Optimal execution in a limit order book and an associated microstructure market impact model*

Costis Maglaras† Ciamac C. Moallemi‡ Hua Zheng §

May 13, 2015

Abstract

We model an electronic limit order book as a multi-class queueing system under fluid dynamics, and formulate and solve a problem of limit and market order placement to optimally buy a block of shares over a short, predetermined time horizon. Using the structure of the optimal execution policy, we identify microstructure variables that affect trading costs over short time horizons and propose a resulting microstructure-based model of market impact costs. We use a proprietary data set to estimate this cost model, and highlight its insightful structure and increased accuracy over conventional (macroscopic) market impact models that estimate the cost of a trade based on its normalized size but disregarding measurements of limit order book variables.

1. Introduction

Modern equity markets have, to a large extent, become computerized technological systems. Market participants, including institutional investors, market makers, and opportunistic investors, interact within today's high-frequency marketplace with the use of electronic algorithms. These algorithms differ across participants and trading styles. At a high level, they dynamically optimize where, how often, and at what price to trade taking into account the state of the exchanges and other real-time market information. Our goal in this paper is to develop models based on queueing theory for the dynamics of an electronic market over short time scales, and to understand how features of the market microstructure impact the execution costs that market participants face.

We will focus on markets that are organized as so-called *electronic limit order books* (LOBs). This is the dominant market structure among, for example, exchange-traded U.S. equities. In an electronic limit order book, traders may provide liquidity by submitting limit orders to buy or sell specific quantities of stock at a specified price, or remove liquidity by sending market orders to

*The second author was supported in part by NSF Grant CMMI-1235023.

†Columbia Business School, Columbia University (c.maglaras@gsb.columbia.edu)
‡Columbia Business School, Columbia University (ciamac@gsb.columbia.edu)
§Columbia Business School, Columbia University (hzheng14@gsb.columbia.edu)

buy or sell at the best available prices. When a market order arrives, it will be matched by the exchange to a contra-side resting limit order. These resting orders are first prioritized by price, and then, within each price level, prioritized by their time of arrival. In this way, each price level can be associated with a queue of resting limit orders that await execution according to a first-in-first-out (FIFO) service discipline, and an electronic limit order book can be naturally modeled as a multi-class queueing system.

A simplified view of a typical portfolio manager is as an agent that makes high-level decisions to buy or sell quantities of securities. The outcomes of these investment decisions are then delegated to a 'trader' that executes them, often making use of a so called 'algorithmic trading' system. These systems are developed internally by large institutional investors or, alternatively, offered as a service by a multitude of banks or brokers. Broadly speaking, such algorithmic trading strategies are designed hierarchically. First, they decide how to schedule the parent order, at a high level, over the course of its execution horizon. For example, if an investor seeks to buy a block of shares over the course of a trading day, this might involve scheduling target quantities for purchase in 5-minute intervals. In this way, the trade scheduling phase involves strategic decisions that consider trade-offs that are realized over minutes or hours. Second, they consider each such sub-interval of the longer horizon, and decide how to execute the target quantity over the sub-interval by dividing it into smaller child orders that are tactically directed to the market either as market or limit orders at optimized price levels and time points. This second phase is often referred to as the micro-trader or slicer, and involves tactical decisions that consider trade-offs on the time scale of seconds to minutes; the queueing delay incurred by limit orders is an important consideration in this step.

An essential input to both the portfolio selection decision as well as the algorithmic trade execution process is the so-called *market impact model*. This model estimates the anticipated cost of a trade and takes into account the adverse effect of one's own trading activity to the price of the security — i.e., how much will the price move against a trader that is buying or selling a block of a specific stock over a specified time horizon. The market impact model depends on the characteristics of the security, such as its liquidity, volatility and typical bid-ask spread, as well as the size and timing of the trade itself. In portfolio construction, a market impact model is often used as a penalty term to capture the trading frictions and resulting costs associated with a portfolio transition. In trade scheduling, it is used in the context of deciding how aggressively to trade — aggressive execution will result in high expected execution costs over shorter trading horizons but reduce execution risk due to exposure to fluctuating market prices. In the micro-trader, a market impact model is used in the tactical optimization of order placement decisions.

In this paper, we first formulate and solve a stylized version of the optimal execution problem faced by the micro-trader described above that takes the form of optimally buying (or selling) a pre-specified quantity of stock over a fixed short time horizon, typically in the order of a few minutes. Then, leveraging the solution of the execution problem, we construct a market impact model that

explicitly takes into account the microstructure information that describes the state and queueing dynamics of the limit order book. Specifically, the key contributions of the paper are the following: (a) We develop a model of the LOB as a multi-class queueing network. Using a fluid (deterministic, mean-field) model of the queueing system, we solve the resulting optimal execution problem, that describes what fraction of the trade quantity will be executed using limit and market orders and at what price levels. (b) Our optimal execution problem yields an estimate for the (optimized) execution costs, which suggests a functional form for a market impact model and identifies relevant microstructure variables (e.g., queue lengths, arrival rates, etc.) that impact trading costs. The microstructure market impact model seems to be novel viz the extensive literature on this topic and to be of practical interest in estimating transaction costs and optimizing trading decisions over short time horizons of the order of a few minutes. (c) Finally, we calibrate the microstructure market impact model using a proprietary data set of algorithmic trades and contemporaneous real-time measurements of limit order book variables. We compare the quality of the statistical fit of the microstructure model to what can be achieved using a typical macroscopic market impact model that estimates costs without consideration of limit order book variables. We find that our microstructure impact model yields a factor of four improvement in out-of-sample explanatory power. We further test the robustness of our model over its specification and over the problem primitives. We find our model has the most explanatory power for larger orders (measured as a percentage of overall volume) and for assets with greater market depth (measured through queues sizes capturing available liquidity). These correspond to settings where our fluid model assumptions are most realistic. Further, we note conventional macro models are also more successful in settings with greater market depth, a fact that seems unobserved thus far in the literature.

Literature review. This paper is related to the growing literature that lies on the interface of queueing and the study of limit order book markets. This connection was first illustrated by Cont et al. (2010); see also Cont and Larrard (2013), Lakner et al. (2013), Blanchet and Chen (2013), Stoikov et al. (2011), and Lakner et al. (2014). Our model builds on Cont et al. (2010), recognizing the multiple price levels in a limit order book can be modeled as a multi-class queue. We work directly with the fluid model representation and do not study the stochastic dynamics of the multi-class queue. The majority of the papers above focus on characterizing the performance of the limit order book, in many cases involving fluid or diffusion approximations. Our emphasis is on optimization of tactical trading decisions, and specifically in optimizing how to execute a block of shares in a limit order book over a predetermined time horizon that is of the same order as that of queueing delays in the order book, and as such modeling of queueing effects becomes important. Related work includes that of Guo et al. (2013), who study a problem of optimizing when to send limit orders and market orders in the market, taking into account, in a stylized manner, the limit order book dynamics but excluding a careful consideration of queueing delays and order cancellation effects. Cont and Kukanov (2013) study the smart order routing problem,

specifically taking into account the fact that there are multiple exchanges to which one can post a limit order, so the control decision becomes how much to post and to which exchange. Our work considers one consolidated limit order book, like Guo et al. (2013), but models explicitly the queueing dynamics, order cancellations, and the ability to trade aggressively on multiple price levels with market orders. Apart from optimizing limit order placement, we find that the optimized routing of market orders over the optimization horizon is an important ingredient that affects the overall execution cost; in particular, it is typically not optimal to send all market orders to trade at the end of the time horizon. The resulting execution cost motivates the microstructure market impact model.

A separate set of papers deal with the longer horizon trade scheduling problem. Bertsimas and Lo (1998) solved this problem when optimizing the expected cost, and Almgren and Chriss (2000) considered the mean-variance criterion; see also Almgren (2003) and Huberman and Stanzl (2005). These papers use a market impact model to capture the cost of the execution expressed as a function of the speed of trading, but do not explicitly model the interaction in a limit order book, or the state variables of the order book. Obizhaeva and Wang (2006), Rosu (2009), Alfonsi et al. (2010) treat the market as one limit order book and use an aggregated and stylized model of market impact to capture how the price moves as a function of trading intensity. These references address the trade scheduling problem, whose longer time horizon allows one to abstract away the queueing effects that are inherent in the limit order book.

Market impact models estimate the expected transaction cost of a trade. They take various functional forms, and typically deconstruct the price impact into temporary and permanent components, and further specify the decay behavior of the temporary contribution. They depend on specific characteristics of the stock as well as the speed of trading, often assumed to be a constant participation rate – e.g., an order executed at 10% participation rate would aim to trade 100 shares for every 1,000 shares traded in the market across all participants. Huberman and Stanzl (2004) showed using a no-arbitrage argument that the permanent price impact must be a linear function of the quantity traded; see also Gatheral (2010). The functional form and decay kernel of the temporary impact term is not as simple to characterize analytically. The simplest assumption treats that decay as being instantaneous. Other alternatives typically allow for exponential or power decay functions. The functional form that specifies the magnitude of the temporary cost is itself typically assumed to be linear or sub-linear function of the speed of trading; stylized analytical arguments and statistical evidence suggest a sub-linear functional form. For example, Chacko et al. (2008) provide empirical evidence that the expected price impact is proportional to the square root of the quantity traded; see also Bouchaud et al. (2008).

We refer to the class of models described above as macroscopic (or "macro") models in the sense that they do not take into account microstructure variables that can be gleaned from the limit order book, and typically try to give cost estimates over long time durations, minutes to hours

to days. These models are typically estimated through large scale cross-sectional regressions based on the realized costs of a proprietary set of algorithmically executed trades. Almgren et al. (2005) describe an econometric approach for that problem, while Rashkovich and Verma (2012) provide important insights that improve the estimation procedure, and allow for more accurate de-trending of the trade data. Moallemi et al. (2014) extend the above approach to include a short term alpha fixed effect associated with the identity of the trader.

In contrast to the above mentioned papers, our analysis proposes a temporary price impact model that explicitly depends on limit order book variables. It is best suited over short time horizons of the order of minutes (the same order of magnitude as that of queueing delays encountered by limit orders until they execute in the market). Recently, Cont et al. (2014) studied a price impact model expressed as a function of the so-called order flow imbalance that measures the difference between events (arrivals, trades and cancellations) on the two sides of the limit order book. Imbalance should be normalized by the queue depth, which is something that emerges in our work as well in capturing the effect of market orders; limit orders have a different relation to depth that we also identify. Cont et al. (2014) did not suggest a model that could be used to explain and predict trading costs, but such an extension may be possible.

The remainder of the paper is organized as follows. Section 2 models the operation of a limit order book as a multi-class queueing system and studies its fluid dynamics. Section 3 states the optimal execution problem. Section 4 characterizes the optimal strategy, on which a microstructure cost function we provide in Section 5 is predicated. Section 6 reports on the empirical performance of our model and provides a comparison with some benchmark models in the literature.

2. The Limit Order Book

An electronic limit order book (LOB) can be modeled as a multi-class queueing system. In broad terms, we will associate queues at each price point where buy or sell limit orders can wait until executed or canceled by the respective traders. We model and track cumulative arrivals of limit orders into the various queues, model the arrival and execution behavior of market orders, and subsequently discuss the dynamics of this queueing system. Figure 1 provides a useful schematic to visualize the various aspects of the LOB.

This paper studies an optimal execution problem and explores how this provides the basis of a microstructure-based transaction cost function. The specific problem that we analyze is one of optimally buying C shares of a security at the lowest possible price over a given time horizon T. In our setting, we typically imagine T to be of the order of a few minutes.

This transient optimal control problem motivates the use of a deterministic fluid model (sometimes known as a "mean field" model) for the evolution of the LOB, where the discrete and stochastic primitive processes (e.g., order arrivals, cancellations) are replaced by continuous and deterministic

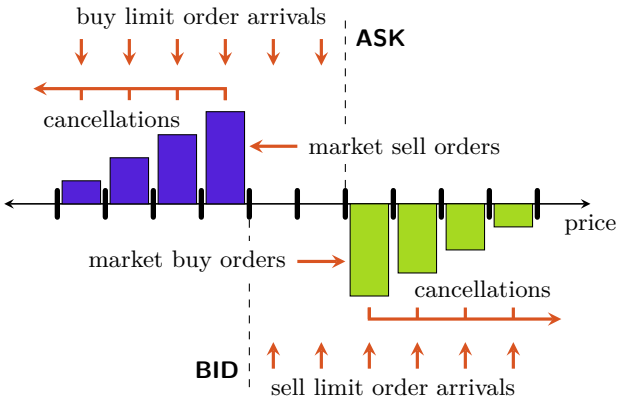

Figure 1: An illustration of an electronic limit order book.

