of $\partial^2 L/\partial N \partial x_j$.

By differentiating $\partial L(Q, N^*, x)/\partial N$ with respect to Q, we obtain

$$\frac{\partial^2 L}{\partial N \partial Q} = \begin{cases} -\dfrac{\lambda(1+\alpha)}{N^2}Q < 0 & \text{if } R = 0 \\ -2\tilde{b}\ln r \dfrac{(r^2-1)r^{2N-1}}{(r^{2N}-1)^2}Q < 0 & \text{if } R > 0 \end{cases}, \tag{39}$$

where $\tilde{b} = \sqrt{b}$, b and r are given as in (26), implying that $\partial N^*/\partial Q > 0$. To verify the individual signs of $\partial N^*/\partial x$, we first calculate the second-order cross-partial derivatives for the case $R = 0$ and then for $R > 0$. If $R = 0$, then $\partial^2 L/\partial N \partial \alpha = -\lambda Q^2/(2N^2) < 0$ and $\partial^2 L/\partial N \partial \lambda = -(1+\alpha)Q^2/(2N^2) < 0$, having $\partial N^*/\partial \alpha > 0$ and $\partial N^*/\partial \lambda > 0$ as a consequence. If $R > 0$,

$$\begin{aligned} \frac{\partial^2 L}{\partial N \partial x_j} &= -k'(N^*)\left(\frac{1}{\tilde{b}}\frac{\partial \tilde{b}}{\partial x_j} + W\frac{\partial r}{\partial x_j}\right), \\ W &= \frac{r - r^{-1} + 2r\ln r}{(r^2-1)\ln r} - \frac{(2N+1)r^{2N} + 2N - 1}{r(r^{2N}-1)} < 0, \end{aligned} \tag{40}$$

and thus we have to evaluate the derivatives $\partial\tilde{b}/\partial x_j$ and $\partial r/\partial x_j$ for each case. The sign of $\partial N^*/\partial \alpha$ is positive because $\partial \tilde{b}/\partial \alpha > 0$ and $\partial r/\partial \alpha < 0$. Regarding the permanent price-impact parameter, $\partial r/\partial\lambda$ is negative when $\lambda^2\sigma_\eta^2$ is sufficiently small relative to σ_ε^2, whereas $\partial\tilde{b}/\partial\lambda$ is always positive. Hence $\partial N^*/\partial\lambda > 0$ under the maintained assumption. The remaining claims of this proposition can be proved in exact the same fashion. □

References

Almgren, R. and Chriss, N. (2000) Optimal execution of portfolio transactions, *Journal of Risk* **3**, 5–39.

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

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

Chan, K., Chung, Y. P., and Johnson, H. (1995) The intraday behavior of bid-ask spreads for NYSE stocks and CBOE options, *Journal of Financial and Quantitative Analysis* **30**(3), 329–346.

Chan, L. K. and Lakonishok, J. (1995) The Behavior of Stock Prices around Institutional Trades. *Journal of Finance* **50**(4), 1147–1174.

Choi, J. Y., Salandro, D., and Shastri, K. (1988) On the estimation of bid-ask spreads: Theory and evidence, *Journal of Financial and Quantitative Analysis* **23**(2), 219–230.

Chordia, T., Roll, R., and Subrahmanyam, A. (2001) Market liquidity and trading activity, *Journal of Finance* **56**(2), 501–530.

Conrad, J., Johnson, K., and Wahal, S. (2001) Institutional trading and soft dollars, *Journal of Finance* **56**, 397–416.

Coppejans, M., Domowitz, I., and Madhavan, A. (2003) Dynamics of liquidity in an electronic limit order book market, Unpublished working paper, Duke University.

Degryse, H., de Jong, F., van Ravenswaaij, M., and Wuyts, G. (2004) Aggressive orders and the resiliency of a limit order market, Unpublished working paper, Leuven University.

Foucault, T., Kadan, O., and Kandel, E. (2005) Limit order book as a market for liquidity, forthcoming in the *Review of Financial Studies*.

Hausman, J. A., Lo, A. W., and MacKinlay, A. C. (1992) An ordered probit analysis of transaction stock prices, *Journal of Financial Economics* **31**, 319–379.

Huberman, G., Jones, C. M., and Lipson, M. L. (2004) The dynamics of institutional trading, Unpublished working paper in preparation, Columbia University.

Huberman, G. and Stanzl, W. (2004) Price manipulation and quasi-arbitrage, *Econometrica* **72**(4), 1247–1275.

ITG Inc. (2004) ITG ACE$^{\text{TM}}$ document, http://www.itginc.com.

Jones, C. M. and Lipson, M. L. (1999) Execution costs of institutional equity orders, *Journal of Financial Intermediation* **8**, 123–140.

Jones, C. M. and Lipson, M. L. (2001) Sixteenths: Direct evidence on institutional execution costs, *Journal of Financial Economics* **59**, 253–278.

Keim, D. B. and 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. B. and Madhavan, A. (1996) The upstairs market for large block transactions: Analysis and measurement of price effects, *Review of Financial Studies* **9**, 1–36.

Keim, D. B. and 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. and Korn, O. (1999) Market depth and order size, *Journal of Financial Markets* **2**(1), 29–48.

Kyle, A. S. (1985) Continuous auctions and insider trading, *Econometrica* **53**(6), 1315–1336.

Madhavan, A., Richardson, M., and Roomans, M. (1997) Why do security prices change? A transaction-level analysis of NYSE stocks, *Review of Financial Studies* **10**(4), 1035–1064.

Vayanos, D. (2001) Strategic trading in a dynamic noisy narket, *Journal of Finance* **56**(1), 131–171.