UNIVERSITY OF CALIFORNIA
Santa Barbara

Optimal Execution with Order Flow

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Statistics and Applied Probability

by

Kyle Bechler

Committee in Charge:

Mike Ludkovski, Chair

Jean-Pierre Fouque

Tomoyuki Ichiba

September 2015

The Dissertation of
Kyle Bechler is approved:

Jean-Pierre Fouque

Tomoyuki Ichiba

Mike Ludkovski, Committee Chairperson

July 2015

Optimal Execution with Order Flow

by

Kyle Bechler

To my Wife

Acknowledgements

I first want to extend my gratitude to my advisor and committee chair, Michael Ludkovski. His insight and guidance were invaluable and without his support, producing this research would surely not have been possible.

I would also like to thank my committee members Jean-Pierre Fouque and Tomoyuki Ichiba for their willingness to work with me and for the role they played in increasing my knowledge and appreciation for the subject matter. Additionally, the department of Statistics and Applied Probability faculty were also instrumental in my experience and learning at UCSB.

I am grateful to my parents for their urging me to pursue this research and degree and for their continuing support.

Most importantly, I am thankful for my wife and her unending patience, encouragement and support over the past five years. Her sacrifices and amazing ability to hold our family and life together made it possible for me to pursue this degree.

Curriculum Vitæ

Kyle Bechler

EDUCATION

PhD, Statistics and Applied Probability July 2015
University of California, Santa Barbara

Master's degree in Mathematical Statistics June 2012
University of California, Santa Barbara

Bachelor's degree in Mathematics, minor in Economics May 2005
Westmont College, Santa Barbara, CA

RESEARCH

Stochastic Control, Algorithmic Trading, Empirical analysis of limit order books.

PROFESSIONAL EXPERIENCE

Senior Analyst November 2014 - current
CBRE | Whitestone - Santa Barbara, CA

Portfolio Risk Analyst June 2006 - November 2014
Peritus Asset Management, LLC - Santa Barbara, CA

PUBLICATIONS

K. Bechler, M. Ludkovski. *Optimal Execution with Dynamic Order Flow Imbalance.* Submitted (2015).

EXTRA-CURRICULAR

Member of SIAM 2014 - current
Member of Financial Mathematics and Engineering activity group

Abstract

Optimal Execution with Order Flow

Kyle Bechler

In this thesis we examine optimal execution models that take into account both market microstructure impact and informational costs. Informational footprint is related to order flow and is represented by the trader's influence on the expected order flow process, while microstructure influence is captured by instantaneous price impact. Indeed, a key piece of information missing from many execution models in the literature is the temporal summary of recent order flow which is known to have an impact on the behavior of liquidity providers. Instead, execution and limit order book models often consider only the limited information summarized by a snapshot of the limit order book. Excluded then, are the important mesoscopic trends in market order flow as well as the informational impact made when an executed order perturbs the expected order flow process.

In the following chapters, we propose several continuous-time stochastic control problems that balance between microstructure and informational costs. Incorporating the trade imbalance leads to the consideration of the current market state and specifically whether one's orders lean with or against the prevailing order flow. Several objective functions are treated, as we account for both symmetric and asymetric

execution costs that arise when trading under differing market conditions. We then initiate statistical analysis on Nasdaq limit order book data to investigate the links between market order flow, price impact and liquidity at the mesoscopic timescale. We find that temporal measures of order flow play a key role in the price formation process and show how these features can be incorporated into an execution model for which closed-form solutions can be obtained.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

More than half of the markets in today's financial world are electronic in nature and utilize some form of *limit order book* (LOB) mechanism to facilitate trading. The speed, efficiency and transparency resulting from this market structure means that market participants are provided real-time access to the full LOB, and with it an extensive level of detail. This has contributed to the significant growth of automated or algorithmic trading, which now accounts for the majority of executed orders in many markets. Algorithms are developed via quantitative methods in order to satisfy the particular objective of some market participant. Possible objectives include the optimal liquidation of an asset while minimizing market impact, or providing liquidity

to the market (market making) while controlling for various risks. Once deployed, an algorithm incorporates a variety of real-time inputs from markets and executes orders automatically, without direct human involvement.

In practice, the development of strategies and algorithms relies heavily on the approach and corresponding assumptions in modeling market dynamics. The complexity of today's market micro-structure has led to a wide variety of approaches with ideas from a number of disciplines including economics, statistics, mathematics and physics. Countless contributions have been made to the literature that focus on only a narrow slice of the overall LOB system. Such a limited scope is a necessary concession resulting from the high dimensionality of the LOB, complex interactions between many heterogeneous agents and interconnectedness of numerous exchanges and other trading venues, and other factors.

Work from the economics literature including early work Foucault et al. [29], Parlour et al. [47] and Easley et al. [25] tended to focus on the behavior of individual agents making rational choices, i.e. utility maximization, thus presenting the LOB as a type of sequential game. Not surprisingly, it is often difficult to recover many stylized features of actual markets when following this convention. An alternative approach from a math/physics perspective directly assigns stochastic dynamics to order flows or prices providing a convenient structure from which to analyze the features of the resulting system. The seminal work by Almgren and Chriss [5] for example originated

the framework of diffusive mid-price models, and others including Cont et al. [18], Alfonsi et al. [1] and Cartea et al. [17] have modeled price or order flow directly in lieu of building from the ground up with rational individuals. This second approach, which draws on a wealth of empirical studies into the statistical properties of price and order flow data, is often helpful in reducing the state-space complexity of the LOB and proves convenient for the purpose of designing algorithms.

Generally with a particular problem in view, one proposes a quantitative micro-structure model with features such as price, order arrival or LOB shape governed by stochastic processes, and then standard control techniques are utilized to solve for optimal strategies. There are two main classes of problems. The first is optimal liquidation, which is concerned with realizing the best value for the asset traded via optimization with respect to the incurred trading costs which are highly dependent on the method of execution. A traditional institutional investor with a large position in some asset would be acutely interested in the liquidity profile, average volume and price volatility of the asset over the course of multiple trading days. The second, is the problem facing a liquidity provider. Market making firms providing liquidity for a given asset are concerned with issues such as latency, and intricacies of order flow on a very short time scale on the order of a millisecond or less. The majority of quantitative academic work in algorithmic or high frequency trading are concerned with some variation of one of these two problems. A notable third objective is the

issue of market stability (see work by Kirilenko et al. [37] and Easley et al. [23]). The present work fits within the optimal liquidation context, but also explores in detail the assumptions and previous results from the literature related to the optimal behavior of liquidity providers.

1.2 The Limit Order Book

Prior to the introduction of LOB markets financial transactions took place in quote-driven environments in which a few designated market makers posted their bid and ask levels, respectively those prices at which they were willing to buy or sell. The market makers set the ask price higher than the bid price in exchange for supplying liquidity to the market and therefore assuming the risk of an unwanted long or short inventory position and the risk of *adverse selection*. Adverse selection has traditionally been described as a market maker's encounter with an informed trader, one with better information about the fundamental value of the asset, but can also be described more generally as the instance when the price moves against the market maker following a trade. In a quote-driven market, traders are only able to trade immediately at the posted bid/ask levels. The LOB market structure on the other hand offers all market participants the ability to place buy and sell orders at any price.

Buy Orders		Sell Orders	
Price	Volume	Price	Volume
49.99	200	50.01	130
49.98	220	50.02	240
49.97	190	50.03	230
49.96	140	50.04	180
49.95	170	50.05	150
49.94	180	50.06	200

Table 1.1: Hypothetical limit order book.

A limit order represents the interest to buy or sell a specific quantity of an asset at a specific price. The LOB is defined as the collection of all active limit orders and is discretized in terms of the *tick size*, typically 1 cent for US equities. The state of the LOB changes throughout the trading day as participants place new limit orders, cancel existing limit orders, or as market orders arrive and are matched to active limit order(s) through execution. Table 1.1 shows a snapshot of the synthetic LOB and Figure 1.1 shows the corresponding graphical representation.

While a limit order specifies the agent's desired quantity and price, the timing of the execution is left uncertain if it occurs at all. On the other hand, a market order

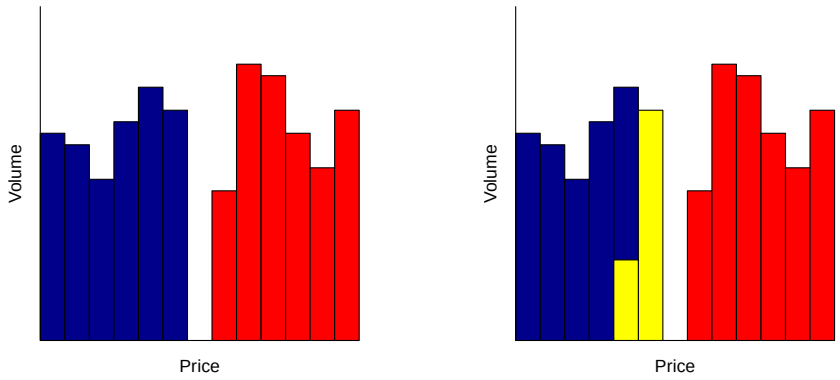

Figure 1.1: Left: Graphical depiction of the limit order book. Right: A market sell order (yellow) is matched against limit orders at the best bid levels.

represents the agent's desire to buy or sell a certain quantity of the asset immediately. Upon arrival of a buy (respectively, sell) order for O shares, the trade-matching mechanism governing the LOB matches the order against the lowest (respectively, highest) priced active limit orders. Figure 2 illustrates one such interaction between a market sell order and the first O shares worth of active limit buy order(s).

For the right to guarantee immediate execution, the agent placing a market order, referred to as *liquidity taker*, receives slightly worse than the mid-price, at least one half of the bid-ask spread, depending on the volume of the order. The bid-ask spread is the difference between the highest priced limit buy order and the lowest priced limit sell order. Conversely, for supplying liquidity to the market, the agent placing the

limit order, the *liquidity provider*, receives slightly better than the mid-price, which in theory is compensation for taking on the associated adverse selection costs.

Beyond the idealized LOB framework presented above, various complicating nuances exist depending on the trading venue. Many exchanges allow participants to place fully or partially hidden orders, subject to certain rules varying by venue. This hidden liquidity is not visible to other participants but can be executed against arriving market orders. Other examples include the method for establishing priority for limit orders with the same price (eg price-time, pro-rata etc) and rebates, in which a small fixed fee is earned or paid for liquidity providers or takers respectively upon an execution. While each of these factors can have significant impact on trading strategies and the profit/loss of market participants, it is not uncommon to ignore many of these issues for the sake of mathematical tractability.

The LOB and related quantities will be defined more carefully in the empirical study in Chapter 3. For the moment it suffices to understand the general framework of the LOB mechanism and note that most problems on the topic are concerned with solving for optimal placement of limit and market orders given a specified model and objective.

1.3 Liquidity and Price Impact

The study of LOB dynamics is the study of trading frictions that are largely ignored by classical mathematical finance models. Though market dynamics clearly operate in discrete quantities, both in time (order arrivals, executions) and space (LOB, tick sizes), it is common to assume that trading, prices or liquidity dynamics are continuous. This approach often simplifies the mathematics and allows for tractable strategies and solutions to be computed via stochastic calculus methods.

For a given asset, the amount of shares available for immediate purchase or sale at a given price, is limited. The result is that trading a large volume of the asset moves the price, usually in an unfavorable direction. The ease with which an asset can be bought or sold is referred to as the liquidity of the asset.

In the LOB context, an asset is said to have a high level of liquidity if a relatively large amount can be purchased or sold immediately (*depth*), the difference between the best bid and best ask prices is small (*spread*), and one can expect to buy or sell a large portion of the asset over time at a price not too much worse than the current price (*resiliency*) (Kyle, [39]). Yet, even for the most liquid of assets, transacting efficiently requires one to consider and model the asset's liquidity dynamics.

There are several approaches in the math-finance literature to modeling liquidity from the perspective of the execution trader. For a large class of stocks, the bid-ask spread remains relatively constant (often at one tick) and is often assumed fixed in

order to simplify the model (See Dayri and Rosenbaum [20]). The depth and resilience of the LOB are typically captured through the model assumptions made related to price impact. Price impact links the volume of an executed order to the resulting change in price, and is often decomposed in the following categories:

Temporary Impact

Temporary or instantaneous price impact consists of the immediate effects of an executed market order. This impact is called temporary because it is assumed that the LOB will replenish quickly leaving the future price of the asset unchanged. Figure 1.1 highlights the consumption of limit buy orders that takes place upon arrival of a large market sell order. Clearly the average price received per share decreases as the order size increases, with the functional relationship dependent on the precise shape of the LOB. It is common for execution models [5] [4] [31] [16] to assume a continuous box shaped LOB (linear impact) which gives rise to the following transaction price

$$\check{P}_t = P_t - \kappa\alpha_t,$$

where P_t is the fundamental price, α_t is the order size and κ is a nonnegative constant. Note that impact in α affects only the current order, and does not affect P_t. The total revenue of the trade is $\alpha_t\check{P}_t$, yielding an execution "cost" of $g(\alpha_t) = \alpha_t(P_t - \check{P}_t) = \kappa\alpha_t^2$. This assumption is roughly consistent with empirical studies which have concluded that g, is well approximated by $g(\alpha) = |\alpha|^{1+\gamma}$, where $\gamma \in [.2, .6]$, see [7],

[41]. Therefore, g must be convex to encourage the splitting of large orders into smaller pieces to be executed incrementally over time to reduce execution costs.

Permanent Impact

Permanent price impact on the other hand refers to the long term effects on the asset price. The size and frequency of market order arrivals conveys information that causes other traders to behave differently in the future. Of course, it is not possible to precisely measure the long-term impact of an action as it would require a comparison between a scenario that happened with one that did not. Studies focusing on the impact of individual trades or metaorders[1] report concave (e.g. the so-called "square root law") impact in the size of the order, see for example Lillo et al. [41], [45]. An alternate approach approximates permanent impact on an aggregate basis by relating trade imbalance, the difference between executed market buy and sell volume, with the corresponding price change over a set time interval. Price and trade imbalance often exhibit a linear relationship [13], [16]. Indeed, within the well known Almgren-Chriss framework, permanent impact must in fact be assumed linear for the model to be free of *price manipulation*[2] [30]. For these reasons and for tractability it is common for optimal execution models to incorporate linear permanent price impact.

[1] A large order executed incrementally over time

[2] Defined as a round trip trade, that is a series of trades with sum zero, with negative expected costs.

Transient Impact

Transient impact refers to price impact that decays over time. In contrast to the above, transient impact evolves over time capturing elements of both temporary and permanent impacts. Implicit in the temporary framework discussed above, is that the LOB recovers instantly to its previous shape and subsequent transactions are not affected. Transient impact relaxes the instant recovery assumption by directly modeling the resilience of the limit order book. Linear transient impact was first considered by Obizhaelva and Wang [46] with several later extensions; Alfonsi et al. [3], Gatheral et al. [32] considered various decay kernels and Alfonsi et al. in [1] allowed for general LOB shape functions.

To sum up, most models that address the "optimal execution of a large investor" problem incorporate one or more of these price impacts. On the other side of the trade is the problem of how best to *supply* liquidity to the market. Several recent studies focus on the optimal order posting strategy for liquidity providers, including Cartea et al. [17] and Guilbaud and Pham [34]. In general, HFT market makers profit through the posting of buy and sell limit orders simultaneously and earning the spread whilst avoiding adverse price movements. A market maker strategy then might consider the depth (relative to the mid-price) or size of her limit order, the duration until cancellation (assuming execution does not occur immediately), and might adapt to the changing state of the LOB or the arriving market order activity. As market

maker strategy determines the shape and behavior of the LOB, understanding key factors that contribute to this strategy are crucial to properly model liquidity.

1.4 Information and Order Flows

It is well established in classic microstucture research [39], [25] that arriving order flows[3] convey information to other market participants. Indeed, Bouchaud et al. [13] argue that endogenous changes in supply and demand, which are manifested in order flows, are more influential in the price formation process than exogenous information. Thus an understanding of how order flow is received and processed by markets and consequently asset prices is a requirement for efficient execution. The two base classes of trades are market orders and limit orders. Market orders indicate actual transactions taking place and hence ultimately drive traders' P&L. Due to their intrinsic nature of "putting money on the table", they carry the most information and are typically viewed as influential by other participants. Limit orders are posted by liquidity providers and serve as reference points in the price discovery process.

Several pathways have been proposed between order flow and liquidity/price movement. First, one-sided market order flow (MOF) is typically associated with a moving market. Indeed, heavy market selling will intrinsically tend to depress prices, mechanically by consuming the top queues of the LOB, and potentially by reveal-

[3]The cumulative volume of buy/sell market and/or limit orders

ing new information about the fundamental value of the asset. As a result, extreme MOF would increase adverse selection and tends to reduce liquidity provision by market makers. Consequently, extreme MOF is expected to lead to scarce liquidity and increased price impact. Consider the following from Easley et al. [27]:

> "...market makers adjust the range at which they are willing to provide liquidity based on their estimates of the probability of being adversely selected by informed traders. Easley, Lopez de Prado and OHara [2012] show that, in high frequency markets, this probability can be accurately approximated as a function of the absolute [*market*] order imbalance...suppose that we are interested in selling a large [*position*] in a market that is imbalance towards sells. Because our order is leaning with previous orders, it reinforces market makers fears that they are being adversely selected, and that their current (long) inventory will be harder to liquidate without incurring a loss. As a result, market makers will further widen the range at which they are willing to provide liquidity, increasing our orders market impact...Thus, order imbalance and market maker behavior set the stage for understanding how orders fare in terms of execution costs."

Market maker strategies must incorporate expected order flows (specifically MOF) and make adjustments in order to account for adverse selection costs or else risk lower profits. A similar result is obtained by Cartea et al. [17], who solve the optimal order placement problem for an HFT market maker. In response to rising adverse selection costs that coincide with increasingly one-sided MOF, the optimal strategy for the HFT market maker is to cancel any current orders and re-post deeper in the LOB.

Second, there are opinions that order flows summarize market sentiment and changing information set. Consequently, market news are encoded in order flow and the latter can be used to measure the influence of a given trade. For example, sudden

"news surprises" might manifest itself in rapid reversal of MOF, which in turn tends to depress liquidity as liquidity providers step back to reduce risk. This concept has been the motivation for defining *toxicity* indices, namely Easley et al. [23],[24], [26].

Third, order flow may indicate manipulative behavior of other participants. For example, the widely cited practice of "order fading" supposedly consists of rapid posting and cancellation of limit orders, to create a mirage of market activity and depth, in order to bait market orders and ultimately generate extra profits from round-trip profits. Similar strategies can be employed to front-run slower traders if the agent has speed advantages. The proliferation of latency arbitrage (see Kirilenko et al. [38]) suggests that these actions are quite profitable. In consequence, regular traders (and market regulators) are urged to monitor order flows to detect such patterns and manipulative actions.

To sum up, order flows can be used to estimate market trends (which would create asymmetric price impact), to detect increased market risk (that increases volatility or symmetric price impact), and to avoid manipulation (which implies that the LOB snapshot is not necessarily indicative of true market state due to latency arbitrage). All of this is crucially important to the execution trader who wishes to efficiently liquidate or acquire many shares via execution of a metaorder. Rather than simply assuming static liquidity or price impact function(s) described above, an execution algorithm ought to consider the typical behavior of HFT market makers and the state

of the liquidity provision process at the time of execution. How the trader's executed metaorder compares to, and assimilates with the existing MOF process appears to be quite relevant to the liquidity and price impact trades will encounter.

1.5 Outline and Contributions

In Chapter 2, taken primarily from the recently submitted paper *Optimal Execution with Dynamic Order Flow Imbalance* [11], the above concepts of information and order flows are bridged within a combined dynamic model. The basics of the Almgren-Chriss setup are maintained, including continuous trading and instantaneous price impact that arises from market microstructure. However, also included is a novel stochastic factor (Y_t) for (expected) market order flow, which is similarly impacted from executed trades. The transient impact on Y represents the informational footprint of the trader and introduces a feedback loop into the problem. It allows the trader to react to changing market conditions, in particular markets changing from being buy-driven to ask-driven and vice versa. The model is used to examine the dynamic problem of *Optimal Execution Horizon* that was introduced in a 1-period version by Easley et al. [27]. Thus, in contrast to Almgren-Chriss and typical execution models, the execution horizon is endogenized, optimally chosen depending on market liquidity.

Intuitively, trading should slow down when informational costs are high, and speed up when they are low. From that point of view, rather than a pure optimization problem, optimal execution is about trading-off price impact and information leakage against timing risk. Moreover, endogenizing execution horizon generates dynamic execution strategies even if the underlying asset value is a martingale. This allows for inherently adaptive trading in contrast to early models with deterministic strategies (e.g. [2, 5, 8]) that consider only price impact. Some more recent extensions by Gatheral and Schied [31] and Almgren and Lorenz [6, 43] do produce dynamic strategies which are "aggressive-in-the-money", accelerating when the price is rising. The approach in Chapter 2 on the other hand adapts to the changing state of liquidity and informational costs.

Chapter 3 then investigates several of the assumptions related to liquidity and execution costs made in [27] and Chapter 2 through an empirical study on Nasdaq data. Much of the recent literature on LOB dynamics operates in the extremely short timescale ($\ll 1$ second), analyzing the predictive power available by conditioning on the state of the LOB (e.g. Huang et al. [35], Donnelly [21], Cont et al. [18] and Lipton et al. [42]. Indeed, part of what makes the analysis of LOBs so challenging are the multiple timescales that apply, ranging from millisecond dynamics up to inter-day trends which span days or weeks. The analysis in Chapter 3 focuses on the minutes timescale at which optimal execution scheduling takes place. At this level, rather

than the state of the LOB, order flows are the main driver in price formation. This chapter aims, through statistical analysis, to understand the link between order flows and liquidity, with a particular focus on periods of scarce liquidity.

Finally, Chapter 4 revisits the optimal execution problem with several of the empirical results from Chapter 3 in mind. When the asset is a liquid stock, the model in Chapter 2 can be refined so that the informational costs related to expected trade imbalance (Y_t) can be made more precise. Where costs are initially left general and not directly mapped to the profit/loss of the trader, in Chapter 4 we explicitly include the asset price and solve the optimal control problem with the trader's wealth process as the objective function. Expected order flow enters the model in two places: (1) Through an additional cost when competing for scarce liquidity, and (2) In price dynamics, capturing the effect on the mid-price of limit order flow and LOB resilience/fading.

Ultimately we find that limit order flow is a key element in the price formation process. In contrast to the picture of a stationary LOB that replenishes following each execution, we instead observe periods of strong resilience (limit additions) and other periods characterized by little to no resistance in the LOB (rapid cancellations). It is shown that limit order flow is strongly linked with the arriving market order flow. Therefore, incorporating the expected trade imbalance Y_t offers key insights into the liquidity provision process yielding a tractable and more realistic execution model.

Chapter 2

Optimal Execution with Expected Trade Imbalance

The concept of optimal execution in financial markets is concerned with realizing the best value for the asset traded via optimization with respect to the incurred trading costs. These costs are driven by two fundamental components: market microstructure frictions and informational asymmetries. Market microstructure implies that market liquidity is finite and trades generate price impact. Informational costs reflect the fact that trades are observed by other participants who will then adjust their own strategies and views of the asset value and create adverse selection.

Specifically, trading in a LOB leaves a double footprint, both in space and in time. In space, an executed sell order consumes the matching standing limit orders on the

bid side. (In our framework since there is no fill risk all trades are assumed to be market orders.) This shortens the respective queues on the bid side of the book and hence can move the best-bid. This is the previously mentioned temporary price impact represented by $g(\cdot)$. Temporally, the executed order is recorded by other participants on the exchange message ticker affecting the observed order *flow*. The effect is both direct (the immediate fact that a sell order of α_t shares was executed), and indirect (the fact that market participants adjust their statistical view of the order flow over time). Thus, to the extent that market participants monitor the ticker (rather than just observe the snapshots of the LOB), an order generates informational footprint. There is a lot of anectodal evidence that many HFT algorithms indeed track the time series of orders placed (for example to detect temporal trading patterns) and hence will react to this footprint. This implies that information costs are at least as important as instantaneous liquidity consumption.

Modelling order flow remains in its early stages. Indeed, while spatially the LOB can be easily summarized as a collection of queues (modeled via say the depth function [2]), the time series of the exchange ticker are much more complicated as market participants process multiple streams of information. There are both executed trades (i.e. trades triggered from market orders) and limit orders, which themselves can be added, cancelled, or modified in other ways depending on the exchange. Orders further carry volume, time stamp and possibly participant type (and limit orders can

be entered at any level of the LOB). These multi-dimensional data is moreover coupled in nontrivial ways, with temporal links both within series (e.g. auto-correlation in the inter-order durations) and across series (e.g. executed market orders tend to increase the arrival rate of limit orders at the touch). See for example a recent model of Cartea et al. [17] who fitted cross-exciting Hawkes processes to the basic order flow at the best-bid and best-ask queues. See also [36] and Cartea and Jaimungal [16] which we revisit in Chapter 4.

In the optimal liquidation literature, very few models consider costs related to temporal order flows and the resulting informational costs. A notable exception is the paper by Easley et al. [27], in which the optimal execution horizon is obtained by minimizing information leakage (the amount by which the trader perturbs the expected market order flow process) subject to timing cost (see Section 2.3.3 for further discussion). The current chapter aims to extend this model [27] to a dynamic, continuous-time framework comparable to popular approaches such as Almgren-Chriss, while capturing both microstructure frictions and informational costs.

2.1 The Optimal Execution Problem

The problem in view is liquidation of a position of size $x_0 = x$. We assume a continuous-time setup, with trading taking place continuously and via infinitesimal amounts. Namely, the trader trades $\dot{x}_t dt$ shares at time t, so that his inventory x_t

follows the dynamics

$$dx_t = \dot{x}_t \, dt. \tag{2.1}$$

Execution ends at the random horizon

$$T_0 := \inf\{t \geq 0 : x_t = 0\},$$

whereupon inventory is exhausted. Throughout, time is supposed to be in traded-volume units.

