Tactical Liquidity Trading and Intraday Volume

Merrell Hora*
Credit Suisse Advanced Execution Services
merrell.hora@credit-suisse.com

PRELIMINARY
COMMENTS WELCOME

September 20, 2006

Abstract

This paper analyzes liquidity trading within the context of the institutional execution process. The problem facing a risk averse, transaction cost minimizing trader who must execute a relatively large trade over a finite horizon is cast as a simple dynamic program. It is shown that the conventional specification of trader preferences is inconsistent with the objective of minimizing costs and results in solutions that are counter-intuitive and easily dominated. A new, time separable cost function is introduced, resulting in a linear-quadratic problem. A complete characterization of the solution and optimal execution policy, including comparative statics, is provided. It is shown that even when price changes and exogenous order flow are martingales, it is optimal for the liquidity trader to engage in tactical trading by altering the rate of execution in response to exogenous changes in prices. For sufficiently large trades, the expected execution path has a non-monotonic, U-shaped profile. This provides a simple and complete explanation for the empirical regularity of the U-shaped pattern of average intraday volume and (unconditional) price volatility. The solution for the portfolio generalization is provided, indicating that this analysis can support portfolio trading and directly incorporate constraints often encountered in practice.

*The views expressed herein are strictly those of the author and are not necessarily those of Credit Suisse Group.

1 Introduction

While the significance of the higher frequency aspects of the investment process cannot be underestimated from a practical and regulatory perspective, the rigorous analysis of these problems has evolved at a disproportionately slow rate. This has been partially rectified by the rapid expansion of the market microstructure literature. However, this literature is generally focused upon the role of asymmetric information in the trading process, addressing other trading activity in the form of exogenous factors that are necessary to prevent market breakdowns, or the no-trade outcomes as in Milgrom and Stokey (1982). Thus, trading that is not motivated by asymmetric information, typically called liquidity trading, is often a secondary concern. While asymmetric information is a fundamental problem of great significance, when specified in a realistic manner, liquidity trading is exposed as a non-trivial problem that is equally fundamental. Given its universal nature and the monetary value of associated transactions, it is easy to argue that liquidity trading is of even greater economic significance. This paper provides evidence supporting this claim by analyzing liquidity trading in a new and generalized manner. By treating liquidity trading as a dynamic optimization problem with realistic trader preferences, the resulting optimal execution policy is shown to be novel, consistent with observed trader behavior, and relevant for actual trading practice. In addition, the nontrivial dynamic properties of this policy provide a new explanation for the well documented patterns of intraday volume in equity markets.[1]

The liquidity trading, or execution, problem corresponds to the problem facing a trader who must execute a relatively large, exogenously specified quantity Q within a fixed time interval of length T, under the objective of minimizing expected transaction costs. These costs include fees, commissions, spreads, and the price impact that is caused by the consumption of liquidity. There is ample empirical evidence dating from Kraus and Stoll (1972) to Chan and Lakonishok (1995) and Chan and Lakonishok (1997), indicating that price impact is the dominant component of transaction costs. Clearly, as markets become increasingly automated, fixed costs are driven down, and this effect is more pronounced. Thus, there is no loss in generality in equating transaction costs with impact. This leads to a definition of cost as "slippage" or implementation shortfall, which is the difference between share weighted average execution price and the arrival price of the order.

Since prices respond to the trader's order flow and exogenous, stochastic shocks, costs are also stochastic. Given the nature of the institutional execution process, it is appropriate to assume that traders are risk averse. This article introduces a new trader cost (disutility) function that supports risk aversion in a simple and coherent manner, and also results in time separable preferences for the dynamic problem. This, in turn, allows the problem to be specified as a simple, finite horizon stochastic dynamic program. A complete characterization

[1] Liquidity trading has been treated as the result of an optimization problem elsewhere, but typically in a context that is not representative of the institutional trading process. See, for example, Bhushan (1991) and Mendelson and Tunca (2004).

of the optimal execution policy is provided, indicating that it is optimal to engage in tactical liquidity trading by responding to exogenous changes in price. This means that the liquidity trader should respond to favorable price shocks by increasing the rate of execution, and decreasing the rate for unfavorable shocks. It is important to note that the optimality of price responsiveness results without assuming that prices are mean reverting or have non-zero drift.

To allow for price responsiveness, the optimal policy utilizes the entire time horizon to complete the trade. Initially, the trades are of moderate size, decreasing as the trader's own (adverse) impact on price is large enough to equate expected costs with the potential benefits of favorable price shocks. The intermediate periods of the execution path are, on average, characterized by steady-state behavior where small, positive trades are made, but depending upon the exogenous shocks, trade size can either increase or decrease. As the terminal period approaches, the rate of execution increases to ensure that the trade is completed on time. Thus, on average, the execution profile is U-shaped, and in this sense, the solution provides a new and very simple explanation for the observed U-shaped volume profiles.[2]

The basic dynamic program easily generalizes to the portfolio case while preserving the simple structure of the single security problem. This is critical for applications given that portfolio trading constitutes a significant percentage of daily volume in U.S. equity markets. It is shown how portfolio constraints encountered in practice can be incorporated into the model, and given that the computational simplicity is preserved, this indicates great potential for applications.

Liquidity trading has been analyzed as a dynamic optimization problem before, although in each case the resulting execution policy exhibits features that are incongruent with actual trading practice and economic intuition. Bertsimas and Lo (1998) model the problem as a simple cost (impact) minimization problem, but do not allow for trader risk aversion. Bertsimas, Hummel, and Lo (1999) provide the extension to portfolios under the same assumptions. Almgren and Chriss (2000) allow for trader risk aversion by specifying mean-variance preferences, although the subsequent analysis is not dynamic and the optimization is performed in a static manner prior to the initial period. Huberman and Stanzl (2005) also allow for risk aversion through mean-variance preferences, and attempt to provide a fully dynamic solution, but their solutions are essentially identical to those provided by Almgren and Chriss (2000).[3]

In each of these cases, the optimal execution policy is independent of ex-

[2]This result is complementary to alternative explanations provided by Admati and Pfleiderer (1988) and Admati and Pfleiderer (1989), that are partial in nature and rely upon asymmetric information to generate trading activity.

[3]This equivalence follows from the mis-specification of the recursive form of the problem. With cost function c, state variable s_t, and state transition function f, the functional equation in Huberman and Stanzl (2005) has the form $v_t(s_t) = \min\{E[c(s_t) + v_{t+1}(f(s_t))] + \frac{R}{2}Var[c(s_t) + v_{t+1}(f(s_t))]\}$. The suggested solution indicates value functions that are quadratic in s_t. Given the linear state transitions and Gaussian innovations, if v_{t+1} is quadratic, then the variance of v_{t+1} results in fourth order terms. Thus, the right-hand side of their functional equation is not quadratic in the state variable, resulting in a contradiction.

ogenous changes in execution price. Thus, for a large buy order, even if the price drops precipitously for reasons unrelated to the liquidity trade, there is no change in the planned execution path. This result is counterintuitive and problematic, as the objective is to minimize costs but it indicates an indifference to cost. Furthermore, under risk aversion, the policies are characterized by extremely large initial trade sizes that diminish at an exponential rate, effectively completing the entire trade well before the allotted time. Both features are in sharp contrast to the behavior of actual institutional traders and institutional trading algorithms, as well as observed patterns in intraday volume. In fact, the numerical examples presented here indicate that these price independent solutions are not optimal and are stochastically dominated by the new solutions provided in this article.

A somewhat different approach is pursued in Obizhaeva and Wang (2005), where a limit order market is used, allowing for more realistic liquidity dynamics. In addition, in the continuous time limit, the trader is allowed to not trade at a given instant. This feature also arises in the optimal policy provided in this paper, although in this case it results in discrete time as a response to adverse price shocks. The optimal policy from Obizhaeva and Wang (2005) exhibits large, discrete trades at the initial and terminal times, and execution at a constant rate otherwise. However, this policy is also independent of price change and subject to similar limitations indicated in the previous paragraph.

The remainder of the paper is organized as follows. The next section presents the dynamic representation of the general liquidity trading problem. Section 3 discusses trader preferences, arriving at the new cost function. The optimal policy (solution) under this function is given in section 4 along with results characterizing the optimal execution policy. In section 5 a sensitivity analysis is performed and numerical examples compare the behavior and performance of the optimal policy with the typical solution obtained elsewhere. Section 6 provides the generalization to portfolios, followed by the conclusion. All of the proofs and secondary results are contained in the appendix.

2 The General Liquidity Trading Problem

For simplicity, attention is restricted to the buyer's problem, as the seller's case is completely analogous.

The problem is based upon a single asset that is traded over a sequence of markets indexed by t. Each market consists of our primary subject (the trader), a population of exogenous noise traders, and a competitive market maker. The sequence of events for each market initiates with an indicative price quote $\tilde{p}_t$ from the market maker, who is assumed to be perfectly competitive in the sense that prices are generated according to the appropriate zero profit condition. Independently, the exogenous traders submit market orders that are aggregated into their net order flow ω_t, where $\{\omega_t\}$ is an *i.i.d.* stochastic process with mean 0 and variance σ_ω^2. The trader cannot observe ω_t, but does know the value of σ_ω^2. Concurrent with the generation of ω_t, the trader selects the amount q_t

to buy in period t as a market order, where $q_t = 0$ indicates that the trader chooses not to transact. To be consistent with market regulations, short sales, or negative values for q_t, are not permitted. Also, with x_t defined to be the number of shares remaining to be bought at the beginning of period t, it must be the case that $q_t \leq x_t$.

Price dynamics follow the specification given in Huberman and Stanzl (2005). First, given the price quote determined at the end of the prior period, $\tilde{p}_t$, the market maker observes the current period net order flow $\omega_t + q_t$, and then sets the transaction price p_t in accordance with the zero (market maker) profit conditions. This response is represented by a temporary price impact function f taking the initial quote and net order flow as arguments. Thus, the transaction price is determined in the following manner

$$p_t = f(\tilde{p}_t, \omega_t + q_t). \tag{1}$$

The market is assumed to be sufficiently liquid so that additional order size constraints or boundary restrictions on f are not required. In most applications, f is simply the sum of $\tilde{p}_t$ and a linear order flow term. For now it suffices to require that f is increasing in both arguments and continuously differentiable, but a precise specification is provided in section 4.