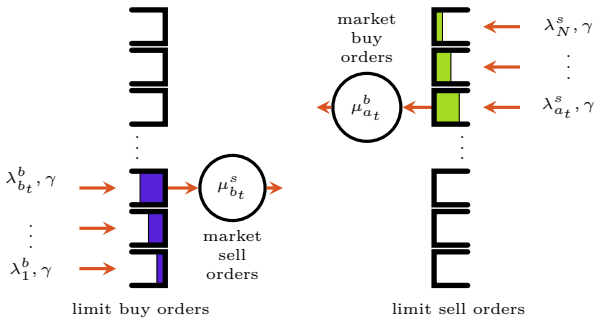

Figure 2: An illustration of the coupled, multi-class priority queueing network associated with an electronic limit order book and its fluid dynamics.

analogues, where infinitesimal orders arrive continuously over time at a rate that is equal to the instantaneous intensity of the underlying stochastic processes. This model can be justified as an asymptotic limit using the functional strong law of large numbers in settings where the rates of order arrivals grow large but the size of each individual order is small relative to the overall order volume over any interval of time.[1] It is well-suited for characterizing transient dynamics in such systems, which roughly correspond to the time scale over which queues drain or move from some initial configuration to an equilibrium state; this is also the relevant time horizon for our optimal execution problem. Indeed, our model is oriented towards liquid securities, where orders arrive on a time scale measured in milliseconds to seconds, while we will consider a time horizon on the order of minutes.

2.1. Multiclass Queueing Network

Our multiclass queueing network model of the LOB is defined as follows:

Prices. We will consider a discrete price grid indexed by $i \in \{1, \dots, N\}$, refer to the ith price point by p_i, and assume that prices are labeled so that $p_1 < p_2 < \cdots < p_N$; it is natural to think that this price sequence is in uniform increments of an underlying minimum tick-size.

Queues. At each price point p_i we associate two queues for buy and sell limit orders, respectively. Specifically, at each time $t \geq 0$, denote by $Q_i^b(t), Q_i^s(t) \in \mathbb{R}_+$ the total quantity of shares available for purchase or sale, respectively, at price level p_i. We define the *best-bid* queue $b_t \in \{1, \dots, N\}$ to be the non-empty queue of buy orders of highest price, i.e.,

$$b_t := \min \left\{ 1 \leq i \leq N \ : \ Q_j^b(t) = 0, \text{ for all } i < j \leq N \right\},$$

and the *best-ask* queue $a_t \in \{1, \dots, N\}$ to be the non-empty queue of sell orders of lowest price,

[1] Mandelbaum and Pats (1995) provides a framework that could be adapted into this setting to justify such a limit.

i.e.,

$$a_t := \max \left\{ 1 \le i \le N \ : \ Q^s_j(t) = 0, \text{ for all } 1 \le j < i \right\}.$$

We denote the overall state of the LOB by $Q(t) := (Q^b(t), Q^s(t)) \in \mathbb{R}^N_+ \times \mathbb{R}^N_+$, where

$$Q^b(t) := (Q^b_1(t), \ldots, Q^b_N(t)) \in \mathbb{R}^N_+ \quad \text{and} \quad Q^s(t) := (Q^s_1(t), \ldots, Q^s_N(t)) \in \mathbb{R}^N_+.$$

We will require that queue length vectors satisfy $b_t < a_t$, or, equivalently, that $p_{b_t} < p_{a_t}$, i.e., the best-bid price is strictly less than the best-ask price. This will be made clearer through the equations of dynamics. Further, we require that both sides of the limit order book be non-empty, i.e., the best bid and best ask levels are well defined and $Q^b_{b_t}(t) \neq 0$ and $Q^s_{a_t}(t) \neq 0$. Denote by $\mathcal{Q} \subset \mathbb{R}^N_+ \times \mathbb{R}^N_+$ the set of such feasible queue length vectors.

Limit order arrivals. *Limit orders* seek to buy (resp., sell) a certain quantity of shares at any price up to and including a limit price that is below (resp., above) the best-bid (resp., best-ask) price in the market.[2] Limit orders cannot be filled upon their arrival, but instead join FIFO queues associated with their limit prices and wait until they are filled or canceled.

Market order arrivals. *Market orders* seek to buy (resp., sell) a certain quantity of shares at the "best" available price. Market orders trade instantaneously against posted limit orders on the contra-side of the order book according to a *price-time* priority rule: when matching a market order to buy (resp., sell) against resting limit orders to sell (resp., buy), the resting orders are first considered in increasing (resp., decreasing) order of price; within each price level, resting limit orders are considered in a first-in-first-out (FIFO) order. The resting limit order shares that are matched to and filled by a market order are subsequently removed from the order book.

Limit order cancellations. Resting limit orders may be canceled at any point. When a cancellation occurs, the canceled shares are removed from their corresponding queue in the order book.

In queueing parlance, a limit order book corresponds to a coupled multiclass queueing network; cf. Figure 2. Job arrivals correspond to the arrival of limit orders, service completions correspond to the arrival of market orders, and abandonments correspond to the arrival of limit order cancellations. The price-time priority rule creates a service discipline where queues are assigned priority classes based on their prices and where each queue is served in FIFO.

2.2. Fluid Model Dynamics

The *fluid model* approximation of the LOB replaces stochastic and discrete arrival and cancellation processes by continuous and deterministic flows.

Limit order arrivals. At time t, we assume that buy and sell limit orders arrive at each price level p_i with rates $\lambda^b_i \cdot \mathbf{1}(i \le b_t)$ and $\lambda^s_i \cdot \mathbf{1}(i \ge a_t)$, respectively, given two vectors $\lambda^s, \lambda^b \in \mathbb{R}^N_+$. In

[2] These are commonly known as non-marketable limit orders. In our setting, limit orders that do not satisfy this price condition (i.e., marketable limit orders) are equivalent to market orders and thus considered as such.

other words, limit orders arrive at price levels that are at the top-of-the-book, i.e., at the current best-bid and best-ask, or at prices inside the book, i.e., buy orders at prices below the best-bid and sell orders at prices higher than the best-ask.[3]

Market order arrivals. Market orders to sell or to buy arrive at rates that are dependent on the current best-bid and best-ask prices, respectively, denoted by $\mu^s_{b_t}$ and $\mu^b_{a_t}$. The two vectors $\mu^s, \mu^b \in \mathbb{R}^N_+$ define the market order arrival rates at different price levels for the best-bid and best-ask, respectively.

Limit order cancellations. We assume that resting limit orders are canceled at a uniform rate $\gamma > 0$, which implies that the cancellation rate per unit time in a queue of size Q is γQ.

Combining the above, we obtain the following ODEs for the order book state process:

$$\dot{Q}^b_i(t) = \lambda^b_i \cdot \mathbf{1}(i \le b_t) - \mu^s_i \cdot \mathbf{1}(i = b_t) - \gamma Q^b_i(t), \qquad \forall\ 1 \le i \le N, \tag{1}$$

$$\dot{Q}^s_i(t) = \lambda^s_i \cdot \mathbf{1}(i \ge a_t) - \mu^b_i \cdot \mathbf{1}(i = a_t) - \gamma Q^s_i(t), \qquad \forall\ 1 \le i \le N. \tag{2}$$

We will make the following assumption regarding the arrival rate parameters:

Assumption 1. *The arrival rate of limit orders at any price level exceeds the arrival rate of contra-side market orders associated with that price level. That is, $\lambda^s_i \ge \mu^b_i$ and $\lambda^b_i \ge \mu^s_i$ for all $1 \le i \le N$.*

The following lemma characterizes the unique stationary point of the fluid dynamics (1)–(2).

Lemma 1. *Given an arbitrary initial condition $Q(0) \in \mathcal{Q}$, there exists a unique solution $Q\colon [0,\infty) \to \mathcal{Q}$ to the fluid model ODEs (1)–(2). This solution satisfies:*

(i) $b_t = b_0$, $a_t = a_0$, for all $t \ge 0$,

(ii) As $t \to \infty$, $Q(t) \to q^*$, where $q^* := (q^{*,b}, q^{*,s})$ is given by

$$q^{*,b}_i := \begin{cases} \lambda^b_i/\gamma & \text{if } 1 \le i < b_0, \\ \dfrac{\lambda^b_i - \mu^s_i}{\gamma} & \text{if } i = b_0, \\ 0 & \text{if } b_0 < i \le N, \end{cases} \qquad q^{*,s}_i := \begin{cases} 0 & \text{if } 1 \le i < a_0, \\ \dfrac{\lambda^s_i - \mu^b_i}{\gamma} & \text{if } i = a_0, \\ \lambda^s_i/\gamma & \text{if } a_0 < i \le N, \end{cases}$$

(All proofs can be found in the Appendix.) Part (i) of Lemma 1 states that starting from any initial condition, the best-bid and best-ask prices remain constant. This is a direct consequence of Assumption 1.[4] Part (ii) of Lemma 1 identifies the long-run equilibrium configuration of the limit order book in terms of the rate parameters and the initial condition.

[3]The rates λ^b_i and λ^s_i are specified as functions of the price level p_i, and and these limit order flows turn off depending on the price level as compared to the prevailing best-bid and best-ask prices. A more complex model would allow for the rates at p_i to depend on the distances of p_i from b_t and a_t, and possibly on the queue lengths, especially these at the best-bid and best-ask. Given our end goal of extracting a transaction cost model which is parsimonious and easily estimable using data, we will not consider these extensions herein.

[4]If Assumption 1 is relaxed, then there may be a short term transient that one would need to consider, e.g., the event rates λ_i, μ_i may be imbalanced in a way that the best-bid or the best-ask would change.

3. The Optimal Execution Problem

We consider a trader that seeks to buy C shares over a given time interval $[0, T]$ by posting limit and market orders over time and at various price levels in the limit order book. The trader's objective is to minimize the average buying price. We describe this problem in detail as follows:

Limit orders. Given Lemma 1, any limit orders posted at price levels p_i with $i < b_t$ (i.e., strictly below the best-bid price) will never trade and can therefore be excluded from consideration, without loss of generality. The following assumption also disallows limit orders strictly above the best-bid price:

Assumption 2 (No Limit Orders Inside Spread)**.** *We restrict attention to execution policies that, at each time t, submit no limit orders at price level i, if $i > b_t$. In other words, no limit orders are submitted inside the current best-bid and best-ask prices.*

We make this assumption for tractability reasons. It disallows the trader from setting a new best-bid price. Under Assumption 2, the limit order placement decision is reduced to selecting how much quantity to submit at the best-bid price level p_{b_t}. In our model, again without loss of generality, we can assume that all limit orders are placed in a single block at time $t = 0$.[5] Thus, we will restrict attention to policies which place all limit orders (if any) at time $t = 0$ at the best-bid price level b_0. We denote by S_L the aggregate size of this limit order, and require that $0 \leq S_L \leq C$.

Market orders. The trader may also place market orders. We denote by $S(t)$ the cumulative number of market orders placed over the interval $[0, t]$.

Assumption 3 (Regularity of Market Orders)**.** *The market order process $S(\cdot)$ must satisfy:*

(i) $S(\cdot)$ is nondecreasing and right continuous with left limits. Denote by $S(t^-)$ the left limit of function $S(\cdot)$ at $t \in (0, T]$ and define $S(0^-) := 0$.

(ii) $S(\cdot)$ has finitely many jump discontinuities and is absolutely continuous on the intervals between jumps.

Given the above assumption, the process $S(\cdot)$ can be rewritten as a combination of discrete jumps or "block" trades, and continuously emitted orders or "flow" trades. Specifically, denote the times of the jump discontinuities by $0 \leq t_1 \leq \cdots \leq t_K \leq T$. Denote by J_k the size of the kth jump or block trade. Then, there exists a Lebesgue integrable instantaneous rate function $r \colon [0, T] \to \mathbb{R}_+$ such that

$$S(t) = \sum_{k=1}^{K} \mathbf{1}\{t_k \leq t\} \cdot J_k + \int_0^t r(s)\, ds, \quad \forall\ t \in [0, T]. \tag{3}$$

[5] We will not provide a proof of that assertion. Intuitively, any policy that submits limit orders at some time $t > 0$ can be weakly improved by submitting the same quantity of limit orders at $t = 0$, which due to the FIFO priority rule, will now execute sooner.

Constraints on the policy. An execution policy is specified via a quantity of limit orders S_L and a market order process $S(\cdot)$ that comprises of block trades $\{J_k\}$ and flow trades $r(\cdot)$.