Beyond the inventory x_t, the main state variable of our model is the expected trade imbalance Y_t. The trade imbalance captures the intrinsic fluctuations among supply and demand for the security realized by the unequal amounts of buyer- and seller-initiated trades. On the short time-scale (intra-day to several days) it is empirically quasi-stationary, in the sense that the observed volume is several orders of magnitude larger than the deviations in net imbalance, cf. Section 2.4.1. Moreover it is highly noisy and appears to be mean-reverting to zero. Therefore, we choose to model (Y_t) in terms of a mean-zero stationary process.

Let Y^0 represent the flow imbalance in the absence of the trader. As a starting point we take (Y_t^0) to be an Ornstein-Uhlenbeck process with mean-reversion parameter β,

$$dY_t^0 = -\beta Y_t^0 \, dt + \sigma \, dW_t. \tag{2.2}$$

The mean-reversion strength β controls the time-scale of the memory in flow imbalance, while the volatility σ controls the size of fluctuations in flow imbalance. Since imbalance is intuitively in the range $[-1, 1]$ (representing markets with 100% buyers, and 100% sellers respectively), the fraction $\sigma^2/(2\beta)$, which is the stationary variance of Y^0, should be on the order of $\sigma^2/(2\beta) \in [0.01, 0.2]$.

The execution program of the trader introduces a downward pressure on the expected trade imbalance process as a result of his selling. The information leaked by the trader's action creates a drift in the realized order imbalance Y_t, pushing it below Y_t^0. By displacing other orders, the trader impacts expectations regarding future order flows and generates adverse selection. A more precise description of this mechanism using (more realistic) discrete-time setup and discrete trades is presented in Section 2.4.3. We postulate that

$$dY_t = -\beta Y_t dt + \phi(\dot{x}_t)dt + \sigma dW_t, \tag{2.3}$$

where $\phi(\cdot)$ captures the information leakage. Observe that

$$Y_t = Y_t^0 + \int_0^t e^{\beta(s-t)}\phi(\dot{x}_s)\, ds, \tag{2.4}$$

so the execution program generates an exponentially decaying impact on Y^0. We assume that $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing with $\phi(0) = 0$. The three main cases we consider are: $\phi(\dot{x}_t) \equiv 0$ corresponding to zero informational footprint; $\phi(\dot{x}_t) = \phi_t$ corresponding to deterministic (but non-zero) impact; and proportional (linear) impact

$\phi(\dot{x}_t) = \eta\dot{x}_t$. Linear impact is computationally convenient, though not necessarily the most realistic reflection of how a trader's activity might influence expectations regarding order flow. Impact that depends on the current value of the flow imbalance is investigated in Section 2.4.3.

Remark 1. Below we restrict our attention to pure selling strategies, so that x_t is non-increasing. As such all our cost functionals are only defined for positive selling rates. Depending on the strength of information leakage, it is possible that the constraint $-\dot{x}_t \geq 0$ is binding, i.e. execution is suspended until market conditions improve. It is beyond the scope of this paper to extend the framework to two-sided trading algorithms that raise the issue of potential market manipulation [3, 32].

The information leakage is "abstract" in the sense that it does not generate trading costs per se. However, in line with [27] we assume that there are adverse selection costs associated with trading in an unbalanced market. Here we assume that this cost is symmetric in Y_t (but note that agent's actions induce only one-sided effects of Y_t); for tractability we take it quadratic. Whether the sign of Y_t should more heavily influence costs is a valid question that we revisit in Chapter 4.

In addition, we carry through two usual costs from the literature. The Almgren-Chriss model [5], detailed in Lemma 2.2.1 below, serves as a baseline strategy in our analysis. This classical model is characterized by two execution costs which we also adopt: instantaneous impact $g(\dot{x}_t)$ of trading at rate $\dot{x}_t$, and inventory cost $\lambda(x_t)$ for

carrying a position of x_t at t. In this chapter we make the assumption (as in [5]) that the asset price is a martingale and therefore does not enter the proceedings. Permanent market impact is modelled via the informational effect on Y_t rather than on the asset price directly.

The continuous-time execution problem is to minimize the sum of the corresponding expected execution costs

$$\inf_{(x_t)\in\mathcal{X}(x)} \mathbb{E}_{x,Y}\left[\int_0^{T_0} \left(g(\dot{x}_s) + \kappa Y_s^2 + \lambda(x_s)\right)\, ds\right], \tag{2.5}$$

over admissible execution strategies $(x_t) \in \mathcal{X}(x)$. The above expectation is conditional on an initial value $Y_0 = Y$ and initial inventory x which induce the measure $\mathbb{P}_{x,Y}$. The horizon T_0 is part of the solution, so that the optimization is formally taking place on the whole future $s \in [0, \infty)$. We assume for the duration that $g(\dot{x}) = \dot{x}^2$. This first term in (2.5) incentivizes the trader to slow down in order to reduce his immediate liquidity costs while the next two terms of the cost functional are such that under certain market conditions, it may be optimal to accelerate trading in order to exit the market sooner.

Let $\mathcal{F}_t = \sigma(Y_s : s \leq t)$ denote the filtration generated by Y. Admissible strategies $(x_t) \in \mathcal{X}(x)$ consist of $(\mathcal{F}_t)$-progressively measurable, absolutely continuous trajectories $t \mapsto x_t$, such that $x_0 = x$, $\lim_{t\to\infty} x_t = 0$ and $\int_0^\infty \dot{x}_s^2\, ds < \infty$ $\mathbb{P}$-a.s. We also require the following assumptions on the model ingredients:

- Instantaneous impact function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is strictly convex;

- The informational cost parameter $\kappa \geq 0$;

- Inventory risk $\lambda : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is non-decreasing in x.

The cost functional in (2.5) is consistent with other approaches taken in the literature. The assumption that $g(\dot{x})$ is convex matches the empirical fact that market participants like to divide a large "parent" order into smaller orders in order to reduce trading costs. In LOBs, $g(\cdot)$ represents the depth of the LOB on the ask-side. If this depth is constant, the instantaneous trading cost is quadratic $g(\dot{x}) = \dot{x}^2$ (by rescaling κ and λ we assume without loss of generality that the leading coefficient is 1). This assumption also appears in [5, 15, 31] among others. Because our primary focus is on informational costs, we assume for the moment that there are no other transient/permanent impacts on the asset value P_t and further posit that strategies are independent of asset dynamics. In Section 2.4.2 we return to this issue and discuss extensions that allow for positive correlation between asset price dynamics (P_t) and order flow (Y_t).

Our second cost term κY_t^2 is motivated primarily by the model presented in [27] and captures the cost of information leakage. The main premise is that liquidity costs (e.g. likelihood of adverse selection) are higher in markets with unbalanced flow. Unlike [27], our model incorporates stochasticity and mean reversion in the flow imbalance process (Y_t), meaning that the impact modelled with this term is transient.

Other models incorporating transient impact have focused on LOB resilience [2, 32]. Also see [4] for another example of stochastic liquidity costs.

The last term $\lambda(x_t)$ in (2.5) represents timing risk, penalizing the trader for leaving his position exposed to adverse price movements. Several risk terms have been applied within the execution literature. The seminal work by Almgren and Chriss, which optimizes over a mean-variance cost functional, reduces to a calculus of variations problem and the risk term $\lambda(x) = cx^2$. Gatheral and Schied [31] investigated a time-weighted value-at-risk measure proportional to $\lambda(x) = cx$. In the context of timing risk, $\lambda(x) = c$ generates costs that are proportional to execution time which is a non-trivial modification once the horizon T_0 is not fixed.

Remark 2. Other authors refer to order imbalance for a different object, namely "spatial" order imbalance. Namely, motivated by queueing notation, they mean the net difference between standing limit orders at the best-bid and best-ask. As shown by [42] and [21], order imbalance is predictive of the next price move (i.e. correlated with the probability of the next price to be an up-tick or a down-tick) and is closely monitored by most HFT algorithms. In this chapter we focus on the temporal order flow and its associated imbalance that was conjectured by Easley et al. [23, 24, 27] to be related to market toxicity. While the LOB depth is related to the history of submitted limit orders, the relationship is highly complicated (due to shifting mid-price, hidden orders, etc.). Consequently, our trade imbalance is not meant to be tied directly to

the LOB depth or any immediate LOB properties, but rather provide a temporal summary of recent orders submitted. Further, we make the important distinction that process (Y_t) is the expected trade imbalance, that is, the trade imbalance that market participants are expecting. This quantity is closely related to the observed trade imbalance that summarizes recent order flows, but the exact relationship between the two is difficult to pinpoint. Section 2.4.1 discuss further.

2.1.1 HJB Formulation

To minimize (2.5) we adopt the standard stochastic control approach, utilizing the dynamic programming principle and Hamilton-Jacobi-Bellman (HJB) PDE. Within this framework strategies are defined by their rate of selling, $\alpha_t := -\dot{x}_t$ and the class of admissible strategies $\mathcal{A}(x)$ consists of all nonnegative $(\mathcal{F}_t)$-progressively measurable processes $(\alpha_t)_{0\leq t\leq T_0}$ for which

$$x_t^\alpha := \left(x - \int_0^t \alpha_s \, ds\right)_+, \quad 0 \leq t,$$

belongs to $\mathcal{X}(x)$. The value function of our problem can be expressed as

$$v(x, Y) = \inf_{(\alpha_t)\in\mathcal{A}(x)} \mathbb{E}_{x,Y}\left[\int_0^{T_0} \left(g(\alpha_s) + \kappa Y_s^2 + \lambda(x_s^\alpha)\right) \, ds\right]. \tag{2.6}$$

If it exists, we define the corresponding optimal strategy as $\alpha^*(x, Y)$. For each path of underlying Y_t^0, α^* induces the realized execution horizon $T_0(x, Y) = \inf\{t : x_t^{\alpha^*} = 0\}$ which is a random variable taking values in $[0, \infty)$.

One key point of interest is the realized execution horizon that results from the optimal dynamic strategy, which can only be obtained by solving (2.6). To this end, we assume that v is sufficiently smooth, and apply the dynamic programming principle which says that

$$t \mapsto v(x_t^\alpha, Y_t) + \int_0^t \left(\alpha_s^2 + \kappa Y_s^2 + \lambda(x_s^\alpha)\right)\, ds$$

ought to be a submartingale for all α and a martingale when α is optimal. Then, an application of Itô's formula suggests that the value function $v(x, Y)$ will satisfy a Hamilton-Jacobi-Bellman equation of the form

$$0 = \frac{1}{2}\sigma^2 v_{YY} - \beta Y v_Y + \kappa Y^2 + \lambda(x) + \inf_{\alpha \geq 0}\{g(\alpha) - \alpha v_x - \phi(\alpha) v_Y\}, \tag{2.7}$$

with the boundary condition $v(0, Y) = 0$ for all Y. We observe that (2.7) is a nonlinear parabolic PDE in (x, Y) for which the corresponding theory (e.g. regarding existence of classical solutions) is rather limited.

When instantaneous price impact cost is quadratic $g(\alpha) = \alpha^2$ (assumed for the remainder) and information leakage is linear $\phi(\alpha) = \eta\alpha$, the candidate optimizer in feedback form is

$$\alpha^*(x, Y) = \frac{v_x + \eta v_Y}{2}. \tag{2.8}$$

Substituting this feedback control into the PDE (2.7) we have

$$0 = \frac{1}{2}\sigma^2 v_{YY} - \beta Y v_Y + \kappa Y^2 + \lambda(x) - \left(\frac{v_x + \eta v_Y}{2}\right)^2. \tag{2.9}$$

Due to the state dependence of the class of admissible strategies $\mathcal{A}(x)$, the problem (2.6) is a finite-fuel control problem. As a result, there does not appear to be a tractable closed form solution which satisfies the zero boundary condition along $x = 0$. In Section 2.5 we illustrate a relatively straightforward method for solving (2.9) numerically via a finite difference scheme.

To understand the feedback strategy in (2.8), we pause to consider the derivatives v_x and v_Y. As we will see, v_x is always positive but v_Y can be either positive or negative. Consequently, the candidate in (2.8) may fail to be non-negative.

Lemma 2.1.1. *The map $x \mapsto v(x, Y)$ is strictly increasing for any Y.*

Proof. Fix $x < x' = x + \epsilon$ for a strictly positive ϵ and consider an (ϵ-optimal) strategy α^ϵ for $v(x', Y)$. Let $T_x := \inf\{t : x'_t = \epsilon\}$ be the random period to sell x shares using α^ϵ. Then by absolute continuity of $t \mapsto x'_t$, $T_x < T_0(x', Y)$. Moreover, $\alpha'(x, Y) := \alpha^\epsilon_t(x', Y)1_{\{t \leq T_x\}}$ is an admissible strategy for the initial conditions (x, Y) since it liquidates exactly $x' - \epsilon = x$ shares. Using the fact that α' is sub-optimal for $v(x, Y)$ and that the second and third terms in (2.5) are strictly positive almost surely, we find $v(x, Y) \leq v(x', Y; \alpha') < v(x', Y)$. □

In (2.9), the horizon is indefinite and ultimate liquidation is only modelled through the boundary condition. Thus, understanding the realized execution horizon $T_0(x, Y)$ is only possible implicitly. Moreover, the nonlinearities in (2.9) make analysis intractable. To achieve tractability we consider an approximate two-stage procedure.

Thus, we first fix a horizon T by imposing the constraint $T_0 = T$. We then solve the resulting fixed-horizon problem to find the best strategy $\alpha^*(T, x, Y)$ and value function $v(T, x, Y)$. In the second step, we optimize over T, to find the statically optimal horizon $T^*(x, Y)$. Finally, we build the semi-dynamic strategy $\tilde{\alpha}(x_t, Y_t) = \alpha^*(T^*(x_t, Y_t), x_t, Y_t)$. Thus, $\tilde{\alpha}$ recomputes T^* as the state variables (x_t, Y_t) evolve and uses the corresponding static trading rate. This approach is analogous to the receding-horizon setup in nonlinear control [48]. Indeed, the initial use of $\alpha^*(T^*(x, Y), x, Y)$ at $t = 0$ corresponds to model predictive control and $\tilde{\alpha}(x_t, Y_t)$ then continuously rolls the initial condition because of the stochastic fluctuations encountered. The above plan is implemented in Sections 2.2 and 2.3 respectively. In the latter section we also compare the execution trajectories and resulting costs from the various strategies.

2.2 Linear Quadratic Setup on Finite Horizon

Fix $T < \infty$. We consider the analogue of (2.6) on $[0, T]$. To avoid confusion we let u denote the value function when defined on the fixed horizon:

$$u(T, x, Y) = \inf_{(\alpha_t) \in \mathcal{A}(T,x)} \mathbb{E}_{x,Y} \left[\int_0^T \alpha_s^2 + \kappa Y_s^2 + \lambda(x_s^\alpha)\, ds \right]. \tag{2.10}$$

For expository purposes, we use the time-to-maturity parametrization for u so that the first argument T represents time *until* the deadline. Strategies on $[0, T]$ are defined in similar fashion to those in Section 2.1 but with a constraint $x_T = 0$ at the terminal

time T. Forced liquidation by T is achieved by leveling an infinite penalty if not completed, leading to a singular initial condition of the form

$$\lim_{T\downarrow 0} u(T,x,Y) = \begin{cases} 0 & \text{if } x = 0 \\ \\ +\infty & \text{if } x \neq 0. \end{cases} \tag{2.11}$$

To obtain explicit solutions to (2.10), the next section treats the case in which ϕ is independent of α over a fixed horizon. In other words, the trader may or may not impact the order flow process, but any impact can be modelled in a deterministic fashion. Section 2.2.2 then addresses the proportional footprint $\phi(\alpha) = \eta\alpha$ case, still over fixed time horizon. It will be shown that the strategies obtained in Sections 2.2.1-2.2.2 are not too suboptimal compared to the indefinite-horizon model laid out in Section 2.1.

2.2.1 Myopic Execution Strategies

In classical optimal execution models [2], [5] and [8], optimal execution rates are deterministic, i.e. α_t is pre-determined. In this scenario, informational costs would also be deterministic. Therefore, we examine the case where $\phi(\alpha)$ is independent of α (but possibly depends on time t). So (Y_t) takes on the dynamics

$$dY_t = -\beta Y_t dt - \phi_t dt + \sigma dW_t.$$

Under this assumption, we can separate the two terms in (2.5) since the dynamics of (Y_t) are not directly affected by the trader; the performance criterion simplifies to

$$\inf_{(x_t)\in\mathcal{X}(x)} \left(\int_0^T \dot{x}_s^2 + \lambda(x_s) ds \right) + \int_0^T \kappa \mathbb{E}_Y[Y_s^2] 1_{\{x_s>0\}} ds. \tag{2.12}$$

Because (Y_t) is independent of the control α, optimal strategies are defined only by the first term in (2.12). Consequently, the resulting (α_t) is independent of Y_t and hence $t \mapsto x_t^*$ is deterministic. Thus, strategies based on (2.12) are myopic in the sense that they entirely ignore the potential "footprint" left by the trader's actions, instead focusing solely on instantaneous cost and inventory risks. The following Lemma provides the solution to (2.12) for popular choices of inventory risk.

Lemma 2.2.1. *Consider the calculus-of-variations problem of finding*

$$\mathcal{I}(T,x) := \inf_{(x_t)} \int_0^T (\dot{x}_t^2 + \lambda(x_t))\, dt$$

where the minimization is over all absolutely continuous curves $t \mapsto x_t$ with $x_0 = x$, $x_T = 0$ and under the constraint that x_t is non-increasing. Then the optimal "myopic"

strategies (with $\alpha_t \equiv -\dot{x}_t$) are

$$\left\{\begin{array}{c} x_t^{ML} = \dfrac{x(T-t)}{T}; \\ \alpha_t^{ML} = \dfrac{x}{T}; \\ \mathcal{I}^{ML}(T,x) = \dfrac{x^2}{T}, \end{array}\right\} \quad \textit{if } \lambda(x) = 0; \tag{2.13}$$

$$\left\{\begin{array}{c} x_t^{MH} = \dfrac{x\sinh(\sqrt{c}(T-t))}{\sinh(\sqrt{c}T)}; \\ \alpha_t^{MH} = \dfrac{\sqrt{c}x\cosh(\sqrt{c}(T-t))}{\sinh(\sqrt{c}T)}; \\ \mathcal{I}^{MH}(T,x) = \sqrt{c}x^2\coth(\sqrt{c}T), \end{array}\right\} \quad \textit{if } \lambda(x) = cx^2; \tag{2.14}$$

$$\left\{\begin{array}{c} x_t^{MQ} = \left(\dfrac{ct^2}{4} - t\left(\dfrac{c\hat{T}}{4} + \dfrac{x}{\hat{T}}\right) + x\right)\mathbb{1}_{\{t<\hat{T}\}}; \\ \alpha_t^{MQ} = \left(\dfrac{c\hat{T}}{4} + \dfrac{x}{\hat{T}} - \dfrac{ct}{2}\right)\mathbb{1}_{\{t<\hat{T}\}}; \\ \mathcal{I}^{MQ}(T,x) = \left(-\dfrac{c^2\hat{T}^3}{48} + \dfrac{c\hat{T}x}{2} + \dfrac{x^2}{\hat{T}}\right); \\ \textit{where } \ \hat{T} := \min(T, \dfrac{2\sqrt{x}}{\sqrt{c}}), \end{array}\right\} \quad \textit{if } \lambda(x) = cx. \tag{2.15}$$

Proof. See Section 2.5 □

The superscripts ML, MQ, MH respectively stand for the Myopic Linear, Quadratic and Hyperbolic models. The first case ML yields linear selling and the TWAP strategy, or if time is parametrized in volume time, the classic VWAP trading strategy. The second strategy x_t^{MH} and its corresponding rate α_t^{MH} represents exponential selling, and is the optimal strategy presented in the original Almgren-Chriss model [5].

This risk term results from the trader's effort to minimize the variance of liquidation cost. From the perspective of an inventory risk measure, one natural alternative is $\lambda(x_s) = cx_s$, which has the attractive property of being proportional to value-at-risk and results in a selling strategy x_t^{MQ} that is quadratic in t. Yet as explained for a similar problem in [31], buying might result with position size small relative to T. Imposing the constraint that x_t^{MQ} is decreasing then leads to the modified solution (2.15) which in the case $\hat{T} < T$ causes liquidation to end prior to the terminal time T.

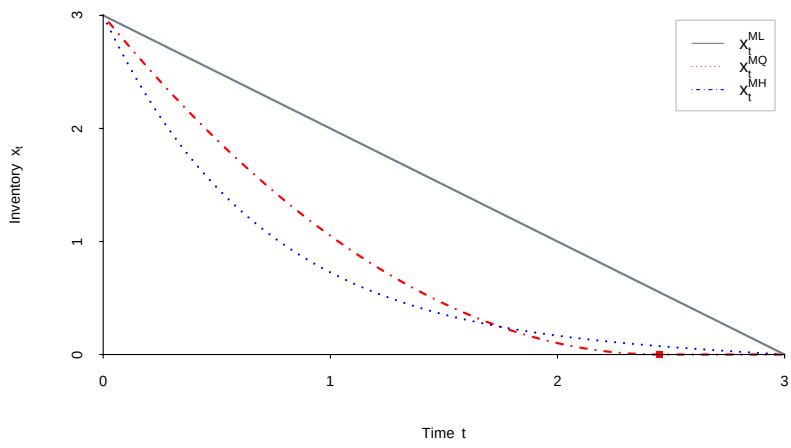

Figure 2.1: Optimal trajectories from Lemma 2.2.1. Figure drawn for initial inventory $x = 3$, horizon $T = 3$, and $c = 2$. This leads to $\hat{T} = \frac{2\sqrt{x}}{\sqrt{c}} = 2.45$ in the quadratic scenario x^{MQ} of (2.15).

Figure 2.1 details the three execution curves described in Lemma 2.2.1 for $x = 3$ and $T = 3$. Compared to the VWAP strategy x^{ML}, the non-zero inventory risk terms in strategies x^{MH} and x^{MQ} lead to higher rates of selling initially. The execution

rate in x^{MH} is proportional to x, cf. (2.14) and liquidation occurs exactly at T. In contrast, as x_t^{MQ} becomes small, the linear risk term cx_t^{MQ} becomes more punitive and it may be optimal to end liquidation prior to time T. Figure 2.1 shows inventory x_t^{MQ} reaching 0 at time $\hat{T} = \sqrt{6}$ as defined in (2.15).

We now turn our attention to the expected execution costs which arise from the trade imbalance. Fixing (α_t^*), we can view the corresponding information impact $\phi(\alpha_t^*)$ also as a deterministic function of t, allowing direct evaluation of the second term in (2.12) using the explicitly available Gaussian distribution of Y_t.

Lemma 2.2.2. *Given a deterministic, time-dependent flow impact $\phi(\alpha_t) = \phi_t$, and $Y_0 = y$, Y_t has the second moment*

$$\mathbb{E}_y\left[Y_t^2\right] = \mu_t^2 + \sigma_t^2, \tag{2.16}$$

where

$$\begin{cases} \mu_t = ye^{-\beta t} - \displaystyle\int_0^t e^{-\beta(t-s)}\phi_s \, ds, \\ \sigma_t^2 = \dfrac{\sigma^2}{2\beta}(1 - e^{-2\beta t}). \end{cases} \tag{2.17}$$

Proof. To solve the SDE (2.3), we start with

$$f(Y_t, t) = Y_t e^{\beta t}$$

which after applying Itô's formula gives

$$\begin{aligned} df(Y_t, t) &= \beta Y_t e^{\beta t} + \sigma e^{\beta t} dY_t \\ &= e^{\beta t}\phi_t dt + \sigma e^{\beta t} dW_t. \end{aligned}$$

Then integrating each side from 0 to t and dividing through by Y_t on each side we have

$$Y_t = ye^{-\beta t} + \int_0^t e^{\beta s}\phi_s ds + e^{-\beta t}\int_0^t \sigma e^{\beta s} dW_s$$

So Y_t is clearly Gaussian with mean and variance as given in (2.17).

□

Putting everything together we obtain the expected total cost for the family of myopic execution strategies in Proposition 2.2.3. We reiterate that while realized costs instantaneously depend on the stochastic process (Y_t), strategies in this section are purely deterministic and do not adapt to (Y_t). The solutions below are labeled according to the form of $\lambda(x)$; while formally the two terms in (2.12) are decoupled, it is of course logical to match the resulting solution $\mathcal{I}$ of the instantaneous price-impact part with the corresponding expectation of $\mathbb{E}_y[\int_0^{T_0} Y_s^2\, ds]$, which is the convention we follow in Proposition 2.2.3.