Once p_t has been determined and the current period transactions are completed, the initial quote for the next period is selected. To allow for exogenous shocks including informational events this quote is subject to an *i.i.d.* disturbance term ϵ_t with mean 0 and variance σ_ϵ^2. In addition, to allow order flow to have a permanent price impact $\tilde{p}_{t+1}$ also depends upon a weighted average of $\tilde{p}_t$ and p_t, where $\alpha \in [0, 1]$ indicates the relative weight. Thus, the next period's quoted price is

$$\tilde{p}_{t+1} = \alpha \tilde{p}_t + (1 - \alpha) p_t + \epsilon_{t+1}. \tag{2}$$

Note that the price impact of order flow is permanent if $\alpha = 0$, and temporary if $\alpha = 1$.

This price formation process clearly indicates how the trader's transaction costs are determined as a function of order flow. As indicated in the introduction, the transaction cost for the entire trade is measured as the weighted average execution price relative to the arrival price of the order, also known as slippage. Thus, the component transaction cost is driven by $\Delta_t = p_t - \tilde{p}_0$, the difference between the period t transaction price and the initial quote. With Q as the target position, the period t component transaction cost is

$$\frac{q_t}{Q} \Delta_t.$$

Given the uncertainty in prices these costs are random, and it is appropriate to incorporate risk aversion into the trader's problem.

To reflect the trader's attitude towards execution risk, additional non-monetary costs are incorporated into the problem through the use of a cost function c. Throughout, c is assumed to be a continuously differentiable, convex function of period t transaction costs and the period t state vector which is defined below.

Convexity and the additional argument reflect the fact that the cost function accounts for additional factors beyond simple transaction costs. From the perspective of the trader, c is a function of q_t and x_t, treating any parameters and prices as given. Since Δ_t can be positive or negative, it is also possible for the monetary costs to be positive or negative even with q_t constrained to be non-negative. Thus, c must be defined over $[-\tilde{p}_0, \infty)$, and not just $\Re_+$. For the remainder of the paper, the term *cost* is used to indicate the comprehensive, utility based costs incurred by the trader, including both monetary and non-monetary components.

With these elements in place, employing a discount factor $\beta \in (0, 1]$, and using C to denote expected total costs, the sequential representation of the trader's problem is as follows

$$\min_{\{q_t\}_{t=0}^T} C(q_0, q_1, \ldots, q_T) = E\left[\sum_{t=0}^{T} \beta^t c(q_t, x_t)\right]$$

subject to

$$0 \leq q_t \leq x_t$$
$$x_{t+1} = x_t - q_t$$
$$x_0 = Q, x_{T+1} = 0$$

where prices are implicit in c and follow equations (1) and (2).

Even though this problem is a well defined, finite horizon convex minimization problem, for computational purposes it is convenient to consider the corresponding recursive representation. In this case, let the state vector be given by $s_t = [x_t, \tilde{p}_t]'$, and let $\Gamma(s_t)$ denote the set of feasible current period quantities, where $\Gamma(s_t)$ is defined by the constraints given above. Then, given the assumptions regarding c (continuously differentiable, convex), f (strictly increasing), Γ, and the stochastic innovations, there exists a unique sequence of functions $v^* = \{v_t^*\}_{t=0}^T$ that satisfy the following functional (Bellman) equation

$$v_t(s_t) = \min_{q_t \in \Gamma(s_t)} E\big\{c(q_t, x_t) + \beta v_{t+1}(s_{t+1})\big\}. \tag{3}$$

Furthermore, under v^* the minimizing value of the right hand side of (3) corresponds to a unique sequence of optimal policy functions $g^* = \{g_t^*\}_{t=0}^T$. It follows that the quantities generated by g^* also attain the minimum for original sequential problem stated above. In this sense, the solution v^* is equivalent to the solution to the sequential problem.[4]

This is a standard finite horizon dynamic optimization problem. The analysis should be straightforward, and well within the reach of standard tools. However, when allowing for risk aversion, previous research has relied upon trader preferences that are time inseparable, thereby precluding the equivalence between the sequential problem and the dynamic program described by equation (3). The

[4] The value and policy functions, v_t and g_t, are indexed by period to emphasize the fact that with a finite time horizon it is not possible to focus on steady-state solutions, and thus, time-varying parameter values must be supported.

following section examines trader preferences in detail, providing a simple, time separable specification that supports the valid use of the recursive approach.

3 Trader Preferences

Two important facts relating to the specification of the trader's problem are that (1) traders can and (2) often do change strategies at any point in time. Typically, strategies are revised, replaced or just canceled when conditions change or when execution is not proceeding as expected. Obviously the decision problem is dynamic and must allow for strategies that are responsive to changes in market conditions. In the present context, this requires preferences that can be clearly defined for any individual component of costs. Since the transaction cost for the entire trade is simply the sum of weighted periodic costs, time separable preferences are the most simple way to satisfy this objective.[5] In addition, it is difficult to imagine how time inseparability could be consistent with the behavior of a rational, cost-minimizing trader.

This point is relevant because, as indicated above, when allowing for risk aversion previous research has employed time inseparable preferences. In these cases, the trader's problem is specified as one of minimizing the sum of the expectation and variance of *total* transaction costs

$$C(q_0, q_1, \ldots, q_T) = E\left[\sum_{t=0}^{T} p_t q_t\right] + \frac{\lambda}{2} Var\left[\sum_{t=0}^{T} p_t q_t\right], \tag{4}$$

where λ is a risk aversion parameter. This specification was introduced by Almgren and Chriss (2000), motivated by their approach of treating the trader's problem as analogous to the standard *static* portfolio optimization problem. Of course, this approach cannot account for any non-trivial dynamic behavior, and it also creates problems that are analogous to those associated with risk aversion and intertemporal substitution in dynamic portfolio problems addressed in Giovannini and Weil (1989) and Kocherlakota (1990). In particular, these preferences inadvertently represent risk aversion and attitudes towards the temporal resolution of uncertainty (Kreps and Porteus 1978) with the same term, creating a clear preference for rapid execution.

To see this, it suffices to derive the covariance between any two distinct cost terms under the assumption of non-stochastic trade sizes. Assuming a linear price impact function f with impact parameter γ as in the following section, the period t transaction price can be written as

$$p_t = \tilde{p}_0 + (1-\alpha)\gamma \sum_{i=1}^{t-1} (\omega_i + q_i) + \gamma(\omega_t + q_t) + \sum_{i=1}^{t} \epsilon_i.$$

[5] Recursive preferences are also consistent with this behavior. However, since the time separable case results in a very simple model and yields interesting results, it is not clear that the added complexity is necessary.

It follows that the simplified unconditional covariance between the period i cost and period j cost, where $i < j$, can be expressed as

$$cov(p_i q_i, p_j q_j) = q_i q_j \big[((i-1)(1-\alpha)^2 + 1)\gamma^2 \sigma_\omega^2 + (i-1)\sigma_\epsilon^2)\big].$$

Since this is increasing in i, expressing risk aversion in terms of the variance of total costs, or any sum of future costs, implicitly assumes a (potentially strong) preference towards rapid execution. Allowing for stochastic quantities only increases the covariance. This effect increases with the number of periods increases, generating solutions that execute in an extremely rapid manner as demonstrated in section 5. This indicates a willingness to realize larger transaction costs in order to complete the trade as soon as possible.

Utilizing preferences that express such a willingness is not a problem, and in fact, captures important aspects of trader behavior. The problem is that by doing this implicitly, it is usually not possible to determine the extent to which a given factor influences the results. This in turn complicates any subsequent conclusions and trading policy implications. For the preferences given above, λ indicates both the sensitivity to variance and time, and any conclusion identifying risk aversion as a factor could also be interpreted as deriving from a preference for earlier resolution of uncertainty.

To avoid these problems and yet still support interesting preferences regarding temporal matters, holding costs are incorporated into the problem with the introduction of a parameter $\rho \geq 0$. Then, time separable preferences are specified through the following stage cost function

$$c(q_t, x_t) = \Delta_t \frac{q_t}{Q} + \frac{\lambda}{2}\left[\Delta_t \frac{q_t}{Q} - E\big(\Delta_t \frac{q_t}{Q}\big)\right]^2 + \rho x_t^2 \tag{5}$$

where E indicates the expected value with respect to the appropriate conditional distribution. In terms of expected utility, the first two terms correspond to simple mean-variance preferences for period t costs. Thus, risk aversion is treated as variance aversion for each component of total costs. This approach to risk aversion is preferred because it clearly articulates the trader's preferences using just the choice (control) variable and state vector. To be consistent with a well behaved preference ordering, the cost function (4) requires quantities to be selected as entire vectors, precluding a sequential or recursive approach.

The primary innovation in (5) is the inclusion of the term ρx_t^2. This term represents holding costs in utility terms, where ρ is the degree of impatience regarding the rate of completion. Thus, risk aversion and attitudes regarding temporal resolution are clearly segmented in this cost function.

It is also important to note that time separability allows for discounting, which explicitly expresses preferences regarding the timing of non-random payoffs. Even though discounting is standard in dynamic decision problems, it has yet to be used in the analysis of the liquidity trader's problem. The following section indicates that this has a material impact upon the qualitative features of the results.

4 The Solution

In this section the general problem specified in section 2 is addressed using the preferences defined by equation (5) in the preceding section, and a simple linear specification of the market impact function f. This allows for a complete analytical derivation of the optimal execution policy and the associated dynamics. These results provide for a clear description of the nature of price dependence in the optimal execution policy and an explanation for the intra-day pattern of volume.