Definition 1 (Admissible Policy). *Given an initial order book state $Q(0^-) \in \mathcal{Q}$, an execution policy $(S_L, S(\cdot))$ with representation (3) is said to be admissible if it satisfies*

(i) A total of C shares is purchased by the end of the time horizon.

(ii) For each block trade J_k occurring at time t_k, with $k = 1, \cdots, K$, the sizes of block trade does not exceed the available liquidity on the ask side of the order book, i.e.,

$$J_k \leq \sum_{i=a_{t_k^-}}^{N} Q_i^s(t_k^-).$$

Denote by $\mathcal{P}(Q(0^-))$ the set of admissible policies given an initial condition $Q(0^-) \in \mathcal{Q}$. For simplicity, we will further assume that ask queues outside of the best-ask price start at their stationary queue lengths specified in Lemma 1. Specifically:

Assumption 4 (Initial Conditions). *$Q(0^-) \in \mathcal{Q}^{eq}$, where*

$$\mathcal{Q}^{eq} := \{q \ : \ q \in \mathbb{R}_+^N, \ q_i = \lambda_i^s/\gamma \text{ for } i = a_0 + 1, \cdots, N\}.$$

Price movement and the effect on book dynamics. We need to augment the dynamics specified in Section 2, to incorporate the effect of the trader's actions:

(a) Buy market orders submitted by the trader may empty queues on the ask side of the LOB, which would induce a price change in the order book. We will assume that the the order book maintains a constant bid-ask spread after a price shift, formalized in Assumption 5.

(b) Buy limit orders submitted by the trader to the best-bid price must be tracked separately from other limit orders at the best-bid price, so as to maintain their queue position and priority to execute relative to other orders at the same price level. Specifically, the total quantity of buy limit orders $Q_{b_t}^b(t)$ at the best-bid price level at time t can be decomposed as follows

$$Q_{b_t(t)}^b = Q^0(t) + Q_L(t) + Q^1(t),$$

where $Q^0(t)$ is quantity of limit orders still in the queue that were submitted at $t = 0^-$; $Q_L(t)$ is quantity of limit orders still in the queue submitted by the trader at $t = 0$; and $Q^1(t)$ is quantity of limit orders submitted by other participants after $t = 0$. These orders are placed in the queue as illustrated in Figure 3: $Q^0(t)$ is in the front of the queue, followed by $Q_L(t)$ and then by $Q^1(t)$.

The trader's market order policy may deplete price levels on the ask side of the book. Let τ_i be

time when the aggregate queue lengths up to price p_i, for $i = a_0, \ldots, N$, are depleted, i.e.,

$$\tau_i := \inf\left\{t \in [0,T] \mid Q^s_j(t) = 0,\ \forall\ j = 0, \ldots, i\right\}, \tag{4}$$

and set $\tau_i = \infty$ if the condition is not satisfied at any time in $[0,T]$.

Note that we have suppressed the dependence of these times on the initial conditions and the execution policy in our notation. By their definition, $0 \le \tau_{a_0} \le \cdots \le \tau_N$. The best ask process a_t, for $t \in [0,T]$, can be expressed in terms of these depletion times by

$$a_t = a_0 + \sum_{i=a_0}^{N} \mathbf{1}\left\{\tau_i \le t\right\}. \tag{5}$$

The next assumption describes the order book behavior when an ask queue is depleted. We assume that the bid-side queues shift to higher price points as needed to ensure that the bid-ask spread $a_t - b_t$ is constant over time.

Assumption 5 (Constant Bid-Ask Spread). *Denote by $k_t := a_t - a_{t-}$ the price jump at the ask at a time $t \in \{\tau_{a_0}, \ldots, \tau_N\}$. We assume that the bid-side of the book shifts up by the same amount k_t at each such time t. In other words,*

$$Q^b_i(t) = \begin{cases} Q^b_{i-k_t}(t^-) + \mathbf{1}\{t = 0, i = b_0\} \cdot S_L & \text{for } i = 1 + k_t, \ldots, b_t, \\ \lambda^b_i/\gamma & \text{for } i = 1, \cdots, k_t, \end{cases} \tag{6}$$

for $t \in \{\tau_{a_0}, \ldots, \tau_N\}$. Further, queue priority at the best-bid price level is not affected by the price change, i.e., $Q^0(t) = Q^0(t^-),\ Q_L(t) = Q_L(t^-),\ Q^1(t) = Q^1(t^-)$, for $t \in \{\tau_{a_0}, \ldots, \tau_N\}$.

System dynamics. Under Assumptions 1–5, and for an admissible policy the evolution of buy limit orders at the best-bid price are as follows:

$$Q^0(0) = Q^b_{b_0}(0^-), \qquad \dot{Q}^0(t) = \begin{cases} -\mu^s_{b_t} - \gamma Q^0(t) & \text{if } Q^0(t) > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

$$Q_L(0) = S_L, \qquad \dot{Q}_L(t) = \begin{cases} -\mu^s_{b_t} \cdot \mathbf{1}\{Q^0(t) = 0\} & \text{if } Q_L(t) > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

$$Q^1(0) = 0, \qquad \dot{Q}^1(t) = \lambda^b_{b_t} - \mu^s_{b_t} \cdot \mathbf{1}\{Q^0(t) = Q_L(t) = 0\} - \gamma Q^1(t). \tag{9}$$

Specifically, the orders submitted by other participants before $t = 0$ or after $t = 0$ may get canceled at rate γ, whereas the block of orders submitted by the trader at $t = 0$ will not get canceled. At

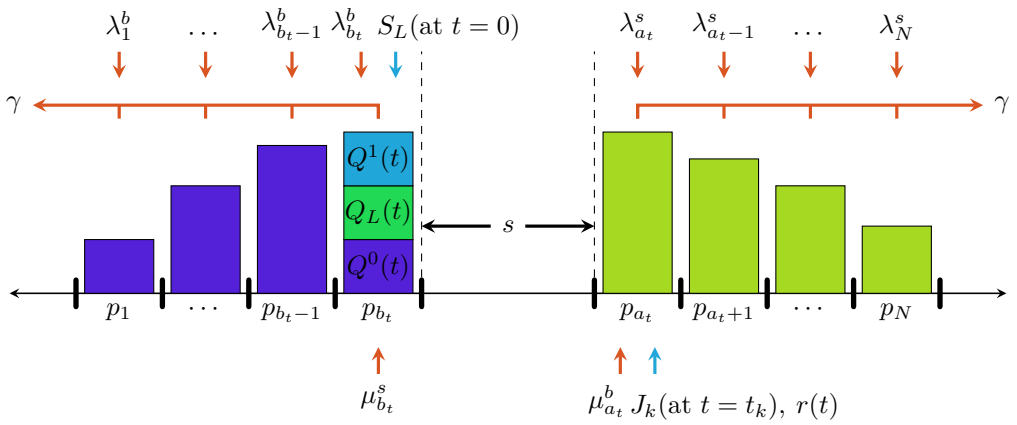

Figure 3: Illustration of system dynamics.

times $t \in \{\tau_{a_0}, \ldots, \tau_N\}$, the bid-side queues will shift price levels according to (6). Further,

$$\dot{Q}^b_i(t) = \lambda^b_i \cdot \mathbf{1}(i < b_t) - \gamma Q^b_i(t) \quad \text{for } 1 \le i < b_t,\ t \notin \{\tau_{a_0}, \ldots, \tau_N\}.$$

The ask-side queues evolve, for $1 \le i \le N$ as follows: for $t \in \{t_1, \ldots, t_K\}$,

$$Q^s_i(t) = \begin{cases} \left(Q^s_i(t^-) - \left(J_k - \sum_{j=a_{t^-}}^{i-1} Q^s_j(t^-)\right)^+\right)^+ & \text{if } i \ge a_{t^-}, \\ 0 & \text{otherwise,} \end{cases}$$

and for $t \notin \{t_1, \ldots, t_K\}$,

$$\dot{Q}^s_i(t) = \lambda^s_i \cdot \mathbf{1}\{i \ge a_t\} - \left(\mu^b_i + r(t)\right) \cdot \mathbf{1}\{i = a_t\} - \gamma Q^s_i(t) \quad \text{for } a_t \le i \le N.$$

Objective function. The optimal execution problem is to pick an admissible policy $(S_L, S(\cdot))$ to minimize the total purchase cost

$$\begin{aligned} P(S_L, S(\cdot)) := &\int_0^T p_{b_t} \cdot \mu^s_{b_t} \mathbf{1}\left\{Q^0(t) = 0, Q_L(t) > 0\right\} dt + \int_0^T p_{a_t} \cdot r(t)\, dt \\ &+ \sum_{k=1}^K \left(\sum_{j=a_{t_k^-}}^{a_{t_k}-1} p_j Q^s_j(t_k^-) + p_{a_{t_k}} \left(J_k - \sum_{j=a_{t_k^-}}^{a_{t_k}-1} Q^s_j(t_k^-) \right) \right), \end{aligned} \tag{10}$$

under Assumptions 1–5, and where the first term is the cost of the executed limit orders, and the second and third terms are the costs due to the flow and block market order trades, respectively.

4. The Optimal Execution Policy

The characterization of the optimal execution policy involves three steps: (a) We identify the execution policy that uses only market orders and minimizes the time needed to fill a target quantity at a given price level. (Lemma 2.) (b) We characterize the optimal execution policy that would complete a target quantity within the specified time horizon again using only market orders. (Lemma 3.) (c) Steps (a)–(b) will ultimately guarantee that the market order execution path will maintain the current price level (b_0, a_0) for all $t < T$, and then push the price at T as needed to complete the target quantity. This property allows us to compute the maximum number of shares that can be executed via limit orders at the best bid, b_0, taking into account the queue priority of orders posted into that best-bid queue prior to $t = 0$ and their respective cancellations over the execution horizon. (Lemma 4.) Jointly these results characterize the optimal policy in Theorem 1.

We first consider the problem of executing in minimum time a target quantity C_{a_0} using market orders only at p_{a_0}, i.e., the (highest priority) best-ask queue that is non-empty at time $t = 0$. In studying this problem we impose the constraint that the queue cannot be depleted prior to finishing the target quantity, and, specifically, that the queue length stays above some arbitrary value $\varepsilon > 0$. This is imposed for mathematical tractability and to guarantee the existence of an optimal policy; without that minimum quantity, the control will strive to take the queue length arbitrarily close to zero, yet without actually depleting the queue that would trigger a price change. This assumption is useful in deriving the structural insight of the next lemma, and will be relaxed later on.

Lemma 2 (Market Orders at One Price)**.** *Without loss of generality we focus at the price level p_{a_0}. Let C_{a_0} be the target number of shares to trade using market orders only at p_{a_0} and let $Q^s_{a_0}(0^-) > 0$ be the initial queue length. Consider the minimum time control problem:*

$$\text{minimize} \quad \{\tau : \ S(\tau) = C_{a_0}\}, \tag{11}$$

over admissible market order control trajectories $\{S(t) : t \in [0, \tau]\}$ that satisfy the following constraints

$$Q^s_{a_0}(t) \geq \varepsilon, \quad t \in [0, \tau) \quad \text{and} \quad S(\tau) - S(\tau^-) \leq Q^s_{a_0}(\tau^-). \tag{12}$$

The optimal control trajectory $\{S^*(t), t \in [0, \tau]\}$ for (11)–(12) is the following:

$$S^*(0) = \begin{cases} Q^s_{a_0}(0^-) - \varepsilon, & \text{if } C_{a_0} > Q^s_{a_0}(0^-), \\ C_{a_0}, & \text{otherwise}, \end{cases} \tag{13}$$

and

$$\dot{S}^*(t) = r^*(t) = \kappa_{a_0}, \quad S^*(t) - S^*(t^-) = 0, \quad \textit{for } t \in (0,\tau), \quad \tau = \frac{\left(C_{a_0} - Q^s_{a_0}(0^-)\right)^+}{\kappa_{a_0}}, \tag{14}$$

where $\kappa_{a_0} := \lambda^s_{a_0} - \mu^b_{a_0} - \gamma\varepsilon$, *and*

$$S^*(\tau) - S^*(\tau^-) = \begin{cases} \varepsilon, & \textit{if } C_{a_0} > Q^s_{a_0}(0^-), \\ C_{a_0}, & \textit{otherwise.} \end{cases} \tag{15}$$

The intuition behind the lemma is simple: we trade as much as possible without depleting the queue at $t = 0$ to avoid the effect of order cancellations at the best-ask queue; if the order is not completed, we trade with a continuous submission of market orders until we fill $C_{a_0} - \varepsilon$ shares; we finish the trade with a small block trade of size ε. Note that the value of κ_{a_0} is such that the queue length will remain constant at ε during $(0, \tau)$. The total duration of the execution is 0 if the target quantity is less than the displayed depth, and is otherwise determined by the length of the interval that is needed to continuously trade at rate κ_{a_0} until the order is completed.