Proposition 2.2.3. *Suppose that $\phi_t(\alpha) = \phi_t$. The corresponding value function u is given by*

$$u^M(T, x, Y) = \mathcal{I}(T, x) + \mathcal{O}(\hat{T}, x, Y), \tag{2.18}$$

where $\mathcal{I}$ *is defined in* (2.13)-(2.15) *and* $\mathcal{O} = \kappa \int_0^{\hat{T}} (\mu_t^2 + \sigma_t^2)\, dt$ *(with* $\hat{T}$*,* μ_t *and* σ_t^2 *from Lemma 2.2.2) is*

$$\mathcal{O}^0(T, x, Y) = \frac{\kappa y^2}{2\beta}\left(1 - e^{-2\beta T}\right) + \frac{\kappa\sigma^2}{4\beta^2}\left(2\beta T + e^{-2\beta T} - 1\right) \quad \text{if } \phi_t \equiv 0; \tag{2.19}$$

$$\begin{aligned}\mathcal{O}^{ML}(T, x, Y) = \mathcal{O}^0(T, x, Y) &+ \frac{\kappa\eta x y}{\beta^2 T}\left(2e^{-\beta T} - 1 - e^{-2\beta T}\right) \\ &+ \frac{\kappa\eta^2 x^2}{2\beta^3 T^2}\left(2\beta T + 4e^{-\beta T} - e^{-2\beta T} - 3\right) \qquad \text{if } \phi_t = \eta\alpha_t^{ML}\end{aligned} \tag{2.20}$$

Closed form expressions are also available for $\mathcal{O}^{MQ}$ *and* $\mathcal{O}^{MH}$*, see Section 2.5.*

Thus, the overall cost of liquidation has two components: the $\mathcal{I}(T, x)$ term that depends only on (T, x), and the informational footprint term $\mathcal{O}(T_0, x, Y)$ that also depends on y. Recall that informational costs accrue only up to T_0; in the linear and hyperbolic cases we always have $T_0 \equiv T$, but in the quadratic case liquidation may be completed early, $T_0 = \hat{T} < T$, see Figure 2.1. In terms of x, $\mathcal{O}$ is constant if $\phi_t = 0$, linear if $\phi_t = \eta\alpha^{ML}$, and quadratic otherwise. As a function of y, $\mathcal{O}$ is quadratic thanks to the linear dynamics of (Y_t) and quadratic informational cost. Financially, the y^2 term represents higher costs due to trading in an unbalanced market, while the y-term adjusts to the fact that selling in a market dominated by buyers is favorable to competing with other sellers for scarce liquidity. The following Corollary shows that the "best" level of y is positive (or 0 if $\phi_t = 0$). Intuitively, it is best to begin trading in an environment with positive order flow so that the trader's selling activity pushes the order imbalance towards 0 and reduces informational costs.

Corollary 2.2.4. *Suppose that $\phi_t(\alpha) = \phi_t \geq 0$. Then the flow imbalance that minimizes expected execution cost is non-negative,* $\arg\min_Y \{u(T, x, Y)\} \geq 0$ *for any* T, x.

Proof. As already discussed, $u^M(T, x, Y)$ is quadratic in y and the coefficient of y^2 is $\frac{\kappa}{2\beta}\left(1 - e^{-2\beta T}\right)$. By inspection it is positive. The dependence on y comes from the $\mathcal{O}$ terms that are of the form

$$\mathcal{O}(T, x, Y) = \kappa \int_0^T (Ye^{-\beta t} - A_t)^2 + \sigma_t^2 \, dt$$

where $A_t \geq 0$ (strictly positive as soon as $\phi_t > 0$ on an interval of positive measure, cf. (2.42)). It follows that the coefficient of Y^2 in $u^M(T, x, Y)$ is $\kappa \int_0^T e^{-2\beta t} dt > 0$ and of Y is $-\int_0^T 2\kappa e^{-\beta t} A_t dt \leq 0$. Thus, setting $\partial_Y u^M(T, x, Y) = 0$ and solving for Y yields a non-negative result. □

2.2.2 Dynamic Execution Strategies

We now return to the problem in (2.10), letting $\phi(\alpha_t) = \eta\alpha_t$. The optimal dynamic strategy is adapted to the expected trade imbalance process (Y_t) and the trader's rate of selling directly influences (Y_t). To avoid confusion we will denote dynamic strategies and expected costs by α_t^D and u^D, respectively. The HJB PDE for $u^D(T, x, Y)$ is

$$u_T^D = \frac{1}{2}\sigma^2 u_{YY}^D - \beta Y u_Y^D + \kappa Y^2 + \lambda(x) + \inf_{\alpha \geq 0} \left\{ g(\alpha) - \alpha u_x^D - \eta\alpha u_Y^D \right\}, \tag{2.21}$$

with $u^D(0, x, Y) = +\infty$ unless $x = 0$. Note that (2.21) is identical to (2.6) but for the time derivative on the left hand side of the equation which is introduced due

to the time-dependence arising from the constrained horizon T. Additionally, with the fixed horizon T, there is no boundary condition in x, meaning it is possible that trading could continue beyond the point at which inventory first reaches 0. Assuming $g(\alpha) = \alpha^2, \lambda(x) = cx^2$, and inserting the feedback control as in (2.9) yields a semi-linear, parabolic PDE

$$u_T^D = \frac{1}{2}\sigma^2 u_{YY}^D - \beta Y u_Y^D + \kappa Y^2 + cx^2 - \left(\frac{u_x^D + \eta u_Y^D}{2}\right)^2. \tag{2.22}$$

with initial condition (2.11). To obtain (2.22), it is necessary to let α be unconstrained and allowed to become negative. This allows us to find the following candidate solution by exploiting the linear-quadratic structure. The motivation comes from u^M in Proposition 2.2.3 where we find a similar result: quadratic in x and Y with an xY term that adds additional costs when $Y < 0$.

Proposition 2.2.5. *The solution of* (2.22) *has the form*

$$u^{DH}(T, x, Y) = x^2 A(T) + y^2 B(T) + xY C(T) + xD(T) + Y E(T) + F(T), \tag{2.23}$$

where $D(T) = E(T) \equiv 0$, A, B, C, F *solve the matrix Riccati ordinary differential equations (ODEs)*

$$\begin{cases} A'(T) &= -A^2 - \eta AC - \frac{\eta^2}{4}C^2 + c \\ B'(T) &= -\eta^2 B^2 - B(\eta C + 2\beta) + \kappa - \frac{1}{4}C^2 \\ C'(T) &= -\frac{\eta}{2}C^2 - C(\eta^2 B + A + \beta) - 2\eta AB \\ F'(T) &= \sigma^2 B, \end{cases} \tag{2.24}$$

and we have the following initial conditions

$$\begin{cases} \lim_{T \downarrow 0} A(T) = +\infty \\ B(0) = C(0) = F(0) = 0. \end{cases} \tag{2.25}$$

The optimal rate of liquidation is

$$\alpha_t^{DH}(T-t, x_t, Y_t) = \frac{x_t(2A(T-t) + \eta C(T-t)) + Y_t(C(T-t) + 2\eta B(T-t))}{2}. \tag{2.26}$$

Proof. See Section 2.5. □

We reiterate that in (2.4.2) A, B, C, F are functions of time remaining and that we have simplified the notation by omitting the time argument (i.e. $A' = A'(T)$, etc.) on the right side of (2.4.2). Close to the deadline T, the impact from impacting Y disappears, and (2.22) converges to the myopic linear case of (2.13). This can be

observed by formally linearizing the Riccati system (2.4.2) in the regime $T - t = \epsilon$ and using the initial conditions (4.11). We obtain the following expansions in ϵ:

$$\begin{cases} A(\epsilon) = \frac{1}{\epsilon} + O(\epsilon) \\ B(\epsilon) = \kappa\epsilon + O(\epsilon^2); \\ C(\epsilon) = -\eta\kappa\epsilon + O(\epsilon^2); \\ F(\epsilon) = \frac{\sigma^2\kappa\epsilon^2}{2} + O(\epsilon^3). \end{cases} \tag{2.27}$$

Inserting into (2.26) gives the short term trading rate $\alpha^{DH}(\epsilon, x_t, Y_t) = \frac{x_t}{\epsilon} + O(\epsilon)$. This heuristically confirms that the strategy (2.26) is admissible which can also be observed in Figure 2.4 below: as $t \to T$, the dynamic trading rate stabilizes, resembling a VWAP strategy.

Remark 3. It is also possible to set up and solve linear quadratic problems for other functional forms of inventory risk $\lambda(x)$. In the constant case $\lambda(x) = c$ the Riccati system is almost the same as (2.4.2) (again $E(T) = D(T) \equiv 0$) except that the c term moves to the fourth line: $F'_{DL}(T) = \sigma^2 B(T) + c$. In the linear case $\lambda(x) = cx$, the resulting Riccati system is

$$\begin{cases} A'(T) &= -A^2 - \eta AC - \frac{\eta^2}{4}C^2 \\ B'(T) &= -\eta^2 B^2 - B(\eta C + 2\beta) + \kappa - \frac{1}{4}C^2 \\ C'(T) &= -\frac{\eta}{2}C^2 - C(\eta^2 B + A + \beta) - 2\eta AB \\ D'(T) &= c - A(D + \eta E) - \frac{\eta C}{2}(\eta E + D) \\ E'(T) &= -\beta E - \eta B(D + \eta E) - \frac{\eta C}{2}(\eta E + D) \\ F'(T) &= \sigma^2 B - \frac{1}{4}(\eta E + D)^2, \end{cases}$$

with initial conditions as in (4.11) and $D(0) = E(0) = 0$. However, dynamically satisfying the constraint $x \geq 0$ is not tractable (cf. $\hat{T}$ in (2.13)) and we found that the resulting unconstrained strategies tend to lead to wild buying-and-selling. Notably, inventory x_t often becomes negative in which case the inventory risk term loses its meaning.

The equations in (2.4.2) can be dealt with using a software package such as `R`. It is however necessary to replace the singular initial condition (2.11) with the condition

$$\lim_{T \downarrow 0} u^D(T, Y, x) = \begin{cases} 0 & \text{if } x = 0 \\ M & \text{if } x \neq 0 \end{cases} \tag{2.28}$$

for a constant M large, essentially allowing for a non-zero position at time T which must then be liquidated in a single order at some additional cost. This is equivalent

to introducing a boundary layer $[0, \epsilon]$ and solving on $T \in [\epsilon, \infty]$ whereupon $M = 1/\epsilon$ is the right choice based on (2.27).

The optimal trading rate α^D in (2.26) is linear in both x_t and Y_t. The former feature is similar to the hyperbolic situation in (2.14) where α_t^{MH} is also linear in x_t. We next illustrate how the dynamic strategy α_t^D compares to its myopic counterpart α_t^M. With fixed terminal time T, the incentive for the trader to speed up or slow down under strategy α_t^D arises from the trader's desire for more balanced order flow. Note that there is no incentive to accelerate one's trading in order to exit the market prior to time T since costs from Y accrue until T. For positive trade imbalance, trading more quickly in the present results in lower execution costs in the future because (Y_t) will be closer to 0 as a result of his activity. Likewise, if imbalance is negative, it is better to reduce trading speed so as not to pull (Y_t) further from 0. For negative Y and large enough T (or large enough κ, β), α_t^D may become negative (i.e. it may be optimal to begin buying), however this happens only under extreme parameters.

Figure 2.2 illustrates the results of Proposition 2.2.5 for a simulated path of (Y_t^0) comparing the myopic α^{MH} versus the adaptive α^{DH}. As can be observed, both strategies have a broadly similar shape, with α^D "fluctuating" around α^{MH}. We also observe that α^D is less aggressive initially, starting out slower and then speeding up (relative to α^{MH}) after $t > 1.5$.

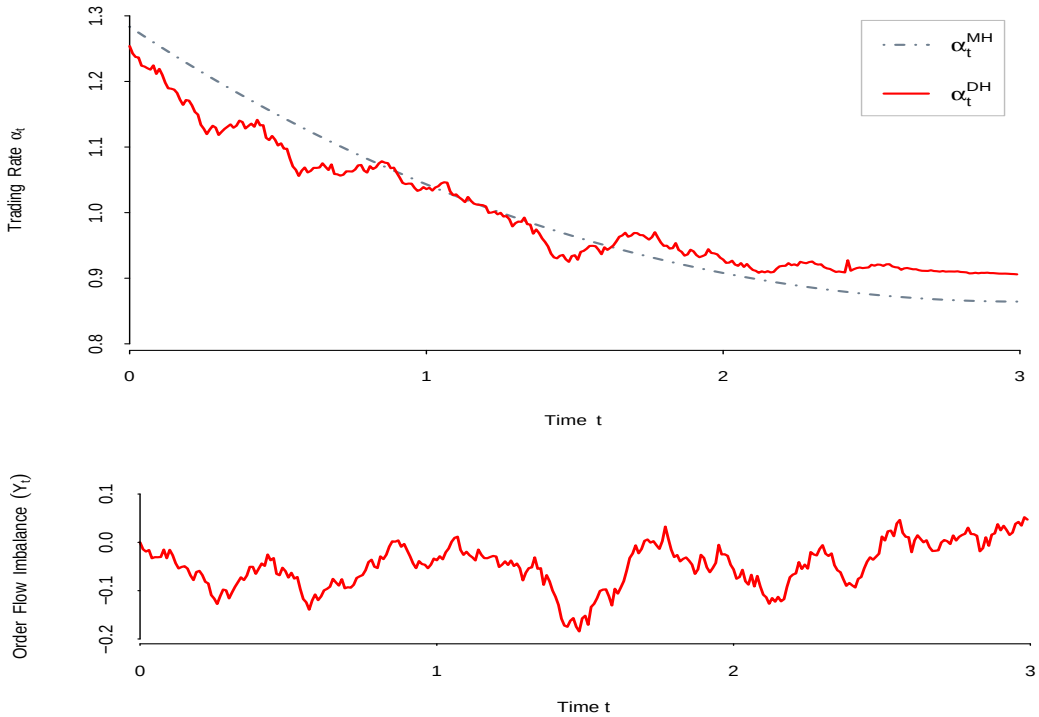

Figure 2.2: Trading rates α_t^{DH} and α_t^{MH} for a sample simulated path of (Y_t) shown in the bottom panel. The figure is drawn for parameter values $T = 3$, $\kappa = 10$, $\sigma = .14$, $\beta = .05$, $\eta = .05$, $\lambda(x) = 0.1x^2$, and initial condition $x_0 = 3, Y_0 = 0$.

2.3 Optimizing Execution Horizon

We now move to the second step of the approximate solution scheme and remove the fixed horizon constraint. Given $u(T, x, Y)$, define

$$T^* := \arg\min_T \; u(T, x, Y).$$

The next Lemma shows that T^* is finite in all the cases considered so far.

Lemma 2.3.1. *For any fixed x, there exists $\bar{T}$ such that $\partial_T u(T, x, Y) > 0$ for all $T > \bar{T}$ and all Y.*

Proof. Recall that $u = \mathcal{I} + \mathcal{O}$, cf. (2.18). As $T \to \infty$, the variational problem (2.12) for $\mathcal{I}$ becomes independent of T. By inspection, $\lim_{T\to\infty} \mathcal{I}^{ML}(T, x) = 0$ and $\lim_{T\to\infty} \mathcal{I}^{MH}(T, x) = \sqrt{c}x^2$. In the quadratic case, for T large enough, $\hat{T} = \frac{2\sqrt{x}}{\sqrt{c}}$ so that $\lim_{T\to\infty} \mathcal{I}^{MQ}(T, x) = \frac{4}{3}x^{3/2}\sqrt{c}$ for some function of the initial inventory. In contrast, $\mathcal{O}(T, x, Y) = \int_0^T \mu_t^2 + \sigma_t^2\, dt$ grows at least linearly in T since $\lim_{t\to\infty} \sigma_t^2 = \frac{\sigma^2}{2\beta}$. Also, the first term is non-negative and it follows that $\partial_T \mathcal{O}(T, x, Y) \geq \frac{\sigma^2}{2\beta}$ asymptotically as $T \to \infty$ and for any y. Hence, $\partial_T u > 0$ for all T large enough. □

Figure 2.3 illustrates Lemma 2.3.1 for the dynamic hyperbolic strategy with value function u^{DH}. We observe that $u^{DH}(T, x, Y)$ appears to be convex in T with a unique global minimum $T^*(x, Y)$. Moreover, $T^*(x, Y)$ is largest for $Y \simeq 0$ and smallest for negative y. This matches the intuition that trading is slowest in balanced markets where informational costs are large, and fastest in sell-driven markets where further information leakage is minimal.

Given initial (x, Y) and corresponding $T^*(x, Y)$, let $\alpha_t^M(T^*, x, Y)$ (and similarly $\alpha_t^D(T^*, x, Y)$) be the resulting strategy over the fixed horizon $[0, T^*)$. This provides a static or open-loop optimal execution strategy, since T^* is fixed and not adjusted as order flow Y_t changes. We can also construct a "dynamic" strategy by continually recomputing $T^*(x_t, Y_t)$ using the latest datum (x_t, Y_t). We denote the latter as

$$\tilde{\alpha}_t^M(x, Y) := \alpha^M(T^*(x_t, y_t), x_t, y_t)$$

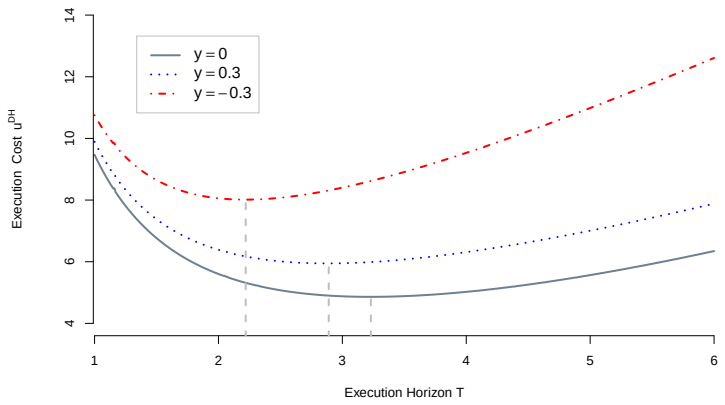

Figure 2.3: Expected execution cost $u^{DH}(T, x, Y)$ as a function of T for different values of trade imbalance Y. The dashed line indicate the value of T achieving the minimum. Figure drawn for $\beta = .05$, $\sigma = .14$, $\eta = .05$, $\kappa = 10$, $\lambda(x) = 0.1x^2$ and inventory $x = 3$.

with associated value function $\tilde{u}^M$. The corresponding $\tilde{\alpha}^D(x, Y)$ and $\tilde{u}^D$ are defined in similar fashion. The approach of "rolling" the horizon T^* as the underlying stochastic state changes is known as receding horizon control or model predictive control, see e.g. [48].

Remark 4. Recomputing T^* can be done at any frequency. Namely, given a (stochastic) set $0 = t_0 < t_1 < \ldots$, one can construct the strategy $\alpha(\mathfrak{T}(t), x_t, Y_t)$ where $\mathfrak{T}(t) := T^*(x_{t_k}, Y_{t_k})$ and $t_k = \max\{t_i : t_i < t\}$. For example, one can take $t_k = \inf\{t : x_t \leq (K-k)x/K\}$, giving K rebalancing periods, during each of which $1/K$ of total inventory is liquidated.

Before moving forward we pause briefly to summarize the various strategies that have been defined. The fully dynamic strategy which solves the original indefinite-

horizon control problem (2.7) in Section 2.1 is denoted $\alpha^*(x, Y)$. In Section 2.2.1 we defined a family of myopic strategies on a fixed horizon, generally denoted $\alpha^M(T, x, Y)$ and specific cases addressed in Lemma 2.2.3. Continually optimizing T^* then yields the receding horizon strategy $\tilde{\alpha}^M(x, Y)$. In Section 2.2.2 we introduced the dynamic strategy $\alpha^D(T, x, Y)$, which adapts to changing flow imbalance over a fixed horizon, as well as the corresponding receding $\tilde{\alpha}^D(x, Y)$.

We proceed to compare execution cost statistics across the described strategies. For easier interpretation we consider the case of zero inventory penalization, $\lambda(x) = c$ independent of x, so that the benchmark strategy (without informational costs) is VWAP, i.e. constant trading rate. The precise parameters were: timing risk $\lambda(x) = c = .1$ (i.e. linear timing costs), initial inventory and initial trade imbalance $x = 3$ and $y = 0$ respectively, $\eta = .075$, $\kappa = 10$, $\beta = .05$ and $\sigma = .14$. Against the original strategy $\alpha^*(x, Y)$, we also compare the adaptive $\tilde{\alpha}^D$ and $\tilde{\alpha}^{ML}$. Both of these adjust the execution horizon by optimizing T in the fixed-horizon solution. Practically, this was achieved by discretizing in time ($\Delta t = .01$) and recomputing $T^*(x_t, Y_t)$ at each time step, see Remark 4. Recall that $\alpha_t^{ML} = x_t / T^*(t)$. To understand the frequency of above "rebalancing", we also show results for the strategy $\alpha_t^{ML}(T^*_{(2)}, x, Y)$ which recomputes the horizon midway through the liquidation process, when $x_t = \frac{x}{2}$. The corresponding path of x_t is therefore piecewise linear with two pieces, see Figure 2.4. This is a convenient compromise in the $VWAP$ setting and nicely illustrates the

advantage gained when the trader is allowed to adjust the execution horizon. Finally, to understand the importance of adaptively adjusting T^*, we also compare to the static $\alpha^{DL}(T^*, x, Y)$ and $\alpha^{ML}(T^*, x, Y) = x/T^*(x, Y)$.

Table 3.1 shows some summary statistics about the distribution of realized costs $J(\alpha) := \int_0^{T_0} (\alpha_s^2 + \kappa Y_s^2 + c)\, ds$. The results were produced with 2000 simulated paths of (Y_t^0). The actual realized trade imbalance paths for each strategy reflect the assumption that $\phi(\alpha_t) = \eta\alpha_t$ represents the true form of information leakage. Beyond the average expected costs $u(x, Y) := \mathbb{E}_{x,Y}[J(\alpha)]$, we also report the standard deviation SD and quantiles $q.$ (at the 5% and 95% level) of $J(\alpha)$ which are important for risk-management perspective. Lastly, we also report the average realized horizon $\mathbb{E}[T_0]$. Of course for non-adaptive strategies, $T_0 \equiv T^*(x, Y)$ is constant. Comparing each $\tilde{\alpha}$ to its respective fixed horizon counterpart demonstrates the importance of utilizing "adaptive" execution horizon. Similarly, comparing respective myopic to dynamic strategies shows that the modelling of deterministic information leakage in the former is not too suboptimal compared to the fully dynamic proportional information leakage strategy of the latter. The cost improvements achieved through optimizing the horizon tend to dominate those obtained through adopting a dynamic strategy in lieu of a myopic strategy.

Of particular interest is that the closed form solution computed for u^{ML} and resulting strategy $\tilde{\alpha}_t^{ML}$ form a reasonable approximation for the difficult indefinite

Optimal Execution Strategy						
	v	$\tilde{u}^D$	$\tilde{u}^{ML}$	$u^{ML}_{(2)}$	u^D	u^{ML}
$\mathbb{E}[J(\alpha)]$	4.257	4.264	4.317	4.411	4.483	4.547
$SD(J(\alpha))$	1.50	1.45	1.39	1.49	1.77	1.84
$q_{.05}(J(\alpha))$	2.70	2.76	2.83	2.96	3.11	3.12
$q_{.95}(J(\alpha))$	7.33	7.28	7.10	7.50	8.19	8.42
$\mathbb{E}[T_0]$	3.87	3.70	3.48	3.44	3.43	3.43

Table 2.1: Statistics for six execution strategies including average realized cost $J(\alpha)$, standard deviation, .05− and .95−quantiles of realized costs as well as average realized execution horizon. Left-to-right the strategies are: dynamic (in Y) with indefinite horizon (v), dynamic with adaptive horizon ($\tilde{u}^D$), myopic with adaptive horizon ($\tilde{u}^{ML}$), myopic with two-step adaptive horizon ($u^{ML}_{(2)}$), dynamic with fixed horizon (u^D) and myopic with fixed horizon (u^{ML}).

horizon setup in (2.6) - (2.9). The cost improvement of the fully dynamic α^* over VWAP strategy α^{ML} is approximately 6.8%. This appears somewhat modest, but note that the latter strategy is applied to the horizon $[0, T^*]$, which is the statically optimal horizon computed at $t = 0$ with the value function u^{ML}.

Figure 2.4 illustrates the various strategies for a sample simulated trade imbalance path (Y^0_t). The one-sided order flow in this particular simulation causes the adaptive horizon strategies to accelerate trading and shorten the horizon relative to the fixed horizon strategies. However, as Table 3.1 shows, on average $\mathbb{E}[T_0]$ actually tends to be longer when using adaptive execution horizon.

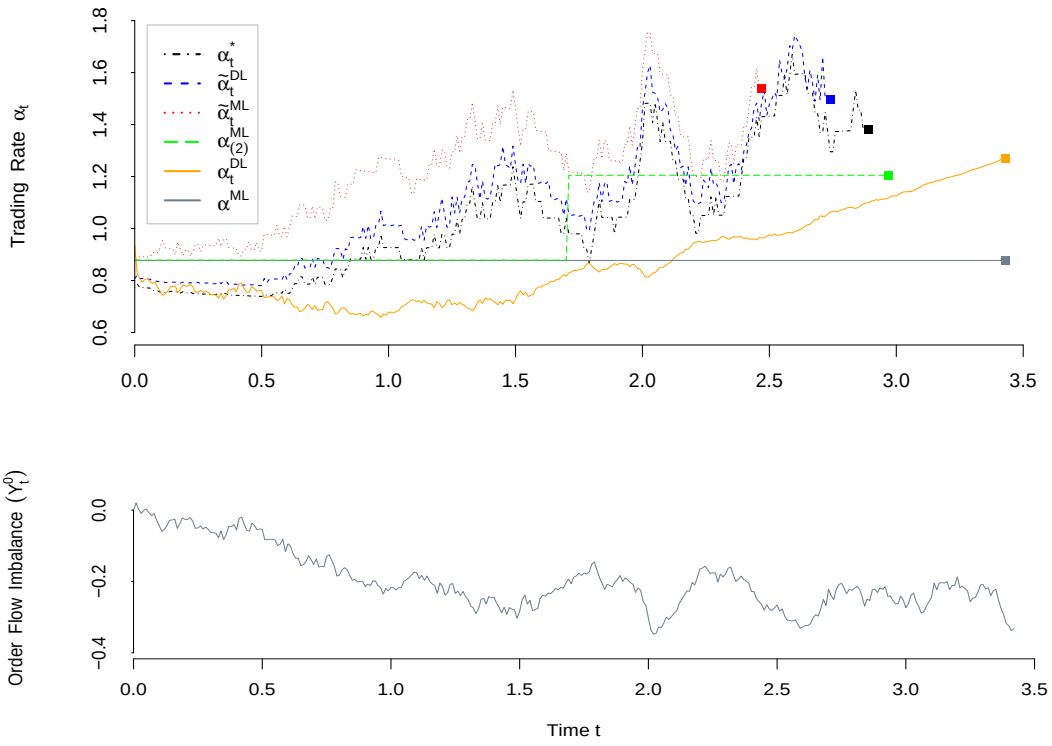

Figure 2.4: Comparison of trading rates (α_t) for each of six strategies in Table 3.1 given the shown simulated path of (Y_t^0) (The realized (Y_t) depends on the strategy chosen). Note that each strategy terminates at a different T_0 indicated with a square.