As is standard in the literature, it is assumed that the price impact of order flow has the following linear structure

$$p_t = f(\tilde{p}_t, \omega_t + q_t) = \tilde{p}_t + \gamma(\omega_t + q_t) \tag{6}$$

where $\gamma > 0$. Although the trader's strategies are constrained to be either strictly buy or sell orders, thereby eliminating the ability to manipulate prices by trading both sides of the market, linearity can be justified by Huberman and Stanzl (2004). Using an environment similar to the one analyzed here, Huberman and Stanzl (2004) demonstrates that price manipulation can only be ruled out if the permanent impact function is linear. In addition, since price updates are driven by order flow in discrete time, the function f can be interpreted as the average response to order flow over short, non-degenerate intervals. In this sense, linearity is a reasonable compromise between the mixed empirical evidence regarding the behavior of price impact functions.[6]

Define $\delta_t = \tilde{p}_t - \tilde{p}_0$, the difference between the initial quotes in period t and period 0. With f as in (6) and given the assumptions regarding $\{\epsilon_t\}$ and $\{\omega_t\}$, it follows that $E(\Delta_t|q_t, \tilde{p}_t) = \delta_t + \gamma q_t$, and $Var(\Delta_t|q_t, \tilde{p}_t) = \gamma^2\sigma_\omega^2$. Thus, with preferences as in (5), the stage expected cost can be computed and the functional equation (3) becomes

$$v_t(s_t) = \min_{q_t \in \Gamma(s_t)} \left\{ \left[\frac{\gamma}{Q} + \frac{\lambda\gamma^2\sigma_\omega^2}{2Q^2} \right] q_t^2 + \frac{\delta_t}{Q} q_t + \rho x_t^2 + \beta E(v_{t+1}(s_{t+1})) \right\}. \tag{7}$$

With the boundary conditions $x_0 = Q$ and $x_{T+1} = 0$, and the state transition equations, this is a complete specification of the liquidity trader's problem. Since the set of feasible controls, Γ, is convex for all s_t, two additional restrictions on the parameters ensure that this problem is a well behaved finite horizon linear-quadratic problem. Thus, it is a simple matter to derive the sequence of optimal policy functions, which are necessarily linear in the state variables (see Ljungqvist and Sargent (2000)).

To ensure that the problem is well behaved, the following parameter restrictions are required. To simplify the notation, collect the terms involving q_t^2 in (7) and define

$$\theta = \gamma/Q + \lambda\gamma^2\sigma_\omega^2/Q^2.$$

[6] Hora (2005) provides evidence that as the trade size and trade duration increase, the price response is weakly concave. That analysis examines the impact of orders executed using trading algorithms, which are special cases of the solutions considered here.

The parameters defining the terminal costs must correspond to a convex function, or

$$2(\theta + \rho) + \Delta_T/Q > 0. \quad (8)$$

Given that $\theta > 0$ and $\rho > 0$, (recall that Δ_T can be negative) this is clearly not restrictive since the focus is on large Q. In addition, the following restriction on holding costs is required to avoid trivial cases where these costs dominate all other factors in a deterministic manner

$$\rho < \theta\left(\frac{1-\beta}{\beta}\right). \quad (9)$$

To facilitate the derivation of the solution, the problem is recast in the standard format of an optimal linear regulator problem. First, note that the period t expectation of the cost function (5) can be rewritten as

$$c(q_t, s_t) = \begin{bmatrix} s_t' & q_t \end{bmatrix} \begin{bmatrix} R & W \\ W' & \theta \end{bmatrix} \begin{bmatrix} s_t \\ q_t \end{bmatrix},$$

where $R = \begin{bmatrix} \rho & 0 \\ 0 & 0 \end{bmatrix}$, and $W = \begin{bmatrix} 0 \\ 0.5Q^{-1} \end{bmatrix}$. The state transitions are

$$s_{t+1} = s_t + Bq_t + z_{t+1} \quad (10)$$

where $B = \begin{bmatrix} -1 \\ \eta \end{bmatrix}$, $\eta = (1-\alpha)\gamma$, and $z_{t+1} = \begin{bmatrix} 0 \\ \epsilon_{t+1} + (1-\alpha)\gamma\omega_t \end{bmatrix}$.

The following result is immediate. (The proofs for all results are presented in the appendix.)

Proposition 1. *The functions $\{v_t\}$ satisfying the functional equation (7) have the form $v_t(s) = s'V_t s + d_t$, where $\{V_t\}$ are matrices satisfying the following matrix Riccati difference equation*

$$V_t = R + \beta V_{t+1} - (\beta B'V_{t+1} + W')'(\theta + \beta B'V_{t+1}B)^{-1}(\beta B'V_{t+1} + W'), \quad (11)$$

$V_T = \begin{bmatrix} \theta + \rho & 0.5Q^{-1} \\ 0.5Q^{-1} & 0 \end{bmatrix}$, *and* $d_t = \beta(\sigma_\epsilon^2 + \eta^2\sigma_\omega^2)(v_{t+1,12} + v_{t+1,22} + d_{t+1})$. *Furthermore, the optimal policy functions are defined by the following vectors*

$$g_t = -(\theta + \beta B'V_{t+1}B)^{-1}(\beta B'V_{t+1} + W'), \quad (12)$$

where the optimal trade size is given by $q_t^* = max\{0, g_t s_t\}$.

The maximum operator results from the constrained optimization and is required because the exogenous price shocks are generally unconstrained, potentially creating an incentive for the trader to engage in selling. For the remainder of this article, g_t and g will be used to indicate both the policy function, inclusive of the maximum operator, and the vector (12).[7]

[7] Since the certainty equivalence principle applies, the interpretation is clear from the context. It is necessary to consider stochastic behavior of prices only when addressing particular price paths.

A unique feature of the optimal policy obtained here is that it preserves the trader's incentive to minimize costs by responding to favorable price movements. This property is established by the following theorem which provides some key comparative static results for the liquidity trading problem.

Theorem 1 (Comparative statics of the policy functions)**.** *The optimal buying policy is increasing in trade size, and if $\beta < 1$, then the optimal buying policy is decreasing in price. If $\beta = 1$ (no discounting), then the optimal policies are independent of price.*

Thus, an increase in Q, x_t, or a decrease in $\tilde{p}_t$ results in an increase in q_t^*, and increases in the quoted price result in a reduction of the rate of execution. It is important to note that this behavior is completely driven by the trader's incentive to minimize transaction costs, and not by any assumptions regarding price dynamics, such as mean-reversion or the potential for non-zero drift. Both mean-reversion and non-zero drift are commonly viewed as dominant factors driving trader or trading algorithm behavior. This result indicates that this need not be the case, and that price responsiveness is completely consistent with cost minimization, which is fortunate given that neither mean-reversion nor non-zero drift are consistent with the absence of arbitrage or the empirical evidence.[8]

The exception to the price responsive case is when there is no discounting. In this case, changes in $\tilde{p}_t$ have no effect upon q_t^*. This result is due to the linear transition function $x_{t+1} = x_t - q_t$, and the fact that prices enter into the cost function linearly. When $\beta = 1$, these two terms are offsetting, eliminating prices from the execution problem. This means that even though the cost function explicitly indicates an incentive to increase trade size as the price quote decreases for the stage problem, in the dynamic representation there is no such incentive. So if there is an exogenous, negative shock to prices, thereby resulting in a lower price quote, the trader will not alter the execution path. Essentially, the problem is degenerate in the sense that it is reduced to one of price impact and variance minimization from the initial specification of cost minimization. Since discounting has not been considered before, (i.e. the case of $\beta < 1$ has not been analyzed before) price independence is common to the policies provided elsewhere in the literature.[9]

Since the trader's problem is necessarily finite horizon, any analysis of the dynamics of the optimal execution path is complicated by the fact that it is necessary to consider the full path to convergence (of the policy functions), as well as the asymptotic properties of the problem. Given the martingale specification on all exogenous shocks, the dynamics can be described in terms of the expected execution path. This is the path that results from holding

[8] Even if one could detect mean-reversion or non-zero drift in real time, which would be required for actual trading, it is not likely that such a skill would be employed strictly for execution purposes.

[9] It also follows that the justification for price independence provided elsewhere, that the current price only represents known information and sunk costs and is therefore decision irrelevant, is incorrect.

all innovations at their conditional mean of 0, and thus, changes are price are completely due to the quantities executed by the trader.

In order to examine the behavior of the execution path over the initial periods, it is helpful to consider convergence of the policy functions. Since convergence is not necessary for optimality or required by applications, it is of interest mostly as a matter of convenience. Given this, and the focus on the behavior of the *expected* execution path, the criterion for convergence of the policy functions employed here refers to the equivalence of policy recommendations given the same state vector. To reflect the expected behavior, define the subset, S_e, of feasible state vectors by $s_1 \in [0, Q]$ and $s_2 \in [0, \eta Q]$. The policy functions $\{g_t\}$ are said to *converge relative to the minimum trade size* q_{min} if there exists some $t^* \in (0, T]$ such that for any $t_1, t_2 \in (0, t^*)$, and any $s \in S_e$, $|g'_{t_1} s - g'_{t_2} s| < q_{min}$. Note that due to the finite horizon, convergence refers to t decreasing to 0.

The potential convergence of $\{V_t\}$ and $\{g_t\}$ is guaranteed by lemmas 3 and 4 in the appendix. With a finite time horizon, the problem is that there may not be enough periods to obtain the limit. The following lemma establishes conditions that guarantee the limit is reached within the given time horizon, and also provides a range for the limit of a linear transformation of V_t. This requires the following restriction on β, relative to T and Q,

$$\beta^{T-1}(1-\beta) < \frac{q_{min}}{2Q}. \tag{13}$$

Lemma 1 (Convergence of $\{V_t\}$ and $\{g_t\}$)**.** *If, conditions (8) and (9) are satisfied, and* β*,* T *and* Q *also satisfy (13), then (i)* $\{V_t\}$ *and* $\{g_t\}$ *converge to* V *and* g *as* $t \downarrow 0$*, and (ii)* $v_{11} - \eta v_{12} \in \big(\rho, \rho/(1-\beta)\big)$

Note that the range for $v_{11} - \eta v_{12}$ implies that if $\rho = 0$, then g_{t1} converges to 0, indicating that without holding costs, initial trading is completely driven by price movements. It is also the case that $v_{12} - \eta v_{22} \in \big(0, 0.5Q^{-1}\big)$, which follows immediately from the initial conditions and the monotonicity of $\{V_t\}$. With lemma 1, it is possible to characterize the dynamics of the expected execution path in a precise manner.