Based on Lemma 2, the length of the execution interval $l_i := \tau_i - \tau_{i-1}$ to execute C_i shares at price p_i, for $i = a_0, \cdots, N$, is

$$l_i = \frac{(C_i - Q^s_i(0^-))^+}{\lambda^s_i - \mu^b_i - \gamma\varepsilon} \approx \frac{(C_i - Q^s_i(0^-))^+}{\kappa_i}, \tag{16}$$

where we redefine $\kappa_i := \lambda^s_i - \mu^b_i$, and the approximation occurs when ε is small; recall that $Q^s_i(0^-) = \bar{Q}^s_i$ for $i > a_0$. We adopt the above approximation for the remainder of this paper. Let $C_{a_0}, C_{a_0+1}, \cdots, C_N$ denote the amount of market orders to execute at prices $p_{a_0}, p_{a_0+1}, \cdots, p_N$, respectively. Given the relationship in equation (16), the optimal execution problem described in Section 3 can be simplified into the following control problem:

$$\underset{S_L, C_{a_0}, \cdots, C_N}{\text{minimize}} \quad \int_0^T p_{b_t} \cdot \mu^s_{b_t} \mathbf{1}\left\{Q^0(t) = 0, Q_L(t) > 0\right\} \, dt + \sum_{i=a_0}^{N} C_i \cdot p_i, \tag{17}$$

subject to

$$S_L + \sum_{i=a_0}^{N} C_i = C, \quad S_L, C_{a_0}, \cdots, C_N \geq 0, \tag{18}$$

$$\int_0^T \mu_{b_t}^s \mathbf{1}\left\{Q^0(t) = 0\right\} dt \geq S_L, \qquad \text{(limit order time)} \tag{19}$$

$$b_t = b_0 + \min\left\{0 \leq j \leq N - a_0 : \sum_{i=a_0}^{a_0+j} l_i > t\right\}, \qquad \text{(limit order dynamics)} \tag{20}$$

$$Q^0(t) \text{ satisfies (7)}, \ Q_L(t) \text{ satisfies (8)}, \text{ for } t \in [0,T], \qquad \text{(limit order dynamics)} \tag{21}$$

$$\sum_{i=a_0}^{N} l_i \overset{(16)}{\approx} \sum_{i=a_0}^{N} \frac{(C_i - Q_i^s(0^-))^+}{\kappa_i} \leq T, \qquad \text{(market order time)} \tag{22}$$

$$C_i \geq Q_i^s(0^-), \quad \text{for } i < n, \qquad \text{(market order dynamics)} \tag{23}$$

$$n = \min\left\{a_0 \leq j \leq N : C_k = 0 \text{ for all } k > j\right\}. \qquad \text{(market order dynamics)} \tag{24}$$

Constraint (19) upper bounds the number of shares that can be traded using limit orders within time T, taking into account the execution priority of limit orders resting in book before time $t = 0$. Constraint (22) ensures that the total time taken trading using market orders at different price levels is upper bounded by the specified time horizon T. Condition (24) identifies the highest price queue in which market orders will be executed, indexed by n, at price p_n, and (23) ensures the time-price priority rule that ensures that all lower priced queues (that have higher priority) will be depleted.

For the remainder the paper we make the following simplifying assumption on κ_i:

Assumption 6. *Assume that $\kappa_i = \lambda_i^s - \mu_i^b = \kappa$ for all i.*

κ_i captures the rate at which the trader can continuously execute with market orders when the best-ask is at price p_i, and without causing a price change. One would expect the continuous trading rate κ_i increases as the price moves up, because more limit orders to sell get submitted at these more favorable price levels. The solution of the optimal execution problem is more involved in that case, and we will not consider it in this paper, given our ultimate interest in specifying a parsimonious microstructure market impact model.

Lemma 3 studies a subproblem of (17)–(24) that seeks to optimize over how to execute C' shares over a time horizon of length T at minimum cost using only market orders, allocated according to $C_{a_0}, \cdots, C_N$ across price levels.

Lemma 3 (Market Orders Across Price Levels). *Given initial queue lengths $Q_{a_0}^s(0^-) > 0$ and $Q_k^s(0^-) = \bar{Q}_k^s$ for $k = a_0 + 1, \cdots, N$ as assumed in Section 3. Consider the problem of minimizing the total*

execution cost of C' shares of market orders over a time horizon of length T

$$
\begin{aligned}
\min_{\{C_k \geq 0,\ k=a_0,\ldots,N\}} \quad & \sum_{k=a_0}^{N} C_k \cdot p_k \\
s.t. \quad & \sum_{k=a_0}^{N} C_k = C', \quad \sum_{k=a_0}^{N} l_k \overset{(16)}{=} \sum_{k=a_0}^{N} \frac{(C_k - Q_k^s(0^-))^+}{\kappa} \leq T \\
& C_i \geq Q_i^s(0^-), \quad \textit{for } i < n, \\
& n = \min\left\{a_0 \leq j \leq N : C_k = 0 \textit{ for all } k > j\right\}.
\end{aligned} \tag{25}
$$

Then, the optimal solution to (25) *is* $\{C_k^*, k = a_0, \ldots, N\}$ *given by*

(26)

$$
C_{a_0}^* = \min\left\{Q_{a_0}^s(0^-) + \kappa T, C'\right\} \textit{ and } C_k^* = \min\left\{Q_k^s(0^-),\ \left(C' - \sum_{m=a_0}^{k-1} C_m^*\right)^+\right\},\ k = a_0+1, \ldots, N.
$$

Under Assumption 6, the above problem admits a simple solution where the trader only applies this continuous submission of market orders at rate κ at the best-ask queue at price a_0, and then submits a block order (as needed) to deplete higher price level queues at T. This is the cheapest price at which the trader can accumulate up to κT shares. A consequence of Lemma 3 is that the best-bid and the best-ask remain equal to (b_0, a_0) for all $t \in [0, T)$, which simplifies the determination of the limit order placement decision, $S_L \in [0, C]$.

Lemma 4 (Limit Orders). *In the optimal solution of problem* (17)–(24),

$$
S_L = \min\left\{\mu_{b_0}^s \left(T - \frac{1}{\gamma}\log\left(1 + \frac{\gamma}{\mu_{b_0}} Q^0(0)\right)\right)^+, C\right\}. \tag{27}
$$

The above expression is intuitive, and crucially depends on the quantity $t_{\text{drain}} := \frac{1}{\gamma}\log\left(1 + \frac{\gamma}{\mu_{b_0}} Q^0(0)\right)$, which is derived from a transient analysis of a fluid queue with abandonments and is equal to the length of time required for the initial queue length $Q^0(0)$ to get depleted either due to cancellations or trades (service completions); this is increasing in the initial queue length and decreasing in the trading rate μ_{b_0} and the cancellation rate γ.

The next theorem characterizes the optimal strategy.

Theorem 1 (Optimal Policy). *Fix the target size $C > 0$, execution horizon $T > 0$, and consider an arbitrary initial condition $Q(0) \in \mathcal{Q}^{eq}$. The optimal execution policy for* (17)–(24) *is the following:*

(a) set the limit order execution quantity S_L according to (27)*;*

(b) for $C' = C - S_L$, set the market order execution quantities $C_{a_0}, C_{a_0+1}, \ldots, C_N$ according to (26)*;*

(c) for $i = a_0$ and C_{a_0} specified above, set the market order execution trajectory $\{S(t) : t \in [0, \tau_{a_0})\}$ according to (13)–(15)*;*

(d) for $i = a_0 + 1, \cdots, N$, according to Lemma 3, $\tau_i = \tau_{a_0} \leq T$. That is, market order executions at higher prices happen with block trades at $t = \tau_{a_0}$. We will refer to this aggregate block as the "cleanup" trade.

In Part (c), the solution uses the infinitesimal $\varepsilon > 0$ to denote the minimum queue length to be maintained in $Q^s_{a_0}$ while submitting a continuous stream of market orders (i.e., service completions) at rate κ.

5. A Microstructure Market Impact Cost Model

In this section, we exploit the solution of the execution problem studied thus far in order to propose a microstructure market impact model. Such a model estimates the trading cost of an order as a function of microstructure limit order book variables, including, for example, real-time measurements of queue lengths and trading rates. We will propose a series of approximations that will yield a parsimonious microstructure market impact model that can be easily and robustly estimated through trade data.

The optimal value of the control problem studied in the previous two sections provides an estimate of the cost of purchasing C shares in T time units. given by

$$\begin{aligned} \text{Total cost} &= p_{b_0} \cdot S_L + p_{a_0} \cdot C_{a_0} + \sum_{i=a_0+1}^{N} p_i \cdot C_i \\ &= (p - s/2) \cdot S_L + (p + s/2) \cdot C_{a_0} + \sum_{k=1}^{N-a_0} (p + s/2 + k\delta) \cdot C_{a_0+k} \\ &= (p + s/2) \cdot C - s \cdot S_L + \sum_{k=1}^{N-a_0} k\delta \cdot C_{a_0+k}, \end{aligned} \tag{28}$$

where p is the arrival price, i.e., the mid-price at the start time of the execution, and the last expression accounts for the execution cost relative to the (contra side or far side) price $p + s/2 = p_{a_0}$. The implementation shortfall, or average purchase price relative to the arrival price, is

$$\overline{IS} := \frac{\text{Total cost}}{C} - p = s/2 - s \cdot \frac{S_L}{C} + \sum_{k=1}^{N-a_0} k\delta \cdot \frac{C_{a_0+k}}{C}. \tag{29}$$

In this formula, the first term accounts for the cost relative to the best-ask price p_{a_0} (the far side), which is half the spread ($s/2$) above the mid-price p. The second term then subtracts the spread for the shares traded using limit orders at the lower price $p_{b_0} = p_{a_0} - s$. The final term adds price increments (a multiple of the tick size) for the higher priced queues that were used in the cleanup

trade. In order to simplify the subsequent empirical analysis, we will make several approximations to the final two terms:

(i) The limit order cost compensation term depends on $S_L = \min\left\{\mu^s_{b_0}(T - t_{\text{drain}})^+, C\right\}$. We will disregard cancellations and approximate the draining time t_{drain} of the orders posted on the near side of the market prior to $t = 0$ by $t_{\text{drain}} \approx Q^0(0)/\mu^s_{b_0}$. Subsequently, we approximate S_L as follows

$$S_L \approx \min\left\{\left(\mu^s_{b_0}T - Q^b_{b_0}(0)\right)^+, C\right\}.$$

(ii) For the cleanup cost term, we will first assume that the stationary queue lengths $\bar{Q}^s_i$, $a_0 < i \leq N$, as defined in Assumption 4, are all equal to some value $\bar{Q}^s$.[6] In that case, it follows from Lemma 3 that $C_{a_0+k} = \bar{Q}^s$ for $0 < k < n$, where

$$n := \left\lceil \frac{(C' - C_{a_0})^+}{\bar{Q}^s} \right\rceil = \left\lceil \frac{(C - S_L - Q^s_{a_0}(0) - \kappa T)^+}{\bar{Q}^s} \right\rceil \tag{30}$$

denotes the number of additional price levels needed in the cleanup trade. We will further simplify the expression by dropping S_L from its calculation, i.e., we set $n \approx (C - Q^s_{a_0}(0) - \kappa T)^+ / \bar{Q}^s$, and subsequently approximate the average price penalty per share due to market order executions relative to the far side to be

$$\frac{\sum_{i=0}^{n} i\delta \cdot \bar{Q}^s}{C_{a_0} + n\bar{Q}^s}. \tag{31}$$

The effect of C_{a_0} diminishes as n increases. When n is large, the average price per share in (31) can further be approximated by

$$\frac{n+1}{2}\delta \approx \frac{\delta}{2} \cdot \frac{(C - Q^s_{a_0}(0) - \kappa T)^+}{\bar{Q}^s} + \frac{\delta}{2}.$$

Combining (i)-(ii), the resulting simplified expression of the implementation shortfall is

$$\overline{IS} = s/2 - s \cdot \frac{\min\left\{\left(\mu^s_{b_0}T - Q^b_{b_0}(0)\right)^+, C\right\}}{C} + \frac{\delta}{2} \cdot \frac{(C - Q^s_{a_0}(0) - \kappa T)^+}{\bar{Q}^s} + \frac{\delta}{2}. \tag{32}$$

This expression depends on the microstructure variables such as trading rates on either side of the book, queue depths, spread, tick size, as well as the trade quantity and time horizon. Specifically,

(a) *Effect of limit orders:* The execution cost is decreasing in S_L, the volume that can be traded using limit orders. The latter is decreasing in the queue length on the near side of the book,

[6] This is certainly an idealization. Typically, one would expect to see the limit order arrival rates λ^s_i increase with price levels i, which then suggests $\bar{Q}^s_i := \lambda^s_i/\gamma$ should also increase with i. Nevertheless, we find in the empirical tests that using a uniform estimate of the stationary queue lengths performs reasonably well.