2.3.1 Comparative Statics

Focusing on a single strategy, $\tilde{\alpha}_t^{DL}$, we briefly discuss how adjusting the values for parameters c, η, and κ affect the trading rate and realized horizon. Increasing c corresponds to lower tolerance for timing risk and intuitively leads to a shorter realized horizon and an increase in trading rate across all values of Y. Choices for κ and η depend respectively on the trader's assessment of the added cost of transacting when

order flow is unbalanced and exactly how susceptible one is to revealing information to other participants. Increasing κ raises sensitivity to the trade imbalance which leads to an increase in trading speed and shortened execution horizon, particularly when (Y_t^2) moves away from 0. At the other extreme, setting $\kappa = 0$ leads to the strategies addressed in Lemma 2.2.1 with zero informational cost. The effects of increasing η depend on the market state, increasing the trading rate when order flow tilts towards buy orders and slowing when order flow is balanced or sell orders dominate. Specifically, an increase in η means a stronger trade impact on the order flow process, and thus it is beneficial in a buy market to tolerate somewhat higher instantaneous costs because the trader can more efficiently capture the savings that result from more balanced order flow in the future (and vice-versa in a sell-tilted market). Figure 2.5 illustrates these comparative statics for $\tilde{\alpha}^{DL}$ in terms of κ and ν. Note that while theoretically α^{DL} from (2.26) could be negative, in all our plots α^{DL} remains far from zero and well-behaved.

2.3.2 Realized Execution Horizon

We now explore some features of the realized execution horizon $T_0(x, Y)$ when following the dynamic strategy $\tilde{\alpha}_t^D(x_t, Y_t)$. The left panel of Figure 2.6 highlights the distribution of $T_0(x, Y)$ for different initial market states. We observe that T_0 tends to be longest in a balanced market. Indeed, in that case the trader pays the

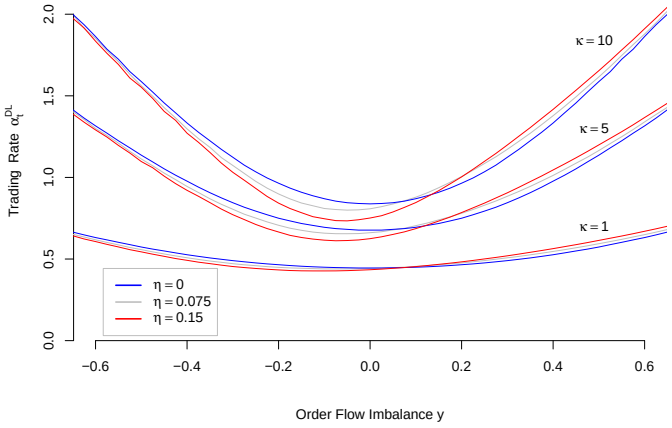

Figure 2.5: Trading rates $\tilde{\alpha}^{DL}(x, Y)$ plotted as a function of flow imbalance y for different values of informational cost κ and information leakage strength η. Inventory level is fixed at $x = 3$.

most attention on minimizing his footprint and instantaneous execution costs, and therefore trades slowly. With positive imbalance $Y_t > 0$, he is incentivized to trade at a more rapid pace in order to bring the market into a more balanced state. On the other hand, when a market is dominated by sell orders $Y_t < 0$, the trader finds himself competing for liquidity and trading occurs at an even faster pace. The asymmetric effect of these effects creates a skew even with a symmetric informational cost κy^2. This phenomenon is further shown in the right panel of Figure 2.6 that shows a scatterplot of T_0 against terminal Y_{T_0}. It is also clear that the issue is not only whether order flow is balanced versus unbalanced, rather the side of the trade is very pertinent to the optimal strategy and realized horizon. Generally, imbalanced order flow results in higher trading costs and a shorter horizon, but clearly as one would

expect, trading against the prevailing order flow (selling when order flow is dominated by buy orders) is preferable.

We remark that with T fixed, execution rate decreases with Y_t in hopes that the order flow process will revert to a more balanced state and the cost of liquidity will decline in the future. However, allowing the horizon to be adjusted brings a new incentive for accelerating execution in order to exit the market altogether and stop information leakage. This is a phenomenon seen especially in times of panic or capitulation when minimizing the footprint is less important than finding liquidity, even at greater cost.

We also observe a strong correlation between realized execution cost and realized execution horizon. Unbalanced order flow results in higher costs from the Y_t^2 term. In addition, as lopsided order flow causes trading to accelerate, the trader also incurs higher costs from the instantaneous cost term α_t^2. So costs tend to be lower for longer realized execution horizon. Figure 2.7 provides another perspective on this feature by highlighting several specific inventory trajectories and the corresponding realized order flow paths. It also shows that the spread in realized horizon T_0 is significant and can be up to 50% of the static T^*.

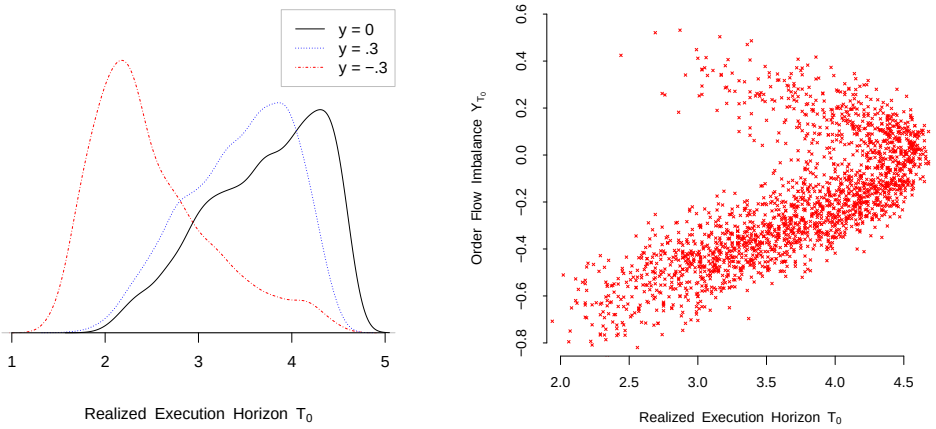

Figure 2.6: Left: Distribution of realized execution horizon T_0 following strategy $\tilde{\alpha}^{DL}$ for different values of initial flow imbalance Y_0. Statistics corresponding to $Y_0 = 0$ are given in Table 3.1. Right: Realized execution horizon T_0 against final trade imbalance Y_{T_0} when initial imbalance $Y_0 = 0$.

2.3.3 Static Information Leakage

One of the motivations for the present analysis was the work of Easley et al. [27] (ELO), who considered a related static optimal execution horizon model. Through the lens of our setup, [27] treated the case where informational costs are measured only through Y_T rather than through the integral term in (2.5). Specifically, ELO equated informational footprint to the absolute value of the terminal flow imbalance $|Y_T|$. Also, ELO (implicitly) assumed a VWAP execution strategy on $[0, T]$ which is equivalent to taking zero inventory risk $\lambda(x) = 0$ and instantaneous impact $\int_0^T \dot{x}_s^2 \, ds$ and translates to the myopic strategy α^{ML} of constant trading rate. Finally, timing risk was modelled directly as $\Lambda(T) = c\sqrt{T}$, motivated by the same structural form

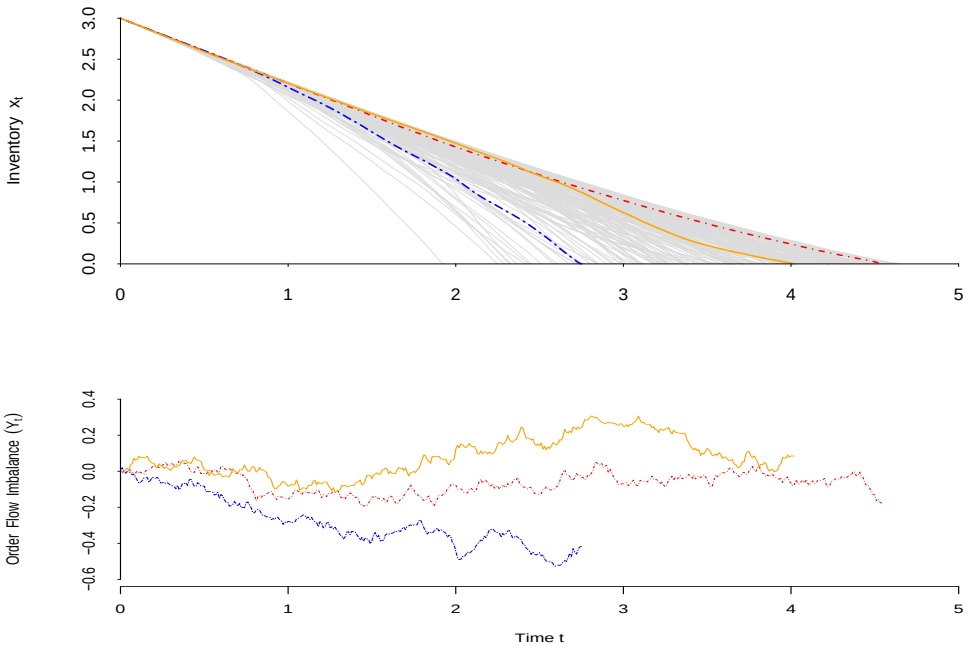

Figure 2.7: Top: 200 simulated trajectories (x_t) from dynamic adaptive strategy $\tilde{\alpha}_t^{DL}$. Highlighted are three trajectories resulting from different realized trade imbalance (Y_t) paths. Bottom: Corresponding realizations of trade imbalance $t \mapsto Y_t$.

for volatility of P_T. The overall problem in [27] was therefore

$$\min_{T \geq 0} \left\{ \mathbb{E}[|Y_T^{\alpha}|] + c\sqrt{T} \right\}, \qquad \alpha_t = x/T. \tag{2.29}$$

Our framework allows treatment of (2.29) in a dynamic setup, i.e. beyond the myopic strategies and beyond a static optimization to obtain $T^*(x, Y)$. The associated HJB equation for (2.29) is

$$u_T = \frac{1}{2}\sigma^2 u_{YY} - \beta y u_Y + \inf_{\alpha \geq 0}\{\alpha^2 - \alpha u_x - \phi(\alpha) u_Y\}, \tag{2.30}$$

with initial condition

$$\lim_{T\downarrow 0} u(T,x,Y) = \begin{cases} |Y| & \text{if } x = 0 \\ \\ +\infty & \text{if } x \neq 0. \end{cases} \tag{2.31}$$

The singular initial condition (2.31) and α^2 cost term (2.30) suggest the VWAP benchmark strategy, which was assumed in [27]. The possibility of directly incorporating a timing cost of the form $\Lambda(T)$ can be easily handled in more generality within (2.5) since the latter term makes no difference to the fixed-horizon problems in Sections 2.2.1-2.2 and hence only shows up in the second-step optimization over T. Based on numerical experiments, replacing running Y-costs with a terminal cost $\propto Y_{T_0}^2$, tends to slow the optimal trading strategy in sell dominated markets which allows the mean reversion in the order flow process to kick in and lower the terminal cost. In the presence of positive order flow, the change in trading rate can be in either direction depending on $\phi(\alpha)$ and the trade-off between instantaneous costs and informational costs.

2.4 Calibration and Extensions

To implement the proposed execution strategies, the trader must continuously measure the expected trade imbalance state Y_t. Moreover, they need to be able to calibrate the parameters of Y_t. This requirement is different from typical execution

strategies that operate in "open-loop" settings, i.e. without any immediate input of market data. Of course, most empirical trading is "closed-loop" and dynamically responds to market messages. In this section we briefly discuss such calibration and translation of market information into model inputs.

2.4.1 Empirical Order Flow

To connect the modeled Y_t to market data, we begin by considering the *executed* orders, which from the flow point of view can be summarized as a sequence $O_1, O_2, \ldots$, where O_i is the signed market order volume (positive for buys and negative for sells) for the security in question. We assume that trading is in volume time, so there is no separate time-stamp component. A raw order flow would then be the cumulative sum $\sum_i O_i$. Assuming that the participants focus on recent trades (i.e. market memory is limited) leads to consideration of moving averages of O_i. By analogy to the discussed Ornstein-Uhlenbeck dynamics we therefore introduce the following exponentially weighted moving average (EWMA) trade imbalance process ($TIMA_i$) that is defined recursively via

$$TIMA_{i+1} = e^{-\beta|O_i|} TIMA_i + (1 - e^{-\beta|O_i|}) \operatorname{sign}(O_i), \tag{2.32}$$

where the memory parameter β is a proxy for the time-scale of market participants persistency of beliefs about order flow. Intuitively, if all trades were of unit volume, we would have $TIMA_{i+1} = e^{-\beta} TIMA_i + (1 - e^{-\beta}) \operatorname{sign}(O_i)$; treating a single trade

of $|O_i|$ as that many unit-volume trades leads to (3.10). We suggest that $\beta = a/V_{daily}$ where V_{daily} is the average daily volume and $a \in [10, 100]$ is the intra-day mesoscopic time-scale of order flow. By construction, O_i takes values in $[-1, 1]$, with $TIMA_i = 0$ representing a balanced market, and positive and negative values of $TIMA_i$ representing a market tilted towards buying and selling respectively. Figure 2.8 shows a typical daily path of $TIMA_i$ for two different values of β. To draw the figure we considered all executed Nasdaq ITCH trades between 9:40am and 3:55pm on a fixed trading day and initialized $TIMA_0 = 0$ in the beginning (note that if one fully adheres to the concept of moving averages, $TIMA_0$ should include flow from the previous day, but this is rather problematic to properly implement). The value of β controls the volatility of $TIMA_i$.

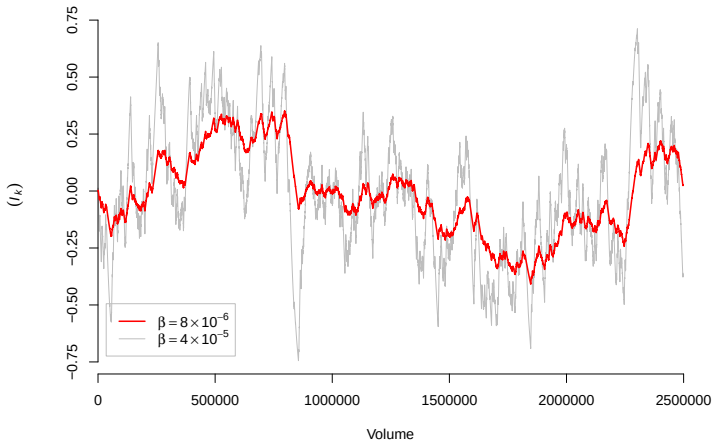

Figure 2.8: The EWMA trade imbalance metric for Teva Pharmaceutical (ticker: TEVA) for a single day 5/3/2011. The data includes executed orders from Nasdaq, BATS and Direct Edge exchanges which accounted for $2,497,623$ of the $8,059,668$ total traded shares on the day. We also show the VPIN-like metric that used $V = 25,000$ and $n = 20$ in (2.33) and (2.34) respectively.

An alternative way to define empirical trade imbalance is based on bucketing. This approach is more in line with a discrete model, such as the one in ELO [27] and indexes flow by equally-sized volume slices rather than by individual trades. Namely, consider (executed) volume slices of size $V = V_k^B + V_k^S$ where (V_k^B) and (V_k^S) represent the buy and sell volume respectively for the k-th slice. The bucket trade imbalance $\widetilde{TI}_k$ is

$$\widetilde{TI}_k := \frac{V_k^B - V_k^S}{V} = 2V_k^B - 1. \tag{2.33}$$

Compared to (3.10), we have the link $V \simeq 1/\beta$ to achieve same time-scale for $(TIMA_k)$ and $(\widetilde{TI}_k)$.

The defined $TIMA_i$ and $\widetilde{TI}_k$ are directly observed, and one could attempt to use them as the basis for the expected flow process (Y_t) used in the previous sections. According to [23, 24, 27, 26], informational costs arise from order flow *toxicity* which is in turn tied to the participants' beliefs about probability of adverse selection. Hence, translating past information contained in TI into Y requires making a judgement on how such beliefs about future order flows are formed. Indeed it is well known that signed market order flow exhibit positive autocorrelation and long memory [13], [12], [49]. Therefore it naturally follows that TI and Y_t would be very closely related.

It is also commonly accepted that certain trades are influential or informative while others have little to no impact on the market. With this in mind, it follows that a trade may have little influence on a market maker's expected flow imbalance

even if the associated trade volume O_i was large. Thus, the private information leaked to the market by a trade is the product of numerous factors beyond trade size: spacing of successive orders, prevailing market state, LOB shape, etc. Consequently, the exact relation between (TI) and (Y_t) remains open. Further questions about the most relevant time scale or how the recent history of observed order imbalance might influence the expected future trade imbalance are investigated in Chapter 3. Another related problem is calibrating the functional form of the information leakage function $\phi(\alpha)$. While in our example we worked with a linear $\phi(\alpha) = \alpha$, a more realistic specification would probably require a convex relationship to trading rate (and possibly zero impact for α small). Nonlinear information leakage is easily accommodated through the use of myopic strategies of Section 2.2.1 which are agnostic about ϕ.

Remark 5. In [26], ELO contend that order flow toxicity is linked to the level of "informed" trading and can be approximated via the following VPIN metric based on (2.33),

$$\text{VPIN}_k = \frac{1}{n} \sum_{i=k-n}^{k-1} |\widetilde{TI}_i|, \tag{2.34}$$

where n, chosen along with bucket size V represents the window relevant for persistency of order flow. Thus, VPIN is a moving average of observed values of $\tilde{TI}_k$ for n latest volume slices. According to ELO, VPIN is a good approximation to the market-makers' expectations of the current imbalance $\text{VPIN}_k \approx \mathbb{E}|\widetilde{TI}_k|$ and hence can be used as a proxy for Y_t in (2.29). In other words, the traders should act to

minimize their impact on VPIN, which is the expected absolute trade imbalance. Because VPIN is directly based on observed traded volumes, this also allows an explicit definition of trader's informational impact, see (2.37) below. ELO suggest to take $V = V_{daily}/50$ and $n = 50$ which makes VPIN a daily moving average of flow imbalance. This seems rather long and in Figure 2.8 we make n smaller to focus on intra-day scale. More recently, concerns about VPIN's usefulness as a predictive indicator have surfaced, for example see Anderson and Bondarenko [9], [10] and ELO [26].

2.4.2 Correlated Price Process

An important aspect that is missing from the presented models is *price risk*. With a fixed horizon, the assumption that the unperturbed asset price is a martingale makes realized revenue only depend on execution risk. However, once the agent has the liberty to extend the execution horizon, this is no longer the case and the trader could also chase higher revenues. Explicit modeling of such objectives would necessarily increase the dimensionality since another stochastic state variable, the (mid-)price P_t, must be added. Nevertheless, under certain assumptions this more general setup could still be tractable. In particular, one could maintain the linear-quadratic structure by adopting the assumptions of [31]. Gatheral and Schied [31] assume that S is described by a geometric Brownian motion and inventory risk is measured by time-averaged

value-at-risk (VAR), i.e. $\lambda_t = \lambda x_t P_t$. Making this adjustment in (2.5) leads to the problem of minimizing

$$\check{u}(T,P,x,Y) := \mathbb{E}_{p,x,y}\left[\int_0^T \left(\dot{x}_t^2 + \kappa Y_t^2 + \lambda x_t P_t\right)\, dt\right]. \tag{2.35}$$

Assuming linear information leakage and that P and Y have correlation $\rho \in [-1,1]$, $dW_t^{(P)} dW_t^{(Y)} = \rho dt$ allows for the solution

$$\check{u}(T,P,x,Y) = x^2\check{A}(T) + Y^2\check{B}(T) + xY\check{C}(T) + P^2\check{D}(T) + Px\check{E}(T) + PY\check{F}(T) + \check{G}(T)$$

where the coefficients $\check{A},\ldots,$ solve again a Riccati ODE

$$\begin{cases} A'(T) &= -A^2 - \eta AC - \frac{\eta^2}{4}C^2 \\ B'(T) &= -\eta^2 B^2 - B(\eta C + 2\beta) + \kappa - \frac{1}{4}C^2 \\ C'(T) &= -\frac{\eta}{2}C^2 - C(\eta^2 B + A + \beta) - 2\eta AB \\ D'(T) &= \sigma_P^2 - \frac{1}{4}(\eta F + E)^2 \\ E'(T) &= c - A(E+\eta F) - \frac{\eta C}{2}(\eta F + E) \\ F'(T) &= -\beta F - \eta B(E+\eta F) - \frac{\eta C}{2}(\eta F + E) \\ F'(T) &= \sigma_Y^2 B - \rho, \end{cases}$$

with similar initial conditions as before. Moreover, a fully closed-form solution is possible for strategies that are myopic with respect to Y_t (see [31] for details). As in Remark 3, the main difficulty is that an inventory risk that is linear in x cannot

guarantee $x_s \geq 0$ and hence is likely to include buying, making the optimization over T delicate.

In the dynamic case the execution strategy $\check{\alpha}^D(t, P_t, x_t, y_t)$ is given by

$$\frac{1}{2}\Big\{x_t(2\check{A}(T-t)+\eta\check{C}(T-t))+Y_t(\check{C}(T-t)+2\eta\check{B}(T-t))+P_t(\check{E}(T-t)+\eta\check{F}(T-t))\Big\}$$

and is therefore linear in price P_t. (The myopic strategy is also linear in P_t). Note that only the execution cost $\check{u}$ is impacted by ρ, while the strategies themselves are independent of the correlation between asset prices and order flow. In this setup, the strategy adapts to the fluctuations in price throughout the execution process. Arguably, the current value of Y might affect the *future* asset price. This issue is revisited in Chapter 4. In reality, the joint behavior of order flow and asset prices remains poorly understood. In fact, the positive correlation between market-order flow and asset price is a direct indication of adverse selection affecting liquidity providers, see e.g. the very recent preprint [14]. Further investigation into how the co-movement of order flow and asset prices might affect execution costs is found in Chapter 3.

2.4.3 Discrete Time Formulation

Building on (2.33) and (2.34) one can construct a discrete model for optimal execution. This involves reinterpreting the strategy α as a participation rate based on the observation that an executed sell trade inherently affects the next bucket

imbalance $\tilde{I}_\ell$ since it physically displaces some of the other volume from that bucket. A discrete-time model also allows to examine more general costs and model dynamics.

Fixing a volume bucket V, we assume that the trader chooses a participation rate α_k at each step, where

$$x_{k+1} = x_k - \alpha_k V$$

and k indexes trade volume. The trader's participation influences the flow imbalance at the next step via

$$Y_{k+1} = F(Y_k, \epsilon_{k+1}) - \phi(\alpha_k, Y_k) \tag{2.36}$$

where ϵ_{k+1} are independent random perturbations, $F(Y, \cdot)$ models the dynamics in Y_k that happen apart from the trader and $\phi(\alpha, Y_k)$ is the information leakage given previous imbalance Y_k. One motivation to generalize the leakage function is the trade influence proposed in [27],

$$Y_{k+1} = \psi(\alpha_k)(Y_k(1-\alpha_k) - \alpha_k) + (1 - \psi(\alpha_k))Y_k + \epsilon_{k+1}, \tag{2.37}$$

where we still have $\alpha_k \in [0, 1]$ and the function $\psi \in [0, 1]$ is monotonic increasing. The new expected trade imbalance is a convex combination of two extreme outcomes: full leakage according to ψ (first term) and no leakage (second term). (2.37) simplifies to

$$\mathbb{E}[Y_{k+1}] = Y_k - \alpha_k \psi(\alpha_k)(Y_k + 1) =: Y_k - \phi(\alpha_k, Y_k).$$

In analogue to Section 2.1, the trader's goal is to minimize total expected costs until the entire position has been liquidated,

$$v(x,Y) := \inf_{\alpha} \mathbb{E}_{x,Y}\left[\sum_{k=0}^{T_0-1} g(\alpha_k) + \kappa Y_k^2 + \lambda(x_k)\right], \quad T_0 = \min\{k : x_k = 0\}. \tag{2.38}$$

The control α_k is constrained so that $\alpha_k \in (0,1]$ and is assumed to be in feedback form, $\alpha_k = \alpha(x_k, Y_k)$.

The indefinite-horizon control problem (2.38) can be solved by introducing an auxiliary "time" variable t such that execution stops after t steps (if it did not terminate already) and remaining inventory at t incurs a terminal cost $H(x_t) = Ax_t^2$ (assuming immediate one-step liquidation after t, cf. the VWAP strategy in (2.13)). This auxiliary problem has value function $v^{(t)}$ defined in (2.39). As $t \to \infty$, this execution horizon constraint vanishes and we expect to recover the time-stationary solution $v(x,Y)$ of (2.38).

Using t as time-to-maturity, we have the discrete-time dynamic programming equations

$$\begin{aligned} v^{(0)}(x,Y) &= H(x), \\ v^{(t)}(x,Y) &= \inf_{\alpha\in(0,1]} \mathbb{E}_Y\Big[g(\alpha) + \kappa Y^2 + \lambda(x) + v^{(t-1)}(x - \alpha V, Y_1)\Big] \end{aligned} \tag{2.39}$$

where $Y_1 = Y_1^\alpha$ is defined by (2.36) and $v^{(t)}(0,Y) = 0 \forall Y$. The one-stage problem in (2.39) can be readily solved using a Markov-chain approximation method by discretizing the state space of Y and the bounded control space of α (which also makes the

state space of x_t discrete). The above procedure allows arbitrary dynamics for (Y_k) beyond (2.37) since one can always use Monte Carlo or other methods to compute the transition density $p_{Y_1}(\cdot|Y_0 = y, \alpha)$.