Theorem 2 (The optimal execution path)**.** *If* $\rho > 0$*,* $\beta \in (0, 1)$*, and conditions (8), (9) and (13) are satisfied, then there exist times* $\underline{t}$ *and* $\bar{t}$*, where* $0 \le \underline{t} < \bar{t} < T$*, such that the expected optimal execution path is decreasing for* $t < \underline{t}$*, constant for* $\underline{t} \le t \le \bar{t}$*, and increasing for* $\bar{t} < t \le T$*.*

That is, on average for orders of sufficient size and duration, the optimal execution path exhibits a non-monotonic U-shaped size profile. This is because initially, the cost associated with remaining size, ρ, dominate any volatility or price related concerns. Yet as the trade executes, there is a price impact, and eventually the price increases to the point where the holding costs are equal to the other costs associated with impact and risk. During this stage, execution is dominated by the exogenous factors driving prices. As time elapses, due to the terminal conditions that the trade must be completed, price sensitivity diminishes to 0 and size sensitivity increases to 1. Consequently, any unfinished portions are executed at an increasing rate.

This result provides a new, more complete and simple explanation for the well documented patterns observed in the intra-day volume of U.S. equity markets. Given that institutional activity dominates trading in these markets, and that such trades are more commonly executed on an agency basis through a broker, the liquidity trading problem is clearly a very reasonable representation of the actual execution process of live markets. This is still true if the execution strategy is determined by the same agent that makes the investment decision. This is a problem that must be solved by all large market participants on a very regular, if not daily, basis. The fact that the U-shaped patterns in intraday volume are so pervasive and persistent indicates that there must be some regular process driving this behavior. Alternative explanations based upon asymmetric information are difficult to extend to a large universe of securities on a repeated, daily basis.

Using theorem 1, the proof of theorem 2 can be extended to the case for $\beta = 1$, indicating that this model can generate the behavior exhibited by alternative strategies. When $\beta = 1$, the optimal policy is independent of price, and only market impact, holding costs and risk matter. Thus, there is no incentive for the trader to delay execution, waiting to see if there are favorable exogenous shocks to price. Impact is minimized by executing at a constant rate for the entire horizon. As holding costs and/or risk aversion increase, the trader will shift execution to the earlier periods, creating the exponentially decreasing profiles given elsewhere. To the extent that the behavior of other models can be replicated, the current model is seen to be a generalization of these models.

5 Sensitivity Analysis and Numerical Examples

The following examples demonstrate the typical behavior and execution performance characteristics of the optimal policy (12). Where applicable, results for the execution policy provided in Huberman and Stanzl (2005) (the front-loaded strategy) are also considered. Since closed form expressions for their optimal policy are only available when $\alpha = 0$ (price impact is strictly permanent), comparisons are restricted to this case. Note that no such restriction is required for the solutions provided here.

5.1 Estimation and Selection of Parameter Values

To obtain parameter values that are representative of actual trading environments and consistent with the relevant empirical literature, parameter estimates and other empirical values are derived from the intra-day trade and quote data for the constituents of the S&P 500 index as of August 14, 2006. The median closing price for this population on that date was 41.2, and this value will be used wherever an initial price is required. The median average daily volume (ADV) for the 60 trading days prior to August 14, 2006 is 2.1 million shares.

To obtain values for σ_ϵ^2, σ_η^2, and γ, refer to the simple model of intra-day returns in Hora (2005),

$$r_t = \beta_t \omega_t + \epsilon_{r,t} \tag{14}$$

where $r_t = \ln(p_t/p_{t-1})$, and ω_t is the latent order flow for period t. Order flow is defined as the difference between buyer and seller initiated volume, which is the same definition used throughout this paper. To allow for time varying liquidity, the order flow parameter is constrained to be positive, but otherwise follows a random-walk, $\beta_t = \beta_{t-1} + \epsilon_{\beta,t}$, where $\epsilon_{\beta,t}$ is an iid Gaussian innovation. The return innovation, $\epsilon_{r,t}$ follows a Markov scale mixture of normals process. Thus, $\epsilon_{r,t}$ has mean zero, but variance that follows a finite state Markov chain. This specification allows for time varying and stochastic volatility with limited memory.

Model (14) provides for direct estimates of the volatility of order flow σ_η through the treatment of order flow as a stochastic latent variable. Since (14) allows for intraday variation, stock specific estimates for the volatility of order flow are taken to be the average of the intraday estimates over the 390 minutes in a normal trading day. The value used in the examples is the median of the averages for the S&P 500 constituents, which is 17.77, expressed in units of 100 shares. This indicates that the volatility of order flow over a typical one minute period is about 1800 shares.

When $\alpha = 0$, $\epsilon_{r,t}$ maps directly to the innovation ϵ_t in the current model. For other values of α, the correspondence is not direct, but since this relates to the distinction between the permanent and temporary impact of order flow, and the focus in upon high-frequency data, the difference as negligible. Since equation (14) uses continuously compounded returns and not price changes, the delta method is used to obtain an approximate value for σ_ϵ,

$$\sigma_\epsilon^2 \approx \text{var}(\epsilon_{r,t}) \exp(r_t) p_{t-1} \approx E\big[\text{var}(\epsilon_{r,t}) p_{t-1}\big]$$

where the last relation follows by taking mean returns to be zero. Values for each stock are obtained by taking the average of $\text{var}(\epsilon_{r,t})p_{t-1}$ for each 1-minute period in the trading day, using posterior means as estimates for $\text{var}(\epsilon_{r,t})$ and the observed standing mid-quote at the end of each period. The value for σ_ϵ is taken to be the (square-root) of the median of these averages, 3.69e-3, corresponding to a standard deviation of less than one penny.

The estimate for γ follows from the two expressions for $\partial p_t/\partial \omega_t$ resulting from equations (6) and (14), yielding $\gamma = p_{t-1}\beta_t$. Following the same procedure used to estimate σ_ϵ, the estimate for γ is taken to be the median of the intraday averages of $p_{t-1}\beta_t$, which is 6.05e-6. This value can be interpreted as indicating that a 10000 share *market* order executed in less than one minute is expected to have an impact of 6.05 cents. This value is about 15 basis points based upon the median price of 41.2, and is consistent with actual execution performance data that covers the same universe of stocks.

Regarding the coefficient of absolute risk aversion λ, there is little guidance that is directly related to the current problem. The large body of work related to the equity premium puzzle (Mehra and Prescott 1985) suggests a neighborhood around 2. However, this literature is concerned with the low frequency asset allocation problem facing a representative consumer over long horizons. Institutional execution traders addressing the high frequency liquidity trading problem are likely to exhibit different preferences, at least in terms of the magnitude of risk aversion. Haigh and List (2005) provide experimental evidence with professional futures and option traders as subjects, indicating that these professionals exhibit certain behavioral characteristics (myopic loss aversion) related to risk aversion to a greater extent than is typically found with non-professionals. This particular group of traders typically trades for their own accounts, while execution traders work on an agency basis. Given the fiduciary responsibility of the execution trader and long term client relationships, it is reasonable to allow for greater levels of risk aversion. With this in mind, a value of 2 is employed as a reference point and where appropriate values in the range of 1 to 3 are considered.

This range for the coefficient of absolute risk aversion is materially different than the values employed by previous research. Huberman and Stanzl (2005) use a values between 0 and 10e-4, while Almgren and Chriss (2000) use values no greater than 10e-6. The reason for these low values relates to the problem indicated in section 3, where lack of time separability results in over-weighting the variance of costs. In this case, variance aversion strongly dominates all other factors. Consequently, using the range given above for the coefficient of risk aversion results in degenerate execution paths in the sense that they attempt to complete the order in as little as 3 minutes, even for trades as large as 50% of average daily volume. Clearly such strategies are infeasible in actual markets. To allow for a more meaningful comparison, when computing the execution path for the front-loaded strategy λ is divided by the number of periods T.

The holding cost parameter ρ is intended to reflect the cost associated with the incomplete portion of the entire order, loosely analogous to the cognitive dissonance associated with procrastination. One approach to arriving at a numerical value is to consider the overnight holding, or inventory, costs that an investment bank might face. Overnight repo rates refer to similar risks, albeit for different instruments in different markets. Expressing a rate of 4.25% on a per-minute basis results in a value of 8.77e-6. Dividing by 100 to reflect the fact that this is a non-monetary cost borne by the trader on an intra-day basis, and not an overnight monetary cost, and then rounding the result yields a value of 10e-8. Since this a per share cost, ρ is taken to be 10e-8/Q.

Finally, to reflect situations that are consistent with non-trivial institutional trading activity, the overall trade size Q is set at 525000 shares, which is 25% of the median average daily volume for the population. Similarly, the time horizon T is 180 minutes. In all cases, β is set equal to 0.95.

5.2 Results

With the parameter values as specified above, it is a simple matter to compute the sequence of optimal policy functions under a variety of conditions. In order to provide an understanding of the default behavior and to demonstrate sensitivities to certain parameters, the analysis begins with a consideration of the expected policy functions. This isolates trader behavior from the exogenous order flow and price innovations, and thus, indicates the responses due to the deterministic impact of the trader's own actions. Given the linearity of the optimal policy functions, the expected policies are easily computed by holding exogenous order flow, ω_t, and the price innovation ϵ_t at their common mean of 0.