$Q^b_{b_0}(0)$ (the bid side when buying, or ask side when selling), and is increasing in the the arrival rate of market orders to the near side (market orders to trade against the trader's posted limit orders), and in the execution horizon T. The expression for S_L also indicates that the execution cost will be decreasing in the cancellation rate, although this dependence has been suppressed in the simplified cost formula. The limit order effect is independent of the trade quantity C (assuming the latter is larger than S_L).

(b) *Market order effect at the top-of-book:* This depends on $C - S_L$, the residual quantity to be traded using market orders, and on $Q^s_{a_0}(0) + \kappa T$. The latter is increasing in the displayed depth $Q^s_{a_0}(0)$, the time horizon T, and the continuous trading rate κ that, as discussed earlier, captures the rate at which one can continuously trade with market orders at a given price level without depleting the respective queue and moving the price.

(c) *Market orders at higher prices:* the residual quantity that needs to get executed at higher price levels is decreasing in S_L (see (a)), $Q^s_{a_0}(0)$, κ, and T (see (b)). Its effect is inversely proportional to the equilibrium depth $\bar{Q}^s$ in each of these queues, since that is used to compute the number of price levels n that the trader will have to deplete.

6. Empirical Results

The microstructure market impact model of equation (32) identifies several important microstructure variables that may affect execution costs. While this model was based on a number of simplifying assumptions, it is our belief that these variables are nevertheless important. In order to demonstrate this, in the remainder of this paper, we will calibrate this model using a proprietary dataset of algorithmic trades executed in the US equities market in the third quarter of 2013. Specifically, we will calibrate weights for the different microstructure variables identified in (32) via a regression analysis, and then validate that the resulting microstructure market impact model can help to explain more of the variability in observed trading costs.

Our data set consists of short time horizon slices of executions arising from algorithms based on TWAP, VWAP, and POV[7] policies. The execution logic used in those trades differs from the optimal policy derived in our stylized analysis in Section 4. Nevertheless, our findings will indicate that the microstructure market impact model leads to improved statistical fits, specifically in explaining the realized costs of execution in this dataset (attribution), when compared with conventional "macro" market impact models. Moreover, the coefficients of the explanatory variables postulated by our analysis are significant and have the right signs. The microstructure market impact model also exhibits improved predictive statistical accuracy, e.g., when used to make real-time predictions of future trading costs based on available information at the beginning of each trade.

[7]See, for example, Sotiropoulos (2013) for a description of these policies.

6.1. The Dataset

We use a proprietary dataset of US equities trades from July to September of 2013. This dataset is itself a random sample of a larger set of algorithmic orders executed over that time period. For each parent order (e.g., a full day execution according to the VWAP strategy), the data is summarized in 1-minute intervals. For each such interval we have execution statistics as well as measurements of various limit order book variables. The data has 980,000 active trade records (i.e., 1-minute summaries of execution activity), and represents a sample of 1,800 different securities.

Most of the analysis is performed in rolled-up 5-minute slices. Parent orders that lasted less than 5 minutes or parent order residuals that lasted less than 5 minutes are discarded. Intervals over which there were no executions are also discarded. We further filter according to the following criteria: (a) keep only slices that correspond to VWAP, TWAP, and POV strategies;[8] (b) remove orders for illiquid securities that have an average daily trading volume lower than 300,000 shares; (c) discard the last slice of each parent order to avoid special considerations and cleanup logic associated with the respective algorithmic strategy, apart from POV orders; (d) discard slices in the opening 15 minutes of the trading day, 9:30am–9:45am, and the last 15 minutes of the day, 3:45pm–4:00pm; (e) discard slices for which the realized implementation shortfall exceeds 200 basis points, where the daily volatility within the period exceeds 4%, or where the trade volume exceeded 5 times the volume of the immediately preceding slice; (f) restrict attention to slices with realized participation rate[9] between 1% and 30%. Table 1 reports monthly descriptive statistics of the filtered dataset.

6.2. Calibration of Auxiliary Model Parameters

There are three quantities in the market impact equation (32) that are not directly observable in the data: the equilibrium queue length $\bar{Q}^s$, the effective tick size δ, and the rate of continuous trading κ.

The parameter κ captures the rate at which one can execute with a continuous stream of market orders at the best-ask without causing any price change. Motivated by Assumption 6 and the discussion after it, we will think of κ as a constant multiple of market order rate μ^b. Specifically, we postulate that κ can be expressed in the form of $\theta \cdot \mu$, where μ is the nominal trading rate and θ is a parameter between 0 and 1. We assume that θ is the same on the bid side and ask side of the book, and across all securities.

Returning to our dataset, we identify the set of slices for which: (a) the average queue length on the far side (i.e., the ask when buying and the bid when selling) was small, specifically less than or

[8] Such strategies tend to follow a fairly consistent rate of trading over short periods of time. The composition of the sample after the various filters were applied was roughly uniform across the three strategies and across months.

[9] The participation rate is the ratio of the execution quantity of the slice over the total volume traded in the corresponding time interval by all market participants.

	JUL 2013	AUG 2013	SEP 2013
Sample Size			
5min Slices	27,760	30,054	29,226
Parent Orders	3,396	3,607	3,882
Distinct Securities	988	896	885
Characteristics			
Average Daily Volume (shares)			
mean	3,014,000	2,595,000	2,509,000
3rd quantile	2,585,000	2,689,000	2,626,000
1st quantile	554,300	578,500	544,000
Size of 5min Slices (shares)			
mean	1,294	1,043	849
3rd quantile	1,000	1,000	700
1st quantile	81	100	82
# 5min Slices in Parent Order			
mean	8.2	8.3	7.5
3rd quantile	10	9.5	8
1st quantile	1	1	1
Average Queue Length			
mean	10,280	21,730	17,750
3rd quantile	2,278	4,078	5,148
1st quantile	434	477	536
Realized Participation Rate			
mean	9.60%	9.40%	8.39%
3rd quantile	17.70%	16.20%	14.19%
1st quantile	2.20%	2.26%	1.90%
Price ($)			
mean	46.80	38.16	41.41
3rd quantile	57.41	52.23	51.64
1st quantile	15.35	13.31	13.33
Spread ($)			
mean	0.031	0.025	0.025
3rd quantile	0.032	0.028	0.024
1st quantile	0.010	0.010	0.010
Daily Volatility			
mean	2.23%	1.90%	1.94%
3rd quantile	2.39%	2.31%	2.34%
1st quantile	1.03%	0.97%	0.90%
Implementation Shortfall (bps)			
mean	3.04	3.09	3.48
3rd quantile	7.25	7.86	7.19
1st quantile	(2.62)	(2.53)	(1.84)

Table 1: Descriptive statistics of the filtered dataset, aggregated into 5-minute slices. *Average queue length* represents the aggregated per side, time-averaged queue length at the best-bid or best-ask over the 5-minute interval. *Price* is the average trading price. *Implementation Shortfall (bps)* = (average trading price - arrival price)*side/arrival price$*10^4$; arrival price is the mid-price at the beginning of the respective 5-minute slice. The above are straight arithmetic averages as opposed to volume or notional weighted. (See Section 6.3)

	JUL 2013	AUG 2013	SEP 2013
Critical ratio θ_{month}	0.112 (0.006)	0.104 (0.004)	0.091 (0.006)

Table 2: Estimates of the critical ratio of trading rate to nominal volume for July-September 2013.

equal to 1/3 of the nominal queue length for the corresponding security; and (b) there was no price impact, i.e., the respective price level did not change. For each such slice we know the quantity that was executed as part of that order. We also generate a forecast for the nominal trading rate μ. We first estimate the fraction of the total daily volume that is forecast to trade over the corresponding time interval, and then re-scale by the average daily volume of the corresponding security.[10] The trading rate estimate μ is set equal to half the forecast volume. The ratio of the executed quantity by the slice and of the corresponding forecast provides a point estimate for θ that is normalized relative to stock-specific characteristics. We average these estimates for each month and report the sample estimates together with the standard errors in Table 2. The estimated parameter can be interpreted as follows: over short time durations, one could trade at a rate that is 10% of the bid volume or ask volume, respectively, or, equivalently, at a 5% participation rate while avoiding any price impact. The order of magnitude of this estimate seems plausible but its precise value is likely to be slightly optimistic, especially for less liquid securities as well as securities that trade with few shares at the best-bid and best-ask.

For the equilibrium queue length $\bar{Q}^s$ and the effective tick size δ, we proceeded as follows. Our dataset contains execution information for the trades described earlier, and we also have access to Trade-And-Quote (TAQ) data for each of the securities included in the dataset over the period of July to September of 2013. Our dataset does not include depth of book information, i.e., information about the price levels and the corresponding queue lengths at the price levels that are not at the best-bid and best-ask price levels at a given point in time. As a result we did not have access to information that would allow us to estimate directly the queue length $\bar{Q}^s$, but instead we approximated it as the average of the queue lengths at the best-bid and best-ask, time averaged over the time interval of each 5-minute execution slice. Similarly, the effective tick size δ is meant to capture the change in price necessary to accumulate $\bar{Q}^s$ shares in the limit order book. Since this was not observable, we will use the volatility, σ^* as a proxy for the tick size δ^*; σ^* is the volatility estimate based on intraday data for the time interval of the respective slice and accounts for the strong time-of-day pattern exhibited by the intraday volatility profile.

6.3. Estimation of the Microstructure and "Macro" Market Impact Models

Microstructure Market Impact Model (In-Sample Regressions). We start by estimating the microstructure market impact model in equation (32) using a linear regression analysis. Let IS_k

[10] The forecast makes use of a cross-sectional liquidity profile depicted in Figure 4 in the Appendix.

denote the implementation shortfall of the kth observation (5-minute slice) in the trade data described in Section 6.1. Implementation shortfall is defined as the normalized difference between the average execution price and the arrival price, denoted as P_k and P_k^0, respectively. It is expressed in basis points. The arrival price is defined as the mid-price, i.e., the average between the best-bid and best-ask prices at the start time of the slice. The start and end times include millisecond timestamps. Specifically,

$$IS_k := (P_k - P_k^0)/P_k^0 \cdot d_k \cdot 10^4,$$

where the trade direction $d_k = 1$ for orders to buy and $d_k = -1$ for orders to sell. Normalizing both sides of (32) by the arrival price we get that

$$IS = \frac{1}{2} \cdot s^* - \frac{\min\left\{C, \left(\mu_{b_0}^s T - Q_{b_0}^b(0)\right)^+\right\}}{C} \cdot s^* + \frac{1}{2} \cdot \frac{(C - Q_{a_0}^s(0) - \kappa T)^+}{\bar{Q}^s} \cdot \delta^* + \frac{1}{2} \cdot \delta^*, \tag{33}$$

where $s^* := s/p \cdot 10^4$, $\delta^* := \delta/p \cdot 10^4$ are the normalized spread and tick size, respectively. Define

$$R^L := \frac{\min\left\{C, \left(\mu_{b_0}^s T - Q_{b_0}^b(0)\right)^+\right\}}{C}, \quad R^M := \frac{(C - Q_{a_0}^s(0) - \kappa T)^+}{\bar{Q}^s}, \tag{34}$$

for the price adjustments due to limit order executions and market orders at higher price levels, respectively. Expressions (33)–(34) are written for buy orders. The corresponding expressions for sell orders would replace in the first term $\mu_{b_0}^s$ with $\mu_{a_0}^b$ and $Q_{b_0}^b(0)$ with $Q_{a_0}^s(0)$, in the second term $Q_{a_0}^s(0)$ with $Q_{b_0}^b(0)$ and $\bar{Q}^s$ with $\bar{Q}^b$. We will estimate the following linear model:

$$IS = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot (R^L s^*) + \beta_3 \cdot (R^M \delta^*) + \beta_4 \cdot \delta^*. \tag{35}$$

The regression results can be found in Table 3. We find consistently good performance for our model, represented by the high R^2 values, the fact that the coefficients are all statistically significant, and that the signs of the coefficients are all in line with our predictions. The month-to-month variability is partially due to the modest sample size and variations in the set of securities and parent orders included in our data set as well as variations in market conditions. If, instead of lower bounding the realized participation rate by 1%, we only allowed slices whose realized participation rate was greater than 3%, then the explanatory power of the model increased to an R^2 of 12.30%, 11.94% and 15.45% for July, August and September, respectively.