2.5 Proofs

Finite Difference Approach to (2.9)

We employ an explicit finite difference scheme to find an approximate solution to (2.9). Denote $v_{i,j} := v(x_i, y_j)$, where $x_i = i\Delta x$ for $i = 0, 1, \ldots, N$ and $y_j = y_0 + j\Delta y$ for $j = 0, 1, \ldots, M$. Derivatives of v are approximated as

$$\begin{aligned}
\frac{\partial}{\partial x} v(x_i, y_j) &= \frac{v_{i+1,j} - v_{i,j}}{\Delta x}; \\
\frac{\partial}{\partial y} v(x_i, y_j) &= \frac{v_{i,j+1} - v_{i,j-1}}{2\Delta y}; \\
\frac{\partial^2}{\partial y^2} v(x_i, y_j) &= \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{(\Delta y)^2},
\end{aligned}$$

and we apply the boundary condition $v(0, y) = v_{0,j} = 0 \forall j$. In y we use the boundary conditions for $v_{i,0}$ and $v_{i,M}$ via $\frac{\partial^2 v(x_i, y_1)}{\partial y^2} = \frac{\partial^2 v(x_i, y_{M-1})}{\partial y^2} = 0$ and choose y_0 and y_M such that $\mathbb{P}((Y_t) \notin [y_0, y_M]) \approx 0$. Substituting into (2.9) we have

$$0 = \frac{1}{2}\sigma^2 \frac{\partial^2 v_{i,j}}{\partial y^2} - \beta y_j \frac{\partial v_{i,j}}{\partial y} + \kappa y_j^2 + \lambda(x_i) - \frac{1}{4}\left(\frac{\partial v_{i,j}}{\partial x} + \eta \frac{\partial v_{i,j}}{\partial y}\right)^2$$

and rearranging terms yields

$$v_{i+1,j} = v_{i,j} + \Delta x \Bigg(2 \left(\kappa y_j^2 + \lambda(x_i) - \beta y_j \frac{v_{i,j+1} - v_{i,j-1}}{2\Delta y} + \frac{\sigma^2}{2} \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{(\Delta y)^2} \right)^{1/2} - \eta \frac{v_{i,j+1} - v_{i,j-1}}{2\Delta y} \Bigg).$$

So we have a so called "time-marching" scheme in x, where v at each inventory level $i+1$ can be approximated by values from the previous inventory level i. This explicit approach is available for appropriate parameter values in which $\alpha^* > 0$ holds. For more extreme parameter values we may have $\alpha^* = 0$ (recall that we constrain $\alpha^* \geq 0$). In this case our explicit method would fail for certain grid points and it would be necessary to utilize an alternative method.

Proof of Lemma 2.2.3

Proof. We skip the trivial case $\lambda(x) = 0$. See [33] for details when $\lambda(x) = cx^2$. For $\lambda(x) = cx$, when x is unconstrained, the problem is a straightforward application of the Euler-Lagrange equation. For cost functional $F(x, \dot{x}, t) = \dot{x}_s^2 + cx_s$, the optimal trajectory x_t^{MQ} must satisfy

$$\left(\frac{dF}{dx} - \frac{d}{dt} \frac{dF}{d\dot{x}} \right) = 0 \tag{2.40}$$

Applying the constraint that x is decreasing introduces the boundary condition at $x = 0$. Then the optimal trajectory for the constrained minimization problem either lies on the curve which satisfies the Euler-Lagrange equation or lies along the boundary,

with the transition from the former to the latter occurring where x_t^{MQ} is tangent to the line $x = 0$. In order to find the point at which this transition takes place, we find T' such that $\frac{dx_t^*}{dt}\big|_{t=T'} = 0$. It is easily verified that $T' = \frac{2\sqrt{x}}{\sqrt{c}}$ satisfies the requirement and so we have $\hat{T} = \min(T, \frac{2\sqrt{x}}{\sqrt{c}})$.

Having computed x_t^M for each choice inventory risk term, α_t^M and $\mathcal{I}^M$ are computed respectively by differentiating with respect to t and integrating over the interval $[0, T]$ ($[0, \hat{T}]$ for $\mathcal{I}^{MQ}$). □

Proof of Proposition 2.2.1

Proof. An application of Fubini's theorem, permits the interchange of expectation and integration leaving us with a straight-forward but lengthy integral computation. The general expression for $\mathcal{O}$ is

$$\mathcal{O}(T, x, Y) = \int_0^T (Ye^{-\beta t} - A_t)^2 + \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t})dt \tag{2.41}$$

where

$$A_t := \int_0^t e^{-\beta(t-s)}\phi_s ds \tag{2.42}$$

captures all the information leakage. Integrating the other two terms gives

$$\begin{aligned}\mathcal{O}^0(T, x, Y) &= \kappa \int_0^T (Ye^{-\beta t})^2 + \sigma_t^2 dt \\ &= \frac{\kappa Y^2}{2\beta}\left(1 - e^{-2\beta T}\right) + \frac{\kappa\sigma^2}{4\beta^2}\left(2\beta T + e^{-2\beta T} - 1\right).\end{aligned}$$

Computing the integral for the remaining three cases is straight-forward but tedious. We provide the values of A_t for each case. For $\phi_t = \eta\alpha_t^{ML}$, we have $A_t^{ML} = \frac{\eta x}{\beta T}(1 - e^{-\beta t})$. When $\phi_t = \eta\alpha_t^{MQ}$ we have

$$\begin{aligned} A_t^{MQ} &= \eta \int_0^t \left(\frac{c\hat{T}}{4} + \frac{x}{\hat{T}} - \frac{ct}{2} \right) e^{-\beta(t-s)}\, ds \\ &= \frac{\eta}{\beta^2} \left(c(1 - \beta t + e^{-\beta t}) + \beta(1 - e^{-\beta t}) \left(\frac{c\hat{T}}{4} + \frac{x}{\hat{T}} \right) \right). \end{aligned}$$

Lastly, when $\phi_t = \eta\alpha_t^{MH}$ and $\sqrt{c} \neq \beta$, we have

$$\begin{aligned} A_t^{MH} &= \eta \int_0^t \frac{\sqrt{c}x \cosh(\sqrt{c}(T-s))}{\sinh(\sqrt{c}T)} e^{-\beta(t-s)}\, ds \\ &= \frac{\eta\sqrt{c}x}{(c-\beta^2)\sinh(\sqrt{c}T)} (\lambda \sinh(\sqrt{c}(T-t)) + \sqrt{c}e^{-\beta t}\sinh(\sqrt{c}T) \\ &\quad - \beta\cosh(\sqrt{c}(T-t)) + \beta e^{-\beta t}\cosh(\sqrt{c}T)). \end{aligned}$$

If $\sqrt{c} = \beta$ then the final expression simplifies to

$$A_t^{MH'} = \frac{\eta x e^{-\beta t} e^{-\beta T} \left(2\beta t e^{2\beta T} + e^{2\beta t} - 1 \right)}{4\beta \sinh(\beta T)}.$$

One can finally integrate (using a symbolic integration software for example) (2.41) over $t \in [0, T]$ to obtain closed-form expressions for $\mathcal{O}^{MQ}$ and $\mathcal{O}^{MH}$. □

Proof of Proposition 2

Proof. Substituting (2.23) into the HJB PDE (2.22) we have

$$\begin{aligned}
u_T^D &= \sigma^2 B(T) + \kappa Y^2 + c^2 x^2 - \beta Y(2YB(T) + xC(T)) \\
&\quad - \left(\frac{2xA(T) + YC(T) + \eta(2YB(T) + xC(T))}{2}\right)^2 \\
&= x^2\left(c^2 - (A(T))^2 - \eta A(T)C(T) - (\eta C(T))^2\right) \\
&\quad + Y^2\left(\kappa - 2\beta B(T) - \left(\frac{C(T)}{2}\right)^2 - \eta B(T)C(T) - (\eta B(T))^2\right) \\
&\quad + xY\left(-\beta C(T) - A(T)C(T) - 2\eta A(T)B(T) - \frac{\eta(C(T))^2}{2} - \eta^2 B(T)C(T)\right) + \sigma^2 B(T).
\end{aligned}$$

On the other hand, $u_T^D = x^2A'(T) + Y^2B'(T) + xYC'(T) + xD'(T) + YE'(T) + F'(T)$. Matching the appropriate powers of x, Y on either side of the equality yields the system of Riccati differential equations (2.4.2), with the boundary conditions (2.11) translating into (4.11). As previously mentioned, the first order terms $D(T)$ and $E(T)$ in x and Y are not required if the inventory risk term is of order x^2 or 1. □

Chapter 3

Order Flows and Limit Order Book Resiliency

Electronic trading marketplaces match liquidity providers and consumers via the limit order book (LOB). At any given moment, the LOB summarizes the current state of trading by listing the resting limit orders entered into the matching engine. Since price formation is essentially mechanical given the LOB state (namely, the LOB directly determines the mid-price), on the instantaneous time-scale the LOB is fundamental for understanding price dynamics. Similarly, the LOB is fundamental for explaining price impact, namely the effect of trader's actions on the book and price. The concept of price impact is typically quantified in terms of *liquidity*, with higher liquidity indicating smaller price impact and scarce liquidity indicating larger

impact. Questions of liquidity are key for market participants planning their actions, such as market making or order execution.

However, on a longer time-scale, the enormous volume of data and noise captured by the LOB becomes a major challenge that obscures the link between price evolution, liquidity, and participant orders. Indeed, dynamically, rather than the static snapshots offered by the LOB, the main drivers of price formation are arguably the order flows. In this chapter, we initiate statistical analysis and modeling of such mesoscopic quantities. One of our main aims is to understand the link between order-flow and liquidity, which could be of interest to execution traders for purposes of scheduling in a manner that adapts to the changing state of liquidity and thus varying expected price impact.

The remainder of the chapter is structured as follows. In Section 3.1 we fix notation, and discuss LOB evolution, highlighting the primary motivations for the current study. Section 3.2 describes our data and approach, explaining how we choose to aggregate order flows to analyze at the intermediate time scale ($1-10$ minutes). Section 3.3 explains regression models and results focusing on price trend, liquidity and scarce liquidity. Lastly, Section 3.4 discusses ramifications from the previous sections and additional anecdotal findings.

3.1 Price Impact

Price impact directly corresponds to liquidity, a somewhat murky concept that has become increasingly difficult to pinpoint with the steady rise in the speed of today's market activity. To help frame the discussion, consider an execution trader who wishes to sell x shares. Because there are a finite number of resting buy orders in the LOB, she follows convention and divides the order into smaller "child" orders so as to disrupt the market as little as possible. Nevertheless her orders consume liquidity and reveal information to other market participants. Moreover, there is the possibility for latency arbitrage. Because execution is typically over multiple time slices, rather than a static cost dictated by the shape of the LOB (e.g. instantaneous impact), expected execution costs are driven by expected price change, which depends on the liquidity state i.e. how the LOB evolves through time.

3.1.1 Limit Order Book Notation

To fix ideas, we revisit the LOB structure and set notation for its quantities. The LOB is constructed by aggregating the history of all submitted trading orders. The two base classes of trades are market orders and limit orders. Market orders, which indicate actual transactions taking place are denoted as $\mathfrak{M} := \{(T_i^M, O_i^M)\}$ where T_i^M are the execution times and O_i^M are corresponding execution volumes. O_i^M's are signed, with positive indicating a buy order and negative a sell order. Limit orders

are $\mathfrak{L} := \{(T_i^L, O_i^L, S_i^L)\}$ where T_i^L are the message time stamps, O_i^L are (signed) order volumes and S_i^L is the limit order price. For limit orders positive O_i^L indicates adding a new limit order, while negative O_i^L indicates a *cancellation*.

At time t, the LOB consists of the vector $(p_i^j(t), v_i^j(t))$, listing the volume $v_i(t) \geq 0$ of resting limit orders at price $p_i(t)$. Limit orders reside in the LOB until executed by an incoming market order, or being cancelled by participant who initially placed the order. The superscript $j \in A, B$ denotes the Ask and Bid sides respectively. We separately index each side of the LOB, starting from the best-bid and best-ask levels and moving consecutively. Thus, $p_1^A(t)$ is the best-ask (at the touch) price, $p_2^A(t) = p_1^A(t) + \Delta p$ is the second price level, etc. The LOB levels are discretized in terms of the tick size Δp, typically 1 cent for US equities. Note that the indexing is consecutive and some queues can be empty, $v_i^j(t) = 0$. By definition $v_1^j(t) > 0$ is always strictly positive and the spread is at least 1 tick $p_1^A(t) > p_1^B(t)$.

The midprice is defined by

$$P(t) := \frac{p_1^A(t) + p_1^B(t)}{2}, \tag{3.1}$$

and the bid-ask spread is defined $Spr(t) = p_1^A(t) - p_1^B(t)$. Figure 3.1 illustrates a hypothetical LOB example.

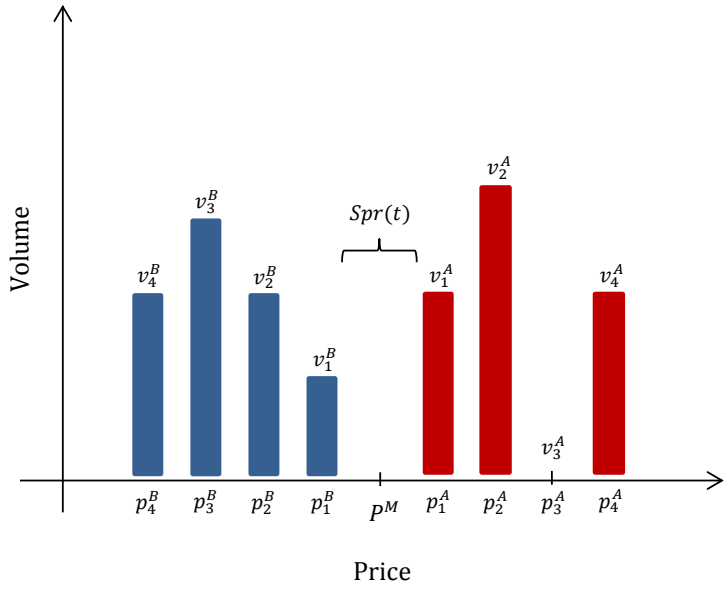

Figure 3.1: Stylized limit order book.

3.1.2 Static Measures of Liquidity

A starting point for understanding the expected price impact of an order is the information contained in a static snapshot of the LOB. Quantities of interest include spread $Spr(t)$, depth, LOB shape and LOB imbalance. Traditionally, liquidity has been measured in terms of such quantities and more recently several studies have shown the predictive information found in an LOB snapshot[35, 18, 42, 22] Importantly, static measurements and the predictions derived may be useful but at *very* short time scales. We briefly consider some examples.

The most basic measure of liquidity is the spread $Spr(t)$. The informativeness of the bid-ask spread varies significantly by asset class. Generally US equities are considered very liquid and spreads in today's markets tend to be small relative to other assets and historical standards. Within the class of US equities "large tick stocks are such that the bid-ask spread is almost always equal to one tick"[28]. Therefore $Spr(t)$ is not a useful measure of liquidity for a large class of stocks because it is not sensitive to changing market conditions.

Another commonly used measure is the volume-at-the-touch $v_1^j(t)$. Over very short time horizons, volume at the touch may be sufficient for understanding the direction and likelihood of a change in mid-price. Volume imbalance, defined

$$VI(t) = \frac{v_1^A(t) - v_1^B(t)}{v_1^A(t) + v_1^B(t)}, \tag{3.2}$$

has been shown to be predictive of the next order side and price move [22, 18]. At the slicing/routing level of the execution process, $VI(t)$ is often used as an indicator of when to employ limit orders and when to cross the spread and place a market order [42]. Similarly, in [35] order arrival rates are modeled as a function of relative queue size (i.e. imbalance).

Beyond the first queue, efficiently summarizing the LOB shape or depth profile (on even a single side of the LOB) is a non-trivial matter. Depth over the first n queues $D_n^j(t) = \sum_{i=1}^n v_i^j(t)$ for example is a blunt measurement as it provides no information about the shape of the book. A better view is provided by the theoretical price impact

of quantity N shares. Let $i^N = \min\{i : \sum_i v_i \geq N\}$, then we define

$$PI_N^j(t) := N^{-1}\left(\sum_{i=1}^{i^N-1} v_i^j(t)(p_i^j(t) - P(t)) + (N - \sum_{i=1}^{i^N-1} v_i^j(t))(p_{i^N}^j(t) - P(t))\right), \quad (3.3)$$

where $v_i^j(t)$ and $p_i^j(t)$ are the depth and price at the ith tick from the mid-price, and N is some fixed quantity of shares. This succinct measure is the weighted average cost per share of immediately executing N shares and therefore contains information on the shape and depth of the LOB, as well as the spread.

Theoretical price impact $PI_N(t)$ can be used to formulate an estimate of the instantaneous execution cost of executing N shares. A recent paper by Cartea and Jaimungal [16] computes $PI(t)$ for various volumes (by walking through the LOB) up to N shares and then fits a simple linear regression model of the form

$$PI_i(t) = \frac{1}{2}\Delta p + \hat{k}i + \epsilon_i, \quad (3.4)$$

where $i = 1, 2, ..., N$, spread is assumed constant $Spr(t) = \Delta p$ and $\hat{k}$ is the estimated slope coefficient. $\hat{k}$ estimates temporary price impact per share when price impact is assumed linear. Both $PI(t)$ and corresponding slope $\hat{k}$ effectively capture the LOB state at time t. However, it is well known that most individual market orders have zero observable impact, meaning that orders rarely "walk the book", instead usually consuming at most the standing liquidity at the first level.

To sum up, static liquidity measures derived from the state of the LOB at a time t make sense primarily at very short time scales ($\ll$ 1 second). In practice,

large orders are split into smaller orders and executed over time to avoid walking the book. Therefore the static picture of the LOB is not enough when it comes to optimal execution. At the scheduling level, $(1 - 10$ minutes) LOB evolution, in particular trends in order-flow and behaviors exhibited by liquidity providers become more relevant.

3.1.3 Order Flows and LOB Evolution

Most traditional execution models in the literature decompose price impact into instantaneous and permanent impacts. The former (discussed in the previous section) models impact only on the current transaction price, and usually carries the assumptions that the LOB recovers infinitely fast to its previous state. The latter captures the impact on the mid-price that persists and affects trading in the future. An alternative approach applied in so-called resilience models [2, 32, 3] instead assumes transient impact. That is, the LOB has some general shape, market orders arrive and impact the mid-price by consuming a portion of the LOB and finally limit orders "refill" over time, often exponentially. In reality, following a market execution, the LOB response varies widely, at times bouncing back immediately in a resilient fashion, while other times falling through and retreating.

Figure 3.2 illustrates the evolution of the LOB for TEVA on 2/18/2011 over a period of 90 seconds. To visualize the book, we focus on the top 2 queues with

volumes $v_1^A(t)$ and $v_1^B(t)$ (left axis). A vertical line represents the jump caused by an arriving order and horizontal lines represent periods of inactivity at each respective queue. Limit additions and cancellations and market orders (orange) are plotted event-by-event so that the mechanics and sequence of order arrivals are more clear. The right axis corresponds to the mid price (upper lines) with dotted vertical lines marking a change in bid- or ask-price.

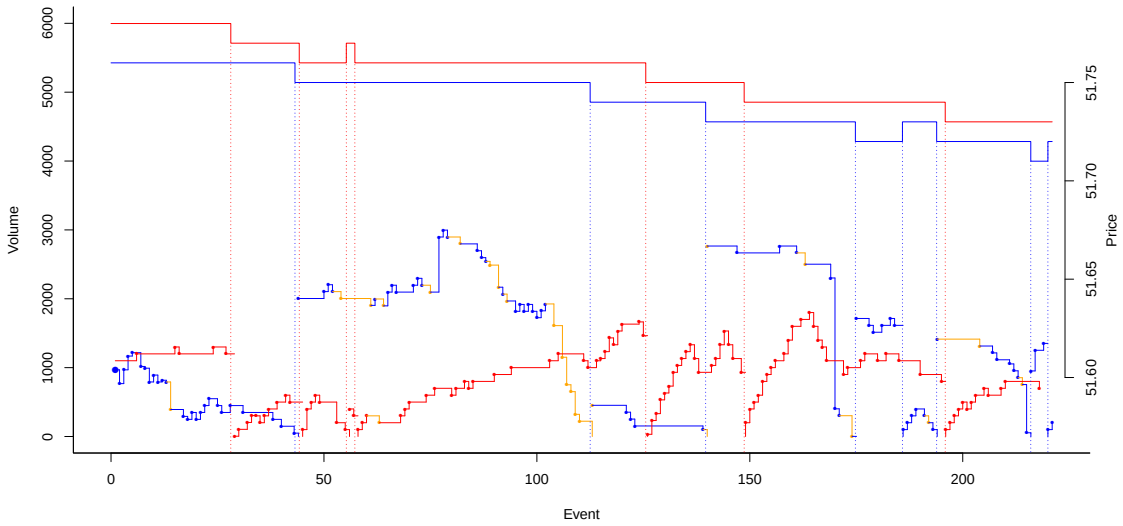

Figure 3.2: Best bid/ask queues $v_1^j(t)$ for TEVA along with bid/ask price $p_1^j(t)$ (top). Data taken from a 90 second window beginning at 2:30pm on 2/18/2011. Cancellations exceed additions at the best bid. Limit orders are in red and blue depending on side, market executions are orange.

One key element that is showcased in Figure 3.2 is the interaction between limit and market orders. We observe several typical "regimes", for example periods where market orders are counteracted with added limit orders (so that $v_1^j(t)$ stays roughly constant over time), and other periods where market orders are accompanied primar-

ily by LO cancellations, creating a strong negative trend in $v_1^j(t)$. The latter situation would correspond to scarce liquidity, as the book is "retreating" along with executed trades. In contrast, in a deep or resilient market, executed orders do not impact the book which bounces back through fresh LO. Visually, one can imagine in Figure 3.2, that excluding market orders, the drift of the queue size at the bid/ask would trend positively through time. Scarce liquidity on the ask (bid) side would be characterized by the alternative, either negative drift or near zero-drift combined with a high quantity of buy (sell) executions.

The microstructure behavior in Figure 3.2 is on the short time-scale and can be analyzed directly using queue-theoretic methods found in [44, 18, 35]. In that context, factors such as static volume imbalance, queue priority and sign of last market orders are the main drivers. For example, Huang et al. [35] model limit order, cancellation and market order arrival rates as a function of the queue sizes $v_i^j(t)$. Under this Markovian assumption tractable formulas can be often be computed for interesting quantities such as the probability of move up or down in the mid-price, but the historical order flow is ignored. Here, we aim to lift these features to a mesoscopic time-scale by documenting and modeling some of the persistent behavior of book liquidity/resilience that are driven by order flows and revealed on the minutes-scale.

Figure 3.3 nicely illustrates this point, albeit with a very extreme example. On the July 12, 2012, four large cap US stocks exhibited an unusual trading pattern;

heavy buying following by heavy selling in a predictible fashion at 30 minute intervals. One expects that LOB volume imbalance $VI(t)$ should be positively correlated with price movement. This is typically the case. Here we see the opposite is true: As the price moves higher (respectively, lower) there is more depth at the best ask (bid). So while the static look indicates plentiful liquidity, market maker actions, apparently responding to the incoming order flow, leads to significant price slippage for the aggressive buyers/sellers. In their research note [40] Lehalle et al. conclude that one likely cause of the very unusual trading behavior was a derivative hedging strategy that entirely ignored common knowledge about market microstructure. For our purposes, the event is a clear example of weak LOB resilience as liquidity providers anticipate 1-sided MOF.

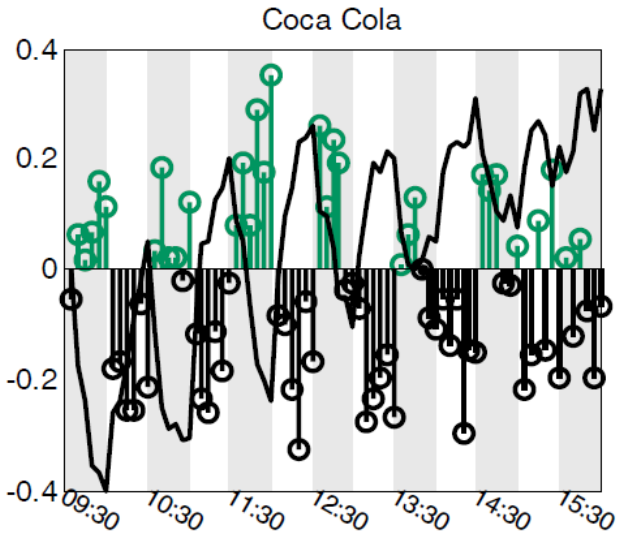

Figure 3.3: Plot taken from Lehalle et al. [40]. Stock price $P(t)$ for Coca-Cola and LOB volume imbalance $VI(t)$ at the touch aggregated over 5 minute time bars (green/black).

3.2 Data and Methodology

Our main dataset consists of Nasdaq ITCH data from the first five calendar months of 2011 for three tickers (MSFT, TEVA, BBBY). Nasdaq ITCH data contains rich information on all order book activities, including limit orders and cancellations, direction and size of market executions, and LOB data out to 30 levels. Pre-processing included removing all executions against hidden orders (less than 10% of executed volume) and aggregating executed orders where necessary. Market orders are often matched against several smaller limit orders and are reported in the data in terms of those matching limit orders. We re-created the size of the original market order by aggregating orders that were consecutive, in the same direction and with equal time stamps. Furthermore, to avoid erratic behaviors often seen near the open or close, we consider only activity between 10am and 3 : 45pm in all statistical analysis. Summary statistics are provided below for the three stocks. The following section first outlines how the vast amount of data (e.g. up to 1,000,000 daily messages), was aggregated for the analysis.

3.2.1 Volume Slices

To move from the micro- to meso-scale, capturing both dynamic and static LOB measures in our analysis, we divide the trading day into buckets and aggregate order flows. Rather than dividing the day according to clock time, we choose to divide

the day into equally sized trade volume slices. While there are difficulties with any method of aggregation of activity, using volume slices has some attractive properties for our purposes.