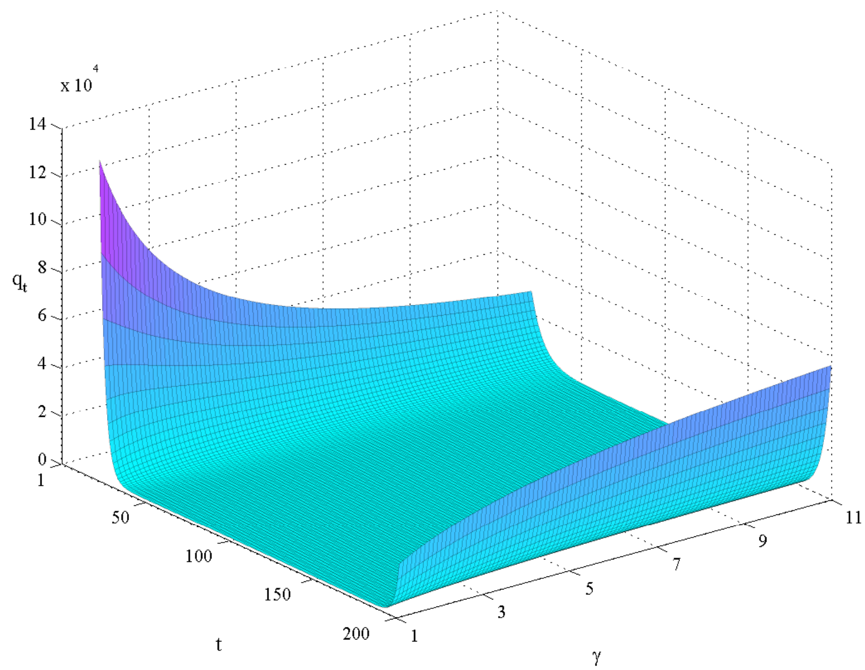

Figure 1: The expected optimal policy as the impact parameter γ increases from 1.0e-6 to 11e-6, holding $\alpha = 0.5$ and $\lambda = 2$.

Figure 5.2 displays the expected optimal policy holding α at 0.5 and λ at 2, but varying the impact parameter γ from 1.0e-6 to 11e-6. Of course, as theorem 2 indicates, the expected transaction size is positive in the initial periods, decreasing to a small positive (steady-state) value, and then increasing again into the terminal period. It is important to note that this does not mean that the optimal policy will remain constant in the intermediate period. Rather, during this portion, the policy will react to exogenous price movements. With lower impact costs, the trader will execute more in the initial periods and less in the terminal periods as there is less expected benefit from delay. Alternatively, with higher impact costs, there is a greater expected benefit from delay, and the trader will execute less in the early periods and more in the final periods. (This asymmetry raises an interesting empirical question. Does the percentage of volume executed in the closing minutes increase as liquidity, as characterized by order flow impact, decreases?)

Another key feature of the optimal policies (12) is that they are responsive to favorable innovations in price, as described in theorem 1. Since this feature is unique to this model, it is informative to compare the actual execution behavior of the optimal policies with the behavior of the front-loaded strategies. The latter case can be characterized as heavily front-loaded strategies in the sense that they complete the entire order well before the allotted time, starting with a very large size that decreases exponentially. Thus, these strategies are also monotonically decreasing, and as indicated above, they are independent of price shocks. Essentially these strategies are minimizing variance by realizing costs through the deterministic impact in the initial periods.

To visually compare the behavior of the two strategies the time horizon is reduced to 60 minutes and the total trade size is reduced to 100,000 shares. All other parameters are as above. Similar to the previous example, the exogenous order flow shocks are held at their mean of 0, and the exogenous price shocks, ϵ_t are held at their mean of 0 except for periods 20 and 40. In period 20, ϵ_t is set to -0.20, while in period 40, it is set to +0.20. Thus, the period 20 shock is favorable to the trader because it reduces prices by 20 cents, while the opposite is true for period 40.

Figure (5.2) displays the results for the optimal policy. Note that the initial trade sizes are positive, but decreasing as the trader impacts prices. In period 21, after the negative price shock, the trade size increases from less than 500 shares to 5000 shares, taking advantage of the reduced prices. The initial pattern repeats until period 40, where the adverse price shock occurs. At this point, trade size is reduced to 0 until the approaching terminal time forces the trader to finish the order. The total cost for this execution scenario is \$0.0858.

Figure (5.2) displays the results for the front-loaded strategy. Even with the reduction in λ, this strategy still executes nearly all of the order by the 10th period, a result that is consistent with the examples provided in Huberman and Stanzl (2005) and Almgren and Chriss (2000). This results in a very steep price impact trajectory, incurring large costs in the early periods. Of course, since this strategy is independent of the current price, exogenous shocks never have any effect on the execution path. Even if the strategy were price dependent, since the order is essentially complete before the 20th period, it would be unable to effectively take advantage of any favorable price changes. Not surprisingly, the total cost for this scenario is \$0.186, a 210% increase over the costs for the optimal policy.

This example provides a simple demonstration of the price responsiveness of the optimal policy. The appropriate approach to comparing the performance of the two strategies is based upon a consideration of the associated distribution of costs. To this end, the path of prices and order flow is simulated by sampling ϵ_t and ω_t from normal distributions with mean zero and standard deviations of 3.69e-3 and 1773 respectively. The execution path and associated cost for each strategy is readily computed from the optimal policy functions and common draws for ω_t and ϵ_t. Sampling 10,000 scenarios, the optimal policy given here

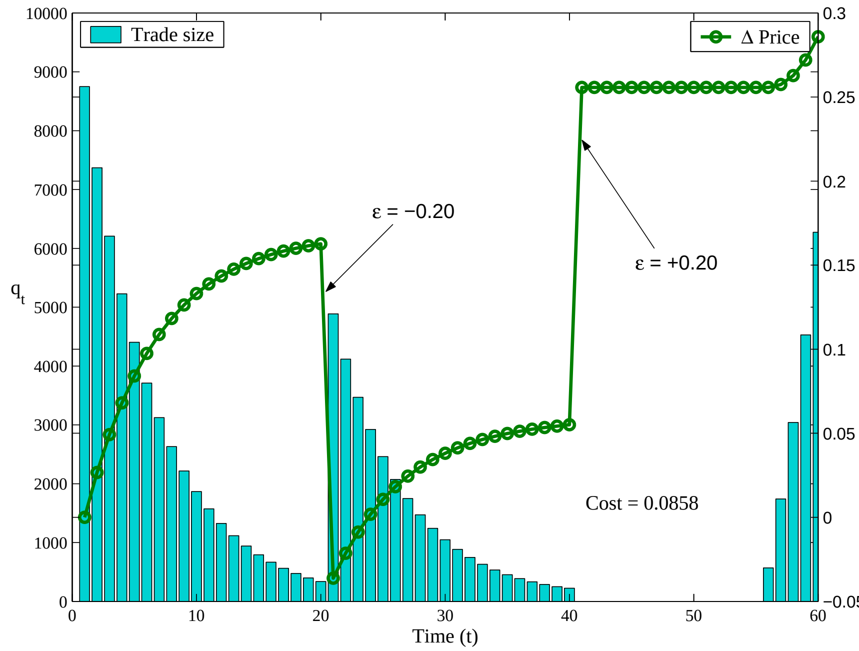

Figure 2: Sample execution path for the optimal policy with Q=100000 shares, T=60, and fixed price shocks at $t = 20$ and $t = 40$. Total cost is \$0.0858.

results in a mean cost of 1.70 and a standard deviation of 0.10, while the front-loaded strategy has a mean cost of 1.95 and a standard deviation of 0.04. The probability that the cost under the optimal policy is greater than the cost for the front-loaded strategy is 0.0060.

Figure (5.2) displays density estimates for the total cost associated with each strategy. This figure affirms the characterization of front-loaded strategies as strategies that realize excessive costs deterministically in order to avoid risk, which is (obviously) two-sided. Clearly, responding to price movements allows for substantial cost savings, so much so that the incremental "risk" is irrelevant. This result also supports the result that front-loaded strategies are not optimal for risk averse traders.

6 The Portfolio Problem

In practice, the single security approach has proven to be adequate for most applications. However, the portfolio execution problem is both of theoretical and practical significance. In general, the dynamics of order flow exhibit a clear, market wide factor, but very little is understood about the nature of this factor. In spite of this, portfolio trading comprises a significant percentage of daily volume in U.S. equity markets. For example, during the week of August

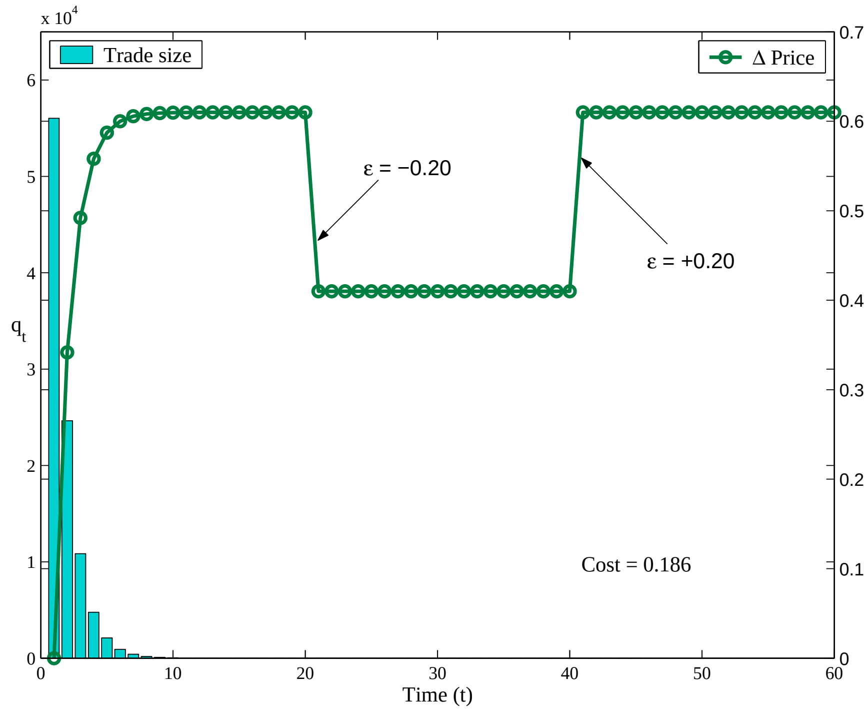

Figure 3: Sample execution path for the front-loaded strategy with Q=100000 shares, T=60, and fixed price shocks at $t = 20$ and $t = 40$. Total cost is \$0.186.

14, 2006, the NYSE reported that "program trades" accounted for 29.1 percent of the volume traded on the exchange, with a total of over 3 billions shares. The NYSE defines program trades as strategies consisting of the simultaneous purchase or sale of at least 15 different stocks with a value of at least \$1 million. This definition is very conservative, as the exchange cannot be aware of portfolio strategies that are disaggregated prior to execution.