Benchmark "Macro" Market Impact Model. Most transient market impact models in the literature express the execution cost as a function of the normalized size of the order, expressed as a percentage of the overall volume that trades in the market in the respective time interval, and

	JUL 2013	AUG 2013	SEP 2013
(intercept)			
coefficient	-0.6888***	-0.6941***	-0.5832**
std. error	0.1232	0.1140	0.1076
spread (bps): s^*			
coefficient	0.3187***	0.3905***	0.3950***
std. error	0.0069	0.0077	0.0070
limit order: $R^L s^*$			
coefficient	-0.3027***	-0.3415***	-0.3658***
std. error	0.0107	0.0100	0.0099
add. tick to pay: $R^M \sigma^*$			
coefficients	0.0991***	0.1480***	0.1486***
std. error	0.0234	0.0225	0.0348
tick size: σ^*			
coefficients	2.3238***	1.8508***	2.4290***
std. error	0.1098	0.0997	0.0996
R-squared	9.91%	10.62%	13.48%

Significance: *** p<0.001, ** p<0.01, * p<0.05

Table 3: Monthly linear regression results for microstructure market impact model of (35).

	JUL 2013	AUG 2013	SEP 2013
(intercept)			
coefficient	0.3204***	0.5495***	0.7799***
std. error	0.1238	0.1148	0.1091
(percent of market vol.)$\cdot\sigma^*$			
coefficients	10.3835***	9.0038***	9.5916***
std. error	0.6445	0.6067	0.6922
volatility: σ^*			
coefficients	1.5498***	1.4778***	1.9781***
std. error	0.1127	0.1026	0.1046
R-squared	3.24%	3.02%	3.75%

Significance: *** p<0.001, ** p<0.01, * p<0.05

Table 4: Monthly linear regression of benchmark model in (36) with $\alpha = 1$ (linear).

suggest the use of functions of the form:

$$IS = \beta_0 + \beta_1 \cdot (\text{Percent of Market Vol.})^{\alpha} \sigma^* + \beta_2 \cdot \sigma^*, \tag{36}$$

where typically $\alpha = 0.5$ or 1.[11]

Table 4 and 5 illustrate the quality of these fits. Note that, as for the microstructure market impact model estimate, σ^* is the intraday volatility estimate for the time interval of the respective slice. A simpler model would use a static volatility estimate, prorated to the duration of the slice,

[11] We have examined a finer grid of $\alpha = 0.1, 0.2, \ldots, 1$. The performance does not vary much with the selection of α, and $\alpha = 0.5$ or $\alpha = 1$ oftentimes have the best performance. We focus on explaining the market impact of short duration slices and we will disregard the decay kernel that is sometimes included in transient market impact models.

	JUL 2013	AUG 2013	SEP 2013
(intercept)			
coefficient	0.3235**	0.5480***	0.7839***
std. error	0.1238	0.1148	0.1091
(percent of market vol.)$^{0.5} \cdot \sigma^*$			
coefficients	6.4110***	5.5267***	5.8011***
std. error	0.3913	0.3685	0.4132
volatility: σ^*			
coefficients	0.7626***	0.8033***	1.2844***
std. error	0.1429	0.1320	0.1367
R-squared	3.27%	3.04%	3.77%

Significance: *** p<0.001, ** p<0.01, * p<0.05

Table 5: Monthly linear regression of benchmark model in (36) with $\alpha = 0.5$ (square root).

but independent of the time-of-day. This reduces the explanatory power of the "macro" models from around 3% to about 1%, underscoring the importance of incorporating this effect.

Cross-Validation. Next we compare the out-of-sample performance of our model against that of the benchmark models. We perform a 3-fold cross-validation using the three monthly samples of data from July to September in 2013.[12] We proceed as follows: in each round, we select one monthly sample among the three as the testing data. On the data of the other two months, our model, the linear benchmark model, and the square root benchmark models are fit. Then, the calibrated models are applied to the test set to evaluate how much of the variability in market impact can be explained by each model. Three rounds of training and testing are performed by rotating through the different months as the test set. Finally, the prediction performance of each model takes an average among the three rounds of cross-validation.

When evaluating the out-of-sample accuracy of the different models, we compare their mean squared error with that of the mean predictor to define a generalized R^2 as:

$$\text{generalized } R^2 := 1 - \frac{\text{Mean Squared Error (selected model)}}{\text{Mean Squared Error (mean predictor)}}. \tag{37}$$

There are two candidate mean predictors to use: the mean of the train set, or the mean of the test set. The former is more popular in the literature and has the interpretation that the mean predictor itself is a model that is trained together with other models on the train dataset in each round. In Table 6, we report the average generalized R^2 values based on both mean predictors.

The microstructure market impact model has an average out-of-sample R^2 of around 11%, explaining a factor of 2.5 more of the out-of-sample variability in realized trading costs relative to the "macro" models when compared to the mean predictor; the "macro" market impact models had

[12] Usually a k-fold cross-validation requires dividing all data randomly into equal size subsets. Here we take the natural monthly division of data instead. We expect the result, in particular, the comparison between the two models, be of similar quality when we trisect randomly.

	Model eq. (35)	Benchmark model eq. (36)		Mean predictor
		$\alpha = 1$	$\alpha = 0.5$	
Avg. out-of-sample R^2 (vs. predicted mean)	11.03%	3.11%	3.12%	0.00%
relative improvement	0.00%	255%	254%	Inf
Avg. out-of-sample R^2 (vs. current mean)	10.97%	3.04%	3.06%	-0.08%
relative improvement	0.00%	261%	258%	Inf

Table 6: Average out-of-sample R^2 and relative improvements for a 3-fold cross-validation comparison between our model and the linear/square root benchmark models under two mean predictors. [13]

an average out-of-sample R^2 of around 3.1%. The performance improvement is consistent across the three separate test sets, and, as we will see below, fairly robust to various changes to the way we construct and estimate the microstructure market impact model. The microstructure model treats separately the limit order effect on the execution cost and suggests that measuring trade size as a multiple of queue depth is useful in explaining execution costs. The latter suggests a further segmentation of the data by security characteristics, which we will explore in the next subsection.

The microstructure model adjusts its cost estimate to real-time limit order book conditions, including trading rates on the bid and ask side of the book, and the depths of the best bid and ask queues. To numerically illustrate this feature, we randomly generated 4-tuples for the variables $(Q^b_{b_0}(0), Q^s_{a_0}(0), \mu^b_{a_0}, \mu^s_{b_0})$ to be within a factor of 3 of their nominal values, and evaluated the market impact cost estimate for a trade of size 3 times the nominal depth; we sampled 10 securities of medium ADV and medium depth. The nominal cost is the one that corresponds to the average values of these order book variables. Figure 5 shows that cost estimates generated by the microstructure model may differ by $\pm 50\%$ from the nominal cost, essentially predicting higher costs when conditions are unfavorable, and lower costs when conditions are favorable.

6.4. Robustness Checks

Order & Security Segmentation. First, we grouped the dataset into three sets depending on their realized participation rate. We used the following segments: $[1\%, 10\%]$, $(10\%, 20\%]$, $(20\%, 30\%]$. Table 7 reports the out-of-sample performance[14] of the microstructure model and the linear/square root benchmark models in each of these segments. The microstructure model continues to statistically outperform the "macro" benchmark models for all of these trade groups, but the explanatory power of all models improves as the participation rate increases, since, as expected, in these settings

[13] The above analysis could be repeated to include orders that are traded at lower participation rates, i.e., below 1% which we used as a filter thus far. When including slices with realized participation greater or equal to .25%, the R^2 of the microstructure market impact model drops to 9%; the "benchmark" linear and square root models exhibit an R^2 of about 3%. When we fit a model exclusively to lower participation rates, say in the interval $[.25\%, 1\%]$, the microstructure model explains 4.4% of the realized cost variability, while the benchmark models explain 1% of the variability.

[14] Out-of-sample results in this section are with respect to the predicted mean unless otherwise indicated.

	Model eq. (35)	Benchmark model eq. (36)		Sample size
		$\alpha = 1$	$\alpha = 0.5$	
Percent of market vol.				
[1%,10%]	8.82%	1.87%	1.89%	55,337
(10%,20%]	14.10%	5.34%	5.21%	19,974
(20%,30%]	15.08%	4.23%	4.24%	11,729
overall: [1%,30%]	11.03%	3.11%	3.12%	87,040

Table 7: Out-of-sample performance when clustering by market participation rate.

the statistical signature of the trading slice is likely to be a key driver of the price movement.

Second, following on the observation of the previous subsection, we segmented the trade observations according to the stock characteristics, and specifically, their average daily volume (ADV) and average queue length. We divided the dataset according to the 33% and 66% ADV percentiles, and further segmented according to average queue length at the 30%, 60%, and 90% percentiles. Table 8 reports the out-of-sample results based on these 12 segments of the data. For 9 out of the 12 segments we have enough observations to perform cross-validation tests. Again, within each of these segments, the average out-of-sample R^2 of our model has consistently significant improvement over those of the "macro" models. Moreover, we see (as one would expect) that model accuracy improves as queue depth increases that correspond to settings where the queueing model used in our analysis may be more relevant. The results are qualitatively similar if we segment with respect to queue lengths expressed in notional dollars rather than shares.

Last, we examined the quality of the models in explaining trading costs for less liquid securities, specifically with average daily volumes between 50,000 shares and 300,000 shares. Table 9 reports the out-of-sample performance of the microstructure and benchmark models on the respective sample of the trading data. The explanatory power of all models improves, but so does the relative difference in performance in favor of the microstructure model.

Effect of Nonlinearity. The structural form of the microstructure model involves two non-linear terms that are not a concern when using the model to produce cost estimates or in attributing trade execution performance, but they may affect computational tractability in the context of an optimization model, either for stock selection or for scheduling how to execute a large trade during the course of a longer time horizon. A drastic simplification of the model would remove the non-linearities, as in

$$IS = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot \frac{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)}{C} \cdot s^* + \beta_3 \cdot \frac{(C - Q^s_{a_0}(0) - \kappa T)}{\bar{Q}^s} \cdot \delta^* + \beta_4 \cdot \delta^*. \tag{38}$$

Using this simplified model in (38) in the cross-validation tests, we see that the out-of-sample R^2 of

		Low depth	Mid depth	High depth	Ultra deep	Overall
Model eq. (35)	Low ADV	6.26%	10.23%	17.14%	too few obs.	
	Mid ADV	5.38%	8.12%	12.62%	too few obs.	11.03%
	High ADV	too few obs.	5.56%	10.32%	24.84%	
		Low depth	Mid depth	High depth	Ultra deep	Overall
Model eq. (36)	Low ADV	2.37%	3.28%	5.10%	too few obs.	
($\alpha = 1$)	Mid ADV	2.23%	2.64%	4.62%	too few obs.	3.11%
	High ADV	too few obs.	3.03%	3.84%	6.64%	
		Low depth	Mid depth	High depth	Ultra deep	Overall
Model eq. (36)	Low ADV	2.39%	3.25%	5.13%	too few obs.	
($\alpha = 0.5$)	Mid ADV	2.27%	2.63%	4.59%	too few obs.	3.12%
	High ADV	too few obs.	3.10%	3.90%	6.68%	
		Low depth	Mid depth	High depth	Ultra deep	Overall
Sample size	Low ADV	14,775	9,503	4,589	133	
	Mid ADV	9,712	10,617	8,083	614	87,040
	High ADV	1,625	5,992	13,440	7,957	

Table 8: Out-of-sample performance when clustering by (average daily volume, average queue length).

	Model eq. (35)	Benchmark model eq. (36)		Mean predictor
		$\alpha = 1$	$\alpha = 0.5$	
Avg. out-of-sample R^2 (vs. predicted mean)	23.26%	4.72%	4.91%	0.00%
relative improvement	0.00%	393%	374%	Inf

Table 9: Out-of-sample performance for the sample of securities with low daily volumes.

the microstructure model drops to an average of 8.19%, yet still outperforming the "macro" models; this comparison held across segments of the data by participation rates or security characteristics.