Most importantly, even though market orders account for only $2 - 4\%$ of total trades, they indicate actual transactions taking place and hence ultimately drive traders' P&L. Due to their intrinsic nature of "putting money on the table", they carry the most information and are typically viewed as influential by other participants. Slicing by trade volume rather than time means more consistency in information across buckets. Under clock-time bucketing, one can easily appreciate the difficulty of comparing buckets with dramatically different activity levels. A related advantage is that working in volume-time helps reduces the intra-day seasonality effects such as volume and volatility clustering. It also allows for a partial recovery of Normality and IID assumptions.

Lastly, one of our key objectives is better understanding LOB behavior by studying interaction in order flows. LOB evolution is driven by volume. Comparing limit order activity across buckets is natural and most consistent when measuring in (trade) volume time. One would reasonably expect for example, that limit additions to the touch depend more on quantity of market orders than passing minutes.

Practically, slices are delineated by slice times τ_k yielding

$$V_k = \sum_{i:\tau_k \leq T_i^M \leq \tau_{k+1}} |O_i^M| \tag{3.5}$$

$$VM_k^B = \sum_{i:\tau_k \leq T_i^M \leq \tau_{k+1}} |O_i^M| 1_{O_i^M < 0} \tag{3.6}$$

$$VL_k^B = \sum_{i:\tau_k \leq T_i^L \leq \tau_{k+1}} O_i^L 1_{S_i^L = p_1^B(T_i^L).} \tag{3.7}$$

In the study, volume slices $V_k = V$ are held constant at 1% of average daily volume (ADV) for each stock. This requires that market orders are split as one bucket "fills" up and the next begins. Slice times τ_{k+1} are taken to be the time stamp of the final trade included (or partially included) in slice k. On average, volume slices span just under four minutes but of course fluctuate and range from one second to much longer depending on trade arrivals.

Note that by definition, $VM^j \geq 0$ is non-negative (and measures total market buys/sells during the k-th slice), whereas VL^j can be of either sign. Indeed, VL^j counts the net additions/cancellations of limit orders at the best ask/bid during the slice. In fact we expect VL^j to also be positive, since by definition liquidity providers ought to add more than they cancel to balance liquidity consumptions captured by VM^j. Co-movement between VM and VL on the same side of the book is what defines resilience in the LOB. Finally, we define for each volume slice, the normalized trade imbalance

$$TI_k = \frac{VM_k^B - VM_k^A}{V}. \tag{3.8}$$

Trade imbalance captures the supply and demand for the asset and is often used as the basis for measuring price impact [13], [16]. We now analyze properties of these mesoscopic time-series TI, VL^j and VM^j.

3.2.2 Data Summary

All three assets are categorized as large cap stocks, with 2011 market caps of approximately \$237B, \$48B and \$12B respectively. Table 3.1 provides summary info for each stock, with the first four lines pertaining to static LOB quantities and the next four lines to order dynamics after bucketing by traded volume. Mid-price and ADV (includes only Nasdaq trading) information are shown for reference. There are also some notable differences which can be seen in the typical LOB snapshot quantities.

For each stock, the spread $Spr(t)$ spends most of the trading day at .01, especially MSFT which rarely sees a spread beyond .01 and neatly fits the definition of a "large tick asset". Another key difference between assets is the relative depth. The second line in the table shows the average volume-at-the-touch v_1 and the third line shows the ratio of volume-at-the-touch to average market order size $v_1/|O^M|$. This ratio shows mechanically some of the reasons for higher volatility observed in TEVA and BBBY compared to MSFT, which exhibits much higher depth relative to market order size. The fourth line of the table shows price impact PI_N^j where N is a fixed

	MSFT		TEVA		BBBY	
	mean	stdev	mean	stdev	mean	stdev
Spr	.011	.003	.014	.006	.015	.007
v_1	$20,006$	$17,956$	$1,019$	$1,393$	586	575
$v_1/\|O^M\|$	12.52	–	3.87	–	3.44	–
PI_N	.017	.012	.019	.012	.016	.011
ΔP_k	4.21E-05	.027	9.09E-04	.054	4.59E-04	.058
TI_k	−8.18E-03	.371	5.27E-03	.359	4.66E-03	.337
$\|TI_k\|$	.295	.225	.282	.222	.267	.206
VL	$83,510$	$102,738$	$14,534$	$10,471$	$5,777$	$5,942$
ADV	$13,702,322$	$6,947,343$	$1,672,942$	$1,169,025$	$846,328$	$305,071$
V_k	$135,000$	–	$17,000$	–	$8,500$	–
P	26.34	1.28	50.38	2.48	50.78	3.61

Table 3.1: Summary statistics for trading days from 1/1/2011 to 5/31/2011.

number of shares. For the remainder we set N equal to the average volume at the top three queues of the LOB $N \approx 3\bar{v}$ for each stock. We find PI_N^j to be the most statistically significant static measure of liquidity/price impact at the time scale induced by volume slices of $V = 1\%$ADV, more so than the other measures described in 3.1.2.

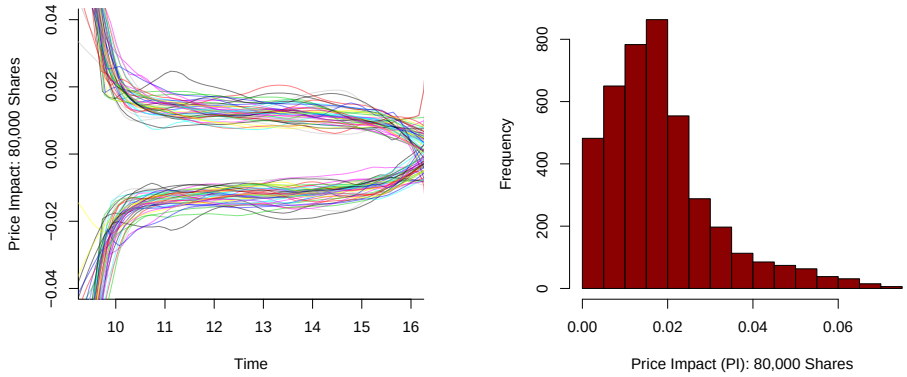

Figure 3.4: Left: daily evolution of PI_N^A (upper) and PI_N^B (lower) for MSFT, $N = 80,000$ over the first 50 trading days of 2011. Right: Histogram of PI^A, measured at each execution time for a single day 1/4/11.

Intraday seasonality is observed in both spread $Spr(t)$ and especially PI_N^j. Note that the intra-day seasonality caused by trade clustering is removed by measuring in volume slices, but there still exists seasonality in the LOB which can be seen in the left plot of Figure 3.4. Each line corresponds to a single trading day, where PI_N^j is computed at each execution time and then smoothed using a lowess approximation. The high price impact values during the first $30-60$ minutes of trading explain much of the increased volatility at this time relative to the rest of the day. PI_N^j compresses significantly near the open and then gradually declines throughout the day reaching its minimum around the close. This pattern is consistent with other studies e.g. [16], and is observed in a number of assets.

Figure 3.5 shows a histogram of net limit order-flow VL_k^B at the bid-side touch (left) and elapsed time (in seconds) for volume slices of $1\%ADV$ (right). Notably,

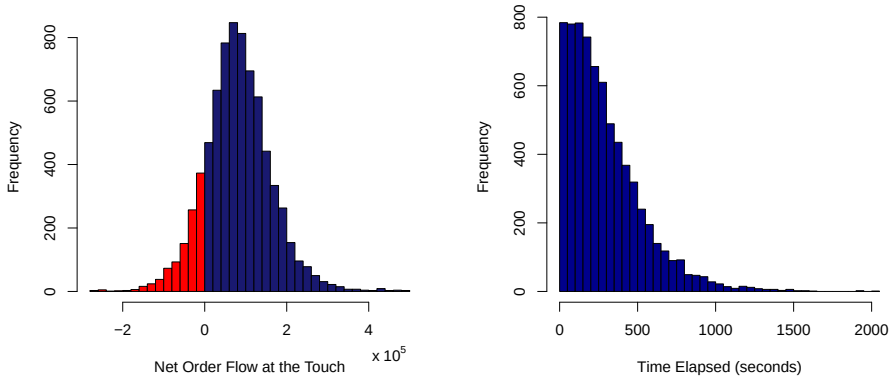

Figure 3.5: Left: Net order flow at the best bid level VL_k^B. Right: Time elapsed during volume slices. Figures drawn for MSFT over 103 trading days.

the correlation coefficient for VL_k and time spanned by the kth slice, $\tau_{k+1} - \tau_k$, is approximately 0. As expected, average limit flow at the touch is typically positive, but does become negative in about 14% of observed volume buckets for MSFT (shown). VL_k is negative less frequently in BBBY ($< 10\%$) and TEVA ($< 4\%$). One likely reason for the discrepancy can be observed in the measure of depth relative to order size $v_1/|O^M|$. Low or negative net order-flow at the touch has less effect on the LOB (in terms of mid-price change) for MSFT than for TEVA or BBBY.

Figure 3.6 shows a histogram of price change ΔP_k (left) and trade imbalance TI_k/V (middle) by volume slice for TEVA over the entire sample of trading days. Also shown are the two plotted against one another showing a positive relationship as should be expected. One feature that immediately stands out is the dispersion in the observed price change ΔP_k conditional on concurrent trade imbalance TI_k. This

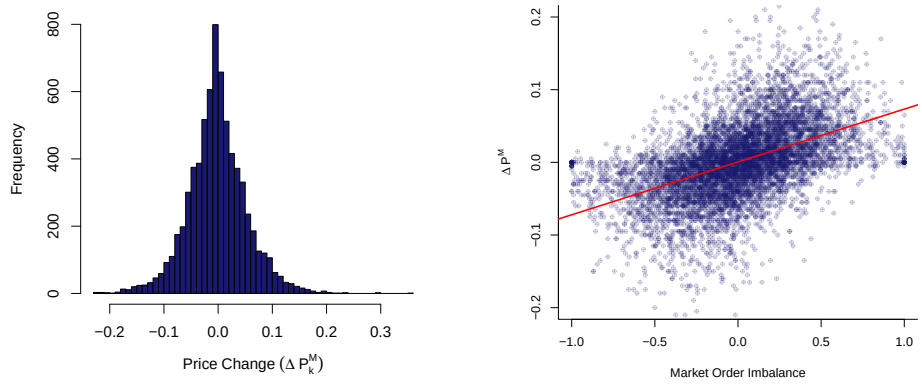

Figure 3.6: Summary plots for TEVA for all 103 trading days. Left: Histogram of ΔP_k. Right: Price change ΔP_k plotted against trade imbalance TI_k fitted-least squares regression line.

highlights that market order flow is only part of the story, and that additional factors play a key role in price formation.

3.3 Empirical Results

Figure 3.6 suggests the natural decomposition of price formation into two components: (1) *price trend*, which is primarily about the supply and demand for the asset captured in trade imbalance TI_k, and (2) *liquidity*, which is seen in the deviation of the observed ΔP_k from the best fit curve between the two variables. This section applies a series of regression models aimed at disentangling price trend from liquidity. We seek to identify key variables important for explaining the magnitude of ΔP at

the minutes scale (Sections 3.3.1 and 3.3.2) and predicting periods of scarce liquidity that lead to out-sized price impact (Section 3.3.3).

3.3.1 Price Trend

In the first step, we seek to explain price change over each volume slice $\Delta P_k = P_{k+1} - P_k$ using only a single explanatory variable, the concurrent trade imbalance TI_k. We fit the following model for each stock on volume slices from the entire sample period

$$\Delta P_k = g(TI_k) + \epsilon_k \tag{3.9}$$

where ϵ_k is the normally distributed error term (assumed iid: autocorrelation in error terms was negligible) and g is a smooth function computed using penalized regression splines (See [50] for details). In contrast to [19, 16] we do not assume a linear relationship between net market flow TI_k and ΔP_k. Instead a generalized additive model (GAM) is assumed to help account for the non-linear dependence that we observe.

Figure 3.7 highlights two key features. The first is the non-linearity in the relationship for extreme values of trade imbalance TI_k. At first glance this result appears somewhat puzzling. However, it is well-known that execution traders take into account the state of the LOB when placing orders. In other words, $135,000$ shares, executed consecutively in the same direction, is an infrequent occurrence for MSFT

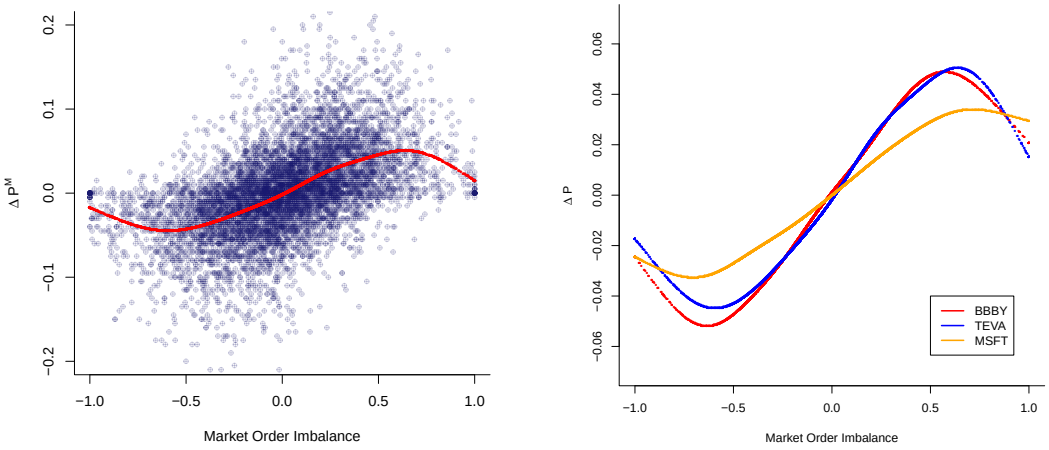

Figure 3.7: Left: ΔP against TI for TEVA over volume slices of $V = 17,000$. Right: Non-linear curves for MSFT, TEVA, BBBY were computed using penalized regression splines. The flatter curve in MSFT is due to its higher depth relative to the size of volume slice V.

and is usually precipitated by a very large quantity of resting limit orders at or near the touch.

The second important feature of Figure 3.7 is the significant amount of dispersion around the best fit curve, which can be verified in the relatively low R^2 levels for the model (3.9) for each stock: 48.4%, 27.4% and 24.9% for MSFT, TEVA and BBBY respectively. Unlike the time slicing approach, each bucket contains the identical amount of traded volume. Thus the difference between observed price change ΔP for two buckets with similar trade imbalance TI must be due to differences in LOB depth/shape or resilience. Lastly, the normal assumption on the error terms ϵ_k is reasonable, although we do observe slightly heavier tails which we can attribute to occurrences of low depth/resiliency.

Because the daily trading session provides convenient break in activity, it is often assumed that LOB behavior is confined to this same daily scale. Therefore, one approach is to fit a separate regression model for each trading day. Indeed, fitting individual daily models does reveal significant day-to-day variation in the shape of the best fit curve (also see [16]). However, the average R^2 across our entire sample when following this convention is only marginally better than when fitting a single model to all days (e.g. the average R^2 across all individual days for MSFT improves to 52%). So the significant noise in (3.9) is not simply the result of inter-day variation, but rather shows that liquidity provision is stochastic on an intra-day basis.

3.3.2 Liquidity

Having removed the effect of trade imbalance (i.e. the price trend), we fit a new regression model to explain the residuals $\hat{\epsilon}_k$ from equation (3.9). For example, in the case of TEVA (Figure 3.7), $\Delta P_k | TI_k = 0.5$ ranges from -0.10 to .20 for buckets across our sample period. Explaining this variation in realized price change for a given trade imbalance is of practical importance to execution traders. In this section we test both static LOB measures and dynamic quantities relating to liquidity provision. Recall, that a static quantity measures something physical related to the LOB at some specific time t. By dynamic, we mean that some quantity related to a process (e.g. limit order flow) is measured over time.

The static variables tested in our regression models included the measures of LOB depth and shape described in Section 3.1 including PI_N, $\hat{k}$, D_i for $i = 1, 2$ and volume imbalance VI. Volume imbalance VI and D_1 were not significant in the regression tests. Price impact slope $\hat{k}$ improved the model slightly but is, as expected, highly correlated with price impact PI_N^j and proved less significant to the model than did PI_N^j. All static variables were measured at the start of the volume slice i.e. the kth residual $\hat{\epsilon}_k$, computed with activity during $[\tau_k, \tau_{k+1}]$ was regressed against static variables measured at time τ_k. Table 3.3.2 summarizes the variables tested here and in the following section.

Dynamic quantities initially tested were the net limit flow at each touch VL^j, trailing limit flow VL_ℓ^j price change ΔP_ℓ, trade imbalance TI_ℓ over the previous ℓ volume slices, where $\ell \in 1, 5, 10, 20$. Also tested was the exponentially weighted moving average of recent trade imbalance, first introduced in Chapter 2,

$$TIMA_{i+1} = e^{-\beta|O_i^M|} TIMA_i + (1 - e^{-\beta|O_i^M|}) \operatorname{sign}(O_i^M), \tag{3.10}$$

where i indexes arrival of market orders of size $|O_i^M|$ and β dictates how quickly the process decays towards 0. We set β equal to 1%ADV and 10%ADV and found that when $\beta = 10\%\text{ADV}$, $TIMA$ is highly significant as a predictor for each stock. Each dynamic quantity was measured at time τ_k except for VL^j which captured the concurrent flows from the kth bucket.

We fit several general linear models and settled on model 3.11 which includes the most important predictors for each stock.

$$\hat{\epsilon}_k = \beta_0 + \beta_1 PI_k^A + \beta_2 PI_k^B + \beta_3 VL_k^A + \beta_4 VL_k^B + \beta_5 TIMA_i + \eta_k \tag{3.11}$$

where η_k are assumed iid and normally distributed. Coefficients and corresponding R^2 levels are shown in Table 3.2. There were additional variables that were to a lesser extent, statistically significant for one or two of the stocks, but these showed negligible improvement to overall model fit.

	MSFT			TEVA			BBBY		
	coeff	stdev	R^2	coeff	stdev	R^2	coeff	stdev	R^2
PI_k^A	4.202E-01	3.074E-02		4.987E-01	4.896E-02		4.242E-01	4.055E-02	
PI_k^B	−4.705E-01	3.029E-02	4.1%	−5.313E-01	4.773E-02	1.6%	−5.179E-01	3.899E-02	0.8%
VL_k^A	−5.727E-08	1.932E-09		−2.053E-06	4.278E-08		−4.401E-06	8.642E-08	
VL_k^B	6.785E-08	1.992E-09		1.817E-06	3.827E-08		2.521E-06	6.894E-08	
$TIMA$	−1.876E-02	1.610E-03	36.5%	−2.964E-02	3.178E-03	42.5%	−5.134E-02	3.805E-03	39.3%

Table 3.2: Least-squares regression Results for model 3.11. Volume slices $V = 1\%ADV$: $135,000$, $17,000$, $8,500$ shares respectively.

Static Variables Given the meso-scale and the amount of activity in each volume slice it is not surprising that "shallow" LOB measures capturing only at the top of the LOB are not all that useful. Instead, the longer time scale in this analysis requires

information from deeper in the LOB. Theoretical price impact PI_N^j ($N \approx 3\bar{v}$), is the LOB measure that provides the most explanatory power for each of the three stocks. On average, for a given trade imbalance TI_k, ΔP_k increases (respectively, decreases) by approximately 0.01 for each increase in PI_N^A (respectively, PI_N^B) of 0.02. PI_N^j does vary throughout the day, but this is largely a seasonal effect (see Figure 3.4).

In the case of MSFT, the depth at the first two queues D_2 is also significant to the model. This is a consequence off MSFT's larger depth profile relative to both the average market order size $|O_i^M|$ and ADV. In other words, the state of the LOB at the start of the volume slice is more relevant to price formation for MSFT than for TEVA or BBBY. The first set of model R^2 values in table 3.2 assumes $\beta_3 = \beta_4 = \beta_5 = 0$, showing the relatively weak explanatory power of PI_N^j.

Dynamic Variables. The net limit order flow processes at each touch VL^j are by far the most significant explanatory variables in the regression. For a given trade imbalance TI, net order flow at the touch measures the resilience of the LOB. When net order flow at the best bid VL^B is large the price is unlikely to move lower even if trade imbalance is negative. On the other hand, if cancellations at the bid dominate additions, the price can move lower on very little volume. For TEVA, an additional $10,000$ shares added to the best bid, on average leads to an increase of 0.02 in ΔP_k conditional on trade imbalance TI_k.

Somewhat less intuitively, the trailing moving average (prior to the kth bucket) trade imbalance $TIMA$ is also a key variable in the model. The coefficient for $TIMA$ is consistently negative across all three stocks and highly significant to the model, of approximately equal importance as PI_N^j. The inclusion of $TIMA$ appears to be capturing the tendency of stock prices, to more often than not, retrace recent movement. More precisely, when TI_k leans with the prevailing trend ($TIMA$) the impact on price is on average less than expected. Lastly, recent price trend ΔP_ℓ proved significant for TEVA and MSFT for $\ell = 20$ and $\ell = 1$ respectively with negative coefficients in both cases.

The primary objective of the last two sections was to identify those quantities that contribute most to price formation at the intermediate time scale. To that end, we also tested for variable interaction and again applied a GAM model to account for non-linear dependence. These revealed nearly identical results in terms of relative variable importance. The moderate improvement in goodness of fit obtained by allowing for non-linear dependence is mostly the result of over-fitting to the very few extreme values of VL^j. We have opted to show only the linear regression results to keep the presentation compact and because the key explanatory variables remain the same. In the following section we choose to focus specifically on periods of low LOB resiliency i.e. scarce liquidity.

Remark 6. The regressions performed in Sections 3.3.1 and 3.3.2 are qualitatively similar to the work by Cont et al [19]. We similarly find that including concurrent limit order flow at-the-touch along with trade imbalance TI_k dramatically improves the statistical fit of the model when predicting ΔP_k. But there are several key differences between the studies. First, in time scale (we test at minutes scale on average compared to their 10 second buckets) and bucketing (we apply volume buckets instead of time bucketing). Second, we separate limit flow and trade imbalance (rather than combining into one variable as in [19] in an effort to understand how the two interact with one another. Third, the goal of this study is to measure and predict periods of scarce liquidity, which is where we now turn our attention.

3.3.3 Scarce Liquidity

Within a given trading day small variation in realized $\hat{\epsilon}$ is to be expected. Pinpointing the confluence of factors that lead to or at least coincide with scarce liquidity is of practical importance due to the potentially large ramifications of triggering, or executing during, a period of heavy cancellation activity in the LOB. We define scarce liquidity $SL \in 0, 1$ in a binary fashion in terms of the realized outcome $\hat{\epsilon}$ so that SL^j

Variable	Description
Static	
PI_N^j	Price impact, $N \approx 3\bar{v}$ shares, $j \in A, B$, (3.3)
$\hat{k}^j$	Slope of linear price impact function, (3.4)
D_i^j	Total shares at top i levels of the ask/bid side of LOB, $i \in 1, 2$
VI	Volume imbalance,(3.2)
τ	Hours from midnight
Dynamic	
VL^j	Concurrent net limit order flow at each touch, (3.7)
VL_ℓ^j	Net limit order flow at the touch during previous ℓ volume slices
ΔP_ℓ	Price change over the previous ℓ volume slices, $\ell \in 1, 2, 5, 10$
$\overline{TI}_\ell$	Mean trade imbalance over the previous ℓ volume slices
$\lvert TI\rvert_\ell$	Mean absolute trade imbalance over the previous ℓ volume slices, (3.13)
ρ_t^j	Correlation between market flow and net limit flow totals over 30-second intervals at the ask/bid side touch, over previous 2.5 hours. (3.14)
$TIMA$	Exponentially weighted moving average of recent trade imbalance, (3.10), $\beta \in 1\%ADV, 10\%ADV$
ρ^{Tox}	Correlation between signed market orders and price change over previous 200 order arrival, Equation (3.15)

Table 3.3: Static and dynamic quantities tested in regression models (3.11) and (3.16).

occurs on at most one side of the LOB per volume slice. Specifically, SL^A is defined

$$SL_k^j = \begin{cases} 1 & \text{if } \hat{\epsilon}_k \geq M \\ \\ 0 & \text{if } \hat{\epsilon}_k < M \end{cases} \tag{3.12}$$

where SL_k^B is defined similarly in the opposite direction and M depends on the asset's liquidity profile. We choose M as a function of the standard deviation of the residuals $\hat{\epsilon}$, namely $M = 1.5\hat{\sigma}_{\hat{\epsilon}}$ so that scarce liquidity occurs on one side of the LOB

in approximately $5-8\%$ of volume slices. Applying a logistic regression model to each side of the LOB we test the predictor variables described in Table 3.3.2.

Alongside the covariates tested in the previous section, we also include additional variables which may contribute to price formation process but do not affect the sign of said price moves. For example, we now include as an explanatory variable the time of the volume slice τ, which was not applicable in explaining signed price change but is useful as a predictor for the magnitude of a price change. We also test three so-called *toxicity* indicators. Toxicity is often used in reference to incoming market order flow from "informed" traders who possess better knowledge of the future asset price. In theory, liquidity should become more expensive in a toxic market as market makers try to avoid being adversely selected.

The first such measure is based on ELO's *VPIN* metric 2.34, which is simply the average of the absolute trade imbalance across the ℓ most recent volume slices. In the current notation the quantity can be expressed

$$|\overline{TI}|_\ell = \ell^{-1} \sum_{i=1}^{\ell} |TI_i|. \tag{3.13}$$

Large values of $|\overline{TI}|_\ell$, are supposed to lead to less liquidity and more volatility. Amongst the three large cap stocks we investigated, this measure was not statistically significant to the regression models.