Clearly, portfolio execution strategies are widely employed in practice. This is in spite of the lack of an adequate theoretical basis. Fortunately, the results for the single security case presented here are easily extended to the portfolio case, preserving the key features of representing risk aversion and holding costs while still allow for tactical price responsiveness. Equally important, is the fact that the recursive, linear-quadratic structure is very easy to compute and does not rely upon any numerical optimization.

The generalization to the n security case is straightforward, with most elements of the single security problem mapping into the obvious n-dimensional counter-parts. For brevity, *vector* indicates an n-dimensional vector, and *matrix* indicates an n by n matrix. All matrices are assumed to be of full rank. Thus, Q indicates the vector of target positions, where the elements of Q can be positive or negative. The vectors of shares remaining to be executed and the quantity

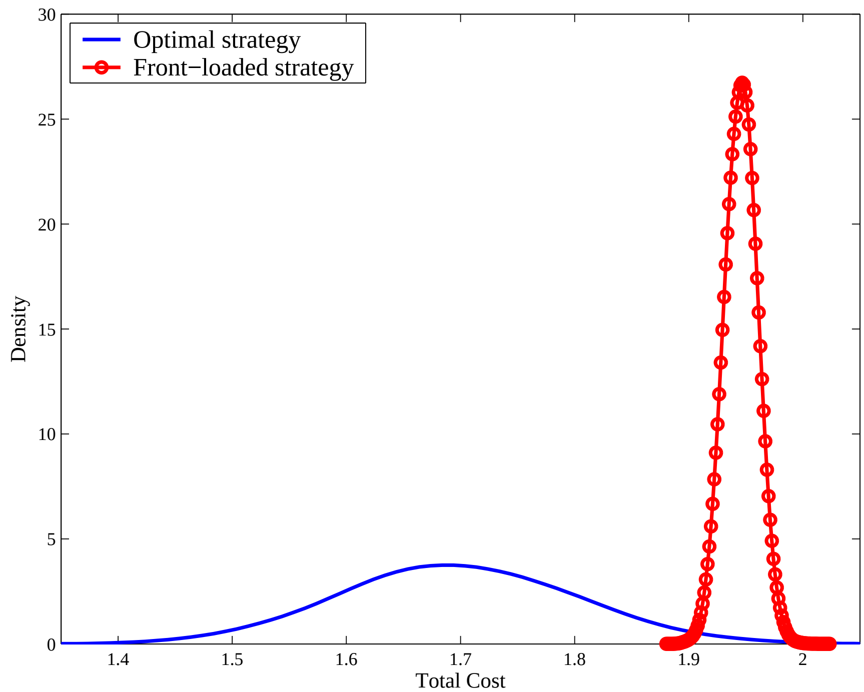

Figure 4: Total cost density estimates for the optimal policy and the front-loaded strategy.

executed in period t are x_t and q_t.

Exogenous order flow, ω, is a random vector with mean 0 and variance matrix Σ_ω, assumed to be symmetric and positive definite (spd). Using a linear price impact function with positive definite parameter matrix $\Gamma = [\gamma_{ij}]$, where $\gamma_{ii} > 0$, transaction prices are given by

$$p_t = \tilde{p}_t + \Gamma(\omega_t + q_t)$$

and $\tilde{p}_t$ is the vector or price quotes at the beginning of period t. The dynamics for price quotes are

$$\tilde{p}_{t+1} = \alpha\tilde{p}_t + (I - A)p_t + \epsilon_t$$

where ϵ_t is a random vector with mean 0 and variance matrix Σ_ϵ, also spd. The matrix $A = diag(\alpha_i)$ corresponds to the persistence of order flow shocks on a security specific basis.

The parameters defining trader preferences do not require any generalization, as risk aversion and holding costs should be common across securities given our high-frequency focus. Thus, defining $Q^{-1} = diag(Q_i^{-1})$, and $\delta_t = \tilde{p}_t - \tilde{p}_0 \in \Re^n$, expected utility becomes

$$E_t(u(q_t, x_t)) = \delta_t' Q^{-1} q_t + q_t' Q^{-1} \Gamma q_t + \frac{\lambda}{2} q_t' \Gamma' \Sigma_\omega \Gamma q_t + \rho x_t' x_t$$

The parameter θ in the single security problem corresponds to the matrix Θ, where

$$\Theta = Q^{-1}\Gamma + \frac{\lambda}{2}\Gamma'\Sigma_\omega\Gamma.$$

The state vector s_t consists of $2n$ elements, the first n corresponding to the number of shares remaining for each security, and the second n elements indicating the current price deviation from the initial price for each security, δ_t. To describe the state transitions, let ι_n be an n dimensional vector of 1s, define the 2 by n matrix B and the vector z_{t+1} as follows

$$B = \begin{bmatrix} -\iota'_n \\ (I-A)\Gamma \end{bmatrix}, \qquad z_{t+1} = \begin{bmatrix} 0_n \\ \epsilon_{t+1} + (I-A)\Gamma\omega_t \end{bmatrix}.$$

Then the state transition equation is

$$s_{t+1} = s_t + Bq_t + z_{t+1}.$$

Thus, the portfolio problem is also linear-quadratic, and completely described by the preceding expressions. Proposition 1 and its proof easily extend to the current problem, with the obvious generalizations providing recursive expressions for the value and optimal policy functions.

Portfolio execution can involve some very interesting and complex constraints that are often critical for effective risk management. Since even simple linear constraints can have a severe, negative impact on computational performance in applications, the simplicity of the linear-quadratic structure suggests great potential. Two examples are as follows.

First, let $B = \{i|Q_i > 0\}$ and $S = \{i|Q_i < 0\}$, the set of indices corresponding to buys and sells respectively. Then *strict dollar-neutrality* is given by the constraint

$$\sum_{i\in B}\sum_{j=1}^{t} p_{ji}q_{ji} = \sum_{i\in S}\sum_{j=1}^{t} p_{ji}q_{ji},$$

equating the dollar value of the realized quantities of buys and sells. Similarly, *strict (market) beta-neutrality* is

$$\sum_{i\in B}\beta_i x_{ti} = \sum_{i\in S}\beta_i x_{ti},$$

where β_i is the usual market-beta for the *i-th* security. Both restrictions can used in applications (hopefully, not simultaneously) by allowing for a deviation from neutrality within a given range.

7 Conclusion

This paper provides a complete analysis of the problem facing a risk averse, transaction cost minimizing liquidity trader. A critical feature of this problem is its finite time horizon, which necessitates careful consideration of the dynamic properties of the solution. Ultimately, it is shown that the finite horizon results in a very interesting, non-monotonic execution path that can be linked to the empirical features of intraday volume and volatility.

By employing a time separable cost function and a linear price adjustment function, the problem is easily expressed as a linear-quadratic dynamic program. Hence, closed form expressions for the value and optimal policy functions are easily obtained, allowing for a detailed analysis of the structure of the problem as well as the dynamics of the expected execution path. It is shown that, by incorporating holding costs in addition to risk aversion, and allowing for discounting, the optimal policy functions are sensitive to exogenous changes in price, and that it is optimal for the liquidity trader to engage in a particular form of tactical trading. It is also shown that, as long as the target position is relatively large in a particular manner, the expected optimal execution path exhibits a U-shaped profile, completely consistent with the U-shaped average intraday volume and volatility profiles observed in many asset markets in an essentially uniform manner. Thus, this paper provides a new, very simple and complete explanation for these empirical regularities. An examination of how the expected execution path varies with liquidity also suggests a very interesting analysis of the relationship between liquidity and the relative importance of opening and closing auctions.

These results also explain the curious features of alternative solutions provided elsewhere. Due to the lack of holding costs and discounting, the alternative approaches produce execution strategies that are independent of price changes. Thus, even though the objective is to minimize transaction costs, the strategy will not alter its rate of execution in the face of even extreme changes in price. This is clearly inconsistent with cost minimization and with observed trader behavior. Another problematic feature of most alternative solutions is that the trader will complete the trade well before the allotted time, executing a substantial percentage in the first few periods. It is shown that this results from the lack of time separability in preferences that reflect variance aversion. Effectively, these alternative solutions are solutions to the problem of minimizing the variance of transaction costs, and not the stated objective. Consequently, it can be argued that the solution provided here is actually the first solution to the problem of minimizing risk adjusted transaction costs. Not surprisingly, a numerical experiment demonstrates that the alternative solutions not only perform worse than the solution provide here, but are stochastically dominated by the solution.

Finally, the generalization to portfolios is provided. It is shown that the linear-quadratic structure can be preserved, and the solutions follow in an analogous manner. This generalization is most interesting for applications and warrants considerable attention in future research.

Appendix

The proofs of the results presented in section 4 rely upon certain well-known properties of a general class of dynamic programs, of which the liquidity trader's problem is a special case. These properties are summarized in the following lemmas, which require additional definitions.

First, let C denote the space of real, convex functions defined on $\Re^2$. Then, define the operator J on C by

$$(Jf)(s) = \max_{q \in \Gamma(s)} E\{c(s,q) + \beta f(s)\}$$

where s, Γ, and β are previously defined, and c is the cost function (5).

Lemma 2. *Given the definitions of c and Γ, and $\beta \in (0,1)$, (i) the operator J maps C into itself, $J : C \rightarrow C$, and (ii) there exists a unique bounded continuous function satisfying the functional equation (7).*

Proof. These results are standard, see theorems 9.6 and 9.8 in Stokey, Lucas, and Prescott (1989). □

Note that the solution to equation (7) can only obtained in the limit as $t \downarrow 0$. Throughout, the iteration of the functional equation proceeds backwards from the fixed terminal time T. While asymptotic results are not necessary for the general results, they are useful in refining the characterization of the optimal policy functions. In any event, lemma 1 establishes conditions that guarantee convergence within relative to a minimum trade size.

Proof of Proposition 1. This result is standard (see Ljungqvist and Sargent (2000), which is followed very closely.)