Effect of Time Horizon. The microstructure variables fluctuate over time, and one could expect that the model accuracy depends on the time horizon of the trade slices. Queue length measurements are likely to be more representative over shorter time intervals, but trading rate measurements will be more noisy over short time intervals. Table 10 summarizes our statistical results when instead of using 5-minute trade slices we organize the data sample in 1-minute slices, and illustrate that the statistical significance (out-of-sample) of the microstructure model improves in shorter horizons that may be relevant in the context of dynamic execution algorithms used to optimize over tactical order placement decisions. Tables 11–12 report the out-of-sample performance in segmented data samples of the 1-minute slices, and should be contrasted to Tables 7–8.

The explanatory power of these models improves if one adds lagged residuals of the past two periods (where each residual is the difference between the realized cost and the predicted cost).

	Model eq. (35)	Benchmark model eq. (36)		Mean predictor
		$\alpha = 1$	$\alpha = 0.5$	
Avg. out-of-sample R^2 (vs. predicted mean)	16.57%	2.67%	2.81%	0.00%
relative improvement	0.00%	521%	490%	Inf
Avg. out-of-sample R^2 (vs. current mean)	16.52%	2.61%	2.75%	-0.06%
relative improvement	0.00%	533%	501%	Inf

Table 10: Out-of-sample performance for the sample of 1-min trade slices.

	Model eq. (35)	Benchmark model eq. (36)		Sample size
		$\alpha = 1$	$\alpha = 0.5$	
Percent of market vol.				
[1%,10%]	13.53%	0.94%	0.96%	73,166
(10%,20%]	19.24%	2.26%	2.26%	40,631
(20%,30%]	21.51%	3.59%	3.59%	19,830
overall: [1%,30%]	16.57%	2.67%	2.81%	133,627

Table 11: Out-of-sample performance when clustering by market participation rate (1-min trade slices).

Their respective coefficients are positive and statistically significant, and they seem to capture short-term price momentum. The explanatory power improves by about 2% when explaining realized costs of 1-minute trading slices, and by about 0.6% for 5-minute slices. The "macro" model also improves by about 1% in terms of its explanatory power if one includes the lagged residual variables. One expects that similar improvements may be realized if one included short-term price signals that essentially added a short-term drift component in the regression models.

Cost prediction versus attribution. Market impact models are often used to compute pre-trade cost estimates that may be used as part of a portfolio selection process, or as part of a dynamic trade execution algorithm. In such settings, the models are used to make cost predictions, e.g., at the beginning of a trading slice, and they use information available at that time, as opposed to contemporaneous information that is available in explaining realized costs. This includes snapshots of the queue lengths as well as trailing averages of the queue lengths and the bid side and ask side volume. Specifically, when making a prediction for a trading slice that commences at some time t, we will use exponentially smoothed trailing averages of the relevant limit order book variables computed over the duration of the previous 5-minute (or 1-minute) trading slice. We discard the first slice of each parent order in our dataset when we study the predictive accuracy of the market impact model, since itself was missing prior information needed for the above estimation; this removes 6.5% of the sample of 5-minute trade slices and 5.6% of the sample of 1-minute slices.

Table 13 reports the resulting average out-of-sample R^2 in comparison with the attributive

		Low depth	Mid depth	High depth	Ultra deep	Overall
Model eq. (35)	Low ADV	12.18%	13.81%	23.12%	too few obs.	
	Mid ADV	9.41%	10.84%	18.78%	too few obs.	16.57%
	High ADV	too few obs.	3.91%	20.74%	28.98%	

Table 12: Out-of-sample performance when clustering by (average daily volume, average queue length) (1-min trade slices).

	Model eq. (35)		Model eq. (36) ($\alpha = 1$)		Model eq. (36) ($\alpha = 0.5$)	
	predictive	attributive	predictive	attributive	predictive	attributive
5min slices	8.20%	11.07%	2.26%	2.82%	2.25%	2.84%
1min slices	11.93%	16.80%	1.99%	2.62%	2.27%	2.76%

Table 13: Out-of-sample performance using predictive estimates of average queue length, market volumes, and spread, based on the sample of 5-minute trade slices and the sample of 1-minute trade slices. "Predictive" refers to the model that is using information available at the beginning of each trade slice to estimate its cost. "Attributive" is the model that uses information over the slice, such as the realized participation rate, or the realized bid-side and ask-side volume. The attributive results differ from those in Tables 6–10 due to the additional filtering of the first trading slice of each parent order; similarly in Table 14.

models in Section 6.3. The drop in explanatory power is more significant in the microstructure model as opposed to the macro models, given that the former is using real-time information in a more nuanced way. However, in absolute terms, the microstructure model continues to significantly outperform the two benchmark models. A similar comparison is reported in Table 14 where the various microstructure variables are replaced with historical forecasts, which may be practical in settings where real-time information is not readily available. We use the average monthly queue depth and spread for the bid and ask side queues and the spreads, and we use 1/2 of the forecast interval volume for the bid and ask side rate of market orders. We continue to use the volatility forecast that corresponds to the time interval of each trading slice in our data set.

References

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

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

R. Almgren and N. Chriss. Optimal control of portfolio transactions. *Journal of Risk*, 3:5–39, 2000.

R. Almgren, C. Thum, E. Hauptmann, and H. Li. Direct estimation of equity market impact. *Risk*, July 2005.

	Model eq. (35)		Model eq. (36) ($\alpha = 1$)		Model eq. (36) ($\alpha = 0.5$)	
	historical	attributive	historical	attributive	historical	attributive
5min slices	7.35%	11.03%	2.44%	3.11%	2.56%	3.12%
1min slices	9.54%	16.57%	1.61%	2.67%	1.73%	2.81%

Table 14: Out-of-sample performance using monthly estimates of average queue length, market volumes, and spread, based on the sample of 5-minute trade slices and the sample of 1-minute trade slices.

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

J. Blanchet and X. Chen. Continuous-time modeling of bid-ask spread and price dynamics in limit order books. Working paper, 2013.

J.-P. Bouchaud, J. D. Farmer, and F. Lillo. How markets slowly digest changes in supply and demand. In *Handbook of Financial Markets: Dynamics and Evolution*, pages 57–156. Elsevier: Academic Press, 2008.

George C. Chacko, Jakub W. Jurek, and Erik Stafford. The price of immediacy. *The Journal of Finance*, 63(3):1253–1290, 2008. ISSN 1540-6261.

R. Cont and A. Kukanov. Optimal order placement in limit order markets. Working paper, 2013.

R. Cont and A. De Larrard. Price dynamics in a markovian limit order market. *SIAM Journal of Financial Mathematics*, 4(1):1–25, 2013.

R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58:549–563, 2010.

R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2014.

J. Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10(7):749–759, 2010.

X. Guo, A. De Larrard, and Z. Ruan. Optimal placement in a limit order book. Working paper, 2013.

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

G. Huberman and W. Stanzl. Optimal liquidity trading. *Review of Finance*, 9:165–200, 2005.

P. Lakner, J. Reed, and S. Stoikov. High frequency asymptotics for the limit order book. Working paper, 2013.

P. Lakner, J. Reed, and F. Simatos. Scaling limit of a limit order book model via the regenerative characterization of lévy trees. Working paper, 2014.

A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume 71, pages 239–282. Proceedings of the IMA, 1995.

C. Moallemi, M. Saglam, and M. Sotiropoulos. Short-term predictability and price impact. Working paper, 2014.

A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. Working paper, 2006.

V. Rashkovich and A. Verma. Trade cost: Handicapping on PAR. *Journal of Trading*, 7(4), 2012.

I. Rosu. A dynamic model of the limit order book. *Review of Financial Studies*, 22:4601–4641, 2009.

M. Sotiropoulos. Execution strategies in equity markets. In D. Easley, Marcos Lopez de Prado, and M. OÕHara, editors, *High-Frequency Trading: New Realities for Traders, Markets and Regulators*, pages 21–42. Risk Books, 2013.

S. Stoikov, M. Avellaneda, and J. Reed. Forecasting prices from level-i quotes in the presence of hidden liquidity. *Algorithmic Finance, Forthcoming*, 2011.

A. Proofs

Proof of Lemma 1. Without loss of generality, we consider the evolution of the buy limit order queues $Q^b(t) = (Q^b_1(t), \ldots, Q^b_N(t))$.

For an arbitrary initial condition $Q(0) \in \mathcal{Q}$, the fluid model ODEs in (1) are initialized at $Q^b(0) \in \mathbb{R}^N_+$, satisfying

$$Q^b_{b_0}(0) > 0; \quad Q^b_i(0) = 0 \text{ for all } b_0 < i \leq N.$$

Starting with best-bid b_0 at time $t = 0$, at least for small t, the fluid model ODEs in (1) can be specified as follows:

$$\begin{aligned} \forall 1 \leq i < b_0: &\quad \dot{Q}^b_i(t) = \lambda^b_i - \gamma Q^b_i(t), \\ i = b_0: &\quad \dot{Q}^b_{b_0}(t) = \lambda^b_{b_0} - \mu^s_{b_0} - \gamma Q^b_{b_0}(t), \\ \forall b_0 < i \leq N: &\quad \dot{Q}^b_i(t) = 0, \end{aligned} \tag{A.1}$$

which has unique solution

$$
\begin{aligned}
\forall 1 \le i < b_0: \quad & Q_i^b(t) = \frac{\lambda_i^b}{\gamma}\left(1 - e^{-\gamma t}\right) + Q_i^b(0)e^{-\gamma t}, \\
i = b_0: \quad & Q_{b_0}^b(t) = \frac{\lambda_{b_0}^b - \mu_{b_0}^s}{\gamma}\left(1 - e^{-\gamma t}\right) + Q_{b_0}^b(0)e^{-\gamma t}, \\
\forall b_0 < i \le N: \quad & Q_i^b(t) = 0.
\end{aligned} \tag{A.2}
$$

From (A.2), for $b_0 < i \le N$, $Q_i^b(t)$ will stay at 0. Moreover, since $\lambda_{b_0}^b > \mu_{b_0}^s$ from Assumption 1, $Q_{b_0}^b(t)$ will stay positive and never hit the border $Q_{b_0}^b(t) = 0$. Therefore, $b_t = b_0$ for all $t \ge 0$. Analogously, $a_t = a_0$ for all $t \ge 0$.

As a result, (A.1) holds for all $t \ge 0$. Subsequently, (A.2) is the unique solution to the fluid model ODEs in (1) for all $t \ge 0$.

Since $Q(0) \in \mathcal{Q}$, $b_t = b_0 < a_0 = a_t$ for all $t \ge 0$. And we have shown that $Q_{b_0}^b(t) > 0$, and analogously $Q_{a_0}^s(t) > 0$, for all $t \ge 0$. Hence, $Q(t) \in \mathcal{Q}$ for all $t \ge 0$.

Finally, as $t \to \infty$, $e^{-\gamma t} \to 0$. From (A.2), we have $Q^b(t) \to q^{*,b}$, with $q^{*,b}$ as given in (ii). ■

Proof of Lemma 2. If $C_{a_0} \le Q_{a_0}^s(0^-)$, we have that $\{S^*(t), t \in [0,\tau]\} = \{S^*(0) = C_{a_0}\}$ and it satisfies the constraints in (11) - (12). Executing immediately with one block trade is feasible and thus is the optimal solution to the minimum time problem.

If $C_{a_0} > Q_{a_0}^s(0^-)$, we start with the feasibility of the proposed control trajectory. From (13),

$$S^*(0) = Q_{a_0}^s(0^-) - \varepsilon,$$

and then $Q_{a_0}^s(0) = \varepsilon$. From (14), $\dot{Q}_{a_0}^s(t) = 0$ for all $t \in (0,\tau)$, which guarantees the queue length stays at $Q_{a_0}^s(t) = \varepsilon$. Furthermore, $\dot{S}^*(t) = \kappa_{a_0}$ for the length of the execution interval, which is determined as $\tau = \left(C_{a_0} - Q_{a_0}^s(0^-)\right)/\kappa_{a_0}$. As a result,

$$S^*(\tau^-) = S^*(0) + \int_0^\tau r^*(t)dt = C_{a_0} - \varepsilon.$$

Finally, from (15), we have that $S^*(\tau) - S^*(\tau^-) = \varepsilon = Q_{a_0}^s(\tau^-)$ and $S^*(\tau) = C_{a_0}$.

We prove the optimality of the proposed trajectory by contradiction. Under control trajectory $\{S^*(t), t \in [0,\tau]\}$, we have that $\tau = \left(C_{a_0} - Q_{a_0}^s(0^-)\right)/\kappa_{a_0}$. Suppose there exists another feasible trajectory that executes C_{a_0} shares within time $\tau' < \tau$.