The co-movement in arriving market VM^j and net limit flow VL^j at one side of the LOB precisely defines the concept of resiliency. We define the correlation on

either side of the LOB as

$$\rho_t^j = \text{corr}(VM_{[t-s,t]}^j, VL_{[t-s,t]}^j), \tag{3.14}$$

where the empirical correlation is taken over a sliding time horizon of length $s = 2.5$ hours with time buckets of 30 seconds. We choose the shorter intervals for this indicator in an effort to obtain more observations for intra-day use. In theory, positive values should indicate strong resilience (e.g. positive drift for $v^j(t)$ in Figure 3.2). We find that ρ_t^j is moderately significant for two stocks BBBY and TEVA. This measure is discussed in more detail in the following section. Lastly, ρ^{Tox}, first described as a useful toxicity measure in [14], is defined

$$\rho^{Tox} = \text{corr}(\Delta P_i, O_i^M) \tag{3.15}$$

where the empirical correlation is computed between the signed volume of the ith market order and subsequent change in price is taken over a sliding window that includes the previous 200 order arrivals. This quantity measures amount of "toxic" flow, that is, order flow that predicts/causes a price change in the same direction.

Each variable in Table 3.3.2 was tested on each side of the LOB for all three stocks. Equation (3.16) includes the key variables in predicting the log-odds of scarce liquidity,

$$\text{logit}(\pi_k^A) = \theta_0 + \theta_1 VL_k^A + \theta_2 VL_k^B + \theta_3 \tau_k + \theta_4 PI_k^A + \theta_5 VL_1^A + \theta_6 \rho_t^A + \theta_7 TIMA_i \tag{3.16}$$

where logit$(\cdot)$ is the log-odds function and we make the usual assumption that

$$SL_k^A \sim Bin(1, \pi_k)$$

with probability π_k. We fit the identical model to SL_k^B except for switching the superscripts on variables PI_N, VL_1 and ρ_t from A to B.

As in the previous section, for each stock and side of the LOB (six models in all), limit order flow at each touch VL^j was of key importance to the model. Unlike the previous section, theoretical price impact PI_N^j on the same side of the book (PI_N^j if testing π^j) was of approximately equal significance as VL^j in all six models. The time of the volume slice τ, was also highly significant, especially for TEVA and BBBY, and with negative coefficient, indicating that scarce liquidity tends to occur less frequently later in the trading session. The trade imbalance moving average $TIMA$, was significant for all six models and affected results in the same manner in the previous model (3.11). The coefficient for $TIMA$ was negative for $j = A$ and positive for $j = B$ showing that scarce liquidity often occurs in the opposite direction of the longer term trend in market order flow, i.e. price tends to mean revert on average.

The three "toxicity" measures added little in terms of predictive power to the model. Both $|TI|_\ell$ and $\hat{\rho}_i$ were not significant for any of the models. In three of the six models the resilience measure ρ_t did prove significant at the 1% level with negative coefficient. So some persistent trend in the interaction between limit and

market flow does appear to exist but our metric ρ_t surely needs further improvement or calibration.

Across all six estimated models (3.16), the percentage of deviance explained ranged between $30\% - 35\%$. Percentage of deviance explained is analogous to R^2 and is one goodness of fit measure for generalized linear models.

To this point each model has included concurrent limit order flow VL^j, and is therefore of limited practical use. Upon removing the concurrent indicators VL^j to get a truly predictive model, the deviance explained by each model (3.16) drops to $7-10\%$. Without including the concurrent additions and cancellations summarized in VL^j, the model fit materially worsens. Further study into these processes on more robust data including multiple assets and trading venues is warranted. Further, defining a model-free dependence structure between the four asynchronous point processes VL^A, VL^B, VM^A and VM^B is an interesting open problem with practical implications.

3.4 Limit and Market Flow

We have shown that at the minutes-scale, the flow of incoming additions and cancellations rather than LOB depth is the primary driver of the price impact of market orders. Further the observed outcomes upon arrival of several market orders on one side varies widely, from strong resilience that leaves the mid-price unchanged, to rapid cancellations that lower the mid-price by several ticks. While predictors

VL^j_{k-1}, ρ^j_t and $\hat{\rho}_i$ attempt to capture persistent trends in limit order arrival processes the predictive power in equation (3.16) when excluding contemporaneous limit orders $\beta_1 = \beta_2 = 0$ is limited. This is not an altogether surprising result. Moving from the event-by-event timescale to the intermediate scale of course brings the effects of not only intra-day but also inter-day trends and broader market themes as well. However, even with the limited data and scope under consideration, we note some interesting features of limit flow and related quantities.

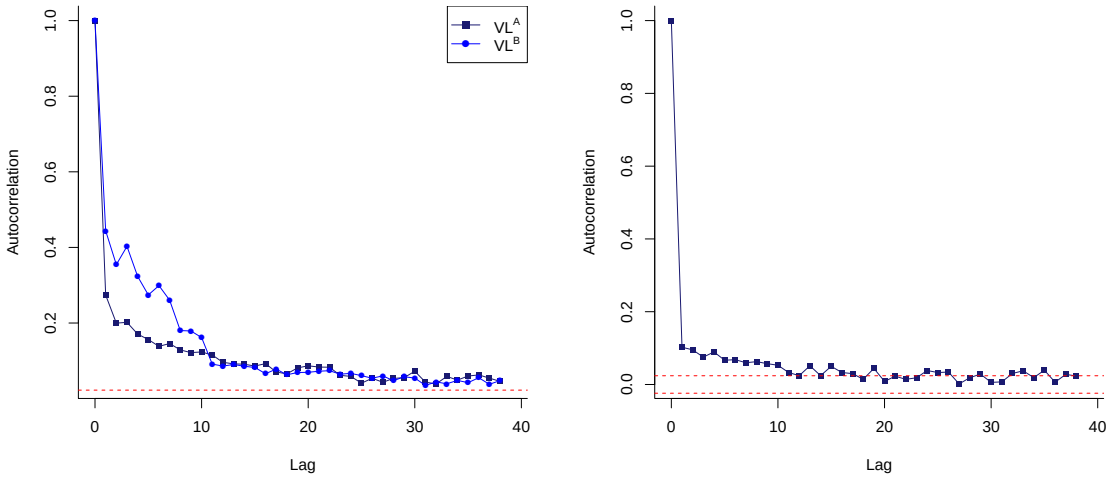

Figure 3.8: Autocorrelation plots at minutes-scale $V = 1\%ADV$, 103 trading days. Left: VL^A and VL^B for BBBY. Right: Occurrence of scarce liquidity for TEVA.

Figure 3.8 illustrates that some non-zero autocorrelation exists in the flow of limit orders to the touch, even at the intermediate time-scale. The left side shows autocorrelation for processes VL^B and VL^A for BBBY and the right shows autocorrelation in the occurrence of TEVA volume slices with scarce liquidity in either direction. While

the figures show some persistence in limit flow, analyzing limit arrivals apart from market order arrivals, makes little sense. Rather the key issues are how the two move together, how this relationship evolves over time and importantly, whether there exist persistent "regimes".

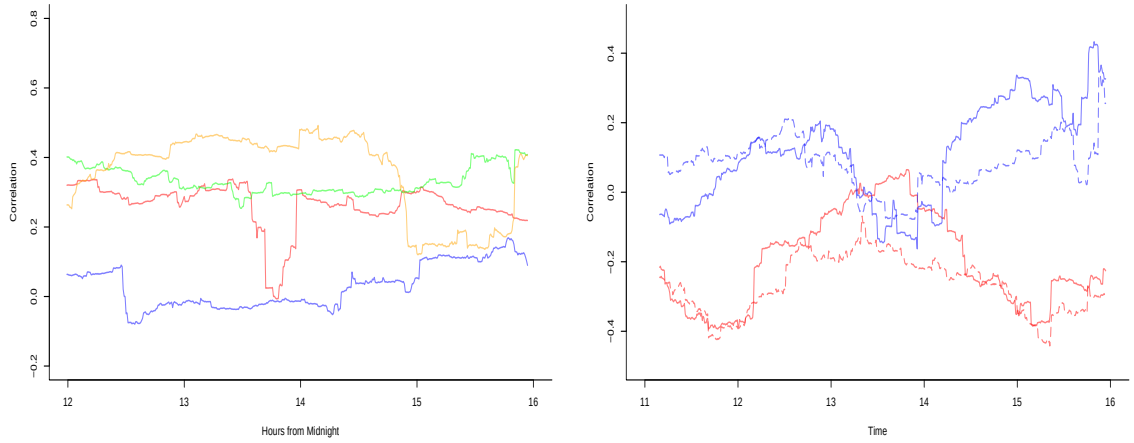

Figure 3.9: Smoothed contemporaneous correlation between VL and VM over the past 2.5 hours using 30-second buckets with colors indicating different trading days. Left: TEVA. Right: MSFT. Solid: bid-side. Dashed: ask-side

Figure 3.9 plots correlation between observed values of VL and VM (ρ^j) over a sliding time interval of 2.5 hours and with 30-second time buckets. The left side shows interaction between VL^B and VM^B for four separate trading days for TEVA while the right shows each side of the book on two separate trading days for MSFT. We observe distinct "regimes" in the correlation value on both an intra- and inter-day basis and interpret lower values as corresponding to increased cancellations and scarce liquidity, while positive values are characteristic of strong resilience. From

one point of view rapidly declining ρ^j could be an indication of the cumulative effect of persistent and aggressive buying/selling that leads market makers to adjust their reference prices to account for added inventory risk. Alternatively, ρ^j could highlight instances of "information leakage" and/or "front running". In any case, the existence of persistent states in interactions of order flows seem to deviate from the traditional model of resilience, in which a well-defined LOB shape is perturbed by arriving market orders and then "refilled" over time.

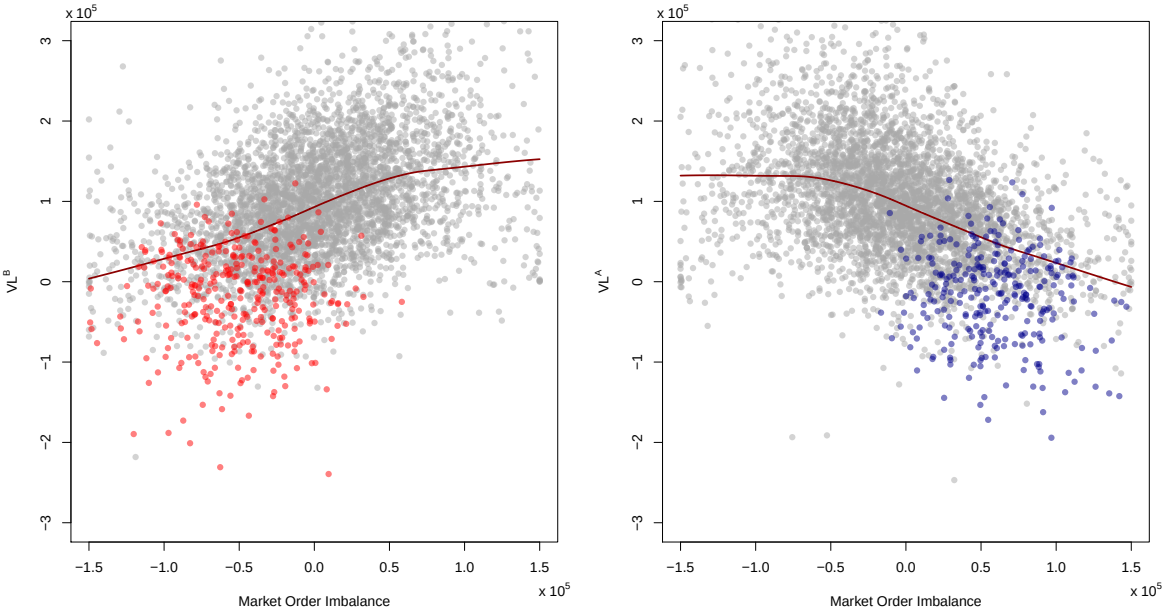

Figure 3.10: TI plotted against net limit order flow at the bet bid (left) and best ask (right). Red (blue) points indicate volume slices with price decrease (increase) of at least .05. MSFT, covering the first 50 trading days of 2011.

By construction, ρ^j ignores the activity taking place at the opposite side of the LOB. Alternatively we could simply analyze the joint movement between VL and TI.

Figure 3.10 plots the two quantities at the $V = 1\%$ADV volume slice scale. There is a clear positive (negative) relationship between TI and VL^B (VL^A), especially for MSFT, but which also holds for TEVA and BBBY. Given the very large number of heterogeneous players participating in trading activity, it is difficult to draw any conclusions as to the causality. Whether HFT market makers are predicting one-sided market flow or reacting to it, these periods persist long enough to appear very clearly at the minutes-scale. Table 3.4 shows that volume slices with low resiliency are anywhere from $2-5$ times as likely to occur when market order flow is one sided (at least 75% in buy/sell direction).

	MSFT		TEVA		BBBY	
	$VL^A < q_{.10}$	$VL^B < q_{.10}$	$VL^A < q_{.10}$	$VL^B < q_{.10}$	$VL^A < q_{.10}$	$VL^B < q_{.10}$
$TI > V/2$	35.4%	0.6%	23.0%	11.2%	27.7%	11.0%
$-V/2 < TI < V/2$	7.4%	8.6%	3.0%	3.9%	10.7%	14.1%
$TI < -V/2$	0.7%	35.7%	11.1%	23.0%	8.7%	27.1%

Table 3.4: Observed probability of low resilience by trade imbalance.

To summarize, we have discussed several key issues in the modeling and measurement of LOBs and price impact. At the minutes scale, order flows are the primary driver of the price formation process and not the measures of depth given by the LOB. Additionally, it is clear that market makers anticipate and/or react to one-

sided MOF which can lead to periods of rapid limit order cancellations or weak LOB resiliency. Along these lines, modeling LOB dynamics or price impact under a Markovian assumption ignores a significant component of the price formation process: the historical order flow. Further research into the multi-scale dynamics of LOBs is certainly warranted.

For the large tick stocks studied here, periods of scarce liquidity appear to induce higher *one-sided* execution costs, which contrasts with Easley et al. [24, 27] and the model in Chapter 2. In the following, we propose a new execution model that incorporates several of the findings from the empirical study.

Chapter 4

Optimal Execution with Expected Trade Imbalance: Liquid Stocks

4.1 Motivation

We now return to the question of how to best liquidate a large position of an asset while considering order flows and the informational costs associated with impacting the order flow process. In Chapter 2, execution costs stemming from the expected trade imbalance (Y_t) were largely symmetric (cost term Y_t^2) and abstract in the sense that it was not specified how exactly the trader's profit and loss were affected. We now assume that the asset in question is a liquid stock, similar to those investigated in Chapter 3. Important features are that the bid-ask spread is usually one tick and

that the posted volume is almost always plentiful at the first several levels of the LOB. In this setting, limit order flow appears to be closely associated with trade imbalance, lending support to the assumption that market maker strategies incorporate expectations about MOF. Further, the effects on liquidity and resulting execution costs appear to be 1-sided in nature (see Section 3.4). In the current chapter optimal executions strategies are calculated when costs in Y_t are asymmetric.

Remark 7. Modeling symmetric execution costs in the LOB setting is an interesting issue. ELO [27] claim that extreme values of trade imbalance indicate two sided effects on the liquidity provision process, which could be easier to detect in less liquid assets, such as *E-Mini* contracts, one of their primary applications. Even for large-tick assets like MSFT, we do find some evidence of longer (e.g. daily) regimes with elevated symmetric costs as measured by the daily average of $PI^A + PI^B$ or frequency of observed instances of scarce liquidity SL[1]. Determining the contributing factors to these regimes is difficult however. First, at the daily scale overall market trends and related assets (Futures, ETFs etc) play a large role. Second, order flows and execution activity for the individual asset must be aggregated across all trading venues to gain a complete picture of market activity. This is certainly a non-trivial issue. Lastly, periods of low liquidity are episodic in nature, occurring infrequently.

[1]Given the limited scope (both length of time and breadth of assets) of the data in Chapter 3 drawing conclusions at this scale proved very difficult.

In this Chapter we focus on one-sided, trade imbalance driven, execution costs that are more commonly characteristic of LOBs for liquid stocks.

Continuing in the Almgren-Chriss framework as in Chapter 2, the goal is now to tie costs in Y_t directly to the trader's wealth process. This can be accomplished in two ways: First, by introducing a liquidity cost in Y_t that affects the immediate execution price, and second, by allowing Y_t to impact the future price. It is important to remember that the process (Y_t) is the trade imbalance that "the market" (especially market makers) expects. As discussed in Section 2.4.1 the link between recently observed trade imbalance and expected trade imbalance is an open question, but the two are undoubtedly closely linked.

Liquidity Cost As in the model in Chapter 2, trading in the same direction as prevailing flow means that the execution trader is competing for liquidity. Therefore, in this case liquidity is potentially more costly that it would otherwise be if trading in a balanced market. Namely, it is more likely that inventory at the first level will be exhausted during the execution process by competing market orders or canceled limit orders. To capture this effect we introduce an additional instantaneous price impact term $h(\alpha_t, Y_t)$ which affects only the current transaction price. When the rate of execution is very high, we posit that $h(\cdot)$ is dominated by the existing quadratic cost term α^2 (2.6). This reflects the fact that we do not observe radical changes to

the entire LOB coinciding with trade imbalance. Therefore we let $h(\alpha_t, Y_t) = \gamma Y_t$ for some constant $\gamma > 0$, so that the resulting cost is linear in α_t and Y_t (see (4.4)).

Impact on Future Price To capture the effect of the expected trade imbalance on the fundamental price, we allow the drift term $f(Y_t)dt$ into the stock price dynamics. Again, we note that Y_t is expected trade imbalance and $f(Y_t)$, the flow driven mid-price drift, captures the fluctuations in the stock price caused by *limit* order additions and cancellations that result from expected trade imbalance. In contrast to the execution model by Cartea et al. [16] (see Remark 8), $f(Y_t)$ is not meant to model the impact of actual arriving MOF such as Equation (3.9) seeks to do, but rather the effect on price of LOB resiliency/fading that occur when MOF is one-sided. The functional form of $f(\cdot)$ is likely to be non-linear and possibly discontinuous as well. For example, it may be the case that $f(Y_t)$ resembles a step function, with $f(Y_t) \approx 0$ for most values of Y_t and stepping up materially when Y_t becomes extreme in either direction. From Section [Last in Ch3] we observed that rapid cancellation resulting in scarce liquidity and large mid-price change is much more likely when imbalance is high. In Section 4.2 we solve the problem assuming $f(\cdot)$ is linear and in Section 4.3 we address the case where $f(\cdot)$ is non-linear.

Remark 8. Cartea and Jaimungal [16] present an Almgren-Chriss extension where the expected future net order flow serves as a linear drift term in the mid-price. It is assumed that future order flow will impact the price in the same way as is observed in

the price trend analysis in Section 3.3.1 for example. However, in reality, the current price (and market makers) typically take into account the expected trade imbalance. As the model [16] is presented, the implicit assumption is either 1) that the expected net order flow process is information only known to the execution trader, or 2) that current prices (and market makers) do not consider expected net order flow process. In contrast, the approach presented in this Chapter models the effect of the expected trade imbalance on market maker behavior, and subsequently the mid-price (through resiliency/fading of the LOB). This is a subtle but important difference and motivates the non-linear effect of Y_t in Section 4.3.

In contrast to Chapter 2 we now focus exclusively on the fixed-horizon execution problem. It is possible to apply similar steps as in Chapter 2 to investigate optimizing the execution horizon but that is not the primary objective here. Additionally, we assume that information leakage is modeled deterministically $\phi(\alpha_t) = \phi(t)$ as in Section 2.2.1. As will be shown, even with deterministic information leakage, we still obtain dynamic strategies that adapt to fluctuations in the expected trade imbalance.

4.2 Optimal Execution with Assymetric Costs

As in Chapter 2 the trader needs to liquidate a position of size $x_0 = x$. We assume a continuous-time setup, with trading taking place continuously and via infinitesimal

amounts. The trader controls the speed of liquidation $\alpha_t = (\alpha_t)_{\{0 \leq t \leq T\}}$ and inventory x_t again follows the dynamics

$$dx_t = -\alpha_t dt. \tag{4.1}$$

The class of admissible strategies $\mathcal{A}(T, x)$ is defined as in Chapter 2, as is the expected trade imbalance process (Y_t) which follows the dynamics

$$dY_t = -\beta Y_t dt - \phi(t)dt + \sigma_Y dW_t^{(Y)}, \tag{4.2}$$

where ϕ, which controls information leakage, is modelled as a deterministic function of time t. Recall that the stationary variance $\sigma_Y^2/(2\beta)$ is chosen such that $\mathbb{P}(Y_t \notin [-1, 1]) \approx 0$. The mid-price follows the SDE

$$dP_t = f(Y_t)dt + \sigma_P dW_t^{(P)}, \tag{4.3}$$

where $f(Y_t)$ captures the fluctuations in the mid-price caused by *limit* additions and cancellations resulting from the expected trade imbalance. In the current section it is assumed that $f(Y_t) = \theta Y_t$ for some nonnegative constant θ. Also note that the noise terms in (4.2) and (4.3) could be correlated but this does not affect the execution strategy or costs.

To account for the limited amount of liquidity resting at the best bid we also define the execution price $\check{P}_t$, which depends on both α_t and Y_t. Trading an order of

size $\alpha_t dt$ obtains the execution price

$$\check{P}_t = P_t - \xi\alpha_t + h(Y_t), \tag{4.4}$$

where $h(\cdot)$ is a decreasing function that captures the effect of trading with or against the prevailing market order flow and ξ is a nonnegative constant . Trading with the flow, the trader competes for liquidity with other participants and may on balance expect to exhaust the first queue more often during execution. For the remainder, we assume the cost is linear in Y_t so that $h(Y_t) = \gamma Y_t$. Lastly, the investor's wealth process, Q_t follows the dynamics

$$dQ_t := \alpha_t \check{P}_t dt. \tag{4.5}$$

The performance criteria and value function for the continuous-time execution problem are written

$$u(t, P, q, x, Y) = \sup_{(\alpha_t)\in\mathcal{A}(T,x)} \mathbb{E}_{P,x,Y}\left[Q_T - \int_0^T \lambda(x_s^\alpha)\, ds\right]. \tag{4.6}$$

Equation (4.6), the analog (2.10) in Chapter 2, no longer has a quadratic cost term in Y_t. Also, rather than minimizing execution costs, the problem is framed such that the objective is to maximize revenues Q_t subject to the running inventory penalty $\int_0^T \lambda(x_s^\alpha)\, ds$. The functional form of $\lambda(x_s)$ dictates the "benchmark" strategy as described in Section 2.2.1. Here, we focus on $\lambda(x_s) = cx_s^2$ which gives as a benchmark

[1]Recall that $\xi = 1$ in Chapter 2.

the Almgren-Chriss strategy when $c > 0$ and the VWAP strategy when $c = 0$. As before, the terminal condition

$$\lim_{t \to T} u(t, P, q, x, Y) = \begin{cases} 0 & \text{if } x = 0 \\ -\infty & \text{if } x \neq 0. \end{cases} \tag{4.7}$$

guarantees that the liquidation process is completed by the terminal time. As a matter of notation, we now fix the terminal time T, and (4.7) and (4.8) are parameterized in terms of t rather than in terms of remaining time (called T in Section 2.2.2). Then, the HJB PDE for $u(t, P, q, x, Y)$ is

$$-u_t = \frac{1}{2}\sigma_Y^2 u_{YY} - \beta Y u_Y - \phi(t) u_Y + \theta Y u_P - \lambda(x) + \sup_{\alpha} \left\{ \alpha(P - \xi\alpha + \gamma Y) u_q - \alpha u_x \right\}, \tag{4.8}$$

with $f(Y) = \theta Y$ and $u(T, P, q, x, Y) = -\infty$ unless $x = 0$. Also, as in (2.22), α is unconstrained and allowed to become negative, though this is an unlikely occurrence for reasonable parameter choices.

Proposition 4.2.1. *The solution of* (4.8) *has the form*

$$u(t, P, q, x, Y) = q + xP + x^2 A(t) + Y^2 B(t) + xY C(t) + xD(t) + Y E(t) + F(t), \tag{4.9}$$

where A, B, C, D, E, F solve the system of ordinary differential equations (ODEs)

$$\begin{cases} A'(t) &= \frac{A^2}{\xi} - c \\ B'(t) &= -2\beta B + \frac{1}{4\xi}(C-\gamma)^2 \\ C'(t) &= \theta - \beta C + \frac{A}{\xi}(C-\gamma) \\ D'(t) &= -\phi(t)C + \frac{AD}{\xi} \\ E'(t) &= -2\phi(t)B - \beta E + \frac{D}{2\xi}(C-\gamma) \\ F'(t) &= \sigma_Y^2 B - \phi(t)E + \frac{D^2}{4\xi}, \end{cases} \tag{4.10}$$

and we have the following terminal conditions

$$\begin{cases} \lim_{t\to T} A(t) = -\infty \\ B(T) = D(T) = E(T) = F(T) = 0 \\ C(T) = \gamma \end{cases} \tag{4.11}$$

For the case $c = 0$ (VWAP benchmark strategy) we choose the corresponding constant information leakage term $\phi(t) = \phi^{VWAP}$ and have the following closed form solution

$$
\begin{aligned}
A(t) &= -\frac{\xi}{T-t} \\
B(t) &= \frac{-\chi^+(t)e^{-2\beta(T-t)}}{8\beta^4\xi(T-t)}\left(\gamma\beta-\theta\right)^2\left(2\chi^+(t)+\beta(T-t)(\chi^+(t)+2)\right) \\
C(t) &= \frac{1}{\beta^2(T-t)}\left(\theta(\beta(T-t)+\chi^-(t))+\beta\gamma\chi^-(t)\right) \\
D(t) &= \frac{\phi^{VWAP}}{2\beta^3(T-t)}\left(2\theta\chi^-(t)-\beta^2(T-t)(2\gamma+\theta(T-t))+2\beta(\theta(T-t)-\gamma\chi^-(t))\right) \\
E(t) &= \frac{\phi^{VWAP}e^{-2\beta(T-t)}}{4\xi\beta^5(T-t)}(\beta\gamma-\theta)(2\chi^+(t)+\beta(T-t)(\chi^+(t)+2))(\theta\chi^+(t) \\
&\quad -\beta(\gamma\chi^+(t)+\theta(T-t)e^{\beta(T-t)})) \\
F(t) &= \int\left(\sigma_Y^2B(t)-\phi^{VWAP}E(t)+\frac{D(t)^2}{4\xi}\right)dt,
\end{aligned}
\tag{4.12}
$$

with

$$
\begin{aligned}
\chi^+(t) &= 1-e^{\beta(T-t)} \\
\chi^-(t) &= 1-e^{-\beta(T-t)}.
\end{aligned}
\tag{4.13}
$$

Proof. See Section 4.4. □

Remark 9. The system (4.9) consists of a Riccati ODE, $A(t)$ and several first order ODEs that can be solved directly in a sequential manner. In the case where $c > 0$ the benchmark strategy follows the Almren-Chriss curve (2.14) and certain integrals in (4.12) can not be easily expressed in explicit form. However lengthy expressions can be obtained by applying a computer algebra program like Mathematica.