The conjecture is that each function v_t has the form $s_t' V_t s_t + d_t$, where d_t is a scalar. In this case, substituting the proposed solution into the right-hand side of equation (7) yields

$$v(s_t) = \min_{q_t \in \Gamma(s_t)} \Big\{ s_t' R s_t + 2 s_t' W q_t + \theta q_t^2 + \beta E([s_t + Bq_t + z_{t+1}]' V_{t+1} [s_t + Bq_t + z_{t+1}] + \beta d_{t+1} \Big\}. \quad (15)$$

Since $E(v_{t+1}|s_t) = 0$, the first-order condition for the minimization on the right-hand side of the functional equation is,

$$q_t^* = -(\theta + B' V_{t+1} B)^{-1} (\beta B' V_{t+1} + W') s_t.$$

The fixed terminal costs imply that $v_T(s) = s' V_T s$, where $V_T = \begin{bmatrix} \theta + \rho & 0.5Q^{-1} \\ 0.5Q^{-1} & 0 \end{bmatrix}$, and $d_T = 0$.

Substituting this expression for q_t^* into (15) results in the following recursive expression for V_t

$$V_t = R + \beta V_{t+1} - (\beta B' V_{t+1} + W')'(\theta + \beta B' V_{t+1} B)^{-1}(\beta B' V_{t+1} + W'). \quad (16)$$

This also provides the recursion for d_t

$$\begin{aligned} d_t &= \beta\big[\mathrm{tr} V_{t+1} E(z_{t+1} z_{t+1}') + d_{t+1}\big] \\ &= \beta(\sigma_\epsilon^2 + \eta^2 \sigma_\omega^2)(v_{t+1,12} + v_{t+1,22} + d_{t+1}). \end{aligned}$$

□

Lemma 3 (Additional properties of $\{V_t\}$)**.** *The sequence of matrices $\{V_t\}$ defined by equation (11) is (a) symmetric, (b) monotonically increasing with t, (c) converges to a limit V as t decreases, and (d) satisfies $\theta + \beta B' V_{t+1} B > 0$ for all t.*

Proof. (a) The symmetry of V_t follows from the symmetry of V_T and R, and from equation (16).

(b) $\{V_t\}$ is increasing in the sense that $s' V_t s \leq s' V_{t+1} s$ for all $s \in \Re^2$. The following parallels Stokey, Lucas, and Prescott (1989), although in this case the recursion initiates with the terminal period T and decreases to the initial period.

Let $v_t(s) = s' V_t s$ (recall that $d_T = 0$). Note that $v_T(s) = F(s, s)$, where F is the expected cost function, and with $q_t = s_1$, $s + Bq_t = \begin{bmatrix} 0 & s_2 + \gamma s_1 \end{bmatrix}'$, so $s' V_T s = 0$. Thus, for any feasible state vector s,

$$v_T(s) = F(s, s_1) + \beta v_T(g_T(s, s_1)). \quad (17)$$

Thus, v_{T-1} satisfies

$$\begin{aligned} v_{T-1}(s) &= \min_{q \in \Gamma(s)} \big\{F(s, q) + \beta v_T(g(s, q))\big\} \\ &\leq \big\{F(s, s_1) + \beta v_T(g(s, s_1))\big\} \\ &= v_T(s) \end{aligned}$$

where the last equality follows from (17). Similarly, for v_{T-2}

$$\begin{aligned} v_{T-2}(s) &= \min_{q \in \Gamma(s)} \big\{F(s, q) + \beta v_{T-1}(g(s, q))\big\} \\ &\leq \min_{q \in \Gamma(s)} \big\{F(s, q) + \beta v_T(g(s, q))\big\} \\ &= v_{T-1}(s). \end{aligned}$$

Repeating this process for $t = T-3, \ldots, 0$ proves monotonicity.

The monotonicity of v_t carries over to each element $v_{t,ij}$ of V_t. To see this, first take $s = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and then $s = \begin{bmatrix} 0 & 1 \end{bmatrix}$, demonstrating that $v_{t,ii} < v_{t+1,ii}$ for $i = 1, 2$. Then, for any $s_1 < 0$ and $s_2 < 0$,

$$2 s_1 s_2 (v_{t,12} - v_{t+1,12}) \leq s_1 (v_{t+1,11} - v_{t,11}) + s_2 (v_{t+1,22} - v_{t,22}) \leq 0$$

from which it follows that $v_{t,12} \leq v_{t+1,12}$.

(c) From lemma 1, and theorem 9.12 in Stokey, Lucas, and Prescott (1989), the existence of the solution to the functional equation (7), is equivalent to the convergence of $\{V_t\}$ as $t \to -\infty$.

(d) From lemma 1, the operator J preserves convexity. Thus, from the second order conditions corresponding to the minimization in (7) it follows that

$$\theta + \beta B' V_{t+1} B > 0$$

for all t.

□

Proof of Theorem 1. The result follows if $g_{t,1} > 0$ and $g_{t,2} < 0$ for all $t = 0, \ldots, T$. Define $\tilde{v}_{ij} = v_{t+1,ij}$. The elements of g_t are as follows

$$g_t = (\theta + \beta B' V_{t+1} B)^{-1} (\beta B' V_{t+1} + W')' \tag{18}$$

$$= (\theta + \beta B' V_{t+1} B)^{-1} \begin{bmatrix} \beta(\tilde{v}_{11} - \eta \tilde{v}_{12}) \\ \beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1} \end{bmatrix}. \tag{19}$$

Since $(\theta + \beta B' V_{t+1} B) > 0$ for all t (lemma 2), the result follows if $\tilde{v}_{11} - \eta \tilde{v}_{12} > 0$ and $\beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1} < 0$ for $t = 0, \ldots, T-1$.

This is proved by reverse induction. To establish the initial conditions, note that, $v_{T,11} - \eta v_{T,12} > 0$ is equivalent to $\theta + \rho > \eta 0.5 Q^{-1}$, which is true for all feasible parameter values defined in the text. Similarly, $\beta(v_{T,12} - v_{T,22}) + 0.5Q^{-1} < 0$ is equivalent to $\beta \in [0, 1)$. Thus, the initial conditions are satisfied for all feasible parameter values.

The induction hypothesis is that $\tilde{v}_{11} - \eta \tilde{v}_{12} > 0$ and $\beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1} < 0$.

From equations (16) and (19)

$$V_t = \begin{bmatrix} \rho + \beta \tilde{v}_{11} - \beta g_{t,1}(\tilde{v}_{11} - \eta \tilde{v}_{12}) & \beta \tilde{v}_{12} - \beta g_{t,2}(\tilde{v}_{11} - \eta \tilde{v}_{12}) \\ \beta \tilde{v}_{21} - g_{t,1}(\beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1}) & \beta \tilde{v}_{22} - g_{t,2}(\beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1}) \end{bmatrix} \tag{20}$$

Then,

$$v_{11} - \eta v_{12} = \rho + \beta(\tilde{v}_{11} - \eta \tilde{v}_{12})[1 - G_{t1} + \eta G_{t2}] \tag{21}$$

$$v_{12} - \eta v_{22} - 0.5Q^{-1} = \big(\beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1}\big)(1 - g_{t,1} + \eta g_{t,2}) \tag{22}$$

Again from equation (19)

$$\begin{aligned} (\theta + \beta B' V_{t+1} B)(1 - g_{t,1} + \eta g_{t,2}) \\ &= \theta + \beta B' V_{t+1} B - \beta(\tilde{v}_{11} - \eta \tilde{v}_{12}) + \eta \big[\beta(\tilde{v}_{12} - \eta \tilde{v}_{22}) - 0.5Q^{-1}\big] \\ &= \theta - \eta 0.5 Q^{-1} > 0. \end{aligned} \tag{23}$$

Since each $\theta + \beta B' V_{t+1} B > 0$, it must be the case that $1 - g_{t,1} + \eta g_{t,2} > 0$, and from equations (21), (22), it follows that $v_{11} - \eta v_{12} > 0$ and $\beta(v_{12} - \eta v_{22}) - 0.5Q^{-1} < 0$ for t, proving the desired result.

The result for $\beta = 1$ follows directly from (19) and (20). If $\beta = 1$, then given V_T, from (19) $g_{T-1,2} = 0$, and from (20) this implies that $v_{T-1,22} = 0$ and $v_{T-1,12} = 0.5Q^{-1}$. Proceeding recursively with (19) and (20), it is clear that $v_{t,12} = 0.5Q^{-1}$, $v_{t,22} = 0$, and $g_{t,2} = 0$ for all $t = 0, \ldots, T-1$.

□

Lemma 4 (Additional properties of $\{g_t\}$). *With g_t defined by equation (12), the sequence $\{1 - g_{t,1} + \eta g_{t,2}\}$ (a) monotonically decreases to 0 with t, and (b) is within $[0, 1)$ for all t. In addition, (c) $\{g_t\}$ is non-decreasing.*

Proof. (a) Lemma 3 establishes that $\{V_t\}$ is monotonically increasing with t. Thus, equation (23) indicates that $1 - g_{t,1} + \eta g_{t,2}$ is monotonically decreasing with t.

(b) The proof of theorem 1 establishes that $1 - g_{t,1} + \eta g_{t,2} \geq 0$. To see that $1 - g_{t,1} + \eta g_{t,2} < 1$, first note that by theorem 1, $g_{t,1} > 0$ and $g_{t,2} < 0$. Since $\eta \geq 0$, $1 - g_{t,1} + \eta g_{t,2} > 1$ implies that $1 > 1 + g_{t,1} - \eta g_{t,2}$, which is not possible.

(c) From equations (21) and (22), and the previous result from part (b), the numerator for each component of g_t is an increasing sequence. The denominator, $c + \beta B' V_{t+1} B$ is increasing, but since it is comprised of the difference of the sequences in the numerators, the rate of increase is no greater than the rate of increase in the numerators.

□

Proof of Lemma 1. The convergence of $\{V_t\}$ and $\{g_t\}$ as t decreases from T has already been established in lemma A.2. It remains to demonstrate that if conditions (8), (9), and (13) are satisfied, then this convergence occurs in a finite number of periods.

Let $a_t = v_{t,11} - \eta v_{t,12}$, $b_t = v_{t,12} - \eta v_{t,22}$, $a = v_{11} - \eta v_{12}$, $b = v_{12} - \eta v_{22}$ and $k = \theta - \eta 0.5Q^{-1}$. Define $\tilde{g}_t = 1 - g_{t,1} + \eta g_{t,2}$. Lemma A.3 indicates that $\tilde{g}_t$ monotonically increases towards some limit $\tilde{g} \in (0, 1)$ as $t \downarrow 0$.