Within time $[0,\tau']$, the total amount of newly arriving sell limit orders into price level p_{a_0} is $\lambda_{a_0}^s \tau'$. From the first constraint in (12), $Q_{a_0}^s(t) \ge \varepsilon$ for all $t \in [0,\tau')$. The total amount of departed sell limit orders from price level p_{a_0} is greater than or equal to

$$\mu_{a_0}^b \tau' + \gamma\varepsilon\tau'.$$

From the constraints in (12), any feasible trajectory can only submit market orders at price level p_{a_0}. Accordingly, the completed number of shares C_{a_0} is constrained by the available liquidity at price level p_{a_0} in the interval $[0, \tau']$, and thus is upper bounded as follows,

$$C_{a_0} \leq Q^s_{a_0}(0^-) + \lambda^s_{a_0}\tau' - \mu^b_{a_0}\tau' - \gamma\varepsilon\tau'. \tag{A.3}$$

As a result, $\tau' \geq \left(C_{a_0} - Q^s_{a_0}(0^-)\right)/\kappa_{a_0} = \tau$, which contradicts with the fact that $\tau' < \tau$. ■

Proof of Lemma 3. If $C' \leq Q^s_{a_0}(0^-) + \kappa T$, we have that

$$C^*_{a_0} = C', \quad C^*_i = 0 \text{ for } i = a_0 + 1, \cdots, N.$$

It is easy to verify that $C^*_{a_0}, \cdots, C^*_N$ is feasible. Furthermore, the resulting total price satisfies

$$\sum_{i=a_0}^{N} C^*_i \cdot p_i = C' \cdot p_{a_0} \leq \sum_{i=a_0}^{N} C_i \cdot p_i,$$

for any feasible $C_{a_0}, \cdots, C_N$, as $p_i \geq p_{a_0}$ for $i = a_0, \cdots, N$.

If $C' > Q^s_{a_0}(0^-) + \kappa T$, we have that

$$C^*_{a_0} = Q^s_{a_0}(0^-) + \kappa T, \quad C^*_i = Q^s_i(0^-) \text{ for } i = a_0 + 1, \cdots, n^* - 1,$$

where n^* is defined as $n^* := \min\left\{a_0 \leq j \leq N : \ \kappa T + \sum_{k=a_0}^{j} Q^s_k(0^-) \geq C'\right\}$, and

$$C^*_{n^*} = C' - \kappa T - \sum_{i=a_0}^{n^*-1} Q^s_i(0^-), \quad C^*_i = 0 \text{ for } i > n^*.$$

In this execution policy, price p_{n^*} will be the highest price at which the trader should submit market orders. It is easy to verify that $C^*_{a_0}, \cdots, C^*_N$ is feasible.

Furthermore, we prove by contradiction that there does not exist an optimal solution with lower total price. Suppose $C_{a_0}, \cdots, C_N$ is such an optimal solution, in which p_n is the highest price to be used by the trader, i.e.,

$$C_i \geq Q^s_i(0^-) \text{ for } i < n, \quad C_n > 0, \quad C_i = 0 \text{ for } i > n.$$

We first show that $n = n^*$. On one hand, if $n < n^*$, from the definition of n^*, we will have

$$\kappa T < C' - \sum_{i=a_0}^{n} Q^s_i(0^-) = \sum_{i=a_0}^{n} \left(C_i - Q^s_i(0^-)\right) \leq \sum_{i=a_0}^{n} \left(C_i - Q^s_i(0^-)\right)^+,$$

which contradicts with the time constraint. Hence, $n \geq n^*$. On the other hand, if $n > n^*$, and at the same time $\sum_{i=a_0}^{n} l_i < T$, then there exists $\eta > 0$ that simultaneously satisfies

$$C_n - \eta > 0, \quad \sum_{i=a_0}^{n} l_i + \frac{\eta}{\kappa} \leq T, \quad \text{and } \eta \cdot (p_n - p_{a_0}) > 0. \tag{A.4}$$

In contrast to the original policy, let the trader submit η less market orders at price p_n, and continuously submit market orders for η/κ time more at price p_{a_0}. The latter policy is still feasible yet has strictly lower price, which contradicts with the fact that $C_{a_0}, \cdots, C_N$ is an optimal solution. Therefore, in this case we should have

$$\sum_{i=a_0}^{n} \kappa l_i = \sum_{i=a_0}^{n-1} \left(C_i - Q_i^s(0^-)\right) + \left(C_n - Q_n^s(0^-)\right)^+ = \kappa T,$$

Subsequently, since $n > n^*$, we have that

$$\sum_{i=a_0}^{n-1} C_i + \left(C_n - Q_n^s(0^-)\right)^+ = \kappa T + \sum_{i=a_0}^{n-1} Q_i^s(0^-) \geq \kappa T + \sum_{i=a_0}^{n^*} Q_i^s(0^-) \geq C'.$$

However, since $\left(C_n - Q_n^s(0^-)\right)^+ < C_n$, the left hand side of the above inequality is strictly less than C', which results in contradiction. Therefore, $n = n^*$.

For the policy $C_{a_0}, \cdots, C_N$, when $n = n^*$, the resulting total price satisfies

$$\begin{aligned}
\sum_{i=a_0}^{N} C_i \cdot p_i &= \sum_{i=a_0}^{n^*-1} \left(Q_i^s(0^-) + \kappa l_i\right) p_i + \left(C' - \sum_{i=a_0}^{n^*-1} \left(Q_i^s(0^-) + \kappa l_i\right)\right) \cdot p_{n^*} \\
&= C' \cdot p_{n^*} - \sum_{i=a_0}^{n^*-1} Q_i^s(0^-) \cdot (p_{n^*} - p_i) - \kappa \sum_{i=a_0}^{n^*-1} l_i \cdot (p_{n^*} - p_i) \\
&\geq C' \cdot p_{n^*} - \sum_{i=a_0}^{n^*-1} Q_i^s(0^-) \cdot (p_{n^*} - p_i) - \kappa T \cdot (p_{n^*} - p_{a_0}) \\
&= \sum_{i=a_0}^{N} C_i^* \cdot p_i,
\end{aligned}$$

which contradicts with the fact that it is an optimal solution with lower total price than that of the solution $C_{a_0}^*, \cdots, C_N^*$. ■

Proof of Lemma 4. Recall that $Q^0(t)$ denotes the quantity of limit orders at the best-bid with higher priority than the trader's order. Its dynamics have been given in (7). Under the assumptions in Section 4, from Lemma 3, we have that $b_t = b_0$ for all $t \in [0, T]$. As a result, until it gets depleted,

the dynamics of $Q^0(t)$ can be simplified to

$$\dot{Q}^0(t) = -\mu^s_{b_0} - \gamma Q^0(t).$$

This ODE has a unique solution for $t \geq 0$ given by

$$Q^0(t) = -\frac{\mu^s_{b_0}}{\gamma} \cdot \left(1 - e^{-\gamma t}\right) + Q^0(0) \cdot e^{-\gamma t}.$$

Thus, the draining time of $Q^0(0)$ is

$$t_{\text{drain}} = \frac{1}{\gamma} \log\left(1 + \frac{\gamma}{\mu^s_{b_0}} Q^0(0)\right).$$

If $T \leq t_{\text{drain}}$, no limit orders submitted by the trader can be executed before the higher priority limit orders get depleted. In this event, $S_L = 0$.

If $T > t_{\text{drain}}$, for $t \in (t_{\text{drain}}, T]$, we have that $Q^0(0) = 0$. Recall that Q^L denote the number of shares left in the trader's limit order. Its dynamics have been given in (8). For $t \in (t_{\text{drain}}, T]$, $\dot{Q}_L(t) = \mu_{b_t}$ if $Q_L(t) > 0$. Therefore, the maximum size of limit order S_L the trader can execute within time $t \in (t_{\text{drain}}, T]$ is $\mu^s_{b_t} \cdot (T - t_{\text{drain}})$.

Moreover, since $S_L \leq C$,

$$S_L = \min\left\{\mu^s_{b_t} \cdot (T - t_{\text{drain}})^+, C\right\}.$$

■

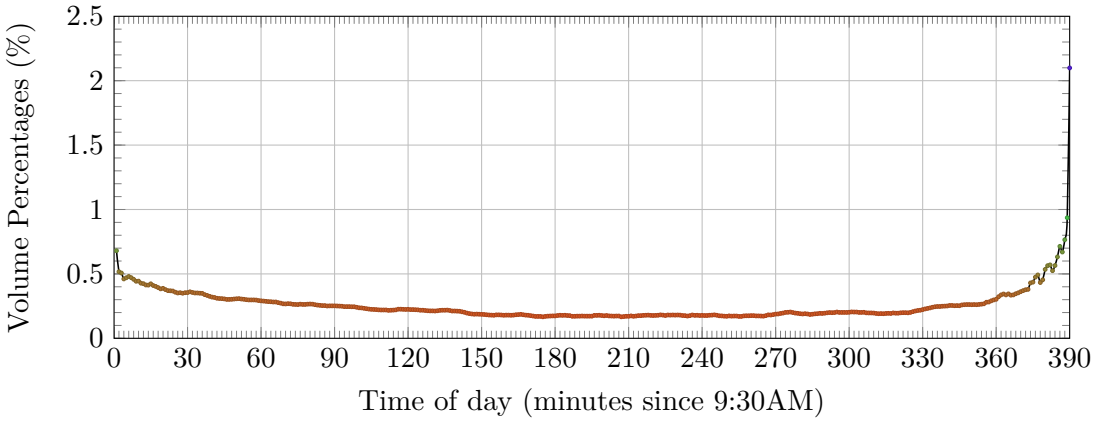

Figure 4: S&P500 cross-sectional, smoothed intraday trading volume profile (min-by-min). Averaged across 5 consecutive trading days. A trading day in the US equities market starts at 9:30am and closes at 4:00pm, i.e., it has 390 minutes. This profile is indicative of "typical" days and it should be adjusted for special occasions such as option expirations, end of month, end of quarter, index rebalancing, Fed announcements, etc.; we do not include that level of granularity in our forecasts but instead apply the typical profile throughout the period of our sample and for all securities, including the ones that are not in the S&P500 and ETFs.

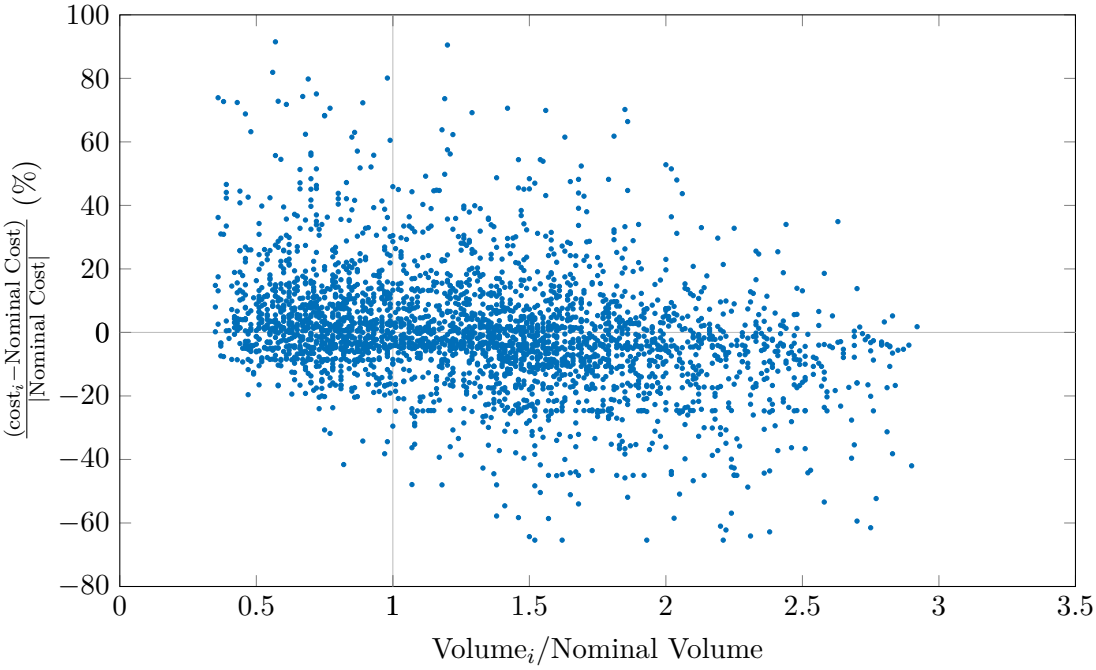

Figure 5: Simulated costs as microstructure variables are varied. Order size = 3 × nominal queue length. Microstructure variables including queue lengths and market order arrival rates vary by a random multiplier in (1/3, 1) w.p. .5 and (1, 3) w.p. .5.