Proposition 4.2.1 yields a candidate solution to the control problem in (4.6). The first two terms in (4.9) are the value of the proceeds from all sales to date plus the value of the remaining inventory, "marked-to-market" at the current price. The remaining

terms, which are independent of P_t, represent the value of optimally liquidating for the remainder of the trading horizon. The optimal strategy is given by the following theorem.

Theorem 4.2.2. ***Verification.*** *The candidate value function given in Proposition 4.2.1 is the solution to the optimal control problem* (4.6)*. The corresponding optimal rate of liquidation when* $c = 0$ *is*

$$\alpha_t^* = \frac{x_t}{(T-t)} + \frac{Y_t}{2\xi}\left(\gamma - \frac{1}{\beta^2(T-t)}\left(\theta(\beta(T-t) + \chi^-(t)) + \beta\gamma\chi^-(t)\right)\right) - \frac{\phi^{VWAP}}{2\beta^3(T-t)}\left(2\theta\chi^-(t) - \beta^2(T-t)(2\gamma + \theta(T-t)) + 2\beta(\theta(T-t) - \gamma\chi^-(t))\right). \tag{4.14}$$

Proof. See Section 4.4. □

The first term of the optimal trading strategy (4.14) corresponds to the VWAP strategy and is independent of Y_t. The functional form of this term depends on the choice of inventory risk penalty $\lambda(x_t)$, with alternatives to the VWAP strategy given by $\mathcal{I}$ in Section 2.2.1. How α_t^* depends on Y_t is dictated by the values chosen for constants γ and θ. When γ is small relative to θ, then the additional liquidity cost of trading with the prevailing market flow is outweighed by the dropping mid-price. In this case the trading rate is decreasing in Y_t. On the other hand, the liquidity cost in Y_t dominates the influence on mid-price when γ is large compared to θ. The third term, which adjusts the trading rate based on the deterministic information leakage $\phi(t)$, is positive and approaches 0 as $t \to T$.

Corollary 4.2.3. *For all values of T, the optimal rate of trading is non-decreasing in Y_t if the following condition holds*

$$\gamma \geq \frac{\theta}{\beta}$$

and decreasing otherwise.

Proof. See Section 4.4. □

Corollary 4.2.3 shows that when both $f(\cdot)$ and $h(\cdot)$ are linear, the effects on the strategy of each function partially offset each other and the resulting strategy is linear in Y_t. The amount by which the market's expected trade imbalance Y_t affects the trading rate depends on the time remaining $T - t$. As $t \to T$, the trading rate stabilizes as the first term in (4.14) dominates the second. In the next section we investigate the more interesting set-up where expected order flow has a non-linear effect on the price P_t.

4.3 Optimal Execution with Non-Linear Flow Driven Mid-Price Drift

The motivation for allowing $f(Y_t)$ to be a non-linear and in particular a convex function is quite intuitive. During relatively calm market conditions, we might observe balanced to moderate values of expected trade imbalance. Under these conditions,

liquidity providers would assume that adverse selection costs are low, or put another way, that most MO arrivals are sent by non-informed traders. Increasingly one sided MOF would be ignored as not indicative of new information. On the other hand, if one-sided MOF persists or MOF is viewed as carrying new information about future prices, expected trade imbalance would spike, causing liquidity providers to quickly adjust their orders. Strong autocorrelation in limit order signs also suggest potential "herding" amongst liquidity providers which would support the non-liner effect of Y_t on the mid-price P_t.

As described in the previous section, when both h and f are linear, the resulting strategy is also linear in Y with the direction and sensitivity determined by the parameter choices. When $f(Y_t)$ is convex, the liquidity cost $h(Y_t)$ drives the trading strategy in a balanced market, and the effect of $f(Y_t)$ on the mid-price and trading rate are negligible. In an unbalanced market, the mid-price drift $f(Y_t)$ heavily influences the trading rate. Choices for the functional form of f that maintain the tractability from Section 4.2 are limited. However, it turns out that $f(Y_t) = \theta Y^3$ can be solved in closed-form when information leakage is modelled deterministically $\phi(\alpha_t) = \phi_t$.

We now have the mid-price drift $f(Y_t) = \theta Y^3$ and the corresponding HJB PDE and conditions from the previous section 4.2. The following proposition gives the form of the solution and explicit optimal trading rate when $\phi_t = 0$. Allowing for

non-zero ϕ_t requires additional higher order terms in Y but does not change the form of (4.15) or the spirit of the optimal trading strategy.

Proposition 4.3.1. *Under the assumptions that $f(Y_t) = \theta Y^3$ and $\phi_t = 0$, the solution of (4.8) has the form*

$$u(t,P,q,x,Y) = q + xP + x^2A(t) + Y^2B(t) + xYC(t) + xY^3D(t) + Y^6E(t) + Y^4F(t) + G(t), \quad (4.15)$$

where A, B, C, D, E, F, G solve the system of ordinary differential equations (ODEs)

$$\begin{cases} A'(t) &= \frac{A^2}{\xi} - c \\ B'(t) &= -2\beta B + \frac{1}{4\xi}(C-\gamma)^2 + 12\sigma_Y^2 F \\ C'(t) &= \sigma_Y^2 D - \beta C + \frac{A}{\xi}(C-\gamma) \\ D'(t) &= \theta - 3\beta D + \frac{AD}{\xi} \\ E'(t) &= \frac{D^2}{4\xi} - 6\beta E \\ F'(t) &= 30\sigma_Y^2 E - 4\beta F + \frac{D}{2\xi}(C-\gamma) \\ G'(t) &= 2\beta\sigma_Y^2. \end{cases} \quad (4.16)$$

and we have the following terminal conditions

$$\begin{cases} \lim_{t\to T} A(t) = -\infty \\ B(T) = D(T) = E(T) = F(T) = G(t) = 0 \\ C(T) = \gamma \end{cases} \quad (4.17)$$

For the case $c = 0$ (VWAP benchmark strategy), the optimal rate of liquidation is

$$\begin{aligned}\alpha_t^* &= \frac{1}{2\xi}\left(\gamma Y_t - 2x_t^2 A(t) - Y_t C(t) - Y_t^3 D(t)\right) \\ &= \frac{x_t}{(T-t)} + \frac{\theta Y_t^3}{9\beta^2(T-t)}\left(e^{-3\beta(T-t)} - 1 + 3\beta(T-t)\right) \\ &+ \frac{Y_t}{18\beta(T-t)}\left(18\gamma\beta^2(1-e^{-\beta(T-t)}) + \theta\sigma_Y^2(9e^{-\beta(T-t)} - e^{-3\beta(T-t)} + 6\beta(T-t) - 8)\right)\end{aligned} \tag{4.18}$$

The proof of Proposition (4.3.1) is identical to Proposition (4.2.1). Functions $C(t)$ and $D(t)$ are positive and approach γ and 0 respectively as $t \to T$, so that strategy's dependence on Y goes to 0 as the horizon approaches. In the following we discuss features and financial interpretation of the optimal liquidation strategies (4.14) and (4.18). To avoid confusion, we denote the linear case ($f(Y_t) = \theta Y_t$) as α_t^ℓ and the convex case ($f(Y_t) = \theta Y_t^3$) as α_t^c. Figure 4.1 illustrates each trading strategy with simulated paths of (Y_t). In the simulation, parameters are set to

$$x_0 = 3 \quad , \quad T = 3 \quad , \quad Y_0 = 0 \quad , \quad \sigma_Y^2 = .1 \quad , \quad \beta = .25$$

$$\xi = .2 \quad , \quad c = .08 \quad , \quad \gamma = .6 \quad , \quad \theta = .3.$$

Each side of Figure 4.1 shows 500 simulations of the optimal strategy (left: (α_t^ℓ) and right: (α_t^c)) along with the baseline Almgren-Chriss strategy (bold solid line) and colored bands corresponding to the 75%/25%, 90%/10% and 99%/.01% quantile ranges. It is clear in both plots that the rate of trading tends to fluctuate with Y_t in the early stages with of trading before stabilizing as the terminal time T nears.

Also evident is that most often, α_t^c paths are much more tightly condensed around the baseline strategy compared to α_t^ℓ. At the same time, α_t^c also experiences more extreme fluctuations in the trading rate in a very small number of paths. This is a straight-forward consequence of extreme values of expected trade imbalance having a very large impact on trading due to the convexity of $f(\cdot)$.

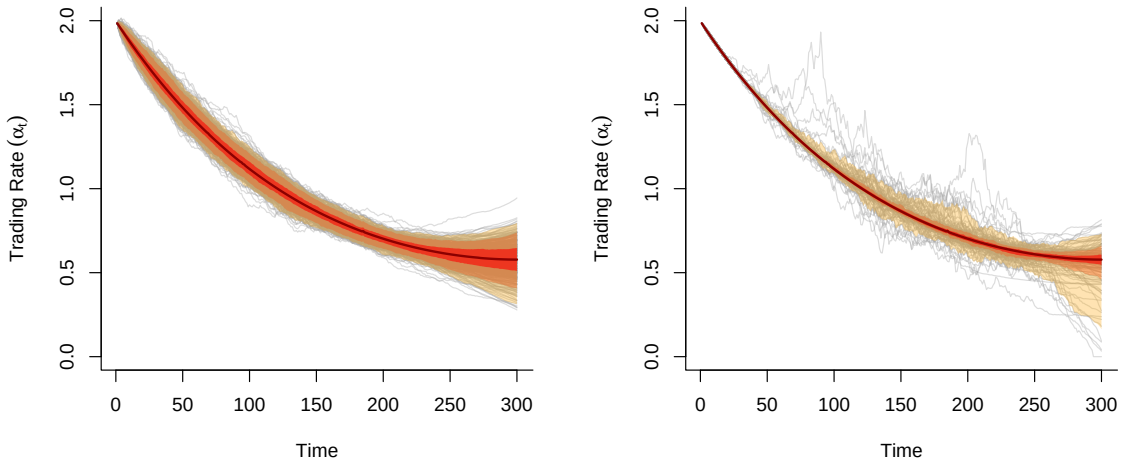

Figure 4.1: Optimal liquidation strategies (α_t^ℓ) (left) and (α_t^c) (right) for 500 simulated paths of expected trade imbalance (Y_t). Baseline Almgren-Chriss strategy is shown (bold solid line) along with colored bands corresponding to the 75%/25%, 90%/10% and 99%/1% quantile ranges.

While the cubic formulation of f may not be the most realistic representation of price dynamics with expected trade imbalance, there are some interesting financial interpretations of the resulting strategy worth noting. First, the current price of the asset most often reflects the available information in the market, including the expected trade imbalance. From [12] and as previously discussed in Remark 8, this

must generally be true given the presence of autocorrelated MOF and the common observation/assumption that the mid-price follows a random walk. Therefore most of the time, and for most values of Y_t, its impact on future asset prices should be minimal. As shown in Chapter 3, scarce liquidity resulting in a noticeable mid-price move is a relatively rare event, occurring perhaps a few times a day. Thus, a strategy considering expected trade imbalance should most often closely follow the benchmark strategy but would occasionally deviate from this benchmark dramatically under certain market conditions.

Second, in a similar vein, during tumultuous market conditions that often go hand-in-hand with extreme values of observed and expected trade imbalance, market participants often accelerate trading in an effort to exit the market and reduce risk. The strategy α_t^c captures this behavior, and explains how the seemingly "irrational" trading could be in fact be optimal in certain cases. Of course with the assumption of a fixed terminal time T, accelerated trading early in the execution process means that trading must slow down later. As shown in Chapter 2, this issue can be resolved by endogening the execution horizon, though it must be handled numerically.

4.4 Proofs

Proof of Proposition 4.2.1

Proof. We first re-write (4.9) in the form

$$u(t, P, q, xY) = q + xP + \hat{u}(t, x, Y). \tag{4.19}$$

Substituting into (4.8) we find the optimal trading rate in feedback control form

$$\alpha^* = \frac{\gamma Y - \hat{u}_x}{2\xi}.$$

. Upon substitution of α^* in the HJB PDE (4.8) we have

$$-\hat{u}_t = \frac{1}{2}\sigma_Y^2 \hat{u}_{YY} - \beta Y \hat{u}_Y - \phi(t)\hat{u}_Y + \theta Y \hat{u}_p - \lambda(x) + \frac{(\gamma Y - \hat{u}_x)^2}{4\xi}. \tag{4.20}$$

Applying the ansatz (4.19) and grouping like terms as in (2.5) yields (4.10). To obtain the expressions for the ODEs 4.10 one begins by solving for $A(t)$ which is a simple example of a Riccati ODE. Substituting the expression for $A(t)$ one can then solve for $C(t)$ and so forth. When $c > 0$ the expression for $A(t)$ involves hyperbolic functions that make solving the following ODEs quite messy. □

Proof of Theorem 4.2.2

Proof. Since $q + xp + x^2A(t) + y^2B(t) + xyC(t) + xD(t) + yE(t) + F(t)$ is a classical solution, standard arguments imply that it suffices to show whether the feedback control is an admissible strategy $\alpha^* \in \mathcal{A}(T, x)$. Namely it must be shown that

$\mathbb{E}\int_0^T (\alpha_t^*)^2\,dt < \infty$. We provide details for the case $\phi_t = 0$. Given $dx_t = -\alpha_t dt$, the initial position x_0 and terminal condition $x_T = 0$, we solve the ODE (4.14) and have

$$x_t^* = \frac{T-t}{T}\left(x_0 - \int_0^t Y_s\left(\frac{T(\gamma - C(s))}{2\xi(T-s)}\right)ds\right).$$

Then it follows that

$$\begin{aligned}
|x_t^*| &\leq \frac{x_0(T-t)}{T} + \left(\int_0^t \frac{T(\gamma - C(s))}{2\xi(T-s)}\,ds\right)\sup_{0\leq s\leq T}|Y_s| \\
&= \frac{x_0(T-t)}{T} + \frac{e^{-\beta T}(\beta\gamma-\theta)(T-t)}{2\beta^2\xi T}\Bigg\{(1-e^{\beta T}) + \beta T e^{\beta T}\Bigg(\log\left(\frac{T-t}{T}\right) \\
&+ \int_0^t \frac{e^{-\beta(T-s)}}{T-s}\,ds\Bigg)\Bigg\}\sup_{0\leq s\leq T}|Y_s| \\
&\leq x_0 + \frac{e^{-\beta T}(\beta\gamma-\theta)}{2\beta^2\xi}\Bigg\{(1-e^{\beta T}) + \beta T e^{\beta T-1} + \beta e^{\beta T}\Bigg\}\sup_{0\leq s\leq T}|Y_s|.
\end{aligned}$$

Because Y_t is Gaussian with known moments, it follows that $\mathbb{E}\int_0^T |x_t Y_t|\,dt < \infty$, and further that $\sup_{t\leq T}|x_t| \in L^2$. Therefore we have that $\mathbb{E}\int_0^T (\alpha_t^*)^2\,dt < \infty$.

□

Proof of Corollary 4.2.3

Proof. It suffices to simply show for what values of γ the Y-coefficient in 4.14 is positive, giving

$$\gamma \geq \frac{\theta(\beta T + e^{-\beta T} - 1)}{\beta^2 T - \beta(e^{-\beta T} - 1)} = \frac{\theta}{\beta}.$$

□

Chapter 5

Conclusion

This dissertation investigated the well-known problem of optimally liquidating a large position in an LOB-driven market. The key element introduced in our execution model is the stochastic factor (Y_t), which represents the expected trade imbalance; that is, the imbalance between buy- and sell-initiated market orders. The expected trade imbalance process (Y_t) is closely related to the recent order flow history and allows our model to take into account the informational cost of trading in addition to the usual market microstructure impact. Further, incorporating Y_t leads to the consideration of the current market state and specifically whether one's orders lean with or against the prevailing order flow, key components often ignored by execution models in the literature.

In Chapter 2 we extended recent research in the area of order flow and market toxicity to a dynamic setting. Compared to the prevailing Almgren-Chriss framework, our model incorporates the direction of previaling MOF and the execution trader's impact on the expected trade imbalance. Further, we allow the trade horizon T to be random, which provides more flexibility allowing the trading strategy to accelerate and exit the market earlier than anticipated under adverse conditions. The model presented in Chapter 2 incorporates two-sided execution costs when executing in a unbalanced market, consistent the model put forth by Easley et al. [27].

The primary motivation for the empirical analysis presented in Chapter 3 was to validate the main assumption made in Easley et al. [27] and the model in Chapter 2: that a market with one-sided MOF leads to higher execution costs for the trader looking to liquidate his position. To that end, we estimated several regression models in an effort to separate price trend from liquidity and to pinpoint key variables in the price formation process. At the minutes scale, we found that order flows are the primary driver of the price formation process. Further, it is clear that market makers anticipate and/or react to one-sided MOF which can lead to periods of rapid limit order cancellations or weak LOB resiliency, i.e. higher execution costs. However, one notable conclusion was that for the assets we investigated, these costs were one-sided in nature.

In Chapter 4 we proposed a new model in which the costs associated with extreme values of expected trade imbalance are asymmetric. The resulting closed-form strategies are dynamic, adapting to changing state of expected trade imbalance Y_t, and have an interesting interpretation related to market stability. Namely, when the flow driven mid-price drift $f(Y_t)$ is allowed to be convex in Y_t, tumultuous market conditions that often go hand-in-hand with extreme values of observed and expected trade imbalance can lead dramatic moves in the selling rate. On the other hand, moderately balanced buy/sell markets result in trading that remains very close to the benchmark strategy.

Several interesting problems on this topic remain. First, LOB analysis is typically done at a single time scale, usually at the very short end ($\ll 1$ second), but there are clearly trends at the minutes-scale and longer. A multi-scale approach to LOB analysis appears warranted. Second, while the model presented in Chapter 4 captures the LOB resiliency/fading that occur during one-sided market order flow, it does not fully consider how a surprise reversal in the MOF trend often has a greater impact on the mid-price. This fact was highlighted in Sections 3.3.2 and 3.3.3 by the negative coefficient for the trade imbalance moving average variable $TIMA$ in the regression models. The precise relationship between observed and expected trade imbalance has large ramifications for traders and ought to be explored further. Refining mid-price dynamics to account for these nuances could certainly improve execution models.

Bibliography

[1] Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010.

[2] Aurélien Alfonsi and Alexander Schied. Optimal trade execution and absence of price manipulations in limit order book models. *SIAM J. Financial Math.*, 1:490–522, 2010.

[3] Aurélien Alfonsi, Alexander Schied, and Alla Slynko. Order book resilience, price manipulation, and the positive portfolio problem. *SIAM Journal on Financial Mathematics*, 3(1):511–533, 2012.

[4] Robert Almgren. Optimal trading with stochastic liquidity and volatility. *SIAM Journal on Financial Mathematics*, 3(1):163–181, 2012.

[5] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[6] Robert Almgren and Julian Lorenz. Adaptive arrival price. *Trading*, 2007(1):59–66, 2007.

[7] Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005.

[8] Robert F Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10(1):1–18, 2003.

[9] Torben G Andersen and Oleg Bondarenko. VPIN and the flash crash. *Journal of Financial Markets*, 17:1–46, 2012.

[10] Torben G Andersen and Oleg Bondarenko. Reflecting on the vpin dispute. *Journal of Financial Markets*, 17:53–64, 2014.

[11] Kyle Bechler and Mike Ludkovski. Optimal execution with dynamic order flow imbalance. *arXiv preprint arXiv:1409.2618*, 2014.

[12] Jean-Philippe Bouchaud. Price impact. In *Encyclopedia of Quantitative Finance*. Wiley Online Library, 2010.

[13] Jean-Philippe Bouchaud, J Doyne Farmer, and Fabrizio Lillo. How markets slowly digest changes in supply and demand. In Thorsten Hens IV and Klaus Reiner Schenk-Hoppé, editors, *Handbook of Financial Markets: Dynamics and Evolution: Dynamics and Evolution*, pages 57–160. Elsevier, 2009.

[14] Rene Carmona and Kevin Webster. The self-financing equation in high frequency markets. Technical report, 2013. arXiv preprint arXiv:1312.2302.

[15] Álvaro Cartea and Sebastian Jaimungal. Optimal execution with limit and market orders. *SIAM Journal on Financial Mathematics*, forthcoming, 2014.

[16] Álvaro Cartea and Sebastian Jaimungal. Incorporating order-flow into optimal execution. *Available at SSRN 2557457*, 2015.

[17] Álvaro Cartea, Sebastian Jaimungal, and Jason Ricci. Buy low sell high: a high frequency trading perspective. Technical report, Availale on SSRN, 2011.

[18] Rama Cont and Adrien De Larrard. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1):1–25, 2013.

[19] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2013.

[20] Khalil Dayri and Mathieu Rosenbaum. Large tick assets: implicit spread and optimal tick size. *arXiv preprint arXiv:1207.6325*, 2012.

[21] Ryan Francis Donnelly. *Ambiguity Aversion in Algorithmic and High Frequency Trading*. PhD thesis, University of Toronto, 2014.

[22] Ryan Francis Donnelly. Ambiguity aversion in algorithmic and high frequency trading. *Available at SSRN 2527808*, 2014.

[23] David Easley, Marcos M López de Prado, and Maureen O'Hara. The microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading. *Journal of Portfolio Management*, 37(2):118–128, 2011.

[24] David Easley, Marcos M López de Prado, and Maureen O'Hara. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5):1457–1493, 2012.

[25] David Easley, Nicholas M Kiefer, Maureen O'hara, and Joseph B Paperman. Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, 51(4):1405–1436, 1996.

[26] David Easley, Marcos M López de Prado, and Maureen O'Hara. VPIN and the flash crash: A rejoinder. *Journal of Financial Markets*, 17:47–52, 2014.

[27] David Easley, Marcos Lopez Prado, and Maureen O'Hara. Optimal execution horizon. *Mathematical Finance*, 2013.

[28] Zoltán Eisler, Jean-Philippe Bouchaud, and Julien Kockelkoren. The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance*, 12(9):1395–1419, 2012.

[29] Thierry Foucault. Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial markets*, 2(2):99–134, 1999.

[30] Jim Gatheral. No-dynamic-arbitrage and market impact. *Quantitative finance*, 10(7):749–759, 2010.

[31] Jim Gatheral and Alexander Schied. Optimal trade execution under geometric brownian motion in the Almgren and Chriss framework. *International Journal of Theoretical and Applied Finance*, 14(03):353–368, 2011.

[32] Jim Gatheral, Alexander Schied, and Alla Slynko. Transient linear price impact and fredholm integral equations. *Mathematical Finance*, 22(3):445–474, 2012.

[33] Richard C Grinold and Ronald N Kahn. *Active portfolio management*. McGraw Hill New York, NY, 2000.

[34] Fabien Guilbaud and Huyen Pham. Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 13(1):79–94, 2013.

[35] Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data: The queue-reactive model. *Journal of the American Statistical Association*, (just-accepted):00–00, 2014.

[36] Thibault Jaisson. Market impact as anticipation of the order flow imbalance. *Quantitative Finance*, (ahead-of-print):1–13, 2015.

[37] Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: The impact of high frequency trading on an electronic market. Technical report, Available on SSRN, 2011.

[38] Andrei A Kirilenko and Gui Lamacie. Latency and asset prices. *Available at SSRN 2546567*, 2015.

[39] Albert S Kyle. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335, 1985.

[40] Matthieu Lasnier and Charles-Albert Lehalle. Quant Note: Saw-tooth Effect on the US Market. Technical report, Credit Agricole, 08 2012.

[41] Fabrizio Lillo, J Doyne Farmer, and Rosario N Mantegna. Econophysics: Master curve for price-impact function. *Nature*, 421(6919):129–130, 2003.

[42] Alexander Lipton, Umberto Pesavento, and Michael G Sotiropoulos. Trade arrival dynamics and quote imbalance in a limit order book. *arXiv preprint arXiv:1312.0514*, 2013.

[43] Julian Lorenz and Robert Almgren. Mean–variance optimal adaptive execution. *Applied Mathematical Finance*, 18(5):395–422, 2011.

[44] Ciamac C Moallemi, Mehmet Saglam, and Michael Sotiropoulos. Short-term predictability and price impact. *Available at SSRN 2463952*, 2014.

[45] Esteban Moro, Javier Vicente, Luis G Moyano, Austin Gerig, J Doyne Farmer, Gabriella Vaglica, Fabrizio Lillo, and Rosario N Mantegna. Market impact and trading profile of hidden orders in stock markets. *Physical Review E*, 80(6):066102, 2009.

[46] Anna A Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013.

[47] Christine A Parlour. Price dynamics in limit order markets. *Review of Financial Studies*, 11(4):789–816, 1998.

[48] James A Primbs. Portfolio optimization applications of stochastic receding horizon control. In *American Control Conference, 2007. ACC'07*, pages 1811–1816. IEEE, 2007.

[49] Bence Toth, Imon Palit, Fabrizio Lillo, and J Doyne Farmer. Why is order flow so persistent? *arXiv preprint arXiv:1108.1632*, 2011.

[50] S.N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2006.