By equation (21)

$$\begin{aligned} a_t &= \rho + \beta a_{t+1}(1 - g_{t,1} + \eta g_{t,2}) \\ &= \rho \sum_{i=0}^{T-t+1} \beta^i \prod_{j=1}^{i} \tilde{g}_{t+i} + \beta^{T-t} a_T \prod_{i=1}^{T-t} \tilde{g}_{t+i}. \end{aligned}$$

Clearly, $a_t \geq \rho$ for all $t \geq 0$.

Define the sequence $\{\tilde{a}_t\}$ by

$$\tilde{a}_t = \frac{\rho}{1 - \beta} + \beta^{T-t} a_T,$$

where $a_T = v_{T,11} - \eta v_{T,12}$. Since, by lemma A.3, $(1 - g_{t,1} + \eta g_{t,2}) \in [0, 1)$ for all $t \geq 0$, it is immediate that $a_t < \tilde{a}_t$, and $a = \lim(v_{t,11} - \eta v_{t,12}) \in (\rho, \rho/(1 - \beta))$.

In addition, it follows that $\{a_t\}$ decreases geometrically (as $t \downarrow 0$) with rate bounded by β.

Substituting b_t into equation (22) it follows that

$$b_t = \beta b_{t+1} \tilde{g}_{t+1} + 0.5Q^{-1}(1 - \tilde{g}_{t+1}) < \beta \tilde{g} b_{t+1} + 0.5Q^{-1},$$

and therefore, $\{b_t\}$ is also decreasing geometrically with a rate bounded by $\beta\tilde{g}$.

Given the notion of convergence relative to a minimum trade size, it suffices to establish that there is some $t^* \in (0, T]$ such that if $t + 1 < t^*$, then

$$|g'_{t+1}s - g'_t s| < q_{min} \tag{24}$$

for any state vector $s \in [0, Q] \bigotimes [0, \eta Q]$, where q_{min} is the minimum trade size.

Applying the (rate of) convergence result for a_t yields

$$|a_{t+1} - a_t| \leq \beta^{T-t}(1 - \beta)a_T = \beta^{T-t}(1 - \beta)(\theta + \rho - \eta 0.5Q).$$

Since

$$\theta + \beta\rho - \beta\eta 0.5Q^{-1} < \theta + \beta a_t - \beta\eta b_t < \theta + \beta a_{t+1} - \beta\eta b_{t+1}$$

using the previous inequality, it follows that

$$\left| \frac{\beta a_{t+1} 0.5Q}{\theta + \beta a_{t+1} - \beta\eta b_{t+1}} - \frac{\beta a_t 0.5Q}{\theta + \beta a_t - \beta\eta b_t} \right| < \beta^{T-t+1}(1 - \beta)Q$$

and applying condition ($\beta^{T-t+1}(1 - \beta) < 0.5Q^{-1}$)

$$|g_{t+1,1} - g_{t,1}| < \beta^{T-t+1}(1 - \beta)Q < \frac{q_{min}}{2}. \tag{25}$$

Since $\{b_t\}$ decreases at the same rate as $\{a_t\}$, repeating the preceding steps in a similar manner for $\{b_t\}$ results in

$$|g_{t+1,2} - g_{t,2}| < \beta^{T-t+1}(1 - \beta)$$

and from condition ($\beta^{T-t+1}(1-\beta) < 0.5Q^{-1}$) it follows that, for any $s_2 \in [0, \eta Q]$

$$|g_{t+1,2} - g_{t,2}|s_2 < \beta^{T-t+1}(1 - \beta)s_2 < \frac{q_{min}}{2}. \tag{26}$$

Combining (25) and (26) establishes the desired result.

□

Proof of Theorem 2. The expected optimal execution path is obtained by taking $\omega_t = 0$ and $\epsilon_t = 0$ for all t. Then, $\delta_t = \eta \sum_{i=1}^{t-1} q_i = \eta(Q - x_t)$.

In the limit (as t decreases from T to 0), q^*_t is decreasing as t increases if and only if

$$g_1 x_t + g_2 \delta_t \geq g_1 x_{t+1} + g_2 \delta_{t+1}.$$

Since $x_t \downarrow 0$, $0 \leq \delta_t \leq \delta_{t+1}$, $g_1 > 0$ and $g_2 < 0$, it follows that $g_1 x_t \geq g_1 x_{t+1}$ and $g_2 \delta_t \geq g_2 \delta_{t+1}$. Thus, in the limit, the expected execution size is never increasing. As long as $q_t^* > 0$ for some t, then the expected execution size is strictly decreasing until the price impact, δ_t, has increased to the point where $g_2\delta_t$ offsets $g_1 x_t$. Since $\delta_1 = 0$ and $x_1 = Q > 0$, $q_1^* > 0$, so there is some $\underline{t} \in (1, T)$, such that $t < \underline{t}$ implies $q_t^* > q_{t+1}^*$.

The condition for increasing q_t^* as t approaches T is

$$g_{t,1}x_t + g_{t,2}\delta_t < g_{t+1,1}(x_t - g_{t,1}x_t - g_{t,2}\delta_t) + g_{t+1,2}\delta_{t+1}. \tag{27}$$

To show that this is satisfied, consider separate inequalities for the first and second components of the policy function in the preceding inequality

$$g_{t,1}x_t < g_{t+1,1}(1 - g_{t,1})x_t - g_{t+1,1}g_{t,2}\delta_t \tag{28}$$

and

$$g_{t,2}\delta_t < g_{t+1,2}\delta_{t+1}. \tag{29}$$

Since $g_{t+1,1}g_{t,2}\delta_t < 0$, a sufficient condition for inequality (28) to be satisfied is

$$g_{t,1} < g_{t,1}(1 - g_{t+1,1}). \tag{30}$$

This is immediate for $t = T - 1$ since $g_T = 1$ and $g_{T-1} < 1/2$,

$$g_{T-1,1} < g_{T,1}(1 - g_{T-1,1}) = (1 - g_{T-1,1}).$$

Equivalently, $g_{T-1,1}/g_{T,1} < 1 - g_{T-1,1}$. For $t < T - 1$, consider two sequences $\{g_{t,1}/g_{t+1,1}\}$ and $\{1 - g_{t,1}\}$. In the limit (for small t) with $g_1 > 0$, $g_{t,1}/g_{t+1,1} > 1 - g_t$ or $g_{t,1} > g_{t+1,1}(1 - g_{t,1})$, and (30) is reversed. Since $\{g_{t1}\}$ is monotonic, it follows that $\{g_{t1}/g_{t+1,1}\}$ and $\{1 - g_{t1}\}$ intersect only once. Thus, for some $\bar{t}, 1 < \bar{t} < T$, for $t > \bar{t}$, $g_{t1}/g_{t+1,1} < 1 - g_{t,1}$ and (30) is satisfied, implying that (28) is satisfied.

Inequality (29) follows from the facts that both $\{g_{t,2}\}$ and $\{\delta_t\}$ are monotonically increasing and $\delta_t \geq 0$. Thus, for $t > \bar{t}$, both (28) and (29) are satisfied, and q_t^* is increasing as t approaches T. □

References

Admati, A., and P. Pfleiderer (1988): "A Theory of Intraday Patterns: Volume and Price Variability," *Review of Financial Studies*, 1(1), 3–40.

——— (1989): "Divide and Conquer: A Theory of Intraday and Day-of-the-Week Mean Effects," *Review of Financial Studies*, 2, 189–223.

Almgren, R., and N. Chriss (2000): "Optimal Execution of Portfolio Transactions," *Journal of Risk*, 3(2), 5–39.

Bertsimas, D., P. Hummel, and A. Lo (1999): "Optimal Control of Execution Costs for Portfolios," *Computing in Science and Engineering*, 1(6), 40–53.

Bertsimas, D., and A. Lo (1998): "Optimal Control of Execution Costs," *Journal of Financial Markets*, 1(1), 1–50.

Bhushan, R. (1991): "Trading Costs, Liquidity, and Asset Holdings," *Review of Financial Studies*, 4(2), 343–360.

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

——— (1997): "Institutional Equity Trading Costs: NYSE versus NASDAQ," *Journal of Finance*, 52(2), 713–735.

Giovannini, A., and P. Weil (1989): "Risk Aversion and Intertemporal Substitution in the Capital Asset Pricing Model," NBER Working Paper #2824.

Haigh, M., and J. List (2005): "Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis," *Journal of Finance*, 60(1), 523–534.

Hora, M. (2005): "A Filtering Approach to Market Impact Estimation," Unpublished manuscript.

Huberman, G., and W. Stanzl (2004): "Price Manipulation and Quasi-Arbitrage," *Econometrica*, 72(4), 1247–1275.

——— (2005): "Optimal Liquidity Trading," *Review of Finance*, 9(2), 165–200.

Kocherlakota, N. (1990): "Disentangling the Coefficient of Relative Risk Aversion from the Elasticity of Intertemporal Substitution: An Irrelevance Result," *Journal of Finance*, 45(1), 175–190.

Kraus, A., and H. Stoll (1972): "Price Impacts of Block Trading on the New York Stock Exchange," *Journal of Finance*, 27(3), 569–588.

Kreps, D., and E. Porteus (1978): "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica*, 46(1), 185–200.

Ljungqvist, L., and T. Sargent (2000): *Recursive Macroeconomic Theory*. MIT Press, Cambridge and London.

Mehra, R., and E. Prescott (1985): "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, 15, 145–161.

Mendelson, H., and T. Tunca (2004): "Strategic Trading, Liquidity, and Information Acquisition," *Review of Financial Studies*, 17(2), 295–337.

Milgrom, P., and N. Stokey (1982): "Information, Trade and Common Knowledge," *Journal of Economic Theory*, 26, 17–27.

Obizhaeva, A., and J. Wang (2005): "Optimal Trading Strategy and Supply/Demand Dynamics," NBER Working Paper #11444.

Stokey, N., R. Lucas, Jr., and E. Prescott (1989): *Recursive Methods in Economic Dynamics*. Harvard University Press, Cambridge and London.